VNiVERSiDAD DE SALAMANCA
FACULTAD DE BIOLOGÍA
DEPARTAMENTO DE MICROBIOLOGÍA Y GENÉTICA
ÁREA: GENÉTICA

TESIS DOCTORAL

Estudio de la evolución de un fitopatógeno: Genómica comparada del hongo patógeno de maíz *Colletotrichum graminicola*

GABRIEL EDUARDO RECH

SALAMANCA, 2013

UNIVERSIDAD DE SALAMANCA
Facultad de Biología
Departamento de Microbiología y Genética
Área: Genética
Centro Hispano-Luso de Investigaciones Agrarias

Insight into the evolution of a plant pathogen: Comparative genomic analysis of the fungal maize pathogen *Colletotrichum graminicola*

PhD Thesis

Programa de Doctorado: Agrobiotecnología

Órgano responsable del Programa de Doctorado:
Departamento de Fisiología Vegetal

Gabriel Eduardo Rech

Salamanca, 2013

UNIVERSIDAD DE SALAMANCA
Facultad de Biología
Departamento de Microbiología y Genética
Área: Genética
Centro Hispano-Luso de Investigaciones Agrarias

Estudio de la evolución de un fitopatógeno: Genómica comparada del hongo patógeno de maíz *Colletotrichum graminicola*

Tesis Doctoral

Programa de Doctorado: Agrobiotecnología

Órgano responsable del Programa de Doctorado:
Departamento de Fisiología Vegetal

Gabriel Eduardo Rech

Salamanca, 2013

D. Luis Román Fernández Lago, Director del Departamento de Microbiología y Genética de la Facultad de Biología de la Universidad de Salamanca y **Dña. Berta Dopico Rivela**, Directora del Departamento de Fisiología Vegetal de la Facultad de Biología de la Universidad de Salamanca, órgano responsable del Programa de Doctorado en Agrobiotecnología

CERTIFICAMOS:

Que la presente Memoria titulada **"Estudio de la evolución de un fitopatógeno: Genómica comparada del hongo patógeno de maíz *Colletotrichum graminicola*"**, ha sido realizada en el Departamento de Microbiología y Genética de la Facultad de Biología y el Centro Hispano-Luso de Investigaciones Agrarias de la Universidad de Salamanca por el Licenciado **D. Gabriel Eduardo Rech**, bajo la dirección del Dr. D. Michael Ronald Thon y la Dra. Dña. Serenella Ana Sukno y cumple las condiciones exigidas para optar al grado de Doctor por la Universidad de Salamanca.

Para que así conste, firmamos el presente certificado en Salamanca a 14 de Noviembre de 2013.

Fdo: D. Luis Román Fernández Lago Fdo: Dña. Berta Dopico Rivela

D. Michael Ronald Thon, Profesor Titular del Departamento de Microbiología y Genética de la Facultad de Biología de la Universidad de Salamanca, y **Dña. Serenella Ana Sukno**, Personal Investigador Doctor del Departamento de Microbiología y Genética de la Facultad de Biología de la Universidad de Salamanca

CERIFICAMOS:

Que la presente Memoria titulada **"Estudio de la evolución de un fitopatógeno: Genómica comparada del hongo patógeno de maíz *Colletotrichum graminicola*"**, ha sido realizada en el Departamento de Microbiología y Genética de la Facultad de Biología y el Centro Hispano-Luso de Investigaciones Agrarias de la Universidad de Salamanca, bajo nuestra dirección, por el Licenciado **D. Gabriel Eduardo Rech**, y cumple las condiciones exigidas para optar el grado de Doctor por la Universidad de Salamanca.

Para que así conste, firmamos el presente certificado en Salamanca a 14 de Noviembre de 2013.

Fdo: Dr. D. Michael R. Thon

Fdo: Dra. Dña. Serenella A. Sukno

Fdo: D. Gabriel E. Rech

Este trabajo se ha llevado a cabo en el Laboratorio 1 del Centro Hispano-Luso de Investigaciones Agrarias (CIALE), Departamento de Microbiología y Genética, de la Universidad de Salamanca bajo la dirección del Profesor Dr. Michael R. Thon y de la Dra. Serenella A. Sukno. Durante el desarrollo de la Tesis he disfrutado de una beca de Formación de Personal Investigador (FPI) de la Secretaría de Estado de Investigación, Desarrollo e Innovación (ex Ministerio de Ciencia e Innovación), enmarcada en el Proyecto "Análisis bioinformático y funcional de la evolución de genes de patogenicidad en hongos" (AGL2008-03177/AGR). Dentro del programa FPI realicé dos Estancias Breves que complementaron el trabajo presentado en esta Tesis: una de ellas en el Laboratorio del Prof. David Adelson, del "Centre for Bioinformatics and Computational Genetics" de la Universidad de Adelaida, Australia; y la otra en el Grupo "Computational Biochemistry Research Group" dirigido por el Prof. Gaston Gonnet en la Escuela Politécnica Federal de Zúrich (ETH), Suiza.

Publicaciones científicas relacionadas con los resultados obtenidos en esta Tesis Doctoral y otros trabajos publicados durante el periodo de formación predoctoral:

- Vargas WA, Sanz-Martín JM, **Rech GE**, Armijos-Jaramillo VA, Rivera-Rodriguez LP, Echeverria MM, Díaz-Mínguez, JM, Thon MR and Sukno SA. Delivery of a pathogen-encoded effector protein into the nucleus of *Zea mays*. The power of effectoromics for crop protection (*manuscript in preparation for Nature Biotechnology*).

- **Rech GE**, Sanz Martín JM, Anisimova M, Sukno SA and Thon MR. Natural selection on coding and non-coding DNA sequences is associated with virulence genes in a plant pathogenic fungus (*manuscript submitted to New Phytologist*).

- **Rech GE**, Vargas WA, Sukno SA and Thon MR. Identification of positive selection in disease response genes within members of the Poaceae (2012). *Plant Signaling & Behavior* 7(12):1667-1675.

- O'Connell RJ, Thon MR, Hacquard S, Amyotte SG, Kleemann J, Torres, MF, Damm U, Buiate EA, Epstein L, Alkan N, Altmüller J, Alvarado-Balderrama L, Bauser CA, Becker C, Birren BW, Chen Z, Choi J, Crouch JA, Duvick JP, Farman MA, Gan P, Heiman D, Henrissat B, Howard RJ, Kabbage M, Koch C, Kracher B, Kubo Y, Law AD, Lebrun M-H, Lee Y-H, Miyara I, Moore N, Neumann U, Nordström K, Panaccione DG, Panstruga R, Place M, Proctor RH, Prusky D, **Rech GE**, Reinhardt R, Rollins JA, Rounsley S, Schardl CL, Schwartz DC, Shenoy N, Shirasu K, Sikhakolli UR, Stüber K, Sukno SA, Sweigard JA, Takano Y, Takahara H, Trail F, Does HC van der, Voll LM, Will I, Young S, Zeng Q, Zhang J, Zhou S, Dickman MB, Schulze-Lefert P, Themaat EVL, Ma L-J and Vaillancourt LJ. Lifestyle transitions in plant pathogenic Colletotrichum fungi deciphered by genome and transcriptome analyses (2012). *Nature Genetics* 44(9):1060–1065.

