Universidade Federal de Pelotas

Programa de Pós-Graduação em Biotecnologia

Tese

Estratégias de bioinformática no estudo da patogenômica de bactérias do gênero *Xanthomonas*

Frederico Schmitt Kremer

Pelotas, 2019

Frederico Schmitt Kremer

Estratégias de bioinformática no estudo da patogenômica de bactérias do gênero *Xanthomonas*

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutor em Ciências (área do Conhecimento: Biotecnologia).

Orientador: Dr. Luciano da Silva Pinto
Comissão de Acompanhamento: Dr. Luciano Carlos da Maia e Dr. Moisés João Zotti

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

K92e Kremer, Frederico Schmitt

Estratégias de bioinformática no estudo da patogenômica de bactérias do gênero xanthomonas / Frederico Schmitt Kremer ; Luciano da Silva Pinto, orientador. — Pelotas, 2019.

244 f. : il.

Tese (Doutorado) — Programa de Pós-Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2019.

1. Genômica. 2. Sequenciamento de nova geração. 3. Pangenômica. 4. Fitopatógeno. 5. Tales. I. Pinto, Luciano da Silva, orient. II. Título.

CDD : 575.21

Elaborada por Ubirajara Buddin Cruz CRB: 10/901

Banca Examinadora

Prof. Dr. Luciano da Silva Pinto (UFPel, CDTec)
Prof. Dr. Odir Antônio Dellagostin (UFPel, CDTec)
Profª. Drª. Vanessa Galli (UFPel, CDTec)
Prof. Dr. Moisés João Zotti (UFPel, FAEM)

Aos meus pais, meu irmão e a todos que me auxiliaram nesta jornada. Muito obrigado!

“Nothing in Biology Makes Sense Except in the Light of Evolution”
Theodosius Dobzhansky

Agradecimentos

Acima de tudo, à minha mãe, Sandra, meu pai, Roberto (que Deus o tenha), e meu irmão, Oscar, por todo apoio, carinho e incentivo dado durante toda minha vida.

Aos meus amigos “dos tempos de CEFET”, Luiz Gustavo, Felipe, Cássia, Bruno (Bolaxa), Thais, Gabriel e Rai, pela amizade duradoura, conversas sobre assuntos aleatórios e gordices acompanhadas de filmes de procedência duvidosa.

Ao meu orientador, Professor Dr. Luciano Pinto, por todo apoio, conhecimentos passados desde o meu período de iniciação científica, por toda a liberdade na execução dos projetos e pela confiança em mim depositada.

Aos meus colegas e ex-colegas do Laboratório de Bioinformática e Proteômica (BioPro-Lab) que colaboraram para a realização deste trabalho: Marcus Redü Eslabão, Rafael Woloski, Amanda Munari, Christian Sanchez, Rafael Cagliari e Amilton Seixas. A todos os demais membros do BioPro-Lab pelo ótimo ambiente de trabalho e pelo clima descontraído e de camaradagem, tão importante para o desenvolvimento científico.

Á professora Dr. Andrea Moura e ao Dr. Ismail Souza por disponibilizarem amostras de DNA de *Xanthomonas*, e por todo o *feedback* construtivo durante a condução dos projetos, e aos professores Dr. Luciano Carlos da Maia e Dr. Moisés João Zotti por aceitarem colaborar na condição de membros da comissão de acompanhamento e pelo *feedback* construtivo que foi dado.

A todos os professores do Programa de Pós-Graduação em Biotecnologia e Bacharelado em Biotecnologia, pelos conhecimentos passados durante a minha formação.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

Resumo

KREMER, Frederico Schmitt. **Estratégias de bioinformática no estudo da patogenômica de bactérias do gênero *Xanthomonas***. 2019. 240f. Tese (Doutorado) - Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

O gênero *Xanthomonas* abrange bactérias Gram-negativas, muitas das quais fitopatogênicas e agentes infecciosos de plantas de elevado valor comercial. Com o advento do sequenciamento de DNA de nova geração (NGS), o aumento na disponibilidade de dados genômicos, através de técnicas de bioinformática, permitiu a elucidação de muitas das características moleculares que conferem a estes organismos e capacidade de infectar seus hospedeiros. Na presente tese, composta por três sub-projetos apresentados na forma de artigos, são descritas diferentes aplicações e estratégias de bioinformática no estudo das relações entre genoma e patogenia (patogenômica) no contexto do estudo deste gênero. O primeiro artigo descreve uma análise pangenômica da espécie *X. oryzae*, que tem como hospedeiro preferencial o arroz (*Oryza sativa*). Esta espécie engloba dois patovares distintos, *oryzae* e *oryzicola*, sendo cada um capaz de acarretar uma doença específica e com uma dinâmica de infecção e sintomatologia própria. Através da análise realizada a partir do genoma núcleo destes patovares foi possível identificar genes diferencialmente distribuídos entre eles, alguns dos quais diretamente associados com processos de patogênese, como genes responsáveis pela captura de nutrientes, sistema CRISPR e degradação de parede celular. No segundo artigo é descrito o desenvolvimento de uma ferramenta para análise de genes da família TALEs, que estão amplamente distribuídos neste gênero e que atuam na regulação da expressão gênica na planta alvo. A ferramenta descrita é capaz de identificar, a partir de uma genoma não-anotado de *Xanthomonas*, genes que codificam para proteínas desta família e seus respectivos genes-alvo em diferentes plantas hospedeiras. Por fim, no terceiro artigo é descrito o sequenciamento de uma cepa local obtida no Rio Grande do Sul (Brasil), de *X. fuscans*, a partir de amostras de feijão (*Phaseolus vulgaris*). A análise do genoma desta cepa revelou diversos fatores de virulências e caraterísticas funcionais e estruturais novas em relação a cepas de referência para esta espécie. Sendo assim, através das abordagens de análise realizadas e da ferramenta proposta, foi possível identificar novas características genômicas relevantes no estudo da patogenômica de diferentes espécies de *Xanthomonas*, bem como disponibilizar uma nova ferramenta para o estudo de genes associados à patogênese destas bactérias, bem como para a identificação de genes com potencial biotecnológico.

Palavras-chave: genômica, pangenômica, TALEs, sequenciamento de nova geração, fitopatógeno

Abstract

KREMER, Frederico Schmitt. **Bioinformatics strategies in the study of pathogenomics of the *Xanthomonas* genus**. 2019. 240f. Tese (Doutorado) - Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

The *Xanthomonas* genus comprises Gram-negative bacteria species, many of which are phytopathogens of plants with high economic value. With the advent of next generation sequencing (NGS), the increase in the availability of genomic data allowed the identification, through bioinformatics analysis, of many molecular features that grant to these organisms the ability to infect their host. In the present thesis, constituted by three sub-projects presented in the form of papers, different applications of bioinformatics strategies are described regarding the study of the genome-pathogenesis relationship (pathogenomics), in the context of the study of this genus. The first paper describes a pangenomic analysis of the species *X. oryzae*, which has as preferential host rice (*O. sativa*). This species comprises two distinct pathovars, *oryzae* and *oryzicola*, each one responsible to a specific disease and with a specific infection dynamic and symptomatology. Through the analysis performed based on the core genome of these pathovars it was possible to identify differentially distributed genes among them, some of which are directly associated with the pathogenesis process, such as those associated with the nutrient uptake, the CRISPR system and cell wall degradation. In the second paper we describe the development of a new tool for the analysis of genes from the TALE family, which are widely distributed in the Xanthomonas genus and are able to regulate the gene expression in the host. The tool is able to identify, based on a unannotated genome, the genes that code for proteins of the family and their respective targets on the genome of different plants. Finally, the third paper describes the sequencing of a local strain, obtained in Rio Grande do Sul (Brazil), of *X. fuscans*, derived from beans samples (*Phaseolus vulgaris*). The analysis of the genome of this strain showed many new virulence factors and functional and structural features when comparing to the reference strains for this species. Therefore, based on the analysis approaches and the proposed tool it was possible to identify new genomic features relevant in the study of the pathogenomics of different strains of *Xanthomonas*, and also provide a new tool to the study of genes associates with the pathogenesis of these bacteria and those with promising biotechnological potential.

Keywords: genomics, pangenomics, TALEs, next generation sequencing, phytopathogen

Lista de Abreviaturas

ACT – *Artemis Comparison Tool* (Ferramenta de Comparação do Artemis)
BB – *Bacterial Left Blight* (Mancha bacteriana da folha)
BLS – *Bacterial leaf Streak* (Crestamento bacteriano)
CBB – *Common blight of bean* (Crestamento do feijão)
CC – Cancro cítrico
CDS – *Coding DNA Sequence* (Sequência de DNA codificante)
DNA – *Deoxyribonucleic acid* (Ácido desoxirribonucleico)
GO – *Gene Ontology* (Ontologia Gênica)
INPI – Instituto Nacional de Propriedade Industrial
KEGG – *Kyoto Encyclopedia of Genes and Genomes* (Enciclopédia de genes e genomas de Kyoto)
KO – *KEGG Ontology* (Ontologia do KEGG)
ML - *Maximum Likehood* (Máxima Verossimilhança)
NCBI – *National Center of Biotechnology* Information (Centro Nacional de Informação de Biotecnologia)
ncRNA – *Non-coding RNA* (RNA não-codificante)
NGS – *Next Generation Sequencing* (Sequenciamento de Nova Geração)
NJ - *Neighboor-Join* (Junção de Vizinhos)
PCR – *Polymerase Chain-Reaction* (Reação em Cadeia da Polimerase)
RNA – *Ribonucleic acid* (Ácido Ribonucleico)
rRNA – *Ribosomal RNA* (RNA ribossômico)
RVD – *Repeat variable di-residue* (Di-resíduos variáveis repetidos)
SNP – *Single-Nucleotide Polymorphism* (Polimorfismo de Nucleotídeo Único)
T3SS – *Type III Secretion System* (Sistema de Secreção Tipo III)
TALE – *Transcription Activator-Like Effector* (Efetor Similar a Ativador de Transcrição)
tRNA – *Transfer RNA* (RNA transportador)
TSS – *Transcription Start Site* (Sítio de Inicio de Transcrição)
UPGMA - *Unweighted Pair Group Method with Arithmetic Mean* (Método Não-Ponderado de Agrupamento de Pares com Média Aritmética)

Xag – *Xanthomonas axonopodis* pv. *glycines*

Xcc – *Xanthomonas campestris* pv. *campestris*

Xfau – *Xanthomonas fuscas* subsp. *aurantifolli*

Xff – *Xanthomonas fuscans* subsp. *fuscans*

Xoc – *Xanthomonas oryzae* pv. *oryzicola*

Xoo – *Xanthomonas oryzae* pv. *oryzae*

Sumário

1 INTRODUÇÃO GERAL

O gênero *Xanthomonas* compreende bactérias Gram-negativas, sendo muitas destas fitopatogênicas e agentes infecciosos de plantas de elevado valor comercial. Este gênero engloba mais de 30 espécies, sendo muitas destas sub-classificadas em variantes patogênicas (patovares), de acordo com a espécie hospedeira preferencial, doença e sintomas que causam, tecidos e/ou órgãos que são tidos como alvos, bem como a partir de outras características genéticas.

Dentre as espécies e patovares deste gênero que apresentam maior impacto econômico podemos destacar: *X. oryzae*, cujos patovares *oryzae* (*Xoo*) e *oryzicola* (*Xoc*) infectam arroz (*Oryza sativa*), causando as doenças “mancha foliar bacteriana” (*bacterial leaf blight*, ou BB) e “crestamento bacteriano” (*bacterial leaf streak*, ou BLS); *X. axonopodis* pv. *glycines* (*Xag*), que infecta soja (*Glycine max*); *X. citri*, cujos patovares infectam diferentes espécies de *Citrus*; *X. campestris*, cujos patovares infectam uma variada gama de plantas de amplo consumo, como banana, couve e tomate. Esta diversidade observada em termos de hospedeiros também se reflete no que diz respeito à dinâmica de patogênese, onde mesmo patovares de uma mesma espécie, como *Xoo* e *Xoc*, apresentam vias diferentes de infecção e mecanismos próprios de virulência. Entretanto, é também sabido que há mecanismos moleculares que são, até certo ponto, conservados nestas espécies, como os genes da família TALE. Estes genes codificam proteínas efetoras que são translocadas para o interior da célula hospedeira por meio de sistemas de secreção como o T3SS (Sistema de Secreção Tipo III), e atuam regulando a expressão gênica do hospedeiro, facilitando assim o processo de invasão dos tecidos, escape dos mecanismos de defesa do hospedeiro, sobrevivência e colonização.

O advento do sequenciamento de DNA de nova geração (NGS) promoveu um aumento expressivo na disponibilidade de dados genômicos de diferentes organismos em bases de dados públicas, como o GenBank, sendo isto também observado para o gênero *Xanthomonas*. Atualmente, mais de 3300 genomas estão disponíveis apenas para este gênero, considerando genomas disponíveis na forma de rascunho (*draft*

genomes) e sequências finalizadas (*complete genomes*). Entretanto, este grande volume de informação (*big data*), trás outros desafios, como a necessidade de métodos computacionais para processamento e mineração destas informações.

Neste contexto, a bioinformática se apresenta como uma ferramenta de grande valor para o estudo das características genômicas que estejam associadas ao processo de patogênese destas bactérias. Estas abordagens permitem o processamento destas informações e o cruzamento de dados de diferentes fontes de modo a facilitar a mineração de informações relevantes e identificação de novos *insights* referentes ao entendimento do patógeno. Além disso, a análise dos genomas destes organismos também pode auxiliar na identificação de genes com potencial aplicação biotecnológica, como enzimas capazes de degradar celulose e outros substratos de relevância industrial.

A presente tese é composta por três sub-projetos, apresentados na forma de artigos, sendo em cada um abordadas diferentes aplicações da bioinformática no estudo das relações entre genoma e patogenia no contexto do gênero *Xanthomonas*. No primeiro artigo é descrita uma análise pangenômica da espécie *X. oryzae*, com foco na identificação de genes que discriminem os patovar *Xoo* e *Xoc* em relação à patogênese. No segundo artigo é descrito o desenvolvimento e a validação de uma nova ferramenta para identificação e análise de genes da família TALEs a partir de genomas, denominada TargeTALE. Por fim, o terceiro artigo descreve o sequenciamento e análise do genoma de uma cepa local, obtida na região do Capão do Leão (Rio Grande do Sul, Brasil), de *X. fuscans*. Além disso, este documento também agrega 4 anexos (A-D), sendo os anexos A e B referentes ao registro do programa de computador descrito nos Artigos 1 e 2, e os anexos C e D referentes a artigos publicados durante o período de execução deste projeto que abordam metodologias utilizadas nos sub-projetos 1 e 3.

2 REVISÃO BIBLIOGRÁFICA

2.1 O gênero *Xanthomonas*

O gênero *Xanthomonas* (*Xantho*, que em Grego significa “amarelo”) compreende bactérias Gram-negativas, sendo muitas destas fitopatogênicas e agentes infecciosos de plantas de elevado valor comercial. Este gênero engloba mais de 30 espécies, sendo muitas destas sub-classificadas em variantes patogênicas (patovares, “pv”), de acordo com a espécie hospedeira preferencial, doença e sintomas que são acarretados, tecidos e/ou órgãos que são tidos com alvos, bem como a partir de outras características genéticas e/ou fenotípicas (Vauterin et al., 1995; Vauterin et al., 2000).

2.1.1 X. oryzae

Xanthomonas oryzae é uma fitopatogênicas, sendo considerado uma das principais pragas de arroz (*Oryza sativa*). É subclassificada em patovares, *oryzae*, o agente etiológico da mancha bacteriana (*bacterial blight*, BB), e *oryzicola*, o agente etiológico das estrias bacterianas da folha (*bacterial leaf stroke*, BLS) (Figura 1.A e 1.B), respectivamente). O impacto de cada um destes patovares na produção varia de acordo com o genótipo de planta e a linhagem da bactéria (também denominada “raça”), sendo que o pv. *oryzae* (ou *Xoo*) pode levar a perdas que variam de 20 a 80%, e o pv. *oryzicola* (*Xoc*) acarretar em perdas de 5 a 30% (Vauterin et al., 1995; Niño-Liu et al., 2006). As cepas de *Xoo* são sub-classificadas em linhas asiáticas, africanas e americanas, sendo estas últimas distintas em nível molecular das demais, já as cepas de *Xoc* são classificadas em africanas e asiáticas (Triplett et al., 2011).

BB é uma doença vascular que afeta as folhas da planta. A bactéria inicia sua infecção entrando através dos hidatódios, coloniza o espaço intercelular da epiderme e atinge o xilema, partindo para o parênquima e se movendo verticalmente até as folhas através dos vasos primários. Em alguns dias, *Xoo* coloniza o tecido vascular da folha,

que se torna cheio de células bacterianas e polissacarídeos extracelulares, resultando nos sintomas clínicos das manchas amareladas na superfície das folhas (Mew, 1993; Ou, 1985; Noda & Hisatoshi, 1999). Já a *Xoc*, invade a planta principalmente através dos estômatos, se multiplica através da cavidade sub-estomatal e coloniza o parênquima. O inicio da infecção é marcado pelo aparecimento de lesões com acúmulo de água e coloração amarela, que são delimitadas pelos vasos da planta. À medida que a infecção progride as lesões necrosam e o tecido morre (Mew, 1993; Nyvall, 1999).

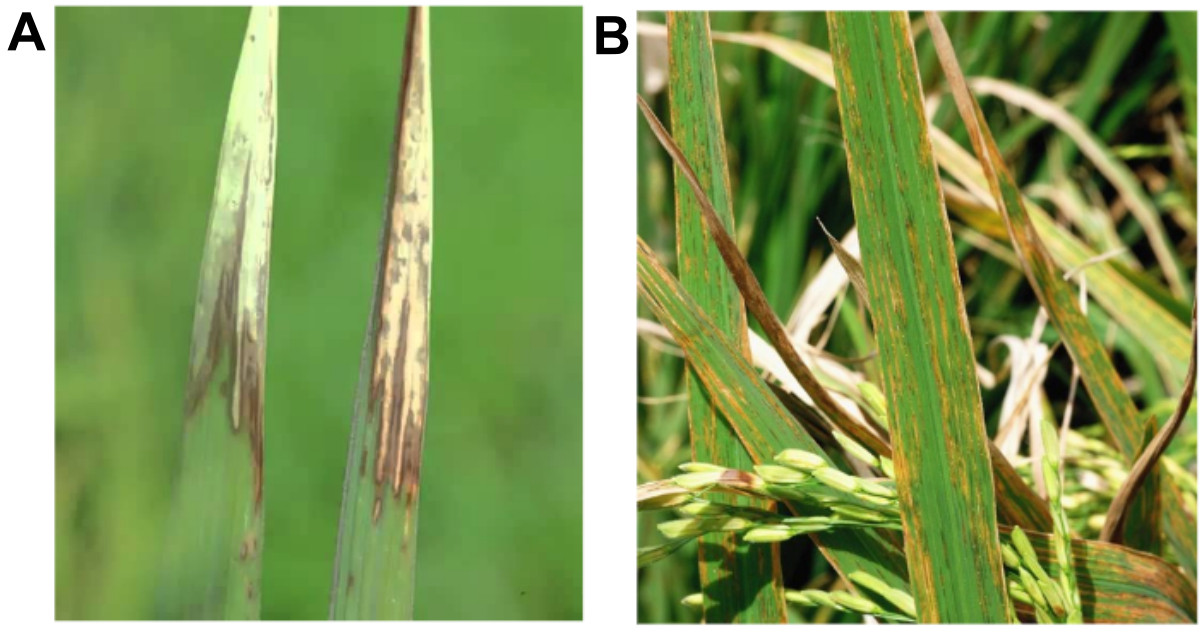


Figura 1. Folhas de arroz (*O. sativa*) infectadas por *Xanthomonas oryzae*. (A) Infecção por *X. oryzae* pv. *oryzae* (*Xoo*) acarretando sintomas de crestamento bacteriano comum. (B) infeccção por *X. oryzae* pv. *oryzae* (*Xoc*) acarretando sintomas das estrias bacterianas.

2.1.2 X. campestris

X. campestris é a espécie mais complexa do gênero, visto que engloba o maior número de patovares já descritos; no entanto, muitos destes foram posteriormente reclassificados como espécie ou patovares de outras espécies (Vauterin et al., 1995). Os patovares de *X. campestris* cobrem uma ampla variedade de hospedeiros, incluindo diferentes cultivares de *Brassica oleracea*, como repolho, couve-flor, brócolis e couve-

de-bruxelas, bem como nabo (*Brassica napa*), rabanete (*Raphanus raphanistrum* subsp. *sativus*), banana (*Musa acuminata*), dentre outras (Williams, 1980; Bonas et al., 2000).

O processo de patogênese desta espécie varia de acordo com o patovar, mas de modo geral inicia com a invasão da planta hospedeira através de lesões ou por meio de estruturas como os hidatódios (ex: *X. campestris* pv. *campestris*) e os estômatos (ex: *X. campestris* pv. *vesicatoria*). No caso de *X. campestris* pv. *campestris* (*Xcc*), a bactéria migra para o tecido vascular e posteriormente para os tecidos foliares, onde são iniciadas as lesões típicas da doença, caracterizadas pelas manchas amarelas nas extremidades (Alvarez, 2000). Já no caso de *X. campestris* pv. *vesicatoria* (*Xcv*), o patógeno não chega a se espalhar pelo tecido vascular, e permanece no tecido intercelular (Bonas et al., 2000). A virulência de *X. campestris*, bem como outras espécies de *Xanthomonas*, está diretamente relacionado com os genes da família *Hrp*, que regulam a resposta de hipersensibilidade no hospedeiro. Estes genes codificam proteínas que interagem com o sistema de secreção tipo III (T3SS), um complexo multiprotêico responsável pela translocação de proteínas denominadas “efetoras” para o interior da célula hospedeira, bem como facilitam a translocação destas para a mesma (Weber & Koebnik, 2005). Além disso, *X. campestris* possui também um amplo arsenal de proteínas que auxiliam no processo de invasão, como celulases, proteases e lipases, bem como fatores de virulência, como a goma xantana, que auxilia na fixação na superfície do vegetal hospedeiro (Solé et al., 2015; Rosseto et al., 2016).

2.1.3 C. citri

Xanthomonas citri pv. *citri* é um dos principais patógenos de variantes asiáticas de *Citrus*, sendo agente etiológico dos cancros cítricos (CC) dos tipos A, A* e A^W, cada um destes acarretados por diferentes linhagens (patotipos) deste patovar. Destes, o patotipo A é o que apresenta maior diversidade de hospedeiros, podendo infectar a maioria das espécies de *Citrus* conhecidas. Já os patotipos A* e A^w limitam-se às espécies *Citrus aurantifolia* e *Citrus macrophylla* (Brunings & Gabriel, 2003). A invasão

da planta hospedeira se dá através dos estômatos e de lesões na superfície, e posteriormente a bactéria migra para as demais partes aéreas (Leduc et al., 2015).

O CC se caracteriza pela formação de lesões na forma de erupções na folha e nos frutos, que iniciam com uma margem com aspecto aquoso e cercados por um halo. Com o tempo a lesão aumenta de tamanho e se torna mais escura, com tom marrom. Lesões nas folhas eventualmente levam à queda destas, o que pode se estender a uma desfolhação completa da planta (Ference et al., 2018). A bactéria também pode ser carreada pelo ar, e desde modo se propagar por quilômetros de distância (Gottwald et al., 2002).

2.1.4 X. fuscans

Xanthomonas fuscans compreende dois patovares: *X. fuscans* subsp. *fuscans* (*Xff*) e *X. fuscans* subsp. *aurantifolii* (*Xantau*). *Xff*, junto com *X. axonopodis* pv. *phaseoli* (*Xap*), é um dos patógenos mais devastadores de feijão e acarreta grandes perdas na produtividade. É o agente etiológico da mancha comum do feijão (*common blight of bean*, ou CBB), que afeta o feijão “comum” (*Phaseolus vulgaris* L.), feijão-lima (*Phaseolus lunatus* L.), dentre outras espécies (Akhavan et al., 2013). Já *Xantau* é o agente etiológico dos cancros tipo B e C, que acometem diferentes espécies de C*itrus* na América do Sul. Por questões de similaridade genômica e sintomatológica (no caso de *Xantau*) em relação a outras espécies, classificações alternativas já foram sugeridas, incluindo a realocação destes como patovares de *X. citri* (Constantin et al., 2016).

A distribuição mundial de CBB é atribuída aos mecanismos eficientes de transmissão entre sementes empregados pela bactéria. Estudos demonstraram um impacto na produção de 40% (Morris & Monier, 2003), com impactos diretos e indiretos para os produtores. Os prejuízos diretos dizem respeito às perdas primárias na produção, qualidade e no custo adicional com pós-colheita, enquanto os impactos indiretos ou efeitos secundários incluem as perdas para os vendedores, exportadores, consumidores

e para os governos. Nas Américas e África, o feijão representa 60% do consumo de proteína diário (Graham & Vance, 2003).

A patogênese de *Xff* e *Xap* envolve a formação de biofilmes bacterianos na superfície da planta, seguido da penetração através dos estômatos, levando por sua vez a colonização do mesofilo, tecidos vasculares e parênquima (Jacques et al., 2005) e por fim na dispersão da bactéria para outras plantas. A doença é caracterizada por pontos de necrose nas folhas, hastes e sementes; apesar de que muitas plantas podem se manter assintomáticas (Ruh et al., 2017). Já a patogênese de *Xantau* é similar á de *X. citri* pv. *citri*, e sua sintomatologia também.

2.2 Fatores de virulência

2.2.1 Sistemas de secreção

Os sistemas de secreção são complexos multiproteicos que efetuam o transporte de proteínas do meio intracelular para o exterior da célula bacteriana. Atualmente, são conhecidos 6 tipos de sistemas de secreção (I-VI), sendo o tipo V sub-classificado em Va, Vb e Cc. No caso dos sistemas III, IV e VI, este transporte é realizado para o interior da célula do organismos hospedeiro, e as proteínas transportadas são denominadas *efetoras*, visto que estas efetuam alterações na fisiologia do organismo hospedeiro de modo a beneficiar o organismo infectante. No caso de *Xanthomonas*, estes sistemas são os de maior relevância para o processo de patogênese. Uma representação destes diferentes sistemas está apresentada abaixo, na Figura 2. O sistema de secreção tipo III (T3SS) é particularmente importante para a translocação dos efetores do tipo TAL para o interior da célula hospedeira (Muñoz Bodnar et al., 2013). Este processo é regulado pelos genes da família *Hrp*, que desencadeiam o processo de translocação quando há o contato direto da célula hospedeira com a do patógeno (Brunings & Gabriel, 2003; Weber & Koebnik, 2005).

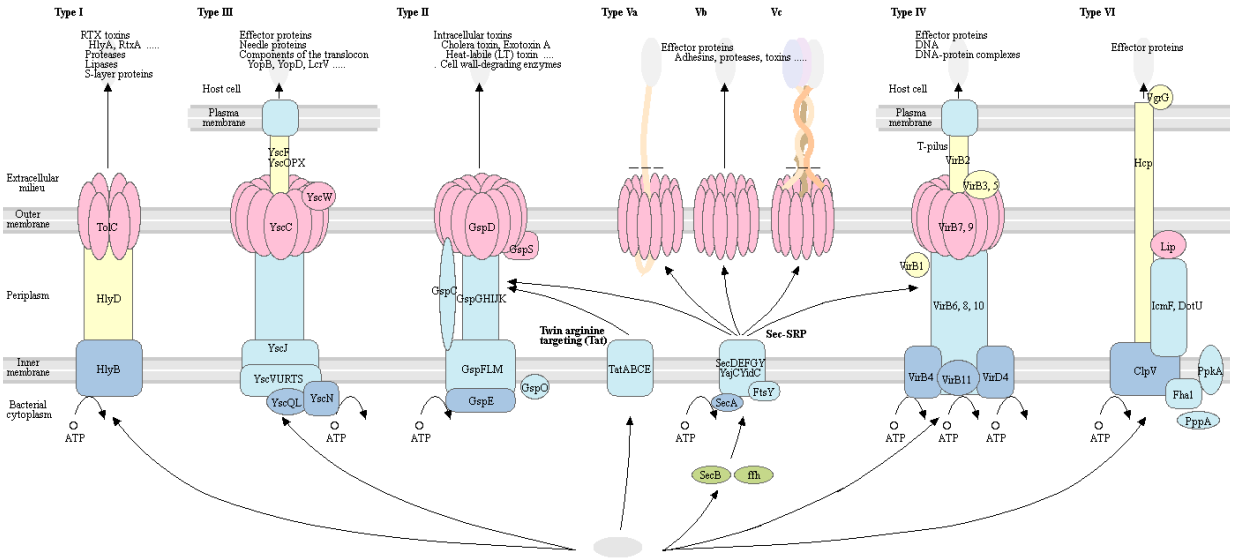


Figura 2. Representação dos diferentes sistemas de secreção presentes em bactérias (I-VI). Imagem obtida a partir do banco de dados do *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (KO: ko03070).

2.2.2 Genes de virulência e avirulência

As interações patógeno-hospedeiro durante a infecção por bactérias do gênero *Xanthomonas* são reguladas por diferentes proteínas destes dois organismos, o que inclui genes de virulência (*vr*) e avirulência (*avr*) do patógeno e genes de resistência (R) da planta. Genes de resistência codificam proteínas capazes de reconhecer proteínas *avr* secretadas pelo patógeno, o que desencadeia uma resposta de hipersensibilidade que é caracterizada pela indução de uma morte rápida das células localizadas próximas ao sítio de infecção de modo a se conter a dispersão do patógeno (Heath, 2000). Já as proteínas *vr* atuam ativamente no processo de virulência e patogênese, o que inclui a regulação de processos celulares e metabólicos do hospedeiro. A presença de genes de avirulência torna a cepa apenas avirulenta em plantas hospedeiras que contenham o gene de resistência cujas proteínas codificadas são capazes de reconhecer estes fatores, resultando assim em uma especificidade “raça-específica” (Brunings & Gabriel, 2003).

2.2.3 TALEs

Os efetores similares a ativadores de transição ("*Transcription Activation-Like Effectors*", ou TALEs), são proteínas produzidas por muitas espécies do gênero *Xanthomonas* e secretadas através do T3SS. Estas proteínas atuam na regulação da expressão gênica nas células da planta hospedeira através do reconhecimento das regiões promotoras dos genes. Este reconhecimento é feito através de domínios repetitivos destas proteínas denominados *repeat di-residues* (RVDs), que são constituídos por ~34 aminoácidos com duas posições hipervariáveis no centro. Uma representação da estrutura de um efetor TALE interagindo com o DNA está apresentada na Figura 3 (Muñoz Bodnar et al., 2013).

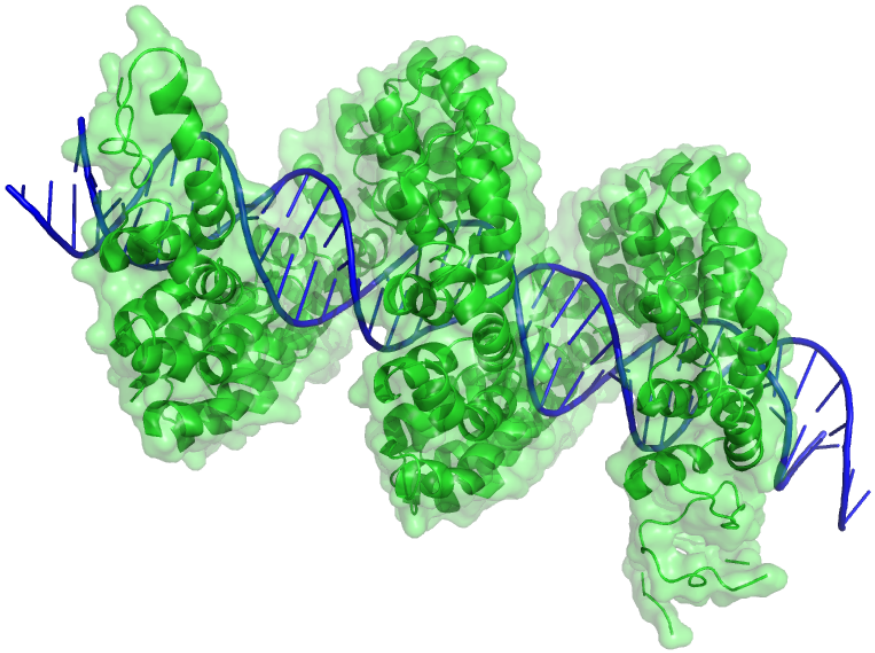


Figura 3. Estrutura de uma proteína da família TALE, PthXo1 de *X. oryzae*, interagindo com a estrutura de um fragmento de DNA derivado do genoma de *O. sativa* (PDB: 3UGM).

O modelo de interação entre os RVDs e o DNA foi elucidado em 2009 (Boch et al., 2009; Moscou & Bogdanove, 2009), sendo possível a partir deste desenvolver ferramentas capazes de predizer, a partir de uma sequência de RVDs, qual sequência

de DNA seria reconhecida. Ferramentas como TargetFinder (Doyle et al., 2012), StoryTeller (Pérez-Quintero et al., 2013), TALgetter (Grau et al., 2013) e TALvez (Pérez-Quintero et al., 2013) permitem que, a partir de um modelo de RVD, esta busca seja efetuada em um banco de dados de sequência providas pelo usuário (no caso de aplicações locais), ou bases padronizadas (quando executando em um servidor remoto). Já a ferramenta AnnoTALE (Grau et al., 2016) permite tanto a busca, quanto a identificação de genes de TALEs em sequência de um genoma, realizando também uma classificação estrutural dos genes identificados através de uma nomenclatura padronizada. Além disso, com uma maior disponibilidade de dados a respeito de TALEs e seus alvos, bases de dados especializadas também foram desenvolvidas, como o daTALbase, que inclui também ferramentas de análise incorporadas (Pérez-Quintero et al., 2018).

De fato, a aplicação deste conhecimento não se limitou ao uso no estudo das proteínas efetoras no processo de patogênese. Posteriormente a possibilidade de prever e desenhar TALEs específicos para uma determinada sequência permitiu o desenvolvimento de metodologias para edição de genomas, como a técnica TALEN (Zhang et al., 2014), que consiste em fusionar uma sequência de TALE com uma endonuclease, permitindo assim que sequências específicas sejam reconhecidas e modificadas em uma determinada célula. No entanto, com o advento de técnicas baseadas em CRISPR, o uso de TALENs acabou decaindo .;, (Ran et al., 2013; Gaj et al., 2013; He et al., 2016).

2.2.4 Enzimas secretadas

O processo de patogênese de bactérias do gênero *Xanthomonas* também é facilitado por enzimas que atuam na degradação de diferentes substratos biológicos, como celulases (Temuujin et al., 2011; Tayi et al., 2018), pectinases (Tayi et al., 2016a), proteases (Solé et al., 2015) e lipases (Mo et al., 2016). Algumas destas enzimas, como as celulases de CelXoA e CelXoB identificadas em *Xanthomonas oryzae* pv. *oryzae* cepa KACC10859, foram demonstradas como necessárias para uma virulência completa em

ensaios com mutantes para os genes que as codificam (Temuujin et al., 2011). Muitas destas enzimas atuam na degradação de parede celular, facilitando assim a invasão dos tecidos da planta e o contato dos sistemas de secreção com a célula vegetal. As enzimas que atuam na degradação de componentes da parede celular, como celulases, hemicelulases e pectinases, são de particular interesse biotecnológico e sua identificação e caracterização em genomas de bactérias fitopatogênicas também serve de bases para o desenvolvimento de insumos aplicáveis em indústrias como a de biocombustíveis na produção do denominado “etanol de segunda geração” (Mendes et al., 2017).

2.3 Genômica de *Xanthomonas*

O advento do sequenciamento de DNA de nova geração (*Next Generation Sequencing*, NGS) representadas pelas plataformas de sequenciamento Roche 454, Illumina HiSeq e MiSeq, Iontorrent PGM e ABI SOLiD, dentre outras, resultou em um aumento expressivo na disponibilidade de dados genômicos em bases de dados públicos para uma ampla variedade de organismos (Mardis, 2008; Liu et al., 2012), sendo isso também observado para o estudo das bactérias do gênero *Xanthomonas*. Em 2005 foram disponibilizados os primeiros genomas de cepas das espécies *X. oryzae* (Lee et al., 2005) e *X. campestris* (Thieme et al., 2005). Ao longo dos anos seguintes, o volume de dados aumentou progressivamente, chegando ao patamar observado hoje, com mais de 3300 genomas de *Xanthomonas* disponíveis no GenBank (Benson et al., 2005). Este grande volume de dados, no entanto, trás novos desafios, uma vez que torna necessário o uso de abordagens que facilitem o processo de mineração de informações relevantes para o estudo destes organismos.

2.4 Abordagens para análise genômica de micro-organismos

2.4.1 Análise estrutural do genoma

A análise estrutural do genoma permite a identificação de rearranjos, translocações e inversões que ocorrem na organização do cromossomo. Esta abordagem tem como base o conceito de sintenia, que consiste em “blocos de genes” (Edwards & Holt, 2013). Estas alterações na estrutura cromossômica são geralmente acarretadas por elementos móveis no genoma, como transposons e fagos, que compõe o mobiloma. Além disso, mecanismos de recombinação homóloga e transferência horizontal de genes também resultam em alterações em nível estrutural da organização do genoma, podendo resultar tanto na incorporação de novos genes quanto na perda de blocos inteiros. Além de terem um efeito aditivo ou subtrativo no repertório de genes presentes em um determinado organismo, essas alterações estruturais também podem resultar em modificações em nível de expressão gênica, visto que durante o processo de translocação é possível que certas regiões não codificantes com efeito regulatório sejam afetadas, o que pode comprometer *clusters* e *operons* inteiros.

Diferentes programas podem ser utilizadas para este tipo de análise, sendo a ferramenta Mauve (Darling et al., 2004) e a ferramenta *Artemis Comparison Tool* (ACT) algumas das mais utilizadas. A ferramenta Mauve permite a identificação de blocos de sintenia através da análise dos chamados blocos localmente co lineares (*locally collinear blocks*, LCB). Estes blocos são sequências conservadas em 2 ou mais genomas que apresentam similaridade maior que um determinado limiar (*threshold*), que são identificadas e refinadas progressivamente através de um algoritmo iterativo, que tenta minimizar o número de eventos de inversão necessários para representar as diferenças estruturais entre os genomas. Esta mesma abordagem foi utilizada, posteriormente, para o desenvolvimento de ferramentas para finalização de genomas a partir de sequências de referência (Rissman et al., 2009). Já a ferramenta ACT utiliza para processamento das regiões sintênicas os resultados do alinhamento gerado pela ferramenta BLASTn em formato tabular. Diferente do Mauve, o ACT não utiliza etapas iterativas para o

refinamento da organização dos blocos, mas permite a filtragem de regiões identificadas como similares através de *threshold* de similaridade a partir dos alinhamentos do BLAST.

A análise de sintenia já foi aplicada na caracterização de *loci* de genes que codificam para proteínas da família TALE em *X. oryzae*, tendo demonstrado que mesmo em cepas geneticamente próximas, estes genes apresentam altas variabilidade posicional no genoma (Grau et al., 2016). Além disso, Cesbron et al (2015) também a empregaram na caracterização de regiões genômicas capazes de discriminar cepas patogênicas e não patogênicas de *X. aboricola*, uma espécie que infecta plantas do gênero *Prunus*, como o pessegueiro (*P. persica*) e diversas espécies de cerejeiras.

2.4.2 Análise funcional a partir de bancos de ontologia

O processo de anotação de um genoma consiste na identificação de seus elementos estruturais e funcionais, sendo geralmente realizado logo após a sua montagem e é atualmente mandatório para o depósito em bancos de dados públicos como o GenBank (Kisand & Lettieri, 2013; Edwards & Holt, 2013). A identificação funcional, no entanto, é geralmente limitada ao nível de produto gênico (proteína codificada), não sendo usual a incorporação informações mais aprofundadas sobre ontologia e processos biológicos (Pareja-Tobes et al., 2012; Sallet et al., 2014; Seemann, 2014; Kremer et al., 2016). A anotação funcional fornece informações valiosas para um maior entendimento dos processos biológicos desempenhados por cada gene do organismo de interesse, permitindo assim que rotas metabólicas e vias de regulação sejam elucidadas (Kasif & Steffen, 2010).

Bancos de dados como o *Cluster of Ortholog Groups* (COG), por exemplo, fornecem uma nomenclatura padronizadas para processos biológicos comuns a muitos microrganismos, permitindo uma maior padronização nas anotações de genomas e um melhor entendimento da distribuição dos diferentes produtos gênicos de um organismo nestes processos biológicos, o que facilita inclusive análises comparativas. Este conceito de organização padronizada de genes foi posteriormente estendido para a criação de

ontologias, que consiste em organizações estruturais de processos e funções compartilhadas por diferentes organismos. Exemplos de ontologias incluem os bancos de dados *Gene Ontology* (GO), uma organização hierárquica de processos, funções e localizações subcelulares (Harris et al., 2004), e o *KEGG Ontology* (KO), que agrega dados de grupos ortólogos associados a rotas metabólicas descritas na base de dados *KEGG Pathways* (Kanehisa & Goto, 2000).

A análise funcional a partir do banco de dados de ontologia do *Gene Ontology* pode ser realizada através de diferentes ferramentas automatizadas para anotação funcional, como o BLAST2GO (Conesa & Götz, 2008), que realiza a anotação de proteínas através de um mapeamento das funções associadas aos seus hits mais relevantes no BLAST (Altschul et al., 1990), e o QuickGO (Binns et al., 2009), que ligam a função a proteínas do UniProt anotadas automaticamente (Apweiler et al., 2004). Além disso, ferramentas como o KAAS (Moriya et al., 2007) e o BlastKOALA (Kanehisa et al., 2016) permitem a anotação de genomas através de ontologias metabólicas do KEGG Pathways.

A anotação funcional permite também análises comparativas entre diferentes conjuntos de genes. No caso de anotações derivadas do *Gene Ontology*, por exemplo, é possível comparar a ocorrência de termos entre dois conjuntos de genes através de testes hipergeométricos ou de contingência (ex: Teste Fischer, Chi-quadrado) de modo a se identificar aqueles que estão enriquecidos ou subrepresentados (Glass & Girvan, 2014).

2.4.3 Análise de variantes genéticas

Variantes genéticas, como polimorfismos de nucleotídeo único (*single-nucleotide polymorphisms*, SNPs) e inserções e deleções (INDELs), são alterações genômicas que afetam um número restritos de bases, sendo geralmente denominadas mutações pontuais. Apesar de ser possível identificar estas alterações a partir do alinhamento direto entre genomas, como o efetuado por ferramentas como o Mummer (Kurtz et al.,

2004), a baixa confiabilidade encontrada sobretudo em genomas rascunho faz esta abordagem pouco confiável. Deste modo, alternativamente, no contexto do NGS a estratégica para análise de variantes consiste em utilizar ferramentas de mapeamento de leituras, como Bowtie (Langmead et al., 2009), BWA (Li & Durbin, 2009) e Segemehl (Hoffmann et al., 2009), para se alinhar as leituras do genoma de interesse contra o genoma de referência, sendo posteriormente os resultados deste alinhamento processados através de ferramentas como SamTools (Li et al., 2009) e GATk (McKenna et al., 2010). O resultado deste processo são as coordenadas dos sítios com variação, bem como métricas de qualidade para cada. Estas variações podem ser então comparadas com a anotação do genoma do organismo de interesse com o uso de ferramentas como o SnpEff (Cingolani et al., 2012), de modo a se identificar o impacto funcional de cada alteração.

A análise de variantes em genomas microbianos é particularmente importante para se identificar mutações pontuais que caracterizem novas cepas, sendo uma técnica promissora para a caracterização de isolados em casos de epidemias de diferentes patógenos (Schürch et al., 2018). Além disso, a análise de SNPs também auxilia no entendimento do processo adaptativo destes organismos, se estendendo também ao que tange o próprio processo de patogênese (Klemm & Dougan, 2016).

No caso de *Xanthomonas citri* pv. *citri*, Gordon et al (2015), realizou uma análise comparativa de 43 isolados utilizando análise de variantes de modo a identificar alterações que estivesses associadas com os três diferentes patotipos (A, A* e A^W), tendo conseguido satisfatoriamente mapear não apenas mutação não-sinonímicas em genes relacionados à patogênese, mas ilhas inteiras de patogênese que caracterizam cada patotipo.

2.4.4 Análise filogenômica

O estudo da filogenética de organismos microbianos em um primeiro momento foi majoritariamente baseada na análise de genes altamente conservados, como os que

codificam para as unidades ribossomais 16S e 18S (Weisburg et al., 1991; Woo et al., 2008; Tringe & Hugenholtz, 2008). Estes marcadores, no entanto, oferecem pouca capacidade de resolução para diferenciar cepas geneticamente próximas, tendo sido então, posteriormente, desenvolvidas outras estratégias que empregam um número maior de genes de modo a se obter capacidade de discriminação. Neste sentido, a estratégia de *Multi-locus Sequence Typing* (MLST), que são baseadas em conjuntos padronizados com múltiplos genes contendo diferentes níveis de variação (alelos) dentro das espécies de interesse, passaram a ser adotado com padrão para muitas espécies microbianas (Maiden et al., 2013; Jolley & Maiden, 2010).

Com o advento do NGS, o grande número de cepas com genomas disponíveis permitiu que a análise filogenética não ficasse limitada apenas a poucos genes, mas sim que utilizasse centenas ou milhares de regiões do genoma de modo a se poder discriminar de forma mais precisa as diferentes linhagens de cada espécie. Ferramentas de alinhamento como o Mugsy (Angiuoli & Salzberg, 2011), permitem a identificação de regiões sintênicas em centenas de organismos, sendo a partir de cada uma destas regiões realizo um alinhamento múltiplo, que posteriormente pode ser utilizado para a construção de árvores filogenéticas. Há também estratégias filogenômicas baseadas em inferência de um genoma núcleo, e posterior reconstrução de, como a ferramenta BlastGraph (Ye et al., 2013) e o pacote Harvest (Treangen et al., 2014).

A técnicas de análise filogenética baseadas em genomas inteiros, entretanto, são também susceptíveis a certos artefatos de sequenciamento, montagem, e em certos casos até de anotação, sobretudo devido à dificuldade de se obter genomas finalizados (sem *gaps*) (Mardis et al., 2002), com as metodologias NGS (Fraser et al., 2002; Branscomb & Predki, 2002; Kremer et al., 2017). Além disso, limitações das plataformas de NGS também dificultam na montagem de regiões repetidas (ex: genes duplicados) ou repetitivas (ex: sequências em tandem, como as *short sequence repeats* (SSR)) do genoma, o que pode comprometer a qualidade de certos marcadores filogenéticos baseados em repetições, como os *variable number tandem repeats* (VNTRs) (Alkan et al., 2010). Em termos de algoritmos, também é necessário adequar o algoritmo de

inferência da árvore filogenética ou tipo de dado que está sendo utilizado. No caso de análises baseadas em genomas inteiros, algoritmos mais sofisticados como *Neighboor-Join* (NJ) e *Maximum Likehood* (ML), amplamente utilizados para a análise de sequências curtas, podem resultar em resultados distorcidos por erros de sequenciamento ou pela distribuição de variabilidade nas diferentes posições do genoma, enquanto que outros mais simples, como o *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA), dado um número grande de genomas, podem ser a única estratégica aplicável por limitações computacionais (Lees et al., 2018).

2.4.5 Análise pangenômica

O termo “pangenoma” foi proposto por Medini *et al* em 2005 para descrever a totalidade de genes presentes em um determinado grupo taxonômico de organismos, sendo geralmente utilizado dentro do contexto de gênero ou espécie. A partir deste conceito são derivados diferentes subconjuntos de genes, de acordo com o seu grau de conservação dentro dos organismos analisados, como genoma núcleo (*core genome*), genoma acessório (*accessory genome*) e genoma exclusivo (*exclusive genome*). O genoma núcleo é composto pelos genes presentes em todas (*hard core*) ou ao menos 95% (*soft core*) dos organismos analisados, sendo este último critério usado como forma de reduzir a interferência de artefatos de sequenciamento, montagem ou anotação nas análises *downstream*, algo que é particularmente relevante ao se utilizar dados de genomas rascunho (*drafts*), que apresenta genes faltantes ou fragmentados.

O pangenoma pode ser classificado de acordo com a sua distribuição cumulativa de genes, sendo para isso necessário se inferir utilizando as diferentes permutações dos genomas analisados o grau de “aumento” no número de genes observado a medida que novos genomas são adicionados à análise. Neste caso, são considerados pangenomas “fechados” aqueles em que não é provável um aumento substancial no repertório de genes quando um novo genoma é adicionado, e “abertos” quando este aumento é provável (Rouli et al., 2015).

A análise do pangenoma tem sido amplamente empregada de modo a se entender os elementos funcionais que definem um determinado grupo de organismos, e o que os diferem de outros grupos próximos. No caso de *Escherichia coli*, esta abordagem foi empregada por Rasko et al (2008) visando entender as características genômicas (*genomic features*) que estão relacionadas com os diferentes patotipos desta bactéria, permitindo a identificação de genes associados aos genótipos enterotóxicos e enteropatogênicos. Esta abordagem também foi utilizada recentemente para o estudo de bactérias do gênero *Leptospira*, de modo a identificar elementos estruturais e funcionais que fossem capazes de diferenciar espécies patogênicas de não-patogênicas (saprófitas) (Fouts et al., 2016).

No gênero *Xanthomonas* a análise pangenômica já foi empregada por Garita-Cambronero et al (2017) para identificar genes responsáveis pela patogênese de *X. arborícola* pv. *pruni*. Este patógeno infecta plantas do gênero *Prunus*, como o pessegueiro (*P. pérsica*), tendo a análise permitido a identificação diversos genes de virulência e efetores secretados pelo T3SS que são característicos de cepas virulentas. A partir dos marcadores identificados, um sistema de caracterização de cepas para a espécie baseado em reação em cadeia da polimerase em tempo real (*Real-Time PCR*).

2.5 Enzimas microbianas de relevância industrial

A alta diversidade genética e de repertório enzimático de bactérias e fungos torna estes organismos uma importante fonte para a obtenção de enzimas para aplicações industriais. Enzimas hidrolíticas, como celulases, pectinases e lipases, são amplamente empregadas na produção de alimentos, fármacos, tecidos e biocombustíveis (Liu & Kokare, 2017), e a identificação de novas proteínas destas classes passa a ser uma demanda constante.

2.5.1 Celulases

Celulases são uma classe de enzimas capazes de degradar a celulose, o principal componente da parede celular de plantas, sendo produzidas por um amplo espectro de espécies de fungos, bactérias e protozoários. Esta classe pode ser sub-classificada de acordo com seu mecanismo em endoglucanases (EC 3.2.1.4), beta-glicosidases (EC 3.2.1.21) e exoglucanases (EC 3.2.1.176). Medie *et al* (2012) demostrou que de um total de 1500 genomas bacterianos, ~40% apresentavam pelo menos um gene codificante para uma destas enzimas. Atualmente, são a terceira classe de enzimas mais consumido pela indústria por conta de sua aplicação no processamento de algodão, papel e na indústria alimentícia, podendo passar ao primeiro lugar com o crescimento na sua demanda pela indústria dos biocombustíveis (Wilson, 2009). Na indústria dos biocombustíveis, as celulases são de grande relevância para a produção do etanol celulósico, permitindo maior aproveitamento da matéria-prima através da hidrólise dos polissacarídeos em açucares simples fermentáveis (Sticklen, 2008).

O interesse na identificação e caracterização de genes que codifique para celulases motivou o sequenciamento de um grande número de genomas microbianos, muitos destes sendo de fitopatógenos ou de bactérias do solo. A cepa PF008 de *Clavibacter michiganensis*, uma bactéria que infecta espécies com o tomate (*Solanum lycopersicum*) e pimenta (*Capsicum* spp) e já reportada como produtora de enzimas desta classe, teve seu genoma sequenciado de modo a se caracterizar o repertório de genes que atuam nestes processos de degradação (Bae et al., 2015). Já no caso de *Thermobifida fusca*, uma bactéria do solo, a análise do genoma permitiu a identificação de mais de 45 enzimas degradadoras de carboidratos, muitas das quais celulases (Lykidis et al., 2007). Em *Xanthomonas*, celulases já foram reportadas como essenciais para a patogênese de certas cepas de *X. oryzae* (Temuujin et al., 2011), *X. campestris* (Rosseto et al., 2016) e *X. citri* (Xia et al., 2016).

2.5.2 Pectinases

Pectinases são uma classe diversificada de enzimas capazes de degradar a pectina, um heteropolissacarídeo ramificado constituído por ácido galacturônico, ramnose, arabinose e galactose. Este polissacarídeo é um dos principais constituintes da parede celular e o principal da lamela média, uma estrutura que mantêm a união entre as células vegetais (Uenojo & Pastore, 2007). Estas enzimas são amplamente utilizadas na indústria alimentícia, principalmente na de sucos, visto que a presença de pectina aumenta a viscosidade e a turbidez dos extratos das frutas, dificultando o processo de filtragem (Alkorta et al., 1998). Da mesma forma que as celulases, as pectinases também são utilizadas no processamento de algodão, e também podem ser empregadas na remoção de impurezas no processamento de purificação da celulose (Pedrolli et al., 2009).

Uma ampla variedade de enzimas desta classe com aplicabilidade industrial já foi reportada em bactérias, como *Bacillus macerans* e *Paenibacillus barcinonensis* e fungos, como *Aspergillus japonicus* e *Penicillium canescens* (Pedrolli et al., 2009). Além disso, estas enzimas estão diretamente relacionadas com o processo de patogênese de muito fitopatógenos (Collmer et al., 1991), sendo já demonstradas como necessárias para a virulência de cepas de *X. oryzae* pv. *oryzae* (Tayi et al., 2016b) e *X. campestris* pv. *campestris* (Wang et al., 2008).

2.5.3 Lipases

Lipases (EC 3.1.1.3) são enzimas que catalisam a hidrólise de triglicerídeos, sendo atualmente uma das classes de enzimas com maior relevância na indústria (Liu & Kokare, 2017). No setor alimentício são amplamente utilizadas na transformação de gorduras, de modo a se obter lipídeos de maior valor comercial e com propriedades mais adequadas para cada aplicação. Já na indústria têxtil, são usadas em diversas etapas de tratamento de tecidos, inclusive de fibras sintéticas, de modo a garantir maior resistência, capacidade de absorção de tinturas e compostos antimicrobianos, dentre

outras propriedades (Hasan et al., 2006). Estas enzimas também são de grande utilidade no setor farmacêutico, onde podem ser utilizadas, devido a sua alta-especificidade, na obtenção e separação de isômeros (enantioseletividade) (Gotor-Fernández et al., 2006). Na indústria de detergentes e outros produtos de limpeza, as lipases também são são adicionadas como constituintes destes produtos para uso doméstico e industrial em conjunto com outras enzimas com atividade hidrolíticas, como celulases e proteases, de modo a maximizar o processo de limpeza (Hasan et al., 2006). Por fim, na indústria de biocombustíveis as lipases são empregadas na hidrólise de triglicerídeos como uma alternativa ao processo de transesterificação alcalina na produção do biodiesel. Este processo possui diversas desvantagens, como a formação de sabão (saponificação), que dificulta o processo de purificação, o que tornou as enzimas uma alternativa atrativa (Ribeiro et al., 2011).

3 HIPÓTESE E OBJETIVOS

3.1 Hipótese

A combinação de diferentes estratégias de bioinformática no estudo da genômica de bactérias do gênero *Xanthomonas* propicia a prospecção e caracterização de genes associados ao processo de patogênese.

3.2 Objetivos Gerais

- Analisar o pangenoma e os genes diferencialmente distribuídos entre genomas núcleo dos patovares *oryzae* e *oryzicola* de *X. oryzae* de modo a identificar possíveis características moleculares relacionadas à patogênese.

- Desenvolver uma ferramenta para análise de genes da família TALE que permita a identificação destes genes em um genoma recém sequenciado e seus respectivos alvos no promoteroma de uma planta alvo.

- Sequenciar, montar, anotar e analisar o genoma de uma cepa local de *Xanthomonas* e identificar características funcionais estruturais relacionadas à patogênese e que as difiram em relação a cepas de referência.

3.3 Objetivos Específicos

De modo a alcançar seus objetivos gerais, a presente tese foi subdividida em três sub-projetos.

3.3.1 Sub-projeto 1

- Obter os dados genômicos de cepas dos patovares *oryzae* e *oryzicola* de *X. oryzae*.
- Re-anotar os genomas.
- Identificar grupos ortólogos no pangenoma e caracterizar sua distribuição usando diferentes algoritmos.
- Analisar o grupo de similaridade entre os genomas dos diferentes patovares.
- Identificar genes diferencialmente distribuídos entre os patovares.
- Identificar processos biológicos associados com os genes diferencialmente distribuídos e analisar seu potencial papel no processo de patogênese.

3.3.2 Sub-projeto 2

- Codificar a aplicação para análise dos TALEs.
- Validar a aplicação usando datasets de expressão gênica.

3.3.3 Sub-projeto 3

- Obter o DNA genômico de uma cepa local de *Xanthomonas*.
- Sequenciar o genoma usando NGS.
- Montar e anotador o genoma.
- Realizar uma análise filogenômica, estrutural e funcional de modo a identificar características relacionadas à patogênese a partir de cepas de referência.

4 CAPÍTULOS

Artigo 1

Formatação: *BMC Genomics* (Fator de impacto: 3,730).
Status do manuscrito: finalização.

Title: A pangenomic analysis of the *X. oryzae* patovars *oryzae* and *oryzicola* as a tool to identify pathovar-specific adaptations

Running Title: Pangenomic analysis of *X. oryzae*

Frederico Schmitt Kremer [1], Amanda Munari Guimarães, Luciano da Silva Pinto [1#]

[1] Laboratório de Bioinformática e Proteômica, Programa de Pós-Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Capão do Leão, Rio Grande do Sul, Brazil

address of correspondence to: ls_pinto@hotmail.com

Abstract

Xanthomonas oryzae is a Gram-negative phytopathogenic bacteria that infects rice (*Oryza sativa*) and is sub-classified in two main pathovars: *oryzae* (*Xoo*), the etiological agent of bacterial blight (BB) and *oryzicola* (*Xoc*), the etiological agent of bacterial leaf streak (BLS). The dynamic of infection, pathogenesis and symptomatology in these two diseases is very different, and comparative studies has been applied to better understand the genomic features that are responsible for each phenotype. In the present work we describe the application of a pangenome-based approach to identify functional features conserved in the genomes of strain from each pathovar, aiming a better description of these functional elements and the pathogenesis process of BB and BSL. Our analysis focused on the identification of genes differentially distributed among the core genome of these two pathovars, as a way to identify genomic features that are characteristic in one pathovar but not in the other.

Keywords: Comparative Genomics; Rice; Phytopathogens; Bacterial Blight; Bacterial Leaf Streak

Full Text

Introduction

The *Xanthomonas* genus comprises Gram-negative bacteria, most of which are phytopathogens and infect plants of high economic value. Species from this genus are usually sub-classified into pathological variants (pathovars, “pv.”) according to their host specificity, tissue that it infects and the disease that it causes. *Xanthomonas oryzae* is phytopathogenic bacteria that infects rice (*Oryza sativa*), been one of the major threads in the rice production worldwide. The species is sub-classified in two patovars, *oryzae* and *oryzicola* (Vauterin *et al.*, 1995). *X. oryzae* pv. *oryzae* (*Xoo*), which causes the bacterial blight (BB), is one of the major threads in the rice production in world-scale, but specially in Asia and Africa, where it was reported as the responsible for loses ranging from 20 to 80% in the production during epidemics. Additionally, although that in relatively smaller scale, *X. oryzae* pv. *oryzicola* (*Xoc*) has also been reported as responsible for expressive losses in the rice production, which might range from 5 to 30% during epidemics (Niño-Liu *et al.*, 2006). Although this species were not reported in all regions of world, such as South America, their study is also relevant for the development of detection and preventive methods that may help countries in these areas to avoid future cases, and some of them, such as Brazil, are major producers and exporters of rice.

Although *Xoo* and *Xoc* belong to the same species and infect the same host, they have very distinct infection dynamics and mechanisms of pathogenesis. BB is a vascular disease that affects the leaves of the plant. The bacteria starts the infection by entering the host usually through the hydrathodes, colonizes the intracellular spaces of the epidermis and achieve the xylem, from where it interacts with the parenchyma and moves vertically to the leaves trough the primary veins. The bacteria also migrate through the commissural veils (Ou, 1985). With a few days, the *Xoo* colonizes the vascular tissue of the leaf, which becomes full of bacteria cells and extracellular polysaccharides (EPS), resulting in the observed symptoms of beads or exudate in the leaf surface. The infection can be detected at the tillering stage, when small green water-soated spots starts to

709 appear in the leaves. The spots expand along the veins, becoming necrotic and lead to
710 the death of the tissue, which becomes white and grey (Curtis, 1943; Mew, 1993; Noda
711 and Hisatoshi, 1999). Different from *Xoo*, *Xoc* invades the plant mainly through the
712 stomata, multiplies in the sub-stomatal cavity and colonizes the parenchyma (Ou, 1985).
713 The early stages of the infection include the formation of water-soaked lesions with yellow
714 tone along the leaves, which are delimited by the veils. As BLS disease progresses, the
715 leaves turn grey and eventually die (Mew, 1993; Nyvall, 1999).
716
717 Comparative studies of *Xoo* and *Xoc* based on microarray (Seo *et al.*, 2008), DNA
718 hybridization (Soto-Suárez *et al.*, 2010) and whole genome-based screening of markers
719 (Feng *et al.*, 2015) have already demonstrated the presence of genomic features that
720 may be used to discriminate these two pathovars. However, this difference is still poorly
721 understood in terms of functional elements in the genome and how they lead to different
722 phenotypes and infection dynamics.
723
724 The advent of Next Generation Sequencing (NGS) has led to an exponential
725 growth in the volume of genomic data available in public databases, such as GenBank
726 (Shendure and Ji, 2008), which allowed the development of multiple-genome comparative
727 studies, such as pangenome and core-genome analysis. The pangenome consists in the
728 whole set of orthologue genes present in a given taxon, such as genus or species, and
729 comprises a wide range of sub-sets, such as core-genome (genes present in all, or at
730 least 95% of the strains), accessory genome (genes present in many, but not all strains)
731 and strain specific genes (present only in one or few strains) (Rouli *et al.*, 2015). The
732 applications of the pangenome analysis has provided ways to identify conserved and
733 discriminatory functional elements that may explain phenotypic differences among
734 pathogenic bacteria, including *Escherichia coli* (Rasko *et al.*, 2008), *Corynebacterium*
735 *pseudotuberculosis* (Soares *et al.*, 2013) and *Leptospira* spp (Fouts *et al.*, 2016). Along
736 with a better understanding of the pathogenesis, pangenomics may also provide
737 important insights regarding proteins with biotechnological applications, such as
738 cellulases and other cell-wall degrading enzymes that may be explored by the industry.
739 Therefore, aiming a better understanding of the molecular features that differentiate these

two pathovars at genomic level, here we report the application of a pangenome analysis for the identification of differentially distributed genes among the core genomes of *Xoo* and *Xoc*, and a prospection of genes associated with the pathogenesis processes in these sub-sets.

Material and Methods

Data collection and Reannotation

The complete genome records of 133 strains of *X. oryzae* were downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank), comprising 119 strains of *Xoo*, 16 of *Xoc* (Table 1). As different annotation pipelines may produce slightly different results, with different levels of reliability, all records were re-annotated using the same pipeline, Genix (Kremer *et al.*, 2016). The original annotations were compared to the new ones using an in house Python script, and annotation discrepancy was calculated based on metrics we suggested previously (Kremer *et al.*, 2016).

Calculation of the average nucleotide identity

The average nucleotide identity (ANI) was calculated using the in house Python script "getANI" (https://github.com/biopro/getANI). Based on the genes identified by Genix, a BLAST all vs. all search was performed using BLASTn (Altschul *et al.*, 1990) from the NCBI-BLAST+ (Camacho *et al.*, 2009) package based on the DNA sequence of protein coding genes, and for each pairwise search, the reciprocal best hit were selected using a minimum identity and coverage threshold of 70%. Based on the ANIs identified for each pair of genomes, a heatmap was generated using the Python Seaborn package (https://seaborn.pydata.org/) with the Minkowski distance measure.

Phylogenetics

CDSs from the *X. oryzae* consensus core genes were filtered and only those ortholog groups present in all strains were selected. Them, multiple sequence alignment was performed for each gene using MUSCLE (Edgar, 2004) and the results were filtered using TrimAl (Capella-Gutiérrez *et al.*, 2009). The processed alignments were merged into a single file and a phylogenetic tree was constructed using the UPGMA algorithm implemented in MEGA X (Kumar *et al.*, 2018) with 10000 steps of bootstrap. The final tree was plotted using the Interactive Tree Of Life (iTOL) web application (Letunic and Bork, 2016).

Identification of orthologue groups

Orthologue groups where identified using GET_HOMOLOGUES (Contreras-Moreira and Vinuesa, 2013), which was executed using the tree different algorithms that are available: BDBH, OrthoMCL and COG triangles. For each cluster in the selected analysis a representative sequence was extracted using CD-HIT (Li and Godzik, 2006), and a multiple sequence alignment (MSA) using MUSCLE (Edgar, 2004). For each algorithm, the result of the predicted groups were processed using PanGP (Zhao *et al.*, 2014) using the “distance guided” method in order to verify the distribution of the pangenome and core genome in the species and in each pathovar, and their respective fitting functions.

Based on the orthologue groups that were identified for each algorithm, those groups present in at least 95% of the strains of each pathovars were considered as part of the core-genome of the specie, and this same threshold was applied to select the core-genome of each pathovar. Finally, those groups present in at least 95% of the strains of a certain pathovar but not in the other were considered differentially distributed.

Functional annotation and metabolic analysis

The representative sequence of each orthologue groups identified by each algorithm were aligned to the Uniprot trEMBL database (http://www.uniprot.org/) using BLASTn from the NCBI-BAST+ (Altschul *et al.*, 1990; Camacho *et al.*, 2009) package. The Uniprot ID from the best hits, when presenting and e-value < 1e-10, was searched against the QuickGO server (Binns *et al.*, 2009) to retrieve Gene Ontology (GO) terms for their Uniprot-GOA annotation (https://www.ebi.ac.uk/GOA). Orthologue groups were also annotated using InterproScan 5. Functional annotation was also performed using RPS-BLAST searches against the Cluster of Ortholog Groups (COG) database (Tatusov, 2000) distributed along with the Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2014) by NCBI (https://www.ncbi.nlm.nih.gov/cdd/). Metabolic analysis based on the KEGG database of pathways (Kanehisa and Goto, 2000) was performed using BlastKOALA (Kanehisa *et al.*, 2016) for the core genomes calculated for *X. oryzae*, *X. oryzae* pv. *oryzae* and *X. oryzae* pv. *oryzicola*, and results were refined using the MinPaths parsimony method (Ye and Doak, 2009).