- Vargas WA, Sanz Martín JM, **Rech GE**, Rivera LP, Benito EP, Díaz-Mínguez JM, Thon MR and Sukno SA. Plant defense mechanisms are activated during biotrophic and necrotrophic development of *Colletotricum graminicola* in maize (2012). *Plant Physiology* 158(3):1342–1358.

A Pamela.

Estamos acostumbrados a contemplar los modelos informáticos como simplificaciones del mundo real, pero en cierto sentido los modelos informáticos de la selección natural no son simplificaciones, sino complicaciones del mundo real.

Richard Dawnkins.
Escalando el monte improbable
(Climbing mount improbable. 1998)

AGRADECIMIENTOS

En primer lugar me gustaría agradecer a mis directores, Mike y Serenella, por haberme dado la oportunidad de realizar este trabajo en su grupo, por permitirme desarrollar mis ideas y brindarme el apoyo necesario para llevarlas a cabo. Gracias también por esas magníficas barbacoas "hispano-argento-estadounidenses"!!!

Gracias a los miembros del Grupo (Walter, Vinicio y José). Ha sido un verdadero placer trabajar junto a ustedes y he aprendido mucho de nuestras charlas (científicas o no) y sin duda esta Tesis no sería la misma sin vuestro aporte.

Un agradecimiento especial a los miembros del CIALE, en particular a los integrantes del Laboratorio 1, por acogerme y brindarme un magnifico ambiente laboral.

Thank you very much to the people who received me in their labs during my stays in Australia and Switzerland. To Professor David Adelson and all people in his Group in the University of Adelaide, especially to Zhipeng Qu for helping me (a lot!) with the ncRNAs pipeline, for making me test Chinese beer and for receive me in the White team FC!! Also to Dr. Maria Anisimova in the ETH, Zurich for helping me with the selection tests and for all your support during my stay.

Me gustaría agradecer también a todos aquellos con los que compartí mi día a día durante el Máster y en el CIALE: Anna, Dany, Raúl, Virginia, Lucía, Pepe, Marta R., Eduardo, Alex, Thais, Sara, Patri, Abe, Luis, Lucho, Irene, Javier, Ángela, Sergio, y muchos otros de los que han pasado por esa estupenda mesa de la sala de becarios, y con quienes he compartido mucho más que las comidas, incluyendo risas y debates trascendentales.

Gracias a los "Galácticos" del CIALE FC!, en especial a su fundador Dov Prusky que con su espíritu aventurero logró juntar un grupo de jóvenes pachangueros que, con unos meses de entrenamiento, llegarían a convertirse en… humm… un grupo de amigos pachangueros que les gusta el fútbol. Una vez a la semana, lo dejaban todo el campo: Rubén (el crack), Dov ("*cuchá*"), Walter (nuestro Pepe), Raúl (la seguridad), Joni (CR7), Tomás ("*fácil, fácil*"), Luis (el talento), José (la organización), Sergio (el aire del equipo), Vinicio (la táctica), Alex (la garra), Javier (el poder de la perseverancia), Eduardo (la experiencia), Isi (el DT) y por supuesto David (el ausente).

A mis entrañables excompañeros y amigos de la carrera de Lic. en Ciencias Biológicas de la Universidad Nacional de Salta: Cris, Ale y Mauri, que a pesar de haber tomado rumbos distintos y de estar separados por miles de kilómetros, sé que puedo contar con ustedes siempre.

Un agradecimiento especial a mi gran amigo y hermano de la vida Maty Fernández, por sus palabras de aliento, su apoyo incondicional y por estar siempre dispuesto a dar una mano.

Muchas gracias a Edgardo y Stella. Nunca olvidaré lo que hicieron durante mi estancia en Zúrich. Hay que tener mucha suerte para conocer personas como ustedes que, con el mayor desinterés, abrieron las puertas de su casa, lo cual inició una amistad que estoy seguro que mantendremos por muchos años.

A los grandes amigos que he tenido la suerte de conocer en Salamanca y que han estado siempre presentes en las buenas y en las malas, especialmente a Walter, Esther, Marcos, Joni, Anita y Marta. Gracias por vuestros consejos y apoyo en todo momento y por ser los responsables de las mejores experiencias que he vivido durante estos cuatro años. Nunca os olvidaré y espero que el futuro nos vuelva a reunir en algún momento.

A mi familia: a mis padres y hermanos por haberme dado todo, por el apoyo constante y por estar siempre presentes a pesar de la distancia. A mis suegros, cuñados, primas y a la extraordinaria familia que descubrimos en Pamplona, por preocuparse siempre por mí y por todo lo que nos han ayudado y apoyado durante estos años.

Por último, quisiera agradecer a la persona más importante de mi vida, Pame, por todo lo que hemos superado juntos durante estos años, por tu apoyo y cariño incondicional y por darme fuerzas cuando más lo necesitaba. Por esto y mucho más, esta Tesis va dedicada a vos.

INDEX

0. RESUMEN

Se estima que la población mundial para el año 2.050 será de unos 9.100 millones de personas y uno de los mayores desafíos que enfrentaremos para ese tiempo será el de ser capaces de alimentar a cada uno de los integrantes del planeta. Incluso en la actualidad, el crecimiento poblacional y una sociedad cada vez más demandante hacen que las prácticas agrícolas tradicionales no sean suficientes para abastecer las exigencias de productos más nutritivos y variados, como así también los requerimientos de otras materias primas derivadas de las plantas utilizadas para la producción de biocombustibles o pienso. En este contexto, la biotecnología aplicada al desarrollo de nuevos y mejores cultivos, estimulada por las nuevas tecnologías de secuenciación, los avances logrados en las ciencias "ómicas" y el acelerado crecimiento de nuestros conocimientos acerca de los procesos biológicos, aparece como una de las alternativas más prometedoras (sino la única) para abastecer las actuales y futuras demandas agrícolas.

La era de las "ómicas", en referencia a las disciplinas como la genómica, la proteómica, la transcriptómica y la metabolómica, ha llevado a una explosión en la cantidad de datos biológicos, la cual condujo al nacimiento de una nueva disciplina científica conocida como Bioinformática. La bioinformática es un área de investigación interdisciplinaria que surge a partir de campos tan disímiles como matemática, química, estadística, física, biología y ciencias de la computación. El objetivo final de la bioinformática es descubrir la riqueza de la información biológica encubierta dentro de una gran masa de datos y obtener una visión más clara de la biología fundamental de los organismos. Todas las ómicas se basan en el análisis de un gran volumen de datos y, por lo tanto, se valen de las técnicas rápidas y automatizadas que ofrece la bioinformática para almacenar y analizar de manera eficiente dichos datos con el objetivo de extraer información biológicamente significativa.

Las tecnologías de secuenciación de última generación (next generation sequencing technologies) han revolucionado el campo de la genómica, ofreciendo la oportunidad a pequeños grupos de secuenciar genomas completos de una manera rápida y a bajo coste (**Figura 1**). Esta revolución ha tenido profundas consecuencias en numerosos campos de investigación, entre ellos el de la agrobiotecnología. Las secuencias genómicas completas de cientos de

especies de importancia agronómica han sido o están siendo finalizadas en la actualidad. Estos avances proporcionan innumerables herramientas a la comunidad científica agrícola, facilitando el desarrollo de cultivos más saludables y productivos a través de los programas de mejoramiento vegetal, como así también aportando información básica acerca de la biología y evolución de las especies de interés agrícola. La bioinformática cumple un rol fundamental durante este proceso, ya que es el área encargada de la manipulación, almacenaje y procesamiento de toda la información generada a partir de los proyectos de secuenciación. En la actualidad, se ofrecen múltiples aplicaciones y herramientas libremente accesibles destinadas al análisis de secuencias biológicas (**Tabla 1**).