Analysis of conservation

For each orthologue group identified by the different algorithm the resulting MSA was processed using an in house python script, “clean-alignment.py” (https://github.com/biopro/clean-alignment), to remove outlier sequences and other artifacts. Briefly, it reads the MSA file in FASTA format, calculates a score for each sequence in the alignment using EvalMSA (Chiner-Oms and González-Candelas, 2016), removes those with a score smaller than a certain threshold (0.75 in this case), and finally removes any columns composed only by gaps using TrimAl (Capella-Gutiérrez *et al.*, 2009). Based on the processed MSAs, the variability of each orthologue group was calculated using the Shannon entropy by calculating the mean entropy of the alignment for each position for the two pathovars separately, and then extracting the average entropy for each alignment for *Xoo*, *Xoc* and the species (the mean of the *Xoo* and *Xoc* average entropy). For each cluster dataset, the mean entropy of all orthologue groups

was converted to z-scores, which were used to assign a respective *p*-value to the degree of variability of each group, and the p-values were corrected using the Bonferroni method. Groups with z-score positive and *p*-value < 0.05 in the two-tailed test were considered significantly variable.

Functional enrichment analysis (GO)

For each of their subsets of ortholog groups identified by each algorithm, a GO enrichment analysis was performed using the Fischer exact test implemented in the SciPy package (https://www.scipy.org/), and results were corrected using the Bonferroni method. For each case, GO terms were considered as enriched when the relative abundance in each subset were higher than the abundance in the complete set of genes, and the *p*-value was < 0.05 in the two-tailed test. This same approach was also applied to identify GO terms enriched in the sets of genes in the core genome identified as differentially conserved among the different datasets.

Results and Discussion

Genome re-annotation

The results from the comparison of the original annotation to the re-annotated records are displayed in Table 2. In the present analysis, the annotation discrepancy ranged from 8.56 (*X. oryzae* pv. *oryzae* str. MAFF 311018) to 17.09 % (*X. oryzae* pv. *oryzae* str. LMG 9585), indicating that the different procedures adopted in the annotation of these genomic records might reflect in a large portion of these genomes, and consequently, in the genes that will be used in the characterization of the pangenome. Although gene finding programs (for prokaryotes) usually present a high accuracy (in most cases > 95%), there are many factors intrinsically (eg: regions with high differences of CG% such as genomic islands, genes with programmed frameshift, short or very long genes and spurious ORFs) and extrinsically associated (eg: algorithm's heuristics and threshold settings, assembly errors, sequencing artifacts) to the genome that may affect

the annotation result. Additionally, as a wide-variety of automated genome annotation pipeline are available for prokaryote genomes (Aziz *et al.*, 2008; Van Domselaar *et al.*, 2005; Kremer *et al.*, 2016; Seemann, 2014; Tatusova *et al.*, 2016), both for local and web use, how complete the annotation is obtained will depends on many factors, such as the selection of the appropriate databases; each research group will select those that fits better in its lab's routine, or even in its specific research needs. Therefore, when performing a comparative genomics analysis based on data submitted from different research groups, it is important to take into account these factors, and the re-annotation might be a key point before the actual characterization of the pangenome. Other studies have already considered this point instead of just taking the records directly from GenBank.

ANI analysis

The result of the correlation matrix generated by Seaborn using the of the Minkowski (Figure 1). As demonstrated by the pairwise ANI comparison, the intra-pathovar sequence identity in genes from the orthologues identified among each pair of strains is usually higher than 99.3 %, while the identity between strains from *Xoo* and *Xoc* is usually < 98.7%. However, *X. oryzae* strains were clustered into 3 distinct groups, corresponding to those Asia, Africa and America, whose distances to each other are similar to the distance among the strains from pathovar *oryzicola*, supporting previous suggestions on the molecular profile of American *Xoo* (Triplett *et al.*, 2011).

One of the cluster formed includes the African strains of *Xoo* NAI8 515 (from Niger) (GenBank: AYSX01000000.1) and AXO1947 (from Cameroon) (GenBank: CP013666.1), which corroborates with the findings already reported for AXO1947 (Huguet-Tapia *et al.*, 2016), and indicates that they might be as distant from *Xoo* as they are from *Xoc*, although presenting similarities in the pathogenesis and symptoms, at genomic level. The same thing was observed for the American strains of *Xoo*, X8-1A (GenBank: AFHL01000000) and X11-5A (GenBank: LHUJ01000000.1), already suggested as a different pathovar, but also reported as etiological agent of BB. It is interesting that in the case of *Xoc*, the

892 genomic and evolutionary distance between African and Asian strains is also evident
893 when analyzing only the ANI matrix and is in accordance with previous works have
894 already demonstrated that these lineages present clear divergences at genomic and
895 molecular level.
896
897 It was previously reported that African and Asian strains of *Xoc* present different
898 sets of TALE effectors (and their respective targets) (Wilkins *et al.*, 2015), which illustrates
899 the adaptation of the pathogen to a different environment and host's genotypes. Finally,
900 previous studies also demonstrated that African strains of *Xoc*, as exemplified by the
901 strains obtained in Burkina Faso, have a high genomic variability, so not only they are
902 different from other pathovars, or even strains from the same pathovar but from different
903 regions, but also at certain level from themselves. Therefore, although the analysis of the
904 ANI results does not point a genomic distance as higher as the observed in *Xoo*, *Xoc*
905 strains from distant geographic locations might also exemplify the genomic-level effect of
906 isolation and differential evolutionary pressure in the pathogen, and this information might
907 corroborate for a better understanding of the evolution and adaptation of these
908 pathogens. Interestingly, the race-level classification, usually applied to strains of *X.*
909 *oryzae*, is not clear observed using the ANI representation, as exemplified by the Asian
910 strains of *Xoo*, which present a quasi-equal pattern in the final matrix.
911
912 *Phylogenetics*
913
914 The UPGMA phylogenetic tree generated from the alignment of the core genes
915 identified in the genome of strains of *X. oryzae* pv. *oryzae*, *X. oryzae* pv. *oryzicola* and in
916 the outgroup *X. campestris* is displayed in Figure 2. The final tree could clearly
917 discriminate the different geographical groups of *X. oryzae* and *X. oryzicola*, and also
918 points to higher evolutionary distance of the American strains of *Xoo* to the other strains
919 of *Xoo* or even *Xoc* than the observed in the ANI-based analysis. Although the UPGMA
920 method is usually avoided when constructing trees based on short fragments of DNA or
921 protein sequence and other methods, such as Neighbor-Join (NJ) or Maximum-Likelihood

(ML) as preferred, it has already been demonstrated as appropriate in the construction of whole-genome alignment-based phylogenetic trees (Klima *et al.*, 2016; Lees *et al.*, 2018).

Interestingly, a recent analysis of the phylogenetics and adaptation of *X. oryzae* pv. *oryzae* (Midha *et al.*, 2017) has already suggested that *Xoc* might represent a variation of *Xoo*, and not a distinct pathovar, due to the same distribution of the strains of *Xoo* and *Xoc* observed in our results. However, this distribution is particularly a consequence of the particular characteristics of the American strains of *Xoo*, and therefore may only enforce that idea that it might represent, in fact, a distinct pathovar. In a same way, our analysis shows that Asian and African strains of *Xoo* may be clustered in four major groups (1 African and 3 of Asian), while *Xoc* can be organized in 3 major groups (1 African and 2 Asian). Race level classification, however, is unclearly distinguished through core-genome-based phylogenomics analysis, and strains for different lineages were clustered in the same group.

Pangenome analysis

The distribution of the pangenome and core genome sizes of *Xoo* and *Xoc*, based on the different algorithms, is displayed in Figure 3. As demonstrated, the tendency of the cumulative distribution of the number of orthologue groups, using different sub-samplings, in the pan and core genome vary drastically depending on the algorithm that was is used. In the case of COG triangles, for example, the number of groups was expressively higher both for *Xoo* and *Xoc*, when comparing to the results obtained using OMCL and BDBH. Moreover, in this case, the increase in the sampling size did not lead to an equilibrium in the final count of orthologue groups, indicating that a strict criteria was applied to group sequenced into the same group. The over-fragmentation of the pangenome as a consequence of the lack of tolerance in the algorithm during the sequence clustering might hinder the characterization of conserved group; thus, in this particular case, COG may not be an appropriate algorithm. In part, it is possible that higher fragmentation was a consequence of genes that may fit into different orthologue groups, and as the algorithm tries to consistently arrange sequences following a consistent criteria, the presence of

ambiguities might avoid the appropriate clustering and results in a higher number of groups, as already reported (Kristensen *et al.*, 2010). When considering the results generated by OMCL and BDBH, on the other hand, it is possible to see that, as more genes are considered in the sampling, smaller is the probability of a new groups of orthologue genes appears, which might indicate a closed pangenome content in both pathovars (Rouli *et al.*, 2015).

The number of orthologue groups identified as part of the different sub-sets of *X. oryzae* and the pathovars *oryzae* and *oryzicola*, based on the different algorithms, is displayed in Table 3. The ratio core-genome / pangenome (Table 4), which might also be used to evaluate the "closeness" of the pangenome (Rouli *et al.*, 2015), widely varies depending on the algorithms, as demonstrated early, and reinforces the impact of the algorithm choice in the modelling of the pangenome. However, it is also important to note that the number of orthologue groups in the core genome sets did not varied between the algorithms as much as in the pangenome sets, and for both pathovars the results obtained using BDBH, OMCL and COG were, till a certain point, consistent.

As observed by analyzing the differential distribution of orthologue groups, we were able to demonstrate that the pathovar *Xoo* have a larger set of conserved genes when compared to *Xoc*, although the core genome of *Xoc* also contains some groups of genes absent in the *Xoo* core genome too. Although the size of these subsets (in this work named "exclusive core genomes") may vary depending on the algorithm, it illustrates the degree difference at genomic level that might reflect the adaptation of these two pathovars, and the analysis of these genes might provide useful information regarding the characterization of the molecular factors associated with the difference in the molecular pathogenesis of BB and BSL. The sequences of the clusters identified in each sub-set by algorithm are available in the GitHub repository https://github.com/biopro/x.oryzae-pangenomic-analysis/.

Functional analysis, metabolic analysis and gene conservation

The analysis of the distribution of COG categories in the core genome of *X. oryzae*, *X. oryzae* pv. *oryzae* and *X. oryzae* pv. *oryzicola* is presented in Figure 4, and the comparison of these sub-sets based on the KEGG pathways identified using BlastKOALA and MinPaths is presented in Figure 5. BlastKOALA was able to annotate 56.74% of the core genome of *X. oryzae*, 52.12% of the core of *Xoo* and 55.56% of the core of *Xoc*. Based on the analysis of BlastKOALA, the higher difference was observed in the number of identified enzymes present in the core-genome. Some of these enzymes, as further discussed, were identified as associated with pathogenesis processes, such as cellulase and lipase activity, which are important functions on many phytopathogens (Have *et al.*, 2002; Subramoni *et al.*, 2010). Additionally, the results points that the major differences between *Xoo* and *Xoc* in terms of gene content are proteins with no correspondence in these two databases (COG and KEGG). From the 592 clusters at belong the exclusive core of *Xoo*, only 206 presented hits on KEGG. When comparing these results with those from analysis of GO enrichment based on the differentially conserved proteins (Table 7), however, it is possible that the difference between *Xoo* and *Xoc* at metabolic and housekeeping level may also be caused by mutations in primary biological processes, as already demonstrated for other species, such as in the adaptation of *Salmonella* serovars (Felten *et al.*, 2017).

Exclusive core genome Xoo core and Xoc core

When analyzing the differences in the core genome of *Xoo* and *Xoc*, we were able to identify 576 genes that are conserved in *Xoo* but not in *Xoc* and 106 that are conserved in *Xoc* but not in *Xoo*. These groups of genes were called “exclusive core genome”, although these genes might be present in the other pathovar, they are only conserved in 95% of the strains of only one of them. A complete list of these proteins along with their functional annotation from InterProScan (Zdobnov and Apweiler, 2001) is available in Supplementary Data 1.A (for *Xoo*) and 1.B (for *Xoc*).

The identification of genes conserved in one pathovar and absent in the other is important in the characterization of the pathogenesis processes and adaptation of each pathovar and also might be explored as potential molecular markers, and the explored in the development of diagnostic methods based on DNA amplification, such as traditional PCR (Song *et al.*, 2014) and Loop-mediated isothermal amplification (LAMP) (Larrea-Sarmiento *et al.*, 2018). The development of this techniques might be particularly important in regions were both pathovars are endemic, such as Asia and Africa, and specially to quickly distinguish these pathovars in late stages of the infection. Finally, the identification of conserved genes in each pathovars might allow the development of transgenic-based strategies that aim the obtaining of resistant cultivars of *O. sativa*, in a similar way than previous studies with other species (Horvath *et al.*, 2012).

Membrane proteins

A wide variety of outer-membrane proteins (OMPs) were identified in the exclusive core both of *Xoo* and *Xoc*. In *Xoo*, the differential analysis revealed that genes that encode a TonB dependent receptor (cluster:218253) and two efflux pumps (cluster:220850 and cluster:220212) are preserved in its core genome. In the case of *Xoc*, other two genes from two outer membrane efflux proteins (cluster:219575 and cluster:220511) were identified as differentially distributed, along with a sugar transporter (cluster:218425) and a sodium/hydrogen exchanger (cluster:220172).

TonB dependent transporters have already been demonstrated as required for the pathogenesis of many phytopathogens, including *Rastonia solanaceacum*, *X. axonopodis* pv. *citri* and *X. campestris* pv. *campestris* (Aini *et al.*, 2010; Brito *et al.*, 2002; Vorhölter *et al.*, 2012). These proteins are located in the outer-membrane of the bacteria and usually act as transporters and / or coupled to signal transduction systems, and their role is particularly important in the surviving of the pathogen in certain niches, such as the xylem, which is mainly constituted by dead cells and has low availability of nutrients (Fatima and Senthil-Kumar, 2015). In the case of cluster:218253, a manual analysis indicated that it is a Vitamin B12 transporter BtuB-like family, which also includes iron-uptaking

transporters (InterPro: IPR039426), and iron was also pointed as the preferable ligand by Phyre2 (Kelley *et al.*, 2015). Iron is a vital nutrient for *Xanthomonas* and the maintenance of its homeostasis is a crucial process for many pathogens (Cassat and Skaar, 2013). The need of its uptake even in environments with low availability of nutrients, such as the vascular system, may be one the evolutionary reason to the high conservation of this transporters in *Xoo*.

Cluster:220850 encodes a multidrug efflux pump, which is associated with the efflux of molecules that may be harmful to the bacteria, including antibiotics and certain metabolites. Cation efflux proteins, such as cluster:220212, play important roles in the resistance to certain metal ions, such as copper, and their acquisition through some events of horizontal transferring was already reported as responsible for a higher resistance to it (Ryan *et al.*, 2007).

A higher tolerance to salinity, as conferred by sodium/hydrogen exchanger, such as cluster:220172, was also already linked to pathogenesis in animal and plant pathogens (Kosono *et al.*, 2005). In the case of BSL, it may be particularly important as sodium tends to accumulate in the leaves during salt stress (Yeo and Flowers, 1982), and the stomata, located in the leaves, are one of the preferential sites of invasion in *Xoc*.

Xoc contains a conserved gene that provides resistance to bacteriocins

Bacteriocins are ribosomal synthesized antimicrobial peptides secreted by some bacteria which can kill other close-related or non-related strains but would not harm the producing bacteria due to specific immunity proteins (Yang *et al.*, 2014). A putative bacteriocin-protecting protein, which is encoded by the “cluster:219578”, was identified in *Xoc* exclusive core, and although absent in the final set of *Xoo* core genome, is also present in 97 strains of *Xoo* including the two American strains. The annotation of this genes indicates that contains a “YdeI/OmpD” domain (Pfam:PF13376) which is associated with resistance to Polymyxin B (Pilonieta *et al.*, 2009), a bacteriocin produced by *Paenibacillus polymyxa*, a nitrogen fixing Gram-negative bacteria that is usually found

in soil and associated with the plants roots (Yegorenkova *et al.*, 2013). Polymyxin B has emerged as a promising candidate for antibiotics development due to the rise of multi-drug resistance to traditional antibiotics (Yuan and Tam, 2008). However, the identification of resistance genes in phytopathogens also represent a risk to public health, as they may be transferred through horizontal gene transferring to animal and human pathogens (Wright, 2010). Therefore, the identification of these genes in the core genome of *Xoc* not only brings new information regarding the pathogen itself, but also some concerns regarding the eventual use of some bacteriocins in clinical applications.

Cellulases, Pectinases and Lipases are differentially distributed in *Xoo* and *Xoc*

We also have identified two ortholog groups associated with potential cellulase activity, one in the exclusive core of *Xoo* (cluster:218867) and another in the exclusive core of *Xoc* (cluster:218163). Many of cell-wall degrading proteins have already been characterized in *X. oryzae*, including endoglucanase (ClsA), cellobiohydrolase (CbsA), xylanase (XynB) and lipase A (LipA).“cluster:218867” comprises the *cbsA* gene, which encodes the CbsA cellobiohydrolase, a cell wall hydrolase secreted from *Xoo* and other pathogens that colonize the xylem, and that is already reported as absent or non-functional in pathogens that infect the parenchymatous tissue, including *Xoc* (Ryan *et al.*, 2011). The production of secreted cellulases and other proteins able to degradate the cell wall is usually required for the full pathogenesis of many phytopathogens, including those from the genera *Xanthomonas* (Xia *et al.*, 2016) and *Pseudomonas* (Roberts *et al.*, 1988).These proteins, however, usually stimulate the immunity of the plant, and in the case of *Xoo* eventually lead to a higher resistance of the plant in further infections (Tayi *et al.*, 2018). In the case of cluster:218163, however, although many of the genes from this cluster are already annotated as a putative hydrolase in public databases, such as xoc_1076 from *X. oryzae* pv. *oryzicola* str. BLS256 (GenBank: CP003057.2), there is still a lack of information regarding its role in the pathogenesis processes during BSL.

Another group of genes potentially encodes a pectinesterase (cluster:218594), which is responsible for the degrading of pectin, a component of the plant cell wall.

Interestingly, this gene showed no orthologs in *Xoc*, although present in other species of *Xanthomonas* according to a BLAST search against the UniProt database, including *X. campestris* (UniProt: A0A0M1KY28), *X. pisi* (UniProt: A0A2S7D1F9), *X. arboricola* (UniProt: A0A2S6YYE9) and *X. prunicola* (UniProt: A0A2N3RHW0). The identification and characterization of pectin-degrading enzymes in *Xoo* is relatively recent, and although they have been large reported as playing a minor role in the pathogenesis when comparing to other cell wall degrading proteins such as cellulases (Tayi *et al.*, 2016), the presence of this carbohydrate in the tissues and organs that are infected by *Xoo*, such as the vascular system, where it is observed not only in the cell wall but also in the inner of the secondary xylem (Yadeta and J Thomma, 2013), might indicate why they are conserved in this pathovar.

Lipases, such as lipase A (LipA), were already reported as secreted by *Xanthomonas oryzae*, and although their role in the pathogenesis process is poorly understood, they were already demonstrated as required for full virulence and able to stimulate defense response in *O. sativa*. Along with its role in the understanding of the pathogenesis process, the identification of lipases in *Xanthomonas* has, as they present thermostability, turning this enzyme in a promising candidate for industrial applications (Aparna *et al.*, 2007; Mo *et al.*, 2016). In the pangenome analysis, we have identified a gene cluster (cluster:220899) exclusively conserved in *Xoo* which encodes a LipA ortholog, indicating that this enzyme might be required for the pathogenesis during BB, as already demonstrated in previous studies, but not necessarily during BSL.

The CRISPR system is conserved in *Xoo*, but not in *Xoc*

Most of the genes from the CRISPR/Cas system were identified as conserved in *Xoo* but not in *Xoc*, including Cas1(cluster:218903), Cas2 (cluster:218902), Cas5 (cluster:218907) and Cas7 (cluster:218905), which is in accordance with the early studies that showed that the CRISPR system in *X. oryzae* is conserved in Asian strains of *Xoo*, but absent in African and American strains of *Xoo* and Asian strains of *Xoc* (Midha *et al.*, 2017). As our dataset is majorly constituted by Asian strains, however, the system was

identified as conserved in the whole pathovar. The CRISPR/Cas system is widely distributed in bacteria and acts as a defense mechanism against phages, but also was already linked to the pathogenesis processes in many species due to its potential role in the regulation of gene expression during certain stress conditions (Louwen *et al.*, 2014). In the case of *Pseudomonas aeruginosa*, for example, the expression of a certain spacer of the CRISPR leads to the inhibition of the biofilm formation and regulation of the cell motility (Zegans *et al.*, 2009). In other species, it was already demonstrated as also required for pathogenesis, including those from the genera *Leptospira* (Fouts *et al.*, 2016) and *Salmonella* and the species *Yersinia Pestis* and *Escherichia coli*. Therefore, it is possible that some of the phenotypic differences between *Xoo* and *Xoc* are due to gene expression regulations by the CRISPR system.

Conclusion

Here we reported that characterization of the *X. oryzae* pangenome and a subtractive analysis of the core genome of *Xoo* and *Xoc* aiming the identification of molecular features that might provide a better understanding of the phenotypic differences observed between these two pathovars. The sets of genes identified by us as differentially distributed might be furtherly used as reference in the studies of virulence factors of *Xoo* and *Xoc* and provides to researchers in the field not only new insights in pathogenesis itself, but also a framework to evaluate the identification of pathovar-specific genes, as the methods described in this paper might be furtherly applied for more species of *Xanthomonas* or other bacteria species.

References

Aini, L.Q., Hirata, H. and Tsuyumu, S. (2010) A TonB-dependent transducer is responsible for regulation of pathogenicity-related genes in *Xanthomonas axonopodis* pv. *citri*. *J. Gen. Plant Pathol.* **76**, 132–142.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10.

Aparna, G., Chatterjee, A., Jha, G., Sonti, R. V. and Sankaranarayanan, R. (2007) Crystallization and preliminary crystallographic studies of LipA, a secretory lipase/esterase from *Xanthomonas oryzae* pv. *oryzae*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **63**, 708–710.

Aziz, R.K., Bartels, D., Best, A.A., et al. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C. and Apweiler, R. (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–6.

Brito, B., Aldon, D., Barberis, P., Boucher, C. and Genin, S. (2002) A Signal Transfer System Through Three Compartments Transduces the Plant Cell Contact-Dependent Signal Controlling *Ralstonia solanacearum hrp* Genes. *Mol. Plant-Microbe Interact.* **15**, 109–119.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–3.**Cassat, J.E. and Skaar, E.P.** (2013) Iron in Infection and Immunity. *Cell Host Microbe* **13**, 509–519.

Chiner-Oms, A. and González-Candelas, F. (2016) EvalMSA: A Program to Evaluate Multiple Sequence Alignments and Detect Outliers. *Evol. Bioinforma.* **12**, EBO.S40583.

Contreras-Moreira, B. and Vinuesa, P. (2013) GET_HOMOLOGUES, a versatile

software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* **79**, 7696–701.

Curtis, L.C. (1943) Deleterious Effects of Guttated Fluids on Foliage. *Am. J. Bot.* **30**, 778.

Domselaar, G.H. Van, Stothard, P., Shrivastava, S., et al. (2005) BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res.* **33**, W455-9.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–7.

Fatima, U. and Senthil-Kumar, M. (2015) Plant and pathogen nutrient acquisition strategies. *Front. Plant Sci.* **6**, 750.

Felten, A., Vila Nova, M., Durimel, K., Guillier, L., Mistou, M.-Y. and Radomski, N. (2017) First gene-ontology enrichment analysis based on bacterial coregenome variants: insights into adaptations of *Salmonella* serovars to mammalian- and avian-hosts. *BMC Microbiol.* **17**, 222.

Feng, W., Wang, Y., Huang, L., Feng, C., Chu, Z., Ding, X. and Yang, L. (2015) Genomic-associated Markers and comparative Genome Maps of Xanthomonas oryzae pv. oryzae and *X. oryzae* pv. *oryzicola*. *World J. Microbiol. Biotechnol.* **31**, 1353–1359.

Fouts, D.E., Matthias, M.A., Adhikarla, H., et al. (2016) What Makes a Bacterial Species Pathogenic?:Comparative Genomic Analysis of the Genus *Leptospira* Small, P.L.C., ed. *PLoS Negl. Trop. Dis.* **10**, e0004403.

Have, A. ten, Tenberge, K.B., Benen, J.A.E., Tudzynski, P., Visser, J. and Kan, J.A.L. van (2002) The Contribution of Cell Wall Degrading Enzymes to Pathogenesis of Fungal Plant Pathogens. In Agricultural Applications., pp. 341–358. Berlin, Heidelberg: Springer Berlin Heidelberg.

Horvath, D.M., Stall, R.E., Jones, J.B., Pauly, M.H., Vallad, G.E., Dahlbeck, D., Staskawicz, B.J. and Scott, J.W. (2012) Transgenic Resistance Confers Effective Field Level Control of Bacterial Spot Disease in Tomato Park, S., ed. *PLoS One* **7**, e42036.

Huguet-Tapia, J.C., Peng, Z., Yang, B., Yin, Z., Liu, S. and White, F.F. (2016) Complete Genome Sequence of the African Strain AXO1947 of *Xanthomonas oryzae* pv.

oryzae. *Genome Announc.* **4**, e01730-15.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30.

Kanehisa, M., Sato, Y. and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J.E. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–58.

Klima, C.L., Cook, S.R., Zaheer, R., et al. (2016) Comparative Genomic Analysis of Mannheimia haemolytica from Bovine Sources Lin, B., ed. *PLoS One* **11**, e0149520.

Kosono, S., Haga, K., Tomizawa, R., Kajiyama, Y., Hatano, K., Takeda, S., Wakai, Y., Hino, M. and Kudo, T. (2005) Characterization of a multigene-encoded sodium/hydrogen antiporter (sha) from *Pseudomonas aeruginosa*: its involvement in pathogenesis. *J. Bacteriol.* **187**, 5242–8.

Kremer, F.S., Eslabão, M.R., Dellagostin, O.A. and Pinto, L. da S. (2016) Genix: A New Online Automated Pipeline for Bacterial Genome Annotation. *FEMS Microbiol. Lett.* **11**.

Kristensen, D.M., Kannan, L., Coleman, M.K., Wolf, Y.I., Sorokin, A., Koonin, E. V and Mushegian, A. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* **26**, 1481–7.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. (2018) MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol.*

Larrea-Sarmiento, A., Dhakal, U., Boluk, G., et al. (2018) Development of a genome-informed loop-mediated isothermal amplification assay for rapid and specific detection of Xanthomonas euvesicatoria. *Sci. Rep.* **8**, 14298.

Lees, J.A., Kendall, M., Parkhill, J., Colijn, C., Bentley, S.D. and Harris, S.R. (2018) Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a

simulation based study. *Wellcome open Res.* **3**, 33.

Letunic, I. and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242-5.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–9.

Louwen, R., Staals, R.H.J., Endtz, H.P., Baarlen, P. van and Oost, J. van der (2014) The Role of CRISPR-Cas Systems in Virulence of Pathogenic Bacteria. *Microbiol. Mol. Biol. Rev.* **78**, 74–88.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., et al. (2014) CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222-6.

Mew, T.W. (1993) Focus on Bacterial Blight of Rice. *Plant Dis.* **77**, 5.

Midha, S., Bansal, K., Kumar, S., et al. (2017) Population genomic insights into variation and evolution of *Xanthomonas oryzae* pv. *oryzae*. *Sci. Rep.* **7**, 40694.

Mo, Q., Liu, A., Guo, H., Zhang, Y. and Li, M. (2016) A novel thermostable and organic solvent-tolerant lipase from *Xanthomonas oryzae* pv. *oryzae* YB103: screening, purification and characterization. *Extremophiles* **20**, 157–165.

Nikaido, H. and Takatsuka, Y. (2009) Mechanisms of RND multidrug efflux pumps. *Biochim. Biophys. Acta - Proteins Proteomics* **1794**, 769–781.

Niño-Liu, D.O., Ronald, P.C. and Bogdanove, A.J. (2006) *Xanthomonas oryzae* pathovars: model pathogens of a model crop. *Mol. Plant Pathol.* **7**, 303–24.

Noda, T. and Hisatoshi, K. (1999) Growth of *Xanthomonas oryzae* pv. *oryzae* in planta and in guttation fluid of rice. *Ann. Phytopathol. Soc. Japan* **65**, 9–14.

Nyvall, R. (1999) Field Crop Diseases Third edit., Ames: iowa State University Press.

Ou, S.H. (1985) Rice diseases, Commonwealth Mycological Institute.

Pilonieta, M.C., Erickson, K.D., Ernst, R.K. and Detweiler, C.S. (2009) A protein important for antimicrobial peptide resistance, YdeI/OmdA, is in the periplasm and interacts with OmpD/NmpC. *J. Bacteriol.* **191**, 7243–52.

Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., et al. (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* **190**, 6881–93.

Roberts, D.P., Denny, T.P. and Schell, M.A. (1988) Cloning of the egl gene of *Pseudomonas solanacearum* and analysis of its role in phytopathogenicity. *J. Bacteriol.* **170**, 1445–51.

Rouli, L., Merhej, V., Fournier, P.-E. and Raoult, D. (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New microbes new Infect.* **7**, 72–85.

Ryan, R.P., Ryan, D.J., Sun, Y.-C., Li, F.-M., Wang, Y. and Dowling, D.N. (2007) An acquired efflux system is responsible for copper resistance in *Xanthomonas* strain IG-8 isolated from China. *FEMS Microbiol. Lett.* **268**, 40–46.

Ryan, R.P., Vorhölter, F.-J., Potnis, N., Jones, J.B., Sluys, M.-A. Van, Bogdanove, A.J. and Dow, J.M. (2011) Pathogenomics of *Xanthomonas*: understanding bacterium–plant interactions. *Nat. Rev. Microbiol.* **9**, 344–355.

Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–9.

Seo, Y.-S., Sriariyanun, M., Wang, L., et al. (2008) A two-genome microarray for the rice pathogens Xanthomonas oryzae pv. oryzae and *X. oryzae* pv. *oryzicola* and its use in the discovery of a difference in their regulation of hrp genes. *BMC Microbiol.* **8**, 99.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–45.

Soares, S.C., Silva, A., Trost, E., et al. (2013) The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar ovis and equi strains. *PLoS One* **8**, e53818.

Song, E.-S., Kim, S.-Y., Noh, T.-H., Cho, H., Chae, S.-C. and Lee, B.-M. (2014) PCR-based assay for rapid and specific detection of the new *Xanthomonas oryzae* pv. *oryzae* K3a race using an AFLP-derived marker. *J. Microbiol. Biotechnol.* **24**, 732–9.

Soto-Suárez, M., González, C., Piégu, B., Tohme, J. and Verdier, V. (2010) Genomic

comparison between *Xanthomonas oryzae* pv. *oryzae and Xanthomonas oryzae* pv. *oryzicola*, using suppression-subtractive hybridization. *FEMS Microbiol. Lett.* **308**, 16–23.

Subramoni, S., Suárez-Moreno, Z.R. and Venturi, V. (2010) Lipases as Pathogenicity Factors of Plant Pathogens. In Handbook of Hydrocarbon and Lipid Microbiology., pp. 3269–3277. Berlin, Heidelberg: Springer Berlin Heidelberg.**Tatusov, R.L.** (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36.

Tatusova, T., DiCuccio, M., Badretdin, A., et al. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624.

Tayi, L., Kumar, S., Nathawat, R., Haque, A.S., Maku, R. V., Patel, H.K., Sankaranarayanan, R. and Sonti, R. V. (2018) A mutation in an exoglucanase of *Xanthomonas oryzae* pv. *oryzae* , which confers an endo mode of activity, affects bacterial virulence, but not the induction of immune responses, in rice. *Mol. Plant Pathol.* **19**, 1364–1376.

Tayi, L., Maku, R. V., Patel, H.K. and Sonti, R. V. (2016) Identification of Pectin Degrading Enzymes Secreted by *Xanthomonas oryzae* pv. *oryzae* and Determination of Their Role in Virulence on Rice Wilson, R.A., ed. *PLoS One* **11**, e0166396.

Triplett, L.R., Hamilton, J.P., Buell, C.R., Tisserat, N.A., Verdier, V., Zink, F. and Leach, J.E. (2011) Genomic Analysis of *Xanthomonas oryzae* Isolates from Rice Grown in the United States Reveals Substantial Divergence from Known *X. oryzae* Pathovars. *Appl. Environ. Microbiol.* **77**, 3930–3937.

VAUTERIN, L., HOSTE, B., KERSTERS, K. and SWINGS, J. (1995) Reclassification of Xanthomonas. *Int. J. Syst. Bacteriol.* **45**, 472–489.

Vorhölter, F.-J., Wiggerich, H.-G., Scheidle, H., Sidhu, V.K., Mrozek, K., Küster, H., Pühler, A. and Niehaus, K. (2012) Involvement of bacterial TonB-dependent signaling in the generation of an oligogalacturonide damage-associated molecular pattern from plant cell walls exposed to *Xanthomonas campestris* pv. *campestris* pectate lyases. *BMC Microbiol.* **12**, 239.

Wilkins, K.E., Booher, N.J., Wang, L. and Bogdanove, A.J. (2015) TAL effectors and

activation of predicted host targets distinguish Asian from African strains of the rice pathogen *Xanthomonas oryzae* pv. *oryzicola* while strict conservation suggests universal importance of five TAL effectors. *Front. Plant Sci.* **6**, 536.

Wright, G.D. (2010) Antibiotic resistance in the environment: a link to the clinic? *Curr. Opin. Microbiol.* **13**, 589–594.

Xia, T., Li, Y., Sun, D., Zhuo, T., Fan, X. and Zou, H. (2016) Identification of an Extracellular Endoglucanase That Is Required for Full Virulence in *Xanthomonas citri* subsp. *citri* Wang, Z., ed. *PLoS One* **11**, e0151017.

Yadeta, K.A. and J Thomma, B.P.H. (2013) The xylem as battleground for plant hosts and vascular wilt pathogens. *Front. Plant Sci.* **4**, 97.

Yang, S.-C., Lin, C.-H., Sung, C.T. and Fang, J.-Y. (2014) Antibacterial activities of bacteriocins: application in foods and pharmaceuticals. *Front. Microbiol.* **5**, 241.

Ye, Y. and Doak, T.G. (2009) A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.* **5**, e1000465.

Yegorenkova, I. V., Tregubova, K. V. and Ignatov, V. V. (2013) *Paenibacillus polymyxa* Rhizobacteria and Their Synthesized Exoglycans in Interaction with Wheat Roots: Colonization and Root Hair Deformation. *Curr. Microbiol.* **66**, 481–486.

Yeo, A.R. and Flowers, T.J. (1982) Accumulation and localisation of sodium ions within the shoots of rice (*Oryza sativa*) varieties differing in salinity resistance. *Physiol. Plant.* **56**, 343–348.

Yuan, Z. and Tam, V.H. (2008) Polymyxin B: a new strategy for multidrug-resistant Gram-negative organisms. *Expert Opin. Investig. Drugs* **17**, 661–668.

Zdobnov, E.M. and Apweiler, R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–8.

Zegans, M.E., Wagner, J.C., Cady, K.C., Murphy, D.M., Hammond, J.H. and O'Toole, G.A. (2009) Interaction between Bacteriophage DMS3 and Host CRISPR Region Inhibits Group Behaviors of *Pseudomonas aeruginosa*. *J. Bacteriol.* **191**, 210–219.

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., Wu, J. and Xiao, J. (2014) PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* **30**, 1297–9.

Table 1. Genomes of *X. oryzae* pv. *oryzae* and pv. *oryzicola* used in the pangenome analysis.

GenBank	Status	pathovar	Strain
AE013598	Complete	*oryzae*	KACC 10331
AP008229	Complete	*oryzae*	MAFF 311018
AYSX00000000	Draft	*oryzae*	NAI8 515
CP000967	Complete	*oryzae*	PXO99A
CP007166	Complete	*oryzae*	PXO86
CP011532	Complete	*oryzae*	XF89b
CP012947	Complete	*oryzae*	PXO83
CP013666	Complete	*oryzae*	AXO1947
CP013670	Complete	*oryzae*	PXO71
CP013674	Complete	*oryzae*	PXO211
CP013675	Complete	*oryzae*	PXO236
CP013676	Complete	*oryzae*	PXO282
CP013677	Complete	*oryzae*	PXO524
CP013678	Complete	*oryzae*	PXO563
CP013679	Complete	*oryzae*	PXO602
CP013961	Complete	*oryzae*	PXO145
JTKM00000000	Draft	*oryzae*	LMG 9585
JXDM00000000	Draft	*oryzae*	BXO1
JXDN00000000	Draft	*oryzae*	BXO2
JXDO00000000	Draft	*oryzae*	BXO6
JXDP00000000	Draft	*oryzae*	BXO8
JXDQ00000000	Draft	*oryzae*	BXO25
JXDR00000000	Draft	*oryzae*	BXO33
JXDS00000000	Draft	*oryzae*	BXO34
JXDT00000000	Draft	*oryzae*	BXO407
JXDU00000000	Draft	*oryzae*	BXO416
JXDV00000000	Draft	*oryzae*	BXO432
JXDW00000000	Draft	*oryzae*	BXO439
JXDX00000000	Draft	*oryzae*	BXO447
JXDY00000000	Draft	*oryzae*	BXO454
JXDZ00000000	Draft	*oryzae*	BXO471
JXEA00000000	Draft	*oryzae*	BXO512
JXEB00000000	Draft	*oryzae*	BXO554
JXEC00000000	Draft	*oryzae*	BXO557
JXED00000000	Draft	*oryzae*	BXO558
JXEE00000000	Draft	*oryzae*	BXO559
JXEF00000000	Draft	*oryzae*	BXO571
JXEG00000000	Draft	*oryzae*	BXO582
JXEH00000000	Draft	*oryzae*	BXO589
JXEI00000000	Draft	*oryzae*	BXO590
JXEJ00000000	Draft	*oryzae*	DXO-012
JXEK00000000	Draft	*oryzae*	DXO-015

JXEL00000000	Draft	*oryzae*	DXO-27
JXEM00000000	Draft	*oryzae*	DXO-044
JXEN00000000	Draft	*oryzae*	DXO-50
JXEO00000000	Draft	*oryzae*	DXO-052
JXEP00000000	Draft	*oryzae*	DXO-089
JXEQ00000000	Draft	*oryzae*	DXO-091
JXER00000000	Draft	*oryzae*	DXO-116
JXES00000000	Draft	*oryzae*	DXO-122
JXET00000000	Draft	*oryzae*	DXO-129
JXEU00000000	Draft	*oryzae*	DXO-131
JXEV00000000	Draft	*oryzae*	DXO-133
JXEW00000000	Draft	*oryzae*	DXO-150
JXEX00000000	Draft	*oryzae*	DXO-165
JXEY00000000	Draft	*oryzae*	DXO-170
JXEZ00000000	Draft	*oryzae*	DXO-174
JXFA00000000	Draft	*oryzae*	DXO-181
JXFB00000000	Draft	*oryzae*	DXO-200
JXFC00000000	Draft	*oryzae*	DXO-203
JXFD00000000	Draft	*oryzae*	DXO-206
JXFE00000000	Draft	*oryzae*	DXO-216
JXFF00000000	Draft	*oryzae*	DXO-226
JXFG00000000	Draft	*oryzae*	DXO-233
JXFH00000000	Draft	*oryzae*	DXO-242
JXFI00000000	Draft	*oryzae*	DXO-246
JXFJ00000000	Draft	*oryzae*	DXO-248
JXFK00000000	Draft	*oryzae*	DXO-331
JXFL00000000	Draft	*oryzae*	DXO-369
JXFM00000000	Draft	*oryzae*	DXO-397
JXFN00000000	Draft	*oryzae*	IXO35
JXFO00000000	Draft	*oryzae*	IXO74
JXFP00000000	Draft	*oryzae*	IXO89
JXFQ00000000	Draft	*oryzae*	IXO90
JXFR00000000	Draft	*oryzae*	IXO92
JXFS00000000	Draft	*oryzae*	IXO93
JXFT00000000	Draft	*oryzae*	IXO97
JXFU00000000	Draft	*oryzae*	IXO98
JXFV00000000	Draft	*oryzae*	IXO99
JXFW00000000	Draft	*oryzae*	IXO134
JXFX00000000	Draft	*oryzae*	IXO141
JXFY00000000	Draft	*oryzae*	IXO151
JXFZ00000000	Draft	*oryzae*	IXO159
JXGA00000000	Draft	*oryzae*	IXO189
JXGB00000000	Draft	*oryzae*	IXO220
JXGC00000000	Draft	*oryzae*	IXO221
JXGD00000000	Draft	*oryzae*	IXO222
JXGE00000000	Draft	*oryzae*	IXO278

JXGF00000000	Draft	*oryzae*	IXO365
JXGG00000000	Draft	*oryzae*	IXO367
JXGH00000000	Draft	*oryzae*	IXO390
JXGI00000000	Draft	*oryzae*	IXO411
JXGJ00000000	Draft	*oryzae*	IXO414
JXGK00000000	Draft	*oryzae*	IXO493
JXGL00000000	Draft	*oryzae*	IXO597
JXGM00000000	Draft	*oryzae*	IXO599
JXGN00000000	Draft	*oryzae*	IXO603
JXGO00000000	Draft	*oryzae*	IXO608
JXGP00000000	Draft	*oryzae*	IXO620
JXGQ00000000	Draft	*oryzae*	IXO621
JXGR00000000	Draft	*oryzae*	IXO627
JXGS00000000	Draft	*oryzae*	IXO630
JXGT00000000	Draft	*oryzae*	IXO639
JXGU00000000	Draft	*oryzae*	IXO644
JXGV00000000	Draft	*oryzae*	IXO645
JXGW00000000	Draft	*oryzae*	IXO651
JXGX00000000	Draft	*oryzae*	IXO675
JXGY00000000	Draft	*oryzae*	IXO685
JXGZ00000000	Draft	*oryzae*	IXO704
JXHA00000000	Draft	*oryzae*	IXO725
JXHB00000000	Draft	*oryzae*	IXO792
JXHC00000000	Draft	*oryzae*	IXO812
JXHD00000000	Draft	*oryzae*	IXO842
JXHE00000000	Draft	*oryzae*	IXO884
JXHF00000000	Draft	*oryzae*	IXO1088
JXHG00000000	Draft	*oryzae*	IXO1104
JXHH00000000	Draft	*oryzae*	IXO1221
AFHL00000000	Draft	*oryzae*	X8-1A
LHUJ00000000	Draft	*oryzae*	X11-5A
AYSY00000000	Draft	*oryzicola*	MAI10
CP003057	Complete	*oryzicola*	BLS256
CP007221	Complete	*oryzicola*	CFBP7342
CP007810	Complete	*oryzicola*	YM15
CP011955	Complete	*oryzicola*	B8-12
CP011956	Complete	*oryzicola*	BLS279
CP011957	Complete	*oryzicola*	BXOR1
CP011958	Complete	*oryzicola*	CFBP7331
CP011959	Complete	*oryzicola*	CFBP7341
CP011960	Complete	*oryzicola*	L8
CP011961	Complete	*oryzicola*	RS105
CP011962	Complete	*oryzicola*	CFBP2286
JXHI00000000	Draft	*oryzicola*	BXOR1
MVFS00000000	Draft	*oryzicola*	BAI15
MVFT00000000	Draft	*oryzicola*	BAI20

MVFU00000000	Draft	*oryzicola*	BAI21

1368

1369 **Table 2.** Overview of the results from the genome re-annotation using Genix.

Pathovar	Strain	CDSs identified		Found	New	Missing	Discrepancy (%)
		GenBank	Genix				
oryzae	KACC 10331	4538	4821	4018	803	520	14.14
oryzae	MAFF 311018	4372	4736	4164	572	208	8.56
oryzae	NAI8 515	0	3873	0	3873	0	100
oryzae	PXO99A	4951	4969	4285	684	666	13.61
oryzae	PXO86	4500	4754	4105	649	395	11.28
oryzae	XF89b	4177	4742	3978	764	199	10.8
oryzae	PXO83	4294	4763	4082	681	212	9.86
oryzae	AXO1947	3706	4229	3526	703	180	11.13
oryzae	PXO71	4236	4703	4012	691	224	10.24
oryzae	PXO211	4319	4785	4102	683	217	9.89
oryzae	PXO236	4273	4702	4063	639	210	9.46
oryzae	PXO282	4258	4791	4035	756	223	10.82
oryzae	PXO524	4279	4751	4063	688	216	10.01
oryzae	PXO563	4282	4730	4058	672	224	9.94
oryzae	PXO602	4257	4747	4045	702	212	10.15
oryzae	PXO145	4571	4812	4186	626	385	10.77
oryzae	LMG 9585	4150	4011	3383	628	767	17.09
oryzae	BXO1	3823	4025	3439	586	384	12.36
oryzae	BXO2	3601	3822	3268	554	333	11.95
oryzae	BXO6	3650	3849	3274	575	376	12.68
oryzae	BXO8	3661	3882	3301	581	360	12.48
oryzae	BXO25	3661	3847	3292	555	369	12.31
oryzae	BXO33	3631	3861	3299	562	332	11.93
oryzae	BXO34	3670	3880	3304	576	366	12.48
oryzae	BXO407	3759	3964	3368	596	391	12.78
oryzae	BXO416	3601	3821	3254	567	347	12.31
oryzae	BXO432	3734	3896	3324	572	410	12.87
oryzae	BXO439	3605	3831	3263	568	342	12.24
oryzae	BXO447	3631	3871	3311	560	320	11.73
oryzae	BXO454	3594	3811	3254	557	340	12.11

oryzae	BXO471	3638	3807	3277	530	361	11.97
oryzae	BXO512	3623	3848	3288	560	335	11.98
oryzae	BXO554	3658	3883	3307	576	351	12.29
oryzae	BXO557	3604	3840	3285	555	319	11.74
oryzae	BXO558	3620	3853	3289	564	331	11.98
oryzae	BXO559	3643	3831	3275	556	368	12.36
oryzae	BXO571	3653	3852	3275	577	378	12.72
oryzae	BXO582	3654	3847	3299	548	355	12.04
oryzae	BXO589	3632	3854	3290	564	342	12.1
oryzae	BXO590	3612	3842	3264	578	348	12.42
oryzae	DXO-012	3628	3840	3271	569	357	12.4
oryzae	DXO-015	3635	3861	3302	559	333	11.9
oryzae	DXO-27	3635	3806	3250	556	385	12.65
oryzae	DXO-044	3634	3828	3284	544	350	11.98
oryzae	DXO-50	3750	3986	3388	598	362	12.41
oryzae	DXO-052	3594	3813	3273	540	321	11.62
oryzae	DXO-089	3665	3881	3330	551	335	11.74
oryzae	DXO-091	3682	3860	3314	546	368	12.12
oryzae	DXO-116	3646	3837	3289	548	357	12.09
oryzae	DXO-122	3618	3821	3282	539	336	11.76
oryzae	DXO-129	3604	3855	3282	573	322	12
oryzae	DXO-131	3606	3807	3252	555	354	12.26
oryzae	DXO-133	3643	3885	3323	562	320	11.72
oryzae	DXO-150	3635	3798	3273	525	362	11.93
oryzae	DXO-165	3680	3894	3317	577	363	12.41
oryzae	DXO-170	3727	3920	3373	547	354	11.78
oryzae	DXO-174	3620	3906	3305	601	315	12.17
oryzae	DXO-181	3709	3857	3310	547	399	12.5
oryzae	DXO-200	3589	3840	3277	563	312	11.78
oryzae	DXO-203	3641	3837	3292	545	349	11.96
oryzae	DXO-206	3788	4002	3438	564	350	11.73
oryzae	DXO-216	3637	3853	3293	560	344	12.07
oryzae	DXO-226	3688	3841	3312	529	376	12.02

oryzae	DXO-233	3713	3926	3353	573	360	12.21
oryzae	DXO-242	3670	3860	3283	577	387	12.8
oryzae	DXO-246	3636	3874	3293	581	343	12.3
oryzae	DXO-248	3596	3799	3276	523	320	11.4
oryzae	DXO-331	3756	3928	3378	550	378	12.08
oryzae	DXO-369	3638	3841	3295	546	343	11.89
oryzae	DXO-397	3662	3841	3274	567	388	12.73
oryzae	IXO35	3708	3949	3375	574	333	11.85
oryzae	IXO74	3716	3905	3337	568	379	12.43
oryzae	IXO89	3634	3878	3299	579	335	12.17
oryzae	IXO90	3732	3909	3334	575	398	12.73
oryzae	IXO92	3713	3909	3342	567	371	12.31
oryzae	IXO93	3711	3860	3328	532	383	12.09
oryzae	IXO97	3748	3957	3381	576	367	12.24
oryzae	IXO98	3658	3831	3278	553	380	12.46
oryzae	IXO99	3645	3873	3306	567	339	12.05
oryzae	IXO134	3827	4043	3426	617	401	12.94
oryzae	IXO141	3785	3968	3417	551	368	11.85
oryzae	IXO151	3764	3964	3328	636	436	13.87
oryzae	IXO159	3720	3894	3335	559	385	12.4
oryzae	IXO189	3680	3876	3293	583	387	12.84
oryzae	IXO220	3699	3887	3323	564	376	12.39
oryzae	IXO221	3618	3849	3296	553	322	11.72
oryzae	IXO222	3782	3960	3355	605	427	13.33
oryzae	IXO278	3647	3889	3300	589	347	12.42
oryzae	IXO365	3814	3994	3395	599	419	13.04
oryzae	IXO367	3661	3831	3295	536	366	12.04
oryzae	IXO390	3789	3980	3395	585	394	12.6
oryzae	IXO411	3679	3877	3302	575	377	12.6
oryzae	IXO414	3587	3823	3273	550	314	11.66
oryzae	IXO493	3791	3983	3395	588	396	12.66
oryzae	IXO597	3673	3875	3289	586	384	12.85
oryzae	IXO599	3686	3906	3317	589	369	12.62

oryzae	IXO603	3726	3884	3327	557	399	12.56
oryzae	IXO608	3704	3886	3321	565	383	12.49
oryzae	IXO620	3715	3948	3363	585	352	12.23
oryzae	IXO621	3727	3903	3339	564	388	12.48
oryzae	IXO627	3722	3911	3321	590	401	12.98
oryzae	IXO630	3803	4023	3409	614	394	12.88
oryzae	IXO639	3668	3854	3308	546	360	12.04
oryzae	IXO644	3781	3989	3389	600	392	12.77
oryzae	IXO645	3830	4002	3387	615	443	13.51
oryzae	IXO651	3693	3921	3343	578	350	12.19
oryzae	IXO675	3648	3900	3309	591	339	12.32
oryzae	IXO685	3728	3977	3371	606	357	12.5
oryzae	IXO704	3687	3915	3324	591	363	12.55
oryzae	IXO725	3765	3949	3347	602	418	13.22
oryzae	IXO792	3776	3961	3374	587	402	12.78
oryzae	IXO812	3701	3925	3327	598	374	12.75
oryzae	IXO842	3753	3915	3339	576	414	12.91
oryzae	IXO884	3659	3877	3318	559	341	11.94
oryzae	IXO1088	3776	3972	3368	604	408	13.06
oryzae	IXO1104	3780	4007	3392	615	388	12.88
oryzae	IXO1221	3751	3910	3323	587	428	13.25
oryzae	X8-1A	4029	4375	3563	812	466	15.21
oryzae	X11-5A	0	4170	0	4170	0	100
oryzicola	MAI10	0	3962	0	3962	0	100
oryzicola	BLS256	4473	4298	3903	395	570	11
oryzicola	CFBP7342	4016	4697	3849	848	167	11.65
oryzicola	YM15	3583	3987	3415	572	168	9.78
oryzicola	B8-12	3744	4267	3592	675	152	10.32
oryzicola	BLS279	3737	4267	3582	685	155	10.49
oryzicola	BXOR1	3656	4183	3512	671	144	10.4
oryzicola	CFBP7331	3911	4517	3752	765	159	10.96
oryzicola	CFBP7341	3925	4520	3762	758	163	10.91
oryzicola	L8	3733	4260	3586	674	147	10.27

oryzicola	RS105	3730	4274	3571	703	159	10.77
oryzicola	CFBP2286	3902	4461	3739	722	163	10.58
oryzicola	BXOR1	3543	3894	3249	645	294	12.63
oryzicola	BAI15	3973	3800	3392	408	581	12.72
oryzicola	BAI20	4159	3973	3527	446	632	13.26
oryzicola	BAI21	3919	3843	3401	442	518	12.37

1370

1371

Table 3. Sub-sets of the pangenome calculated for the *X. oryzae* species and for the pathovar *oryzae* and *oryzicola* based on the results of each algorithm used in GET_HOMOLOGUES. Core-genome comprises all genes present in at least 95% of the strains in both pathovars (for the species), or in each pathovar. Exclusive core-genome are those genes that are conserved in at least 95% of the strain in one pathovar but not in the other. Pathovar conserved genes are those present in at least 95% of the strain in one pathovar but present in only 5% or less in the other pathovar.

Sub-sets	Algorithm			
	BDBH	**COG**	**OMCL**	**Consensus**
X. oryzae pangenome	3724	14599	8230	3292
X. oryzae core genome	2342	2391	2385	2342
Xoo core genome	3019	3059	3055	2826
Xoc core genome	2486	2892	2903	2356
Xoo exclusive core genome	677	668	670	576
Xoc exclusive core genome	144	501	518	106
Xoo exclusive	286	279	289	256
Xoc exclusive	0	305	309	0

Table 4. Ratio core genome / pangenome calculated based on the results of each algorithm implemented in GET_HOMOLOGUES that used in the inference of the pangenome of *X. oryzae*.

BDBH	COG	OMCL	Consensus
62.88%	16.37%	28.97%	71.14%
2342	2391	2385	2342
62.88%	16.37%	28.97%	71.14%

Table 5. Gene Ontology terms identified as enriched in core exclusive genome subsets of *X. oryzae pv. oryzae* and *X. oryzae pv. oryzicola* based on the orthologue groups identified by different algorithms implemented in the GET_HOMOLOGUES pipeline.

Pathovar	Sub-set	GO Term	Namespace	Name	*p*-value [a]			
					BDBH	COG	OMCL	Consensus
oryzae	Core exclusive	GO:0000049	molecular function	tRNA binding	0.01			0.03
oryzae	Core exclusive	GO:0000272	biological process	polysaccharide catabolic process	0.01			
oryzae	Core exclusive	GO:0000724	biological process	double-strand break repair via homologous recombination		0.03		
oryzae	Core exclusive	GO:0000906	molecular function	6,7-dimethyl-8-ribityllumazine synthase activity		0.03		
oryzae	Core exclusive	GO:0002949	biological process	tRNA threonylcarbamoyladenosine modification		0.03		
oryzae	Core exclusive	GO:0003677	molecular function	DNA binding	0.04			0.03
oryzae	Core exclusive	GO:0003684	molecular function	damaged DNA binding		0.03		
oryzae	Core exclusive	GO:0003735	molecular function	structural constituent of ribosome	0.01			
oryzae	Core exclusive	GO:0003755	molecular function	peptidyl-prolyl cis-trans isomerase activity	0.04			0.04
oryzae	Core exclusive	GO:0003908	molecular function	methylated-DNA-[protein]-cysteine S-methyltransferase activity		0.01		0.05
oryzae	Core exclusive	GO:0003937	molecular function	IMP cyclohydrolase activity		0.03	0.03	
oryzae	Core exclusive	GO:0003994	molecular function	aconitate hydratase activity		0.05	0.05	
oryzae	Core exclusive	GO:0004020	molecular function	adenylylsulfate kinase activity			0.03	
oryzae	Core exclusive	GO:0004107	molecular function	chorismate synthase activity		0.03	0.03	
oryzae	Core exclusive	GO:0004149	molecular function	dihydrolipoyllysine-residue succinyltransferase activity		0.03		0.05
oryzae	Core exclusive	GO:0004315	molecular function	3-oxoacyl-[acyl-carrier-protein] synthase activity		0.05		
oryzae	Core exclusive	GO:0004386	molecular function	helicase activity			0.02	
oryzae	Core exclusive	GO:0004399	molecular function	histidinol dehydrogenase activity			0.03	
oryzae	Core exclusive	GO:0004413	molecular function	homoserine kinase activity			0.03	0.05
oryzae	Core exclusive	GO:0004478	molecular function	methionine adenosyltransferase activity		0.01	0.03	
oryzae	Core exclusive	GO:0004519	molecular function	endonuclease activity		0.04		
oryzae	Core exclusive	GO:0004521	molecular function	endoribonuclease activity		0.03		
oryzae	Core exclusive	GO:0004527	molecular function	exonuclease activity			0.02	
oryzae	Core exclusive	GO:0004553	molecular function	hydrolase activity, hydrolyzing O-glycosyl compounds		0.03		
oryzae	Core exclusive	GO:0004643	molecular function	phosphoribosylaminoimidazolecarboxamide formyltransferase activity		0.03	0.03	

oryzae	Core exclusive	GO:0004673	molecular function	protein histidine kinase activity	0.02			0.01
oryzae	Core exclusive	GO:0004816	molecular function	asparagine-tRNA ligase activity		0.03	0.03	
oryzae	Core exclusive	GO:0004821	molecular function	histidine-tRNA ligase activity		0.03	0.03	
oryzae	Core exclusive	GO:0005351	molecular function	carbohydrate:proton symporter activity		0.03	0.05	
oryzae	Core exclusive	GO:0005622	cellular component	intracellular	0.02			0.01
oryzae	Core exclusive	GO:0005840	cellular component	ribosome			0.02	
oryzae	Core exclusive	GO:0005975	biological process	carbohydrate metabolic process			0.02	
oryzae	Core exclusive	GO:0005978	biological process	glycogen biosynthetic process		0.03		
oryzae	Core exclusive	GO:0006099	biological process	tricarboxylic acid cycle		0.01	0.05	
oryzae	Core exclusive	GO:0006189	biological process	'de novo' IMP biosynthetic process		0.03		
oryzae	Core exclusive	GO:0006207	biological process	'de novo' pyrimidine nucleobase biosynthetic process		0.03		
oryzae	Core exclusive	GO:0006281	biological process	DNA repair			0.02	
oryzae	Core exclusive	GO:0006352	biological process	DNA-templated transcription, initiation		0.04		
oryzae	Core exclusive	GO:0006364	biological process	rRNA processing	0.03			
oryzae	Core exclusive	GO:0006556	biological process	S-adenosylmethionine biosynthetic process		0.01	0.03	
oryzae	Core exclusive	GO:0006567	biological process	threonine catabolic process		0.03	0.05	
oryzae	Core exclusive	GO:0006812	biological process	cation transport			0.05	
oryzae	Core exclusive	GO:0006878	biological process	cellular copper ion homeostasis		0.02	0.05	
oryzae	Core exclusive	GO:0006974	biological process	cellular response to DNA damage stimulus		0.01		
oryzae	Core exclusive	GO:0008152	biological process	metabolic process			0.02	
oryzae	Core exclusive	GO:0008168	molecular function	methyltransferase activity		0.04		
oryzae	Core exclusive	GO:0008233	molecular function	peptidase activity		0.04		
oryzae	Core exclusive	GO:0008270	molecular function	zinc ion binding			0.01	
oryzae	Core exclusive	GO:0008483	molecular function	transaminase activity		0.04		
oryzae	Core exclusive	GO:0008616	biological process	queuosine biosynthetic process		0.01		
oryzae	Core exclusive	GO:0008643	biological process	carbohydrate transport		0.01	0.03	
oryzae	Core exclusive	GO:0008658	molecular function	penicillin binding			0.01	
oryzae	Core exclusive	GO:0008697	molecular function	4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase activity	0.02		0.02	
oryzae	Core exclusive	GO:0008705	molecular function	methionine synthase activity		0.01	0.03	0.05
oryzae	Core exclusive	GO:0008810	molecular function	cellulase activity		0.02		
oryzae	Core exclusive	GO:0008854	molecular function	exodeoxyribonuclease V activity			0.02	0.05
oryzae	Core exclusive	GO:0008967	molecular function	phosphoglycolate phosphatase activity		0.01		
oryzae	Core exclusive	GO:0009002	molecular function	serine-type D-Ala-D-Ala carboxypeptidase activity		0.02	0.03	
oryzae	Core exclusive	GO:0009011	molecular function	starch synthase activity		0.03	0.03	
oryzae	Core exclusive	GO:0009103	biological process	lipopolysaccharide biosynthetic process			0.03	
oryzae	Core exclusive	GO:0009228	biological process	thiamine biosynthetic process		0.04		

oryzae	Core exclusive	GO:0009252	biological process	peptidoglycan biosynthetic process			0.02	
oryzae	Core exclusive	GO:0009274	cellular component	peptidoglycan-based cell wall		0.02	0.05	
oryzae	Core exclusive	GO:0009306	biological process	protein secretion		0.01		
oryzae	Core exclusive	GO:0009401	biological process	phosphoenolpyruvate-dependent sugar phosphotransferase system		0.01	0.02	
oryzae	Core exclusive	GO:0015649	molecular function	2-keto-3-deoxygluconate:proton symporter activity		0.01	0.03	
oryzae	Core exclusive	GO:0015891	biological process	siderophore transport		0.03		
oryzae	Core exclusive	GO:0016021	cellular component	integral component of membrane			0.05	
oryzae	Core exclusive	GO:0016042	biological process	lipid catabolic process		0.01	0.04	0.05
oryzae	Core exclusive	GO:0016740	molecular function	transferase activity			0.02	
oryzae	Core exclusive	GO:0016757	molecular function	transferase activity, transferring glycosyl groups		0.03		
oryzae	Core exclusive	GO:0016758	molecular function	transferase activity, transferring hexosyl groups	0.02			0.02
oryzae	Core exclusive	GO:0016772	molecular function	transferase activity, transferring phosphorus-containing groups		0.02		
oryzae	Core exclusive	GO:0016779	molecular function	nucleotidyltransferase activity	0.04			0.04
oryzae	Core exclusive	GO:0016787	molecular function	hydrolase activity			0.04	
oryzae	Core exclusive	GO:0016788	molecular function	hydrolase activity, acting on ester bonds		0.02		
oryzae	Core exclusive	GO:0016798	molecular function	hydrolase activity, acting on glycosyl bonds			0.03	
oryzae	Core exclusive	GO:0016799	molecular function	hydrolase activity, hydrolyzing N-glycosyl compounds		0.05		
oryzae	Core exclusive	GO:0016810	molecular function	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds		0.03		
oryzae	Core exclusive	GO:0016829	molecular function	lyase activity		0.02		
oryzae	Core exclusive	GO:0016887	molecular function	ATPase activity	0.01			
oryzae	Core exclusive	GO:0018106	biological process	peptidyl-histidine phosphorylation	0.02			0.01
oryzae	Core exclusive	GO:0019354	biological process	siroheme biosynthetic process	0.02			0.02
oryzae	Core exclusive	GO:0019556	biological process	histidine catabolic process to glutamate and formamide		0.05		
oryzae	Core exclusive	GO:0030529	cellular component	intracellular ribonucleoprotein complex	0.01			
oryzae	Core exclusive	GO:0030599	molecular function	pectinesterase activity		0.03	0.03	
oryzae	Core exclusive	GO:0031419	molecular function	cobalamin binding		0.01	0.03	0.05
oryzae	Core exclusive	GO:0033512	biological process	L-lysine catabolic process to acetyl-CoA via saccharopine			0.03	
oryzae	Core exclusive	GO:0033818	molecular function	beta-ketoacyl-acyl-carrier-protein synthase III activity		0.03	0.03	

oryzae	Core exclusive	GO:0034028	molecular function	5-(carboxyamino)imidazole ribonucleotide synthase activity		0.05	0.03	
oryzae	Core exclusive	GO:0036104	biological process	Kdo2-lipid A biosynthetic process		0.01	0.03	
oryzae	Core exclusive	GO:0042545	biological process	cell wall modification		0.03	0.03	
oryzae	Core exclusive	GO:0042834	molecular function	peptidoglycan binding			0.03	0.05
oryzae	Core exclusive	GO:0042919	biological process	benzoate transport		0.05		
oryzae	Core exclusive	GO:0042925	molecular function	benzoate transmembrane transporter activity		0.05		
oryzae	Core exclusive	GO:0043115	molecular function	precorrin-2 dehydrogenase activity		0.03		
oryzae	Core exclusive	GO:0045330	molecular function	aspartyl esterase activity		0.03	0.03	
oryzae	Core exclusive	GO:0045493	biological process	xylan catabolic process			0.02	
oryzae	Core exclusive	GO:0046294	biological process	formaldehyde catabolic process			0.01	
oryzae	Core exclusive	GO:0046656	biological process	folic acid biosynthetic process		0.02	0.05	
oryzae	Core exclusive	GO:0046872	molecular function	metal ion binding			0.04	
oryzae	Core exclusive	GO:0050511	molecular function	undecaprenyldiphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase activity		0.01	0.03	
oryzae	Core exclusive	GO:0050660	molecular function	flavin adenine dinucleotide binding	0.03			
oryzae	Core exclusive	GO:0051301	biological process	cell division	0.02			0.04
oryzae	Core exclusive	GO:0051903	molecular function	S-(hydroxymethyl)glutathione dehydrogenase activity		0.03	0.03	
oryzae	Core exclusive	GO:0055085	biological process	transmembrane transport		0.03		
oryzae	Core exclusive	GO:0071555	biological process	cell wall organization		0.03		
oryzae	Core exclusive	GO:2000142	biological process	regulation of DNA-templated transcription, initiation		0.04		
oryzae	Exclusive	GO:0000103	biological process	sulfate assimilation			0.03	
oryzae	Exclusive	GO:0000160	biological process	phosphorelay signal transduction system	0.02			
oryzae	Exclusive	GO:0000166	molecular function	nucleotide binding		0.02	0.01	
oryzae	Exclusive	GO:0000287	molecular function	magnesium ion binding	0.03			0.05
oryzae	Exclusive	GO:0003676	molecular function	nucleic acid binding		0.03		
oryzae	Exclusive	GO:0003723	molecular function	RNA binding		0.04		
oryzae	Exclusive	GO:0003779	molecular function	actin binding		0.04		
oryzae	Exclusive	GO:0003905	molecular function	alkylbase DNA N-glycosylase activity		0.04		
oryzae	Exclusive	GO:0004020	molecular function	adenylylsulfate kinase activity	0.02			
oryzae	Exclusive	GO:0004222	molecular function	metalloendopeptidase activity		0.01		
oryzae	Exclusive	GO:0004252	molecular function	serine-type endopeptidase activity		0.02		0.04
oryzae	Exclusive	GO:0004414	molecular function	homoserine O-acetyltransferase activity		0.04		
oryzae	Exclusive	GO:0004553	molecular function	hydrolase activity, hydrolyzing O-glycosyl compounds			0.05	
oryzae	Exclusive	GO:0004629	molecular function	phospholipase C activity	0.04			0.05

oryzae	Exclusive	GO:0004781	molecular function	sulfate adenylyltransferase (ATP) activity			0.04	
oryzae	Exclusive	GO:0006351	biological process	transcription, DNA-templated	0.01		0.03	
oryzae	Exclusive	GO:0006355	biological process	regulation of transcription, DNA-templated			0.04	
oryzae	Exclusive	GO:0006412	biological process	translation			0.02	
oryzae	Exclusive	GO:0008643	biological process	carbohydrate transport	0.03			0.03
oryzae	Exclusive	GO:0008810	molecular function	cellulase activity	0.04		0.04	
oryzae	Exclusive	GO:0009086	biological process	methionine biosynthetic process	0.02			0.04
oryzae	Exclusive	GO:0009252	biological process	peptidoglycan biosynthetic process	0.02			0.02
oryzae	Exclusive	GO:0009279	cellular component	cell outer membrane		0.04		
oryzae	Exclusive	GO:0015031	biological process	protein transport	0.05			0.05
oryzae	Exclusive	GO:0015074	biological process	DNA integration			0.04	
oryzae	Exclusive	GO:0015293	molecular function	symporter activity	0.04			0.04
oryzae	Exclusive	GO:0015649	molecular function	2-keto-3-deoxygluconate:proton symporter activity	0.03			0.03
oryzae	Exclusive	GO:0016020	cellular component	membrane	0.01			
oryzae	Exclusive	GO:0016052	biological process	carbohydrate catabolic process		0.04		
oryzae	Exclusive	GO:0016491	molecular function	oxidoreductase activity			0.01	
oryzae	Exclusive	GO:0016787	molecular function	hydrolase activity		0.01		
oryzae	Exclusive	GO:0016798	molecular function	hydrolase activity, acting on glycosyl bonds	0.04		0.02	
oryzae	Exclusive	GO:0016874	molecular function	ligase activity	0.03			
oryzae	Exclusive	GO:0019700	biological process	organic phosphonate catabolic process		0.04		
oryzae	Exclusive	GO:0019843	molecular function	rRNA binding	0.02			0.02
oryzae	Exclusive	GO:0022857	molecular function	transmembrane transporter activity	0.05			
oryzae	Exclusive	GO:0030245	biological process	cellulose catabolic process		0.04		
oryzae	Exclusive	GO:0030529	cellular component	intracellular ribonucleoprotein complex	0.05			0.05
oryzae	Exclusive	GO:0030599	molecular function	pectinesterase activity	0.03			0.03
oryzae	Exclusive	GO:0042545	biological process	cell wall modification	0.03			0.03
oryzae	Exclusive	GO:0045330	molecular function	aspartyl esterase activity	0.03			0.03
oryzae	Exclusive	GO:0046411	biological process	2-keto-3-deoxygluconate transport		0.04		
oryzae	Exclusive	GO:0046872	molecular function	metal ion binding	0.01			
oryzae	Exclusive	GO:0050023	molecular function	L-fuconate dehydratase activity		0.04		
oryzae	Exclusive	GO:0051301	biological process	cell division	0.05			
oryzae	Exclusive	GO:0051536	molecular function	iron-sulfur cluster binding	0.02			0.02
oryzae	Exclusive	GO:0051539	molecular function	4 iron, 4 sulfur cluster binding	0.03			0.03
oryzicola	Core exclusive	GO:0000155	molecular function	phosphorelay sensor kinase activity		0.03		
oryzicola	Core exclusive	GO:0000160	biological process	phosphorelay signal transduction system		0.03		
oryzicola	Core exclusive	GO:0000271	biological process	polysaccharide biosynthetic process		0.04		
oryzicola	Core exclusive	GO:0003676	molecular function	nucleic acid binding			0.03	
oryzicola	Core exclusive	GO:0003677	molecular function	DNA binding			0.01	

oryzicola	Core exclusive	GO:0003824	molecular function	catalytic activity		0.04		
oryzicola	Core exclusive	GO:0003866	molecular function	3-phosphoshikimate 1-carboxyvinyltransferase activity		0.02	0.05	
oryzicola	Core exclusive	GO:0004222	molecular function	metalloendopeptidase activity		0.03		
oryzicola	Core exclusive	GO:0004360	molecular function	glutamine-fructose-6-phosphate transaminase (isomerizing) activity		0.02	0.03	
oryzicola	Core exclusive	GO:0004412	molecular function	homoserine dehydrogenase activity		0.05		
oryzicola	Core exclusive	GO:0004553	molecular function	hydrolase activity, hydrolyzing O-glycosyl compounds			0.03	
oryzicola	Core exclusive	GO:0004601	molecular function	peroxidase activity			0.02	
oryzicola	Core exclusive	GO:0004648	molecular function	O-phospho-L-serine:2-oxoglutarate aminotransferase activity		0.02	0.02	
oryzicola	Core exclusive	GO:0004871	molecular function	signal transducer activity		0.04		
oryzicola	Core exclusive	GO:0005215	molecular function	transporter activity		0.02		
oryzicola	Core exclusive	GO:0005506	molecular function	iron ion binding			0.02	
oryzicola	Core exclusive	GO:0005840	cellular component	ribosome	0.03			
oryzicola	Core exclusive	GO:0005886	cellular component	plasma membrane		0.02		0.05
oryzicola	Core exclusive	GO:0006534	biological process	cysteine metabolic process		0.02	0.05	
oryzicola	Core exclusive	GO:0006564	biological process	L-serine biosynthetic process	0.01	0.03	0.03	
oryzicola	Core exclusive	GO:0006596	biological process	polyamine biosynthetic process	0.04			
oryzicola	Core exclusive	GO:0007165	biological process	signal transduction		0.01		
oryzicola	Core exclusive	GO:0008080	molecular function	N-acetyltransferase activity		0.03		
oryzicola	Core exclusive	GO:0008239	molecular function	dipeptidyl-peptidase activity		0.02	0.05	
oryzicola	Core exclusive	GO:0008410	molecular function	CoA-transferase activity		0.03		
oryzicola	Core exclusive	GO:0008616	biological process	queuosine biosynthetic process	0.01	0.05		
oryzicola	Core exclusive	GO:0008677	molecular function	2-dehydropantoate 2-reductase activity		0.04		
oryzicola	Core exclusive	GO:0008705	molecular function	methionine synthase activity			0.01	
oryzicola	Core exclusive	GO:0008750	molecular function	NAD(P)+ transhydrogenase (AB-specific) activity			0.02	
oryzicola	Core exclusive	GO:0008800	molecular function	beta-lactamase activity			0.03	
oryzicola	Core exclusive	GO:0008915	molecular function	lipid-A-disaccharide synthase activity	0.01	0.03	0.03	
oryzicola	Core exclusive	GO:0009252	biological process	peptidoglycan biosynthetic process			0.05	
oryzicola	Core exclusive	GO:0009435	biological process	NAD biosynthetic process		0.02		
oryzicola	Core exclusive	GO:0015940	biological process	pantothenate biosynthetic process	0.01			
oryzicola	Core exclusive	GO:0016260	biological process	selenocysteine biosynthetic process		0.01	0.03	
oryzicola	Core exclusive	GO:0016614	molecular function	oxidoreductase activity, acting on CH-OH group of donors			0.03	
oryzicola	Core exclusive	GO:0016765	molecular function	transferase activity, transferring alkyl or aryl (other than methyl) groups		0.03		
oryzicola	Core exclusive	GO:0016779	molecular function	nucleotidyltransferase activity			0.05	

oryzicola	Core exclusive	GO:0016853	molecular function	isomerase activity	0.04	0.04		
oryzicola	Core exclusive	GO:0016998	biological process	cell wall macromolecule catabolic process		0.04		
oryzicola	Core exclusive	GO:0018659	molecular function	4-hydroxybenzoate 3-monooxygenase activity			0.02	
oryzicola	Core exclusive	GO:0019076	biological process	viral release from host cell			0.02	
oryzicola	Core exclusive	GO:0019835	biological process	cytolysis		0.01	0.03	
oryzicola	Core exclusive	GO:0020037	molecular function	heme binding		0.02		
oryzicola	Core exclusive	GO:0022857	molecular function	transmembrane transporter activity	0.05			0.04
oryzicola	Core exclusive	GO:0023014	biological process	signal transduction by protein phosphorylation		0.03		
oryzicola	Core exclusive	GO:0030245	biological process	cellulose catabolic process		0.02		
oryzicola	Core exclusive	GO:0031071	molecular function	cysteine desulfurase activity			0.01	
oryzicola	Core exclusive	GO:0031419	molecular function	cobalamin binding			0.01	
oryzicola	Core exclusive	GO:0042597	cellular component	periplasmic space			0.05	
oryzicola	Core exclusive	GO:0042742	biological process	defense response to bacterium			0.02	
oryzicola	Core exclusive	GO:0043190	cellular component	ATP-binding cassette (ABC) transporter complex	0.03			0.02
oryzicola	Core exclusive	GO:0044659	biological process	cytolysis by virus of host cell			0.02	
oryzicola	Core exclusive	GO:0046429	molecular function	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase activity		0.01	0.05	
oryzicola	Core exclusive	GO:0046677	biological process	response to antibiotic		0.01		
oryzicola	Core exclusive	GO:0046872	molecular function	metal ion binding	0.02			0.05
oryzicola	Core exclusive	GO:0051536	molecular function	iron-sulfur cluster binding		0.04		
oryzicola	Core exclusive	GO:0097056	biological process	selenocysteinyl-tRNA(Sec) biosynthetic process		0.01	0.03	
oryzicola	Core exclusive	GO:0098869	biological process	cellular oxidant detoxification		0.03		
oryzicola	Exclusive	GO:0003676	molecular function	nucleic acid binding		0.02		
oryzicola	Exclusive	GO:0003700	molecular function	DNA binding transcription factor activity		0.02		
oryzicola	Exclusive	GO:0003723	molecular function	RNA binding			0.02	
oryzicola	Exclusive	GO:0004096	molecular function	catalase activity		0.03	0.05	
oryzicola	Exclusive	GO:0004412	molecular function	homoserine dehydrogenase activity		0.02		
oryzicola	Exclusive	GO:0004601	molecular function	peroxidase activity		0.01	0.03	
oryzicola	Exclusive	GO:0004803	molecular function	transposase activity			0.01	
oryzicola	Exclusive	GO:0005506	molecular function	iron ion binding			0.02	
oryzicola	Exclusive	GO:0005524	molecular function	ATP binding			0.04	
oryzicola	Exclusive	GO:0005737	cellular component	cytoplasm		0.02		
oryzicola	Exclusive	GO:0005840	cellular component	ribosome			0.03	
oryzicola	Exclusive	GO:0005975	biological process	carbohydrate metabolic process		0.05		
oryzicola	Exclusive	GO:0006313	biological process	transposition, DNA-mediated			0.01	
oryzicola	Exclusive	GO:0006412	biological process	translation			0.02	

oryzicola	Exclusive	GO:0006950	biological process	response to stress	0.02	
oryzicola	Exclusive	GO:0008239	molecular function	dipeptidyl-peptidase activity		0.02
oryzicola	Exclusive	GO:0008299	biological process	isoprenoid biosynthetic process		0.01
oryzicola	Exclusive	GO:0008410	molecular function	CoA-transferase activity	0.01	0.03
oryzicola	Exclusive	GO:0008697	molecular function	4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase activity	0.05	
oryzicola	Exclusive	GO:0009086	biological process	methionine biosynthetic process	0.03	
oryzicola	Exclusive	GO:0009253	biological process	peptidoglycan catabolic process	0.03	
oryzicola	Exclusive	GO:0009435	biological process	NAD biosynthetic process		0.03
oryzicola	Exclusive	GO:0016740	molecular function	transferase activity	0.04	
oryzicola	Exclusive	GO:0016787	molecular function	hydrolase activity	0.03	
oryzicola	Exclusive	GO:0016998	biological process	cell wall macromolecule catabolic process	0.02	0.03
oryzicola	Exclusive	GO:0019357	biological process	nicotinate nucleotide biosynthetic process	0.04	
oryzicola	Exclusive	GO:0019835	biological process	cytolysis		0.01
oryzicola	Exclusive	GO:0020037	molecular function	heme binding		0.01
oryzicola	Exclusive	GO:0030245	biological process	cellulose catabolic process		0.02
oryzicola	Exclusive	GO:0043565	molecular function	sequence-specific DNA binding	0.03	
oryzicola	Exclusive	GO:0043639	biological process	benzoate catabolic process	0.04	
oryzicola	Exclusive	GO:0046917	molecular function	triphosphoribosyl-dephospho-CoA synthase activity	0.04	
oryzicola	Exclusive	GO:0047042	molecular function	androsterone dehydrogenase (B-specific) activity	0.04	
oryzicola	Exclusive	GO:0090305	biological process	nucleic acid phosphodiester bond hydrolysis	0.04	0.05

a = p-value is shown only for those terms identified as enriched in the sub-set for the given algorithm and when the value is equal or smaller than 0.05.

Table 6. Distribution of gene considered as significantly variable based on the z-score normalization derived from the number of “number of variable sites” / “gene length” calculated for each dataset of orthologues.

Dataset	*X. oryzae*	Pathovars	
		X. oryzae* pv. *oryzae	***X. oryzae* pv. *oryzicola***
BDBH	91	119	109
COG	355	573	304
OMCL	269	528	262
Consensus	64	92	93

1394 **Table 7.** Gene Ontology terms identified as enriched in each pathovar in the genes selected as differentially variable.

Pathovar	GO Term	Category	Name	p-values			
				BDBH	COG	OMCL	Consensus
					0.00	0.00	
						0.00	
oryzae	GO:0000166	nucleotide binding	molecular function		0.00	0.00	0.00
oryzae	GO:0000287	magnesium ion binding	molecular function		0.00	0.00	
oryzae	GO:0003700	DNA binding transcription factor activity	molecular function	0.02	0.00	0.00	
oryzae	GO:0003723	RNA binding	molecular function	0.00	0.00	0.00	
oryzae	GO:0003824	catalytic activity	molecular function	0.03	0.00	0.00	
oryzae	GO:0003924	GTPase activity	molecular function			0.01	
oryzae	GO:0004252	serine-type endopeptidase activity	molecular function			0.00	
oryzae	GO:0004386	helicase activity	molecular function		0.00	0.00	
oryzae	GO:0004518	nuclease activity	molecular function		0.00	0.00	
oryzae	GO:0004527	exonuclease activity	molecular function		0.03	0.05	
oryzae	GO:0004812	aminoacyl-tRNA ligase activity	molecular function		0.00	0.00	
oryzae	GO:0005506	iron ion binding	molecular function		0.00		
oryzae	GO:0005524	ATP binding	molecular function		0.00	0.00	0.03
oryzae	GO:0005622	intracellular	cellular component		0.00	0.00	
oryzae	GO:0005623	cell	cellular component		0.00		
oryzae	GO:0005737	cytoplasm	cellular component	0.00	0.00	0.00	0.00
oryzae	GO:0005840	ribosome	cellular component				0.03
oryzae	GO:0005886	plasma membrane	cellular component	0.04	0.00	0.00	
oryzae	GO:0005887	integral component of plasma membrane	cellular component		0.00	0.00	
oryzae	GO:0006260	DNA replication	biological process		0.02	0.04	
oryzae	GO:0006281	DNA repair	biological process		0.00	0.00	
oryzae	GO:0006351	transcription, DNA-templated	biological process	0.00	0.00	0.00	
oryzae	GO:0006355	regulation of transcription, DNA-templated	biological process	0.00	0.00	0.00	
oryzae	GO:0006412	translation	biological process		0.00	0.00	

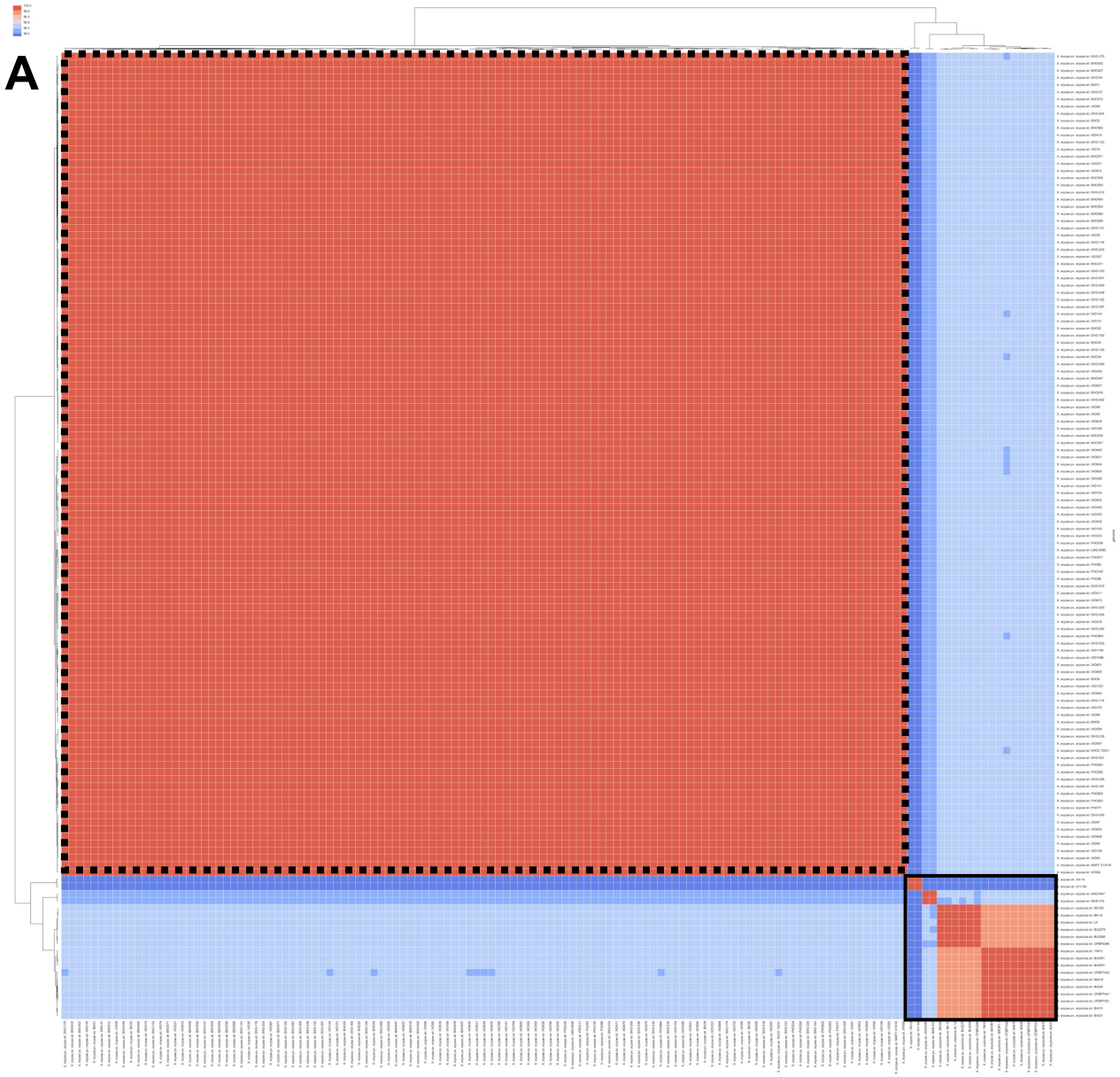
oryzae	GO:0006418	tRNA aminoacylation for protein translation	biological process		0.00	0.00	
oryzae	GO:0006457	protein folding	biological process			0.05	
oryzae	GO:0006508	proteolysis	biological process			0.00	
oryzae	GO:0006629	lipid metabolic process	biological process		0.00		
oryzae	GO:0006807	nitrogen compound metabolic process	biological process		0.03	0.05	
oryzae	GO:0006811	ion transport	biological process			0.00	
oryzae	GO:0007049	cell cycle	biological process		0.00	0.00	
oryzae	GO:0008033	tRNA processing	biological process			0.00	
oryzae	GO:0008152	metabolic process	biological process			0.00	
oryzae	GO:0008168	methyltransferase activity	molecular function		0.01	0.00	
oryzae	GO:0008237	metallopeptidase activity	molecular function			0.00	
oryzae	GO:0008270	zinc ion binding	molecular function		0.00	0.00	
oryzae	GO:0008360	regulation of cell shape	biological process		0.00	0.00	
oryzae	GO:0008652	cellular amino acid biosynthetic process	biological process		0.00	0.00	
oryzae	GO:0009055	electron transfer activity	molecular function			0.00	
oryzae	GO:0009058	biosynthetic process	biological process		0.00	0.00	
oryzae	GO:0009073	aromatic amino acid family biosynthetic process	biological process		0.00	0.01	
oryzae	GO:0009252	peptidoglycan biosynthetic process	biological process		0.00	0.00	
oryzae	GO:0009279	cell outer membrane	cellular component			0.00	
oryzae	GO:0009432	SOS response	biological process		0.03	0.05	
oryzae	GO:0015031	protein transport	biological process		0.00	0.00	
oryzae	GO:0015992	proton transport	biological process		0.03	0.05	
oryzae	GO:0016301	kinase activity	molecular function		0.00	0.00	
oryzae	GO:0016310	phosphorylation	biological process		0.01	0.00	0.02
oryzae	GO:0016491	oxidoreductase activity	molecular function	0.00	0.00	0.00	0.00
oryzae	GO:0016740	transferase activity	molecular function	0.00	0.00	0.00	0.00
oryzae	GO:0016746	transferase activity, transferring acyl groups	molecular function		0.00	0.00	

oryzae	GO:0016747	transferase activity, transferring acyl groups other than amino-acyl groups	molecular function			0.05	
oryzae	GO:0016757	transferase activity, transferring glycosyl groups	molecular function		0.00	0.00	
oryzae	GO:0016779	nucleotidyltransferase activity	molecular function		0.03	0.00	
oryzae	GO:0016787	hydrolase activity	molecular function		0.00	0.00	
oryzae	GO:0016829	lyase activity	molecular function	0.00	0.00	0.00	
oryzae	GO:0016853	isomerase activity	molecular function		0.02	0.00	0.04
oryzae	GO:0016874	ligase activity	molecular function		0.00	0.00	
oryzae	GO:0020037	heme binding	molecular function		0.00	0.00	
oryzae	GO:0022900	electron transport chain	biological process		0.00		
oryzae	GO:0030170	pyridoxal phosphate binding	molecular function		0.02		
oryzae	GO:0030288	outer membrane-bounded periplasmic space	cellular component		0.01		
oryzae	GO:0032259	methylation	biological process		0.01	0.00	
oryzae	GO:0042597	periplasmic space	cellular component		0.00		
oryzae	GO:0045454	cell redox homeostasis	biological process		0.00		
oryzae	GO:0046872	metal ion binding	molecular function	0.00	0.00	0.00	0.00
oryzae	GO:0050660	flavin adenine dinucleotide binding	molecular function			0.00	
oryzae	GO:0050661	NADP binding	molecular function			0.00	
oryzae	GO:0051287	NAD binding	molecular function		0.00		
oryzae	GO:0051301	cell division	biological process		0.00	0.00	
oryzae	GO:0051536	iron-sulfur cluster binding	molecular function		0.00	0.00	
oryzae	GO:0051539	4 iron, 4 sulfur cluster binding	molecular function		0.00	0.00	
oryzae	GO:0055114	oxidation-reduction process	biological process	0.00	0.00	0.00	0.00
oryzae	GO:0071555	cell wall organization	biological process		0.00	0.00	
oryzae	GO:0098869	cellular oxidant detoxification	biological process		0.00		
oryzicola	GO:0000049	tRNA binding	molecular function			0.02	
oryzicola	GO:0000166	nucleotide binding	molecular function		0.04	0.00	0.00
oryzicola	GO:0000287	magnesium ion binding	molecular function		0.01		

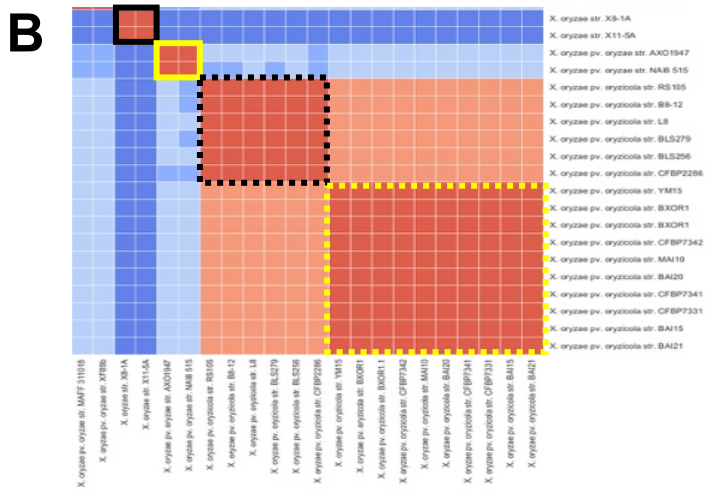
oryzicola	GO:0003700	DNA binding transcription factor activity	molecular function		0.00	0.00	
oryzicola	GO:0003723	RNA binding	molecular function		0.00	0.00	
oryzicola	GO:0003735	structural constituent of ribosome	molecular function	0.02	0.00	0.00	
oryzicola	GO:0004518	nuclease activity	molecular function			0.01	
oryzicola	GO:0005524	ATP binding	molecular function			0.00	0.00
oryzicola	GO:0005622	intracellular	cellular component			0.00	
oryzicola	GO:0005737	cytoplasm	cellular component	0.00	0.00	0.00	0.00
oryzicola	GO:0005840	ribosome	cellular component	0.01	0.00	0.00	0.03
oryzicola	GO:0005886	plasma membrane	cellular component	0.00	0.00	0.00	0.00
oryzicola	GO:0006281	DNA repair	biological process			0.00	
oryzicola	GO:0006351	transcription, DNA-templated	biological process	0.00	0.00		
oryzicola	GO:0006355	regulation of transcription, DNA-templated	biological process	0.02	0.02	0.00	
oryzicola	GO:0006412	translation	biological process	0.00	0.00	0.00	0.00
oryzicola	GO:0006974	cellular response to DNA damage stimulus	biological process			0.00	
oryzicola	GO:0008033	tRNA processing	biological process			0.01	
oryzicola	GO:0008270	zinc ion binding	molecular function		0.00	0.00	
oryzicola	GO:0008652	cellular amino acid biosynthetic process	biological process				0.02
oryzicola	GO:0015031	protein transport	biological process		0.00	0.00	
oryzicola	GO:0016310	phosphorylation	biological process			0.02	0.02
oryzicola	GO:0016491	oxidoreductase activity	molecular function		0.00		0.03
oryzicola	GO:0016740	transferase activity	molecular function	0.00	0.00	0.00	0.00
oryzicola	GO:0016746	transferase activity, transferring acyl groups	molecular function		0.01		
oryzicola	GO:0016787	hydrolase activity	molecular function		0.00	0.00	
oryzicola	GO:0016829	lyase activity	molecular function	0.00	0.00	0.00	0.01
oryzicola	GO:0016874	ligase activity	molecular function		0.00	0.00	0.02
oryzicola	GO:0030529	intracellular ribonucleoprotein complex	cellular component		0.00	0.00	

oryzicola	GO:0046872	metal ion binding	molecular function	0.00	0.00	0.00	0.00
oryzicola	GO:0051536	iron-sulfur cluster binding	molecular function		0.00	0.00	
oryzicola	GO:0051539	4 iron, 4 sulfur cluster binding	molecular function		0.00	0.00	
oryzicola	GO:0055114	oxidation-reduction process	biological process		0.00		
oryzicola	GO:0006355	regulation of transcription, DNA-templated	biological process	0.02	0.02	0.00	
oryzicola	GO:0006412	translation	biological process	0.00	0.00	0.00	0.00
oryzicola	GO:0006974	cellular response to DNA damage stimulus	biological process			0.00	
oryzicola	GO:0008033	tRNA processing	biological process			0.01	
oryzicola	GO:0008270	zinc ion binding	molecular function		0.00	0.00	
oryzicola	GO:0008652	cellular amino acid biosynthetic process	biological process				0.02
oryzicola	GO:0015031	protein transport	biological process		0.00	0.00	
oryzicola	GO:0016310	phosphorylation	biological process			0.02	0.02
oryzicola	GO:0016491	oxidoreductase activity	molecular function		0.00		0.03
oryzicola	GO:0016740	transferase activity	molecular function	0.00	0.00	0.00	0.00
oryzicola	GO:0016746	transferase activity, transferring acyl groups	molecular function		0.01		
oryzicola	GO:0016787	hydrolase activity	molecular function		0.00	0.00	
oryzicola	GO:0016829	lyase activity	molecular function	0.00	0.00	0.00	0.01
oryzicola	GO:0016874	ligase activity	molecular function		0.00	0.00	0.02
oryzicola	GO:0030529	intracellular ribonucleoprotein complex	cellular component		0.00	0.00	

a = only shown for those case were the GO term was identified as enriched in the sub-set


A

1399


B
X. oryzae str. X8-1A
X. oryzae str. X11-5A
X. oryzae pv. oryzae str. AXO1947
X. oryzae pv. oryzae str. NAIB 515
X. oryzae pv. oryzicola str. RS105
X. oryzae pv. oryzicola str. B8-12
X. oryzae pv. oryzicola str. L8
X. oryzae pv. oryzicola str. BLS279
X. oryzae pv. oryzicola str. BLS256
X. oryzae pv. oryzicola str. CFBP2286
X. oryzae pv. oryzicola str. YM15
X. oryzae pv. oryzicola str. BXOR1
X. oryzae pv. oryzicola str. BXOR1
X. oryzae pv. oryzicola str. CFBP7342
X. oryzae pv. oryzicola str. MAI10
X. oryzae pv. oryzicola str. BAI20
X. oryzae pv. oryzicola str. CFBP7341
X. oryzae pv. oryzicola str. CFBP7331
X. oryzae pv. oryzicola str. BAI15
X. oryzae pv. oryzicola str. BAI21

1400

Figure 1. Average Nucleotide Identity (ANI) matrix calculated based on the pairwise comparison of the different strains of *X. oryzae* from the pathovars *oryzae* and *oryzicola*. Calculations were based on bi-directional best hits that share at least 70% of sequence identity and 70% of coverage in both sequences. The final matrix was sorted at its axis using the Minkowski distance method implemented in Seaborn package. (A) Asian *Xoo* strains are indicated by a dashed black line square, while the others are indicated by the continuous black square. (B) American *Xoo* (continuous black line square), African *Xoo* (continuous black line square), African *Xoc* (dashed black line square) and Asian *Xoc* (dashed yellow line square).

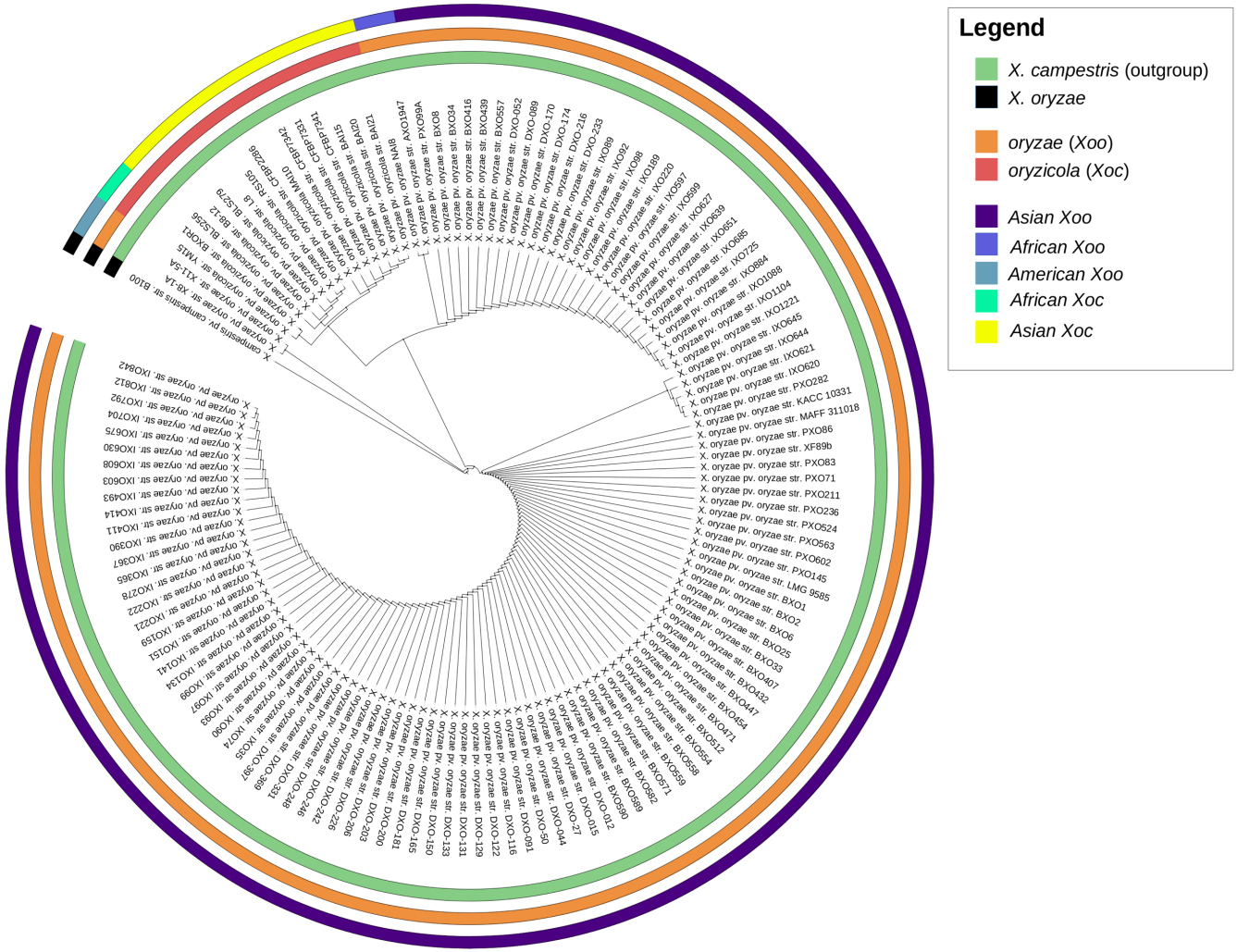


Figure 2. UPGMA phylogenetic tree generated based on core-genome alignment of different strains of *Xanthomonas oryzae*. A genome of *X. campestris* was used as outgroup. The final tree was produced using 1000 steps of Bootstrap and plotted using iTOL.

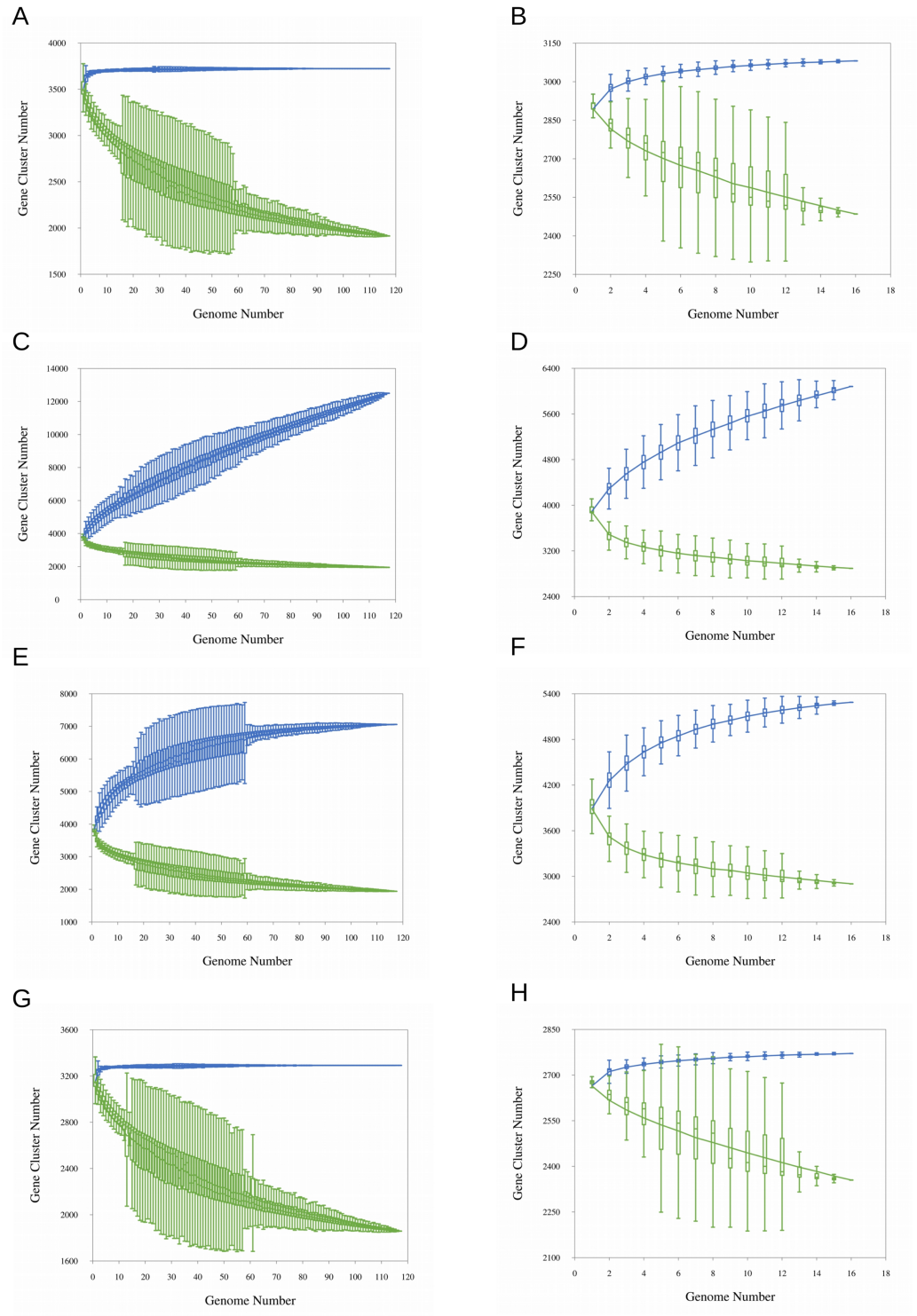
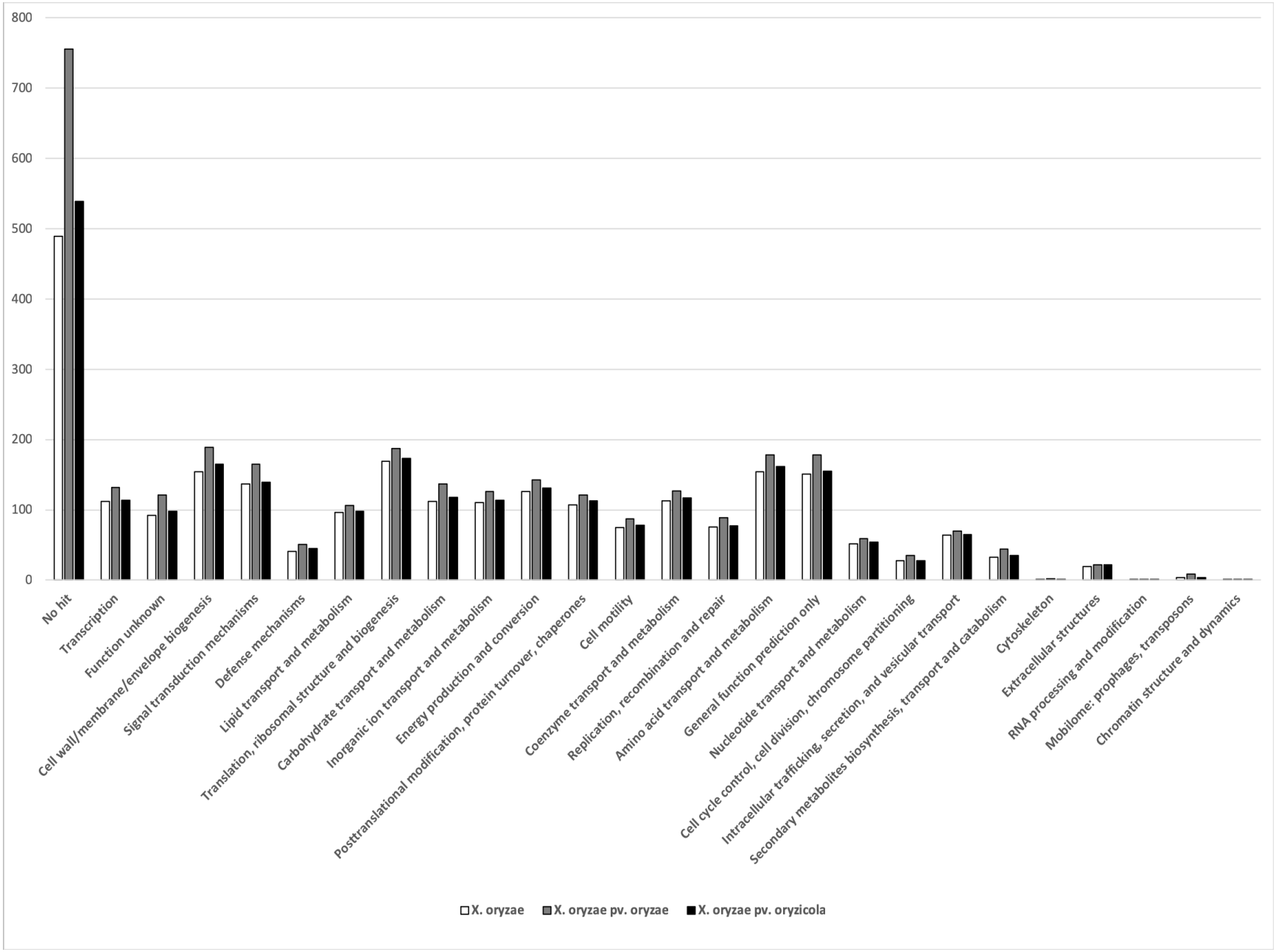


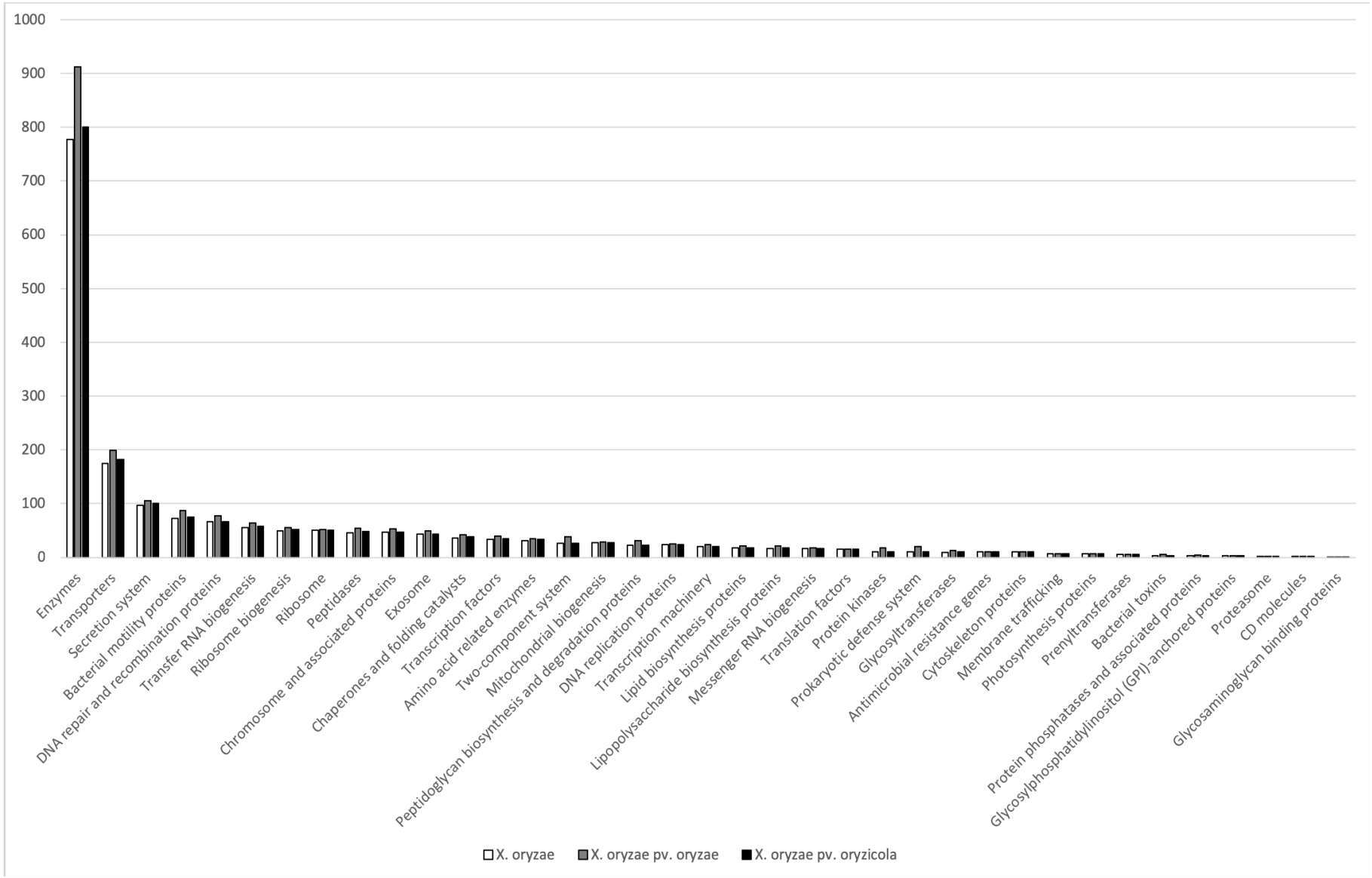
Figure 3. Cumulative distribution of the pangenome (pangenome) and core-genome (green) sizes using different sampling values calculated for the results generated by the algorithms BDBH, COG and OMCL, and their consensus for the pathovars *oryzae* and

oryzicola. (A) BDDH: *oryzae*; (B) BDBH: *oryzicola*; (C) COG: *oryzae*; (D): COG: *oryzicola*; (E) OMCL: *oryzae*; (F) OMCL: *oryzae*; (G) Consensus: *oryzae*; (H) Consensus: *oryzicola*. Plots generated using PanGP.


800
700
600
500
400
300
200
100
0
No hit
Transcription
Function unknown
Cell wall/membrane/envelope biogenesis
Signal transduction mechanisms
Defense mechanisms
Lipid transport and metabolism
Translation, ribosomal structure and biogenesis
Carbohydrate transport and metabolism
Inorganic ion transport and metabolism
Energy production and conversion
Posttranslational modification, protein turnover, chaperones
Cell motility
Coenzyme transport and metabolism
Replication, recombination and repair
Amino acid transport and metabolism
General function prediction only
Nucleotide transport and metabolism
Cell cycle control, cell division, chromosome partitioning
Intracellular trafficking, secretion, and vesicular transport
Secondary metabolites biosynthesis, transport and catabolism
Cytoskeleton
Extracellular structures
RNA processing and modification
Mobilome: prophages, transposons
Chromatin structure and dynamics
X. oryzae
X. oryzae pv. oryzae
X. oryzae pv. oryzicola

1423

Figure 4. Distribution of the annotation of the core genomes of *X. oryzae* (white), *Xoo* (grey) and *Xoc* (black) based on the Cluster of Ortholog Groups (COG) classification.


1000
900
800
700
600
500
400
300
200
100
0
Enzymes
Transporters
Secretion system
Bacterial motility proteins
DNA repair and recombination proteins
Transfer RNA biogenesis
Ribosome biogenesis
Ribosome
Peptidases
Chromosome and associated proteins
Exosome
Chaperones and folding catalysts
Transcription factors
Amino acid related enzymes
Two-component system
Mitochondrial biogenesis
Peptidoglycan biosynthesis and degradation proteins
DNA replication proteins
Transcription machinery
Lipid biosynthesis proteins
Lipopolysaccharide biosynthesis proteins
Messenger RNA biogenesis
Translation factors
Protein kinases
Prokaryotic defense system
Glycosyltransferases
Antimicrobial resistance genes
Cytoskeleton proteins
Membrane trafficking
Photosynthesis proteins
Prenyltransferases
Bacterial toxins
Protein phosphatases and associated proteins
Glycosylphosphatidylinositol (GPI)-anchored proteins
Proteasome
CD molecules
Glycosaminoglycan binding proteins
X. oryzae
X. oryzae pv. oryzae
X. oryzae pv. oryzicola

1426

Figure 5. Distribution of the annotation of the core genomes of *X. oryzae* (white), *Xoo* (grey) and *Xoc* (black) based on the KEGG pathway database BRIDE classification.

Supplementary Data 1.A InterPro annotation of the *X. oryzae* pv. *oryzae* (*Xoo*) exclusive core genome.

Cluster ID	InterPro Annotation
cluster:217820	Alpha-L-arabinofuranosidase, C-terminal (InterPro:IPR010720)
cluster:217824	
cluster:217825	YiaAB two helix (InterPro:IPR008024)
cluster:217826	Glucose/Sorbosone dehydrogenase (InterPro:IPR012938)
cluster:217851	
cluster:217871	
cluster:217878	Domain of unknown function DUF1730 (InterPro:IPR013542)
cluster:217881	N-acetylmuramoyl-L-alanine amidase, catalytic domain (InterPro:IPR002508); AMIN domain (InterPro:IPR021731)
cluster:217890	
cluster:217904	tRNA/rRNA methyltransferase, SpoU type (InterPro:IPR001537)
cluster:217909	
cluster:217912	
cluster:217916	
cluster:217924	
cluster:217927	Poly(R)-hydroxyalkanoic acid synthase, class III, PhaE subunit (InterPro:IPR010123)
cluster:217939	HAMP domain (InterPro:IPR003660); Signal transduction response regulator, receiver domain (InterPro:IPR001789); Cache 3/Cache 2 fusion domain (InterPro:IPR033462)
cluster:217962	Peptidase M23 (InterPro:IPR016047)
cluster:217966	DJ-1/PfpI (InterPro:IPR002818)
cluster:217979	
cluster:217980	
cluster:217989	Transcription regulator HTH, GntR (InterPro:IPR000524)
cluster:217996	EF-hand domain (InterPro:IPR002048)
cluster:217997	
cluster:218008	
cluster:218010	Transcription regulator HTH, LysR (InterPro:IPR000847); LysR, substrate-binding (InterPro:IPR005119)
cluster:218028	
cluster:218029	
cluster:218037	Endopeptidase, NLPC/P60 domain (InterPro:IPR000064)

cluster:218040	
cluster:218041	
cluster:218049	
cluster:218077	
cluster:218078	
cluster:218081	Oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase (InterPro:IPR018020)
cluster:218083	NodB homology domain (InterPro:IPR002509)
cluster:218085	Peptidase M20 (InterPro:IPR002933)
cluster:218092	Predicted virulence protein, SciE type (InterPro:IPR009211)
cluster:218094	Type VI secretion system, TssF-like (InterPro:IPR010272)
cluster:218097	OTT_1508-like deaminase (InterPro:IPR027796)
cluster:218108	Glyoxalase/fosfomycin resistance/dioxygenase domain (InterPro:IPR004360)
cluster:218109	Protein of unknown function DUF3182 (InterPro:IPR021519)
cluster:218111	
cluster:218116	
cluster:218117	
cluster:218118	Thioredoxin domain (InterPro:IPR013766)
cluster:218125	Aconitase/3-isopropylmalate dehydratase large subunit, alpha/beta/alpha domain (InterPro:IPR001030); Aconitase B, HEAT-like domain (InterPro:IPR015933); Aconitase B, swivel (InterPro:IPR015929)
cluster:218129	
cluster:218130	
cluster:218131	AMP-binding enzyme, C-terminal domain (InterPro:IPR025110); AMP-dependent synthetase/ligase (InterPro:IPR000873)
cluster:218140	Protein of unknown function DUF1176 (InterPro:IPR009560)
cluster:218141	Protein of unknown function DUF1176 (InterPro:IPR009560)
cluster:218142	Transposase IS204/IS1001/IS1096/IS1165, helix-turn-helix domain (InterPro:IPR032877)
cluster:218143	
cluster:218147	
cluster:218154	
cluster:218156	
cluster:218173	
cluster:218177	

cluster:218179	
cluster:218181	
cluster:218188	
cluster:218199	Bacterial Ig-related (InterPro:IPR011081); Coagulation factor 5/8 C-terminal domain (InterPro:IPR000421)
cluster:218213	Sulfatase-modifying factor enzyme (InterPro:IPR005532); Protein kinase domain (InterPro:IPR000719)
cluster:218217	Histidine kinase/HSP90-like ATPase (InterPro:IPR003594)
cluster:218220	Tn3 transposase DDE domain (InterPro:IPR002513)
cluster:218240	
cluster:218245	
cluster:218247	
cluster:218251	Beta-L-arabinofuranosidase, GH127 (InterPro:IPR012878)
cluster:218252	
cluster:218253	TonB-dependent receptor-like, beta-barrel (InterPro:IPR000531)
cluster:218254	
cluster:218255	
cluster:218258	Sporulation-like domain (InterPro:IPR007730); RlpA-like protein, double-psi beta-barrel domain (InterPro:IPR009009)
cluster:218260	
cluster:218272	Glycine zipper 2TM domain (InterPro:IPR008816)
cluster:218284	Integrase, SAM-like, N-terminal (InterPro:IPR004107)
cluster:218293	SCP2 sterol-binding domain (InterPro:IPR003033)
cluster:218297	Domain of unknown function DUF2059 (InterPro:IPR018637)
cluster:218302	Transposase IS200-like (InterPro:IPR002686)
cluster:218310	
cluster:218312	Cytochrome c-like domain (InterPro:IPR009056)
cluster:218353	Beta-ketoacyl synthase, C-terminal (InterPro:IPR014031); Beta-ketoacyl synthase, N-terminal (InterPro:IPR014030)
cluster:218354	
cluster:218357	
cluster:218359	Copper resistance B precursor (InterPro:IPR007939)
cluster:218365	ATP-grasp fold, ATP-dependent carboxylate-amine ligase-type (InterPro:IPR003135)

cluster:218366	Band 7 domain (InterPro:IPR001107)
cluster:218371	Protein of unknown function DUF423 (InterPro:IPR006696)
cluster:218387	
cluster:218398	Translocation and assembly module TamB (InterPro:IPR007452)
cluster:218400	
cluster:218402	Radical SAM (InterPro:IPR007197); 4Fe4S-binding SPASM domain (InterPro:IPR023885)
cluster:218404	Thiaminase-2/PQQC (InterPro:IPR004305)
cluster:218409	Bifunctional transglycosylase second domain (InterPro:IPR028166); Penicillin-binding protein, transpeptidase (InterPro:IPR001460); Glycosyl transferase, family 51 (InterPro:IPR001264)
cluster:218410	
cluster:218420	Signal transduction histidine kinase, phosphotransfer (Hpt) domain (InterPro:IPR008207); Histidine kinase CheA-like, homodimeric domain (InterPro:IPR004105); Histidine kinase/HSP90-like ATPase (InterPro:IPR003594); Signal transduction response regulator, receiver domain (InterPro:IPR001789); CheW-like domain (InterPro:IPR002545)
cluster:218434	MetA-pathway of phenol degradation, putative (InterPro:IPR025737)
cluster:218436	Peptidase M20 (InterPro:IPR002933); Peptidase M20, dimerisation domain (InterPro:IPR011650)
cluster:218449	Signal transduction response regulator, receiver domain (InterPro:IPR001789)
cluster:218450	Signal transduction histidine kinase, dimerisation/phosphoacceptor domain (InterPro:IPR003661)
cluster:218460	Biotin/lipoyl attachment (InterPro:IPR000089); 2-oxoacid dehydrogenase acyltransferase, catalytic domain (InterPro:IPR001078); Peripheral subunit-binding domain (InterPro:IPR004167)
cluster:218477	Histidine kinase/HSP90-like ATPase (InterPro:IPR003594)
cluster:218492	
cluster:218512	DNA recombination-mediator protein A (InterPro:IPR003488)
cluster:218515	GYF domain 2 (InterPro:IPR025640); RDD (InterPro:IPR010432)
cluster:218534	
cluster:218539	Acriflavin resistance protein (InterPro:IPR001036)
cluster:218544	
cluster:218545	Methylpurine-DNA glycosylase (InterPro:IPR003180)
cluster:218573	2-keto-3-deoxygluconate permease (InterPro:IPR004684)
cluster:218574	Phosphate-selective porin O/P (InterPro:IPR010870)

cluster:218575	Phosphate-selective porin O/P (InterPro:IPR010870)
cluster:218594	Pectinesterase, catalytic (InterPro:IPR000070)
cluster:218597	Ethanolamine ammonia-lyase light chain (InterPro:IPR009246)
cluster:218602	Tail specific protease (InterPro:IPR005151)
cluster:218607	Aspartate/ornithine carbamoyltransferase, carbamoyl-P binding (InterPro:IPR006132); Aspartate/ornithine carbamoyltransferase, Asp/Orn-binding domain (InterPro:IPR006131)
cluster:218615	YCII-related (InterPro:IPR005545)
cluster:218619	Major facilitator superfamily (InterPro:IPR011701)
cluster:218620	Cytochrome ubiquinol oxidase subunit 2 (InterPro:IPR003317)
cluster:218621	Cytochrome ubiquinol oxidase subunit 1 (InterPro:IPR002585)
cluster:218627	Zinc finger, FPG/IleRS-type (InterPro:IPR010663); Formamidopyrimidine-DNA glycosylase, catalytic domain (InterPro:IPR012319); DNA glycosylase/AP lyase, H2TH DNA-binding (InterPro:IPR015886)
cluster:218638	
cluster:218645	
cluster:218648	Chorismate synthase (InterPro:IPR000453)
cluster:218664	Radical SAM (InterPro:IPR007197)
cluster:218684	Phosphoesterase (InterPro:IPR007312)
cluster:218685	Phosphoesterase (InterPro:IPR007312)
cluster:218686	Bacterial phospholipase C, C-terminal domain (InterPro:IPR008475); Phosphoesterase (InterPro:IPR007312)
cluster:218690	GGDEF domain (InterPro:IPR000160); GAF domain (InterPro:IPR003018)
cluster:218692	CHASE3 (InterPro:IPR007891)
cluster:218696	Outer membrane protein, OmpW (InterPro:IPR005618)
cluster:218697	
cluster:218702	Thioredoxin-like fold (InterPro:IPR012336); Disulphide bond isomerase, DsbC/G, N-terminal (InterPro:IPR018950)
cluster:218714	
cluster:218718	Protein of unknown function DUF2339, transmembrane (InterPro:IPR019286)
cluster:218719	Protein of unknown function DUF3999 (InterPro:IPR025060)
cluster:218728	Tetrapyrrole methylase (InterPro:IPR000878)
cluster:218734	PepSY-associated TM protein (InterPro:IPR005625)
cluster:218749	DHHA1 domain (InterPro:IPR003156); DDH domain (InterPro:IPR001667)

cluster:218750	
cluster:218759	Penicillin-binding protein, dimerisation domain (InterPro:IPR005311); Penicillin-binding protein, transpeptidase (InterPro:IPR001460)
cluster:218761	Rod shape-determining protein MreC (InterPro:IPR007221)
cluster:218795	Protein of unknown function DUF4380 (InterPro:IPR025488)
cluster:218805	Chaperone DnaJ, C-terminal (InterPro:IPR002939); DnaJ domain (InterPro:IPR001623)
cluster:218812	3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III, C-terminal (InterPro:IPR013747); 3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III (InterPro:IPR013751)
cluster:218820	
cluster:218825	MCP methyltransferase, CheR-type, SAM-binding domain, C-terminal (InterPro:IPR022642); Signal transduction response regulator, chemotaxis, protein-glutamate methylesterase (InterPro:IPR000673); Chemotaxis receptor methyltransferase CheR, N-terminal (InterPro:IPR022641)
cluster:218827	
cluster:218838	Cupin 2, conserved barrel (InterPro:IPR013096)
cluster:218839	
cluster:218843	
cluster:218844	Glycosyl hydrolase family 95, N-terminal domain (InterPro:IPR027414)
cluster:218845	Alpha-glucosidase, domain of unknown function DUF4968 (InterPro:IPR032513); Domain of unknown function DUF5110 (InterPro:IPR033403); PA14 domain (InterPro:IPR011658); Glycoside hydrolase family 31 (InterPro:IPR000322)
cluster:218846	
cluster:218851	
cluster:218856	Zinc finger, FPG/IleRS-type (InterPro:IPR010663)
cluster:218860	Avidin/streptavidin (InterPro:IPR005468)
cluster:218867	1, 4-beta cellobiohydrolase (InterPro:IPR016288)
cluster:218874	
cluster:218875	
cluster:218879	Short-chain dehydrogenase/reductase SDR (InterPro:IPR002347)
cluster:218880	
cluster:218882	Major facilitator superfamily (InterPro:IPR011701)
cluster:218885	PhnA protein, N-terminal (InterPro:IPR013987); PhnA protein, C-terminal (InterPro:IPR013988)

cluster:218886	Cation efflux protein (InterPro:IPR002524)
cluster:218897	
cluster:218900	GGDEF domain (InterPro:IPR000160)
cluster:218902	Virulence-associated protein D / CRISPR associated protein Cas2 (InterPro:IPR019199)
cluster:218903	CRISPR-associated protein Cas1 (InterPro:IPR002729)
cluster:218904	Dna2/Cas4, domain of unknown function DUF83 (InterPro:IPR022765)
cluster:218905	CRISPR-associated protein Cas7, subtype I-B/I-C (InterPro:IPR006482)
cluster:218906	CRISPR-associated protein, Csd1-type (InterPro:IPR010144)
cluster:218907	CRISPR-associated protein, Cas5 (InterPro:IPR021124)
cluster:218908	DEAD/DEAH box helicase domain (InterPro:IPR011545)
cluster:218911	Pyrophosphate-energised proton pump (InterPro:IPR004131)
cluster:218912	
cluster:218925	Bacterial lipid A biosynthesis acyltransferase (InterPro:IPR004960)
cluster:218931	
cluster:218942	Type IV / VI secretion system, DotU (InterPro:IPR017732)
cluster:218959	
cluster:219000	Penicillin-binding protein, OB-like domain (InterPro:IPR031376); Glycosyl transferase, family 51 (InterPro:IPR001264); Penicillin-binding protein, transpeptidase (InterPro:IPR001460)
cluster:219002	Fimbrial assembly PilN (InterPro:IPR007813)
cluster:219014	
cluster:219019	
cluster:219034	Protein of unknown function DUF3658 (InterPro:IPR022123)
cluster:219035	Sodium:dicarboxylate symporter (InterPro:IPR001991)
cluster:219037	
cluster:219038	Peptidase M23 (InterPro:IPR016047)
cluster:219039	Peptidoglycan binding-like (InterPro:IPR002477); L,D-transpeptidase catalytic domain (InterPro:IPR005490)
cluster:219045	AMP-dependent synthetase/ligase (InterPro:IPR000873)
cluster:219051	Histidine kinase/HSP90-like ATPase (InterPro:IPR003594); Signal transduction histidine kinase, dimerisation/phosphoacceptor domain (InterPro:IPR003661); Two-component sensor kinase, N-terminal (InterPro:IPR013727)
cluster:219054	
cluster:219061	Putative phosphate transport regulator (InterPro:IPR018445)

cluster:219067	Signal recognition particle, SRP54 subunit, GTPase domain (InterPro:IPR000897)
cluster:219070	GGDEF domain (InterPro:IPR000160); MHYT domain (InterPro:IPR005330); EAL domain (InterPro:IPR001633)
cluster:219088	Glycosyltransferase 2-like (InterPro:IPR001173)
cluster:219089	Methyltransferase FkbM (InterPro:IPR006342)
cluster:219097	DegT/DnrJ/EryC1/StrS aminotransferase (InterPro:IPR000653)
cluster:219114	Flagellar basal-body/hook protein, C-terminal domain (InterPro:IPR010930); Flagellar basal body rod protein, N-terminal (InterPro:IPR001444); Flagellar hook protein FlgE (InterPro:IPR011491)
cluster:219127	
cluster:219135	Leucyl/phenylalanyl-tRNA-protein transferase (InterPro:IPR004616)
cluster:219136	
cluster:219137	
cluster:219138	3-beta hydroxysteroid dehydrogenase/isomerase (InterPro:IPR002225)
cluster:219164	
cluster:219165	PPM-type phosphatase domain (InterPro:IPR001932); Double Cache domain 1 (InterPro:IPR033479); HAMP domain (InterPro:IPR003660)
cluster:219166	Histidine kinase/HSP90-like ATPase (InterPro:IPR003594)
cluster:219167	STAS domain (InterPro:IPR002645)
cluster:219170	Mitochondrial fission protein ELM1-like (InterPro:IPR009367)
cluster:219175	Peptidase S53, activation domain (InterPro:IPR015366)
cluster:219179	SMP-30/Gluconolactonase/LRE-like region (InterPro:IPR013658)
cluster:219180	
cluster:219182	
cluster:219196	RNA polymerase sigma-70 region 2 (InterPro:IPR007627)
cluster:219249	
cluster:219250	
cluster:219259	Protein of unknown function DUF3391 (InterPro:IPR021812)
cluster:219261	
cluster:219262	
cluster:219263	
cluster:219265	Calcineurin-like phosphoesterase, N-terminal (InterPro:IPR032285)
cluster:219272	Histidinol dehydrogenase (InterPro:IPR012131)

cluster:219276	
cluster:219284	RNA-binding S4 domain (InterPro:IPR002942)
cluster:219285	
cluster:219286	CheW-like domain (InterPro:IPR002545); Histidine kinase CheA-like, homodimeric domain (InterPro:IPR004105); Histidine kinase/HSP90-like ATPase (InterPro:IPR003594); Signal transduction histidine kinase, phosphotransfer (Hpt) domain (InterPro:IPR008207)
cluster:219291	
cluster:219293	
cluster:219294	TonB-dependent receptor-like, beta-barrel (InterPro:IPR000531)
cluster:219295	
cluster:219299	
cluster:219319	
cluster:219320	
cluster:219327	Aminotransferase, class I/classII (InterPro:IPR004839)
cluster:219339	Domain of unknown function DUF3971 (InterPro:IPR025263); AsmA-like, C-terminal (InterPro:IPR032712)
cluster:219340	Rhamnogalacturonan lyase, domain III (InterPro:IPR029411); Rhamnogalacturonase B, N-terminal (InterPro:IPR015364); Rhamnogalacturonan lyase, domain II (InterPro:IPR029413)
cluster:219347	Domain of unknown function DUF58 (InterPro:IPR002881)
cluster:219373	4'-phosphopantetheinyl transferase domain (InterPro:IPR008278)
cluster:219385	Benzoate transporter (InterPro:IPR004711)
cluster:219386	Benzoate transporter (InterPro:IPR004711)
cluster:219387	GAF domain (InterPro:IPR003018); PAS fold-2 (InterPro:IPR013654); Phytochrome, central region (InterPro:IPR013515); PAS fold-4 (InterPro:IPR013656)
cluster:219393	
cluster:219420	
cluster:219438	FAD/NAD(P)-binding domain (InterPro:IPR023753); NADH:flavin oxidoreductase/NADH oxidase, N-terminal (InterPro:IPR001155)
cluster:219441	
cluster:219450	Alcohol dehydrogenase, C-terminal (InterPro:IPR013149); Alcohol dehydrogenase, N-terminal (InterPro:IPR013154)
cluster:219453	Peptidase S46 (InterPro:IPR019500)
cluster:219455	