Uno de los mayores problemas a los que se enfrenta la agricultura en todo el mundo está relacionado con las enfermedades producidas por agentes biológicos (virus, bacterias, hongos, nematodos, insectos, etc.). Informes recientes sitúan las pérdidas en los cultivos debido a las enfermedades en torno al 40%, en su gran mayoría causadas por patógenos fúngicos. Uno de los mayores problemas para el desarrollo de estrategias efectivas frente a hongos fitopatógenos es que, al igual que la mayoría de los microorganismos, presentan gran plasticidad fenotípica y una extraordinaria capacidad de adaptarse a nuevos ambientes y/o de infectar nuevos huéspedes. En este contexto, resulta imprescindible determinar qué genes (u otras regiones del genoma) están evolucionando más rápidamente a fin de entender de qué manera puede originarse la enfermedad.

Por otra parte, la identificación en el laboratorio de genes importantes para la patogenicidad, es a menudo una tarea difícil y costosa. Una alternativa más rápida y barata, en ausencia de candidatos a priori, es buscar evidencias de evolución adaptativa a nivel de todo el genoma. La evolución adaptativa puede ser descripta como la retención de cambios en el genoma que, de alguna manera, confieren ventajas adaptativas los individuos que los poseen. Las evidencias de evolución adaptativa pueden ser de distintos tipos: selección positiva actuando en secuencias homólogas, cambios rápidos en el número de genes de una familia génica y/o reordenaciones estructurales en el genoma.

La gran cantidad de genomas fúngicos disponibles, sumado a lo relativamente económico que resulta en nuestros días re-secuenciar genomas completos de hongos, nos brinda una oportunidad única para realizar estudios de genómica comparada y evolución molecular entre estos organismos. Adicionalmente, este crecimiento acelerado de datos biológicos, ha ido acompañado del desarrollo de poderosos algoritmos destinados al análisis a nivel de genomas completos, lo cual nos ofrece la posibilidad de investigar distintos aspectos de la evolución molecular de estos organismos.

Al igual que otros fitopatógenos, los hongos secretan una gran diversidad de proteínas efectoras que modulan la inmunidad innata de las plantas, permitiéndoles infectar y colonizar la planta hospedadora. Hace tiempo que los fitopatólogos describen los efectores en términos diferentes, incluyendo factores de avirulencia, factores de virulencia, elicitores y toxinas. Dentro del campo de la fitopatología, entender la manera en que las proteínas efectoras alteran la fisiología de la planta hospedadora para permitir el desarrollo de infecciones, se ha convertido en uno de los principales temas de investigación. Se sabe que algunas de estas proteínas interactúan directamente con proteínas del hospedador, por lo que están sometidas a una fuerte presión selectiva para evitar su reconocimiento por el hospedador. Existen pruebas de que, por lo menos en ciertos casos, los genes efectores evolucionan rápidamente y están sujetos a selección positiva (**Tabla 2**).

Entre las enfermedades vegetales causadas por los hongos, la antracnosis es una de las más destructivas, causando pérdidas significativas en cultivos en los cinco continentes. Los agentes etiológicos mejor conocidos de la enfermedad son ciertas especies del hongo filamentoso *Colletotrichum*, que infecta prácticamente todas las familias de plantas de interés agronómico u hortícola, con la consecuencia de pérdidas económicas significativas (**Figura 5**). Además, los hongos del género *Colletotrichum* representan importantes modelos experimentales en estudios relacionados con diversos aspectos de la fitopatología, como ser enzimas degradadoras de carbohidratos, procesos infectivos, resistencia del hospedador y biología molecular de las interacciones planta-patógeno. Muchas de las especies del género *Colletotrichum* presentan una forma de nutrición denominada hemibiotrofía, por lo que exhiben

características tanto biotróficas como necrotróficas (**Figura 7**). En las especies hemibiotróficas de *Colletotrichum*, las esporas germinan y forman apresorios en forma de semiesfera que penetran en las células de la planta combinando fuerza mecánica y enzimas degradadoras. Al entrar en el tejido de la planta, el hongo vive como biotrofo durante un breve tiempo en el seno de las células vivas del hospedador y luego cambia a una forma de vida necrotrófica en la que mata las células de la planta, causando lesiones necróticas que van expandiéndose.

Varios grupos de investigación alrededor del mundo utilizan especies del género *Colletotrichum* como modelos para el estudio de la patogénesis vegetal. Dichos esfuerzos han resultado recientemente en la secuenciación del genoma de dos especies pertenecientes a este importante grupo de patógenos, entre ellos el de *Colletotrichum graminicola*, agente causal de la antracnosis del maíz y organismo modelo para el estudio de la antracnosis en gramíneas.

En esta Tesis, se han utilizado numerosas herramientas bioinformáticas para atender a diversas hipótesis relativas a la evolución de la patogenicidad en hongos filamentosos, haciendo hincapié en el patosistema *C. graminicola*-maíz. Para esto, nos hemos basado en el análisis de los genomas de ocho cepas de *C. graminicola*, una de ellas la recientemente secuenciada M1.001 y las restantes siete resecuenciadas por nuestro grupo. En el primer capítulo de la Tesis (**Chapter I**) se describen los procedimientos llevados a cabo para la secuenciación, ensamblaje y anotación de los nuevos genomas, como así también las características fenotípicas de dichos aislados. Por otra parte, por medio de análisis de los genomas, se presentan resultados concernientes a cuatro aspectos fundamentales de la genómica comparada de estos organismos: recombinación, relaciones filogenéticas, variaciones estructurales del genoma y ganancia-pérdida de genes. En el segundo capítulo (**Chapter II**) se analizan los patrones de selección natural actuando sobre secuencias codificantes y no codificantes de proteínas a nivel de todo el genoma. En el tercer capítulo (**Chapter III**) se detallan los resultados de la aplicación de una metodología bioinformática para la detección y anotación de ARNs no codificantes (ncRNAs) fúngicos a partir de secuencias públicamente disponibles. Finalmente, en el cuarto capítulo (**Chapter IV**), se analizan los patrones de selección positiva

actuando sobre genes que codifican proteínas relacionadas con las defensas de las plantas entre miembros de las Poáceas.

En general, la presente Tesis Doctoral ofrece un recurso importante para los fitopatólogos moleculares, ya que podría servir como guía para el análisis bioinformático de genomas de hongos fitopatógenos en busca de dianas para el desarrollo de estrategias de control. A su vez, el presente trabajo contribuye a mejorar nuestros conocimientos acerca de los procesos moleculares y evolutivos que tienen lugar durante las interacciones planta-patógeno. A continuación, se describen los aspectos más importantes de cada uno de los capítulos desarrollados en la Tesis.

Capítulo I. Caracterización, secuenciación y genómica comparada de aislados de *Colletotrichum graminicola*.