cluster:219457	Sporulation-like domain (InterPro:IPR007730)
cluster:219460	Protein of unknown function DUF455 (InterPro:IPR007402)
cluster:219467	Signal transduction histidine kinase, dimerisation/phosphoacceptor domain (InterPro:IPR003661); Histidine kinase/HSP90-like ATPase (InterPro:IPR003594); Phosphate regulon sensor protein PhoR (InterPro:IPR021766)
cluster:219469	Peptidase M48 (InterPro:IPR001915)
cluster:219473	Domain of unknown function DUF4123 (InterPro:IPR025391)
cluster:219484	Aminotransferase class-III (InterPro:IPR005814)
cluster:219492	Glycine cleavage system H-protein/Simiate (InterPro:IPR033753)
cluster:219495	Major facilitator superfamily (InterPro:IPR011701)
cluster:219502	Heat shock protein Hsp33 (InterPro:IPR000397)
cluster:219505	
cluster:219506	CBS domain (InterPro:IPR000644)
cluster:219507	AsmA (InterPro:IPR007844)
cluster:219519	Lipid/polyisoprenoid-binding, YceI-like (InterPro:IPR007372)
cluster:219553	
cluster:219563	
cluster:219566	
cluster:219567	Domain of unknown function DUF2220 (InterPro:IPR024534); Domain of unknown function DUF3322 (InterPro:IPR024537)
cluster:219595	Sporulation-like domain (InterPro:IPR007730)
cluster:219600	Alpha-D-phosphohexomutase, alpha/beta/alpha domain III (InterPro:IPR005846); Alpha-D-phosphohexomutase, alpha/beta/alpha domain I (InterPro:IPR005844); Alpha-D-phosphohexomutase, alpha/beta/alpha domain II (InterPro:IPR005845); Alpha-D-phosphohexomutase, C-terminal (InterPro:IPR005843)
cluster:219621	Protein of unknown function DUF3300 (InterPro:IPR021728)
cluster:219623	Zinc/iron permease (InterPro:IPR003689)
cluster:219631	
cluster:219647	Protein of unknown function DUF1800 (InterPro:IPR014917)
cluster:219651	
cluster:219653	Type III pantothenate kinase (InterPro:IPR004619)
cluster:219656	
cluster:219657	

cluster:219662	CsbD-like domain (InterPro:IPR008462)
cluster:219667	EamA domain (InterPro:IPR000620)
cluster:219673	Peptidase S9, prolyl oligopeptidase, catalytic domain (InterPro:IPR001375)
cluster:219684	Peptidase S8/S53 domain (InterPro:IPR000209)
cluster:219689	
cluster:219691	
cluster:219701	Glycosyltransferase family 28, N-terminal domain (InterPro:IPR004276); Glycosyl transferase, family 28, C-terminal (InterPro:IPR007235)
cluster:219711	Post-segregation antitoxin CcdA (InterPro:IPR009956)
cluster:219718	UspA (InterPro:IPR006016); Signal transduction histidine kinase, osmosensitive K+ channel sensor, N-terminal (InterPro:IPR003852); Sensor protein KdpD, transmembrane domain (InterPro:IPR025201); Signal transduction histidine kinase, dimerisation/phosphoacceptor domain (InterPro:IPR003661); Histidine kinase/HSP90-like ATPase (InterPro:IPR003594)
cluster:219724	Xaa-Pro dipeptidyl-peptidase, C-terminal (InterPro:IPR013736)
cluster:219728	Lumazine/riboflavin synthase (InterPro:IPR002180)
cluster:219733	GNAT domain (InterPro:IPR000182)
cluster:219735	
cluster:219741	Protein of unknown function DUF4386 (InterPro:IPR025495)
cluster:219744	Glutathione-dependent formaldehyde-activating enzyme/centromere protein V (InterPro:IPR006913)
cluster:219745	Alcohol dehydrogenase, N-terminal (InterPro:IPR013154); Alcohol dehydrogenase, C-terminal (InterPro:IPR013149)
cluster:219752	
cluster:219783	TraB family (InterPro:IPR002816)
cluster:219805	
cluster:219813	Abortive infection system protein AbiD/AbiF-like (InterPro:IPR011664)
cluster:219814	EcoEI R protein C-terminal domain (InterPro:IPR013670); Helicase, C-terminal (InterPro:IPR001650); Helicase/UvrB, N-terminal (InterPro:IPR006935); Domain of unknown function DUF4145 (InterPro:IPR025285)
cluster:219828	DNA-binding HTH domain, TetR-type (InterPro:IPR001647)
cluster:219832	
cluster:219838	Bacterial DNA polymerase III, alpha subunit, NTPase domain (InterPro:IPR011708); DNA polymerase, helix-hairpin-helix motif (InterPro:IPR029460); PHP domain (InterPro:IPR004013)

cluster:219848	Cobinamide kinase/cobinamide phosphate guanyltransferase (InterPro:IPR003203)
cluster:219866	Protein of unknown function DUF4870 (InterPro:IPR019109)
cluster:219870	
cluster:219871	Serine proteases, trypsin domain (InterPro:IPR001254)
cluster:219896	
cluster:219907	Protein of unknown function DUF2069, membrane (InterPro:IPR018643)
cluster:219908	Glycosyl hydrolase family 30, beta sandwich domain (InterPro:IPR033452); Glycosyl hydrolase family 30, TIM-barrel domain (InterPro:IPR033453)
cluster:219911	
cluster:219915	
cluster:219921	
cluster:219937	Ribbon-helix-helix protein, CopG (InterPro:IPR002145)
cluster:219938	Ribbon-helix-helix protein, CopG (InterPro:IPR002145)
cluster:219939	Toxin-antitoxin system, RelE/ParE toxin family (InterPro:IPR007712)
cluster:219940	
cluster:219941	
cluster:219950	Integrase, SAM-like, N-terminal (InterPro:IPR004107); Integrase, catalytic domain (InterPro:IPR002104)
cluster:219974	Chaperone DnaJ, C-terminal (InterPro:IPR002939); DnaJ domain (InterPro:IPR001623); Heat shock protein DnaJ, cysteine-rich domain (InterPro:IPR001305)
cluster:219984	Peptidase S8/S53 domain (InterPro:IPR000209)
cluster:219985	
cluster:219990	
cluster:219991	
cluster:220000	TonB-dependent receptor, plug domain (InterPro:IPR012910); TonB-dependent receptor-like, beta-barrel (InterPro:IPR000531)
cluster:220001	Transposase InsH, N-terminal (InterPro:IPR008490)
cluster:220010	Peptidase M50 (InterPro:IPR008915)
cluster:220039	Lytic transglycosylase, superhelical linker (InterPro:IPR012289); Transglycosylase SLT domain 1 (InterPro:IPR008258)
cluster:220067	TonB-dependent receptor, plug domain (InterPro:IPR012910); TonB-dependent receptor-like, beta-barrel (InterPro:IPR000531)
cluster:220077	Transketolase-like, pyrimidine-binding domain (InterPro:IPR005475)

cluster:220082	
cluster:220090	
cluster:220092	Flavodoxin-like fold (InterPro:IPR003680)
cluster:220109	
cluster:220122	
cluster:220129	Rho termination factor, N-terminal (InterPro:IPR011112); ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding domain (InterPro:IPR000194); Rho termination factor, RNA-binding domain (InterPro:IPR011113)
cluster:220131	ATP-dependent RNA helicase RhlB (InterPro:IPR022077); Helicase, C-terminal (InterPro:IPR001650); DEAD/DEAH box helicase domain (InterPro:IPR011545)
cluster:220137	Uncharacterised domain UPF0126 (InterPro:IPR005115)
cluster:220144	Pseudouridine synthase, RsuA/RluB/C/D/E/F (InterPro:IPR006145); RNA-binding S4 domain (InterPro:IPR002942)
cluster:220151	Domain of unknown function DUF2329 (InterPro:IPR019282)
cluster:220170	Polysaccharide biosynthesis protein, CapD-like domain (InterPro:IPR003869)
cluster:220173	
cluster:220187	Protein of unknown function DUF805 (InterPro:IPR008523)
cluster:220193	Helicase/UvrB, N-terminal (InterPro:IPR006935)
cluster:220205	
cluster:220212	Cation efflux protein (InterPro:IPR002524)
cluster:220213	Metal-sensitive transcriptional repressor (InterPro:IPR003735)
cluster:220219	TonB, C-terminal (InterPro:IPR037682)
cluster:220220	Flavin-dependent halogenase (InterPro:IPR006905)
cluster:220223	TonB-dependent receptor, plug domain (InterPro:IPR012910); TonB-dependent receptor-like, beta-barrel (InterPro:IPR000531)
cluster:220226	Protein of unknown function DUF885 (InterPro:IPR010281)
cluster:220228	
cluster:220253	Phage tail collar domain (InterPro:IPR011083)
cluster:220256	Phage tail collar domain (InterPro:IPR011083)
cluster:220260	
cluster:220261	

cluster:220262	UvrD-like DNA helicase, C-terminal (InterPro:IPR014017); UvrD/AddA helicase, N-terminal (InterPro:IPR034739); PD-(D/E)XK endonuclease-like domain, AddAB-type (InterPro:IPR038726)
cluster:220264	Toxin HigB-1 (InterPro:IPR007711)
cluster:220274	
cluster:220275	
cluster:220282	
cluster:220301	POTRA domain, BamA/TamA-like (InterPro:IPR010827); Bacterial surface antigen (D15) (InterPro:IPR000184); TamA, POTRA domain 1 (InterPro:IPR035243)
cluster:220316	Glycosyl hydrolase family 67, C-terminal (InterPro:IPR011099); Glycosyl hydrolase family 67, catalytic domain (InterPro:IPR011100); Alpha glucuronidase, N-terminal (InterPro:IPR005154)
cluster:220317	Glycosyl hydrolases family 2, sugar binding domain (InterPro:IPR006104); Sialate O-acetylesterase domain (InterPro:IPR005181)
cluster:220322	Transposase, IS4-like (InterPro:IPR002559)
cluster:220338	Lactate/malate dehydrogenase, N-terminal (InterPro:IPR001236); Lactate/malate dehydrogenase, C-terminal (InterPro:IPR022383)
cluster:220342	Amidase signature domain (InterPro:IPR023631)
cluster:220348	Dihydroneopterin aldolase/epimerase domain (InterPro:IPR006157)
cluster:220349	Gcp-like domain (InterPro:IPR000905)
cluster:220351	Uncharacterised protein YqeY/AIM41 (InterPro:IPR019004)
cluster:220356	Glutathione S-transferase, N-terminal (InterPro:IPR004045)
cluster:220380	
cluster:220388	
cluster:220389	
cluster:220391	
cluster:220392	Peptidase T2, asparaginase 2 (InterPro:IPR000246)
cluster:220401	Metal-independent alpha-mannosidase (InterPro:IPR008313)
cluster:220411	Phosphatidylglycerol lysyltransferase, C-terminal (InterPro:IPR024320); Lysylphosphatidylglycerol synthetase/glycosyltransferase AglD (InterPro:IPR022791)
cluster:220436	tRNA/rRNA methyltransferase, SpoU type (InterPro:IPR001537); RNA 2-O ribose methyltransferase, substrate binding (InterPro:IPR013123)
cluster:220454	
cluster:220466	Serine aminopeptidase, S33 (InterPro:IPR022742)

cluster:220469	
cluster:220471	FAD-binding domain (InterPro:IPR002938)
cluster:220476	Ankyrin repeat-containing domain (InterPro:IPR020683); Ankyrin repeat (InterPro:IPR002110)
cluster:220489	Methylguanine DNA methyltransferase, ribonuclease-like domain (InterPro:IPR008332); Methylated-DNA-[protein]-cysteine S-methyltransferase, DNA binding (InterPro:IPR014048)
cluster:220497	
cluster:220515	Plasmid pRiA4b, Orf3 (InterPro:IPR012912)
cluster:220523	
cluster:220526	SsuA/THI5-like (InterPro:IPR015168)
cluster:220527	
cluster:220551	Siroheme synthase, central domain (InterPro:IPR028281); Sirohaem synthase, dimerisation domain (InterPro:IPR019478); Tetrapyrrole methylase (InterPro:IPR000878)
cluster:220553	HTH-type transcriptional regulator AraC-type, N-terminal (InterPro:IPR032687)
cluster:220554	
cluster:220564	PEP-utilising enzyme, C-terminal (InterPro:IPR000121); Phosphocarrier protein HPr-like (InterPro:IPR000032); Phosphotransferase system, enzyme I N-terminal (InterPro:IPR008731); PEP-utilising enzyme, mobile domain (InterPro:IPR008279); PTS EIIA type-2 domain (InterPro:IPR002178)
cluster:220566	Phosphotransferase system, EIIB component, type 2/3 (InterPro:IPR003501)
cluster:220568	
cluster:220600	Preprotein translocase SecG subunit (InterPro:IPR004692)
cluster:220618	
cluster:220622	PepSY-associated TM protein (InterPro:IPR005625)
cluster:220623	PepSY-associated TM protein (InterPro:IPR005625)
cluster:220624	
cluster:220625	Protein of unknown function DUF3325 (InterPro:IPR021762)
cluster:220634	Glycoside hydrolase, family 43 (InterPro:IPR006710)
cluster:220635	NADP-dependent oxidoreductase domain (InterPro:IPR023210)
cluster:220636	Amidohydrolase-related (InterPro:IPR006680)
cluster:220638	Fumarylacetoacetase-like, C-terminal (InterPro:IPR011234)
cluster:220639	Enolase C-terminal domain-like (InterPro:IPR029065); Mandelate racemase/muconate lactonizing enzyme, N-terminal domain (InterPro:IPR013341)

cluster:220641	Helicase-associated domain (InterPro:IPR007502); Helicase, C-terminal (InterPro:IPR001650); ATP-dependent RNA helicase HrpB, C-terminal (InterPro:IPR013689)
cluster:220649	Alpha-1,4-glucan:maltose-1-phosphate maltosyltransferase, domain N/S (InterPro:IPR021828); Glycosyl hydrolase, family 13, catalytic domain (InterPro:IPR006047)
cluster:220650	Maltogenic Amylase, C-terminal (InterPro:IPR032091); Glycosyl hydrolase, family 13, catalytic domain (InterPro:IPR006047)
cluster:220652	Calcineurin-like phosphoesterase domain, lpxH type (InterPro:IPR024654)
cluster:220653	
cluster:220654	
cluster:220655	
cluster:220656	
cluster:220658	KduI/IolB isomerase (InterPro:IPR021120)
cluster:220659	S-adenosylmethionine synthetase, N-terminal (InterPro:IPR022628); S-adenosylmethionine synthetase, C-terminal (InterPro:IPR022630); S-adenosylmethionine synthetase, central domain (InterPro:IPR022629)
cluster:220660	LexA-binding, inner membrane-associated putative hydrolase (InterPro:IPR007404)
cluster:220661	
cluster:220662	Methyltransferase type 11 (InterPro:IPR013216)
cluster:220669	STAS domain (InterPro:IPR002645)
cluster:220690	
cluster:220697	THIF-type NAD/FAD binding fold (InterPro:IPR000594)
cluster:220700	Glycine zipper 2TM domain (InterPro:IPR008816)
cluster:220703	
cluster:220705	
cluster:220729	
cluster:220736	Sulfur carrier ThiS/MoaD-like (InterPro:IPR003749)
cluster:220737	Thiazole synthase ThiG (InterPro:IPR033983)
cluster:220748	Orotidine 5'-phosphate decarboxylase domain (InterPro:IPR001754)
cluster:220750	Protein of unknown function DUF480 (InterPro:IPR007432)
cluster:220753	
cluster:220777	Prokaryotic N-terminal methylation site (InterPro:IPR012902); Type IV minor pilin ComP (InterPro:IPR031982)
cluster:220786	Short-chain dehydrogenase/reductase SDR (InterPro:IPR002347)

cluster:220789	ATP-grasp fold, ATP-dependent carboxylate-amine ligase-type (InterPro:IPR003135)
cluster:220793	Domain of unknown function DUF2272 (InterPro:IPR019262)
cluster:220808	Cytochrome b561, bacterial/Ni-hydrogenase (InterPro:IPR011577)
cluster:220811	
cluster:220822	Glycosyltransferase subfamily 4-like, N-terminal domain (InterPro:IPR028098); Glycosyl transferase, family 1 (InterPro:IPR001296)
cluster:220829	Calcineurin-like phosphoesterase domain, ApaH type (InterPro:IPR004843)
cluster:220836	
cluster:220838	
cluster:220850	Outer membrane efflux protein (InterPro:IPR003423)
cluster:220851	LysR, substrate-binding (InterPro:IPR005119)
cluster:220856	
cluster:220859	JmjC domain (InterPro:IPR003347)
cluster:220886	Vitamin B12-dependent methionine synthase, activation domain (InterPro:IPR004223); Cobalamin (vitamin B12)-binding domain (InterPro:IPR006158); Pterin-binding domain (InterPro:IPR000489); Cobalamin (vitamin B12)-binding module, cap domain (InterPro:IPR003759)
cluster:220890	
cluster:220899	Fungal lipase-like domain (InterPro:IPR002921)
cluster:220900	
cluster:220901	Glycosyl hydrolase family 53 (InterPro:IPR011683)
cluster:220924	
cluster:220925	
cluster:220929	Putative L,D-transpeptidase tautomerase (InterPro:IPR032676)
cluster:220931	
cluster:220932	Anticodon-binding (InterPro:IPR004154)
cluster:220935	RNA polymerase sigma-70 region 2 (InterPro:IPR007627); RNA polymerase sigma factor 70, region 4 type 2 (InterPro:IPR013249)
cluster:220936	
cluster:220943	
cluster:220946	
cluster:220952	Regulator of nucleoside diphosphate kinase, N-terminal domain (InterPro:IPR029462); Transcription elongation factor, GreA/GreB, C-terminal (InterPro:IPR001437)

cluster:220968	
cluster:220983	
cluster:220987	
cluster:220992	TonB-dependent receptor, plug domain (InterPro:IPR012910)
cluster:221003	Protein of unknown function DUF2066 (InterPro:IPR018642)
cluster:221016	Sulfatase, N-terminal (InterPro:IPR000917)
cluster:221017	
cluster:221020	Fido domain (InterPro:IPR003812)
cluster:221022	RNA polymerase sigma factor 54 interaction domain (InterPro:IPR002078); Signal transduction response regulator, receiver domain (InterPro:IPR001789); DNA binding HTH domain, Fis-type (InterPro:IPR002197)
cluster:221031	Domain of unknown function DUF4166 (InterPro:IPR025311)
cluster:221039	Protein of unknown function DUF4432 (InterPro:IPR027839)
cluster:221041	
cluster:221043	
cluster:221051	Glycosyltransferase 2-like (InterPro:IPR001173)
cluster:221053	Lipocalin/cytosolic fatty-acid binding domain (InterPro:IPR000566)
cluster:221059	
cluster:221063	Phenazine biosynthesis PhzF protein (InterPro:IPR003719)
cluster:221089	NodB homology domain (InterPro:IPR002509)
cluster:221096	Integral membrane protein TerC (InterPro:IPR005496)
cluster:221102	PucR C-terminal helix-turn-helix domain (InterPro:IPR025736)
cluster:221103	
cluster:221104	
cluster:221110	
cluster:221123	Transposase, IS4-like (InterPro:IPR002559); Transposase, IS4, N-terminal (InterPro:IPR024473)
cluster:221126	
cluster:221127	ZipA, C-terminal FtsZ-binding domain (InterPro:IPR007449)
cluster:221133	Aminoacyl-tRNA synthetase, class II (D/K/N) (InterPro:IPR004364); OB-fold nucleic acid binding domain, AA-tRNA synthetase-type (InterPro:IPR004365)
cluster:221137	
cluster:221138	

cluster:221151	THUMP domain (InterPro:IPR004114); S-adenosylmethionine-dependent methyltransferase (InterPro:IPR019614); Putative RNA methylase domain (InterPro:IPR000241)
cluster:221152	ABC transporter-like (InterPro:IPR003439)
cluster:221155	GNAT domain (InterPro:IPR000182)
cluster:221162	
cluster:221171	
cluster:221177	
cluster:221186	
cluster:221187	
cluster:221188	
cluster:221206	Protein of unknown function DUF839 (InterPro:IPR008557)
cluster:221208	TonB, C-terminal (InterPro:IPR037682)
cluster:221210	Glutamine synthetase, catalytic domain (InterPro:IPR008146)
cluster:221211	MetA family (InterPro:IPR033752)
cluster:221212	Amino acid exporter protein, LeuE-type (InterPro:IPR001123)
cluster:221214	Fatty acid desaturase domain (InterPro:IPR005804)
cluster:221215	Cys/Met metabolism, pyridoxal phosphate-dependent enzyme (InterPro:IPR000277)
cluster:221216	Aminotransferase class V domain (InterPro:IPR000192)
cluster:221218	Histone-like protein H-NS (InterPro:IPR001801)
cluster:221221	Lipocalin/cytosolic fatty-acid binding domain (InterPro:IPR000566)
cluster:221222	Lipocalin/cytosolic fatty-acid binding domain (InterPro:IPR000566)
cluster:221237	Homogentisate 1,2-dioxygenase (InterPro:IPR005708)
cluster:221246	Peptidase M4, C-terminal (InterPro:IPR001570)
cluster:221247	Peptidase M4, C-terminal (InterPro:IPR001570); Peptidase M4 domain (InterPro:IPR013856)
cluster:221248	
cluster:221265	Rhamnose/fucose mutarotase (InterPro:IPR008000)
cluster:221269	
cluster:221271	Domain of unknown function DUF4142 (InterPro:IPR025419)
cluster:221285	Starch synthase, catalytic domain (InterPro:IPR013534); Glycosyl transferase, family 1 (InterPro:IPR001296)
cluster:221289	Glycosyl hydrolase, family 13, catalytic domain (InterPro:IPR006047)
cluster:221293	Di-haem cytochrome c peroxidase (InterPro:IPR004852)
cluster:221298	

cluster:221303	Pyridine nucleotide-disulphide oxidoreductase, dimerisation domain (InterPro:IPR004099)
cluster:221304	Methylated-DNA-[protein]-cysteine S-methyltransferase, DNA binding (InterPro:IPR014048)
cluster:221305	Peptidase S54, rhomboid domain (InterPro:IPR022764)
cluster:221307	
cluster:221308	Transcription elongation factor, GreA/GreB, N-terminal (InterPro:IPR022691); Transcription elongation factor, GreA/GreB, C-terminal (InterPro:IPR001437)
cluster:221310	
cluster:221322	
cluster:221328	Endonuclease/exonuclease/phosphatase (InterPro:IPR005135)
cluster:221343	Amidohydrolase 3 (InterPro:IPR013108)
cluster:221344	
cluster:221346	Restriction endonuclease type IV, Mrr (InterPro:IPR007560)
cluster:221361	
cluster:221380	TonB-dependent receptor, plug domain (InterPro:IPR012910)
cluster:221381	
cluster:221400	
cluster:221412	
cluster:221413	MarR-type HTH domain (InterPro:IPR000835)
cluster:221417	
cluster:221418	
cluster:221419	Peptidase M3A/M3B catalytic domain (InterPro:IPR001567)
cluster:221420	Peptidase M3A/M3B catalytic domain (InterPro:IPR001567)
cluster:221428	
cluster:221429	
cluster:221440	Glutathione S-transferase, N-terminal (InterPro:IPR004045)
cluster:221441	
cluster:221444	
cluster:221451	
cluster:221453	ABC transporter type 1, transmembrane domain MetI-like (InterPro:IPR000515)
cluster:221457	RND efflux pump, membrane fusion protein, barrel-sandwich domain (InterPro:IPR032317); Membrane fusion protein, biotin-lipoyl like domain (InterPro:IPR039562)
cluster:221483	NDUFAF3/Mth938 domain-containing protein (InterPro:IPR007523)

cluster:221490	GTPase HflX, N-terminal (InterPro:IPR025121); GTP-binding protein, middle domain (InterPro:IPR032305); GTP binding domain (InterPro:IPR006073)
cluster:221498	
cluster:221499	
cluster:221501	
cluster:221502	
cluster:221505	Signal transduction response regulator, receiver domain (InterPro:IPR001789); RNA polymerase sigma factor 54 interaction domain (InterPro:IPR002078); DNA binding HTH domain, Fis-type (InterPro:IPR002197)
cluster:221512	Alcohol dehydrogenase, C-terminal (InterPro:IPR013149)
cluster:221513	
cluster:221516	Serine aminopeptidase, S33 (InterPro:IPR022742)
cluster:221525	
cluster:221526	
cluster:221535	EF-hand domain (InterPro:IPR002048)
cluster:221540	Patatin-like phospholipase domain (InterPro:IPR002641)
cluster:221543	
cluster:221544	Methyltransferase domain, putative (InterPro:IPR022744)
cluster:221545	Serine aminopeptidase, S33 (InterPro:IPR022742)
cluster:221546	
cluster:221547	
cluster:221548	Methylglyoxal synthase-like domain (InterPro:IPR011607); Bifunctional purine biosynthesis protein PurH-like (InterPro:IPR002695)
cluster:221555	Major facilitator superfamily (InterPro:IPR011701)
cluster:221558	
cluster:221565	
cluster:221566	Bactofilin A/B (InterPro:IPR007607)
cluster:221584	OmpA-like domain (InterPro:IPR006665); Motility protein B, N-terminal domain (InterPro:IPR025713)
cluster:221596	
cluster:221600	Extradiol ring-cleavage dioxygenase, class III enzyme, subunit B (InterPro:IPR004183)
cluster:221606	Domain of unknown function DUF2236 (InterPro:IPR018713)
cluster:221609	

1431

Supplementary Data 1.B InterPro annotation of the *X. oryzae* pv. *oryzicola* (*Xoc*) exclusive core genome.

Cluster ID	InterPro Annotation
cluster:217852	Signal transduction response regulator, receiver domain (InterPro:IPR001789); Histidine kinase/HSP90-like ATPase (InterPro:IPR003594); PAS domain (InterPro:IPR000014)
cluster:217942	Transcription regulator HTH, LysR (InterPro:IPR000847)
cluster:217973	Aminotransferase class V domain (InterPro:IPR000192)
cluster:217982	RND efflux pump, membrane fusion protein (InterPro:IPR006143); RND efflux pump, membrane fusion protein, barrel-sandwich domain (InterPro:IPR032317)
cluster:218068	Conserved hypothetical protein CHP02448 (InterPro:IPR012661)
cluster:218080	Transthyretin/hydroxyisourate hydrolase domain (InterPro:IPR023416); FAD-dependent urate hydroxylase HpyO, FAD/NAD(P)-binding domain (InterPro:IPR038732)
cluster:218151	Transcriptional regulator, AbiEi antitoxin (InterPro:IPR025159)
cluster:218163	Cell wall hydrolase, SleB (InterPro:IPR011105)
cluster:218180	FMN-dependent dehydrogenase (InterPro:IPR000262)
cluster:218224	
cluster:218393	PAAR motif (InterPro:IPR008727)
cluster:218394	VRR-NUC domain (InterPro:IPR014883)
cluster:218425	Major facilitator, sugar transporter-like (InterPro:IPR005828)
cluster:218430	
cluster:218441	SNARE associated Golgi protein (InterPro:IPR032816); Phosphatidic acid phosphatase type 2/haloperoxidase (InterPro:IPR000326); LssY-like C-terminal domain (InterPro:IPR025902)
cluster:218448	Putative esterase (InterPro:IPR000801)
cluster:218487	
cluster:218514	Fimbrial protein pilin (InterPro:IPR001082); GYF domain 2 (InterPro:IPR025640)
cluster:218553	
cluster:218656	Oxidoreductase FAD/NAD(P)-binding (InterPro:IPR001433); 2Fe-2S ferredoxin-type iron-sulfur binding domain (InterPro:IPR001041); Flavoprotein pyridine nucleotide cytochrome reductase-like, FAD-binding domain (InterPro:IPR008333)
cluster:218657	DNA-binding HTH domain, TetR-type (InterPro:IPR001647)
cluster:218661	
cluster:218662	RNA-binding S4 domain (InterPro:IPR002942)
cluster:218663	
cluster:218757	

cluster:218776	
cluster:218833	
cluster:218876	GAD-related (InterPro:IPR014983)
cluster:218917	KDPG/KHG aldolase (InterPro:IPR000887)
cluster:218988	Rieske [2Fe-2S] iron-sulphur domain (InterPro:IPR017941)
cluster:219029	Domain of unknown function DUF58 (InterPro:IPR002881)
cluster:219030	Protein of unknown function (DUF4381 (InterPro:IPR025489)
cluster:219144	
cluster:219145	ATPase, AAA-type, core (InterPro:IPR003959); AAA C-terminal domain (InterPro:IPR032423); MgsA AAA+ ATPase C-terminal (InterPro:IPR021886)
cluster:219146	Domain of unknown function DUF190 (InterPro:IPR003793)
cluster:219148	
cluster:219149	Chloride channel, voltage gated (InterPro:IPR001807)
cluster:219156	
cluster:219159	Pyrrolo-quinoline quinone repeat (InterPro:IPR002372)
cluster:219160	GTPase Der, C-terminal KH-domain-like (InterPro:IPR032859); GTP binding domain (InterPro:IPR006073)
cluster:219163	
cluster:219176	Queuosine precursor transporter (InterPro:IPR003744)
cluster:219242	
cluster:219302	Serine-tRNA synthetase, type1, N-terminal (InterPro:IPR015866); Aminoacyl-tRNA synthetase, class II (G/ P/ S/T) (InterPro:IPR002314)
cluster:219318	
cluster:219322	Amine oxidase (InterPro:IPR002937)
cluster:219332	
cluster:219392	SEC-C motif (InterPro:IPR004027)
cluster:219406	Amino acid/polyamine transporter I (InterPro:IPR002293)
cluster:219428	Glutamate/phenylalanine/leucine/valine dehydrogenase, C-terminal (InterPro:IPR006096); Glutamate/phenylalanine/leucine/valine dehydrogenase, dimerisation domain (InterPro:IPR006097)
cluster:219429	Protein of unknown function DUF45 (InterPro:IPR002725)
cluster:219443	
cluster:219478	Glycoside hydrolase family 20, catalytic domain (InterPro:IPR015883); Beta-hexosaminidase, bacterial type, N-terminal (InterPro:IPR015882)

cluster:219490	NfeD-like, C-terminal domain (InterPro:IPR002810)
cluster:219575	Outer membrane efflux protein (InterPro:IPR003423)
cluster:219578	
cluster:219580	
cluster:219592	SGNH hydrolase-type esterase domain (InterPro:IPR013830)
cluster:219619	
cluster:219787	
cluster:219859	Major facilitator superfamily (InterPro:IPR011701)
cluster:219877	
cluster:219930	Glyoxalase/fosfomycin resistance/dioxygenase domain (InterPro:IPR004360)
cluster:219952	ABC-2 type transporter (InterPro:IPR013525)
cluster:220005	Glycosyl transferase, family 19 (InterPro:IPR003835)
cluster:220035	
cluster:220036	Inosine/uridine-preferring nucleoside hydrolase domain (InterPro:IPR001910)
cluster:220075	
cluster:220093	
cluster:220158	
cluster:220172	Cation/H+ exchanger (InterPro:IPR006153)
cluster:220207	Ketopantoate reductase, C-terminal domain (InterPro:IPR013752); Ketopantoate reductase, N-terminal domain (InterPro:IPR013332)
cluster:220230	
cluster:220231	NAD(P)-binding domain (InterPro:IPR016040)
cluster:220295	Methyltransferase type 11 (InterPro:IPR013216)
cluster:220396	Killing trait, RebB (InterPro:IPR021070)
cluster:220475	YcgL domain (InterPro:IPR027354)
cluster:220511	Outer membrane efflux protein (InterPro:IPR003423)
cluster:220516	ABC transporter type 1, transmembrane domain (InterPro:IPR011527); ABC transporter-like (InterPro:IPR003439)
cluster:220546	Signal transduction response regulator, receiver domain (InterPro:IPR001789); GAF domain (InterPro:IPR003018); PAS fold-3 (InterPro:IPR013655); Signal transduction histidine kinase, dimerisation/phosphoacceptor domain (InterPro:IPR003661); Histidine kinase/HSP90-like ATPase (InterPro:IPR003594); PAS fold-4 (InterPro:IPR013656)
cluster:220599	NADH:ubiquinone/plastoquinone oxidoreductase, chain 3 (InterPro:IPR000440)

cluster:220648	
cluster:220667	
cluster:220713	
cluster:220731	
cluster:220742	
cluster:220778	PilC beta-propeller domain (InterPro:IPR008707)
cluster:220844	Trigger factor, ribosome-binding, bacterial (InterPro:IPR008881); Trigger factor, C-terminal (InterPro:IPR008880)
cluster:220845	Clp protease proteolytic subunit /Translocation-enhancing protein TepA (InterPro:IPR023562)
cluster:220846	Clp ATPase, C-terminal (InterPro:IPR019489); Zinc finger, ClpX C4-type (InterPro:IPR010603); ATPase, AAA-type, core (InterPro:IPR003959)
cluster:220847	ATPase, AAA-type, core (InterPro:IPR003959); Peptidase S16, Lon proteolytic domain (InterPro:IPR008269); Lon, substrate-binding domain (InterPro:IPR003111)
cluster:220864	CAP domain (InterPro:IPR014044)
cluster:220933	NADP transhydrogenase beta-like domain (InterPro:IPR034300)
cluster:221010	
cluster:221019	Type II/IV secretion system protein (InterPro:IPR001482); General secretory system II, protein E, N-terminal (InterPro:IPR007831)
cluster:221040	
cluster:221071	
cluster:221073	Uncharacterized membrane protein YitT (InterPro:IPR003740)
cluster:221107	
cluster:221238	
cluster:221338	
cluster:221352	Aminotransferase class V domain (InterPro:IPR000192)
cluster:221354	Enolpyruvate transferase domain (InterPro:IPR001986)
cluster:221369	
cluster:221426	
cluster:221613	FAD dependent oxidoreductase (InterPro:IPR006076)

1433

Artigo 2

Formatação: *Molecular Plant-Microbe Interactions* (Fator de impacto: 3,588).
Status do manuscrito: em revisão.

Title: TargeTALE: A web resource to identify TALEs in *Xanthomonas* genomes and their respective targets.

Running Title: TargeTALE

Frederico Schmitt Kremer [1], Amanda Munari Guimarães [1], Christian Domingues Sanchez [1], Luciano da Silva Pinto [1#]

[1] Laboratório de Bioinformática e Proteômica, Programa de Pós-Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Capão do Leão, Rio Grande do Sul, Brazil

#address correspondence to: ls_pinto@hotmail.com

Keywords: phytopathogens, bioinformatics, genome annotation, effector proteins, T3SS

Abstract

The *Xanthomonas* genus, comprises more than 30 species of Gram-negative bacteria, most of which are pathogens of plants with high economic value, such as rice, common bean and maize. Transcription activator-like effectors (TALEs), which act by regulating the hosts gene expression, are some of the major virulence factors of these bacteria. The *in silico* analysis of these genes and their targets provide useful information regarding the pathogenesis these bacteria, and although there are some tools that already allow the prediction of its prediction and identification of potential targets, their usability is very limited, non-intuitive for non-bioinformatician researchers. Additionally, to our Knownledge, there is no straightforward way to identify TALEs genes and targets directly from a *Xanthomonas* genome and a dataset of promoters. Therefore, here we present TargeTALE, a novel tool for both TALE and target prediction, which aims to provide a user-friendly way to analyze these genes and characterize newly sequenced genomes. The analysis of the results obtained by TargeTALE in a proof-of-concept validation demonstrate that, at optimum setting, ~93% of the predicted target genes with available expression data were confirmed as up-regulated during the infection, indicating that the tool might be useful for researchers in the field. TargeTALE can be accessed from the URL http://bioprolab.ufpel.edu.br/targeTALE/.

Introduction

The genus *Xanthomonas* covers over 30 species of Gram-negative bacteria, most of which are phytopathogenic and infect plants of high economic importance (Vauterin et al. 2000). *X. oryzae*, for example, is a major threat to rice (*Oryza sativa*) production, and its pathovars *oryzae* (*Xoo*) and *oryzicola* (*Xoc*), which are etiological agents of bacterial blight (BB) and bacterial leaf streak (BLS), respectively, have re-emerged as worldwide concerns during the past decades. The pathovar *Xoo* has already been reported to be capable of reducing the production yield by up to 80% in endemic regions, while *Xoc,* although less impacting, has also been reported to cause losses of up to 30% during epidemics in certain regions (Niño-Liu et al. 2006). In a similar way, other species of the genus, such as *X. citri*, which infects many species of the *Citrus* genus, *X. citri* subsp. *fuscans*, which infects many species of beans, and *X. axonopodis* pv. *glycines* (*Xag*), which infects soybean (*Glycine max*), also have a drastic impact in the productivity depending on the geographic region, and lead to many economic losses as well.

Transcription activation-like effectors (TALEs) are proteins produced by several species of *Xanthomonas* and are secreted via the type III secretion system (T3SS). These proteins act as modulators of gene expression in the host cells by specifically recognizing promoter regions in the host genome. The promoter sequence is recognized by variable domains comprising repeats of ~34 amino acids with two hypervariable residues, which are called repeat variable di-residues (RVDs) (Muñoz Bodnar et al. 2013). A previously described a model of the relationship between the RVD sequence and DNA specificity (Boch et al. 2009; Moscou and Bogdanove 2009) has made it possible to predict the targeted regions of a certain TALE in a DNA sequence *in silico*, based on the analysis of its RVDs. In fact, some tools such as TargetFinder (Doyle et al. 2012), StoryTeller (Pérez-Quintero et al. 2013), TALgetter (Grau et al. 2013), TALvez (Pérez-Quintero et al. 2013), and AnnoTALE (Grau et al. 2016) have already been developed for such analysis.

These tools usually allow the identification, prediction of targets based on a provided set of sequences, and/or classification of TALEs, but researchers with non-bioinformatics background might find it difficult to use them on genome-wide analysis of TALEs and their respective targets in newly sequenced genomes. In addition, although TALvez provides a web-based interface and databases from various plant hosts, it requires the input of previously annotated RVD sequences and does not accept unannotated genomes in FASTA format, which makes it largely inapplicable for processing newly sequenced genomes. Furthermore, when pre-loaded databases are used, the TALvez output is also mainly limited to previously described promoters, which makes the identification of novel target genes difficult. In order to provide a simple tool for *Xanthomonas* researchers to analyze TALE genes, we have developed a new pipeline named TargeTALE, which integrates TALE predictors, a wide variety of promoter datasets, TALE target predictors, and functional annotations of the identified targets based on the Gene Ontology (GO) terms. Therefore, our approach is intended allow a straightforward way to analyze genes and biological processed target by TALEs directly identified from the genome of *Xanthomonas* strains.

Methodology

The pipeline

TargeTALE was developed using the programming language Python, and the third-party packages BioPython (http://biopython.org/), Bioservices (https://github.com/cokelaer/bioservices) and Requests (https://github.com/requests/requests), and was deployed as a webserver running on top of Apache HTTP server (https://httpd.apache.org). TALE genes are predicted using AnnoTALE (Grau et al. 2016), which is also used to assign a standardized class name to TALE genes and identify the RVD region. Promoter databases were retrieved for a wide variety of model and non-model species from different sources, including Ensembl (https://www.ensembl.org/), JGI (https://jgi.doe.gov/), TAIR (https://www.arabidopsis.org/) and Osiris (Morris et al. 2008).

The targets of each TALE are identified using TALgetter (Grau et al. 2013), and only the hits with p-value < 0.05 and e-value < 1e–10 are selected. Finally, the Uniprot ID of each target is accessed from the Uniprot database (http://www.uniprot.org), using its client implemented in the Python Bioservices package. The Uniprot ID is used to retrieve its Gene Ontology (GO) annotation from QuickGO (Binns et al. 2009). An overview of the pipeline is displayed in Figure 1.

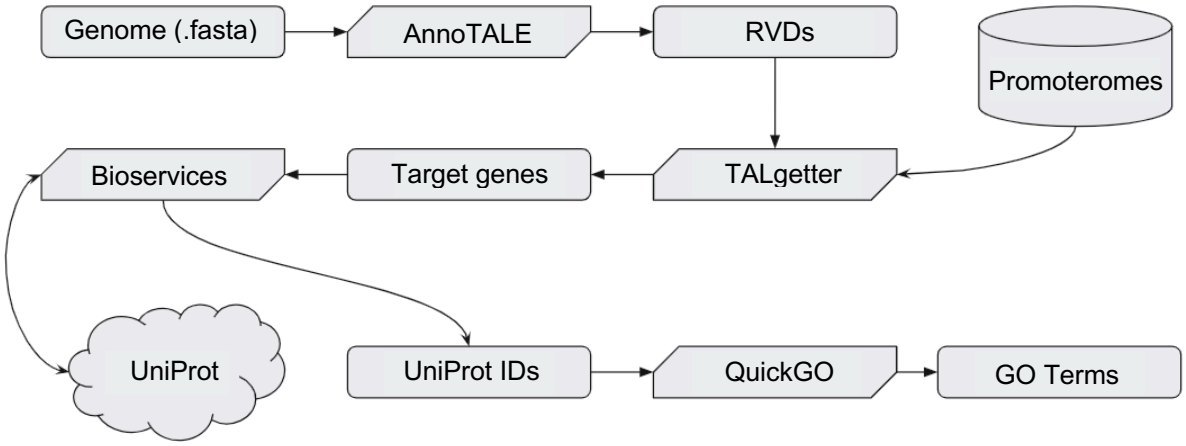


Figure 1. Overview of the TargeTALE main pipeline.

Validation: X. oryzae datasets

As a proof-of-concept, to evaluate the reliability of the pipeline, we analyzed six genomes of *X. oryzae*: two of *Xoo* and four of *Xoc* (GenBank: CP000967.2, CP003057.2, CP007166.1, CP007221.1, CP011955.1, and CP011958.1), using the different promoterome datasets of *O. sativa* derived from Osiris (250, 500, 750, 1000, 1500, 2000 and 3000bp upstream, considering gene models or transcript alignment). In the context of Osiris, gene models-derived promoters were extracted from upstream regions of the gens predicted during the annotation, while those derived from transcript alignment are based on the alignment of mRNA sequence and usually are close to the transcription start sequence (TSS).

Then, the identified target genes in each combination of strain-dataset were matched with strain-specific gene expression data, available at EBI Expression Atlas

(https://www.ebi.ac.uk/gxa/) (EBI-GXA: E-GEOD–36272, E-GEOD–67588, E-GEOD–33411, E-GEOD–67588, E-GEOD–67588, E-GEOD–67588, and E-GEOD–67588). In first analysis, we considered as positive hits those genes whose expression levels were increased (up-regulated) during the infection by the specific strains, and negative those expression was not. As not all genes indexed in the promoterome are covered by the gene sets used in the expression studies we also calculated the percentage of correct hits considering only those genes with data was available in the expression datasets. Genes were considered up-regulated during the infection if the Log_2-fold change is ≥ 2.

Results

TargeTALE

The pipeline was implemented as a webserver and doesn't require any registration to be used. During the submission the user must only upload a FASTA file containing the genome to be analyzed, select one of the available promoterome datasets and inform his email address, and when the job is finished the user will receive an email with the link to access the analysis report. The results might also be downloaded as a ".tar.gz" file, which included tables, in tab-separeted value format (TSV) and, with a summary of the TALEs and targets, GO annotations of the targets and the intermediary results of each program. The results of the evaluation of the tool using gene expression data of *O. sativa* after the infection by different strains of *X. oryzae* are presented in Supplementary Table S1 and S2, where Supplementary Table S1 present the results considering the percentage of targets identified as up-regulated excluding those targets not indexed in the respective study, and Supplementary Table S2 present the results considering the percentage of targets identified as up-regulated including those targets not indexed in the respective study.

Discussion

Currently, TargeTALE provides promoter datasets (promoteromes) for 13 different plant species, covering common hosts of *Xanthomonas,* such as rice (*Oryza sativa*), common bean (*Phaseolus vulgaris*), sweet orange (*Citrus sinensis*), tomato (*Solanum lycopersicum*), soybean (*Glycine max*) and maize (*Zea mays*), and model organisms to study host-pathogen interactions and pathogenesis, such as *Arabidopsis. thaliana* (Buell 2002) and *Brachypodium distachyon* (Fitzgerald et al. 2015). For each host we also provide datasets of promoter sequences derived from different up-stream distances from its genes. In the case of datasets generated using an in house scripts, we provide upstream sequences of 250, 500, 750 and 1000 base-pairs (bp). The datasets of promoters of rice derived from Osiris, are available for upstream distances of 250, 500. 750, 1000, 1500, 2000 and 3000, and includes predictions derived from gene models and transcript data. Finally, datasets for *A. thaliana* were obtained from TAIR (Rhee et al. 2003) and are available for up-stream distances of 500, 1000 and 3000 base-pairs.

In the validation of our pipeline using *X. oryzae* data we were able to obtain a mean confirmation rate of ~95% using the 250 bp upstream derived from genes identified based on transcripts datasets retrieved from EBI Gene Expression Atlas. Using this approach, a slightly difference was observed from strain to strain and when considering different upstream distances. The higher accuracy with 250 bp datasets was also observed when considering the ratio between the genes confirmed to be up-regulated during the infection and full set of predicted target genes were considered, although this value dropped to 35.9%. However, it is unclear whether this was a consequence of discrepancy in the gene expression datasets, although previous studies have already demonstrated that the effective activation of a target gene expression may require additional factors, such as the presence of multiple TALEs binding to the same gene, even in the opposite DNA strand (Streubel et al. 2017). Additionally, it was already reported that TALEs usually bind in positions located 300bp around the transcription start site (TSS) (Grau et al. 2013), which might explain why the 250bp upstream datasets resulted in more accurate results, and why better results were usually obtained when

using data derived from transcripts and not from predicted gene models. Thus, the analysis also indicated that datasets of longer up-stream regions do not provide better results, as they might increase the occurrence of false-positive.

We strongly discourage the submission of draft genomes sequenced using short-read platforms, such as Illumina MiSeq and Ion Torrent PGM, as the repetitive structure of TALE genes may lead to miss-assemblies in this case, so long-read sequenced genomes (Peng et al. 2016), obtained from platforms such as PacBio, Oxford Nanopore or even reference genomes obtained from Sanger-based platforms, must be preferred when possible.

Conclusion

We have herewith presented the TargeTALE, a webserver for the identification of TALE genes in *Xanthomonas* spp. genomes and their respective potential targets many plant promoteromes. The tool might be useful particularly for researchers without bioinformatics background that want to characterize genes targeted by specific strains, specially newly-sequenced genomes. The information provided by TargeTALE might also be applied in comparative studies aiming a better understanding of the adaptation of the pathogen based on the differential selection of target genes in the host, and facilitate future studies in pathogenomics. Finally, with the incorporation of new datasets of promoters from other hosts, this platforms might also be extended to other species of *Xanthomonas*. The webserver is available through the URL http://bioprolab.ufpel.edu.br/targeTALE/ and its source code is available at the GitHub repository https://github.com/biopro/targetale.

References

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. 2009. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics. 25:3045–6

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. 2009. Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. Science (80-.). 326:1509–1512

Doyle, E. L., Booher, N. J., Standage, D. S., Voytas, D. F., Brendel, V. P., VanDyk, J. K., and Bogdanove, A. J. 2012. TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. Nucleic Acids Res. 40:W117–W122

Grau, J., Reschke, M., Erkes, A., Streubel, J., Morgan, R. D., Wilson, G. G., Koebnik, R., and Boch, J. 2016. AnnoTALE: bioinformatics tools for identification, annotation, and nomenclature of TALEs from Xanthomonas genomic sequences. Sci. Rep. 6:21077

Grau, J., Wolf, A., Reschke, M., Bonas, U., Posch, S., and Boch, J. 2013. Computational Predictions Provide Insights into the Biology of TAL Effector Target Sites A. Tresch, ed. PLoS Comput. Biol. 9:e1002962

Morris, R. T., O'Connor, T. R., and Wyrick, J. J. 2008. Osiris: an integrated promoter database for Oryza sativa L. Bioinformatics. 24:2915–2917

Moscou, M. J., and Bogdanove, A. J. 2009. A Simple Cipher Governs DNA Recognition by TAL Effectors. Science (80-.). 326:1501–1501

Muñoz Bodnar, A., Bernal, A., Szurek, B., and López, C. E. 2013. Tell Me a Tale of TALEs. Mol. Biotechnol. 53:228–235

Niño-Liu, D. O., Ronald, P. C., and Bogdanove, A. J. 2006. Xanthomonas oryzae pathovars: model pathogens of a model crop. Mol. Plant Pathol. 7:303–24

Peng, Z., Hu, Y., Xie, J., Potnis, N., Akhunova, A., Jones, J., Liu, Z., White, F. F., and Liu, S. 2016. Long read and single molecule DNA sequencing simplifies genome assembly and TAL effector gene analysis of Xanthomonas translucens. BMC

Genomics. 17:21

Pérez-Quintero, A. L., Rodriguez-R, L. M., Dereeper, A., López, C., Koebnik, R., Szurek, B., and Cunnac, S. 2013. An improved method for TAL effectors DNA-binding sites prediction reveals functional convergence in TAL repertoires of Xanthomonas oryzae strains. PLoS One. 8:e68464

Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. 2003. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res. 31:224–8

Streubel, J., Baum, H., Grau, J., Stuttman, J., and Boch, J. 2017. Dissection of TALE-dependent gene activation reveals that they induce transcription cooperatively and in both orientations. PLoS One. 12:e0173580

Vauterin, L., Rademaker, J., and Swings, J. 2000. Synopsis on the taxonomy of the genus xanthomonas. Phytopathology. 90:677–82

Supplementary Table S1. Ratio of genes predicted as targets that were confirmed as up-regulated during the infection in the different combinations of strain and promoterome datasets, considering only those targets that have gene expression data available for the specific strain. Each dataset of promoterome consist in promoters derived from prediction based on gene models (P) or transcript data (T), considering different upstream distances.

Strain	Promoterome dataset													
	250 bp		500 bp		750 bp		1000 bp		1500 bp		2000 bp		3000 bp	
	P	T	P	T	P	T	P	T	P	T	P	T	P	T
PXO99A	0.90	1.00	0.81	0.89	0.80	0.80	0.72	0.78	0.67	0.78	0.75	0.64	0.78	0.71
BLS256	0.90	0.92	0.75	0.92	0.83	0.70	0.83	0.78	0.80	0.85	0.82	0.67	0.60	0.67
PXO86	1.00	1.00	0.86	0.92	0.92	0.85	1.00	0.86	0.71	0.86	1.00	0.89	1.00	1.00
CFBP7342	1.00	0.88	0.78	0.90	0.86	0.79	0.73	0.92	0.83	0.79	0.86	0.78	0.71	0.82
B8-12	0.82	0.92	0.75	0.78	0.48	0.86	0.78	0.71	0.80	0.86	0.80	0.64	0.56	0.62
CFBP7331	0.88	1.00	0.86	0.89	0.92	1.00	0.70	0.90	0.88	0.93	1.00	0.71	1.00	1.00
Average / subset	0.92	0.95	0.80	0.88	0.80	0.83	0.79	0.82	0.78	0.84	0.87	0.72	0.77	0.80
Average / subset (P + T)	0.93		0.84		0.82		0.81		0.81		0.80		0.79	

Supplementary Table S2. Ratio of genes predicted as targets that were confirmed as up-regulated during the infection in the different combinations of strain and promoterome datasets. Each dataset of promoterome consist in promoters derived from prediction based on gene models (P) or transcript data (T), considering different upstream distances.

Strain	Promoterome dataset													
	250 bp		500 bp		750 bp		1000 bp		1500 bp		2000 bp		3000 bp	
	P	T	P	T	P	T	P	T	P	T	P	T	P	T
PXO99A	0.47	0.31	0.24	0.32	0.24	0.19	0.27	0.21	0.15	0.27	0.24	0.21	0.18	0.25
BLS256	0.45	0.33	0.23	0.25	0.27	0.23	0.19	0.19	0.13	0.17	0.10	0.17	0.11	0.15
PXO86	0.45	0.42	0.21	0.24	0.24	0.30	0.36	0.19	0.17	0.16	0.20	0.24	0.17	0.07
CFBP7342	0.30	0.32	0.14	0.20	0.24	0.23	0.24	0.24	0.20	0.20	0.16	0.24	0.14	0.21
B8-12	0.20	0.42	0.25	0.24	0.12	0.32	0.24	0.17	0.17	0.29	0.18	0.16	0.11	0.11
CFBP7331	0.26	0.36	0.32	0.29	0.33	0.26	0.24	0.36	0.23	0.23	0.30	0.21	0.28	0.17
Average / subset	0.36	0.36	0.23	0.25	0.24	0.25	0.26	0.23	0.17	0.22	0.20	0.21	0.16	0.16
Average / subset (P + T)	0.36		0.24		0.25		0.24		0.20		0.20		0.16	

Artigo 3

Formatação: *Brazilian Journal of Microbiology* (Fator de impacto: 1,810).

Status do manuscrito: aceito.

Title: Genome Sequence of *Xanthomonas fuscans* subsp. *fuscans* strain *Xff49*: A New Isolate Obtained from Common Beans in Southern Brazil

Running Title: Genome of a Brazilian strain of *X. fuscans*

Frederico Schmitt Kremer[1], Ismail Teodoro de Souza Junior[2], Amanda Munari Guimarães[1], Rafael dos Santos Danelon Woloski[1] , Andrea Bittencourt Moura[2], Luciano da Silva Pinto[1#]

[1] Laboratório de Bioinformática e Proteômica, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Capão do Leão, Brazil

[2] Laboratório de Bacteriologia Vegetal, Faculdade de Agronomia Eliseu Maciel, Universidade Federal de Pelotas, Capão do Leão, Brazil

Address correspondence to: ls_pinto@hotmail.com

Abstract

The genus, *Xanthomonas* comprise Gram-negative bacteria, many of which are phytopathogens. *Xanthomonas fuscans* subsp. *fuscans* is one of the most devastating pathogens affecting the bean plant, resulting in the common bacterial blight of bean (CBB). The disease is mainly foliar and affects a wide variety of bean species, thus acting as the yield-limiting factor for the bean crop. Here, we report the whole-genome sequencing of a new strain of *X. fuscans* subsp. *fuscans,* named Xff49, isolated from the infected and symptomatic beans from Capão do Leão, Southern Brazil. The genetic analysis demonstrated the presence of single-nucleotide variants (SNVs) in this strain, potentially affecting the mobilome, cell mobility, and inorganic ion metabolism. In addition, the analysis resulted in the identification of a new plasmid similar to the pAX22 derived from *Achromobacter denitrificans*, which was named plX, along with plA and plC, previously reported in other strains of *X. fuscans* subsp. *fuscans. Xff49* represents the first Brazilian genome of *X. fuscans* subsp. *fuscans* and might provide useful information applicable to the studies of phylogenetics, evolution, and pathogenomics, thereby allowing a better understanding of the genomic features present in the Brazilian strains.

Full Text

Introduction

The genus *Xanthomonas* comprises at least 30 species of Gram-negative bacteria, most of which are phytopathogens of plant species with high economic value, such as rice (*X. oryzae*), beans (*X. fuscans*), citrus (*X. citri*), soybean (*X. axonopodis* pv. *glycines*), and many others[1]. *Xanthomonas* species are usually sub-classified into pathovars and subspecies according to the infected host and tissue, the resulting disease, and/or genomic features. The term pathovar refers to a single or a group of bacterial strains with similar features but differ in host plant specificity from other strains of the same species.

Xanthomonas fuscans subsp. *fuscans (Xff)*, along with *X. axonopodis* pv. *phaseoli* (*Xap*), is one of the most devastating pathogens of beans leading to severe losses in crop yield and productivity. It is the causative agent of the common blight of bean (CBB), which affects not only common bean cultures (*Phaseolus vulgaris* L.), but also lima bean (*P. lunatus* L.) along with many other species [2]. The worldwide distribution of CBB is attributed to the efficient mechanism of transmission among seeds employed by bacteria such that studies report a reduction in production of up to 40% [3] and causing both direct and indirect costs to the farmers. The direct or primary loss may consist of loss of crop yield, quality, and extra costs involved in the post-harvesting processes, whereas indirect or secondary effects may include loss to traders, exporters, consumers, farmers, and government due to low crop productivity. In America and Africa, beans represent an important component of the diet and comprise almost 60% of the daily protein intake [4]. Since beans constitute one of the major sources of protein in these continents, an efficient control of causative agents of CBB is a constant issue. Several genetic factors and quantitative trait loci (QTLs) have been reported to be responsible for providing resistance to CBB [5–7], especially in bean lineages grown in Mesoamerica and Andes. Host resistance to CBB is considered to be a practical approach toward controlling

crop damage. However, the application and successful insertion of these genetic elements into commercial cultivars have been very limited.

The pathogenesis of both *Xff* and *Xap* strains involves the formation of biofilms by bacteria on the plant surface. This surface growth facilitates penetration into the phytosystem through the stomata, leading to the colonization of the mesophyll, vascular tissues, and parenchyma [8] and helping in the dispersal of infection to other plants. The disease is characterized by the formation of spots and necrosis in the leaves, stems, pods, and seeds; however, many infected plants may be asymptomatic [9]. Although the genomic analysis of *X. fuscans* subsp. *fuscans* strain 4834-R, a highly aggressive strain isolated in France [10,11], has identified several genes shared among the genus, the virulence factors contributing to CBB are still poorly understood. Some of these genes include those that encode for effector secretory proteins that function through type III and type IV secretion systems (T3SS and T4SS), adhesins, degradation enzymes, lipopolysaccharides (LSP) biosynthesis, and formation of biofilms (e.g., xanthan gum formation) [11]. Bacterial appendages, such as flagella, play a significant role in biofilm formation by pathogenic bacteria. However, as 4834-R strain does not contain a functional flagellum, usually found in flagellated strains from the species, other strains have been sequenced to provide a better description of the pathogen [12].

In the present study, we report the whole-genome sequencing of a new strain of *X. fuscans* subsp. *fuscans* named *Xff49*, isolated from infected and symptomatic beans from Capão do Leão, Southern Brazil. To our knowledge, Xff49 is the first Brazilian strain of *Xff* to be sequenced and made publicly available.

Material and Methods

Whole-genome shotgun sequencing (WGS) was performed at the Neoprospecta facility using the Illumina MiSeq platform and a 250-bp paired-end library. Raw reads obtained were analyzed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and pre-processed using Fastx-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Trimmomatic [13] was utilized for quality filtering and read trimming.

De novo genome assemblies were generated using A5 [14], Velvet [15], CLC Genome Workbench (https://www.qiagenbioinformatics.com/), SGA [16], and Ray [17]. The assemblies were merged using CISA [18] to generate a consensus assembly. Contigs and scaffolding generated by CISA were aligned using BLASTn [19] from the NCBI-BLAST+ [20] package to a database of *X. fuscans* genome obtained from GenBank (Supplementary data 1) to separate chromosome contigs from plasmid contigs ("profiling").

Unmapped contigs were compared to the National Center for Biotechnology Information (NCBI) nucleotide database. Multiple reference-guided scaffolding was performed for the contigs mapped to the chromosome of the bacteria using MeDuSa [21] and the same dataset of *X. fuscans* genomes used for the contig profiling. The scaffolds produced by this method were re-scaffolded using CAR [22] based on the genome of *X. fuscans* subsp. *fuscans* strain 4834-R (GenBank: NC_022541.1). This genome was identified by BLAST as the closest available complete genome. Contigs and scaffolds mapped to plasmids were ordered using CAR based on the sequence of the *X. fuscans* plasmids, namely plA (GenBank: FO681495.1) and plC (GenBank: FO681497.1), according to their respective best hits. Then, gap closing was performed using IMAGE [23]. As CAR and IMAGE does not estimate the gap distance during the scaffolding or gap-closing process, all contigs were joined using an arbitrary gap span of 300 base-pairs (bps). The final assembly was evaluated using QUAST [24].

Genome annotation was performed using Genix [25] and InterProScan, and manually curated using Artemis [26]. The protein-coding genes present in the chromosome of Xff49 and 4834-R were compared using OrthoVeen [27] to identify those genes present in Xff49 but absent in 4834-R, and effector proteins secreted by T3SS (T3Es) and T4SS (T4Es) were predicted using the EfectiveDB webserver [28]. The genomes of Xff49 and 4834-R were also analyzed using AntiSMASH [29] to identify gene clusters and metabolic pathways associated with secondary metabolites biosynthesis.

A dataset of 20 genomes of *X. fuscans* and 3 from outgroups (Supplementary data 2), including both draft and finished records, was obtained from GenBank and aligned to the genome of Xff49 using Mugsy [30]. The process led to the formation of a MAFF file containing the alignments of different syntenic regions identified in the input genomes. The sub-alignments were processed using trimAl [31] and then merged into a single alignment with functions implemented in the BioPython package [32]. The phylogenetic tree was generated using the implementation of the unweighted pair-group method with arithmetic mean (UPGMA) algorithm available in the MEGA X package using 10,000 cycles of bootstrapping [33].

A variant calling analysis was performed based on the genome of *X. fuscans* subsp. *fuscans* strain 4834-R to determine the nucleotide differences at a given locus. Paired-end reads were mapped using SMALT (https://www.sanger.ac.uk/science/tools/smalt–0), and the alignments were processed using SAMTools, BCFTools, and VCFUtils.pl [34]. The variants identified by VCFUtils.pl were filtered using the PyVCF library (https://github.com/jamescasbon/PyVCF) to select only those single-nucleotide polymorphisms (SNPs) with Q ≥ 30. The functional impact of each variant was predicted using SnpEff [35] based on the reference annotation, and affected genes were annotated using RPS-BLAST+, from the NCBI-BLAST+ package, using the conserved domains database (CDD) [36] to assign molecular functions based on the Clusters of Orthologous Group (COG) database nomenclature [37].

Finally, to identify the presence of genes from the flagellar genes cluster, we have followed the same procedure adopted originally in the characterization of strain 4834-R, which was based on flagellar genes identified in the genome of *X. campestris* pv. *vesicatoria* (**GenBank:** AM039952.1) [11]. In the analysis, we compared the strains Xff49, 4834-R and *X. fuscans* pv. *aurantifolli* str. ICPB 10535 (*Xac*), which was already reported with a functional polar flagellum and motility when cultured in semi-solid media [38].

Results and Discussion

The results of the genome assembly analysis are displayed in Table 1. After *de novo* assembly, profiling, and scaffolding, four replicons were reconstructed, including the chromosome and three plasmids. The chromosome of *X. fuscans* subsp. *fuscans* str. Xff49 has a length of approximately 5.2 megabases (Mb), and its scaffold contains 69 gaps, with a few genome translocations and inversions when compared to the reference sequence of *X. fuscans* subsp. *fuscans* strain 4834-R (Figure 1). We also identified plasmids similar to pLA and pLC from *4834-R* (GenBank: FO681495.1 and FO681497.1, respectively), although no contig with genes homologous to plB was identified (GenBank: FO681496.1). However, another plasmid, similar to the pAX22 from *Achromobacter xylosoxidans* subsp. *denitrificans* (GenBank: NC_022242.1), was identified and named "plX".

Table 1. Overview of the metrics calculated by QUAST for the genome sequence of *X. fuscans* subsp. *fuscans* strain Xff49

Replicon	Gaps	CG%	Size (bp)	# of contigs	N50 (contigs)	L50 (contigs)
Chromosome	69	64.68%	5182257	70	112211	15
Plasmid plA	3	59.49%	41822	4	26697	1
Plasmid plC	4	57.52%	34064	5	5176	2
Plasmid plX	1	63.04%	18734	2	15216	1

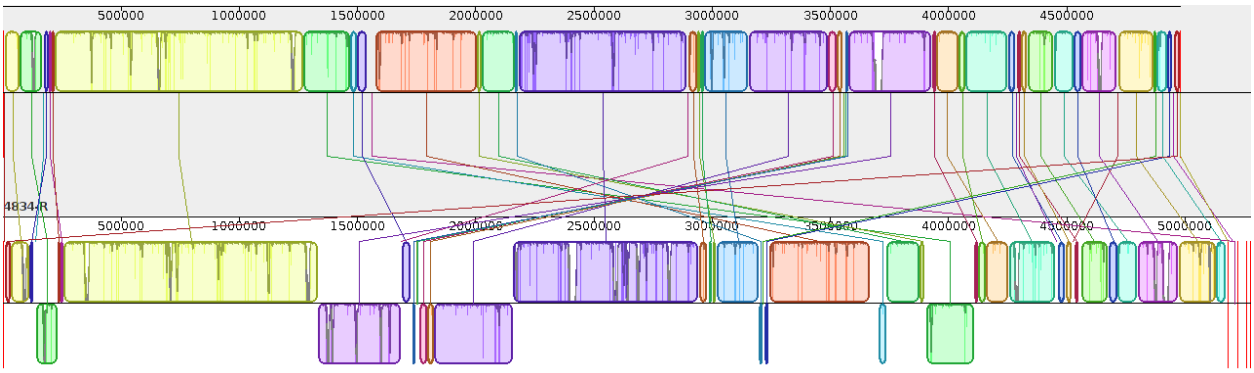
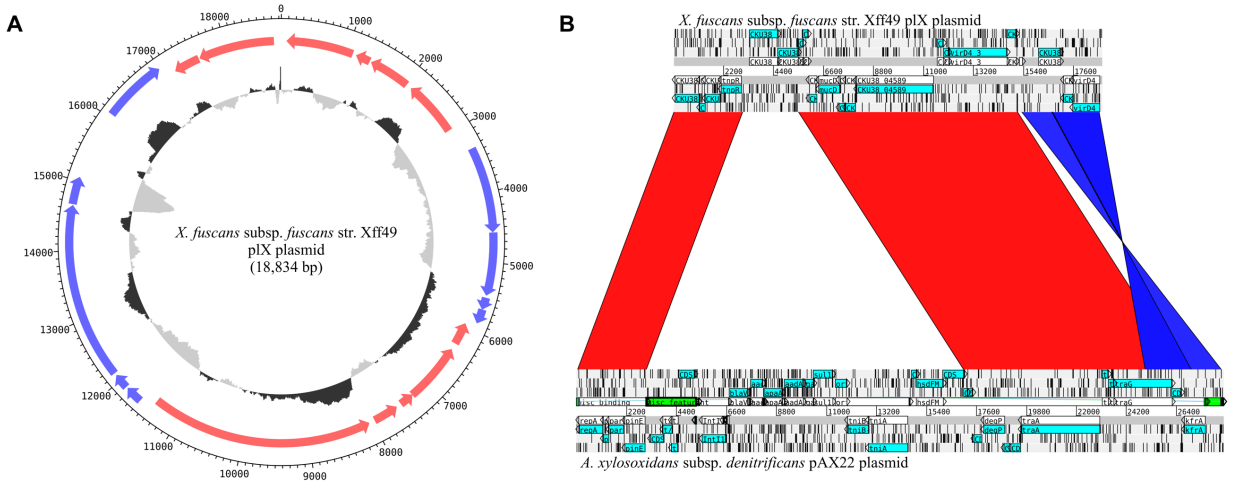


Figure 1. Structural comparison of the scaffolded chromosome of *X. fuscans* subsp. *fuscans* str. Xff49 and the reference genome of *X. fuscans* subsp. *fuscans* str. 4834-R.

A circular map of plX generated using DNAplotter and a synteny comparison between plX and pAX22 generated using Artemis Comparison Tool (ACT) is presented in Figure 2 (Figure 2.A and Figure 2.B, respectively). When comparing to paX22, we have identified a set of 17 genes that are missing in plX, which include transposases, proteins for the toxin / antitoxin system, metallo-beta-lactamase and proteins associated with aminoglycoside metabolism. All these genes are located in a single non-syntenic region in the middle of the paX22, were in plX there are only 2 protein-coding genes, one that encodes a cytosine-specific methyltransferase and other that encodes a restriction endonuclease NotI (CKU38_04582). The complete list of genes differing between paX22 and plX in this locus is presented in Supplementary Data 3.



1901

1902

1903 **Figure 2.** Structure and organization of the plX plasmid from *X. fuscans* subsp. *fuscans*
1904 str. Xff49. (A) Circular map of the plX plasmid constructed using DNAplotter. (B) Synteny
1905 comparison between plX and the pAX22 plasmid from *A. xylosoxidans* subsp.
1906 *denitrificans* constructed using Artemis Comparison Tool.

1907

1908 An overview of the results from the genome annotation is depicted in Table 2 and
1909 the pathogenesis-related genes that were identified on the chromosome and plasmids
1910 are summarized in Supplementary data 4. Based on the annotation of the strain 4834-R,
1911 we could identify many genes associated with the pathogenesis process, including the
1912 xanthan gum biosynthesis pathway, e.g., *gumBCDEFGHIJKLM*, ABC transporters, and a
1913 wide variety of effector proteins, including a transcription activator-like effector (TALE)
1914 protein, encoded by the gene CKU38_04498. The genome-wide analysis of virulence
1915 genes based on the genome of strain *4834-R* allowed the identification of more than 20
1916 genes involved in the pathogenesis, including *AvrB2*, *XopAD*, *XopAE*, *XopAK*, *XopAM*,
1917 *XopC2*, *XopE1*, *XopF1*, *XopF2*, *XopI*, *XopJ5*, *XopN*, *XopK*, *XopL*, *XopP2*, *XopP1*, *XopQ*,
1918 *XopR*, *XopT*, *XopV*, *XopX*, *XopZ* and *XfuTAL2*.

1919

1920

Table 2. Overview of the features identified in the genome sequence of *X. fuscans* subsp. *fuscans* strain Xff49

Replicon	CDS	tRNA	rRNAs	tmRNAs	ncRNAs and regulatory*
Chromosome	4218	61	7	1	48
Plasmid plA	49	0	0	0	1
Plasmid plC	37	0	0	0	0
Plasmid plX	20	0	0	0	0

*= Other families of ncRNAs that do not belong to the tRNA, rRNA or tmRNA families, and regulatory elements (eg: riboswitches).

The annotation of the Xff49 genome allowed the identification of a wide variety of non-coding elements (Supplementary Data 5), such as those from the *Xanthomonas* small RNA families (sX and asX), some of which have already been demonstrated as regulators of the pathogenesis process in other species of the *Xanthomonas* genus [39], such as *X. campestris* and *X. oryzae*, but not in *X. fuscans*. SX13 (CKU38_04192), for example, was already pointed as a major regulator of the virulence in *X. campestris* pv. *vesicatoria* due to its effect on the expression of many genes associated with the *Hrp*-regulon [40].

In the annotation of the plasmid plA we have identified a non-coding element from the Rfam [41] family RF01695, which potentially transcribes to a C4 antisense-like RNA used by many phages that infect bacteria. This ncRNAs are present in purified phages and were also identified in many bacteria genome as well. Although this feature was absent in the original annotation of the plA plasmid of 4834-R, a manual analysis using the cmsearch program for the INFERNAL package [42] revelated that it is also present and overlaps a conserved hypothetical protein with no domains identified using InterproScan [43], Pfam [44] or SMART [45].

The comparative analysis of protein-coding genes shared and differentially distributed among Xff49 and 4834-R revelated that there are at least 3632 ortholog

groups conserved between these two strains (Figure 3), although these number might be higher if considering the genes that might be absent in the Xff49 genome assembly as a consequence of misassembles (eg: transposase genes). Additionally, 7 groups were identified as exclusive to 4834-R, and 37 to Xff49

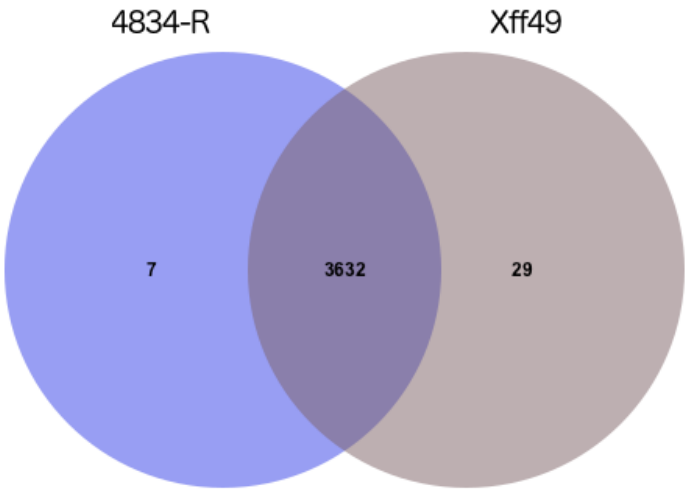


Figure 3. Venn diagram illustrating the genes shared by genomes of Xff49 and 4834-R and those exclusive to each strain as identified by the OrthoVenn program.

From the ortholog groups identified as absent in 4834-R based on the OrthoVenn analysis, we have identified one, which present two copies in the genome (CKU38_01232 and CKU38_00615), that encodes an AbiF-like from the abortive infection system, which is associated with resistance to phage infection. In the context of bacterial plant disease, phages have already been pointed as potential candidates for biocontrol of phytopathogens in many crop species [46]. Therefore, the adaptation of the resistance to phage infection in phytopathogens may become an issue in the future in crop protection. Another ortholog group identified as exclusive in Xff49 encodes a long-chain-specific 3-oxoacyl-acyl carrier protein reductase (CKU38_02114 and CKU38_02115), which has already been demonstrated as acting in the diffusible signal factor (DSF) quorum sensing regulation, a process that is important in the pathogenesis of many phytopathogens, including *X. campestris* [47]. Additionally, we also have identified two genes (CKU38_04417 and CKU38_04433) that encode outer-membrane efflux pump proteins from the

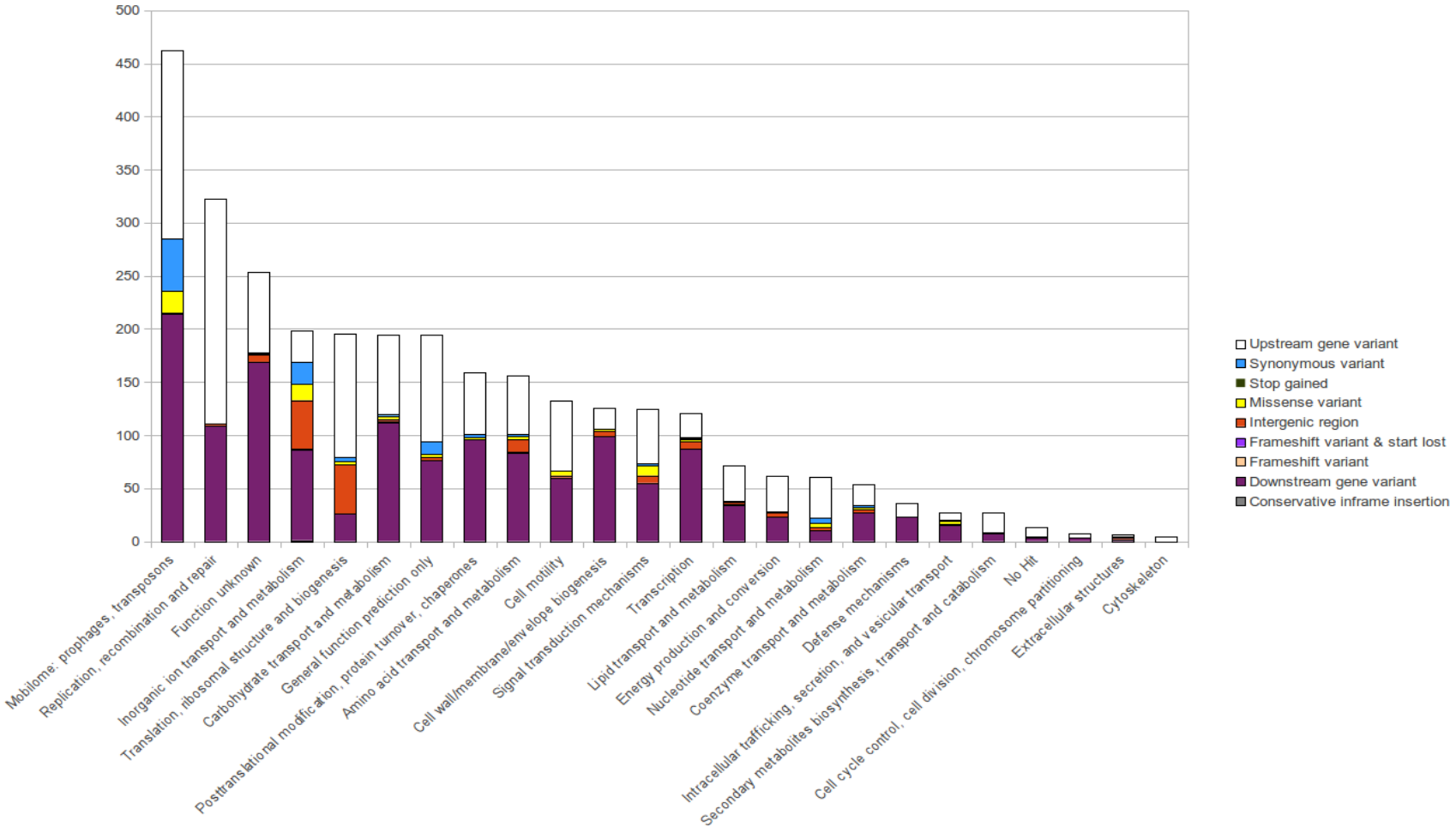
resistance-nodulation-cell division superfamily (RND). The RND family of transporters are Gran-negative bacteria, and usually associated with the active transporting of antibiotics [48]. Finally, two beta-lactamase genes with no correspondence in 4834-R were identified in Xff49 (CKU38_00306 and CKU38_01422). Beta-lactamases confer resistance to beta-lactam antibiotics, which are synthesized by many filamentous fungi such as those from the genera *Aspergillus* and *Penicillium* [49]. The identification of these genes in phytopathogens is particularly important not only because they may provide resistance to molecules secreted by fungi present in the environment and surface of the host plant, but also due to the potential risks of horizontal transferring to animal and human pathogens [50]. The complete output from the OrthoVenn analysis is available in Supplementary Data 6.

Based on the analysis of the Xff49 genome using the EfectiveDB webserver we were able to identify 557 putative T3Es and 47 putative T4Es, from which 67 T3Es and 8 T4Es have no homologue identified in 4834-R (Supplementary data 7). However, a manual revision of these hits demonstrated that all proteins identified as secreted by these secretion systems are located, in fact, in the cytoplasm or in the membrane. Therefore, these results were not used in further analysis.

Based on the AntiSMASH analysis we were able to identify in Xff49 and 4834-R complete gene clusters responsible for the biosynthesis of xanthoferrin, a siderophore molecule secreted by bacteria from the *Xanthomonas* genus that is used in the uptaking of iron from the environment [51]. Gene clusters for xanthan and xanthomonadin biosynthetic pathways were also identified in the genome of these strains, although incomplete. Xanthan gum plays an important role in the formation of biofilm and helps the adhesion of the bacteria to the surface on the host plant, thus, been important in its pathogenesis. The xanthan biosynthesis gene cluster was identified with 61% of coverage both in Xff49 and 4834-R [52]. Xanthomonadin, a yellow pigment produced by many bacteria in the *Xanthomonas* genus, is important in the protection of the DNA from ultraviolet light, improving the ability to survive in the environment [53]. The gene cluster for

the biosynthesis of this molecule was identified with coverage of 64% in Xff49 and 71% in 4834-R.

Our variant calling analysis identified 299 SNPs with QUAL ≥ 40 Q and 19 INDELs with coverage ≥ 30×. However, most of these are located in intergenic regions or do not have a functional impact on protein level (synonymous mutation), indicating that the 4834-R and Xff49 strains share a high similarity at the functional level in shared genes. The distribution of the short variants identified in the variant calling analysis in different groups of biological processes indexed in the COG database is displayed in Figure 4. As indicated, only a few processes, such as those associated with the mobilome (e.g.: transposons), inorganic ion metabolism, and signal transduction, were affected by missense mutations, while most of the other processes might be only affected by intergenic mutations or silent mutations (whose effects are much more difficult to predict considering only genomic data).



2012

2013 **Figure 4.** Distribution of the biological processed affected by single-nucleotide variants identified in the genome of *X.*
2014 *fuscans* subsp. *fuscans* str. Xff49 based on the reference genome of *X. fuscans* subsp. *fuscans* strain 4834-R. Biological
2015 processed are named based on the COG classification.

Phylogenetic analysis of the strains based on the multiple genome comparisons revealed a high similarity between the different strains of *Xff* (Figure 5). It also revealed that the subspecies *fuscans* and *aurantifolii* are relatively phylogenetically distant. It was also observed that the subspecies *aurantifolii* was subdivided into two main branches, indicating that there might be a higher genetic variability inside this subspecies when compared to the subspecies *fuscans*. It is unclear, however, how much these two pathovars differ at a functional level, although this difference might be directly associated with the differences in the preferred host for each group [54].

In the analysis of the flagellar genes, we have identified the ortholog genes of the gene cluster initially characterized in *X. campestris* pv. *vesicatoria* in the genomes of Xff49, 4834-R and *X. fuscans* pv. *aurantifolli* str. ICPB 10535 (*Xac*) (Supplementary data 8). The organization of the gene cluster in the different strains was analyzed using the Artemis Comparison Tool, and the results are presented in Supplementary Data 9. By comparing the results for each strain we were able to identify that Xff49 presents a higher number of genes from these gene cluster when comparing to 4834-R, even been an unfinished genomic record (with some gaps), and its “gene profile” is similar to *X. fuscans* pv. *aurantifolli* str. ICPB 10535, which was already experimentally validated as presenting cell motility and a polar flagellum in previous studies.

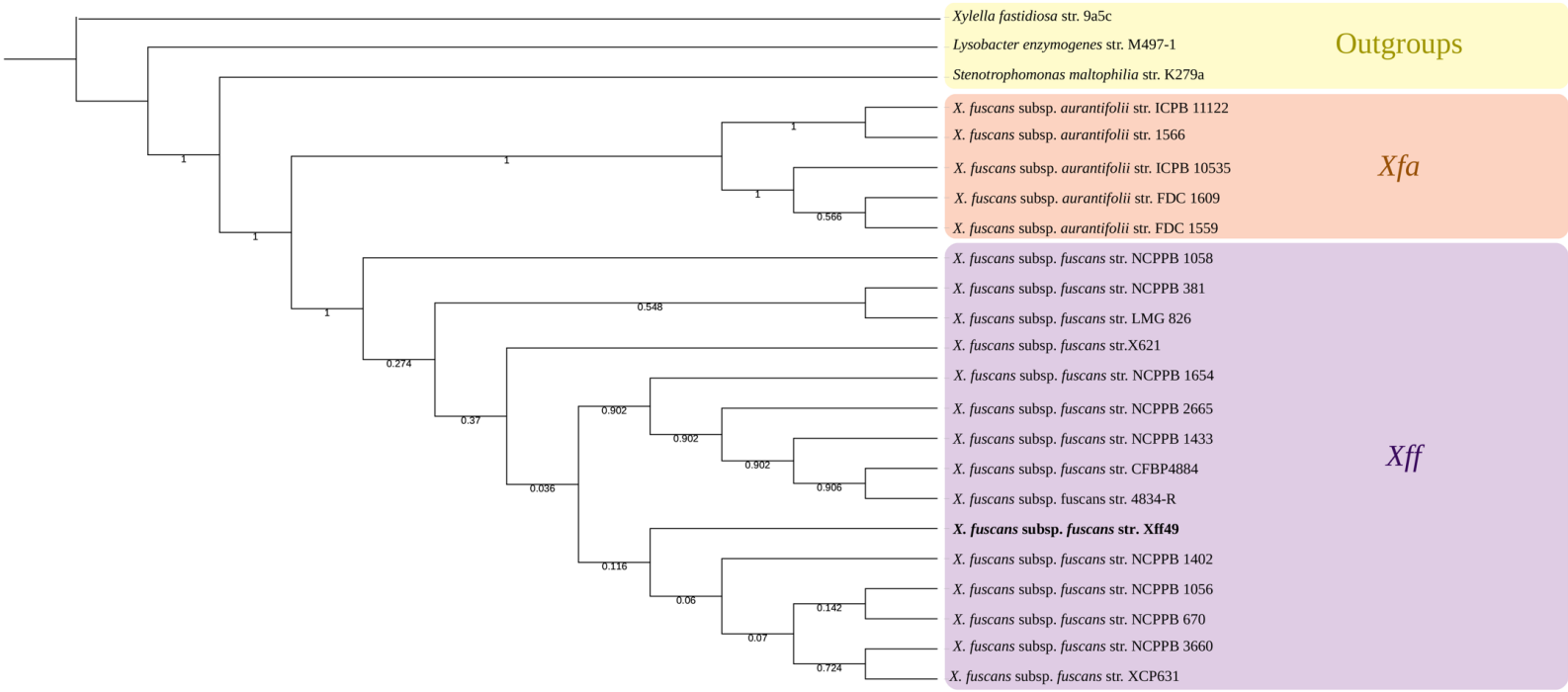


Figure 5. Phylogenetic tree constructed based on the alignment of syntenic blocks identified among different strains of *X. fuscans* subsp. *fuscans* and *X. fuscans* subsp. *aurantifolii*. A genome sequence of *X. oryzae* pv. *oryzicola* was used as outgroup.

Although the phylogenetics and variants calling analysis demonstrated that these strains are closely-related, the comparative analysis based on orthologs groups identified some genes that differ among these two strains, including genes associated with antibiotic resistance and regulation of quorum-sensing, which may provide different capabilities of survining in the environment or in the host. The identification of new antibiotic resistance genes in a phytopathogen may also provide useful information for public health researchers interested in the analysis of the transfer of these genes from environmental microorganisms to human and animal pathogens. Finally, we also demonstrated that the genomic deletion of genes from the flagellar cluster observed in 4834-R is absent in Xff49, and the presence of these genes is similar to the strain ICPB 10535, which was already reported with a functional flagellum.

Conclusion

In the paucity of effective methods to control CBB infection in bean crop, development of genetically resistant cultivars is considered to be an effective management approach. In this regard, the availability of new genomic data of *X. fuscans* subsp. *fuscans* might provide novel and useful information for researchers working on the analysis of its molecular features, pathogenomics, molecular genetics, and phylogenetics of the genus and species. As Xff49 is the first Brazilian strain to be sequenced and made publicly available, it might be used for multiple genome comparisons and genome plasticity by comparing this strain to other strains isolated in South America. These studies, in turn, may pave the way for identification of cultivars providing resistance to CBB.

Sequence data from this article have been deposited with the DDBJ/EMBL/GenBank Data Libraries under Accession Numbers CP023294.1 (*X. fuscans* subsp. *fuscans* Xff49 complete chromosome), CP023295.1 (*X. fuscans* subsp. *fuscans* Xff49 plasmid plA), CP023296.1 (*X. fuscans* subsp. *fuscans* Xff49 plasmid plC) and CP023297.1 (*X. fuscans* subsp. *fuscans* Xff49 plasmid plX).

Funding: This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil).

References

1. Ryan RP, Vorhölter F-J, Potnis N, et al. Pathogenomics of Xanthomonas: understanding bacterium–plant interactions. *Nat Rev Microbiol*. 2011;9(5):344-355. doi:10.1038/nrmicro2558.

2. Akhavan A, Bahar M, Askarian H, Lak MR, Nazemi A, Zamani Z. Bean common bacterial blight: pathogen epiphytic life and effect of irrigation practices. *Springerplus*. 2013;2(1):41. doi:10.1186/2193-1801-2-41.

3. Morris CE, Monier J-M. The ecological significance of biofilm formation by plant-associated bacteria. *Annu Rev Phytopathol*. 2003;41(1):429-453. doi:10.1146/annurev.phyto.41.022103.134521.

4. Graham PH, Vance CP. Legumes: Importance and Constraints to Greater Use. *PLANT Physiol*. 2003;131(3):872-877. doi:10.1104/pp.017004.

5. Shi C, Navabi A, Yu K. Association mapping of common bacterial blight resistance QTL in Ontario bean breeding populations. *BMC Plant Biol*. 2011;11(1):52. doi:10.1186/1471-2229-11-52.

6. Durham KM, Xie W, Yu K, Pauls KP, Lee E, Navabi A. Interaction of common bacterial blight quantitative trait loci in a resistant inter-cross population of common bean. Link W, ed. *Plant Breed*. 2013;132(6):658-666. doi:10.1111/pbr.12103.

7. Zhu J, Wu J, Wang L, Blair MW, Zhu Z, Wang S. QTL and candidate genes associated with common bacterial blight resistance in the common bean cultivar Longyundou 5 from China. *Crop J*. 2016;4(5):344-352. doi:10.1016/j.cj.2016.06.009.

8. Jacques M-A, Josi K, Darrasse A, Samson R. Xanthomonas axonopodis pv. phaseoli var. fuscans Is Aggregated in Stable Biofilm Population Sizes in the

2100 Phyllosphere of Field-Grown Beans. *Appl Environ Microbiol*. 2005;71(4):2008-2015.
2101 doi:10.1128/AEM.71.4.2008-2015.2005.

2102 9. Ruh M, Briand M, Bonneau S, Jacques M-A, Chen NWG. Xanthomonas adaptation
2103 to common bean is associated with horizontal transfers of genes encoding TAL effectors.
2104 *BMC Genomics*. 2017;18(1):670. doi:10.1186/s12864-017-4087-6.

2105 10. Jacques M-A, Josi K, Darrasse A, Samson R. Xanthomonas axonopodis pv.
2106 phaseoli var. fuscans is aggregated in stable biofilm population sizes in the phyllosphere
2107 of field-grown beans. *Appl Environ Microbiol*. 2005;71(4):2008-2015.
2108 doi:10.1128/AEM.71.4.2008-2015.2005.

2109 11. Darrasse A, Carrère S, Barbe V, et al. Genome sequence of Xanthomonas fuscans
2110 subsp. fuscans strain 4834-R reveals that flagellar motility is not a general feature of
2111 xanthomonads. *BMC Genomics*. 2013;14:761. doi:10.1186/1471-2164-14-761.

2112 12. Indiana A, Briand M, Arlat M, et al. Draft Genome Sequence of the Flagellated
2113 Xanthomonas fuscans subsp. fuscans Strain CFBP 4884. *Genome Announc*. 2014;2(5).
2114 doi:10.1128/genomeA.00966-14.

2115 13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
2116 sequence data. *Bioinformatics*. 2014;30(15):2114-2120.
2117 doi:10.1093/bioinformatics/btu170.

2118 14. Tritt A, Eisen JA, Facciotti MT, Darling AE. An integrated pipeline for de novo
2119 assembly of microbial genomes. *PLoS One*. 2012;7(9):e42304.
2120 doi:10.1371/journal.pone.0042304.

2121 15. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using
2122 de Bruijn graphs. *Genome Res*. 2008;18(5):821-829. doi:10.1101/gr.074492.107.

2123 16. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using
2124 compressed data structures. *Genome Res*. 2012;22(3):549-556.
2125 doi:10.1101/gr.126953.111.

2126 17. Boisvert S, Laviolette F, Corbeil J. Ray: simultaneous assembly of reads from a
2127 mix of high-throughput sequencing technologies. *J Comput Biol*. 2010;17(11):1519-1533.
2128 doi:10.1089/cmb.2009.0238.

18. Lin SH, Liao YC. CISA: Contig Integrator for Sequence Assembly of Bacterial Genomes. *PLoS One*. 2013;8(3):1-7. doi:10.1371/journal.pone.0060843.

19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2.

20. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10(1):421. doi:10.1186/1471-2105-10-421.

21. Bosi E, Donati B, Galardini M, et al. MeDuSa: a multi-draft based scaffolder. *Bioinformatics*. 2015;31(15):2443-2451. doi:10.1093/bioinformatics/btv171.

22. Lu CL, Chen KT, Huang SY, Chiu HT. Car: Contig assembly of prokaryotic draft genomes using rearrangements. *BMC Bioinformatics*. 2014;15(1):1-10. doi:10.1186/s12859-014-0381-3.

23. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol*. 2010;11(4):R41. doi:10.1186/gb-2010-11-4-r41.

24. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-1075. doi:10.1093/bioinformatics/btt086.

25. Kremer FS, Eslabão MR, Dellagostin OA, Pinto L da S. Genix: A New Online Automated Pipeline for Bacterial Genome Annotation. *FEMS Microbiol Lett*. 2016;11(16).

26. Rutherford K, Parkhill J, Crook J, et al. Artemis: Sequence visualization and annotation. *Bioinformatics*. 2000;16(10):944-945. doi:10.1093/bioinformatics/16.10.944.

27. Wang Y, Coleman-Derr D, Chen G, Gu YQ. OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res*. 2015;43(W1):W78-84. doi:10.1093/nar/gkv487.

28. Eichinger V, Nussbaumer T, Platzer A, Jehl M-A, Arnold R, Rattei T. EffectiveDB—updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Res*. 2016;44(D1):D669-D674. doi:10.1093/nar/gkv1269.

29. Blin K, Wolf T, Chevrette MG, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res*. 2017;45(W1):W36-W41. doi:10.1093/nar/gkx319.

30. Angiuoli S V, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*. 2011;27(3):334-342. doi:10.1093/bioinformatics/btq665.

31. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972-1973. doi:10.1093/bioinformatics/btp348.

32. Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423. doi:10.1093/bioinformatics/btp163.

33. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol Biol Evol*. May 2018. doi:10.1093/molbev/msy096.

34. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.

35. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92. doi:10.4161/fly.19695.

36. Marchler-Bauer A, Derbyshire MK, Gonzales NR, et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res*. 2014;43(Database issue):D222-6. doi:10.1093/nar/gku1221.

37. Tatusov RL. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. 2000;28(1):33-36. doi:10.1093/nar/28.1.33.

38. Moreira LM, Almeida NF, Potnis N, et al. Novel insights into the genomic basis of citrus canker based on the genome sequences of two strains of Xanthomonas fuscans subsp. aurantifolii. *BMC Genomics*. 2010;11(1):238. doi:10.1186/1471-2164-11-238.

2185 39. Abendroth U, Schmidtke C, Bonas U. Small non-coding RNAs in plant-pathogenic
2186 Xanthomonas spp. *RNA Biol*. 2014;11(5):457-463. doi:10.4161/rna.28240.

2187 40. Schmidtke C, Abendroth U, Brock J, Serrania J, Becker A, Bonas U. Small RNA
2188 sX13: A Multifaceted Regulator of Virulence in the Plant Pathogen Xanthomonas. Waldor
2189 MK, ed. *PLoS Pathog*. 2013;9(9):e1003626. doi:10.1371/journal.ppat.1003626.

2190 41. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA
2191 family database. *Nucleic Acids Res*. 2003;31(1):439-441.
2192 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165453&tool=pmcentrez&ren
2193 dertype=abstract. Accessed March 26, 2015.

2194 42. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments.
2195 *Bioinformatics*. 2009;25(10):1335-1337. doi:10.1093/bioinformatics/btp157.

2196 43. Zdobnov EM, Apweiler R. InterProScan--an integration platform for the signature-
2197 recognition methods in InterPro. *Bioinformatics*. 2001;17(9):847-848.
2198 http://www.ncbi.nlm.nih.gov/pubmed/11590104. Accessed May 12, 2014.

2199 44. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database.
2200 *Nucleic Acids Res*. 2014;42(Database issue):D222-30. doi:10.1093/nar/gkt1223.

2201 45. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain
2202 annotation resource. *Nucleic Acids Res*. 2012;40(Database issue):D302-5.
2203 doi:10.1093/nar/gkr931.

2204 46. Buttimer C, McAuliffe O, Ross RP, Hill C, O'Mahony J, Coffey A. Bacteriophages
2205 and Bacterial Plant Diseases. *Front Microbiol*. 2017;8:34. doi:10.3389/fmicb.2017.00034.

2206 47. Hu Z, Dong H, Ma J-C, et al. Novel *Xanthomonas campestris* Long-Chain-Specific
2207 3-Oxoacyl-Acyl Carrier Protein Reductase Involved in Diffusible Signal Factor Synthesis.
2208 Lindow SE, ed. *MBio*. 2018;9(3). doi:10.1128/mBio.00596-18.

2209 48. Nikaido H, Takatsuka Y. Mechanisms of RND multidrug efflux pumps. *Biochim*
2210 *Biophys Acta - Proteins Proteomics*. 2009;1794(5):769-781.
2211 doi:10.1016/j.bbapap.2008.10.004.

2212 49. Brakhage AA. Molecular regulation of beta-lactam biosynthesis in filamentous

fungi. *Microbiol Mol Biol Rev*. 1998;62(3):547-585. http://www.ncbi.nlm.nih.gov/pubmed/9729600. Accessed October 7, 2018.

50. Wright GD. Antibiotic resistance in the environment: a link to the clinic? *Curr Opin Microbiol*. 2010;13(5):589-594. doi:10.1016/j.mib.2010.08.005.

51. Pandey SS, Patnana PK, Rai R, Chatterjee S. Xanthoferrin, the **α** -hydroxycarboxylate-type siderophore of *Xanthomonas campestris* pv. *campestris* , is required for optimum virulence and growth inside cabbage. *Mol Plant Pathol*. 2017;18(7):949-962. doi:10.1111/mpp.12451.

52. Bianco MI, Toum L, Yaryura PM, et al. Xanthan Pyruvilation Is Essential for the Virulence of *Xanthomonas campestris* pv. *campestris*. *Mol Plant-Microbe Interact*. 2016;29(9):688-699. doi:10.1094/MPMI-06-16-0106-R.

53. Poplawsky AR, Urban SC, Chun W. Biological Role of Xanthomonadin Pigments in Xanthomonas campestris pv. Campestris. *Appl Environ Microbiol*. 2000;66(12):5123. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC92432/. Accessed June 11, 2018.

54. Huang C-L, Pu P-H, Huang H-J, et al. Ecological genomics in Xanthomonas: the nature of genetic adaptation with homologous recombination and host shifts. *BMC Genomics*. 2015;16(1):188. doi:10.1186/s12864-015-1369-8.

Suplementary Data 1. *X. fuscans* genome sequences used in the reference-guided genome scaffolding of *X. fuscans* subsp. fuscans strain Xff49

Name	Accessions	Status
X. fuscans subsp. fuscans str. 4834-R	FO681494.1, FO681495.1, FO681496.1 and FO681497.1	Finished
X. fuscans subsp. *fuscans* str.X621	JXHS00000000.3	Draft
X. fuscans subsp. *fuscans* str. XCP631	JXLW00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 670	JRRE00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 381	JTKK00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 3660	JSEX00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 2665	JSBQ00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1654	JSBR00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1433	JSBT00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1402	JSEW00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1058	JSEY00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1056	JSEV00000000.2	Draft
X. fuscans subsp. *fuscans* str. LMG 826	JPYF00000000.1	Draft
X. fuscans subsp. *fuscans* str. CFBP4884	JPHG00000000.1	Draft

Suplementary Data 2. Genome sequences used in the phylogenetic analysis. The genome of *Xylella fastidiosa* str. 9a5c, *Lysobacter enzymogenes* str. M497-1 and *Stenotrophomonas maltophilia* str. K279a were used as outgroups.

Name	Accessions	Status
X. fuscans subsp. fuscans str. 4834-R	FO681494.1, FO681495.1, FO681496.1 and FO681497.1	Finished
X. fuscans subsp. *fuscans* str.X621	JXHS00000000.3	Draft
X. fuscans subsp. *fuscans* str. XCP631	JXLW00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 670	JRRE00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 381	JTKK00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 3660	JSEX00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 2665	JSBQ00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1654	JSBR00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1433	JSBT00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1402	JSEW00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1058	JSEY00000000.2	Draft
X. fuscans subsp. *fuscans* str. NCPPB 1056	JSEV00000000.2	Draft
X. fuscans subsp. *fuscans* str. LMG 826	JPYF00000000.1	Draft
X. fuscans subsp. *fuscans* str. CFBP4884	JPHG00000000.1	Draft
X. fuscans subsp. *aurantifolii* str. ICPB 11122	ACPX00000000.1	Draft
X. fuscans subsp. *aurantifolii* str. ICPB 10535	ACPY00000000.1	Draft
X. fuscans subsp. *aurantifolii* str. FDC 1559	CP011160.1	Finished
X. fuscans subsp. *aurantifolii* str. FDC 1609	CP011163.1	Finished
X. fuscans subsp. *aurantifolii* str. 1566	CP012002.1	Finished
Xylella fastidiosa str. 9a5c	AE003849.1	Finished
Lysobacter enzymogenes str. M497-1	AP014940.1	Finished

	Stenotrophomonas maltophilia str. K279a	AM743169.1	Finished

2235

Supplementary Data 3. Pathogenesis-related genes identified in the genome of *X. fuscans* subsp. fuscans str. Xff49 based on the genome of str. 4834-R.

Function	Locus tag		Identity (%)
	4834-R	Xff49	
Part of the type 1 pilus formation process	XFF4834R_chr30690	CKU38_01620	99.43
Part of the type 1 pilus formation process	XFF4834R_chr30700	CKU38_01619	100.00
Part of the type 1 pilus formation process	XFF4834R_chr30710	CKU38_01618	99.87
Part of the type 1 pilus formation process	XFF4834R_chr30720	CKU38_01617	100.00
Part of the type 1 pilus formation process	XFF4834R_chr30730	CKU38_01616	100.00
Part of the type 1 pilus formation process	XFF4834R_chr30740	CKU38_01615	100.00
YapH autotransporter (hemagglutinin-like)	XFF4834R_chr22670	CKU38_02338	100.00
YapH autotransporter (hemagglutinin-like)	XFF4834R_chr42170	CKU38_04457	99.87
XadA1 (YadA-like autotransporter)	XFF4834R_chr34400	CKU38_03622	99.52
XadA2 (YadA-like autotransporter)	XFF4834R_chr34420	CKU38_03624	99.81
FhaC (hemagglutinin-like)	XFF4834R_chr19440	CKU38_01979	98.23
Hemagglutinin	XFF4834R_chr39830	CKU38_04198	100.00
gumB (xanthan gun biosynthesis pathway)	XFF4834R_chr26110	CKU38_02687	99.62
gumC (xanthan gun biosynthesis pathway)	XFF4834R_chr26120	CKU38_02688	100.00
gumD (xanthan gun biosynthesis pathway)	XFF4834R_chr26130	CKU38_02689	100.00
gumE (xanthan gun biosynthesis pathway)	XFF4834R_chr26140	CKU38_02690	100.00
gumF (xanthan gun biosynthesis pathway)	XFF4834R_chr26150	CKU38_02691	100.00
gumG (xanthan gun biosynthesis pathway)	XFF4834R_chr26160	CKU38_02692	100.00
gumH (xanthan gun biosynthesis pathway)	XFF4834R_chr26170	CKU38_02693	99.72

gumI (xanthan gun biosynthesis pathway)	XFF4834R_chr26180	CKU38_02694	99.73
gumJ (xanthan gun biosynthesis pathway)	XFF4834R_chr26190	CKU38_02695	100.00
gumK (xanthan gun biosynthesis pathway)	XFF4834R_chr26200	CKU38_02696	99.39
gumL (xanthan gun biosynthesis pathway)	XFF4834R_chr26210	CKU38_02697	100.00
gumM (xanthan gun biosynthesis pathway)	XFF4834R_chr26220	CKU38_02698	100.00
xanA (xanthan gun biosynthesis pathway)	XFF4834R_chr34730	CKU38_03655	99.57
xanB (xanthan gun biosynthesis pathway)	XFF4834R_chr34740	CKU38_03656	99.49
xagA (xanthan gun biosynthesis pathway)	XFF4834R_chr34180	CKU38_03600	100.00
xagB (xanthan gun biosynthesis pathway)	XFF4834R_chr34190	CKU38_03601	99.84
HrpX (control of the expression of T3SS and T3Es[b])	XFF4834R_chr32690	CKU38_01364	100.00
ABC Transporter	XFF4834R_chr29870	CKU38_01681	100.00
ABC Transporter	XFF4834R_chr29880	CKU38_01680	100.00
ABC Transporter	XFF4834R_chr29890	CKU38_01679	99.57
ABC Transporter	XFF4834R_chr24540	CKU38_02550	99.29
ABC Transporter	XFF4834R_chr24550	CKU38_02551	100.00
ABC Transporter	XFF4834R_chr24570	CKU38_02555	100.00
ABC Transporter	XFF4834R_chr24580	CKU38_02555	100.00
ABC Transporter	XFF4834R_chr24590	CKU38_02557	99.75
ABC Transporter	XFF4834R_chr24600	CKU38_02558	100.00
ABC Transporter	XFF4834R_chr35340	CKU38_03713	99.76
ABC Transporter	XFF4834R_chr35350	CKU38_03714	100.00
ABC Transporter	XFF4834R_chr35360	CKU38_03715	100.00
ABC Transporter	XFF4834R_chr35370	CKU38_03716	100.00
ABC Transporter	XFF4834R_chr38590	CKU38_04098	100.00

ABC Transporter	XFF4834R_chr38610	CKU38_04035	99.76
ABC Transporter	XFF4834R_chr38620	CKU38_04099	100.00
ABC Transporter	XFF4834R_chr38630	CKU38_04100	100.00
ABC Transporter	XFF4834R_chr38640	CKU38_04101	100.00
ABC Transporter	XFF4834R_chr40790	CKU38_04325	100.00
ABC Transporter	XFF4834R_chr40800	CKU38_04326	99.73
ABC Transporter	XFF4834R_chr40810	CKU38_04327	100.00
ABC Transporter	XFF4834R_chr17340	CKU38_03185	99.44
AvrB2 (avirulence gene)	XFF4834R_chr00460	CKU38_00088	100.00
XopAD (virulence gene)	XFF4834R_chr40870	CKU38_04332	84.92
XopAE (virulence gene)	XFF4834R_chr38990	CKU38_04137	100.00
XopAK (virulence gene)	XFF4834R_chr35620	CKU38_03744	100.00
XopAM (virulence gene)	XFF4834R_chr33550	CKU38_01267	99.90
XopC2 (virulence gene)	XFF4834R_chr33300	CKU38_01253	100.00
XopE1 (virulence gene)	XFF4834R_chr02600	CKU38_00296	99.75
XopF1 (virulence gene)	XFF4834R_chr03180	CKU38_00367	100.00
XopF2 (virulence gene)	XFF4834R_chr18460	CKU38_03334	99.76
XopI (virulence gene)	XFF4834R_chr07620	CKU38_00805	100.00
XopJ5 (virulence gene)	XFF4834R_chr16310	CKU38_03088	100.00
XopK (virulence gene)	XFF4834R_chr15450	CKU38_03008	98.77
XopL (virulence gene)	XFF4834R_chr15400	CKU38_03004	100.00
XopN (virulence gene)	XFF4834R_chr18430	CKU38_03330	99.86
XopP1 (virulence gene)	XFF4834R_chr33320	CKU38_01291	100.00

XopP2 (virulence gene)	XFF4834R_chr16310	CKU38_03088	100.00
XopQ (virulence gene)	XFF4834R_chr42130	CKU38_04452	100.00
XopR (virulence gene)	XFF4834R_chr25420	CKU38_01427	100.00
XopT (virulence gene)	XFF4834R_chr23790	CKU38_02465	100.00
XopV (virulence gene)	XFF4834R_chr42980	CKU38_04487	100.00
XopX (virulence gene)	XFF4834R_chr42980	CKU38_04487	100.00
XopZ (virulence gene)	XFF4834R_chr21120	CKU38_02168	100.00
XfuTAL2 (virulence gene)	XFF4834R_pla00470	CKU38_04498	100.00

[b] = Type III effector proteins

Supplementary Data 4. Genes from the *A. xylosoxidans* subsp. *denitrificans* paX22 plasmid absent in the *X. fuscans* subsp. *fuscans* plX plasmid.

Locus_tag	Product	Location		
		Start	End	Strand
ST32002_p15	putative transposase	3251	3793	-
ST32002_p05	putative toxic component of a toxin/antitoxin system, T/AT1	3799	4194	-
ST32002_p02	putative transcriptional regulator of atoxin/antitoxin system, T/AT2	4191	4442	-
ST32002_p16	putative recombinase	4507	5190	+
ST32002_p18	class 1 integron integrase	5537	6550	-
ST32002_p19	metallo-beta-lactamase	6717	7517	+
ST32002_p23	aminoglycoside acetyl transferase	7661	8179	+
ST32002_p21	aminoglycoside phosphotransferase	8250	9044	+
ST32002_p22	aminoglycoside adenyltransferase	9161	9952	+
ST32002_p09	QacE-delta1	10076	10423	+
ST32002_p07	Sul1	10417	11256	+
ST32002_p17	hypothetical protein	11384	11884	+
ST32002_p03	TniB-delta3	11941	12849	-
ST32002_p04	TniA	12843	14558	-
ST32002_p24	hypothetical protein	14743	14964	+
ST32002_p20	HsdFM	14957	16105	+
ST32002_p26	hypothetical protein	16105	17034	+

Supplementary Data 5. Non-coding elements identified in the genome of *X. fuscans* subsp. *fuscans* str. Xff49

Locus_tag	Rfam		
	Family	**Accession**	***Xanthomonas* species** [a]
CKU38_00115	sX2 sRNA [b]	RF02222	*X. albilineans* *X. campestris* pv. *campestris* *X. citri* pv. *citri* *X. euvesicatoria* *X. fragariae* *X. translucens* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_00405	asX1 asRNA [c]	RF02235	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* *X. campestris* pv. *campestris* *X. fragariae* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_00570	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_00652	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria*

			X. oryzae pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_00656	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_00765	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_00798	FMN [d] Riboswitch	RF00050	*X. fragariae* *X. albilineans* *X. campestris* pv. *campestris* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_00817	RNase P	RF00011	None
CKU38_00835	Xoo8 sRNA	RF02243	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. fragariae* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*

CKU38_00949	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_01135	sX4 sRNA	RF02223	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. fragariae* *X. oryzae* pv. *oryzae*
CKU38_01146	SRP ^{e} RNA	RF00169	*X. albilineans* *X. campestris* pv. *campestris* *X. fragariae* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_01233	Xcc1 sRNA	RF02221	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. fragariae* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola* *X. translucens* pv. *undulosa*
CKU38_01264	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri*

			X. euvesicatoria *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_01286	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_01298	Glycine riboswitch	RF00504	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_01662	tmRNA [f]	RF00023	*X. albilineans* *X. axonopodis* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. fragariae* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_01801	sX5 sRNA	RF02224	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_01873	sX6 sRNA	RF02225	*X. translucens* pv. *undulosa*

			X. oryzae pv. *oryzae* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria*
CKU38_01901	sX7 sRNA	RF02226	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. fragariae* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_01983	sRNA-Xcc1	RF02221	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. fragariae* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola* *X. translucens* pv. *undulosa*
CKU38_02061	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_02341	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria*

			X. oryzae pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_02367	traJ-II RNA	RF01760	*X. albilineans* *X. fragariae*
CKU38_02448	Xoo5 sRNA	RF02242	*X. translucens* pv. *undulosa* *X. oryzae* pv. *oryzae* *X. campestris* pv. *campestris* *X. albilineans* *X. fragariae*
CKU38_02530	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *oryzicola* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_02577	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_02590	isrK Hfq binding RNA	RF01394	None
CKU38_02625	Xcc1 sRNA	RF02221	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris*

			X. euvesicatoria *X. fragariae* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola* *X. translucens* pv. *undulosa*
CKU38_02822	asX2 asRNA	RF02236	*X. euvesicatoria* *X. axonopodis* pv. *citri*
CKU38_02905	Cobalamin Riboswitch	RF00174	*X. albilineans* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. fragariae* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_03053	SAM [g] riboswitch	RF00634	None
CKU38_03060	Xoo2 sRNA	RF02241	*X. albilineans* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. fragariae* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_03089	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*

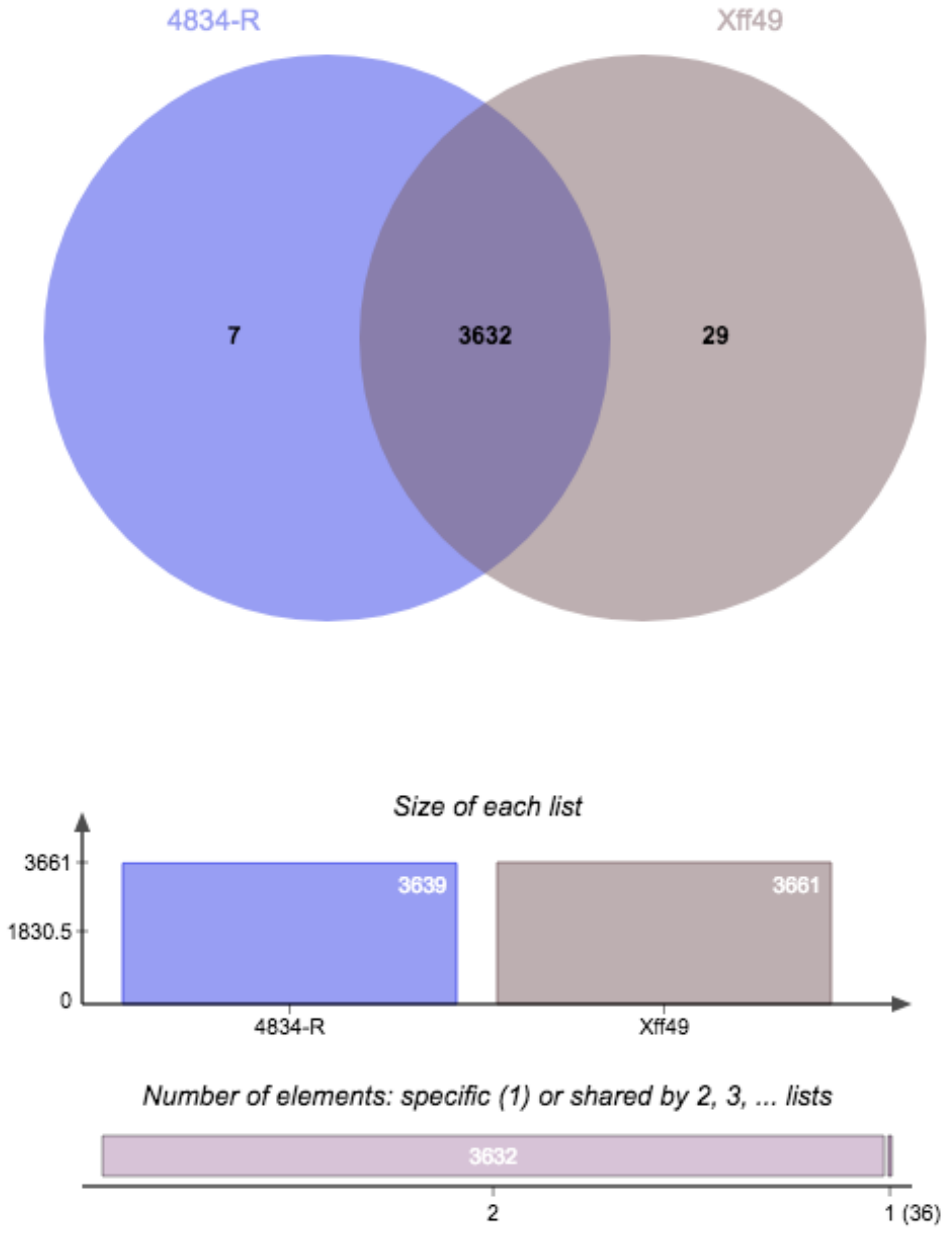
CKU38_03288	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_03459	6S / SsrS RNA	RF00013	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. fragariae* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_03467	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_03486	sX11 sRNA	RF02230	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzae*
CKU38_03498	Xoo1 sRNA	RF02240	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. fragariae* *X. oryzae* pv. *oryzae*

CKU38_03500	TPP [h] riboswitch	RF00059	*X. albilineans* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. fragariae* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_03734	sX12 sRNA	RF02231	*X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris*
CKU38_03873	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_04118	asX4 / Xoo4 sRNA	RF02238	*X. albilineans* *X. axonopodis* pv. *oryzicola* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_04139	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri*

			X. euvesicatoria *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_04145	sX9 sRNA	RF02228	*X. albilineans* *X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. citri* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. oryzae* pv. *oryzicola*
CKU38_04192	sX13 sRNA	RF02232	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*
CKU38_04308	yybP-ykoY manganese riboswitch	RF00080	*X. fragariae* *X. campestris* pv. *campestris*
CKU38_04396	sX14/Xoo3 sRNA	RF02233	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. euvesicatoria* *X. oryzae* pv. *oryzae*
CKU38_04513	C4 asRNA	RF01695	*X. axonopodis* pv. *citri* *X. campestris* pv. *campestris* *X. oryzae* pv. *oryzae* *X. translucens* pv. *undulosa*

[a] = *Xanthomonas* species in which these ncRNAs and non-coding elements were identified according to the Rfam database; [b] sRNA = small RNA; [c] asRNA = antisense RNA; [d] SRP = signal recognition particle; [e] FMN = flavin mononucleotide; [f] tmRNA = transfer-messenger RNA; [g] SAM = S-adenosyl methionine; [h] TPP = thiamine pyrophosphate.

Supplementary data 6. Veen Diagram generated using OrthoVeen representing the groups of orthologous genes shared between the strains 4834-R and Xff49 of *X. fuscans* subsp. *fuscans*.



Supplementary data 7. Genes from Xff49 identified as absent in 4834-R and predicted as potential effectors secreted by the T3SS or T4SS.

Gene	Product	InterPro	Secretion System	
			T3SS	T4SS
CKU38_00065	hypothetical protein		Y	
CKU38_00071	putative hemolysin-III family membrane protein	IPR004254: AdipoR/Haemolysin-III-related	Y	
CKU38_00081	Hypothetical protein		Y	
CKU38_00083	ubiquinone biosynthesis O-methyltransferase	IPR010233: Ubiquinone biosynthesis O-methyltransferase IPR029063: S-adenosyl-L-methionine-dependent methyltransferase	Y	
CKU38_00303	putative methyltransferase		Y	
CKU38_00377	putative Short-chain dehydrogenase/reductase SDR	IPR002347: Short-chain dehydrogenase/reductase SDR IPR036291: NAD(P)-binding domain superfamily	Y	
CKU38_00510	hypothetical protein		Y	
CKU38_00605	putative two-component regulatory system protein	IPR001932: PPM-type phosphatase domain IPR003660: HAMP domain IPR036457: PPM-type phosphatase domain superfamily	Y	
CKU38_00618	hypothetical protein		Y	
CKU38_00621	hypothetical protein		Y	
CKU38_00623	hypothetical protein	IPR025668: Transposase DDE domain	Y	
CKU38_00643	D-alanine/D-serine/glycine transporter	IPR002293: Amino acid/polyamine transporter I IPR004840: Amino acid permease, conserved site IPR004841: Amino acid permease/ SLC12A domain	Y	
CKU38_00937	secreted acyl-CoA synthetase (long-chain-fatty-acid--CoA ligase)	IPR000873: AMP-dependent synthetase/ligase IPR025110: AMP-binding enzyme, C-terminal domain	Y	
CKU38_00938	hypothetical protein		Y	

CKU38_00976	RNA polymerase factor sigma-70	IPR007627: RNA polymerase sigma-70 region 2 IPR013249: RNA polymerase sigma factor 70, region 4 type 2 IPR013324: RNA polymerase sigma factor, region 3/4-like IPR013325: RNA polymerase sigma factor, region 2 IPR014284: RNA polymerase sigma-70 like domain IPR036388: RNA polymerase sigma-B factor	Y
CKU38_00979	putative membrane protein		Y
CKU38_01163	hypothetical protein		Y
CKU38_01234	hypothetical protein		Y
CKU38_01235	hypothetical protein	IPR014944: Toxin SymE-like	Y
CKU38_01254	catalase	IPR002226: Catalase haem-binding site IPR010582: Catalase immune-responsive domain IPR011614: Catalase core domain IPR018028: Catalase, mono-functional, haem-containing IPR020835: Catalase superfamily IPR024708: Catalase active site IPR024712: Catalase, mono-functional, haem-containing, clade 2 IPR029062: Class I glutamine amidotransferase-like IPR037060: Catalase core domain superfamily	Y
CKU38_01351	hypothetical protein		Y
CKU38_01402	putative DNA mismatch repair protein, MutS	IPR000432: DNA mismatch repair protein MutS, C-terminal IPR005748: DNA mismatch repair protein MutS IPR007695: DNA mismatch repair protein MutS-like, N-terminal IPR007696: DNA mismatch repair protein MutS, core IPR007860: DNA mismatch repair protein MutS, connector domain IPR007861: DNA mismatch repair protein MutS, clamp	Y

		IPR016151: DNA mismatch repair protein MutS, N-terminal IPR017261: DNA mismatch repair protein MutS/MSH IPR027417: P-loop containing nucleoside triphosphate hydrolase IPR036187: DNA mismatch repair protein MutS, core domain superfamily IPR036678: MutS, connector domain superfamily		
CKU38_01485	hypothetical protein	IPR019180: Oxidoreductase-like, N-terminal IPR039251: Oxidoreductase-like domain-containing protein 1	Y	
CKU38_01538	Abi-like protein	IPR011664: Abortive infection system protein AbiD/AbiF-like	Y	Y
CKU38_01539	hypothetical protein		Y	
CKU38_01570	putative acetyl-CoA carboxylase, carboxytransferase, alpha subunit	IPR001095: Acetyl-CoA carboxylase, alpha subunit	Y	Y
CKU38_01927	hypothetical protein		Y	
CKU38_02116	putative 3-oxoacyl-[acyl-carrier-protein] synthase III	IPR013747: 3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III, C-terminal IPR013751: 3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III IPR016039: Thiolase-like	Y	
CKU38_02118	putative pyridoxal phosphate-dependent aminotransferase	IPR000653: DegT/DnrJ/EryC1/StrS aminotransferase IPR015421: Pyridoxal phosphate-dependent transferase, major domain IPR015424: Pyridoxal phosphate-dependent transferase	Y	
CKU38_02126	Flagellar hook-associated protein 2 C-terminus	IPR003481: Flagellar hook-associated protein 2, N-terminal IPR010809: Flagellar hook-associated 2, C-terminal IPR010810: Flagellin hook, IN motif	Y	
CKU38_02127		IPR001029: Flagellin, D0/D1 domain IPR001492: Flagellin	Y	

CKU38_02135	Xfu2 transposase	IPR010810: Flagellin hook, IN motif	Y
CKU38_02136	Flagellar hook-associated protein 1	IPR001444: Flagellar basal body rod protein, N-terminal IPR002371: Flagellar hook-associated protein 1 IPR010930: Flagellar basal-body/hook protein, C-terminal domain	Y
CKU38_02140	Flagellar basal-body rod FlgG	IPR001444: Flagellar basal body rod protein, N-terminal IPR010930: Flagellar basal-body/hook protein, C-terminal domain IPR012834: Flagellar basal-body rod FlgG IPR019776: Flagellar basal body rod protein, conserved site IPR020013: Flagellar hook-basal body protein, FlgE/F/G IPR037925: Flagellar hook-basal body protein, FlgE/F/G-like	Y
CKU38_02143	Flagellar basal body protein FlaE	IPR001444: Flagellar basal body rod protein, N-terminal IPR010930: Flagellar basal-body/hook protein, C-terminal domain IPR011491: Flagellar hook protein FlgE IPR019776: Flagellar basal body rod protein, conserved site IPR020013: Flagellar hook-basal body protein, FlgE/F/G IPR037058: Flagellar hook protein FlgE superfamily IPR037925: Flagellar hook-basal body protein, FlgE/F/G-like	Y
CKU38_02144	Flagellar hook capping protein	IPR005648: Flagellar hook capping protein IPR025963: FlgD Tudor-like domain IPR025965: FlgD Ig-like domain	Y
CKU38_02145	Flagellar basal-body rod protein FlgC	IPR001444: Flagellar basal body rod protein, N-terminal IPR006299: Flagellar basal-body rod protein FlgC	Y

		IPR010930: Flagellar basal-body/hook protein, C-terminal domain IPR019776: Flagellar basal body rod protein, conserved site		
CKU38_02303	Helix-destabilising protein	IPR003512: Bacteriophage M13, G5P, DNA-binding IPR012340: Nucleic acid-binding, OB-fold	Y	Y
CKU38_02311	putative EvpA, Type VI Secretion System	IPR008312: Type VI secretion system, VipA, VC_A0107 or Hcp2	Y	
CKU38_02365	Helix-turn-helix domain protein	IPR001387: Cro/C1-type helix-turn-helix domain IPR010982: Lambda repressor-like, DNA-binding domain superfamily	Y	
CKU38_02629	putative TonB-dependent transporter	IPR010916: TonB box, conserved site IPR012910: TonB-dependent receptor, plug domain IPR037066: TonB-dependent receptor, plug domain superfamily	Y	
CKU38_02839	putative methyl-accepting chemotaxis protein	IPR003660: HAMP domain IPR004089: Methyl-accepting chemotaxis protein (MCP) signalling domain IPR004090: Chemotaxis methyl-accepting receptor IPR024478: Chemotaxis methyl-accepting receptor HlyB-like, 4HB MCP domain	Y	
CKU38_03014	putative amino acid-polyamine-organocation (apc) superfamily transporter protein	IPR024478: Amino acid/polyamine transporter I	Y	
CKU38_03211	putative membrane fusion protein	IPR006143: RND efflux pump, membrane fusion protein IPR032317: RND efflux pump, membrane fusion protein, barrel-sandwich domain	Y	
CKU38_03217	hypothetical protein		Y	
CKU38_03218	Abi-like protein	IPR011664: Abortive infection system protein AbiD/AbiF-like	Y	Y
CKU38_03234	hypothetical protein		Y	
CKU38_03413	hypothetical protein		Y	
CKU38_03497	putative transcriptional regulator	IPR001867: OmpR/PhoB-type DNA-binding domain	Y	

		IPR011042: Six-bladed beta-propeller, TolB-like IPR011659: WD40-like Beta Propeller IPR016032: Signal transduction response regulator, C-terminal effector IPR036388: Winged helix-like DNA-binding domain superfamily		
CKU38_03571	hypothetical protein		Y	Y
CKU38_03851	hypothetical protein		Y	
CKU38_03866	putative primosomal protein N' (ATP-dependent helicase priA)	IPR005337: RapZ-like family	Y	Y
CKU38_03897	Transcription termination factor Rho	IPR000194: ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding domain IPR003593: AAA+ ATPase domain IPR004665: Transcription termination factor Rho IPR011112: Rho termination factor, N-terminal IPR011113: Rho termination factor, RNA-binding domain IPR011129: Cold shock domain IPR012340: Nucleic acid-binding, OB-fold IPR027417: P-loop containing nucleoside triphosphate hydrolase IPR036269: Rho termination factor, N-terminal domain superfamily	Y	Y
CKU38_03903	putative peptidoglycan-binding protein, lysM containing domain	IPR018392: LysM domain IPR036779: LysM domain superfamily	Y	
CKU38_03904	NADPH-dependent 7-cyano-7-deazaguanine reductase, QueF type 2	IPR016428: NADPH-dependent 7-cyano-7-deazaguanine reductase, QueF type 2 IPR029139: NADPH-dependent 7-cyano-7-deazaguanine reductase, N-terminal IPR029500: NADPH-dependent 7-cyano-7-deazaguanine reductase QueF	Y	
CKU38_03956	hypothetical protein		Y	
CKU38_04060	putative beta-glucosidase	IPR001764: Glycoside hydrolase, family 3, N-terminal IPR002772: Glycoside hydrolase family 3 C-terminal domain	Y	

		IPR013783: Immunoglobulin-like fold IPR017853: Glycoside hydrolase superfamily IPR019800: Glycoside hydrolase, family 3, active site IPR026891: Fibronectin type III-like domain IPR036881: Glycoside hydrolase family 3 C-terminal domain superfamily IPR036962: Glycoside hydrolase, family 3, N-terminal domain superfamily	
CKU38_04061	putative dihydroneopterin aldolase	IPR006156: Dihydroneopterin aldolase IPR006157: Dihydroneopterin aldolase/epimerase domain	Y
CKU38_04093	catalase	IPR011614: Catalase core domain IPR018028: Catalase, mono-functional, haem-containing IPR020835: Catalase superfamily IPR024168: Catalase, SrpA-type, predicted	Y
CKU38_04176	putative secreted protein, xanthomonadin biosynthesis		Y
CKU38_04205	hypothetical protein		Y
CKU38_04230	putative EvpE, Type VI Secretion System	IPR024168: GpW/Gp25/anti-adapter protein IraD IPR017737: Type VI secretion system, lysozyme-related protein	Y
CKU38_04235	putative AraC family transcriptional regulator	IPR009057: Homeobox-like domain superfamily IPR018060: DNA binding HTH domain, AraC-type	Y
CKU38_04242	putative LysR family transcriptional regulator	IPR000847: Transcription regulator HTH, LysR IPR005119: LysR, substrate-binding IPR036388: Winged helix-like DNA-binding domain superfamily IPR036390: Winged helix DNA-binding domain superfamily	Y
CKU38_04248	Ribosomal protein L28	IPR001383: Ribosomal protein L28 IPR026569: Ribosomal protein L28/L24 IPR034704: L28p-like IPR037147: Ribosomal protein L28/L24 superfamily	Y

CKU38_04258	Xfu2 transposase	IPR008490: Transposase InsH, N-terminal	Y	Y
CKU38_04314	putative cation symporter, GPH family		Y	

2253

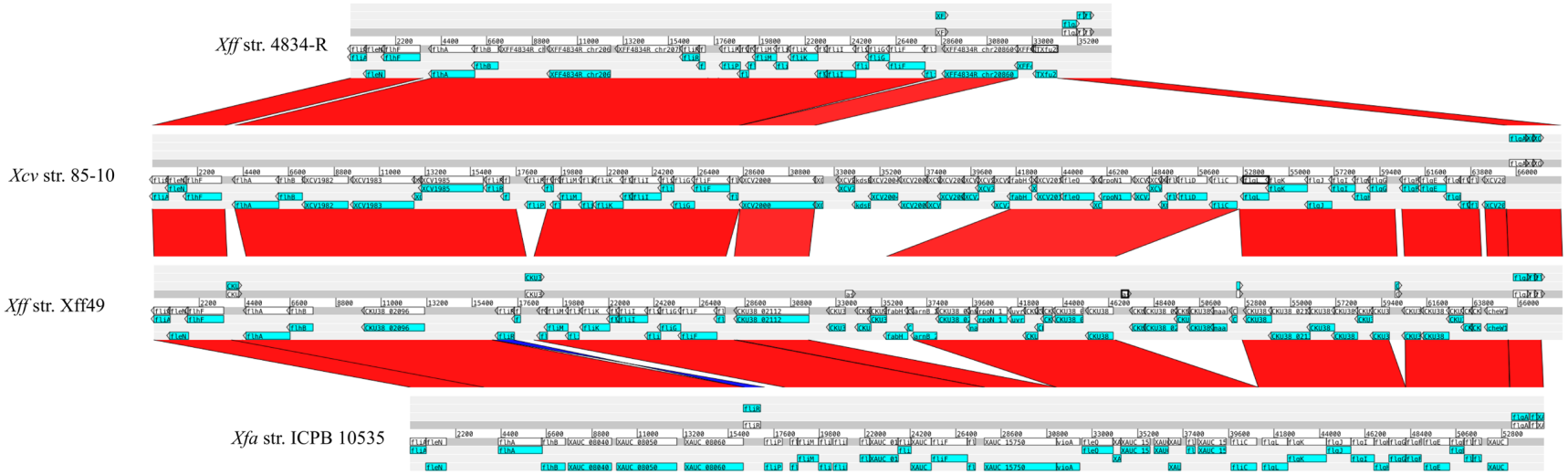
Supplementary Data 8. Presentence of ortholog genes from the flagellar cluster in the genomes of *Xff* str. 4834-R, *Xff* str. Xff49 and *Xfa* str. ICPB10535, based on the flagellum gene cluster identified in the genome of *X. campestris* pv. *vesicatoria* str. 85-10.