Uno de los aspectos fundamentales que debe determinarse cuando se analizan poblaciones de patógenos es su variabilidad genética y aquellos factores que podrían estar favoreciendo su mantenimiento. Entender dicha variabilidad y las principales causas de su origen y persistencia puede ayudarnos a discernir aquellas características genómicas que determinan la habilidad de los patógenos para atacar a otros organismos y para adaptarse rápidamente a nuevos sistemas.

Con el objetivo de caracterizar la variabilidad genética entre asilados de *C. graminicola*, se resecuenciaron los genomas completos de siete asilados de campo procedentes de distintas regiones del mundo y que presentan notables diferencias fenotípicas, principalmente en cuanto a virulencia frente al maíz (**Tabla 3** y **Figura 11**). Para analizar la variabilidad fenotípica de los aislados, se realizaron ensayos *in planta* e *in vitro*. Como indicador de virulencia se utilizó el tamaño medio de lesión (mm^2) causada por inoculaciones del hongo en hojas de maíz cuatro días después de la infección. Este ensayo resultó en la división de cuatro categorías superpuestas de virulencia, que iban desde aislados completamente asintomáticos (i63127 y i51134) hasta aislados altamente virulentos (M1.001 y i318). Por otra parte, se cuantificaron y analizaron tres rasgos fenotípicos vinculados al crecimiento *in vitro* como lo son la tasa de crecimiento (en tres medios de cultivo diferentes), el porcentaje de germinación de esporas y la esporulación, medida por el número de esporas por

placa. Todos los ensayos mostraron diferencias significativas entre los aislados (**Tabla 4 y Figura 12**) y una correlación positiva con el tamaño medio de lesión (**Figura 13**). Entre los caracteres que mostraron mayor correlación se encuentran la tasa de crecimiento en medio "oatmeal agar" y en medio mínimo, indicando que la virulencia de los asilados podría estar relacionada con su capacidad para esporular y para sobrevivir y crecer bajo condiciones de nutrientes limitados. El experimento de porcentaje de germinación de esporas y posteriores ensayos de infección sobre hojas de maíz con heridas demostraron que, si bien los aislados i63127 y i51134 aparentan ser no-patogénicos a cuatro días después de la infección (**Figura 11**), existen diferencias fundamentales entre ambos. En primer lugar, el asilado i51134 desarrolla síntomas a 7-10 días después de la inoculación en hojas sanas y es capaz de generar prácticamente la misma sintomatología que los aislados más virulentos cuando se inocula en hojas con heridas, mientras que i63127 no muestra síntomas incluso en hojas con heridas provocadas artificialmente. Por otra parte, i51134 es capaz de generar apresorios en la superficie de hojas sanas, mientras que i63127 no los desarrolla. Posteriores estudios desarrollados en nuestro grupo han demostrado que el aislado i63127 es capaz de sobrevivir a modo de endófito dentro de células del mesófilo en hojas con heridas provocadas artificialmente, sin realizar el característico cambio en el modo de vida de *C. graminicola* (de biotrofo a necrotrofo) responsable de la aparición de los devastadores síntomas de la antracnosis.

Una vez caracterizados fenotípicamente los aislados, se procedió a realizar cultivos monospóricos para extraer el ADN genómico que sería enviado para secuenciar. La secuenciación de los aislados se realizó mediante el sistema Illumina HiSeq2000, resultando en más de 400 millones de "reads" de 100 pares de bases de longitud. El ensamblaje de los genomas se realizó utilizando dos aproximaciones diferentes (**Figura 8**): Ensamblaje por mapeo de los reads al genoma de referencia M1.001 (*mapping assembly*) y el ensamblaje *de novo*. El ensamblaje basado en el mapeo y alineamiento de los reads al genoma de referencia resultó en la reconstrucción de secuencias genómicas consenso de alta calidad para cada uno de los asilados, con una media de entre 24X a 132X bases cubriendo cada nucleótido en el genoma de referencia (*depth of coverage*) y entre el 85-99% del genoma de M1.001 cubierto con al menos 3X en cada

aislado (**Tabla 5**). Por otro lado, por medio del ensamblaje *de novo* también obtuvimos genomas de alta calidad, con excepción de los genomas iJAB2 y i51134, que mostraron una fragmentación mucho mayor con respecto al resto de los genomas, probablemente debido a una pobre calidad de ADN genómico secuenciado para estos asilados (**Tabla 6**), por lo que estos aislados no fueron considerados en algunos de los posteriores análisis en los que dicha fragmentación podría influir en los resultados y conclusiones obtenidos.

Determinar la existencia de eventos de recombinación en hongos fitopatógenos resulta fundamental a la hora de encarar estrategias para controlar la enfermedad, debido a que este mecanismo contribuye de manera significativa a la generación de nuevas combinaciones alélicas, incrementando las posibilidades de adaptación a nuevos ambientes. Para analizar la presencia de recombinación en los aislados de campo de *C. graminicola* se utilizaron dos aproximaciones diferentes, haciendo uso de las secuencias generadas por ambas estrategias de ensamblaje. En primer lugar, se analizó la presencia de eventos de recombinación a través de la exploración de sitios filogenéticamente incompatibles entre las secuencias cromosómicas consenso generadas a partir del *mapping assembly* (**Tabla 7 y Figura 15**). Como resultado obtuvimos que todos los cromosomas, con excepción del cromosoma 12, mostraban evidencias de la existencia de sitios incompatibles. Posteriormente se aplicó el mismo procedimiento, pero analizando el genoma completo en ventanas (*sliding-windows*) de 1.000 nucleótidos. Se analizaron un total de 65.824 ventanas, de las cuales un 2% mostró valores significativos de recombinación en base a la presencia de sitios filogenéticamente incompatibles (**Figura 16**). Debido a que la evidencia de sitios incompatibles puede también estar determinada por eventos de mutación recurrente, se procedió a aplicar otro test de recombinación basado en la detección de secuencias mosaico. Para esto, inicialmente se realizó un alineamiento múltiple de los genomas ensamblados *de novo*, y se identificaron aquellos bloques relativamente conservados entre todos los aislados (Locally Collinear Blocks o LCBs). De un total de 37.167 LCBs analizados, 1.889 (5,1%) mostraron evidencias de mosaicismo (**Tabla 8**). Posteriormente se analizaron los alineamientos múltiples de secuencias creados a parir del locus cromosómico de cada gen ±500 nucleótidos, resultando en un total de 257 loci (2.2%) con evidencias de mosaicismo (**Tabla 9**). En general,

todos los análisis mostraron evidencias de recombinación entre los genomas de aislados de *C. graminicola*. Estos resultados fueron más consistentes para los cromosomas largos (1 al 10). Por otra parte, los minicromosomas (11 al 13) mostraron resultados más variados. El cromosoma 12, por ejemplo, presentó una alta diversidad genética, dificultando el análisis y la interpretación de los resultados, ya que la mayoría de las secuencias pertenecientes a este cromosoma no pudieron ser evaluadas. Por otra parte, los minicromosomas 11 y 13 muestran resultados variados y algunas veces contradictorios. Precisamente estos cromosomas presentan una gran cantidad de ADN repetitivo (**Tabla 12**), lo cual podría estar afectando los resultados e interpretación de los análisis de recombinación, principalmente debido a su probable implicación en mecanismos que promueven la ocurrencia de mutaciones recurrentes. Más allá de los resultados observados para los minicromosomas, la evidencia de recombinación en cromosomas largos fue predominante. Una de las causas más probables es la presencia de reproducción sexual en aislados de campo de *C. graminicola*. Estas observaciones resultan de gran interés a la hora de diseñar estrategias para el control de la antracnosis del maíz, ya que hasta el momento no se ha encontrado la fase sexual de *C. graminicola* en campo, la cual podría impactar de manera significativa en la biología poblacional de la especie, favoreciendo la evolución acelerada de factores de virulencia o de genes relacionados con la resistencia a antifúngicos. Sin embargo, es importante destacar que la evidencia de recombinación no necesariamente implica que se hayan producido eventos de recombinación meiótica. Numerosos trabajos han demostrado la presencia de recombinación en hongos estrictamente asexuales, por ejemplo a través de la recombinación parasexual. Un análisis más exhaustivo, incluyendo información respecto a la biología reproductiva de los asilados de campo y estudios de genética poblacional a lo largo de más de una generación, podrían arrojar resultados más concluyentes con respecto al comportamiento reproductivo de los aislados de campo de *C. graminicola*.

Con el objetivo de determinar las relaciones genéticas entre los aislados de *C. graminicola*, se construyó un árbol de máxima verosimilitud (*maximun likelihood*) en base al alineamiento y concatenación de 8.288 genes presentes en los 8 aislados (**Figura 18**). Curiosamente, los dos aislados más virulentos y que muestran una gran similitud fenotípica (i318 y M1.001), mostraron también una

gran similitud genética, sugiriendo que i318 podría representar un clon de la cepa M1.001. Estos resultados son aún más llamativos teniendo en cuenta que el aislado i318 fue colectado en Nigeria, mientras que M1.001 fue colectado en Missouri (USA). Esto sugiere que la cepa i318 ha sido probablemente introducida en Nigeria a través de material contaminado proveniente de USA. Más allá de la relación descripta entre los aislados i318 y M1.001, en general la reconstrucción filogenética no mostró una asociación evidente entre similitud genética y virulencia. Sin embargo, hay que tener en cuenta que los genes utilizados para la reconstrucción filogenética representan los más conservados entre los aislados de *C. graminicola*, por lo que los factores determinantes de la variabilidad en virulencia podrían estar relacionados con secuencias que evolucionan muy rápidamente, reordenaciones cromosómicas o la ganancia/pérdida de genes, factores que no fueron incluidos durante la reconstrucción del árbol. Los análisis de reordenaciones cromosómicas y de ganancia/pérdida de genes se detallan a continuación, mientras que un detallado estudio acerca de las secuencias que evolucionan más rápidamente se describe en la siguiente sección (**Capítulo II**).

El análisis de reordenaciones cromosómicas se realizó mediante la exploración del alineamiento de los reads de cada uno de los genomas mapeados al genoma de referencia de M1.001, con excepción de los genomas de iJAB2 e i51134, ya que presentaban una cantidad considerablemente menor de reads mapeados al genoma de referencia, lo que no proporcionaba suficiente confianza a la hora de predecir variaciones estructurales. Se estudiaron distintos tipos de variaciones intra-cromosomales (inversiones, deleciones, inserciones y translocaciones) e inter-cromosomales (translocaciones inter-cromosomales). Para los cinco aislados analizados, el cromosoma 6 y los minicromosomas 11, 12 y 13 mostraron ser los más afectados por las variaciones estructurales (**Tabla 11**, **Figura 20** y **Tabla 12**), lo cual sugiere que estos cromosomas podrían representar una fuente de generación de nuevas variantes genómicas, promovidas principalmente por la presencia de grandes cantidades de ADN repetitivo (**Figura 21**). Para determinar las posibles implicaciones de las variaciones estructurales, se analizaron los genes cercanos a los puntos cromosómicos de ruptura (*breakpoints*) compartidos por todos los genomas (**Figura 19**), los cuales probablemente representan modificaciones ocurridas en

el genoma de la cepa M1.001. De un total de 17 regiones genómicas analizadas, se obtuvieron 374 genes que podrían estar afectados por variaciones cromosómicas en el genoma de M1.001. Este set de genes esta enriquecido con efectores específicos de *C. graminicola* (**Tabla 13**), por lo es probable que dichas variaciones promuevan la variabilidad de estos genes, contribuyendo a una mayor plasticidad para adaptarse y atacar nuevos huéspedes. Por otra parte, también analizamos la función de los genes localizados en regiones cercanas a *breakpoints* que afectaban únicamente al genoma del aislado asintomático i63127 (**Figura 19**). Se analizaron un total de 72 genes, pero no se observó enriquecimiento de ninguna categoría funcional en particular. Sin embargo, se identificaron 4 genes que codifican enzimas involucradas en la degradación de la pared celular vegetal (una celulasa y tres hidrolasas). Variaciones estructurales que modificaran la secuencia proteica final de alguna o varias de estas enzimas, podrían ser responsables de la ausencia de una fase necrotrófica en el aislado i63127, la cual se caracteriza principalmente por la degradación y maceración del tejido vegetal a través de esta clase de enzimas. Por otra parte, otros genes afectados por la ocurrencia de variaciones estructurales podrían ser también los responsables del cambio (switch) en el modo de vida de *C. graminicola* de biotrofo a necrotrofo, por lo que también representan excelentes candidatos para su caracterización funcional.

Como objetivo final de esta sección, se deseaba determinar la existencia de genes únicos en cada uno de los aislados, como así también aquellos genes presentes en todos ellos (*core gene set*). Para esto, en primer lugar se realizó la anotación automática de los genomas ensamblados *de novo*, y posteriormente se procedió a identificar secuencias homólogas por medio de alineamientos globales y locales (**Figura 22**). Como resultado se obtuvo un "*core gene set*" constituido por 10.379 genes (**Figura 23**), muchos de ellos involucrados en funciones celulares básicas, pero también muchos otros que codifican factores de virulencia (**Figura 24** y **Tabla 15**), lo que implicaría que incluso los aislados menos virulentos llevan la maquinaria básica para ser patógenos. Por otra parte, el análisis del contenido génico reveló que el aislado asintomático i63127 carecía de 35 genes presentes en todos los otros aislados. Entre ellos, dos efectores, dos factores de virulencia una cutinasa y otros seis genes sobreexpresados durante la infección de la cepa M1.001 en hojas de maíz. La pérdida de estos genes,

probablemente fundamentales para el completo desarrollo de la antracnosis, representan excelentes candidatos para el diseño de experimentos funcionales, ya que la pérdida de alguno de ellos podría estar directamente involucrada con la ausencia de la formación de apresorios y de una fase necrotrófica, y la consecuente ausencia de síntomas, demostrada por el aislado i63127. Por último, el análisis de genes únicos en cada genoma reveló que una gran fracción de estos codifica pequeñas proteínas secretadas (**Tabla 17**) que podrían actuar como efectores específicos en determinados cultivos huéspedes (variedades) o representar innovaciones evolutivas involucradas en la adaptación a nuevos ambientes.

En general, el primer capítulo de esta Tesis Doctoral se ha orientado a la introducción de algunas características fenotípicas y genómicas de ocho cepas de *C. graminicola*. Además, se presentan evidencias referentes a posibles eventos de recombinación ocurridos entre los aislados de campo y un análisis exhaustivo de translocaciones genómicas y variaciones en el contenido de genes entre los aislados. Muchos de estos eventos estarían directamente involucrados en la patogenicidad, virulencia y especificidad de los aislados, representando un valioso recurso para posteriores análisis funcionales y de genética de poblaciones en este patógeno.