Product	Strain			
	85-10	4834-R	ICPB 10535	Xff49
RNA polymerase sigma factor for flagellar operon FliA	XCV1977	XFF4834R_chr20630	XAUC_08000	CKU38_02088
flagellar synthesis regulator FleN	XCV1978	XFF4834R_chr20640	XAUC_08010	CKU38_02089
flagellar GTP-binding protein FlhF	XCV1979	XFF4834R_chr20650	XAUC_08010	CKU38_02090
flagellar biosynthesis pathway protein FlhA	XCV1980	XFF4834R_chr20660	XAUC_08020	CKU38_02092
flagellar biosynthesis pathway protein FlhB	XCV1981	XFF4834R_chr20670	XAUC_08030	CKU38_02093
Putative sensor protein	XCV1982	XFF4834R_chr19240	XAUC_08040	CKU38_01960
Putative sensor protein	XCV1983	XFF4834R_chr20690	XAUC_08050	CKU38_02096
Hypothetical protein	XCV1984	-	-	-
Putative sensor protein	XCV1985	XFF4834R_chr20690	XAUC_08060	CKU38_03226
flagellar biosynthesis pathway protein FliR	XCV1986	XFF4834R_chr20710	XAUC_43400	CKU38_03227
flagellar biosynthesis pathway protein FliQ	XCV1987	XFF4834R_chr20720	-	CKU38_03228
flagellar biosynthesis pathway protein FliP	XCV1988	XFF4834R_chr20730	XAUC_01960	-
flagellar biosynthesis pathway protein FliO	XCV1989	XFF4834R_chr20740	-	CKU38_02101
flagellar motor switch protein FliN	XCV1990	XFF4834R_chr20750	XAUC_01950	CKU38_02102
flagellar motor switch protein FliM	XCV1991	XFF4834R_chr20760	XAUC_01940	CKU38_02103
flagellar basal body-associated protein FliL	XCV1992	XFF4834R_chr20770	XAUC_01930	CKU38_02104
flagellar hook-length control protein FliK	XCV1993	XFF4834R_chr20780	XAUC_01920	CKU38_02105
flagellar biosynthesis chaperone FliJ	XCV1994	XFF4834R_chr20790	XAUC_01910	CKU38_02106
flagellar biosynthesis pathway protein FliI	XCV1995	XFF4834R_chr20800	XAUC_01900	CKU38_02107
flagellar biosynthesis pathway protein FliH	XCV1996	XFF4834R_chr20810	XAUC_01890	CKU38_02108
flagellar motor switch protein FliG	XCV1997	XFF4834R_chr20820	XAUC_01880	CKU38_02109
flagellar biosynthesis lipoprotein FliF	XCV1998	XFF4834R_chr20830	XAUC_01870	CKU38_02110
flagellar hook-basal body complex protein FliE	XCV1999	XFF4834R_chr20840	XAUC_01860	CKU38_02111
putative glycosyltransferase	XCV2000	XFF4834R_chr20860	XAUC_15750	CKU38_02112
Hypothetical protein	XCV2001	-	-	-
Hypothetical protein	XCV2002	-	-	-

3-deoxy-manno-octulosonate cytidylyltransferase	XCV2003	-	-	-
Hypothetical protein	XCV2004	-	-	-
Hypothetical protein	XCV2005	-	-	-
Hypothetical protein	XCV2006	-	-	-
putative Rieske 2Fe-2S family protein	XCV2007	-	-	-
putative acetyltransferase	XCV2008	-	-	-
short chain dehydrogenase	XCV2009	-	-	-
short chain dehydrogenase	XCV2010	-	-	-
3-oxoacyl-[acyl-carrier protein] synthase III	XCV2011	-	-	-
putative acyl carrier protein	XCV2012	-	-	CKU38_02117
putative aminotransferase	XCV2013	-	XAUC_15840	CKU38_02118
flagellar sigma-54 dependent transcriptional activator FleQ	XCV2014	-	XAUC_15850	CKU38_02119
putative two-component response regulator	XCV2015	-	XAUC_15860	CKU38_02120
RNA polymerase sigma-54 factor	XCV2016	-	XAUC_15870	CKU38_02121
two-component system response regulator, LuxR family	XCV2017	-	XAUC_15880	CKU38_02122
Hypothetical protein	XCV2018	-	XAUC_15890	CKU38_02123
Hypothetical protein	XCV2019	-	-	CKU38_02124
flagellin-specific chaperone FliS	XCV2020	-	XAUC_15900	CKU38_02125
Flagellar hook-associated protein FliD	XCV2021	-	XAUC_15910	CKU38_02126
flagellin and related hook-associated protein FliC	XCV2022	-	XAUC_15920	CKU38_02127
flagellin and related hook-associated protein FlgL	XCV2023	-	XAUC_15930	CKU38_02135
flagellin and related hook-associated protein FlgK	XCV2024	-	XAUC_15940	CKU38_02136
muramidase FlgJ	XCV2025	-	XAUC_15950	CKU38_02137
flagellar basal-body P-ring protein FlgI	XCV2026	-	XAUC_15960	CKU38_02138
flagellar basal body L-ring protein FlgH	XCV2027	-	XAUC_15970	CKU38_02139
flagella basal body rod protein FlgG	XCV2028	-	XAUC_15980	CKU38_02140
flagella basal body rod protein FlgF	XCV2029	-	XAUC_06670	CKU38_02142
flagellin and related hook protein FlgE	XCV2030	-	XAUC_06660	CKU38_02143
flagellar hook capping protein FlgD	XCV2031	-	XAUC_06650	CKU38_02144
flagellar basal body rod protein FlgC	XCV2032	-	XAUC_06640	CKU38_02145
flagellar basal body protein FlgB	XCV2033	-	XAUC_06630	CKU38_02146
chemotaxis signal transduction protein	XCV2034	-	XAUC_06620	CKU38_02147

flagellar basal body P-ring biosynthesis FlgA	XCV2035	XFF4834R_chr20900	XAUC_06610	CKU38_02148
Negative regulator of flagellin synthesis protein	XCV2036	XFF4834R_chr20910	XAUC_06600	CKU38_02149
flagellar protein FlgN	XCV2037	XFF4834R_chr20920	XAUC_06590	CKU38_02150

2257



Supplementary Data 9. Synteny analysis of the flagellum gene cluster identified in the genome of *X. fuscans* subsp. *fuscans* str. 4834-R, *X. fuscans* subsp. str. Xff49 and *X. fuscans* subsp. *aurantifolii* str. ICPB 10535, identified based on the genome of *X. campestris* str. 85-10.

5 DISCUSSÃO GERAL E PERSPECTIVAS

5.1 O pangenoma de X. oryzae

A análise do pangenoma da espécie *X. oryzae* permitiu a identificação de um grande número de genes diferencialmente distribuídos entre os patovares *oryzae* e *oryzicola*, tendo sido possível identificar processos biológicos e funções moleculares potencialmente associadas à patogênese que ilustre a diferença em nível molecular destes grupos distintos. Os resultados destas análises podem servir de base para trabalhos futuros direcionados ao desenvolvimento de métodos de diagnóstico capazes de diferenciar estes patovares de forma mais acurada, como o desenvolvido para *X. arboricola* (Cesbron et al., 2015), bem como os genes identificados como diferencialmente distribuídos podem ser posteriormente alvo de estudos que visem uma maior compreensão de seu papel real no processo de patogênese.

Além disso, também foi possível identificar dentre os genes diferencialmente distribuídos, proteínas com função relacionada à degradação da parede celular, o que inclui celulases, lipases e pectinesterase. Estas proteínas são de particular interesse na indústria, sendo as celulases e as pectinesterases relevantes para setores como o de biocombustíveis, onde são usadas na produção de etanol de segunda geração (Jung et al., 2012). Já lipases são amplamente utilizadas nas indústrias alimentícia, farmacêutica, cosmética, têxtil e de detergentes, onde substituem total ou parcialmente métodos físico-químicos de degradação e / ou transformação de lipídeos e outras moléculas orgânicas, como poliésteres (Hasan et al., 2006).

Em trabalhos futuros será possível clonar estas proteínas e avaliar em laboratório a sua atividade, de modo a verificar seu potencial como insumos industriais. Tendo em vista o grande número de genomas utilizadas na análise, é possível incorporar os dados de variabilidade destas proteínas em algoritmos de engenharia de proteínas *in silico* de modo a produzir também sequência otimizadas com maior estabilidade térmica

ou cinética enzimática (Davidson, 2006). Deste modo, novas sequências podem ser derivadas a partir dos grupos ortólogos identificados na análise do pangenoma.

5.2 *TargeTALE*

A ferramenta desenvolvida para análise de genes TALEs também se mostrou eficiente, atendendo ao seu propósito na identificação de genes desta família e de seus alvos, tendo sido possível obter valores de acurácia altos de acordo com as métricas estabelecidas para determinadas configuração nas condição de teste empregadas. Esta ferramenta poderá ser posteriormente estendida para trabalhar com mais *datasets* de promotoremas, poderá auxiliar no processo de caracterização de novas cepas de *Xanthomonas* que venham a ser sequenciadas, bem como em estudos comparativos de efetores e alvos como os já desenvolvidos por Quibod et al (2016).

5.3 O genoma de X. fuscans subsp. fuscans str. Xff49

Por fim, o sequenciamento e a caracterização do genoma da cepa *Xff49* possibilitou uma maior compreensão de seus fatores de virulência e também a identificação de características genéticas que à diferem da cepa de referência para o patovar, 4834-R, incluindo a presença de um cluster completo para genes do flagelo, ausente na cepa de referência, alterações estruturais, mutações em genes responsáveis por processos biológicos e a identificação de um novo plasmídeo previamente não reportado no gênero, e que apresenta características de conjugação, denominado *plX.*

6 CONCLUSÃO GERAL

Através das abordagens empregadas e desenvolvidas, foi possível identificar genes em bactérias do gênero *Xanthomonas* que codificam para fatores de virulência nos patovares de *X. oryzae* e em uma cepa local de *X. fuscans*. Por fim, também foi possível descrever e validar uma nova ferramenta para o estudo da patogênese de *Xanthomonas* por meio da análise de genes da família a TALE. Além disso, também foi possível identificar genes com potencial biotecnológico.

7 REFERÊNCIAS

AGREN, J., SUNDSTRÖM, A., HÅFSTRÖM, T. & SEGERMAN, B. (2012). Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PloS one*. 7(6). p. E39107. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0039107 [Accessed May 8, 2015].

AKHAVAN, A., BAHAR, M., ASKARIAN, H., LAK, M.R., NAZEMI, A. & ZAMANI, Z. (2013). Bean common bacterial blight: pathogen epiphytic life and effect of irrigation practices. *SpringerPlus*. 2(1). p. 41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23539532 [Accessed September 1, 2017].

ALKAN, C., SAJJADIAN, S. & EICHLER, E.E. (2010). Limitations of next-generation genome sequence assembly. *Nature Methods*. 8(1). p. 61–65. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3115693&tool=pmcentrez&rendertype=abstract [Accessed October 1, 2015].

ALKORTA, I., GARBISU, C., LLAMA, M.J. & SERRA, J.L. (1998). Industrial applications of pectic enzymes: a review. *Process Biochemistry*. 33(1). p. 21–28. Available at: https://www.sciencedirect.com/science/article/pii/S0032959297000460 [Accessed January 24, 2019].

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. & LIPMAN, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*. 215(3). p. 403–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2231712 [Accessed April 28, 2014].

ALVAREZ, A.M. (2000). Black Rot of Crucifers. In *Mechanisms of Resistance to Plant Diseases*. Dordrecht: Springer Netherlands, p. 21–52. Available at: http://link.springer.com/10.1007/978-94-011-3937-3_2 [Accessed January 13, 2019].

ANGIUOLI, S. V & SALZBERG, S.L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics (Oxford, England)*. 27(3). p. 334–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21148543 [Accessed May 25, 2017].

APWEILER, R., BAIROCH, A., WU, C.H., BARKER, W.C., BOECKMANN, S.F., GASTEIGER, E., HUANG, H., LOPEZ, R., MAGRANE, M., MARTIN, M.J., NATALE, D.A., O'DONOVAN, C., REDASCHI, N. & YEH, L.L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*. 32(90001). p. D115-119.

BAE, C., OH, E.-J., LEE, H.-B., KIM, B.-Y. & OH, C.-S. (2015). Complete genome sequence of the cellulase-producing bacterium Clavibacter michiganensis PF008. *Journal of Biotechnology*. 214. p. 103–104. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26410454 [Accessed January 24, 2019].

BENSON, D.A., KARSCH-MIZRACHI, I., LIPMAN, D.J., OSTELL, J. & WHEELER, D.L. (2005). GenBank. *Nucleic acids research*. 33(Database issue). p. D34-8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=540017&tool=pmcentrez&rendertype=abstract [Accessed April 28, 2015].

BINNS, D., DIMMER, E., HUNTLEY, R., BARRELL, D., O'DONOVAN, C. & APWEILER, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics (Oxford, England)*. 25(22). p. 3045–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19744993 [Accessed July 19, 2017].

BOCH, J., SCHOLZE, H., SCHORNACK, S., LANDGRAF, A., HAHN, S., KAY, S., LAHAYE, T., NICKSTADT, A. & BONAS, U. (2009). Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science*. 326(5959). p. 1509–1512. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19933107 [Accessed February 27, 2018].

BONAS, U., VAN DEN ACKERVEKEN, G., BUTTNER, D., HAHN, K., MAROIS, E., NENNSTIEL, D., NOEL, L., ROSSIER, O. & SZUREK, B. (2000). How the bacterial plant pathogen Xanthomonas campestris pv. vesicatoria conquers the host. *Molecular Plant Pathology*. 1(1). p. 73–76. Available at: http://doi.wiley.com/10.1046/j.1364-3703.2000.00010.x [Accessed January 13, 2019].

BRANSCOMB, E. & PREDKI, P. (2002). On the high value of low standards. *Journal of bacteriology*. 184(23). p. 6406–9; DISCUSSION 6409. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=135412&tool=pmcentrez&rendertype=abstract [Accessed December 29, 2014].

BRUNINGS, A.M. & GABRIEL, D.W. (2003). Xanthomonas citri: breaking the surface. *Molecular Plant Pathology*. 4(3). p. 141–157. Available at: http://doi.wiley.com/10.1046/j.1364-3703.2003.00163.x [Accessed January 13, 2019].

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T.L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*. 10(1). p. 421. Available at: http://www.biomedcentral.com/1471-2105/10/421 [Accessed July 9, 2014].

CESBRON, S., BRIAND, M., ESSAKHI, S., GIRONDE, S., BOUREAU, T., MANCEAU, C., FISCHER-LE SAUX, M. & JACQUES, M.-A. (2015). Comparative Genomics of Pathogenic and Nonpathogenic Strains of Xanthomonas arboricola Unveil Molecular and Evolutionary Events Linked to Pathoadaptation. *Frontiers in Plant Science*. 6. p. 1126. Available at: http://journal.frontiersin.org/Article/10.3389/fpls.2015.01126/abstract [Accessed January 13, 2019].

CINGOLANI, P., PLATTS, A., WANG, L.L., COON, M., NGUYEN, T., WANG, L., LAND, S.J., LU, X. & RUDEN, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly*. 6(2). p. 80–92. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3679285&tool=pmcentrez&rendertype=abstract [Accessed December 22, 2014].

COLLMER, A., BAUER, D.W., HE, S.Y., LINDEBERG, M., KELEMU, S., RODRIGUEZ-PALENZUELA, P., BURR, T.J. & CHATTERJEE, A.K. (1991). Pectic Enzyme Production and Bacterial Plant Pathogenicity. In Springer, Dordrecht, p. 65–72. Available at: http://link.springer.com/10.1007/978-94-015-7934-6_10 [Accessed January 24, 2019].

CONESA, A. & GÖTZ, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*. 2008. p. 619832.

CONSTANTIN, E.C., CLEENWERCK, I., MAES, M., BAEYEN, S., VAN MALDERGHEM, C., DE VOS, P. & COTTYN, B. (2016). Genetic characterization of strains named as *Xanthomonas axonopodis* pv. *dieffenbachiae* leads to a taxonomic revision of the *X. axonopodis* species complex. *Plant Pathology*. 65(5). p. 792–806. Available at: http://doi.wiley.com/10.1111/ppa.12461 [Accessed January 12, 2019].

DARLING, A.C.E., MAU, B., BLATTNER, F.R. & PERNA, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*. 14(7). p. 1394–403. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=442156&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

DAVIDSON, A.R. (2006). Multiple Sequence Alignment as a Guideline for Protein Engineering Strategies. In *Protein Design*. New Jersey: Humana Press, p. 171–182. Available at: http://link.springer.com/10.1385/1-59745-116-9:171 [Accessed January 20, 2019].

DOYLE, E.L., BOOHER, N.J., STANDAGE, D.S., VOYTAS, D.F., BRENDEL, V.P., VANDYK, J.K. & BOGDANOVE, A.J. (2012). TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Research*. 40(W1). p. W117–W122. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22693217 [Accessed December 9, 2018].

EDWARDS, D.J. & HOLT, K.E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial informatics and experimentation*. 3(1). p. 2. Available at: http://www.microbialinformaticsj.com/content/3/1/2 [Accessed January 22, 2015].

FERENCE, C.M., GOCHEZ, A.M., BEHLAU, F., WANG, N., GRAHAM, J.H. & JONES, J.B. (2018). Recent advances in the understanding of *Xanthomonas citri* ssp. *citri* pathogenesis and citrus canker disease management. *Molecular Plant Pathology*. 19(6). p. 1302–1318. Available at: http://doi.wiley.com/10.1111/mpp.12638 [Accessed January 13, 2019].

FOUTS, D.E., MATTHIAS, M.A., ADHIKARLA, H., ADLER, B., AMORIM-SANTOS, L., BERG, D.E., BULACH, D., BUSCHIAZZO, A., CHANG, Y.-F., GALLOWAY, R.L., HAAKE, D.A., HAFT, D.H.,

HARTSKEERL, R., KO, A.I., LEVETT, P.N., MATSUNAGA, J., MECHALY, A.E., MONK, J.M., NASCIMENTO, A.L.T., ET AL. (2016). What Makes a Bacterial Species Pathogenic?:Comparative Genomic Analysis of the Genus Leptospira. *PLoS neglected tropical diseases*. 10(2). p. E0004403. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26890609 [Accessed August 30, 2016].

FRASER, C.M., EISEN, J.A., NELSON, K.E., PAULSEN, I.T. & SALZBERG, S.L. (2002). The Value of Complete Microbial Genome Sequencing (You Get What You Pay For). *Journal of Bacteriology*. 184(23). p. 6403–6405. Available at: http://jb.asm.org/content/184/23/6403.full [Accessed December 29, 2014].

GAJ, T., GERSBACH, C.A. & BARBAS, C.F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*. 31(7). p. 397–405. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23664777 [Accessed January 6, 2019].

GARITA-CAMBRONERO, J., PALACIO-BIELSA, A., LÓPEZ, M.M. & CUBERO, J. (2017). Pan-Genomic Analysis Permits Differentiation of Virulent and Non-virulent Strains of Xanthomonas arboricola That Cohabit Prunus spp. and Elucidate Bacterial Virulence Factors. *Frontiers in microbiology*. 8. p. 573. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28450852 [Accessed January 13, 2019].

GLASS, K. & GIRVAN, M. (2014). Annotation enrichment analysis: an alternative method for evaluating the functional properties of gene sets. *Scientific reports*. 4. p. 4191. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24569707 [Accessed January 13, 2019].

GORDON, J.L., LEFEUVRE, P., ESCALON, A., BARBE, V., CRUVEILLER, S., GAGNEVIN, L., PRUVOST, O., SWINGS, J., CIVEROLO, E., LEYNS, F., CLEENE, M., SWINGS, J.-G., LEY, J., BRUNINGS, A., GABRIEL, D., CIVEROLO, E., SUN, X., STALL, R., JONES, J., ET AL. (2015). Comparative genomics of 43 strains of Xanthomonas citri pv. citri reveals the evolutionary events giving rise to pathotypes with different host ranges. *BMC Genomics*. 16(1). p. 1098. Available at: http://www.biomedcentral.com/1471-2164/16/1098 [Accessed June 27, 2016].

GOTOR-FERNÁNDEZ, V., BRIEVA, R. & GOTOR, V. (2006). Lipases: Useful biocatalysts for the preparation of pharmaceuticals. *Journal of Molecular Catalysis B: Enzymatic*. 40(3–4). p. 111–120. Available at: https://www.sciencedirect.com/science/article/pii/S1381117706000750 [Accessed January 24, 2019].

GOTTWALD, T.R., GRAHAM, J.H. & SCHUBERT, T.S. (2002). Citrus Canker: The Pathogen and Its Impact. *Plant Health Progress*. 3(1). p. 15. Available at: https://apsjournals.apsnet.org/doi/10.1094/PHP-2002-0812-01-RV [Accessed January 13, 2019].

GRAHAM, P.H. & VANCE, C.P. (2003). Legumes: Importance and Constraints to Greater Use. *PLANT PHYSIOLOGY*. 131(3). p. 872–877. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/12644639 [Accessed September 1, 2017].

GRAU, J., RESCHKE, M., ERKES, A., STREUBEL, J., MORGAN, R.D., WILSON, G.G., KOEBNIK, R. & BOCH, J. (2016). AnnoTALE: bioinformatics tools for identification, annotation, and nomenclature of TALEs from Xanthomonas genomic sequences. *Scientific reports*. 6. p. 21077. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26876161 [Accessed June 28, 2016].

GRAU, J., WOLF, A., RESCHKE, M., BONAS, U., POSCH, S. & BOCH, J. (2013). Computational Predictions Provide Insights into the Biology of TAL Effector Target Sites A. Tresch (ed.). *PLoS Computational Biology*. 9(3). p. E1002962. Available at: http://dx.plos.org/10.1371/journal.pcbi.1002962 [Accessed September 27, 2017].

HARRIS, M.A., CLARK, J., IRELAND, A., LOMAX, J., ASHBURNER, M., FOULGER, R., EILBECK, K., LEWIS, S., MARSHALL, B., MUNGALL, C., RICHTER, J., RUBIN, G.M., BLAKE, J.A., BULT, C., DOLAN, M., DRABKIN, H., EPPIG, J.T., HILL, D.P., NI, L., ET AL. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*. 32(90001). p. 258D–261. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308770&tool=pmcentrez&rendertype=abstract [Accessed October 14, 2015].

HASAN, F., SHAH, A.A. & HAMEED, A. (2006). Industrial applications of microbial lipases. *Enzyme and Microbial Technology*. 39(2). p. 235–251. Available at: https://www.sciencedirect.com/science/article/pii/S0141022905004606 [Accessed January 20, 2019].

HE, Z., PROUDFOOT, C., WHITELAW, C.B.A. & LILLICO, S.G. (2016). Comparison of CRISPR/Cas9 and TALENs on editing an integrated EGFP gene in the genome of HEK293FT cells. *SpringerPlus*. 5(1). p. 814. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27390654 [Accessed January 6, 2019].

HEATH, M.C. (2000). Hypersensitive response-related death. *Plant Molecular Biology*. 44(3). p. 321–334. Available at: http://link.springer.com/10.1023/A:1026592509060 [Accessed January 13, 2019].

HOFFMANN, S., OTTO, C., KURTZ, S., SHARMA, C.M., KHAITOVICH, P., VOGEL, J., STADLER, P.F. & HACKERMÜLLER, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS computational biology*. 5(9). p. E1000502. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19750212 [Accessed September 19, 2016].

JACQUES, M.-A., JOSI, K., DARRASSE, A. & SAMSON, R. (2005). Xanthomonas axonopodis pv. phaseoli var. fuscans Is Aggregated in Stable Biofilm Population Sizes in the Phyllosphere of Field-Grown Beans. *Applied and Environmental Microbiology*. 71(4). p. 2008–2015. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15812033 [Accessed September 1, 2017].

JOLLEY, K.A. & MAIDEN, M.C.J. (2010). BIGSdb: Scalable analysis of bacterial genome

variation at the population level. *BMC bioinformatics*. 11(1). p. 595. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-595 [Accessed June 8, 2015].

JUNG, S.-K., PARISUTHAM, V., JEONG, S.H. & LEE, S.K. (2012). Heterologous expression of plant cell wall degrading enzymes for effective production of cellulosic biofuels. *Journal of biomedicine & biotechnology*. 2012. p. 405842. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22911272 [Accessed January 20, 2019].

KANEHISA, M. & GOTO, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*. 28(1). p. 27–30. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

KANEHISA, M., SATO, Y. & MORISHIMA, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*. 428(4). p. 726–731. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26585406 [Accessed July 19, 2017].

KASIF, S. & STEFFEN, M. (2010). Biochemical networks: the evolution of gene annotation. *Nature chemical biology*. 6(1). p. 4–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20016491 [Accessed January 13, 2019].

KISAND, V. & LETTIERI, T. (2013). Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. *BMC genomics*. 14. p. 211. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3618134&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2014].

KLEMM, E. & DOUGAN, G. (2016). Advances in Understanding Bacterial Pathogenesis Gained from Whole-Genome Sequencing and Phylogenetics. *Cell Host & Microbe*. 19(5). p. 599–610. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27173928 [Accessed January 13, 2019].

KREMER, F.S., ESLABÃO, M.R., DELLAGOSTIN, O.A. & PINTO, L. DA S. (2016). Genix: a new online automated pipeline for bacterial genome annotation A. Millard (ed.). *FEMS Microbiology Letters*. 363(23). p. FNW263. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27856568 [Accessed July 24, 2017].

KREMER, F.S., MCBRIDE, A.J.A. & PINTO, L.S. (2017). Approaches for in silico finishing of microbial genome sequences. *Genetics and Molecular Biology*. 40(3).

KURTZ, S., PHILLIPPY, A., DELCHER, A.L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S.L. (2004). Versatile and open software for comparing large genomes. *Genome biology*. 5(2). p. R12. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=395750&tool=pmcentrez&rendertype=abstract [Accessed May 9, 2014].

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*. 10(3). p. R25. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2690996&tool=pmcentrez&rendertype=abstract [Accessed April 28, 2014].

LEDUC, A., TRAORÉ, Y.N., BOYER, K., MAGNE, M., GRYGIEL, P., JUHASZ, C.C., BOYER, C., GUERIN, F., WONNI, I., OUEDRAOGO, L., VERNIÈRE, C., RAVIGNÉ, V. & PRUVOST, O. (2015). Bridgehead invasion of a monomorphic plant pathogenic bacterium: *X anthomonas citri* pv. *citri* , an emerging citrus pathogen in Mali and Burkina Faso. *Environmental Microbiology*. 17(11). p. 4429–4442. Available at: http://doi.wiley.com/10.1111/1462-2920.12876 [Accessed January 13, 2019].

LEE, B.-M., PARK, Y.-J., PARK, D.-S., KANG, H.-W., KIM, J.-G., SONG, E.-S., PARK, I.-C., YOON, U.-H., HAHN, J.-H., KOO, B.-S., LEE, G.-B., KIM, H., PARK, H.-S., YOON, K.-O., KIM, J.-H., JUNG, C., KOH, N.-H., SEO, J.-S. & GO, S.-J. (2005). The genome sequence of Xanthomonas oryzae pathovar oryzae KACC10331, the bacterial blight pathogen of rice. *Nucleic Acids Research*. 33(2). p. 577–586. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15673718 [Accessed July 19, 2017].

LEES, J.A., KENDALL, M., PARKHILL, J., COLIJN, C., BENTLEY, S.D. & HARRIS, S.R. (2018). Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome open research*. 3. p. 33. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29774245 [Accessed December 23, 2018].

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*. 25(16). p. 2078–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract [Accessed April 28, 2014].

LI, H. & DURBIN, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*. 25(14). p. 1754–60. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract [Accessed April 29, 2014].

LIU, L., LI, Y., LI, S., HU, N., HE, Y., PONG, R., LIN, D., LU, L. & LAW, M. (2012). Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*. 2012. p. 251364. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3398667&tool=pmcentrez&rendertype=abstract [Accessed July 11, 2014].

LIU, X. & KOKARE, C. (2017). Microbial Enzymes of Use in Industry. *Biotechnology of Microbial Enzymes*. p. 267–298. Available at: https://www.sciencedirect.com/science/article/pii/B978012803725600011X [Accessed January 24, 2019].

LYKIDIS, A., MAVROMATIS, K., IVANOVA, N., ANDERSON, I., LAND, M., DIBARTOLO, G.,

MARTINEZ, M., LAPIDUS, A., LUCAS, S., COPELAND, A., RICHARDSON, P., WILSON, D.B. & KYRPIDES, N. (2007). Genome Sequence and Analysis of the Soil Cellulolytic Actinomycete Thermobifida fusca YX. *Journal of Bacteriology*. 189(6). p. 2477–2486. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17209016 [Accessed January 24, 2019].

MAIDEN, M.C.J., JANSEN VAN RENSBURG, M.J., BRAY, J.E., EARLE, S.G., FORD, S.A., JOLLEY, K.A. & MCCARTHY, N.D. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature reviews. Microbiology*. 11(10). p. 728–36. Available at: http://dx.doi.org/10.1038/nrmicro3093 [Accessed February 22, 2016].

MARDIS, E., MCPHERSON, J., MARTIENSSEN, R., WILSON, R.K. & MCCOMBIE, W.R. (2002). What is finished, and why does it matter. *Genome research*. 12(5). p. 669–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11997333 [Accessed December 15, 2015].

MARDIS, E.R. (2008). Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*. 9. p. 387–402. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18576944 [Accessed April 28, 2014].

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*. 20(9). p. 1297–303. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928508&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

MEDIE, F.M., DAVIES, G.J., DRANCOURT, M. & HENRISSAT, B. (2012). Genome analyses highlight the different biological roles of cellulases. *Nature Reviews Microbiology*. 10(3). p. 227–234. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22266780 [Accessed January 24, 2019].

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. & RAPPUOLI, R. (2005). The microbial pan-genome. *Current opinion in genetics & development*. 15(6). p. 589–94. Available at: http://www.sciencedirect.com/science/article/pii/S0959437X05001759 [Accessed March 13, 2015].

MENDES, F.B., IBRAIM PIRES ATALA, D. & THOMÉO, J.C. (2017). Is cellulase production by solid-state fermentation economically attractive for the second generation ethanol production? *Renewable Energy*. 114. p. 525–533. Available at: https://www.sciencedirect.com/science/article/pii/S0960148117306845 [Accessed January 13, 2019].

MEW, T.W. (1993). Focus on Bacterial Blight of Rice. *Plant Disease*. 77(1). p. 5. Available at: http://www.apsnet.org/publications/PlantDisease/BackIssues/Documents/1993Abstracts/PD_77_5.htm [Accessed November 12, 2017].

MO, Q., LIU, A., GUO, H., ZHANG, Y. & LI, M. (2016). A novel thermostable and organic solvent-tolerant lipase from Xanthomonas oryzae pv. oryzae YB103: screening, purification and characterization. *Extremophiles*. 20(2). p. 157–165. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26791383 [Accessed December 27, 2018].

MORIYA, Y., ITOH, M., OKUDA, S., YOSHIZAWA, A.C. & KANEHISA, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research*. 35(Web Server issue). p. W182-5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1933193&tool=pmcentrez&rendertype=abstract [Accessed February 25, 2015].

MORRIS, C.E. & MONIER, J.-M. (2003). The ecological significance of biofilm formation by plant-associated bacteria. *Annual Review of Phytopathology*. 41(1). p. 429–453. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12730399 [Accessed September 1, 2017].

MOSCOU, M.J. & BOGDANOVE, A.J. (2009). A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science*. 326(5959). p. 1501–1501. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19933106 [Accessed December 8, 2018].

MUÑOZ BODNAR, A., BERNAL, A., SZUREK, B. & LÓPEZ, C.E. (2013). Tell Me a Tale of TALEs. *Molecular Biotechnology*. 53(2). p. 228–235. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23114874 [Accessed November 7, 2017].

NIÑO-LIU, D.O., RONALD, P.C. & BOGDANOVE, A.J. (2006). Xanthomonas oryzae pathovars: model pathogens of a model crop. *Molecular plant pathology*. 7(5). p. 303–24. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20507449 [Accessed April 11, 2016].

NODA, T. & HISATOSHI, K. (1999). Growith of Xanthomonas oryzae pv. oryzae in planta and in guttation fluid of rice. *Annals of the Phytopathological Society of Japan*. 65. p. 9–14.

NYVALL, R. (1999). *Field Crop Diseases* Third edit ed., Ames: iowa State University Press.

OU, S.H. (1985). *Rice diseases*, Commonwealth Mycological Institute. Available at: https://books.google.com.br/books?id=-k3mewv9nMoC&pg=PA92&lpg=PA92&dq=disease+forecasting+for+Xanthomonas+oryzae&source=bl&ots=Zji_vyx99j&sig=oCGcmWcrhP2EBsZwRPb3Add9s48&hl=en&ei=DNSkTqLbBOfj0QGktdmCBQ&sa=X&oi=book_result&ct=result&redir_esc=y#v=onepage&q [Accessed November 12, 2017].

PAREJA-TOBES, P., MANRIQUE, M., PAREJA-TOBES, E., PAREJA, E. & TOBES, R. (2012). BG7: a new approach for bacterial genome annotation designed for next generation sequencing data. *PloS one*. 7(11). p. E49239. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0049239 [Accessed December 24, 2015].

PEDROLLI, D.B., MONTEIRO, A.C., GOMES, E. & CARMONA, E.C. (2009). Pectin and Pectinases: Production, Characterization and Industrial Application of Microbial Pectinolytic Enzymes. *The Open Biotechnology Journal*. 3(1). p. 9–18. Available at: http://benthamopen.com/ABSTRACT/TOBIOTJ-3-9 [Accessed January 24, 2019].

PÉREZ-QUINTERO, A.L., RODRIGUEZ-R, L.M., DEREEPER, A., LÓPEZ, C., KOEBNIK, R., SZUREK, B. & CUNNAC, S. (2013). An Improved Method for TAL Effectors DNA-Binding Sites Prediction Reveals Functional Convergence in TAL Repertoires of Xanthomonas oryzae Strains D. Arnold (ed.). *PLoS ONE*. 8(7). p. E68464. Available at: http://dx.plos.org/10.1371/journal.pone.0068464 [Accessed December 9, 2018].

PÉREZ-QUINTERO, A.L., LAMY, L., ZARATE, C.A., CUNNAC, S., DOYLE, E., BOGDANOVE, A., SZUREK, B. & DEREEPER, A. (2018). daTALbase: A Database for Genomic and Transcriptomic Data Related to TAL Effectors. *Molecular Plant-Microbe Interactions*. 31(4). p. 471–480. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29143556 [Accessed December 9, 2018].

QUIBOD, I.L., PEREZ-QUINTERO, A., BOOHER, N.J., DOSSA, G.S., GRANDE, G., SZUREK, B., VERA CRUZ, C., BOGDANOVE, A.J. & OLIVA, R. (2016). Effector Diversification Contributes to Xanthomonas oryzae pv. oryzae Phenotypic Adaptation in a Semi-Isolated Environment. *Scientific reports*. 6. p. 34137. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27667260 [Accessed November 7, 2017].

RAN, F.A., HSU, P.D., WRIGHT, J., AGARWALA, V., SCOTT, D.A. & ZHANG, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature protocols*. 8(11). p. 2281–308. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24157548 [Accessed June 28, 2016].

RASKO, D.A., ROSOVITZ, M.J., MYERS, G.S.A., MONGODIN, E.F., FRICKE, W.F., GAJER, P., CRABTREE, J., SEBAIHIA, M., THOMSON, N.R., CHAUDHURI, R., HENDERSON, I.R., SPERANDIO, V. & RAVEL, J. (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of bacteriology*. 190(20). p. 6881–93. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2566221&tool=pmcentrez&rendertype=abstract [Accessed May 5, 2015].

RIBEIRO, B.D., DE CASTRO, A.M., COELHO, M.A.Z. & FREIRE, D.M.G. (2011). Production and use of lipases in bioenergy: a review from the feedstocks to biodiesel production. *Enzyme research*. 2011. p. 615803. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21785707 [Accessed January 24, 2019].

RISSMAN, A.I., MAU, B., BIEHL, B.S., DARLING, A.E., GLASNER, J.D. & PERNA, N.T. (2009). Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics (Oxford, England)*. 25(16). p. 2071–3. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723005&tool=pmcentrez&rendertype=abstract [Accessed December 11, 2014].

ROSSETO, F.R., MANZINE, L.R., DE OLIVEIRA NETO, M. & POLIKARPOV, I. (2016). Biophysical

and biochemical studies of a major endoglucanase secreted by Xanthomonas campestris pv. campestris. *Enzyme and Microbial Technology*. 91. p. 1–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27444323 [Accessed January 13, 2019].

ROULI, L., MERHEJ, V., FOURNIER, P.-E. & RAOULT, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New microbes and new infections*. 7. p. 72–85. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26442149 [Accessed April 1, 2018].

RUH, M., BRIAND, M., BONNEAU, S., JACQUES, M.-A. & CHEN, N.W.G. (2017). Xanthomonas adaptation to common bean is associated with horizontal transfers of genes encoding TAL effectors. *BMC Genomics*. 18(1). p. 670. Available at: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4087-6 [Accessed September 1, 2017].

SALLET, E., GOUZY, J. & SCHIEX, T. (2014). EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics (Oxford, England)*. 30(18). p. 2659–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24880686 [Accessed May 15, 2015].

SCHÜRCH, A.C., ARREDONDO-ALONSO, S., WILLEMS, R.J.L. & GOERING, R.V. (2018). Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clinical Microbiology and Infection*. 24(4). p. 350–354. Available at: https://www.sciencedirect.com/science/article/pii/S1198743X17307103 [Accessed January 13, 2019].

SEEMANN, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*. 30(14). p. 2068–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24642063 [Accessed July 10, 2014].

SOLÉ, M., SCHEIBNER, F., HOFFMEISTER, A.-K., HARTMANN, N., HAUSE, G., ROTHER, A., JORDAN, M., LAUTIER, M., ARLAT, M. & BÜTTNER, D. (2015). Xanthomonas campestris pv. vesicatoria Secretes Proteases and Xylanases via the Xps Type II Secretion System and Outer Membrane Vesicles. *Journal of bacteriology*. 197(17). p. 2879–93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26124239 [Accessed January 13, 2019].

STICKLEN, M.B. (2008). Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nature Reviews Genetics*. 9(6). p. 433–443. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18487988 [Accessed May 24, 2017].

TAYI, L., KUMAR, S., NATHAWAT, R., HAQUE, A.S., MAKU, R. V., PATEL, H.K., SANKARANARAYANAN, R. & SONTI, R. V. (2018). A mutation in an exoglucanase of *Xanthomonas oryzae* pv. *oryzae* , which confers an endo mode of activity, affects bacterial virulence, but not the induction of immune responses, in rice. *Molecular Plant Pathology*. 19(6). p. 1364–1376. Available at: http://doi.wiley.com/10.1111/mpp.12620 [Accessed December 26, 2018].

TAYI, L., MAKU, R. V., PATEL, H.K. & SONTI, R. V. (2016a). Identification of Pectin Degrading Enzymes Secreted by Xanthomonas oryzae pv. oryzae and Determination of Their Role in Virulence on Rice R. A. Wilson (ed.). *PLOS ONE*. 11(12). p. E0166396. Available at: https://dx.plos.org/10.1371/journal.pone.0166396 [Accessed December 27, 2018].

TAYI, L., MAKU, R. V., PATEL, H.K. & SONTI, R. V. (2016b). Identification of Pectin Degrading Enzymes Secreted by Xanthomonas oryzae pv. oryzae and Determination of Their Role in Virulence on Rice R. A. Wilson (ed.). *PLOS ONE*. 11(12). p. E0166396. Available at: https://dx.plos.org/10.1371/journal.pone.0166396 [Accessed January 24, 2019].

TEMUUJIN, U., KIM, J.-W., KIM, J.-K., LEE, B.-M. & KANG, H.-W. (2011). Identification of novel pathogenicity-related cellulase genes in Xanthomonas oryzae pv. oryzae. *Physiological and Molecular Plant Pathology*. 76(2). p. 152–157. Available at: https://www.sciencedirect.com/science/article/pii/S0885576511000865 [Accessed January 13, 2019].

THIEME, F., KOEBNIK, R., BEKEL, T., BERGER, C., BOCH, J., BUTTNER, D., CALDANA, C., GAIGALAT, L., GOESMANN, A., KAY, S., KIRCHNER, O., LANZ, C., LINKE, B., MCHARDY, A.C., MEYER, F., MITTENHUBER, G., NIES, D.H., NIESBACH-KLOSGEN, U., PATSCHKOWSKI, T., ET AL. (2005). Insights into Genome Plasticity and Pathogenicity of the Plant Pathogenic Bacterium Xanthomonas campestris pv. vesicatoria Revealed by the Complete Genome Sequence. *Journal of Bacteriology*. 187(21). p. 7254–7266. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16237009 [Accessed September 1, 2017].

TREANGEN, T.J., ONDOV, B.D., KOREN, S. & PHILLIPPY, A.M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome biology*. 15(11). p. 524. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25410596 [Accessed December 21, 2018].

TRINGE, S.G. & HUGENHOLTZ, P. (2008). A renaissance for the pioneering 16S rRNA gene. *Current opinion in microbiology*. 11(5). p. 442–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18817891 [Accessed November 24, 2015].

TRIPLETT, L.R., HAMILTON, J.P., BUELL, C.R., TISSERAT, N.A., VERDIER, V., ZINK, F. & LEACH, J.E. (2011). Genomic Analysis of Xanthomonas oryzae Isolates from Rice Grown in the United States Reveals Substantial Divergence from Known X. oryzae Pathovars. *Applied and Environmental Microbiology*. 77(12). p. 3930–3937. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21515727 [Accessed December 25, 2018].

UENOJO, M. & PASTORE, G.M. (2007). Pectinases: aplicações industriais e perspectivas. *Química Nova*. 30(2). p. 388–394. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40422007000200028&lng=pt&nrm=iso&tlng=pt [Accessed January 24, 2019].

VAUTERIN, L., HOSTE, B., KERSTERS, K. & SWINGS, J. (1995). Reclassification of

Xanthomonas. *International Journal of Systematic Bacteriology*. 45(3). p. 472–489. Available at: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-45-3-472 [Accessed June 28, 2016].

VAUTERIN, L., RADEMAKER, J. & SWINGS, J. (2000). Synopsis on the taxonomy of the genus xanthomonas. *Phytopathology*. 90(7). p. 677–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18944485 [Accessed June 28, 2016].

WANG, L., RONG, W. & HE, C. (2008). Two *Xanthomonas* Extracellular Polygalacturonases, PghAxc and PghBxc, Are Regulated by Type III Secretion Regulators HrpX and HrpG and Are Required for Virulence. *Molecular Plant-Microbe Interactions*. 21(5). p. 555–563. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18393615 [Accessed January 24, 2019].

WEBER, E. & KOEBNIK, R. (2005). Domain structure of HrpE, the Hrp pilus subunit of Xanthomonas campestris pv. vesicatoria. *Journal of bacteriology*. 187(17). p. 6175–86. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16109959 [Accessed January 13, 2019].

WEISBURG, W.G., BARNS, S.M., PELLETIER, D.A. & LANE, D.J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.*. 173(2). p. 697–703. Available at: http://jb.asm.org/content/173/2/697.abstract [Accessed February 23, 2016].

WILLIAMS, P.H. (1980). Black Rot: A Continuing threat to world crucifers. *Plant Disease*. 64(8). p. 736. Available at: https://doi.org/10.1094%2Fpd-64-736.

WILSON, D.B. (2009). Cellulases and biofuels. *Current Opinion in Biotechnology*. 20(3). p. 295–299. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19502046 [Accessed January 24, 2019].

WOO, P.C.Y., LAU, S.K.P., TENG, J.L.L., TSE, H. & YUEN, K.-Y. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*. 14(10). p. 908–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18828852 [Accessed August 7, 2015].

XIA, T., LI, Y., SUN, D., ZHUO, T., FAN, X. & ZOU, H. (2016). Identification of an Extracellular Endoglucanase That Is Required for Full Virulence in Xanthomonas citri subsp. citri Z. Wang (ed.). *PLOS ONE*. 11(3). p. E0151017. Available at: https://dx.plos.org/10.1371/journal.pone.0151017 [Accessed December 26, 2018].

YE, Y., WEI, B., WEN, L. & RAYNER, S. (2013). BlastGraph: a comparative genomics tool based on BLAST and graph algorithms. *Bioinformatics (Oxford, England)*. 29(24). p. 3222–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24068035 [Accessed July 23, 2016].

ZHANG, M., WANG, F., LI, S., WANG, Y., BAI, Y. & XU, X. (2014). TALE: A tale of genome editing. *Progress in Biophysics and Molecular Biology*. 114(1). p. 25–32.

ANEXOS

Anexo A – Certificado de Registro de Programa de Computador da ferramenta TargeTALE

Órgão Responsável: Instituto Nacional de Propriedade Industrial (INPI)

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA INDÚSTRIA, COMÉRCIO EXTERIOR E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512018051653-0**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 29/12/2017, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: TargeTALE

Data de criação: 29/12/2017

Titular(es): UNIVERSIDADE FEDERAL DE PELOTAS

Autor(es): FREDERICO SCHMITT KREMER; LUCIANO DA SILVA PINTO; AMANDA MUNARI GUIMARÃES; CHRISTIAN DOMINGUES SANCHEZ

Linguagem: HTML; SQL; PYTHON; PHP

Campo de aplicação: BL-02; BL-04

Tipo de programa: AP-01

Algoritmo hash: OUTROS

Resumo digital hash: f0df866fa4e974b616b252396b905bd7

Expedido em: 25/09/2018

Aprovado por:
Liane Elizabeth Caldeira Lage
Diretora de Patentes, Programas de Computador e Topografias de Circuitos

Anexo B – Certificado de Registro de Programa de Computador da ferramenta Clean-Alignment

Órgão Responsável: Instituto Nacional de Propriedade Industrial (INPI)

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA INDÚSTRIA, COMÉRCIO EXTERIOR E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512018051751-0**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 29/12/2017, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: Clean-alignment

Data de criação: 29/12/2017

Titular(es): UNIVERSIDADE FEDERAL DE PELOTAS

Autor(es): FREDERICO SCHMITT KREMER; LUCIANO DA SILVA PINTO

Linguagem: PYTHON

Campo de aplicação: BL-02; BL-04

Tipo de programa: AP-01

Algoritmo hash: OUTROS

Resumo digital hash: 1081f3cbe0744035294a9b035001e2d9

Expedido em: 02/10/2018

Aprovado por:
Liane Elizabeth Caldeira Lage
Diretora de Patentes, Programas de Computador e Topografias de Circuitos

Anexo C – Artigo descrevendo a ferramenta Genix, referente ao projeto de Mestrado, e publicado durante o período do Doutorado. A ferramenta foi utilizada nos artigos 1 e 3 desta tese.

Periódico: FEMS Microbiology Letters (Qualis na área de Biotecnologia: B2)

FEMS Microbiology Letters, 363, 2016, fnw263

doi: 10.1093/femsle/fnw263
Advance Access Publication Date: 16 November 2016
Research Letter

RESEARCH LETTER – Taxonomy & Systematics

Genix: a new online automated pipeline for bacterial genome annotation

Frederico Schmitt Kremer*, Marcus Redü Eslabão, Odir Antônio Dellagostin and Luciano da Silva Pinto*

Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, Rio Grande do Sul, Brazil, 96010-610

*Corresponding author: Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, Rio Grande do Sul, Brazil, 96010-610. Tel: +55 53 3275 7350; E-mail: fred.s.kremer@gmail.com, ls_pinto@hotmail.com

One sentence summary: We present Genix, a new web-based tool that may be used to identify genes in the genomic sequence of microorganisms with high accuracy.

Editor: Andrew Millard

ABSTRACT

Next-generation sequencing has significantly reduced the cost of genome-sequencing projects, resulting in an expressive increase in the availability of genomic data in public databases. The cheaper and easier is to sequence new genomes, the more accurate the annotation steps have to be to avoid both the loss of information and the accumulation of erroneous features that may affect the accuracy of further analysis. In the case of bacteria genomes, a range of web annotation software has been developed; however, many applications have yet to incorporate the steps required to improve their result, including the removal of false-positive/spurious and a more complete identification of non-coding features. We present Genix, a new web-based bacterial genome annotation pipeline. A comparison of the results generated by Genix for four reference genomes against those generated by other annotation tools indicated that our pipeline is able to provide results that are closer to the reference genome annotation, with a smaller amount of false-positive proteins and missing functional annotated proteins. Additionally, the metrics obtained by Genix were slightly better than those obtained by Prokka, a state-of-art standalone annotation system. Our results indicate that Genix is a useful tool that is able to provide a more refined result, and may be a user-friendly way to obtain high-quality results.

Keywords: next-generation sequencing; genomics; webserver; prokaryotes; whole-genome shotgun

INTRODUCTION

The next-generation sequencing (NGS) platforms, including Roche 454, Illumina Solexa, Ion Torrent PGM and PacBio SMRT, have provided an accessible way to obtain high-quality whole genome sequences (Mardis 2008). The reduction in the cost per base, combined to the higher throughput, resulted in a notable increase in the availability of genomic data in public databases like GenBank and Uniprot (Metzker 2005).

The sequencing process usually results in millions of short reads, which are fragments of the genome that had their sequences identified. Velvet (Zerbino and Birney 2008), ABySS (Simpson *et al.* 2009), SPAdes (Bankevich *et al.* 2012) and Ray (Boisvert, Laviolette and Corbeil 2010) are examples of *de novo* assembly programs that can be used to reconstruct, at least partially, the chromosome sequence. Although it is hard to achieve a 'finished' (ungapped) genome assembly using NGS (Alkan, Sajjadian and Eichler 2010), the *de novo* programs usually generate reliable 'draft' assemblies that can be used in most of the common applications (Land *et al.* 2015), and it is possible to improve a genome assembly by joining the contigs using

Received: 30 April 2016; **Accepted:** 15 November 2016

scaffolding tools (e.g. SSPACE, SOPRA) (Dayarian, Michael and Sengupta 2010; Boetzer *et al*. 2011), reference-based contig ordering tools (e.g. Mauve, ABACAS and CONTIGuator) (Assefa *et al*. 2009; Rissman *et al*. 2009; Galardini *et al*. 2011) and gap filling programs (e.g. GapFiller, IMAGE, FGAP) (Tsai, Otto and Berriman 2010; Boetzer and Pirovano 2012; Piro *et al*. 2014).

After the assembly, the functional and structural features of the genome may be identified in the 'annotation' process, which usually relies on the integration of programs for *ab initio* gene finding, non-coding RNA (ncRNA) prediction, database search and protein functional annotation (Kisand and Lettieri 2013). *Ab initio* gene finding in prokaryotic genomes can be performed by programs such as GLIMMER (Delcher *et al*. 1999), GeneMark (Borodovsky *et al*. 2003) and Prodigal (Hyatt *et al*. 2010), which identify possible coding-DNA sequences (CDSs) by looking for those open reading frames (ORFs) that match to a given gene model and criteria. The identified CDSs can be compared to a reference sequence database (e.g. Genbank, Uniprot), using tools like BLAST, BLAT and USEARCH, to infer their molecular function based on sequence similarity. The inference of non-coding genes, such as those responsible for tRNAs and rRNAs, can be performed by specific-purpose tools, such as tRNAscan-SE (Lowe and Eddy 1997), RNAmmer (Lagesen *et al*. 2007) and Aragorn (Laslett 2004), or by general-purpose tools, such as Infernal (Nawrocki, Kolbe and Eddy 2009). In the same way as CDS identification, RNA families can also be inferred using pairwise sequence alignment, but covariance models (CMs) and hidden Markov models usually provide a more reliable result as they also consider structural features and sequence conservation, which is important for structural RNAs (Durbin *et al*. 1998).

The accuracy of the annotation process has a strong impact on the studies that will be performed using its results. For pathogenic bacteria, for example, the annotated protein-coding genes may be used to identify new candidates for recombinant vaccine development or targets for antibiotics therapeutics (Sette and Rappuoli 2010), genes responsible for antibiotic-resistance (Köser, Ellington and Peacock 2014) or even *loci* that can be applied in molecular typing for epidemiological studies (Salipante *et al*. 2015). An accurate annotation is also important for RNA-Seq studies, where annotation artifacts may compromise the analysis of differential gene expression.

The genome annotation, even for bacteria, would become a laborious and time-consuming process if executed manually, so it is usually performed by automated pipelines (Beckloff *et al*. 2012). Examples of pipelines for bacteria genome annotation include the webservers RAST (Aziz *et al*. 2008), BASys (Van Domselaar *et al*. 2005) and xBASE (Chaudhuri and Pallen 2006), and the standalone applications BG7 (Pareja-Tobes *et al*. 2012), Prokka (Seemann 2014), Eugene-PP (Sallet, Gouzy and Schiex 2014) and MEGAnnotator (Lugli *et al*. 2016), which integrate different sets of programs and databases, and may provide slightly different results for the same genome.

Although webservers are usually user friendly, they do not typically offer the same level of customization or even a deeper identification of non-coding features, but instead a simplified annotation of coding DNA sequences, rRNAs and tRNAs. As many ncRNAs are associated with gene-expression regulation, their correct identification is important for a better understanding of the resulting transcriptome and molecular physiology of the sequenced organism (Repoila and Darfeuille 2009). Additionally, both RAST, BASys and xBASE use the program Glimmer (Delcher *et al*. 1999) for gene finding, although it has already been demonstrated that the results from this program may contain a considerable number of false-positive genes when compared to newer software like Prodigal (Hyatt *et al*. 2010). As none of these pipelines include additional steps to remove false-positive genes from their annotation (e.g. Antifam; Eberhardt *et al*. 2012), their results tends to be less reliable.

We present Genix, a new web-based automated bacterial annotation pipeline developed to provide a more reliable annotation of both protein-coding and non-coding genes. To evaluate its accuracy, we compared annotations generated by Genix for four reference bacterial genomes to their original annotations and to those generated by RAST, BASys and Prokka.

METHODOLOGY

The pipeline

Genix takes as input a set of sequences (in FASTA format) and a taxonomy identifier (tax id), which is a common identification code that is used by many biological sequence databases (e.g. GenBank, Uniprot) to identify different taxons (https://www.ncbi.nlm.nih.gov/Taxonomy/). Using the tax id, a set of proteins is retrieved from UniProt to generate the 'raw dataset' that is processed by CD-HIT (Li and Godzik 2006), a sequence clustering tool, to generate a final and non-redundant protein dataset for the subsequent annotation steps. The clustering process reduces the size of the protein database by grouping sequences that have an identity equal of greater than a given threshold and electing a representative sequence for each group. The user can adjust the identity threshold according to its needs, although we recommend the use of the default value '95%'.

After clustering, a script parses the CD-HIT output to extract the best annotation for each group. This process is necessary because proteins derived from Uniprot's unreviewed subset (trEMBL) may present an incomplete or even erroneous annotation, as they are not manually reviewed like those from the Swissprot. As CD-HIT just takes into account the sequence identity for defining the representative sequence in each group, it is possible to an unreviewed protein be selected instead of reviewed one, and the protein identification (product name) could be compromised. Additionally, the proteins from Uniprot also have a 'protein score' that indicates the level of 'evidence' (e.g. 'inferred from automatic annotation', 'transcript level evidence', 'protein level evidence') for each entry, and it can also be used to identify the most reliable product nam. Therefore, the extraction of the best annotation is based on two rules: (1) prefer Swissprot product names instead of trEMBL product names (2) If two proteins are from trEMBL, prefer the annotation with the higher score.

For the annotation, ORFs are predicted by Prodigal (Hyatt *et al*. 2010) and then searched against the protein database using BLASTp (Altschul *et al*. 1990), from the NCBI-BLAST+ package (Camacho *et al*. 2009) and against the spurious ORFs database AntiFam (Eberhardt *et al*. 2012) using HMMER (Eddy 2011). Those ORFs that have a hit on the protein database with an e-value below the specified threshold and have not significant hit on Antifam are maintained in the annotation. By default, significant hits are considered when the e-value is smaller than 0.0005, and it is applied both for BLAST and the Antifam HMMER search.

The identification of tRNAs, rRNAs and tmRNAs is performed using tRNAscan-SE (Lowe and Eddy 1997), RNAmmer (Lagesen *et al*. 2007) and Aragorn (Laslett 2004), respectively. Other classes of ncRNAs and some cis-regulatory elements (e.g. riboswitches) are identified using a BLASTn (Altschul *et al*. 1990) search against the RFam database (Griffiths-Jones *et al*. 2003), followed by a CM alignment using INFERNAL (Nawrocki, Kolbe and Eddy 2009). The combination of

BLASTn and INFERNAL is similar to that implemented in the script 'rfam_scan.pl', available at the Rfam's FTP page (ftp://ftp.ebi.ac.uk/pub/databases/Rfam/tools/rfam_scan.pl), and it is applied to speed up the process. Although the INFERNAL algorithm (based on CMs) can accurately identify ncRNAs, it is also very slow when compared to BLAST-based search (heuristic local alignment). However, BLAST doesn't consider structural implications and sequence conservation and it is expected to have a relatively high rate of false positives, so it is only employed to reduce the search space.

In the final data integration step, CDSs that overlap ncRNAs (except antisense RNAs genes) are removed and 5′ or 3′ broken CDSs and genes are corrected if located at the start or end of the sequences (common in draft genomes) by adjusting the 'codon_start' information. The start codon position is also adjusted by checking it in the hit coverage in the BLAST alignment. If the 5′ terminus of the CDS doesn't covers the start of the hit protein, Genix tries to extend the gene by shifting to next allowed start codon in the upstream region between the previous feature. If the extension improves the hit coverage, and the new 5′ doesn't overlap any prohibited feature (e.g. ncRNA) or has a hit in the Antifam database, the changing is maintained. A similar approach is used to reduce the CDS region if its 5′ terminus is not covered by the hit sequence or when an Antifam hit is found on the ORF. However, those Antifam hits that may not be corrected just by changing the 5′ terminus are removed.

If the user provides a Genbank submission template file (.sbt), the Genix webserver uses tbl2asn to generate a submission file for Genbank. An overview of the main pipeline is presented in Fig. 1.

Implementation of the webserver

Genix was developed to run on an Apache HTTP server. The core of the application was written in Python 2.7 with some Bash and Perl scripts, which were designed to process controlling and text processing tasks respectively. MySQL was used as the relational main database managing system, storing all the data submitted by the users, the jobs and user's information. During the annotation, a small SQLite database was also used to store the features identified by each software application before the final data integration step. The genome browser JBrowse (Skinner *et al.* 2009) was also integrated and is available after the annotation be completed.

Validation

To evaluate the accuracy of Genix, we compared its results to those generated by the web-based annotation pipelines RAST and BASys, and to the stand-alone annotation pipeline Prokka, for the genome of *Leptospira interrogans* serovar Copenhageni strain Fiocruz L1-130 (GenBank: NC_000913.3), *Escherichia coli* strain K12 (GenBank: AE016823.1), *Listeria monocytogenes* strain EGD-e (GenBank: AL591824.1) and *Mycobacterium tuberculosis* strain H37Rv (GenBank: AL123456.3). All programs were executed in default settings. For Genix, all genomes were annotated using their respective genus-level tax ids: *Mycobacterium* (1763), *Escherichia* (561), *Leptospira* (171) and *Listeria* (1637).

The annotations generated by Genix and the other pipelines were compared to the reference annotations to identify missing CDSs (present in the reference but absent in the new annotation) and new CDSs (present in the new annotation but absent in the reference). Additionally, all annotations were analyzed and compared to each another to identify exclusive genes (not found in any other). Based on the count of missing and new genes, an 'annotation discrepancy' was calculated using the equation:

$$\frac{(\text{No of missing CDSs}) + (\text{No of new CDSs})}{(\text{No of CDSs in the reference}) + (\text{No of CDSs in the annotation})}$$

Using this formula, a discrepancy equal to 0 would represent an annotation 100% equal to the reference, while 1 would present a complete absence of all reference CDSs, or a scenario where the annotations are completely different, for example. As the discrepancy is affected both by the number of new genes and missing genes, it may be used to evaluate how 'far' is the new annotation from its respective reference, what can also indicate, at least partially, the degree of 'misannotation'. If for a genome with 4500 known genes, a novel annotation program identifies 5000 genes, were 4250 are in the original annotation and 750 are completely new, it's discrepancy will be $(250 + 750)/(4500 + 5000) = 10.52\%$.

To verify the relevance of the discrepant *loci*, all the missing, new and exclusive genes were functionally annotated by BLAST2GO using the Swiss-Prot database for BLASTx searches. Finally, all CDSs from all annotations were aligned by HMMER against the AntiFam database with an e-value threshold of 0.0005 to analyze the occurrence of spurious ORFs.

RESULTS

Genix was implemented as a webserver. During the submission, the user can control the identity threshold used by CD-HIT, BLAST and HMMER, upload submission templates (.sqn), provide linkage information (in case of scaffolds) and source modifiers (e.g. strain, serogroup and serovar), which may be useful in the case of genomes that will be submitted to Genbank after annotation. After submission, a job ID is generated, and the user may use this to access the results. Genix webserver provides the annotation in Genbank (.gb), GFF and Feature Table (.tbl) formats, and FASTA files containing protein sequences (.faa), non-coding features (.ffn) and coding DNA sequences (.fna). Additionally, if a Genbank submission template is uploaded, a compressed folder containing the outputs of tbl2asn is also provided by the server. Finally, it is also possible to submit multiples genomes for annotation using the 'bulk submission portal' that supports collections of genomes in 'tar.gz' compressed files, with a limit size of ∼50 Mb per submission.

The results of the validation using the reference genomes of *Mycobacterium tuberculosis*, *Escherichia coli*, *Leptospira interrogans* and *Listeria monocytogenes* e are shown in Tables 1–4, respectively. A deeper analysis of the percentage of functional CDSs in the new CDSs identified by each program is shown in Table 5. Finally, Table 6 contains a brief description of the non-coding features identified in the reference genomes that were neglected by the RAST and BASys web annotation pipelines.

DISCUSSION

In the comparison to other annotation pipelines, Genix achieved the lowest discrepancy for three of the four reference genomes (Tables 1–4), having been only surpassed by Prokka in the genome of *Escherichia coli* K-12 (Table 2). For two genomes (*Mycobacterium tuberculosis* and *Escherichia coli*), Genix achieved the lowest values of 'missing functional proteins', and for *Leptospira interrogans* it achieved the same value obtained by Prokka, which was smaller than those obtained by RAST and BASys, indicating

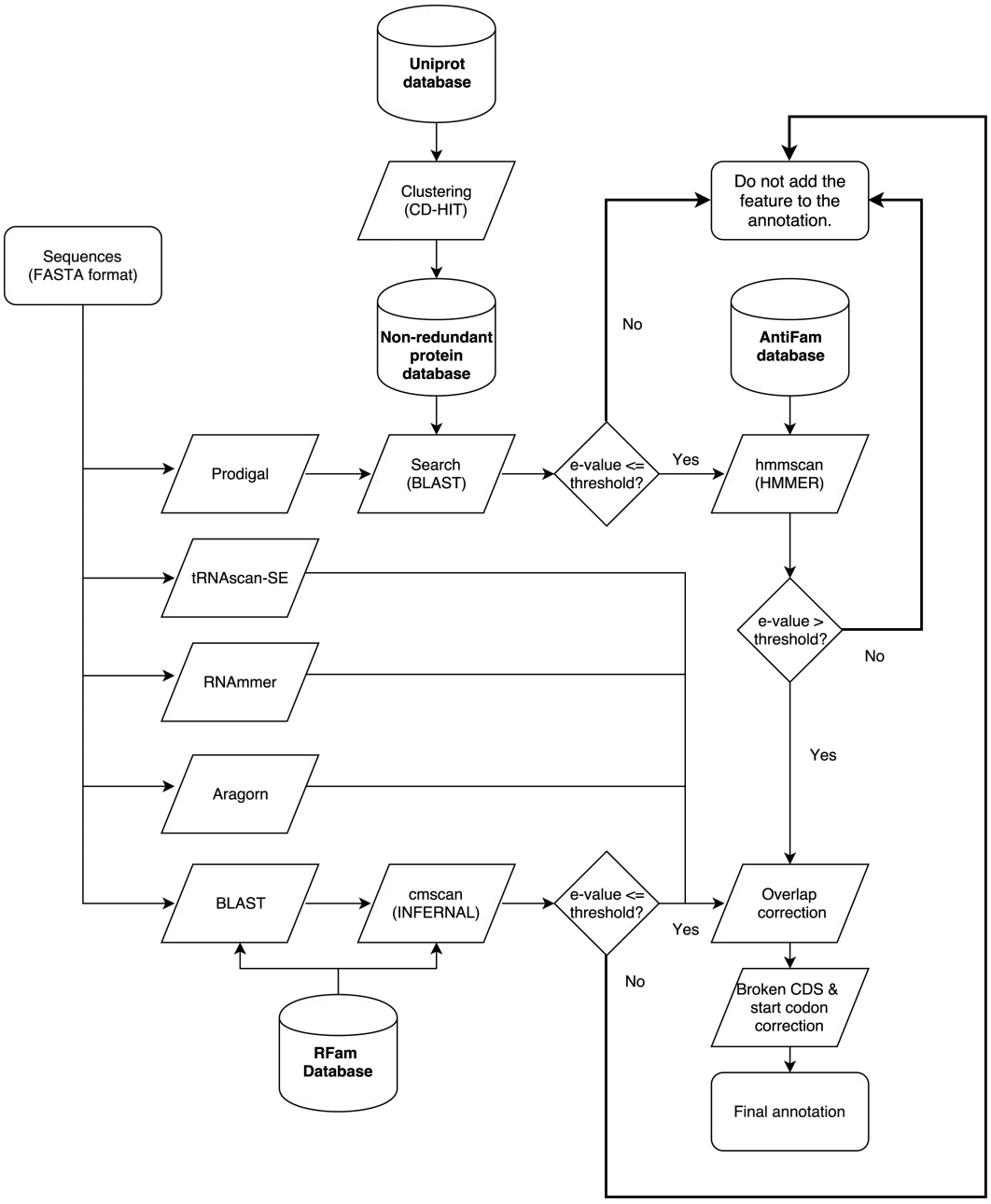


Figure 1. Flowchart representing the structure of the Genix annotation pipeline.

Table 1. Results of the annotations generated by Genix, RAST and BASys for the genome of *M. tuberculosis* strain H37Rv compared to the Genbank annotation.

	CDS								
Software	Total	Missing	New	Exclusive	Missing functional	New functional	Exclusive functional	Discrepancy (%)	AntiFam hits
GenBank	4031	–	–	97	–	–	21	0.00	1
Genix	3907	175	223	7	52	30	1	4.94	0
RAST	4061	251	532	258	87	65	24	9.39	2
BASys	4406	235	845	597	91	66	37	12.45	4
Prokka	3867	204	244	4	70	38	2	5.53	2

Table 2. Results of the annotations generated by Genix, RAST and BASys for the genome of *E. coli* strain K12 compared to the Genbank annotation.

	CDS								
Annotation	Total	Missing	New	Exclusive	Missing functional	New functional	Exclusive functional	Discrepancy (%)	AntiFam hits
GenBank	4319	–	–	150	–	–	63	0.00	0
Genix	4141	176	174	6	79	65	2	4.05	0
RAST	4173	264	382	270	137	95	33	7.38	4
BASys	3861	287	116	25	147	83	19	4.76	1
Prokka	4135	181	178	4	83	69	1	4.16	1

Table 3. Results of the annotations generated by Genix, RAST and BASys for the genome of *L. interrogans* strain L1130 (chromosome I) compared to the Genbank annotation.

	CDS								
Software	Total	Missing	New	Exclusive	Missing functional	New functional	Exclusive functional	Discrepancy (%)	AntiFam hits
GenBank	3394	–	–	78	–	–	3	0.00	40
Genix	3146	227	206	26	5	30	2	6.40	0
RAST	4178	167	1118	638	12	33	3	16.60	285
BASys	3974	194	968	517	13	35	6	15.37	156
Prokka	3172	223	224	20	5	32	2	6.58	37

Table 4. Results of the annotations generated by Genix, RAST and BASys for the genome of *L. monocytogenes* strain EGD-e compared to the Genbank annotation.

	CDS								
Software	Total	Missing	New	Exclusive	Missing functional	New functional	Exclusive functional	Discrepancy (%)	AntiFam hits
Ref	2855	–	–	9	–	–	3	0.00	2
Genix	2834	25	29	4	7	3	0	0.95	0
RAST	2869	32	78	37	6	5	1	1.91	1
BASys	2864	56	121	82	7	4	0	3.06	1
Prokka	2860	15	35	7	3	3	0	0.87	1

Table 5. Percentage of functional CDS in the new CDSs identified by Genix, RAST and BASys in the genomes of *M. tuberculosis*, *E. coli*, *L. interrogans* and *L. monocytogenes*.

Software	*M. tuberculosis*	*E. coli*	*L. interrogans*	*L. monocytogenese*
Genix	13.45	37.35	14.56	10.34
RAST	12.21	24.86	2.9	6.4
BASys	7.8	71.55	3.6	3.3
Prokka	15.57	38.76	14.28	8.5

that the number of trustable proteins that have been neglected in their annotations is smaller than in the annotation generated by those online tools. Therefore, the decrease in the number of identified proteins by these tools may not necessarily imply in an increase in the number of false-negatives, but represent a more refined result.