Capítulo II. Selección natural en secuencias codificantes y no codificantes de proteínas en aislados de *Colletotrichum graminicola*

Uno de los mayores problemas en el desarrollo de estrategias eficaces para combatir hongos fitopatógenos, es que tienen la capacidad de adaptarse a una gran variedad de estilos de vida y por lo tanto, emplean diferentes estrategias para infectar y colonizar sus huéspedes. Entender la manera en que los hongos han evolucionado para ocupar diferentes nichos ecológicos, nos puede ayudar a determinar características genéticas, fisiológicas y bioquímicas exclusivas de patógenos, contribuyendo a la identificación de dianas específicas para el desarrollo de estrategias de control respetuosas con el medio ambiente.

Si bien el origen de los caracteres adaptativos responsables de las diferencias fenotípicas entre los organismos sigue siendo objeto de debate, el papel de la

selección positiva (selección a favor de mutaciones ventajosas) se considera fundamental durante dicho proceso. Las poblaciones de patógenos y huéspedes ejercen entre sí presiones de selección a lo largo de numerosas generaciones, provocando adaptaciones rápidas que pueden dejar huellas en los genomas de los individuos involucrados en la interacción. Se han propuesto dos modelos evolutivos para describir la dinámica poblacional ocurrida durante la interacción huésped-patógeno que se conocen como "Carrera Armamentista" (*arms race*) y "Reina Roja" (*red queen*) (**Figura 2**). En el modelo de carrera armamentista, la interacción huésped-patógeno constantemente selecciona variantes genéticas que aumentan la aptitud o eficacia biológica (*fitness*) en ambas poblaciones. En este caso, la selección es direccional (positiva), ya que los cambios se acumulan en ambas poblaciones con una tendencia a la fijación de la variante ventajosa. En el modelo de la Reina Roja, dos o más variantes génicas pueden coexistir en ambas poblaciones, y son mantenidas por la acción de la selección balanceadora, que promueve la persistencia de dichos polimorfismos dinámicos, favoreciendo las variantes que ofrecen a las poblaciones una mayor eficacia biológica en determinado momento de la interacción. Diferentes modos de selección natural provocan diferentes patrones de variabilidad en las secuencias de ADN. De esta manera, mediante el análisis de los polimorfismos, podemos determinar aquellas secuencias que probablemente juegan un papel funcionalmente importante durante la interacción huésped-patógeno.

En los últimos años, numerosos trabajos han demostrado que los análisis de selección natural pueden ayudar en gran medida en la determinación de secuencias codificantes de proteínas involucradas en la interacción planta-patógeno (**Tabla 2**). Por otra parte, el alto nivel de similitud que muestran las proteínas (en número y función) de organismos fenotípicamente muy diferentes, y el descubrimiento en eucariotas superiores de que una gran proporción del ADN no codificante es funcional, ha llevado a muchos investigadores a preguntarse si la principal causa de la diversidad fenotípica está determinada por mutaciones en las regiones codificantes de proteínas o, por el contrario, en las secuencias reguladoras. Por esta razón, recientemente se le ha prestado mucha atención a la evolución molecular del ADN no codificante, con resultados que demuestran que, en numerosos organismos, una gran

proporción del AND no codificante se encuentra bajo la acción de la selección natural, por lo que estas regiones podrían estar también participando en la adaptación de los organismos al medio.

En el presente capítulo, se han investigado patrones de selección natural actuando a nivel de secuencias codificantes y no codificantes de proteínas utilizando los genomas consenso obtenidos para cada uno de los siete aislados de *C. graminicola* y descriptos en el primer capítulo, sumados al genoma de la cepa M1.001. Inicialmente, se analizaron los patrones de polimorfismo a nivel de todo el genoma en base a la distribución empírica de los valores "D" de Tajima. La prueba D de Tajima se basa en la comparación de dos estimadores de la diversidad genética poblacional: la diversidad nucleotídica y el número total de sitios segregativos en un grupo de secuencias. La relación que existe entre ambos estimadores permite determinar si las secuencias se encuentran bajo el modelo neutral o si se desvían del mismo. Si ambos estimadores dan el mismo resultado en cuanto a variación genética, significaría que el polimorfismo observado es neutro y se encuentra distribuido aleatoriamente. En cambio, si existen diferencias entre ambos, la selección estaría afectando alguno de ellos promoviendo su incremento o decremento: si existe selección balanceadora, se observa un incremento de la diversidad nucleotídica (por lo tanto D>0); si existe un mayor número de sitios deletéreos en la muestra el número total de sitios segregativos se verá incrementado (por lo que D<0), indicando la acción de selección positiva o purificadora. De esta forma, determinando estadísticamente las diferencias entre ambos, se puede detectar de manera indirecta la participación de la selección natural en el mantenimiento o pérdida del polimorfismo en las poblaciones. Un problema potencial de la prueba D de Tajima, es que diferentes procesos demográficos pueden confundirse con los efectos de la selección. Por este motivo, buscar valores inusuales (extremos) en la distribución empírica de los valores D, proporciona evidencias más certeras con respecto a la acción de la selección, dado que se espera que los procesos demográficos afecten a todo el genoma por igual, mientras que la selección afectaría sólo regiones específicas. Una primera aproximación al estudio de la distribución de los polimorfismos entre los aislados de *C. graminicola*, consistió en calcular los valores de D en ventanas de 5Kb, con saltos de 500bp (*sliding-windows*) a lo largo de todo el genoma (**Figura 25**). Al analizar los genes

localizados en ventanas con valores extremos de D, no se observó enriquecimiento para ninguna categoría funcional, lo cual podría deberse a que, mediante esta estrategia, todos los polimorfismos en la ventana se consideran bajo las mismas presiones selectivas, sin diferenciación entre aquellos polimorfismos que son efectivamente funcionales (por ejemplo los que alteran la secuencia de la proteína codificada o los que ocurren en regiones regulatorias) y los que no (por ejemplo los que ocurren en sitios sinónimos de las regiones codificantes).

Para intentar determinar si diferentes clases de secuencias mostraban diferentes patrones de selección, se clasificó el genoma completo de los aislados de *C. graminicola* en diferentes clases de secuencias: intrones, regiones intergénicas (5'upstream, 3'downstream, 5'UTR and 3'UTR) y regiones codificantes (sitios sinónimos y no sinónimos) (**Figura 26A**) y se estimaron los valores de D para cada una de las secuencias en cada clase. Los resultados mostraron que, en general, las regiones intrónicas, UTRs y sitios no-sinónimos presentaban valores de Tajima ligeramente menos positivos que los sitios sinónimos, indicando que la mayoría de estas secuencias se encuentran bajo la acción de la selección purificadora (**Figura 26B** y **Tabla 19**). Un análisis detallado de las regiones no codificantes que mostraron valores extremos de D, reveló que muchas de ellas pertenecían a genes relacionados con patogenicidad (**Figura 27**). Por ejemplo, el set de secuencias 5'UTR con valores extremadamente positivos de D, esta estadísticamente enriquecido con regiones 5'UTR pertenecientes a genes que codifican proteínas secretadas y genes sobreexpresados en *C. graminicola* durante la infección en maíz. Valores positivos de D, usualmente se asocian con la acción de la selección balanceadora. Sin embargo, este modelo ha sido usualmente aplicado a secuencias codificantes. La región 5'UTR suele estar involucrada en la regulación del gen adyacente. Mediante este análisis, se detectaron patrones en la región 5'UTR que se asimilan al modelo de la reina roja, por lo que es probable que el mantenimiento de los polimorfismos en esta región esté contribuyendo a la plasticidad fenotípica, mediante la expresión diferencial de genes importantes para la patogenicidad.

Una segunda aproximación al estudio de los patrones de selección, consistió en utilizar estadística paramétrica para la identificación de secuencias bajo la acción de la selección positiva. La selección positiva a nivel molecular puede ser detectada a través de la comparación de las tasas de substituciones. En forma general, se asume que cuando una secuencia presenta un exceso significativo en la tasa de substituciones, en comparación con la esperada bajo el modelo neutral, está siendo sometida a la selección positiva. En el caso de las secuencias codificantes, se calcularon los valores de ω = dN/dS, donde dN representa la tasa de substituciones no sinónimas por sitio no sinónimo y dS representa la tasa de substituciones sinónimas por sitio sinónimo. Para el cálculo de ω se utilizaron los modelos de codones de Markov implementados en el programa PAML para determinar, mediante máxima verosimilitud, diferencias estadísticas entre los modelos que asumían selección positiva y los que no. Como resultado se obtuvieron 224 secuencias codificantes (1,86%) con evidencias de selección positiva. El set de 224 genes, esta enriquecido con genes involucrados en metabolismo secundario, proteínas secretadas y factores de virulencia (**Figura 27**), es decir, moléculas que se espera que interactúen constantemente con el huésped y que posiblemente estén involucradas en la carrera armamentista.

El análisis de selección positiva en secuencias no codificantes se llevó a cabo mediante el método de Wong y Nielsen (2004). En este análisis, y de forma análoga al análisis de regiones codificantes, la tasa de substitución en la región no codificante (dNC) se compara con la tasa de substitución neutra (dS) por medio de la expresión ζ = dNC/dS. Una de cuestiones más importantes a determinar en análisis de selección positiva en regiones no codificantes está relacionada con la adecuada elección de los sitios putativamente neutrales con los que comparar la tasa de substitución en la región no codificante. En el presente trabajo se utilizó la tasa de substituciones sinónimas en las regiones adyacentes a la secuencia no codificante que se deseaba analizar. Para evitar la influencia de posibles casos de selección a nivel de sitios sinónimos, la tasa de substitución sinónima se calculó utilizando las secuencias codificantes de los tres genes más cercanos a la región en cuestión. Mediante el análisis de la función y expresión de los genes adyacentes a las secuencias con evidencias de selección positiva, se determinó que numerosas regiones 3'UTR de genes

involucrados en patogenicidad mostraban evidencias de selección positiva. Entre ellos, genes codificantes de factores de virulencia y genes sobreexpresados durante la infección, sugiriendo que el proceso adaptativo a nivel molecular podría estar también relacionado con substituciones en las regiones reguladoras que favorecen la variabilidad en la expresión de genes de patogenicidad, ofreciendo una mayor plasticidad fenotípica.

Este capítulo de la Tesis Doctoral representa el primer estudio de selección natural a nivel molecular en regiones codificantes y no codificantes de proteínas a escala genómica en un hongo fitopatógeno. Se determinó que ambos tipos de secuencias se encuentran bajo la acción de la selección, la cual actúa preferentemente en regiones genómicas involucradas en patogenicidad. Como es de esperar bajo el modelo de la carrera armamentista, la selección positiva parece ser predominante en genes que codifican proteínas que interactúan con el huésped (efectores, factores de virulencia y metabolitos secundarios). Por otra parte, otros genes sobreexpresados durante la infección, muestran evidencias de selección en las regiones reguladoras, sugiriendo un rol adaptativo de estas secuencias, probablemente implicado en la plasticidad fenotípica. Curiosamente, las regiones 5'UTR de genes sobreexpresados durante la infección, e implicadas en la regulación de los genes adyacentes, muestran evidencias de selección balanceadora, más acorde al modelo de la reina roja. En base a los resultados obtenidos en el presente capítulo se sugiere que, mientras que las substituciones adaptativas a nivel de secuencias codificantes podrían ser importantes en proteínas que interactúan constantemente con el huésped, los polimorfismos en las regiones regulatorias podrían conferir mayor flexibilidad en el proceso de infección mediante la regulación diferencial (en tiempo y lugar) de genes importantes para la patogénesis.

Capítulo III. Anotación de ARN no codificantes (ncRNAs) fúngicos

Las células de todos los organismos estudiados hasta el momento contienen dos especies de ARN, el ARN mensajero que codifica proteínas (ARNm) y el ARN no codificante o non-coding RNA (ncRNA). En contraste con los ARNm, los ncRNAs no se traducen a proteínas, pero tienen importantes funciones celulares, ya sea solos o formando complejos con proteínas. Algunas funciones conocidas de los ncRNAs son el procesamiento del ARN, la modificación y

regulación de la transcripción, la estabilidad y traducción del ARNm y la secreción de proteínas (**Figura 29**). Los ncRNAs se clasifican como ncRNAs largos (más de 200 nucleótidos) y ncRNAs cortos (menos de 200 nucleótidos-normalmente-entre 20 y 30).

Existen tres clases principales de ncRNAs cortos bien estudiados: los ARNs cortos de interferencia (siRNA), los micro ARNs (miRNAs) y los RNAs que interactúan con proteínas Piwi (piRNAs). A los ncRNAs cortos clásicamente se los ha asociado con la dirección de vías de silenciamiento génico, ya sea por medio de la represión de la traducción o por degradación del ARN mensajero, y se han relacionado también con la formación de heterocromatina (Amaral et. al, 2008). A pesar de que la literatura predominante está referida a los ncRNA cortos, cada vez hay más evidencia de la presencia de ncRNAs largos funcionales en muchos organismos. Los ncRNAs largos están comúnmente asociados a la diferenciación celular y el desarrollo de organismos complejos. Se ha propuesto que las trayectorias epigenéticas de diferenciación, están principalmente programadas por redes reguladoras de ARN.

Los ARNs cortos están muy bien descriptos en muchos eucariotas superiores como en los mamíferos y las plantas, sin embargo, el conocimiento en este campo relacionado con eucariotas simples, como los hongos filamentosos, es muy limitado. Un considerable número de ncRNAs se han descrito en levaduras (*Saccharomyces cerevisiae*). El análisis bioinformático de los genomas de siete especies diferentes de levadura, proporcionó un gran número de ncRNAs estructurales evolutivamente conservados, sugiriendo su importante rol en la regulación post-transcripcional. Muchos de estos ncRNAs son sintetizados por la ARN Polimerasa II e incluyen, entre otros, a los Cryptic Unstable Transcripts (CUTs, ~400nt) que son moléculas de ARN derivados de la transcripción de las regiones intergénicas del genoma y por lo general son rápidamente degradados y a otra clase de transcriptos más estables que se han denominado Stable Unannotated Transcripts (SUTs, ~700nt). En el hongo filamentoso *Neurospora crassa*, recientemente se han descrito varias nuevas especies de ARN cortos, incluyendo algunos milRNAs, ARNs cortos de interferencia independientes de Dicer (disiRNAs) y qiRNAs (que surgen como respuesta a daños en el ADN). En general, estos trabajos demuestran la existencia e importancia de los ncRNAs in

hongos, sin embargo es probable que la gran mayoría de ncRNAs fúngicos y sus funciones sean desconocidas. Por otra parte, el conocimiento acerca del número y las funciones de los ncRNAs podría resultar fundamental para una mejor comprensión de la biología de los hongos, ya que podría abrir nuevas alternativas para el desarrollo de drogas antifúngicas.