By analyzing the percentage of new genes with functional annotation (Table 5), it is also possible to verify the disparity in the annotations generated by BASys. While in the annotation of the *E. coli* K-12 genome, ∼71% of the new genes were identified as functional by BLAST2GO, for the other genomes the percentage was smaller than 8%. It may indicate that gene finding algorithm used by BASys and its database may be optimized and overfitted to work with some specific gene models. Therefore, it also indicate annotation generated by BASys for genomes of newly discovered organisms would present a lower reliability when compared to those generated for widely studied organisms, or by other annotation pipelines. It may also be extended to RAST, as for *M. tuberculosis* and *L. interrogans* its discrepancy was higher than 9% and 16%, respectively.

For the genome of *L. interrogans,* Genix achieved a discrepancy of 6.40% and Prokka 6.58%, while RAST and BASys annotations differed ∼16% in the CDS content comparing to the reference annotation. The analysis using Antifam also revealed that the reference annotation for this genome includes 40 spurious ORFs, while Prokka, RAST and BASys had 37, 158 and 288, respectively. The presence of a high number of Antifam hits in the reference genome of *L. interrogans* strain L1-130 may be a

Table 6. Non-coding features identified by INFERNAL and added by Genix to the annotation for the genomes of *E. coli* K-12, *L. monocytogenese* EGD-e, *L. interrogans* L1-130 and *M. tuberculosis* strain h37Rv.

Specie	Non-coding feature	Occurences	Rfam accession
Escherichia coli strain K12	rdlD antitoxin	4	RF01813
Escherichia coli strain K12	MicF RNA	1	RF00033
Escherichia coli strain K12	RyhB RNA	1	RF00057
Escherichia coli strain K12	RyeB RNA	1	RF00111
Escherichia coli strain K12	STnc150 Hfq binding RNA	1	RF01402
Escherichia coli strain K12	Sok antitoxin	8	RF01794
Escherichia coli strain K12	ArcZ RNA	1	RF00081
Escherichia coli strain K12	IS061 RNA	1	RF00115
Escherichia coli strain K12	Enterobacterial sRNA STnc430	1	RF02053
Escherichia coli strain K12	Bacterial RNase P class A	1	RF00010
Escherichia coli strain K12	sroD RNA	1	RF00370
Escherichia coli strain K12	t44 RNA	1	RF00127
Escherichia coli strain K12	SraC/RyeA RNA	1	RF00101
Escherichia coli strain K12	*Salmonella enterica* sRNA STnc40	1	RF02057
Escherichia coli strain K12	Selenocysteine transfer RNA	21	RF01852
Escherichia coli strain K12	RybB RNA	1	RF00110
Escherichia coli strain K12	DnaX ribosomal frameshifting element	1	RF00382
Escherichia coli strain K12	rseX Hfq binding RNA	1	RF01401
Escherichia coli strain K12	Tryptophan operon leader	1	RF00513
Escherichia coli strain K12	DsrA RNA	1	RF00014
Escherichia coli strain K12	MicC RNA	1	RF00121
Escherichia coli strain K12	SgrS RNA	1	RF00534
Escherichia coli strain K12	C0299 RNA	1	RF00119
Escherichia coli strain K12	Enterobacterial sRNA STnc450	1	RF02051
Escherichia coli strain K12	Fumarate/nitrate reductase regulator sRNA	1	RF01796
Escherichia coli strain K12	Threonine operon leader	2	RF00506
Escherichia coli strain K12	OxyS RNA	1	RF00035
Escherichia coli strain K12	C0465 RNA	1	RF00116
Escherichia coli strain K12	CsrB/RsmB RNA family	1	RF00018
Escherichia coli strain K12	RprA RNA	1	RF00034
Escherichia coli strain K12	CyaR/Rye RNA	1	RF00112
Escherichia coli strain K12	Enterobacterial sRNA STnc70	1	RF02069
Escherichia coli strain K12	Short intergenic abundant RNA	5	RF00113
Escherichia coli strain K12	sraA	1	RF02029
Escherichia coli strain K12	OrzO-P RNA antitoxin family	1	RF02083
Escherichia coli strain K12	tpke11	1	RF02031
Escherichia coli strain K12	GadY	1	RF00122
Escherichia coli strain K12	sroC RNA	1	RF00369
Escherichia coli strain K12	Enterobacterial sRNA STnc180	1	RF02079
Escherichia coli strain K12	nuoG RNA	1	RF01748
Escherichia coli strain K12	Enterobacterial sRNA STnc480	1	RF02068
Escherichia coli strain K12	SymR antitoxin	1	RF01809
Escherichia coli strain K12	RydC RNA	1	RF00505
Escherichia coli strain K12	Enterobacterial sRNA STnc240	1	RF02074
Escherichia coli strain K12	Alpha operon ribosome binding site	1	RF00140
Escherichia coli strain K12	GlmZ RNA activator of glmS mRNA	2	RF00083
Escherichia coli strain K12	DicF RNA	2	RF00039
Escherichia coli strain K12	sroB RNA	1	RF00368
Escherichia coli strain K12	Selenocysteine insertion sequence 3	3	RF01989
Escherichia coli strain K12	Selenocysteine insertion sequence 2	1	RF01988
Escherichia coli strain K12	SraG RNA	1	RF00082
Escherichia coli strain K12	Enterobacteria sRNA STnc130	1	RF02084
Escherichia coli strain K12	Enterobacterial sRNA STnc370	1	RF02064
Escherichia coli strain K12	IS128 RNA	1	RF00125
Escherichia coli strain K12	cspA thermoregulator	7	RF01766
Escherichia coli strain K12	istR Hfq binding RNA	2	RF01400
Escherichia coli strain K12	Archaeal RNase P	1	RF00373
Escherichia coli strain K12	sroH RNA	1	RF00372
Escherichia coli strain K12	Spot 42 RNA	1	RF00021
Escherichia coli strain K12	Glm Y RNA activator of glmS mRNA	2	RF00128
Escherichia coli strain K12	Ribosomal S15 leader	1	RF00114
Escherichia coli strain K12	Leucine operon leader	2	RF00512

Table 6. (*Continued*).

Specie	Non-coding feature	Occurences	Rfam accession
Escherichia coli strain K12	CsrC RNA family	2	RF00084
Escherichia coli strain K12	MicA sRNA	1	RF00078
Escherichia coli strain K12	yybP-ykoY leader	2	RF00080
Escherichia coli strain K12	Phenylalanine leader peptide	3	RF01859
Escherichia coli strain K12	Proteobacterial sRNA sX4	1	RF02223
Escherichia coli strain K12	JUMPstart RNA	3	RF01707
Escherichia coli strain K12	sroE RNA	1	RF00371
Escherichia coli strain K12	RNase E 5' UTR element	1	RF00040
Escherichia coli strain K12	SraB RNA	1	RF00077
Escherichia coli strain K12	Repression of heat shock gene expression	1	RF00435
Escherichia coli strain K12	Enterobacterial sRNA STnc380	1	RF02055
Escherichia coli strain K12	Enterobacterial sRNA STnc550	1	RF02081
Escherichia coli strain K12	tp2	1	RF02030
Escherichia coli strain K12	Enterobacterial sRNA STnc410	1	RF02060
Escherichia coli strain K12	sraL Hfq binding RNA	1	RF01408
Escherichia coli strain K12	ryfA RNA	1	RF00126
Escherichia coli strain K12	IS102 RNA	1	RF00124
Escherichia coli strain K12	rydB RNA	1	RF00118
Escherichia coli strain K12	Histidine operon leader	2	RF00514
Escherichia coli strain K12	Bacterial antisene RNA HPnc0260	1	RF02194
Escherichia coli strain K12	STnc560 Hfq binding RNA	1	RF01407
Escherichia coli strain K12	rncO	1	RF00552
Escherichia coli strain K12	iscR stability element	1	RF01517
Escherichia coli strain K12	Enterobacterial sRNA STnc630	1	RF02052
Escherichia coli strain K12	OmrA-B family	2	RF00079
Escherichia coli strain K12	Gammaproteobacterial sRNA STnc100	2	RF02076
Escherichia coli strain K12	RtT RNA	11	RF00391
Escherichia coli strain K12	Enterobacterial sRNA STnc540	1	RF02082
Escherichia coli strain K12	C0719 RNA	1	RF00117
Listeria monocytogenes egd-e	Ribosomal protein L10 leader	1	RF00557
Listeria monocytogenes egd-e	Listeria sRNA rli56	1	RF01483
Listeria monocytogenes egd-e	Listeria Hfq binding LhrC	5	RF00616
Listeria monocytogenes egd-e	Listeria sRNA rli54	1	RF01491
Listeria monocytogenes egd-e	Listeria Hfq binding LhrA	2	RF00615
Listeria monocytogenes egd-e	Listeria sRNA rli52	1	RF01480
Listeria monocytogenes egd-e	Listeria sRNA rli53	1	RF01481
Listeria monocytogenes egd-e	Listeria sRNA sbrA	1	RF01489
Listeria monocytogenes egd-e	Listeria snRNA rli45	1	RF01475
Listeria monocytogenes egd-e	Ribosomal protein L13 leader	1	RF00555
Listeria monocytogenes egd-e	Listeria sRNA rliH	1	RF01460
Listeria monocytogenes egd-e	Listeria sRNA rliI	1	RF01487
Listeria monocytogenes egd-e	Bacterial antisene RNA HPnc0260	1	RF02194
Listeria monocytogenes egd-e	L17 ribosomal protein downstream element	1	RF01708
Listeria monocytogenes egd-e	Listeria sRNA rliA	3	RF01464
Listeria monocytogenes egd-e	Listeria sRNA rliB	1	RF01471
Listeria monocytogenes egd-e	Listeria sRNA rli59	1	RF01484
Listeria monocytogenes egd-e	Listeria sRNA rliD	2	RF01494
Listeria monocytogenes egd-e	Listeria sRNA rliE	9	RF01459
Listeria monocytogenes egd-e	Listeria sRNA rliF	1	RF01476
Listeria monocytogenes egd-e	Listeria sRNA rli38	2	RF01470
Listeria monocytogenes egd-e	Bacterial RNase P class B	1	RF00011
Listeria monocytogenes egd-e	Bacterial RNase P class A	1	RF00010
Listeria monocytogenes egd-e	Listeria sRNA rli31	1	RF01465
Listeria monocytogenes egd-e	Listeria sRNA rli32	1	RF01468
Listeria monocytogenes egd-e	Listeria sRNA rli33	1	RF01469
Listeria monocytogenes egd-e	Listeria sRNA rli34	2	RF01466
Listeria monocytogenes egd-e	Listeria sRNA rli36	1	RF01467
Listeria monocytogenes egd-e	Listeria sRNA rli37	1	RF01493
Listeria monocytogenes egd-e	Listeria sRNA rli49	1	RF01488
Listeria monocytogenes egd-e	Listeria sRNA rli48	1	RF01479
Listeria monocytogenes egd-e	Listeria sRNA rli62	1	RF01486
Listeria monocytogenes egd-e	Listeria sRNA rli61	1	RF01485
Listeria monocytogenes egd-e	Listeria sRNA rli41	2	RF01473

Table 6. (*Continued*).

Specie	Non-coding feature	Occurences	Rfam accession
Listeria monocytogenes egd-e	Listeria sRNA rli40	3	RF01472
Listeria monocytogenes egd-e	Listeria sRNA rli43	1	RF01477
Listeria monocytogenes egd-e	Listeria sRNA rli47	1	RF01478
Listeria monocytogenes egd-e	T-box leader	16	RF00230
Listeria monocytogenes egd-e	RNA anti-toxin A	4	RF01776
Listeria monocytogenes egd-e	cspA thermoregulator	3	RF01766
Listeria monocytogenes egd-e	Ribosomal protein L20 leader	2	RF00558
Listeria monocytogenes egd-e	Listeria snRNA rli51	1	RF01490
Listeria monocytogenes egd-e	ykoK leader	1	RF00380
Listeria monocytogenes egd-e	PrfA thermoregulator UTR	1	RF00038
Listeria monocytogenes egd-e	PyrR binding site	2	RF00515
Listeria monocytogenes egd-e	Listeria snRNA rli23	3	RF01458
Listeria monocytogenes egd-e	yybP-ykoY leader	1	RF00080
Listeria monocytogenes egd-e	Listeia sRNA rli42	1	RF01474
Listeria monocytogenes egd-e	Listeria sRNA rli28	4	RF01492
Listeria monocytogenes egd-e	Listeria sRNA rli22	2	RF01457
Listeria monocytogenes egd-e	Ribosomal protein L21 leader	1	RF00559
Listeria monocytogenes egd-e	Listeria sRNA rli27	1	RF01463
Listeria monocytogenes egd-e	Listeria sRNA rli26	1	RF01462
Leptospira interrogans L1-130	CRISPR RNA direct repeat element	4	RF01315
Leptospira interrogans L1-130	Gammaproteobacterial sRNA STnc100	2	RF02076
Leptospira interrogans L1-130	Bacterial RNase P class A	1	RF00010
Mycobacterium tuberculosis h37rv	ydaO/yuaA leader	1	RF00379
Mycobacterium tuberculosis h37rv	ykoK leader	2	RF00380
Mycobacterium tuberculosis h37rv	ASpks TB sRNA	5	RF01782
Mycobacterium tuberculosis h37rv	yybP-ykoY leader	1	RF00080
Mycobacterium tuberculosis h37rv	Mycobacterium B11	1	RF01783
Mycobacterium tuberculosis h37rv	mraW RNA	2	RF01746
Mycobacterium tuberculosis h37rv	Bacterial RNase P class A	1	RF00010
Mycobacterium tuberculosis h37rv	ASdes TB sRNA	3	RF01781
Mycobacterium tuberculosis h37rv	6C RNA	1	RF01066
Mycobacterium tuberculosis h37rv	Actino-pnp RNA	1	RF01688

consequence of mistranslations of insertion sequences (IS) in the 5′ terminus of transposases genes and not necessarily real false positive, but genes with wrong start codon identification. However, in the case of RAST and BASys, for this same genome there were also mistranslations of other non-coding and structural features, including repeated DNA elements and shadow ORFs. Therefore, it demonstrates the relevance of the inclusion of methods to detect spurious ORFs in automated genome annotation pipelines.

It was also observed that some annotations generated by RAST and BASys present overlapping CDSs in opposite strands. Although gene overlapping is frequently observed in bacterial genomes (Johnson and Chisholm 2004), usually the overlap is limited to only a few nucleotides (<30 nucleotides), and do not cover a large portion of the coding region. An example of this type of error, usually called 'shadow ORF' (Eberhardt *et al*. 2012), identified in the annotation generated by BASys for the genome of *M. tuberculosis* H37Rv is presented in Fig. 2. In this case, the CDS located in the reverse strand, annotated as a hypothetical protein, is absent in the reference annotation and no hits to non-hypothetical proteins in GenBank, being probably a bias of the gene finding program. On the other hand, the CDS annotated in the plus strand, also present in the GenBank reference annotation, and also found by Genix and RAST, has a InterPro domain (InterPro: IPR010273). A complete report of all overlapping CDSs in opposite strands, identified in the annotations for the genomes of *M. tuberculosis* strain H32Rv, *E. coli* strain K12, *L. interrogans* serovar Copenhageni strain Fiocruz L1-130 (Chromosome I) and *Listeria monocytogenes strain* EDG-e is presented in Supplementary Data 1.

The fact that some genes present in the original (reference) annotations could not be found by any annotation pipeline, or were missed by most of them, may indicates the limitation of *ab initio* gene finding in bacteria, and it is particularly true for those genes with short ORFs, which represent the major part of the reference-exclusive genes in the four tested genomes. Although short ORFs may be a consequence of false positives in gene finding algorithms, some of they are in fact functional, as shown by the results from BLAST2GO, and may be wrongly bypassed by programs such as Prodigal to avoid the errors usually observed in GLIMMER, for example. BASys, which uses GLIMMER as the main gene finding algorithm, and RAST, which uses it as an alternative method for gene prediction, usually tends to misidentify non-coding ORFs as potential genes and even in the annotation generated by these programs a great number of short real genes were missing. It is possible that these genes may only be correctly annotated using additional information, such as gene models constructed using curated information and/or experimental data, such as RNA-Seq and MS-MS data, for example, in a similar way than what is applied for organisms with complex gene structure (e.g. eukaryotes).

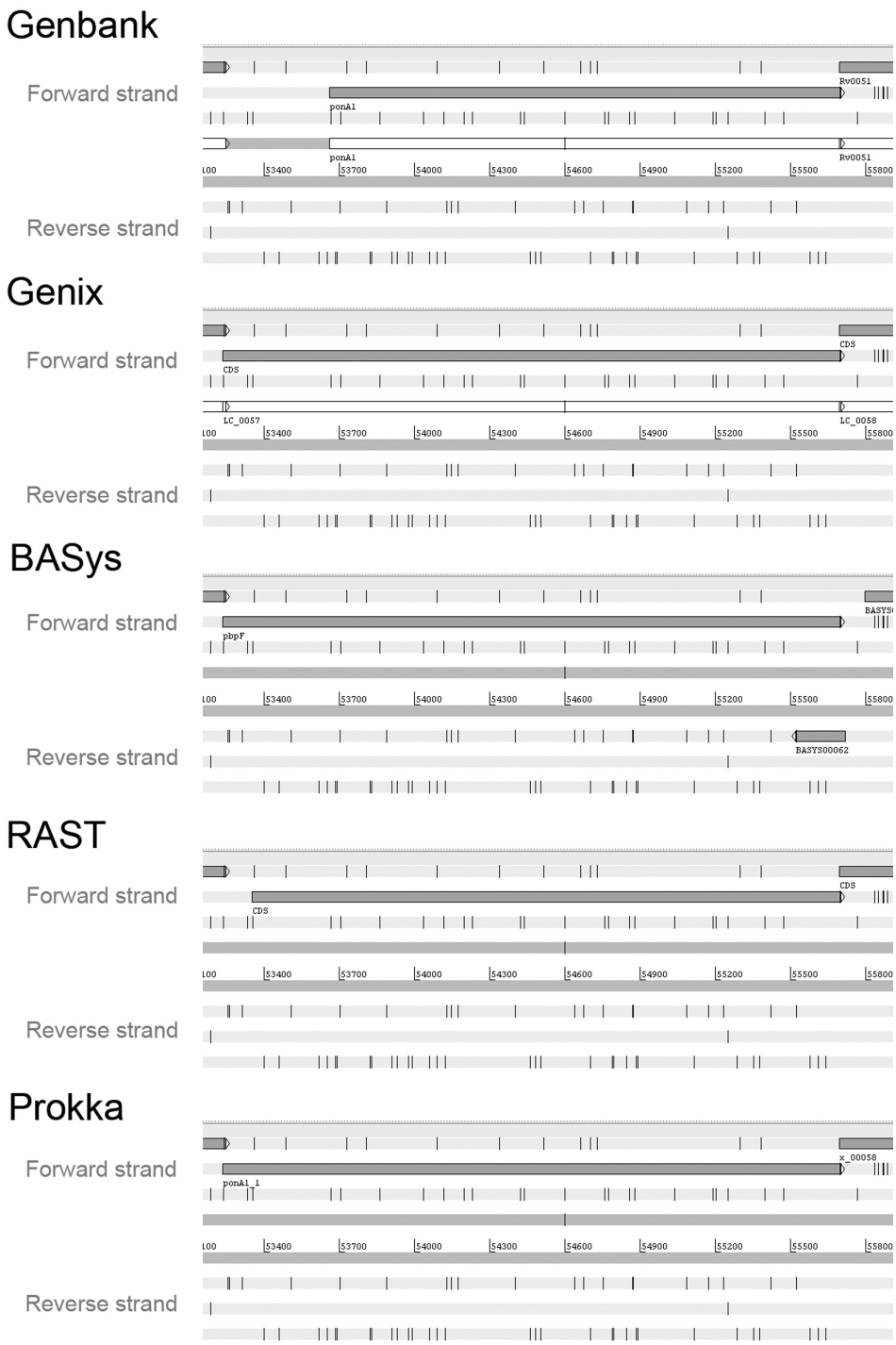


Figure 2. Example of overlapping CDSs in opposite strands predicted by the program BASys in the genome of *M. tuberculosis* H37Rv (GenBank: AL123456.3). The CDS located in the reverse strand (–) has no correspondent in the reference annotation, no domains identified by Interproscan and no hits to non-hypothetical proteins in Genbank, and may be an artifact.

Although Genix and Prokka share many characteristics in common, especially in terms of algorithm and the tools that are used, there are differences in the protein database that is used for the annotation, the codon start refinement, the use of Antifam and the user interface. In the comparison to the other annotation pipelines, these two tools showed a reduced rate of Antifam hits, missing functional proteins and the smallest values of discrepancy for the majority of the tested genomes, resulting in the annotations from Genix being slightly more accurate (Tables 1–4). Additionally, the percentage of functional genes on the new identified genes is slightly higher in the Genix annotations for at least two genomes (Table 5). As Prokka is a very

versatile tool, it is possible to boost its performance by using customized databases and trying different configurations, but in a straightforward analysis, using default settings, Genix may provide a more accurate result, and, due to its web-based interface, is more user friendly for users without a background in bioinformatics.

Finally, the non-coding features identified by BASys and RAST are limited to rRNAs and tRNAs, while Genix and Prokka also included tmRNAs, ribozymes and regulatory RNA motifs, along with many other families of ncRNAs and regulatory elements. TmRNAs are a family of ncRNAs with both mRNA and tRNA function, widely conserved in bacteria genomes, and play an important role in protein translation regulation (Hayes and Keiler 2010). Ribozymes are RNA molecules with catalytic activity and are usually associated with the degradation of other RNA molecules (Wang *et al*. 1996; Walter and Engelke 2002). Regulatory RNA motifs, such as riboswitches and thermoregulators (e.g. *cspA*), are present in the UTR region of some mRNAs and provide an alternative mechanism to control the mRNA expression (Vitreschak *et al*. 2004; Garst, Edwards and Batey 2011). As non-coding RNAs acts in many physiological processes (Repoila and Darfeuille 2009), their correct annotation is important for a better understanding of the sequenced organism's biology and its transcriptome. A report of the non-coding feature identified by INFERNAL and added Genix to the annotation of the four analyzed genomes is presented in Table 6.

The time required by Genix to annotate a bacteria genome is directly affected by the size of the protein database used by BLAST. As Uniprot contains many redundant sequences, the clustering tool CD-HIT was added to reduce the size of the final sequence database. To optimize the database construction, we also recommend the use of species-level or genus-level tax ids for well-studied organisms and more generic tax ids (e.g. phylum) for those species that are still poorly studied.

CONCLUSION

Genix, a web-based automated bacterial genome annotation pipeline, has demonstrated its potential as a useful tool that can provide more filtered and refined data than other annotation pipelines. When tested on the genomes of *Mycobacterium tuberculosis* strain H32Rv, *Escherichia coli* strain K12, *Leptospira interrogans* serovar Copenhageni strain L1-130 and *Listeria monocytogenes* strain EGD-e, Genix produced an annotation that was closer to the reference when compared to the annotations generated by the web-based tools RAST and BASys and to the standalone tool Prokka.

Although the number of new functional proteins is smaller if compared to other annotation pipelines, the percentage of functional proteins in all new CDSs was greater for Genix in three of the four genomes when comparing to the other web-based annotation pipelines. Genix also had the smallest values of missing proteins with functional annotation, indicating that it have a low ratio of relevant false negatives.

By integrating Antifam as part of the pipeline, we also reduced the probability of spurious ORFs, increasing the reliability of the protein identification. Antifam, along with ncRNA annotation, is also useful during the codon start refinement, which helps to improve the reliability of the annotation and provide more accurate data for analysis that may be employed using the annotated genome (e.g. RNA-Seq, reverse vaccinology).

In the comparative tests, BASys was able to identify larger number of new functional CDSs, but also miss-annotated several overlapping ORFs in opposite strands as CDSs, and for two of the four genomes showed the lowest percentage of functional new identified CDSs. Finally, the identification of non-coding RNAs may extend the information about the genomic features of many bacteria, resulting in a better understanding of its biology.

Genix can be accessed through the URL http://labbioinfo.ufpel.edu.br/genix/, and the source code of the genome annotation pipeline is available at the GitHub https://github.com/fredericokremer/genix. We have also uploaded the result of the comparison to the other annotation pipelines to our server and they can be accessed from the URL http://labbioinfo.ufpel.edu.br/genix_benchmarking/.

SUPPLEMENTARY DATA

Supplementary data are available at *FEMSLE* online.

ACKNOWLEDGEMENTS

We are grateful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil), Coordenação de Aperfeiçoamento do Pessoal de Nível Superior (CAPES, Brazil) and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS, Brazil) for providing scholarships for FSK and MRE.

Conflict of interest. None declared.

REFERENCES

Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat Methods* 2010;**8**:61–5.

Altschul SF, Gish W, Miller W *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.

Assefa S, Keane TM, Otto TD *et al*. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* 2009;**25**:1968–9.

Aziz RK, Bartels D, Best AA *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008;**9**:75.

Bankevich A, Nurk S, Antipov D *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77.

Beckloff N, Starkenburg S, Freitas T *et al*. Bacterial genome annotation. *Methods Mol Biol* 2012;**881**:471–503.

Boetzer M, Henkel CV, Jansen HJ *et al*. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 2011;**27**:578–9.

Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. *Genome Biol* 2012;**13**:R56.

Boisvert S, Laviolette F, Corbeil J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol* 2010;**17**:1519–33.

Borodovsky M, Mills R, Besemer J *et al*. Prokaryotic gene prediction using GeneMark and GeneMark.hmm. *Curr Protoc Bioinformatics* 2003, 4.5.1–4.5.16, DOI: 10.1002/0471250953.bi0405s01.

Camacho C, Coulouris G, Avagyan V *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.

Chaudhuri RR, Pallen MJ. xBASE, a collection of online databases for bacterial comparative genomics. *Nucleic Acids Res* 2006;**34**:D335–7.

Dayarian A, Michael TP, Sengupta AM. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics* 2010;**11**:345.

Delcher AL, Harmon D, Kasif S *et al*. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999;**27**: 4636–41.

Durbin R, Eddy SR, Krogh A *et al*. *Biological Sequence Analysis*. Cambridge, UK: Cambridge University Press, 1998.

Eberhardt RY, Haft DH, Punta M *et al*. AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database (Oxford)* 2012;**2012**:bas003.

Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;**7**:e1002195.

Galardini M, Biondi EG, Bazzicalupo M *et al*. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol Med* 2011;**6**:11.

Garst AD, Edwards AL, Batey RT. Riboswitches: structures and mechanisms. *Cold Spring Harb Perspect Biol* 2011;**3**:a003533.

Griffiths-Jones S, Bateman A, Marshall M *et al*. Rfam: an RNA family database. *Nucleic Acids Res* 2003;**31**:439–41.

Hayes CS, Keiler KC. Beyond ribosome rescue: tmRNA and co-translational processes. *FEBS Lett* 2010;**584**:413–9.

Hyatt D, Chen G-L, Locascio PF *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;**11**:119.

Johnson ZI, Chisholm SW. Properties of overlapping genes are conserved across microbial genomes. *Genome Res* 2004; **14**:2268–72.

Kisand V, Lettieri T. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. *BMC Genomics* 2013;**14**:211.

Köser CU, Ellington MJ, Peacock SJ. Whole-genome sequencing to control antimicrobial resistance. *Trends Genet* 2014;**30**: 401–7.

Lagesen K, Hallin P, Rødland EA *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007;**35**:3100–8.

Land M, Hauser L, Jun S-R *et al*. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomic* 2015;**15**:141–61.

Laslett D. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* 2004;**32**: 11–6.

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.

Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997;**25**:955–64.

Lugli GA, Milani C, Mancabelli L *et al*. MEGAnnotator: a user-friendly pipeline for microbial genomes assembly and annotation. *FEMS Microbiol Lett* 2016;**363**, DOI: 10.1093/femsle/fnw049.

Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genom Hum G* 2008;**9**:387–402.

Metzker ML. Emerging technologies in DNA sequencing. *Genome Res* 2005;**15**:1767–76.

Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 2009;**25**:1335–7.

Pareja-Tobes P, Manrique M, Pareja-Tobes E *et al*. BG7: a new approach for bacterial genome annotation designed for next generation sequencing data. *PLoS One* 2012;**7**:e49239.

Piro VC, Faoro H, Weiss VA *et al*. FGAP: an automated gap closing tool. *BMC Res Notes* 2014;**7**:371.

Repoila F, Darfeuille F. Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biol Cell* 2009;**101**:117–31.

Rissman AI, Mau B, Biehl BS *et al*. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* 2009; **25**:2071–3.

Salipante SJ, SenGupta DJ, Cummings LA *et al*. Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *J Clin Microbiol* 2015;**53**:1072–9.

Sallet E, Gouzy J, Schiex T. EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics* 2014;**30**:2659–61.

Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;**30**:2068–9.

Sette A, Rappuoli R. Reverse vaccinology: developing vaccines in the era of genomics. *Immunity* 2010;**33**:530–41.

Simpson JT, Wong K, Jackman SD *et al*. ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009;**19**: 1117–23.

Skinner ME, Uzilov AV, Stein LD *et al*. JBrowse: a next-generation genome browser. *Genome Res* 2009;**19**:1630–8.

Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* 2010;**11**:R41.

Van Domselaar GH, Stothard P, Shrivastava S *et al*. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res* 2005;**33**:W455–9.

Vitreschak AG, Rodionov DA, Mironov AA *et al*. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet* 2004;**20**:44–50.

Walter NG, Engelke DR. Ribozymes: catalytic RNAs that cut things, make things, and do odd and useful jobs. *Biologist (London)* 2002;**49**:199–203.

Wang JY, Qiu L, Wu ED *et al*. RNases involved in ribozyme degradation in Escherichia coli. *J Bacteriol* 1996;**178**:1640–5.

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;**18**: 821–9.

Anexo D – Artigo de revisão de literatura descrevendo ferramentas e metodologias para finalização de genomas publicada durante o período do Doutorado. Muitas destas estratégias foram utilizadas no artigo 3 desta tese.

Periódico: Genetics and Molecular Biology (Qualis na área de Biotecnologia: B1)

Genetics and Molecular Biology, 40, 3, 553-576 (2017)

DOI: http://dx.doi.org/10.1590/1678-4685-GMB-2016-0230

Review Article

Approaches for *in silico* finishing of microbial genome sequences

Frederico Schmitt Kremer[1], Alan John Alexander McBride[1] and Luciano da Silva Pinto[1]

[1]*Programa de Pós-Graduação em Biotecnologia (PPGB), Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, Brazil.*

Abstract

The introduction of next-generation sequencing (NGS) had a significant effect on the availability of genomic information, leading to an increase in the number of sequenced genomes from a large spectrum of organisms. Unfortunately, due to the limitations implied by the short-read sequencing platforms, most of these newly sequenced genomes remained as "drafts", incomplete representations of the whole genetic content. The previous genome sequencing studies indicated that finishing a genome sequenced by NGS, even bacteria, may require additional sequencing to fill the gaps, making the entire process very expensive. As such, several *in silico* approaches have been developed to optimize the genome assemblies and facilitate the finishing process. The present review aims to explore some free (open source, in many cases) tools that are available to facilitate genome finishing.

Keywords: microbial genetics, molecular microbiology, genomics, microbiology, draft genomes.

Received: September 25, 2016; Accepted: March 13, 2017.

Introduction

The advent of second generation of sequencing platforms, usually referred to as Next Generation Sequencing (NGS) technologies, promoted an expressive growth in the availability of genomic data in public databases, mainly due to the drastic reduction in the cost-per-base (Chain *et al.*, 2009). Compared to the Sanger sequencing technique (Sanger *et al.*, 1977), NGS platforms, like Illumina HiSeq, IonTorrent PGM, Roche 454 FLX, and ABI SOLiD, are able to generate a significantly higher throughput, resulting in a high sequencing coverage. However, they typically have a lower accuracy (in terms of average Phred-score in the raw data) and, in most cases, can only generate short-length reads (*short-reads*) (Liu *et al.*, 2012). The term "short-read" is commonly used to refer to the data generated by platforms such as Illumina, IonTorrent and SOLiD, as the length of their reads usually range from 30 bp (eg: SOLiD) to ~120 bp (*e.g.*, Illumina HiSeq), which is smaller than the length usually obtained by Sanger sequencing (~1 kb), and by the PacBio and Oxford Nanopore platforms (commonly referred to as "long-reads").

The higher throughtput obtained by NGS has stimulated the development of new algorithms and tools that are capable of dealing with a larger volume of data and generating genomic assemblies in a reasonable time. Traditional sequence assemblers, like Phrap (http://www.phrap.org/phredphrap/) and CAP3 (Huang and Madan, 1999), were replaced with new ones, such as Velvet (Zerbino and Birney, 2008), ABySS (Simpson *et al.*, 2009), Ray (Boisvert *et al.*, 2010), SPAdes (Bankevich *et al.*, 2012) and SOAPdenovo (Luo *et al.*, 2012), many of which were based on the *De Bruijn* graph algorithm (Pevzner *et al.*, 2001; Compeau *et al.*, 2011). A large number of these software offerings are capable of assembling data from different sequencing platforms, called "hybrid assembly"; however, all of them exhibit some limitations and, even after several corrections and optimization steps, the generation of a finished genome continues to represent a complex task (Alkan *et al.*, 2011).

Genome finishing is achieved by converting a set of contigs or scaffolds into complete sequences that represent the full genomic content of the organism without unknown regions (gaps) (Mardis *et al.*, 2002), and aims a "closed" and full representation of the chromosome organization. On the contrary, a draft genome may be composed of a set of a few or several contigs or scaffolds (Land *et al.*, 2015). Although useful for many studies, the unfinished and fragmented nature of draft genomes may difficult analysis on comparative genomics and structural genomics (Ricker *et al.*, 2012). In addition, some genes may be missed if located in a region without coverage (*e.g.*, edges of contigs/scaffolds), or due to assembly errors (*e.g.*, repetitive regions that are "collapsed" into a single one) (Klassen and Currie, 2012).

Genome finishing represents a relevant step to reduce data loss and lead to a more accurate representation of the genomics features of the organism of interest. The tradi-

Send correspondence to Frederico Schmitt Kremer. Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Campus Universitário, S/N, CEP 96160-000, Capão do Leão, RS, Brazil. E-mail: fred.s.kremer@gmail.com.

tional way to close gapped genomes includes: (1) design primers based on the edge of adjacent contigs, (2) PCR amplification, (3) Sanger sequencing, and (4) local assembly, usually with (5) manual curation. As this process is very time consuming, new *in vitro* approaches were developed to speed it up, including multiplex PCR (Tettelin *et al.*, 1999), optical mapping (Latreille *et al.*, 2007), and hybrid sequencing and assembly (Ribeiro *et al.*, 2012). This, however, usually results in a drastic increase in the cost of the sequencing project. Therefore it is not surprising that from the 87,956 prokaryote genomes available in GenBank until December 2016 (GenBank release 217), only a small fraction (~6,586) is finished (http://www.ncbi.nlm.nih.gov/genbank/).

In the last years, the availability of third-generation sequencing technologies, such as PacBio SMRT and Oxford Nanopore, has also provided another way to achieve finished genomes (Ribeiro *et al.*, 2012). As these platforms usually generate reads with length > 10 kb, the assembly algorithms have to deal with less ambiguities and problematic regions (Koren and Phillippy, 2015). This makes the reconstruction of the chromosome sequences easier, but just like second generation technologies, both platforms have their own limitations. In the earlier versions of the PacBio SMRT platform, for example, it used to present a high error-rate, so it was recommended to correct part of these errors using short-reads data (*e.g.*, Illumina) (Koren *et al.*, 2012; Au *et al.*, 2012; Salmela and Rivals, 2014) before using its result for any analysis. Although this problem has been minimized with the improvements in the sequencing chemistry and base-calling process over the last years, the time required for the sequencing, the cost of each run, and the price of the equipment itself are still drawbacks, and these are some of reasons as to why it is not more frequently applied. In the case of Oxford Nanopore, there is a limited number of tools that can be used to pre-process and analyze its data, and some types of errors (*e.g.*, under-representation of specific k-mers) are still recurrent (Deschamps *et al.*, 2016). Therefore, second-generation platforms are still the most popular ones for a wide variety of applications, and remain as the cheaper alternative to obtain genomic data.

Previous reviews that focused on genome assembly and whole-genome sequencing analysis have already described some of the *in silico* tools that have the potential to improve a genome assembly without the need for experimental data. Edwards and Holt (2013), Dark (2013) and Vincent *et al.* (2016) have provided very comprehensive descriptions of some of the main steps of data analysis in microbial genomes sequenced by next-generation technologies, including *de novo* assembly, reference-based contig ordering, genome annotation and comparative genomics, and these are useful starting points for those that are new to microbial genomics with NGS. Nagarajan *et al.* (2010) have also made some useful considerations regarding genome assembly and presented some methodologies for genome finishing, but limited to a small set of approaches that they have developed to improve assemblies. Ramos *et al.* (2013) exemplified some assembly and post-assembly methods for genomes sequenced with IonTorrent platforms. Finally, Hunt *et al.* (2014) have also made a complete and accurate benchmark analysis of different tools for assembly scaffolding based on paired-end/mate-pair data. However, although very instructive, these reviews provide descriptions of some specific procedures, but do not give an in-depth view of the different finishing strategies, nor of the algorithms that are used in background by different tools from each category. Therefore, a more complete description of these tools may be useful, especially for those researchers that are starting to work with microbial genome analysis and want to achieve better assemblies without relying on re-sequencing or manual gap-closing.

The present review aims at discussing some of the categories of software that may be applied in the process of assembly finishing, especially in the context of microbial genome sequencing projects. A particular focus is placed on those that do not require additional experimental and/or new sequencing data. The review intentionally focuses on tools that are freely available, at least for academic use, and omits proprietary software, like the CLC Genomics Workbench (http://www.clcbio.com/), for example. This was not due to anticipated differences in the performance or quality of the results, but to adhere to the original intention: an overview of tools that could be used by most researchers. As different approaches have been developed to improve genomic assemblies, the description of these programs was divided into four main categories: *scaffolding* (placement of contigs into larger sequences by using experimental data or information for closely-related genomes, and joining them by gap regions)*, assembly integration* (generation of a consensus assembly using multiple assemblies for the same genome), *gap closing* (solving gaps by identifying their correct sequence), *error correction* (removal of misidentified bases or misassembled regions) and *assembly evaluation* (quantification of the reliability of a genome assembly and identification of its erroneous regions) (Table 1). These different categories of programs can be combined, according to the type of data that is available (*e.g.*, sequencing platform and library that was used), and may help to reduce the fragmentation and improve the reliability of a genome assembly. Based on the categories of software that are discussed in the present review we have created the flowchart showed in Figure 1, which may of help in choosing the most appropriate approach for genome finishing, depending on the type of the data that is available. Examples of genome projects that used some of the tools discussed in the present review are shown in Supplementary material (Table S1).

Table 1 - Overview of the tools described in the present review

Category	Tool	Main features	Dependences*	Reference	Download link / webserver
Scaffolding	ABySS	- Paired-end scaffolding. - Scaffolding feature already integrated in the ABySS *de novo* assembly pipeline. - Uses the estimated distances generated by the program DistanceEst (from the same package) as input. - Allows the scaffolding using long-reads, such as those generated by PacBio and Oxford Nanopore platforms.	boost libraries: www.boost.org/ Open MPI: http://www.open-mpi.org sparse-hash library: http://goog-sparsehash.sourceforge.net/	(Simpson *et al.,* 2009)	http://www.bcgsc.ca/platform/bioinfo/software/abyss
Scaffolding	Bambus 2	- Paired-end scaffolding. - Can be easily integrated with assembly projects that are built on top of the AMOS package. - Supports the scaffolding of metagenomes. - Requires experience with the AMOS package and its data formats.	AMOS package (Treangen *et al.,* 2011): http://amos.sourceforge.net/	(Koren *et al.,* 2011)	https://sourceforge.net/projects/amos/
Scaffolding	MIP	- Paired-end scaffolding. - Supports both paired-end and mate-pair (long range) reads.	lpsolve library: http://sourceforge.net/projects/lpsolve/ lemon library: http://lemon.cs.elte.hu/	(Salmela *et al.,* 2011)	https://www.cs.helsinki.fi/u/lmsalmel/mip-scaffolder/
Scaffolding	OPERA	- Paired-end scaffolding. - Identifies potential spurious connections caused by chimeric reads and repetitive genomics elements that may affect the reliability of the scaffolding. - Contigs identified as misassembled may be used in the construction of more than one scaffold, but sometimes it may lead to new assembly errors.	BWA (Li and Durbin 2009): http://bio-bwa.sourceforge.net/ Bowtie (Langmead *et al.,* 2009): http://bowtie-bio.sourceforge.net/ Samtools (Li *et al.,* 2009): http://samtools.sourceforge.net/	(Gao *et al.,* 2011)	https://sourceforge.net/projects/operasf
Scaffolding	SCARPA	- Paired-end scaffolding. - Only uses for scaffolding those contigs with length greater than the N50 of the assembly. - Allows multiple libraries to be used in the same scaffolding project.	None	(Donmez and Brudno, 2013)	http://compbio.cs.toronto.edu/hapsembler/scarpa.html
Scaffolding	SGA	- Paired-end scaffolding. - Scaffolding feature already integrated in the SGA assembly pipeline, which is optimized for Illumina data and large genomes. - Uses the estimated distances generated by the program DistanceEst (from the package ABySS) as input, along with the read mapping file in .BAM format. - Allows multiple libraries to be used in the same scaffolding project.	Bamtools (Barnett *et al.,* 2011): https://github.com/pezmaster31/bamtools BWA (Li and Durbin, 2009): http://bio-bwa.sourceforge.net/ Samtools (Li *et al.,* 2009): http://samtools.sourceforge.net/ Sparse-hash library: http://goog-sparsehash.sourceforge.net/	(Simpson and Durbin, 2012)	https://github.com/jts/sga

Table 1 - continued

Category	Tool	Main features	Dependences*	Reference	Download link / webserver
Scaffolding	SOPRA	- Paired-end scaffolding. - Developed to improve the assemblies generated by Velvet and SSAKE, and required the .AFG files. - Supports data from early Illumina and ABI SOLiD platforms, including paired-end and mate-pair reads. - Is not fully automated, so it is necessary to run different scripts for each step of the scaffolding.	None	(Dayarian *et al*., 2010)	http://www.physics.rutgers.edu/~anirvans/SOPRA/
Scaffolding	SSPACE	- Paired-end scaffolding. - Trims the edge of the contigs as they are more suitable to assembly errors. - Requires information about the paired-end library, including mean size of the insert, standard deviation and the relative orientation of the mates.	None	(Boetzer *et al.,* 2011)	http://www.baseclear.com/genomics/bioinformatics/basetools/
Scaffolding	SSPACE-LongRead	- Paired-end scaffolding. - Allows the scaffolding using long-reads, such as those generated by PacBio and Oxford Nanopore platforms.	None	(Boetzer and Pirovano, 2014)	http://www.baseclear.com/genomics/bioinformatics/basetools/
Scaffolding	MUMmer	- Single reference-based scaffolding. - The result of the alignment must be post-processed to obtain the scaffolds.		(Kurtz *et al.,* 2004)	http://mummer.sourceforge.net/
Scaffolding	ABACAS	- Single reference-based scaffolding. - Useful when the reference and the target genome are closely-related, and the genome to be scaffolded is not larger than the reference genome. - Not optimized for bacteria with two or more replicons/chromosomes (ex: *Leptospira* genus). - Allows the design of primers for gap-closing.	MUMmer (Kurtz *et al.,* 2004): http://mummer.sourceforge.net/ Primer3 (Koressaar and Remm, 2007; Untergasser *et al.,* 2012): http://primer3.ut.ee/	(Assefa *et al.*, 2009)	http://abacas.sourceforge.net/
Scaffolding	CONTIGuator	-Single reference-based scaffolding. -Useful when the target genome is composed by more than one chromosome / replicon. - Allows a more sensitive identification of syntenic regions, if compared to ABACAS, as it applies a BLAST search after MUMmmer.	ABACAS (Assefa *et al.*, 2009): http://abacas.sourceforge.net/ BioPython (Python package): http://biopython.org/ BLAST+ (Altschul *et al.,* 1990; Camacho *et al.,* 2009): ftp://ftp.ncbi.nlm.nih.gov/blast/ MUMmer (Kurtz *et al.,* 2004): http://mummer.sourceforge.net/ Primer3 (Koressaar and Remm, 2007; Untergasser *et al.,* 2012): http://primer3.ut.ee/	(Galardini *et al.,* 2011)	http://contiguator.sourceforge.net/

Table 1 - continued

Category	Tool	Main features	Dependences*	Reference	Download link / webserver
Scaffolding	Mauve	- Single reference- based scaffolding. - Can be used both through a commandline interface (CLI) and a graphical user interface (GUI). - Allows the identification of genomic inversions and translocations. - Not optimized for bacteria with two or more replicons/chromosomes.	Java: https://www.java.com/	(Darling *et al*., 2004; Rissman *et al.,* 2009)	http://darlinglab.org/mauve/mauve.html
Scaffolding	FillScaffolds	- Single reference- based scaffolding. - Not optimized for bacteria with two or more replicons/chromosomes. - Results may require post-processing to reconstruct the sequence of the scaffold.	Java: https://www.java.com/	(Muñoz *et al.,* 2010)	Supplementary data of Muñoz *et al.* (2010). http://dx.doi.org/10.1186/1471-2105-11-304
Scaffolding	SIS	- Single reference-based scaffolding. - Allows the identification of genomic inversions. - Not optimized for bacteria with two or more replicons/chromosomes.	MUMmer (Kurtz *et al*., 2004): http://mummer.sourceforge.net/	(Dias *et al*., 2012)	http://marte.ic.unicamp.br:8747.
Scaffolding	CAR	- Single reference-based scaffolding. - Allows the identification of genomic inversions and translocations. - Also available as a webserver. - Not optimized for bacteria with two or more replicons/chromosomes.	MUMmer (Kurtz *et al*., 2004): http://mummer.sourceforge.net/ PHP: https://php.net/	(Lu *et al.,* 2014)	http://genome.cs.nthu.edu.tw/CAR/
Scaffolding	RACA	- Multiple reference-based scaffolding. - Optimized for large genomes and with multiple chromosomes. - Can also use paired-end data.	None	(Kim *et al.,* 2013):	http://bioen-compbio.bioen.illinois.edu/RACA/
Scaffolding	Ragout	- Multiple reference-based scaffolding. - Uses phylogenetic information to identify the most probable orientation of the contigs / scaffolds.	Networkx (Python package): http://networkx.github.io/ Newick (Python package): http://www.daimi.au.dk/~mailund/newick.html Sibelia: http://github.com/bioinf/Sibelia	(Kolmogorov *et al.,* 2014)	https://github.com/fenderglass/Ragout
Scaffolding	MeDuSa	- Multiple reference-based scaffolding. - Accepts both finished and draft genomes as reference.	BioPython (Python package): http://biopython.org/ Java: https://www.java.com/ MUMmer (Kurtz *et al*., 2004): http://mummer.sourceforge.net/	(Bosi *et al.,* 2015)	https://github.com/combogenomics/medusa
Assembly integration	Minimus	- Can be easily integrated with assembly projects that are built on top of the AMOS package. - Requires experience with the AMOS package and its data formats.	AMOS package (Treangen *et al.,* 2011): http://amos.sourceforge.net/	(Sommer *et al.,* 2007)	https://sourceforge.net/projects/amos/

Table 1 - continued

Category	Tool	Main features	Dependences*	Reference	Download link / webserver
Assembly integration	Reconciliator	- Corrects the misassembled regions in a target assembly by comparing to an alternative assembly for the same genome. - Identifies repetitive regions that suffered compressions or expansions.	MUMmer (Kurtz *et al.,* 2004): http://mummer.sourceforge.net/	(Zimin *et al.,* 2008)	http://www.genome.umd.edu/
Assembly integration	MAIA	- Allows the integration of two or more assemblies. - Accepts reference genome to perform scaffolding, what is useful for those contigs without correspondence in the other assemblies.	Matlab: https://www.mathworks.com/ MUMmer: http://mummer.sourceforge.net/ GAIMC (Matlab toolbox): http://github.com/dgleich/gaimc	(Nijkamp *et al.,* 2010)	http://bioinformatics.tudelft.nl
Assembly integration	CISA	- Allows the integration of three or more assemblies. - Corrects misassembled regions and compressed / expanded repeated regions.	BLAST+ (Altschul *et al*., 1990; Camacho *et al.,* 2009): ftp://ftp.ncbi.nlm.nih.gov/blast/ MUMmer (Kurtz *et al.,* 2004): http://mummer.sourceforge.net/	(Lin and Liao, 2013)	http://sb.nhri.org.tw/CISA/
Assembly integration	GAA	- Uses the alignment between the different contigs in the set of assemblies to generate an assembly graph, which is explored to identify to minimal set of independent paths.	BLAT (Kent, 2002): https://genome.ucsc.edu/ GSMapper: http://454.com/	(Yao *et al.,* 2012)	http://sourceforge.net/projects/gaa-wugi/
Assembly integration	Mix	- Generate an extension graph that represents the connection between the contigs. - Filters the alignment to reduce the ambiguities caused by repetitive sequences.	Networkx (Python package): http://networkx.lanl.gov/ BioPython (Python package): http://biopython.org/ MUMmer(Kurtz *et al.,* 2004): http://mummer.sourceforge.net/	(Soueidan *et al.,* 2013)	https://github.com/cbib/MIX
Assembly integration	GAM / GAM-NGS	- Requires the read files to perform the assembly integration. - One of the assemblies to be merged is defined as "master", while the others are defined as "slaves". - Allows the identification of misassembled regions in the master, which are corrected before the generation of the final assembly.	cmake: https://cmake.org/ zlib library: http://www.zlib.net/ boost libraries: www.boost.org/ sparse-hash library: http://goog-sparsehash.sourceforge.net/	(Casagrande *et al.,* 2009; Vicedomini *et al.,* 2013)	https://github.com/vice87/gam-ngs
Assembly integration	Zorro	- Requires the read files to perform the assembly integration. - Remaps the reads back to the contigs and identifies misassembled and repetitive regions based on the coverage. - Splits the misassembled contigs and performs the assembly integration using Minimus.	AMOS (Treangen *et al.,* 2011): http://amos.sourceforge.net/ BioPerl (Perl module): http://bioperl.org Bowtie (Langmead *et al.,* 2009): http://bowtie-bio.sourceforge.net/ MUMmer (Kurtz *et al.,* 2004): http://mummer.sourceforge.net/	(Argueso *et al.,* 2009)	http://lge.ibi.unicamp.br/zorro/

Table 1 - continued

Category	Tool	Main features	Dependences*	Reference	Download link / webserver
Gap closing	GapCloser	- Gap-closing feature already integrated in the SOAPdenovo *de novo* assembly pipeline - Performs a local reassembly in the gap region using the reads located in the edges of the surrounding contigs.	None	(Li *et al.,* 2010)	http://soap.genomics.org.cn/
Gap closing	IMAGE	- Iteratively performs a remapping of the reads to the contigs, followed by the selection of those that overlap the gap region and a local reassembly.	None	(Tsai *et al.,* 2010)	https://sourceforge.net/projects/image2
Gap closing	GapFiller	- Iteratively performs a remapping of the reads to the contigs, followed by the selection of those that overlap the gap region and a local reassembly. - Requires information about the paired-end library, including mean size of the insert, its standard deviation and the relative orientation of the mates.	None	(Boetzer and Pirovano, 2012)	http://www.baseclear.com/genomics/bioinformatics/basetools
Gap closing	Enly	- Iteratively performs a remapping of the reads to the contigs, followed by the selection of those that overlap the gap region and a local reassembly. - If a reference genome is provided, a new scaffolding step can be performed to improve the assembly.	BioPython (Python package): http://biopython.org/ BLAST and BLAST+ (Altschul *et al.,* 1990; Camacho *et al*., 2009): ftp://ftp.ncbi.nlm.nih.gov/blast/ Cdbfasta/cdbyank: http://compbio.dfci.harvard.edu/tgi/software/ EMBOSS: http://emboss.sourceforge.net/ Minimo assembler (Treangen *et al.,* 2011): http://amos.sourceforge.net/ MUMmer (Kurtz *et al.,* 2004): http://mummer.sourceforge.net/ Phrap: http://www.phrap.org/phredphrapconsed.html	(Fondi *et al*., 2014)	http://enly.sourceforge.net/
Gap closing	FGAP	- Uses alternative assemblies of the target genome to identify regions that overlap the gap.	Matlab: https://www.mathworks.com/	(Piro *et al.,* 2014)	http://www.bioinfo.ufpr.br/fgap/
Gap closing	Sealer	- Performs a local re-assembly of the gap regions using different settings of k-mer, what may help in the solving of regions with repetitive sequences.	boost libraries: www.boost.org/ sparse-hash library: http://goog-sparsehash.sourceforge.net/ Open MPI: http://www.open-mpi.org	(Paulino *et al*., 2015)	https://github.com/bcgsc/abyss/tree/sealer-release

Table 1 - continued

Category	Tool	Main features	Dependences*	Reference	Download link / webserver
Gap closing	GMCLoser	- May use both paired-end reads and alternative assemblies to perform the gap-closing. - Applies a likelihood analysis to avoid the effect of misassemblies in the alternative assemblies.	MUMmer (Kurtz *et al.* 2004): http://mummer.sourceforge.net/ BLAST+ (Altschul *et al*., 1990; Camacho *et al*., 2009): ftp://ftp.ncbi.nlm.nih.gov/blast/ Bowtie (Langmead *et al*., 2009): http://bowtie-bio.sourceforge.net/ YASS (Noé and Kucherov, 2005): http://bioinfo.lifl.fr/yass	(Kosugi *et al.,* 2015)	https://sourceforge.net/projects/gmcloser/
Gap closing	MapRepeat	- Performs a reference-based scaffolding using a closely-related genome provided by the user. - Uses a reference-guided assembly to perform the gap-closing process.	BLAST+ (Altschul *et al.,* 1990; Camacho *et al*., 2009): ftp://ftp.ncbi.nlm.nih.gov/blast/ BioPython (Python package): http://biopython.org/ MIRA: http://mira-assembler.sourceforge.net MUMmer (Kurtz *et al*., 2004): http://mummer.sourceforge.net/	(Mariano *et al*., 2015)	http://github.com/dcbmariano/maprepeat
Gap closing	GapBlaster	- Allows a manual gap-closing using an alternative assembly of the target genome.	BLAST and BLAST+ (Altschul *et al*., 1990; Camacho *et al.,* 2009): ftp://ftp.ncbi.nlm.nih.gov/blast/ MUMmer (Kurtz *et al*., 2004): http://mummer.sourceforge.net/	(de Sá *et al*., 2016)	https://sourceforge.net/projects/gapblaster2015/
Assembly evaluation	REAPR	- Calculates the accuracy of the assembly based on the coverage after remapping the reads back to the scaffolds. - Misassembled regions can be identified as they usually present a discrepant coverage. - A new set of scaffolds is generated by splitting the regions identified as misassembled.	File::Basename, File::Copy, File::Spec, File::Spec::Link, Getopt::Long and List::Util (Perl modules): http://www.cpan.org/ R: https://www.r-project.org/	(Hunt *et al.,* 2013)	http://www.sanger.ac.uk/science/tools/reapr
Assembly evaluation	QUAST	- Calculate several assembly metrics, such as C+G%, N50 and L50. - Can be used to compare different assemblies for the same genome, and / or compare then to a reference genome.	boost libraries: www.boost.org/ Java: https://www.java.com/ Matplotlib (Python package): http://matplotlib.org Time::HiRes (Perl module): http://www.cpan.org/	(Gurevich *et al*., 2013)	http://bioinf.spbau.ru/quast

Table 1 - continued

Category	Tool	Main features	Dependences*	Reference	Download link / webserver
Assembly evaluation	ALE	- Calculates the accuracy of the assembly based on the k-mers and C+G% distribution along the scaffolds. - Doesn't require a reference genome.	Matplotlib (Python package): http://matplotlib.org Mpmath (Python package): http://mpmath.org Numpy (Python package): http://www.numpy.org Pymix (Python package): http://www.pymix.org/pymix Setuptools (Python package): https://github.com/pypa/setuptools	(Clark *et al.,* 2013)	http://www.alescore.org
Assembly evaluation	CGAL	- Calculates the accuracy of the assembly based on the coverage after remapping the reads back to the scaffolds.	None	(Rahman and Pachter, 2013)	http://bio.math.berkeley.edu/cgal/
Assembly evaluation	GMvalue	- Aligns the assembly to a reference genome (or alternative assembly) to identify misassembled regions. - A new set of scaffolds is generated by splitting the regions identified as misassembled.	MUMmer (Kurtz *et al.,* 2004): http://mummer.sourceforge.net/ BLAST+ (Altschul *et al.,* 1990; Camacho *et al*., 2009): ftp://ftp.ncbi.nlm.nih.gov/blast/ Bowtie (Langmead *et al*., 2009): http://bowtie-bio.sourceforge.net/ YASS (Noé and Kucherov, 2005): http://bioinfo.lifl.fr/yass	(Kosugi *et al.,* 2015)	https://sourceforge.net/projects/gmcloser/
Assembly correction	iCORN	- Requires paired-end reads. - Interactively identifies and corrects short misassemblies, such as base-substitutions and short INDELs.	SNP-o-matic (Manske and Kwiatkowski, 2009): https://snpomatic.svn.sourceforge.net/svnroot/snpomatic SSAHA Pileup (Ning *et al*., 2001): ftp://ftp.sanger.ac.uk/pub/zn1/ssaha_pileup/	(Otto *et al.,* 2010)	http://icorn.sourceforge.net/
Assembly correction	SEQuel	- Requires paired-end reads. - Interactively identifies and corrects short misassemblies, such as base-substitutions and short INDELs. - Performs a local reassembly of the misassembled regions using information from k-mers and paired-end reads.	Java: https://www.java.com/ JGraphT (Java library): http://jgrapht.org/	(Ronen *et al*., 2012)	http://bix.ucsd.edu/SEQuel/
Assembly correction	GFinisher	- Doesn't require paired-end reads. - Integrates a reference-guided scaffolding step and gap-closing procedures, along with the assembly correction process. - Identifies misassembled regions based on the GC-Skew distribution.	Java: https://www.java.com/	(Guizelini *et al.,* 2016)	http://gfinisher.sourceforge.net/

* = Considering a computer running UNIX, Linux or Mac OS operating systems (OSs). As Make, sed, awk, GCC, Perl, Bash, Python and the GNU/Unix standard utility set are already included in most of the distributions / versions of these OSs, these programs were not listed as dependences.

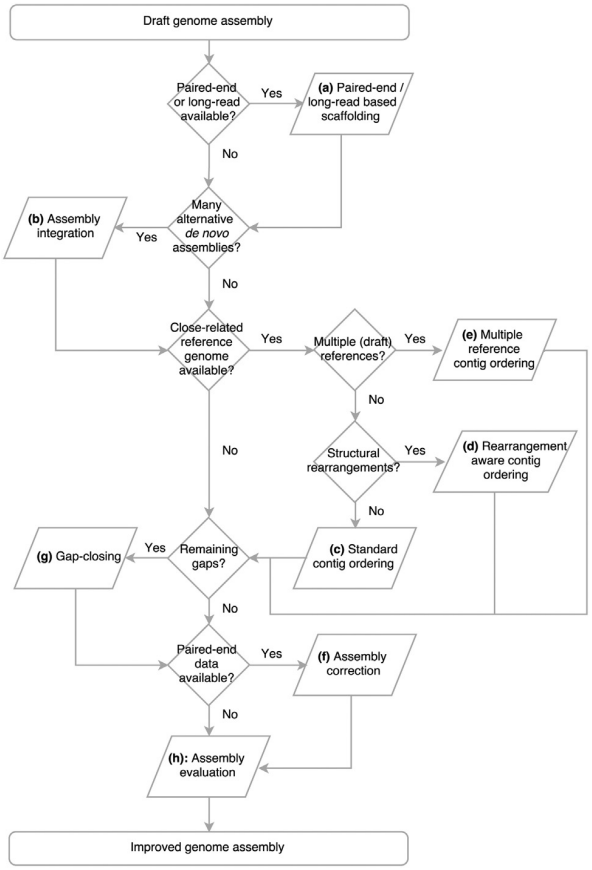


Figure 1 - A flowchart demonstrating how and when the different genome finishing approaches can be combined according to the data that is available for the user. (a) *Scaffolding using paired-end reads or long-reads*, which is directly dependent on the way the genome was sequencing (platform, library), and sometimes performed as part of the *de novo* assembly process. (b) *Assembly integration*, which consists in the combination of different *de novo* assemblies and generation of a consensus/extended assembly. Some programs use only the assemblies as input, while others use also the sequencing reads. (c) The *standard contig-ordering* approach based on a single reference genome, which consists in the identification of synteny blocks that guide the orientation of the contigs in the draft genome, without taking into count the occurrence of genome inversions other rearrangements. (d) The *rearrangement-aware contig-ordering*, that identifies potential sites of inversion and translocations based on signatures on the alignment against the reference genome. (e) *The multiple-reference contig ordering*, that may be more appropriate in those cases where there is no finished reference genome, but there is a relatively high number of close-related drafts, or when there are no apparent closest reference to be used. (f) *Assembly correction*, which consists in the removing of short misassemblies, including base-substitutions and short insertions and deletions. (g) *Gap-closing*, which consists in the joining of adjacent contigs that used to be spaced by a gap. (h) *Assembly evaluation*, which may provide help to access the reliability of the assembly.

Scaffolding

By definition, a contig consists of a contiguous sequence has no unknown regions or assembly gaps (but may contain "N" that represent base-calling errors) (Staden, 1979). On the other hand, a scaffold consists of two or more contigs that have been joined according to some linkage information (*e.g.*, paired-end reads, genome maps) (Huson *et al.*, 2002). Paired-end or mate-pair libraries can be very useful in *de novo* genome assembly, and several tools use the relative position information to connect contigs into scaffolds (Hunt *et al.*, 2014). In a similar way, with the increase in the availability of genomic sequences from a wide variety of species, other scaffolding alternatives were developed to use one or multiple genomes as reference to order the contigs.

Paired-end scaffolding

Most of the *de novo* genome assemblers usually integrate scaffolding steps after the contig constructions, although it is also possible to use third-party tools aiming a more reliable result. The A5 assembly pipeline (Tritt *et al.*, 2012), for example, uses *de novo* assembler IDBA (Peng *et al.*, 2010) to construct the contigs and SSPACE (Boetzer *et al.*, 2011) to generate scaffolds. The scaffolding with paired-end reads usually consist of the alignment of reads to the contigs, followed by the identification of connections between different contigs using the relative-orientation information and the estimated insert-size. ABySS (Simpson *et al.*, 2009), SOPRA (Dayarian *et al.*, 2010), SOAPdenovo (Li *et al.*, 2010), Bambus 2 (Koren *et al.*, 2011), MIP (Salmela *et al.*, 2011), Opera (Gao *et al.*, 2011), SSPACE (Boetzer *et al.*, 2011), SLIQ (Roy *et al.*, 2012), SGA (Simpson and Durbin 2012), SCARPA (Donmez and Brudno 2013), WiseScaffolder (Farrant *et al.*, 2015) and ScaffoldScaffolder (Bodily *et al.*, 2016) are examples of scaffolding tools based on paired-end information. More recently, the use of long-reads was also incorporated into scaffolding tools such as AHA (Bashir *et al.*, 2012) and SSPACE-LongRead (Boetzer and Pirovano, 2014).

ABySS (Simpson *et al.*, 2009)**:** The program abyss-scaffold, which comes with the ABySS assembly package (Simpson *et al.*, 2009), uses the estimated mate-distance distribution in paired-end reads to connect contigs and generate scaffolds. The distance distribution can be calculated by DistanceEst, that is also part of the package, and is also used by other assembly pipelines, such as SGA (Simpson and Durbin, 2012). ABySS was developed to be used both with small and large genomes, and can be executed in a computer clustering by using the Message Passing Interface (MPI), this being useful also in case of high-coverage data and when dealing with multiple libraries. Like most scaffolding programs, abyss-scaffold can be used also in the scaffolding of contigs generated by third-party programs. Finally, ABySS also supports scaffolding with long-reads by using BWA-MEM for read alignment (Li and Durbin, 2009). The source code of the ABySS package is available at the address http://www.bcgsc.ca/platform/bioinfo/software/abyss, and is developed to work on the Linux operating system.

SOPRA (Dayarian *et al.*, 2010): This scaffolding tool was designed to improve assemblies generated by Velvet (Zerbino and Birney, 2008) and SSAKE (Warren *et al.*, 2007), and targets the earlier sequencing platforms from Illumina and ABI SOLiD. The program parses the read-placing file generated by these assemblers and extracts information of paired-end/mate-pair reads, that is used to calculate the mean distance between pairs and the correct orientation. Based on this file, SOPRA also infers the connections between contigs by searches of those pairs of reads where mates are in different contigs. The program is not fully automated, so each step of data processing must be executed by a different script before the main scaffolding process. Another drawback is the limited support for different *de novo* assemblers, as it requires read-placing files in AFG format, and this is only produced by a few assemblers nowadays (*e.g.*, Velvet, Ray and AMOS). SOPRA can be obtained from the website http://www.physics.rutgers.edu/~anirvans/SOPRA/.

Bambus 2 (Koren *et al.*, 2011)**:** It is part of the AMOS package (Treangen *et al.*, 2011) and is both a genome and metagenome scaffolding tool and an updated version of the Sanger-based program Bambus targeting NGS data (Pop *et al.*, 2004). The program requires read-placing information to construct a contig-graph, and explore the graph to find consistent connections between the contigs. As Bambus2 can also be used to scaffold metagenomic assemblies, different from other programs, it considers the effect of DNA samples containing mixes of closely related organisms in the assembly processes and reduces the chance of fragmentation and miss-joining by analyzing the molecular variants. However, the use of Bambus 2 is not as simple as for other scaffolding tools, as it requires some experience with the AMOS tools to generate its input file and processes the outputs (Treangen *et al.*, 2011). The AMOS package can be obtained from the SourceForge repository: https://sourceforge.net/projects/amos/.

MIP (Salmela *et al.*, 2011)**:** uses the concept of mixed integer programming to generate a set of scaffolds from a genome assembly and a set paired-ends/mate-pair reads. First, readaligner (Mäkinen *et al.*, 2010) is used to map the read-pairs back to the contigs. Then, the pairs are filtered to remove inconsistent connections, and the distances between the contigs are estimated based on the mean distance between the mates, which is calculated for each library used in the assembly. The connections in the generated scaffold graph have a minimum and maximum estimated length, derived from the library information. The MIP source code, along with usage instructions, is available at https://www.cs.helsinki.fi/u/lmsalmel/mip-scaffolder/.

Opera (Gao *et al.*, 2011)**:** takes as inputs a collection of contigs and mapped reads and generates a scaffold graph based on the paired-end information. Frist, the program filters the connections between contigs to remove possible miss-joining errors caused by chimeric pairs. The graph is contracted, and the optimum orientation of the contigs inside the scaffolds is inferred by a dynamic programming algorithm that explores the search space. The algorithm can also infer the occurrence of repeated genomic regions, usually assembled into a single contig in case of short-reads. In this case, repeated regions are identified by comparing the coverage of the contigs to the mean coverage of the whole genome and selection those with value greater than 1.5 times the genomic mean. The identification of these regions allows a contig to be used in more than one scaffolds, which can provide a better assembly of repeated regions, but can also result in misassemblies. Opera can be obtained from its SourceForge repository https://sourceforge.net/projects/operasf/.

SGA (String Graph Assembler) (Simpson and Durbin 2012)**:** is a *de novo* genome assembler developed for the memory-efficient assembly of small and large genomes by applying the method proposed by Myers (2005). As part of its assembly pipeline, SGA also provide a scaffolding tool that uses information from read alignment (in .BAM format), that can be generated by a wide variety of mapping tools (Li *et al.*, 2008; Langmead *et al.*, 2009; Lunter and Goodson, 2011), and estimated distance between mates, generated by DistanceEst, from the ABySS package, to connect contigs into scaffolds. SGA also supports scaffolding from multiple libraries, with different insert sizes, and was optimized to work with Illumina data. SGA is available from the GitHub repository https://github.com/jts/sga.

SCARPA (Donmez and Brudno, 2013): uses paired-end information to generate scaffolds, but takes into account that not only chimeric reads may be to responsible for inconsistent linkages between contigs but also misassembled sequences. It estimates the mean and standard deviation of the distance between the mates, but only uses information from those contigs with length greater than the assembly N50. The connections between the contigs are estimated based on the mate information and the calculated metrics, and if more than one library is provided, SCARPA process the scaffolding iteratively starting from the library with smaller insertion size. The program can be obtained from the URL http://compbio.cs.toronto.edu/hapsembler/scarpa.html.

SSPACE (Boetzer *et al.*, 2011) and SSPACE-LongRead (Boetzer and Pirovano, 2014): these scaffolding programs are currently distributed by BaseClear (http://www.baseclear.com), which also distributes the gap-closing program GapFiller (Boetzer and Pirovano, 2012). SSPACE requires information about paired-end library, including mean and standard deviation of distance between the mates and the expected orientation, whose values can be predicted with the script "estimate_insert_size.pl", distributed along with the program. The user may choose between BWA (Li and Durbin, 2009) and Bowtie (Langmead *et al.*, 2009) for read mapping, the minimum number of connections to link two contigs, the num-

ber of bases that will be removed from the border of the contigs (as they usually contain errors), and the number of iterations. For SSPACE-LongRead, the target assembly is aligned to a collection of long-reads using BLASR (Chaisson and Tesler, 2012) and the alignments are filtered and refined to find the best orientation. Both SSPACE and SSPACE-LongRead can be requested from the BaseClear website http://www.baseclear.com/genomics/bioinformatics/basetools/.

Hunt *el al.* (2014) have performed an extensive comparison of the scaffolding tools and demonstrated that the quality of the resulting scaffolds is directly affected by the read-mapping program and the complexity of the genome. For the tested datasets, the best results were obtained by SGA, SOPRA and SSPACE, although all tested tools presented a certain percentage of miss-joined scaffolds in their outputs. Some scaffolding tools (*e.g.*, SGA) use pre-aligned reads as input, so the user is able to test and choose the read mapper. In this case, it is important to try different read mappers, taking into account that platform-specific bias and read quality may have a drastic effect on the quality of the alignment (Hatem *et al.*, 2013; Caboche *et al.*, 2014). When using mate-pair libraries (long-insert paired-end reads) it is also important to check if the scaffolder was designed to support it, or if it was just designed for standard, short-insert, paired-end reads. As mate-pairs may present a relatively high rate of "false mates", special care may be taken when working with this type of data.

Single reference-based scaffolding

In many cases the pairing information is not enough to generate a reliable reconstruction of the genome's structure, or simply, the genome was not sequenced using paired-reads, but with single-end sequencing. In order to overcome this, some tools were developed to use a reference genome as a template to perform the contig ordering and relative positioning (Figure 2). Software like MUMmer (Kurtz *et al.*, 2004), ABACAS (Assefa *et al.*, 2009), CONTIGuator (Galardini *et al.*, 2011) and Mauve (Darling *et al.*, 2004; Rissman *et al.*, 2009) are able to identify the most probable orientation of the contigs, but may generate incorrect results in the case of genome inversions of translocations. On the other hand, SIS (Dias *et al.*, 2012), CAR (Lu *et al.*, 2014), and FillScaffolds (Muñoz *et al.*, 2010) consider the occurrence of changes in the genomic structure and take these phenomena into account during the

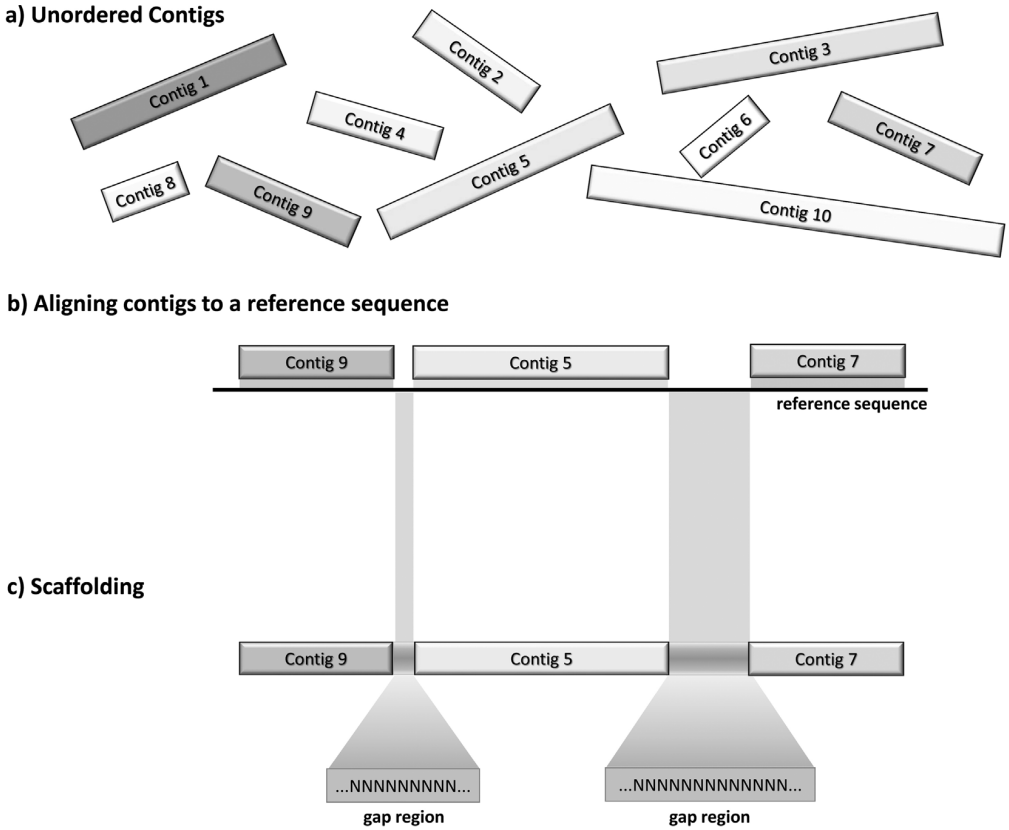


Figure 2 - Reference-based contig ordering. (a) The program takes a set of contigs (or scaffolds) and (b) aligns these to a reference genome to identify the most probable relative orientation of the sequences in the draft genome. (c) Regions not covered by the contigs represent gaps and may be sequencing/assembling artifacts or natural deletions. Based on the relative position of each contig, a scaffold is created.

analysis, generating a more accurate reconstruction. All these tools use information from a single genome as reference, however more recently, some tools, such as Ragout (Kolmogorov *et al.*, 2014) and MeDuSa (Bosi *et al.*, 2015), were developed to use information from multiple reference genomes, allowing an evolutionary-based inference of structural re-arrangements. These multiple reference-based contig ordering tools will be discussed in the next section.

MUMmer (Kurtz *et al.*, 2004): is a genome-scale sequence alignment tool which can be applied to perform the alignment of a set of contigs/scaffolds to a reference genome, allowing a wide variety of applications in genomic analysis and NGS data processing, including reference-guided scaffolding. The two main algorithms of the MUMmer package are NUCmer, which performs a standard DNA-DNA alignment, and PROmer, which performs an alignment of the six reading frames of both sequences (leading to a more sensitive result, especially in the case of highly divergent organisms). The package also includes other tools, such as delta-filter, that can be used to remove the ambiguities in the alignments and select those that are more relevant for the analysis. Many scaffolding tools, like ABACAS (Assefa *et al.*, 2009), CONTIGuator (Galardini *et al.*, 2011) and MeDuSa (Bosi *et al.*, 2015), are built on top of MUMmer and take advatange of its performance, but also add new features to improve the output. MUMmer itself does not provide the sequence of the scaffold, just the positions of the alignments. Therefore, it is necessary to perform a post-processing of the results to obtain the sequence of the scaffolds. MUMmer can be obtained from its SourceForge repository http://mummer.sourceforge.net/.

ABACAS (Algorithm-based Automatic Contiguation of Assembled Sequences) (Assefa *et al.*, 2009): can use NUCmer or PROmer from the MUMmer (Kurtz *et al.*, 2004) package to align the contigs against a reference genome. The regions that do not have an equivalent sequence in the contig set are filled with Ns, indicating gaps. ABACAS can also be used to design PCR primers to amplify the unknown regions by integrating Primer3 (Koressaar and Remm, 2007; Untergasser *et al.*, 2012). ABACAS can be obtained from its SourceForge repository http://abacas.sourceforge.net/, and as part of the PAGIT package (Swain *et al.*, 2012), available at http://www.sanger.ac.uk/science/tools/pagit.

CONTIGuator (Galardini *et al.*, 2011): uses ABACAS (Assefa *et al.*, 2009) to perform the contig ordering, but adds support to multiple references, which may be useful in the case of organisms that have more than one chromosome. BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) is used to align the contigs used as input with the reference sequences to identify the correct reference for each sequence. Then, ABACAS (Assefa *et al.*, 2009) is used, and its results are integrated with the BLAST alignment to generate a final assembly. CONTIGuator can be obtained from its SourceForge repository http://contiguator.sourceforge.net/, and is also available as a webserver http://combo.dbe.unifi.it/contiguator.

Mauve (Darling *et al.*, 2004; Rissman *et al.*, 2009): is an alignment tool that can handle and align multiple genomes and identify regions of high similarity called Locally Collinear Blocks (LCBs). One of the program's features, Mauve Contig Mover, performs contig ordering using the same algorithm (Rissman *et al.*, 2009). The program runs in an iterative mode, generating and optimizing the contig orientations based on the reference until no change is possible that can improve the model. A directory is generated for each iteration that contains inputs to visualize the genome in Mauve and a FASTA file with the sorted contigs. The Mauve aligner can be obtained from the URL http://darlinglab.org/mauve/mauve.html.

FillScaffolds (Muñoz *et al.*, 2010): analyzes the genomic distance between the contig set and a reference genome and generates an ordered sequence through identifying orthologous genes. It considers the effects of the evolutionary distance in the case of missing genes, and then uses the position of the orthologos present in the reference to order the contigs. The source code of FillScaffolds is available as a supplementary data of the Muñoz *et al.* (2010) paper at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-304**.**

SIS (Scaffolds from Inversion Signatures) (Dias *et al.*, 2012): takes as input a set of contigs in FASTA format and a coordinate file generated by NUCmer or PROmer (Kurtz *et al.*, 2004) after these contigs have been aligned with the reference sequence. Using the coordinates, the program searches for inversion signatures and generates a collection of orientations of the sequences that can be used to construct the scaffolds. The source code of SIS can be obtained from the URL http://marte.ic.unicamp.br:8747.

CAR (Contig Assembly using Rearrangements) (Lu *et al.*, 2014): uses NUCmer and PROmer in combination, unlike ABACAS (Assefa *et al.*, 2009) and SIS (Dias *et al.*, 2012), that use the result of only one. Based on the coordinates, CAR uses a block permutation model to generate the contig order by considering not only the effect of the genomic inversions, but also the occurrence of transpositions (Li *et al.*, 2013). CAR can be used from the webserver http://genome.cs.nthu.edu.tw/CAR/, where the source code is also available for download.

Considering the main algorithm of each program, it is important to keep in mind that the most appropriate tool for a given task will depend on the organism and the availability of the reference genomes. ABACAS (Assefa *et al.*, 2009) is very useful if the reference genome is larger than the target genome (considering the sum of the length of all contigs, and that all contigs have a homologous region in the reference), and the primer designing tools might be helpful in some cases; however, its sensibility decreases in cases of structural divergence. In such cases, other tools,

like CONTIGuator (Galardini *et al.*, 2011) and Mauve (Darling *et al.*, 2004; Rissman *et al.*, 2009), may be more effective. Finally, SIS (Dias *et al.*, 2012) and CAR (Lu *et al.*, 2014) are indicated if the draft genome may present genomic inversions or transpositions. For most of the applications, these tools usually provide reliable results, especially for organisms that do not show a very variable genomic organization, and/or when there are enough finished genomes to properly choose the best reference. However, in some situations it may be necessary to evaluate different tools and references to check which one provides the best results. Finally, a single-reference may also lead to an "overfitted" ordering, especially when the reference is smaller, or in case of genomic inversions and translocations.

Multiple reference scaffolding

Sometimes it is very difficult to identify the most appropriate reference genome to use for contig ordering, especially when structural rearrangements are common events in the genus/species of interests. Additionally, when using BLAST to identify the most "close-related" strain from a database of already finished genomes, it is not usual to find different strains as best hit for each contig. Finally, there are also those cases where no finished genome is available, but there are draft genomes of related strains. In these cases it would not be appropriate to use programs that take into account alignments against only one reference, so data from multiple organisms should be considered. The use of multi-references is relatively recent and another consequence of the advent of NGS, as there is more draft genomes than finished ones available in public databases. Examples of algorithms and programs that use this approach are RACA (Kim *et al.*, 2013), Ragout (Kolmogorov *et al.*, 2014) and MeDuSa (Bosi *et al.*, 2015).

RACA (Reference-Assisted Chromosome Assembly) (Kim *et al.*, 2013): uses local sequence alignment to identify co-linear synteny blocks**.** The synteny blocks are filtered using a length threshold, and based on the reference genomes, the probability of each synteny block adjacent to the others is calculated. This probability can also be combined with paired-end information to identify the most probable set of scaffold. The source code of RACA can be obtained from the URL http://bioen-compbio.bioen.illinois.edu/RACA/.

Ragout (Kolmogorov *et al.*, 2014): uses phylogenetic information and synteny blocks to order a set of contigs from a target genome using multiple genome references. First, Sibelia (Minkin *et al.*, 2013) is used to identify synteny blocks shared by the target and the reference sequences. Based on the synteny, the nucleotides of the genomes are represented as sequences of blocks, and the best "block orientation" is identified by a maximum parsimony, taking into account the block order in the reference genomes. The source code of Ragout can be obtained from its repository at GitHub https://github.com/fenderglass/Ragout.

MeDuSa (Multi-Draft based Scaffolder) (Bosi *et al.*, 2015): is a graph-based scaffolder that uses information from multiples references, which can be finished or draft genomes. The program uses NUCmer to alignment the target genomes to the references and construct a weighted graph based on the alignments were the nodes of the graph are connected by identifying those contigs that aligned to the same sequence in the references. In the next step, the orientation of each contig is assigned based on the alignment information and the most-probable ordered identified in the graph. The source code of MeDuSa can be obtained from the repository at GitHub https://github.com/combogenomics/medusa.

Assembly integration

Different assemblers, or even the same assembler executed with different configurations, may produce different results. Minimum coverage, coverage cut-offs, minimum contig length, and k-mer size are examples of just some parameters that can affect the decisions of the assembler during the construction of the contigs (Baker, 2012). The way low-quality reads are treated, or how the correct paths in the assembly graph are constructed, is also different for each program. As different assemblies may present different representations of a given region in the genome, the construction of a "consensus assembly" can be an effective method of reducing assembly errors and generating an optimized set of contigs. This process, which is sometimes called "assembly reconciliation," "assembly merging," or "assembly integration," can receive as input only a set of assemblies, as implemented in Minimus (Sommer *et al.*, 2007), Reconciliator (Zimin *et al.*, 2008), MAIA (Nijkamp *et al.*, 2010), CISA (Lin and Liao 2013), GAA (Yao *et al.*, 2012) and Mix (Soueidan *et al.*, 2013), or both a set of assemblies and the reads used for the assembly, as is the case with GAM-NGS (Vicedomini *et al.*, 2013) and Zorro (Argueso *et al.*, 2009).

Minimus (Sommer *et al.*, 2007): is an assembly tool from the AMOS package (Treangen *et al.*, 2011). Initially conceived to perform assembly of small genomes, it was posteriorly adapted for assembly integration. The main algorithm is based on the overlap-layout-consensus paradigm (Peltola *et al.*, 1984), which involves taking a set of sequences and performing several alignments to identify overlaps. The information provided by the alignments is used to construct a graph that is minimized by a combination of algorithms (Myers, 1995, 2005) to generate a final assembly. Minimus is available as part of the AMOS package, which can be obtained from the SourceForge repository https://sourceforge.net/projects/amos/.

Reconciliator (Zimin *et al.*, 2008): uses NUCmer, from the MUMmer package (Kurtz *et al.*, 2004), to identify assembly errors by comparing a template with a secondary

assembly. With the alignments, the tool is able to identify the regions that have possibly suffered compression or expansion due to assembly errors in repetitive DNA sequences. The source code of Reconciliator is available from the URL http://www.genome.umd.edu/.

MAIA (Multiple Assembly Integrator) (Nijkamp *et al.*, 2010): uses the overlap-layout-consensus paradigm in a similar way to Minimus to construct a graph based on the overlaps identified by MUMmer (Kurtz *et al.*, 2004). The connections in the graph are used to construct a new assembly, and contigs that have no connection can be integrated with the assembly using a reference genome as a template. MAIA was implemented on top of the Matlab programming language and is available as a package for it that can be obtained from the URL http://bioinformatics.tudelft.nl.

GAA (Graph Accordance Assembly) (Yao *et al.*, 2012): is an assembly integration software that is based on a homonymous data structure. Taking a set of contigs as input, the tool uses BLAT (Kent, 2002) to generate alignments, identify overlaps and then generate a graph that represents the connections between the contigs. GAA is available from the SourceForge repository http://sourceforge.net/projects/gaa-wugi/.

CISA (Contig Integrator for Sequence Assembly) (Lin and Liao, 2013): uses a four-step algorithm to generate the merged assembly. First, a set of representative contigs is chosen from the individual assemblies. Assembly errors are identified by aligning all sets to one another, and any regions that are present in only one sequence are considered to be erroneous. In the event of errors, the contigs are broken in the incorrect portion into smaller sequences. The third step consists of generating several alignments using BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009), and NUCmer (Kurtz *et al.*, 2004) to identify the optimal length of repetitive sequences. The information generated in the third step is used to construct the merged assembly in the final stage of the program. CISA can be obtained from the URL http://sb.nhri.org.tw/CISA/.

Mix (Soueidan *et al.*, 2013): uses alignments generated by NUCmer (Kurtz *et al.*, 2004) to generate an extension graph where the contigs are connected by their borders. The alignments are filtered to remove repetitive sequences, and this information is used to generate a graph. Finally, the algorithm parses the graph to identify the Maximal Independent Longest Path Set (MILPS) that represents the final assembly. The Mix source code is available at the GitHub repository https://github.com/cbib/MIX.

GAM (Genomic Assemblies Merger) (Casagrande *et al.*, 2009) and GAM-NGS (Genomic Assemblies Merger for Next Generation Sequencing) (Vicedomini *et al.*, 2013): GAM takes an assembly as a template, which is referred to as the "master", and extends it using one or more sets of auxiliary assemblies (called "slaves") by identifying blocks inside the sequences that were generated by the same reads. GAM was designed to run using contigs that are generated by Sanger sequencing, and it needs an AFG layout file that contains the position information of the reads used to construct the contigs. AFG files are only produced by a few NGS assemblers, like Velvet (Zerbino and Birney, 2008) and Ray (Boisvert *et al.*, 2010). A newer version, called GAM-NGS (Vicedomini *et al.*, 2013), was developed to work in the modern context of genome assembly and to avoid the requirement for a template file by using a read alignment file (BAM) (Li *et al.*, 2009) instead. Additionally, GAM-NGS can also identify misassemblies and perform corrections before generating the final assembly. The source code of GAM-NGS is available from the GitHub repository https://github.com/vice87/gam-ngs.

Zorro (Argueso *et al.*, 2009): Combines a preprocessing step based on the masking of repetitive DNA in the sequences, followed by the split of possible misassembled regions, and the assembly integration performed by Minimus (Sommer *et al.*, 2007). The misassembled regions are identified by using Bowtie (Langmead *et al.*, 2009) to remap the reads used for the assembly back to the contigs and by analyzing the coverage along the sequence. Zorro can be obtained from the URL http://lge.ibi.unicamp.br/zorro/.

The use of read alignment information, usually taken from BAM/SAM files, may help Assembly integrating tools to properly choose and identify bocks that are derived from the same genomic region, but, just like scaffolding, the quality of the process be directly affected by the sequencing platform and settings of the read mapper. Additionally, assembly integration with paired-end reads is usually much more time and memory consuming than scaffolding, so it is important to its important to check the system's availability before choose use this approach.

Gap closing

As described in the first section, different kinds of information may be used to generate scaffolds, like paired sequence, optical and/or genomics maps, and a reference genome. To improve the assembly, some algorithms were designed to close the gaps inside the scaffolds. One approach (Figure 3), which is used by tools like GapCloser (Li *et al.*, 2010; Luo *et al.*, 2012), IMAGE (Tsai *et al.*, 2010), GapFiller (Boetzer and Pirovano, 2012), Enly (Fondi *et al.*, 2014), MapRepeat (Mariano *et al.*, 2015) and Sealer (Paulino *et al.*, 2015), utilizes the information of single-end/paired-end reads to extend, and sometimes locally-reassemble, the contigs and close the gaps. Another approach, which is employed by FGAP (Piro *et al.*, 2014) and GapBlaster (de Sá *et al.*, 2016), uses an auxiliary set of contigs to find sequences that may fill the region based on local alignment. Finally, hybrid approaches, such as GMcloser (Kosugi *et al.*, 2015), may use a combination of paired-end data, alternative assemblies and long reads in the gap-closing process.

GapCloser (Li *et al.*, 2010): is part of the SOAPdenovo software package (Li *et al.*, 2010). This mod-

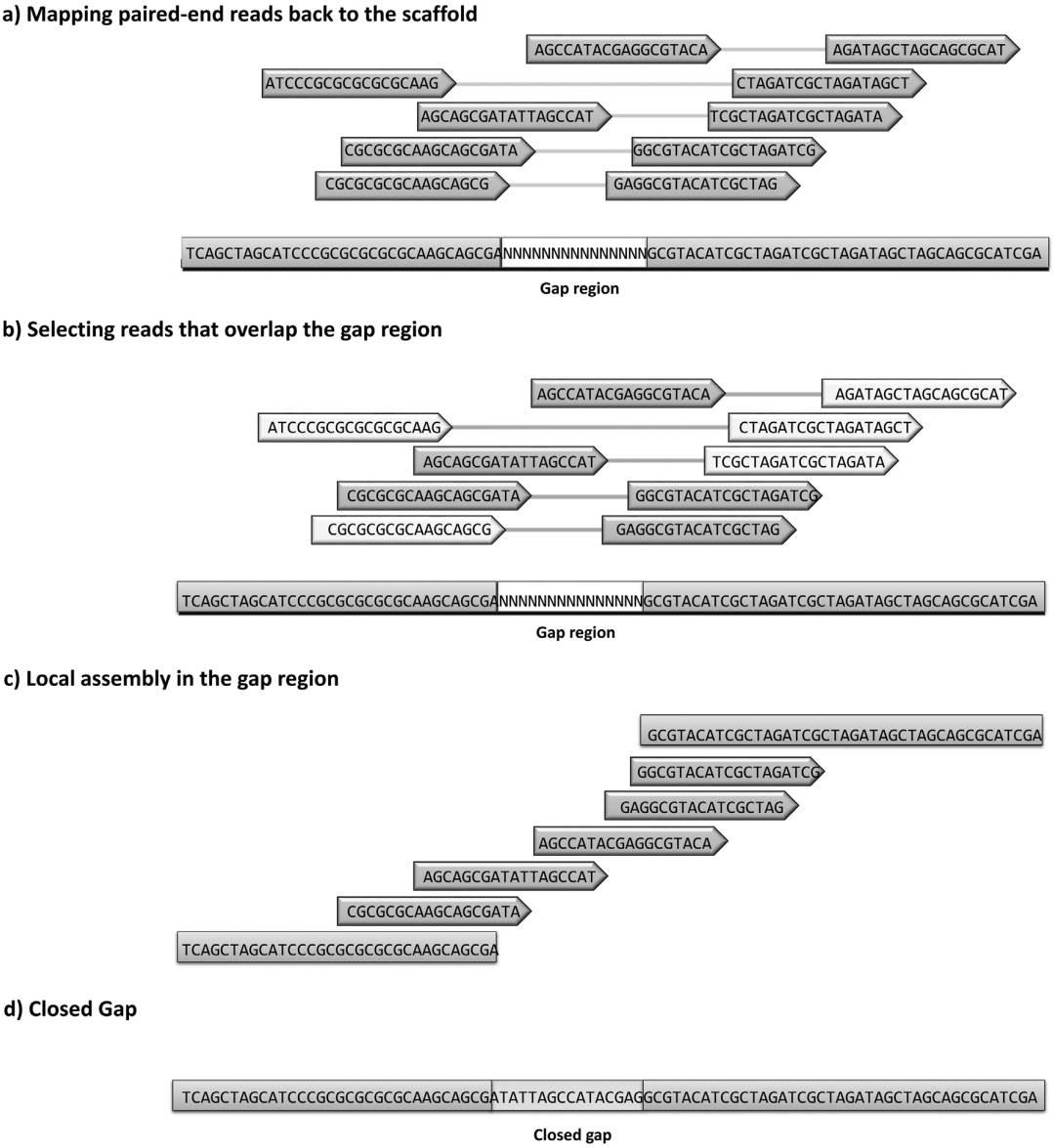


Figure 3 - Example of a gap-closing approach using paired-end reads. (a) Taking as example a scaffold constituted by two contigs joined by an assembly gap (a run of `N's) by remapping the reads back to the contigs (b) it is possible to identify reads that have at least one of the mates in the gap region. Finally, (c) the reads identified inside the gap can be de novo assembled to fill the region, resulting in a (d) closed gap.

ule uses the alignment algorithms from the package to realign the reads to the scaffolds. Based on the reads located near the gap, a local *de novo* assembly is performed constructed using a *De Bruijn* algorithm. In SOAPdenovo2 (Luo *et al.*, 2012), the GapCloser module was updated to deal with the possible errors caused by high divergent reads that might be used during the construction of the consensus sequence. GapCloser is distributed both as a stand-alone application or as part of the SOAPdenovo package, and can be obtained from the URL http://soap.genomics.org.cn/.

IMAGE (Iterative Mapping and Assembly for Gap Elimination) (Tsai *et al.*, 2010): takes as inputs a library of Illumina paired-end reads and a set of scaffolds and performs a remapping of the reads back to the sequences using SSAH2 (Ning *et al.*, 2001). Alternatively, the user can provide only the reads, and a *de novo* assembly is performed by Velvet (Zerbino and Birney, 2008) before the gap-closing step. The program maps the reads to the scaffolds and identifies those that are located at the border. Then, the reads pairs that have at least one of the mates on the border are selected, and a *de novo* assembly is performed by Velvet using the selected reads and contigs where the reads were mapped as inputs. After the assembly, the contigs are extended, and this increases the probability that the mates will

align with the adjacent contigs or the gap will be filled after the assembly increases as more iterations are performed. At the end of the process, the contigs that were linked during the assembly can be ordered according to their previously known relative position. IMAGE https://sourceforge.net/projects/image2/ can be obtained from its SourceForge repository or as part of the PAGIT package www.sanger.ac.uk/science/tools/pagit.

GapFiller (Boetzer and Pirovano 2012): In a similar approach to that used by the previous tools, GapFiller performs a remapping of the paired-end reads back to the scaffolds. However, to run the program, it is necessary to know (or at less have an estimation of) the insert size and the orientation of the library used. If this information is not available, it is possible to calculate it using the script "estimate_insert_size.pl" that is distributed along with SSPACE. To perform the gap filling exercise, the program trims the contigs on both sides to reduce the possible assembly errors caused by low coverage and then applies an iterative process of read remapping and extension of the contigs. The alignment of the reads can be performed using BWA (Li and Durbin, 2009) or Bowtie (Langmead *et al.*, 2009) and the pairs that have one of the mates on the border of the contig and the other inside the gap region are identified. The contig extension uses a *k*-mer based assembly according to the reads found inside the gap. GapFiller, in the same way than SSPACE, is available for download at the BaseClear website http://www.baseclear.com/genomics/bioinformatics/basetools/.

Enly (Fondi *et al.*, 2014): is a simple gap-closing program that works by re-mapping reads, in FASTA format, back to a target assembly using BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009), followed by a local assembly using Phrap (www.phrap.org/) or Minimo, from the AMOS package (Treangen *et al.*, 2011). If a reference genome is provided, a modified version of CONTIGuator (Galardini *et al.*, 2011) is used to order de target assembly and verify the accuracy of the gap-closing. The program can be obtained from the SourceForge repository http://enly.sourceforge.net/.

FGAP (Piro *et al.*, 2014): Unlike the other gap-filling tools, which utilize paired-end reads, FGAP's algorithm uses a supplementary set of long sequences, which can be a library of long-reads or an alternative assembly of the genome. After trimming the contigs, BLASTn, from the NCBI-BLAST+ package (Altschul *et al.*, 1990; Camacho *et al.*, 2009), is used to identify sequences in the supplementary set that overlaps the gap. FGAP can be download and used as a webserver from the URL http://www.bioinfo.ufpr.br/fgap/.

Sealer (Paulino *et al.*, 2015): The program was developed to be applied in large genomes, although can be applied in small prokaryote genomes as well. The main algorithm consists in the selection of nucleotides in the assembly that flank the gap-regions, followed by a local assembly by Konnector (Vandervalk *et al.*, 2014) using data from paired-end reads. Konnector, which takes the paired-end reads and generates pseudo-long reads by applying a combination of Bloom filter and *De Bruijn* graph, is distributed alongside with the recent versions of ABySS (Simpson *et al.*, 2009). Sealer evokes Konnector with different *k*-mers, what allows a more efficient closing of gaps caused by low coverage (usually closed by shorter *k*-mers) and repetitive elements (usually closed by longer *k*-mers). As it was developed aiming large genomes, Sealer tends to be more memory-efficient, but some steps of the program are executed serially, not in parallel, increasing the computation time. Sealer can be obtained from its GitHub repository https://github.com/bcgsc/abyss/tree/sealer-release.

GMCLoser (Kosugi *et al.*, 2015): combines information from paired-end reads and an alternative assembly, or long-reads, to fill gaps inside scaffolds. Paired-end reads are mapped to the target, and alternative assembly with Bowtie (Langmead *et al.*, 2009) and MUMmer (Kurtz *et al.*, 2004) is used to align the alternative assembly to the target assembly. To reduce the effect of misassembled regions that may be present in alternative assemblies, the program uses a likelihood approach to evaluate the contig joining by verifying the consistence based on the mate information. The likelihood algorithm available to evaluate assemblies can be accessed with the GMValue program, distributed alongside with GMCloser. The program can be obtained from its SourceForge repository https://sourceforge.net/projects/gmcloser/.

MapRepeat (Mariano *et al.*, 2015): is both a gap-closing and reference-guided scaffolding program**.** This software receives a genome assembly (in FASTA format), a reference genome (in Genbank format) and the sequencing reads (FASTA or FASTQ format). First, the assembly is ordered using the reference genome using CONTIGuator (Galardini *et al.*, 2011), and for each gap the adjacent contigs are aligned back to the reference using BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009), aiming the identification of the region in the reference that is analogous to the missing region inside the gap in the target assembly. MIRA (www.chevreux.org) is used to align the reads to the reference and generate a consensus sequence, and the regions analogous to the gap are selected and used to close it. The source code of MapRepeat can be obtained from its GitHub repository http://github.com/dcbmariano/maprepeat.

GapBlaster (de Sá *et al.*, 2016): is a gap-closing programs that, in contrast with most of the other tools, allows a manual curation of the gaps in a draft assembly instead of providing an automated gap-closing algorithm. The user may choose between the legacy NCBI-BLAST package , NCBI-BLAST+ (Altschul *et al.*, 1990; Camacho *et al.*, 2009) or Nucmer (Kurtz *et al.*, 2004) to perform alignments between a set of contigs and the draft genome of interest,

and then verify which alignments identified as flanking the gap regions may be considered for gap-closing.

The algorithms implemented in GapCloser (Li *et al.*, 2010; Luo *et al.*, 2012), IMAGE (Tsai *et al.*, 2010) and GapFiller (Boetzer and Pirovano, 2012) are based on paired-end information and are very useful for most of the genomes sequenced using Illumina or Roche 454 platforms. GapFiller requires not only the paired-end reads, but also an estimation of the size, of the insert (and its standard deviation) and the orientation of the read pairs (FR: forward-reverse, RF: reverse-forward, FF: forward-forward or RR: reverse-reverse), however, if such information is not available, it is possible to estimate it using the same script distributed along with SSPACE (Boetzer *et al.*, 2011).

In comparison to Gapfiller, Gapcloser, which also uses paired-end information for gap-closing, requires less information and may provide a more straightforward way to fill gaps. However, GapCloser also has its own limitations, including the lack of support for reads longer than 155 bp (nowadays, 2x250 bp paired-ends reads are very common for whole-genome sequencing, especially for microbial genome sequenced using Illumina MiSeq).

IMAGE is also very easy to be executed. It does not require prior knowledge about the library construction, and only needs the scaffolds, the reads and the number of iterations to be run. As it uses SMALT (based on hashes) instead of BWA or Bowtie (based on Burrows-Wheeler compression), the read-mapping also tends to be more sensitive and able to identify alignments, even if only part of the reads matches the sequence (*e.g.*, edges of contigs and gaps). However, as IMAGE does not handle insert size, it is unable to estimate the size of the gap, so the user must choose an arbitrary length after closing for those gaps that remained in the scaffolds. Although it is not problem when working with genomes assembled using only paired-end reads (with short-insert sizes), it may result in drawbacks if mate-pair reads where used for scaffolding, as long gaps would be confused with short ones.

More recently, FGAP (Piro *et al.*, 2014), GMcloser (Kosugi *et al.*, 2015) and GapBlaster (de Sá *et al.*, 2016) have provided another way to close gaps by using information for alternative assemblies, long reads or merged paired-end reads. In fact, GMcloser may also use information from paired-end reads, but by combining data from alternative assemblies and longs reads, it is possible to achieve a more reliable result. Sealer (Paulino *et al.*, 2015) also uses some of the advantages of long-reads, but by merging paired-end reads in artificial long one. Although it is not even close to PacBio or Oxford Nanopore reads, or not even to Sanger reads (~1 kb), this merging process may be helpful to solve some short unknown regions.

FGAP and GapBlaster are similar in the way they identify potential targets for gap-closing, but different in the sense that FGAP performs the gap-closing automatically whereas GapBlaster requires manual inspection. Although most of the tools are automated, the availability of tools for manual revision is also important, as some regions are very difficult to be assembled or corrected, and it is difficult to define generic rules to solve these situations.

Error-correction and assembly evaluation

The combination of different approaches employed to assemble and finish a genome can result in some artifacts, caused by the limitations of each software and/or the platform used for sequencing. Tools like QUAST (Gurevich *et al.*, 2013), REAPR (Hunt *et al.*, 2013), ALE (Clark *et al.*, 2013) and GMvalue (Kosugi *et al.*, 2015) were developed to evaluate the accuracy of an assembly. Additionally, other softwares, like iCORN (Otto *et al.*, 2010) and SEQuel (Ronen *et al.*, 2012), are able to correct assembly errors, including insertions, deletions and base substitutions.

Assembly evaluation

REAPR (Recognition of Errors in Assemblies using Paired Reads) (Hunt *et al.*, 2013): takes as input a FASTA file containing the scaffolds and a BAM file generated by the remapping of the reads used in the assembly. First, a coverage analysis is performed using SNP-o-matic (Manske and Kwiatkowski 2009) or SMALT (https://www.sanger.ac.uk/resources/software/smalt), for genomes that are smaller or bigger than 100 Mb, respectively. The software analyzes the assembly base-per-base and uses the information per position in a metric called *Fragment Coverage Distribution* (FCD), where the expected coverage distribution is compared to the observed values. The discrepant regions are treated as possible misassemblies and REAPR can generate a new set of scaffolds splitting the erroneous regions into separate sequences. REAPR can be obtained from the URL http://www.sanger.ac.uk/science/tools/reapr.

QUAST (QUality ASsessment Tool for genome assemblies) (Gurevich *et al.*, 2013): can be used to compare different assemblies of the same genome or to simply analyze an assembly. If more than one assembly is loaded, the program uses the MUMmer package to align the sequences and identify possible erroneous regions, count the number of aligned/unaligned bases, and also calculate, for each assembly, metrics like N50, L50, C+G% content, and other useful statistics. QUAST can be obtained from the URL http://bioinf.spbau.ru/quast, where it is also available through a webserver.

ALE (Assembly Likelihood Evaluation) (Clark *et al.*, 2013): uses a combination of statistical analysis that are mainly based on probability distribution and Bayesian inference to determine the accuracy of an assembly without requiring a reference genome. To evaluate the assembly, the program analyzes the *k*-mer distribution, C+G% and the relative orientation of the mates (paired-end reads) in the BAM file. ALE can be obtained from its website http://www.alescore.org.

CGAL (Computing Genome Assembly Likelihood) (Rahman and Pachter 2013): evaluates the accuracy of the assembly using a probability distribution analysis that takes into consideration the expected coverage with that obtained after the reads are remapped to the scaffolds. CGAL can be obtained from the URL http://bio.math.berkeley.edu/cgal/.

GMvalue (Kosugi *et al.*, 2015): is a program distributed along with the gap-closing program GMCloser, but can be used as a stand-alone application. The program uses NUCmer to align the assembly to a reference genome and identify misassemblies, such as insertions and deletions (INDELs). GMvalue can also generate "error-free" assemblies by splitting contigs in their erroneous regions. The program is distributed along with GMCloser and can be obtained from its SourceForge repository https://sourceforge.net/projects/gmcloser/.

When assembling a novel genome, the first metrics that are taken into account are the number of contigs, the length of the assembly and the N50. Although they can provide a good idea of how "contiguous" is the assembly, they do not measure its reliability, and may be easily distorted by inappropriate assembly procedures. Use of information from just one pair of reads to join two contigs into a scaffold, for example, may lead to an assembly error, even if the apparent fragmentation of the assembly is being reduced. If two regions are wrongly joined during the assembly, it may be misinterpreted as a natural biological event, and makes difficult further steps of assembly finishing.

Assembly correction

Assembly evaluation tools are very useful to identify structural inconsistences in a draft of apparently "finished" genome, but sometimes a manual revision and correction is inapplicable. Platform-specific errors, such as base-substitutions in Illumina data, or homopolymeric-sequence errors in IonTorrent data, may affect the annotation process, as some genes may be wrongly identified as frameshifted or mutated. As these types of errors may occur along the genome assembly, it is important to have automated tools to correct them automatically and reduce the chance of potential effects on the downstream analysis. ICORN and SEQuel are examples of programs that can be used to reduce these base-scale errors, and are based in the same principle that is used for variant calling analysis, consisting in the mapping of the reads against the sequence followed by the identification of those regions where there is a discrepancy, such as single-nucleotide polymorphism (SNPs) and Insertions and Deletions (INDELs) (Figure 4).

iCORN (Iterative Correction of Reference Nucleotides) (Otto *et al.*, 2010): is an automated pipeline for assembly correction. Using a paired-end library, the program performs the read remapping using SSAH (Ning *et al.*, 2001) and the variant calling and coverage analysis using SNP-o-matic (Manske and Kwiatkowski, 2009). The correction of the reference is followed by a new coverage analysis. If the correction promoted an improvement in the coverage, a new iteration commences, and the corrected sequence is used as the new reference; otherwise, the program stops and the last corrected sequence is returned as output. ICORN can be obtained from its SourceForge repository http://icorn.sourceforge.net/ and as part of the PAGIT package.

SEQuel (Ronen *et al.*, 2012): uses a modification of the widely used *De Bruijn* graphs called *positional De Bruijn graph*. Using the BWA's output (Li and Durbin, 2009), the main algorithms combine the *k*-mer information with the relative position and orientation information of the mates in the paired-end library to construct a graph that is used to generate a new set of corrected contigs. SEQuel can be obtained from the URL http://bix.ucsd.edu/SEQuel/.

Both iCORN and SEQuel can be used to reduce base-scale errors in genome sequences, such as single nucleotide substitutions errors and small artificial INDELs, but are not able to identify and correct genome-scale misassemblies, such as those identified by REAPR or GMValue. In fact, both REAPR and GMValue can be used to break misassembled genomes and generate a correct set of contigs, but they are not able to generate an improved assembly, in terms of contiguity, and new rounds of scaffolding and gap-closing would be necessary for this purpose. To address the problem of genome-scale assembly correction, Guizelini *et al.* (2016) have developed the tool GFinisher, which integrates the detection and correction of assembly errors with the reference-guided scaffolding and the gap-closing processes.

GFinisher (Guizelini *et al.*, 2016): is an assembly correction tool that also incorporates elements of assembly integration, reference-based scaffolding, and gap-closing in its internal pipeline. First, the program performs a reference-guided scaffolding using the module jContigSort (https://sourceforge.net/projects/jcontigsort/), which applies a HashMap-based algorithm using the *k*-mers from the draft genome and from a reference. Gaps between the contigs are then closed using jFGAP, a Java-implementation of the FGAP algorithm (Piro *et al.*, 2014), by taking information form alternative assemblies of the target genome. To identify misassembled regions, GFinisher uses an adaptation of the GC-skew metric (Lobry, 1996), which allows the analysis of fluctuations in the distribution of C and G along the sequence by using a sliding window. Points in the sequence where the GC-skew seems to be discrepant when compared to its context are considered as assembly errors, and are used to split the contigs. Then, the processes of reference-guided scaffolding, gap-closing and analysis of the GC-skew are repeated, and a final assembly is generated. At the end, all intermediary assemblies and the final one are analyzed by QUAST (Gurevich *et al.*, 2013). GFinisher can be obtained from its SourceForge repository http://gfinisher.sourceforge.net/.

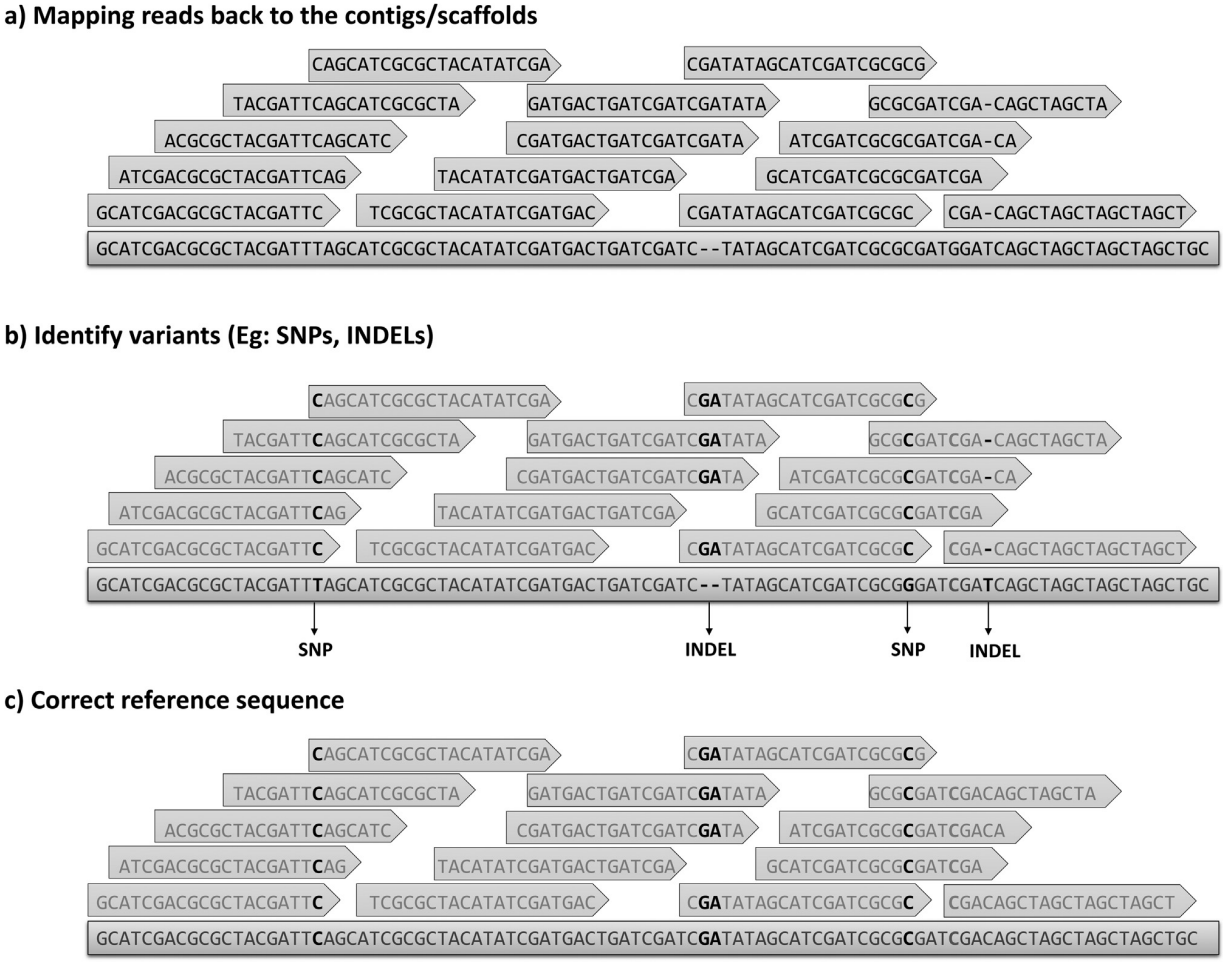


Figure 4 - Example of a simplified assembly correction approach for base substitutions and insertion/deletion misassemblies. The process steps are (a) map the reads to the assembly, (b) identify variants (eg: SNPs and INDELs) in a similar way to the common variant calling analysis pipelines, and finally, (c) correct the regions in the assembly that show discrepancies. These steps may be reiterated several times until no further change be able to improve the assembly.

While base-scale errors may affect the genome annotation by the presence of artificial frameshifts or non-synonymous mutations, the genome-scale misassemblies may lead to the erroneous identification of genome rearrangements or even the loss of relevant genes in the annotation, if they are located in the misassembled region. Therefore, the correction of base-scale errors and genome-scale misassemblies are crucial steps in the generation of high-quality finished assemblies.

Conclusion

NGS platforms have provided new ways to obtain large amounts of genomics data and reduced expressively the cost of the sequencing process itself, but also brought new challenges, especially to the data management and analysis. In the case of genome sequencing projects, for example, depending on the organism and the sequencing platform, the size of the FASTQ files containing the raw reads may vary from less than 1 gigabyte (Gb) to thousands of Gbs. Although nowadays it is perfectly possible to perform the *de novo* assembly of a microbial genome using a desktop computer, or even a notebook, when the volume of data is relatively low, many NGS analysis still require high-performance computing infrastructures (*e.g.*, clusters, cloud computing, distributed systems) to be executed. In fact, even *de novo* genome assembly, which is in a certain way a well-established field, has been constantly renewed as more approaches are being developed to optimize memory usage (*e.g.*, string-graphs, compressed data structures) and the assembly process itself, especially due to the limitations implied by the short reads sequencers and, more recently, the relatively higher error-rate observed in third-generation sequencing platforms.

The development of new bioinformatics tools in the last decade was highly influenced by the new generation of

sequencing platforms, and many software offerings were specifically designed to overcome the limitations of the NGS platforms and the challenges they have brought. The use of whole genome shotgun (WGS) sequencing became a common practice, and the amount of draft genomes available rose exponentially. However, the process by which a finished genome is generated, even in the case of bacteria, is still somewhat flawed and challenging.

The present review aimed to describe some tools that might be useful for improving genome assembly and facilitating the finishing process. For didactical purposes, the available tools were divided into four main groups. However, there are many other forms of improvements that can be used to optimize an assembly. Additionally, *in silico* tools can be very helpful and may reduce the need of re-sequencing. However, they also have some limitations, and it is important to know that their efficiency is directly dependent on the quality of the assembly and the reads.

As there are many aspects that may affect the assembly, from the sequence itself (*e.g.*, repetitive DNA, GC-rich or GC-poor) to technique artifacts (*e.g.*, platform bias, assembler errors), it is very difficult to create a straightforward method to turn millions of short-reads to close and error-free chromosome sequences. Therefore, when working with a newly sequenced organism, it may be useful to try different strategies for *de novo* assembly, assembly integration, scaffolding and gap-closing, and check how they affect the reliability of the genome assembly to choose the best parameters. Knowing the characteristics of the genomic structure of an organism, the sequencing platform and the library construction may also be very useful when choosing the tools, as some of them are designed and optimized to deal with certain specific error patterns.

Although not always applicable, there are many genome finishing tools that are also available as webservers, what may be very convenient for those researchers that are not fully confortable with Linux and commandline interfaces. Additionally, as most of these tools do not directly require the sequencing data or mapping files, the uploading and processing time is usually very fast.

Genome announcement papers that describe sequencing projects using the same technology may facilitate the choice, especially when the organism is closely related to the one you are interested to analyze. As new tools are constantly being developed and released, it may also be useful to check not only bibliography databases, such as PubMed, but also bioinformatics software repositories, such as OMICtools (https://omicstools.com), and web-based discussion forums, such as Biostars (https://www.biostarts.org) and SEQanswers (https://seqanswers.com), to keep updated about new programs and techniques.

The new generation of DNA sequencing platforms, such as PacBio RS II and Oxford Nanopore, will certainly guide the development of new bioinformatics tools and analysis protocols for the next years, and may provide an easier way to generate high-quality finished genomes. These sequencing platforms intended to produce reads that are much longer than the older second-generation platforms, like Illumina, IonTorrent, and SOLiD, and may overcome most of the limitations associated with short-reads. One of the most common issues associated with genome assembly that uses short-reads is that repeated DNA regions, like simple sequence repetitions (SSRs) that are longer than the read length, are considered computationally impossible to assemble "exactly". However, today, the throughput of these new long-read sequencers is still low and has a relatively high error rate compared to alternative platforms. This indicates that many aspects of genome sequencing still require improvement. Additionally, sequencing by PacBio SMRT still very time consuming and expensive when compared to other platforms, such as IonTorrent PGM and Illumina MiSeq. Finally, the size of the raw files generated by PacBio SMRT and Oxford Nanopore are substantially larger than the ones for second-generation platforms, even for microbial genomes, what reflects the growing need for computational and storage resources in the context of NGS.

Fifteen years since the Human Genome Sequencing Consortium released the first draft, and 10 years after Roche released its 454 platform, DNA sequencing has undergone many changes, and many technologies have been released. Furthermore, it is possible that, in the near future, only a *de novo* assembly, without any read preprocessing or post-assembly improvement, will be enough to generate high-quality finished genomes. However, such procedures are nowadays still necessary to achieve a reliable result.

References

Alkan C, Sajjadian S and Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8:61-65.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410.

Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OVC, Missawa SK, Galzerani F, Costa GGL, Vidal RO, Noronha MF, *et al.* (2009) Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. Genome Res 19:2258-2270.

Assefa S, Keane TM, Otto TD, Newbold C and Berriman M (2009) ABACAS: Algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25:1968-1969.

Au KF, Underwood JG, Lee L and Wong WH (2012) Improving PacBio long read accuracy by short read alignment. PLoS One 7:e46679.

Baker M (2012) *De novo* genome assembly: What every biologist should know. Nat Methods 9:333-337.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, *et al.* (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455-477.

Barnett DW, Garrison EK, Quinlan AR, Stromberg MP and Marth GT (2011) BamTools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics 27:1691-1692.

Bashir A, Klammer AA, Robins WP, Chin C-S, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P, *et al.* (2012) A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol 30:701-707.

Bodily PM, Fujimoto MS, Snell Q, Ventura D and Clement MJ (2016) ScaffoldScaffolder: Solving contig orientation via bidirected to directed graph reduction. Bioinformatics 32:17-24.

Boetzer M and Pirovano W (2012) Toward almost closed genomes with GapFiller. Genome Biol 13:R56.

Boetzer M, Henkel CV, Jansen HJ, Butler D and Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578-579.

Boetzer M and Pirovano W (2014) SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 15:211.

Boisvert S, Laviolette F and Corbeil J (2010) Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol 17:1519-1533.

Bosi E, Donati B, Galardini M, Brunetti S, Sagot M-F, Lió P, Crescenzi P, Fani R and Fondi M (2015) MeDuSa: A multi-draft based scaffolder. Bioinformatics 31:2443-2451.

Caboche S, Audebert C, Lemoine Y, Hot D, Soon W, Hariharan M, Snyder M, Fonseca N, Rung J, Brazma A, *et al.* (2014) Comparison of mapping algorithms used in high-throughput sequencing: Application to Ion Torrent data. BMC Genomics 15:264.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL (2009) BLAST+: Architecture and applications. BMC Bioinformatics 10:421

Casagrande A, Del Fabbro C, Scalabrin S and Policriti A (2009) GAM: Genomic Assemblies Merger: A graph based method to integrate different assemblies. IEEE Int Conf Bioinform Biomed 2009:321-326.

Chain PSG, Grafham D V, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, *et al.* (2009) Genomics. Genome project standards in a new era of sequencing. Science 326:236-237.

Chaisson MJ and Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. BMC Bioinformatics 13:238.

Clark SC, Egan R, Frazier PI and Wang Z (2013) ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics 29:435-43.

Compeau PEC, Pevzner PA and Tesler G (2011) How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 29:987-991.

Dark M (2013) Whole-genome sequencing in bacteriology: State of the art. Infect Drug Resist 6:115-123.

Darling ACE, Mau B, Blattner FR and Perna NT (2004) Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394-1403.

Dayarian A, Michael TP and Sengupta AM (2010) SOPRA: Scaffolding algorithm for paired reads via statistical optimization. BMC Bioinformatics 11:345.

de Sá PHCG, Miranda F, Veras A, de Melo DM, Soares S, Pinheiro K, Guimarães L, Azevedo V, Silva A and Ramos RTJ (2016) GapBlaster-A graphical gap filler for prokaryote genomes. PLoS One 11:e0155327.

Deschamps S, Mudge J, Cameron C, Ramaraj T, Anand A, Fengler K, Hayes K, Llaca V, Jones TJ, May G, *et al.* (2016) Characterization, correction and *de novo* assembly of an Oxford Nanopore genomic dataset from *Agrobacterium tumefaciens*. Sci Rep 6:28625.

Dias Z, Dias U and Setubal JC (2012) SIS: A program to generate draft genome sequence scaffolds for prokaryotes. BMC Bioinformatics 13:96.

Donmez N and Brudno M (2013) SCARPA: Scaffolding reads with practical algorithms. Bioinformatics 29:428-434.

Edwards DJ and Holt KE (2013) Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb Inform Exp 3:2.

Farrant GK, Hoebeke M, Partensky F, Andres G, Corre E and Garczarek L (2015) WiseScaffolder: An algorithm for the semi-automatic scaffolding of Next Generation Sequencing data. BMC Bioinformatics 16:281.

Fondi M, Orlandini V, Corti G, Severgnini M, Galardini M, Pietrelli A, Fuligni F, Iacono M, Rizzi E, De Bellis G, *et al.* (2014) Enly: Improving draft genomes through reads recycling. J Genomics 2:89-93.

Galardini M, Biondi EG, Bazzicalupo M and Mengoni A (2011) CONTIGuator: A bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med 6:11.

Gao S, Sung W-K and Nagarajan N (2011) Opera: Reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. J Comput Biol 18:1681-1691.

Guizelini D, Raittz RT, Cruz LM, Souza EM, Steffens MBR and Pedrosa FO (2016) Gfinisher: A new strategy to refine and finish bacterial genome assemblies. Sci Rep 6:34963.

Gurevich A, Saveliev V, Vyahhi N and Tesler G (2013) QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29:1072-1075.

Hatem A, Bozdag D, Toland AE, Çatalyürek ÜV, Flicek P, Birney E, Cokus S, Feng S, Sultan M, Schulz M, *et al.* (2013) Benchmarking short sequence mapping tools. BMC Bioinformatics 14:184.

Huang X and Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868-877.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M and Otto TD (2013) REAPR: A universal tool for genome assembly evaluation. Genome Biol 14:R47.

Hunt M, Newbold C, Berriman M and Otto TD (2014) A comprehensive evaluation of assembly scaffolding tools. Genome Biol 15:R42.

Huson DH, Reinert K and Myers EW (2002) The greedy path-merging algorithm for contig scaffolding. J ACM 49:603-615.

Kent WJ (2002) BLAT: The BLAST-like alignment tool. Genome Res 12:656-664.

Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge R-L, Auvil L, Capitanu B, Zhang G, Lewin HA, *et al.* (2013) Reference-assisted chromosome assembly. Proc Natl Acad Sci U S A 110:1785-1790.

Klassen JL and Currie CR (2012) Gene fragmentation in bacterial draft genomes: Extent, consequences and mitigation. BMC Genomics 13:14.

Kolmogorov M, Raney B, Paten B and Pham S (2014) Ragout - a reference-assisted assembly tool for bacterial genomes. Bioinformatics 30:i302-i309.

Koren S and Phillippy AM (2015) One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 23:110-120.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, *et al.* (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat Biotech 30:693-700.

Koren S, Treangen TJ and Pop M (2011) Bambus 2: Scaffolding metagenomes. Bioinformatics 27:2964-2971.

Koressaar T and Remm M (2007) Enhancements and modifications of primer design program Primer3. Bioinformatics 23:1289-1291.

Kosugi S, Hirakawa H and Tabata S (2015) GMcloser: Closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. Bioinformatics 31:3733-3741.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C and Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12.

Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, *et al.* (2015) Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141-161.

Langmead B, Trapnell C, Pop M and Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Latreille P, Norton S, Goldman BS, Henkhaus J, Miller N, Barbazuk B, Bode HB, Darby C, Du Z, Forst S, *et al.* (2007) Optical mapping as a routine tool for bacterial genome sequence finishing. BMC Genomics 8:321.

Li C-L, Chen K-T and Lu CL (2013) Assembling contigs in draft genomes using reversals and block-interchanges. BMC Bioinformatics 14 Suppl 5:S9.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Li R, Li Y, Kristiansen K and Wang J (2008) SOAP: Short oligonucleotide alignment program. Bioinformatics 24:713-714.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, *et al.* (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265-272.

Lin S-H and Liao Y-C (2013) CISA: Contig integrator for sequence assembly of bacterial genomes. PLoS One 8:e60843.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L and Law M (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364.

Lobry JR (1996) Asymmetric substituion patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660-665.

Lu C, Chen K-T, Huang S-Y and Chiu H-T (2014) CAR: Contig assembly of prokaryotic draft genomes using rearrangements. BMC Bioinformatics 15:381.

Lunter G and Goodson M (2011) Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21:936-939.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, *et al.* (2012) SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1:18.

Mäkinen V, Välimäki N, Laaksonen A and Katainen R (2010) Unified view of backward backtracking in short read mapping. In: Elomaa T, Mannila H and Orponen P (eds) Algorithms and Applications. Springer-Verlag, Berlin, pp 182-195.

Manske HM and Kwiatkowski DP (2009) SNP-o-matic. Bioinformatics 25:2434-2435.

Mardis E, McPherson J, Martienssen R, Wilson RK and McCombie WR (2002) What is finished, and why does it matter. Genome Res 12:669-671.

Mariano DC, Pereira FL, Ghosh P, Barh D, Figueiredo HC, Silva A, Ramos RT and Azevedo VA (2015) MapRepeat: An approach for effective assembly of repetitive regions in prokaryotic genomes. Bioinformation 11:276-9.

Minkin I, Patel A, Kolmogorov M, Vyahhi N and Pham S (2013) Algorithms in Bioinformatics. In: Darling A and Stoye J (eds) Proc. 13th International Workshop, WABI. Springer, Berlin, pp 215-229.

Muñoz A, Zheng C, Zhu Q, Albert VA, Rounsley S and Sankoff D (2010) Scaffold filling, contig fusion and comparative gene order inference. BMC Bioinformatics 11:304.

Myers EW (1995) Toward simplifying and accurately formulating fragment assembly. J Comput Biol 2:275-290.

Myers EW (2005) The fragment assembly string graph. Bioinformatics 21 Suppl 2:i79-i85.

Nagarajan N, Cook C, Di Bonaventura M, Ge H, Richards A, Bishop-Lilly KA, DeSalle R, Read TD and Pop M (2010) Finishing genomes with limited resources: Lessons from an ensemble of microbial genomes. BMC Genomics 11:242.

Nijkamp J, Winterbach W, van den Broek M, Daran J-M, Reinders M and de Ridder D (2010) Integrating genome assemblies with MAIA. Bioinformatics 26:433-439.

Ning Z, Cox AJ and Mullikin JC (2001) SSAHA: A fast search method for large DNA databases. Genome Res 11:1725-1729.

Noé L and Kucherov G (2005) YASS: Enhancing the sensitivity of DNA similarity search. Nucleic Acids Res 33:W540-W543.

Otto TD, Sanders M, Berriman M and Newbold C (2010) Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics 26:1704-1707.

Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD and Birol I (2015) Sealer: A scalable gap-closing application for finishing draft genomes. BMC Bioinformatics 16:230.

Peltola H, Söderlund H and Ukkonen E (1984) SEQAID: A DNA sequence assembling program based on a mathematical model. Nucleic Acids Res 12:307-321.

Peng Y, Leung HCM, Yiu SM and Chin FYL (2010) Research in Computational Molecular Biology. In: Berger B (ed) Proc. 14th Annual International Conference, RECOMB 2010. Springer Berlin, Heidelberg, pp 426-440.

Pevzner PA, Tang H and Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A 98:9748-9753.

Piro VC, Faoro H, Weiss VA, Steffens MBR, Pedrosa FO, Souza EM and Raittz RT (2014) FGAP: An automated gap closing tool. BMC Res Notes 7:371.

Pop M, Kosack DS and Salzberg SL (2004) Hierarchical scaffolding with Bambus. Genome Res 14:149-159.

Rahman A and Pachter L (2013) CGAL: Computing genome assembly likelihoods. Genome Biol 14:R8.

Ramos RTJ, Carneiro AR, Soares S de C, Santos AR dos, Almeida S, Guimarães L, Figueira F, Barbosa E, Tauch A, Azevedo V, *et al.* (2013) Tips and tricks for the assembly of a *Corynebacterium pseudotuberculosis* genome using a semiconductor sequencer. Microb Biotechnol 6:150-156.

Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM, Montmayeur A, Shea TP, Walker BJ, *et al.* (2012) Finished bacterial genomes from shotgun sequence data. Genome Res 22:2270-2277.

Ricker N, Qian H and Fulthorpe RR (2012) The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. Genomics 100:167-175.

Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD and Perna NT (2009) Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics 25:2071-2073.

Ronen R, Boucher C, Chitsaz H and Pevzner P (2012) SEQuel: Improving the accuracy of genome assemblies. Bioinformatics 28:188-196.

Roy RS, Chen KC, Sengupta AM and Schliep A (2012) SLIQ: Simple linear inequalities for efficient contig scaffolding. J Comput Biol 19:1162-1175.

Salmela L, Mäkinen V, Välimäki N, Ylinen J and Ukkonen E (2011) Fast scaffolding with small independent mixed integer programs. Bioinformatics 27:3259-3265.

Salmela L and Rivals E (2014) LoRDEC: Accurate and efficient long read error correction. Bioinformatics 30:3506-3514.

Sanger F, Nicklen S and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463-5467.

Simpson JT and Durbin R (2012) Efficient *de novo* assembly of large genomes using compressed data structures. Genome Res 22:549-556.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM and Birol I (2009) ABySS: A parallel assembler for short read sequence data. Genome Res 19:1117-23.

Sommer DD, Delcher AL, Salzberg SL and Pop M (2007) Minimus: A fast, lightweight genome assembler. BMC Bioinformatics 8:64.

Soueidan H, Maurier F, Groppi A, Sirand-Pugnet P, Tardy F, Citti C, Dupuy V and Nikolski M (2013) Finishing bacterial genome assemblies with Mix. BMC Bioinformatics 14 Suppl 1:S16.

Staden R (1979) A strategy of DNA sequencing employing computer programs. Nucleic Acids Res 6:2601-2610.

Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M and Otto TD (2012) A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat Protoc 7:1260-1284.

Tettelin H, Radune D, Kasif S, Khouri H and Salzberg SL (1999) Optimized multiplex PCR: Efficiently closing a whole-genome shotgun sequencing project. Genomics 62:500-507.

Treangen TJ, Sommer DD, Angly FE, Koren S and Pop M (2011) Next generation sequence assembly with AMOS. Curr Protoc Bioinformatics 33:11.8.1-11.8.18.

Tritt A, Eisen JA, Facciotti MT and Darling AE (2012) An integrated pipeline for *de novo* assembly of microbial genomes. PLoS One 7:e42304.

Tsai IJ, Otto TD and Berriman M (2010) Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol 11:R41.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M and Rozen SG (2012) Primer3 - new capabilities and interfaces. Nucleic Acids Res 40:e115.

Vandervalk BP, Jackman SD, Raymond A, Mohamadi H, Yang C, Attali DA, Chu J, Warren RL and Birol I (2014) Konnector: Connecting paired-end reads using a bloom filter de Bruijn graph. 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2014:51-58.

Vicedomini R, Vezzi F, Scalabrin S, Arvestad L and Policriti A (2013) GAM-NGS: Genomic assemblies merger for next generation sequencing. BMC Bioinformatics 14 Suppl 7:S6.

Vincent AT, Derome N, Boyle B, Culley AI and Charette SJ (2016) Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. J Microbiol Methods 138:60-71.

Warren RL, Sutton GG, Jones SJM and Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500-501.

Yao G, Ye L, Gao H, Minx P, Warren WC and Weinstock GM (2012) Graph accordance of next-generation sequence assemblies. Bioinformatics 28:13-16.

Zerbino DR and Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821-829.

Zimin AV, Smith DR, Sutton G and Yorke JA (2008) Assembly reconciliation. Bioinformatics 24:42-45.

Supplementary material

The following online material is available for this article:

Table S1 - Examples of sequenced microbial genomes for which the tools discussed in the present review were used.

Associate Editor: Guilherme Corrêa de Oliveira