En el tercer capítulo de esta Tesis Doctoral, se presentan los resultados obtenidos a partir de la aplicación de una metodología orientada a la identificación de ncRNAs a partir de secuencias públicamente disponibles de ESTs (Expressed Sequences Tag) pertenecientes a 42 especies de hongos y 4 especies de Oomycetos. Los criterios utilizados para seleccionar las especies que se analizarían fueron: 1) Genoma de la especie completamente secuenciado, 2) Disponibilidad de la anotación completa del genoma, 3) Más de 10.000 ESTs en NCBI (dbEST). Además, se incluyeron algunas especies que, a pesar de no cumplir con el requisito 3), representaban especies de interés económico, agrícola o biológico como ser *Aspergillus oryzae*, *Colletotrichum graminicola*, *Colletotrichum higginsianum*, *Fusarium oxysporum* y *Paxillus involutus* (**Tabla 21**).

La metodología utilizada para la anotación de ncRNAs constó de cuatro pasos principales (**Figura 30**): 1) Obtención de ESTs a partir de bases de datos públicas (NCBI - dbEST) y control de calidad (eliminación de contaminantes y secuencias de baja calidad y complejidad) de dichas secuencias. 2) Agrupamiento (clustering) y ensamblaje de los ESTs, con el objetivo de obtener secuencias únicas. 3) Mapeo de los ESTs a cada genoma, con el objetivo de descartar aquellas secuencias que no pertenezcan a la especie en cuestión, como ser aquellas que pudieran provenir de contaminaciones. 4) Filtración de transcriptos codificantes de proteínas. A partir de todos los unigenes, se realizaron numerosos análisis para descartar aquellos que codifican proteínas.

La metodología se aplicó individualmente para cada una de las 46 especies, resultando en la identificación de 8.251 ncRNAs putativos, a partir del análisis de 2.127.338 ESTs (**Tabla 22**). Posteriormente, se procedió a anotar los ncRNAs identificados. Para esto, se utilizó BLAST con diferentes bases de datos de ncRNAs, como así también otros métodos computacionales utilizados comúnmente para la predicción de ncRNAs. Un total de 590 secuencias (7,15%)

fueron anotadas, representadas principalmente por miRNAs (**Tabla 23**). El bajo porcentaje de secuencias que pudieron ser anotadas en base a comparaciones con RNAs conocidos, sumado al bajo grado de similitud que mostraron las secuencias (**Figura 32**), sugieren que los ncRNAs fúngicos se encuentran pobremente conservados a nivel de secuencia.

Con el objetivo de conocer si existía alguna relación entre la proporción de transcriptos clasificados como ncRNAs y la filogenia, patogénesis y/o estilo de vida de las especies analizadas, se clasificaron las especies de acuerdo a los diferentes criterios y se calculó si existían diferencias significativas entre los diferentes grupos. El árbol filogenético de las especies analizadas se reconstruyó en base al gen de copia única MS456, anteriormente descrito como uno de los más adecuados para reconstrucciones filogenéticas en hongos (**Figura 31**). No se observaron diferencias significativas entre los distintos grupos monofiléticos ni en relación al modo de vida de las especies analizadas (**Tabla 24**). Sin embargo, las especies patogénicas mostraron una leve, pero significativa ($p=0.036$) proporción mayor de transcriptos anotados como ncRNAs putativos con respecto a las no patogénicas, sugiriendo que la complejidad asociada a organismos patógenos podría estar determinada, al menos parcialmente, por un incremento en el número de ARNs funcionales, probablemente involucrados en la regulación de procesos más complejos como la infección y colonización de otros organismos. Sin embargo, es importante tener en cuenta que la correlación (o falta de correlación) entre los diferentes grupos de especies y la proporción de transcriptos clasificados como ncRNAs, podría estar altamente sesgada por la falta de precisión en la anotación de cada genoma, ya que la adecuada anotación de secuencias codificantes es fundamental para la correcta aplicación de la metodología utilizada para identifica ncRNAs.

A fin de determinar si los ncRNAs identificados se estaban expresando particularmente en alguna librería, se calculó la proporción de secuencias que pertenecían a cada una de las librerías de ESTs, tanto para el conjunto de todos los transcriptos únicos mapeados, como así también para las secuencias clasificadas como ncRNAs. Luego se realizó una prueba de Chi-Cuadrado para cada especie con el objeto de determinar si había diferencias en los valores obtenidos de cada conjunto de datos. Para aquellas especies en que la prueba

Chi-Cuadrado dio diferencias estadísticamente significativas, se realizó una prueba Z para proporciones, con el fin de determinar qué librerías estaban diferencialmente representadas entre los dos conjuntos de datos. Numerosas especies mostraron un predominio de ncRNAs producidos durante la fase reproductiva, es decir ESTs generados a partir de estructuras reproductivas como conidios, cleistotecios, peritecios y esporas o durante la creación de los mismos (**Tabla 26**). La producción de estas estructuras consiste básicamente en un proceso de diferenciación celular del cuerpo vegetativo del hongo, por lo que nuestros resultados coinciden con observaciones previas en las que se ha demostrado que los ncRNAs podrían ser fundamentales en dicho proceso. En caso de ser así, estas secuencias podrían convertirse en nuevas dianas en el desarrollo de antifúngicos, dirigidos a evitar la generación de estructuras de dispersión como esporas o conidios.

Por último, en el capítulo anterior (**Capítulo II**) se encontraron evidencias de selección natural actuando a nivel de secuencias no codificantes entre aislados de *C. graminicola* y una de las hipótesis que se planteó, vinculaba la funcionalidad de dichas secuencias con la transcripción de ncRNAs funcionales. Lamentablemente, al momento de realizar la anotación de los ncRNAs, las secuencias disponibles de ESTs eran muy pocas para este organismo, por lo que solo se detectaron 22 transcriptos que putativamente representan ARNs no codificantes. Sin embargo, sorprendentemente cuatro de los 22 ncRNAs de *C. graminicola* coinciden con regiones que presentan evidencias de algún tipo de selección (**Tabla 27**), sugiriendo una relativa importancia funcional con posibles implicaciones en la interacción planta-patógeno. Estas secuencias, como así también el resto de ncRNAs fúngicos identificados en este capítulo, representan excelentes candidatos para su caracterización funcional, debido a que son secuencias que se transcriben durante diferentes estadios del crecimiento o en estructuras especializadas, pero que no son traducidas a proteínas, sugiriendo que podrían representar ARN funcionales involucrados en algunas de las múltiples funciones usualmente asignadas a este tipo de secuencias, ofreciendo la posibilidad para el desarrollo de nuevas estrategias para el control de las enfermedades fúngicas.

Capítulo IV. Selección positiva en genes de respuesta defensiva entre especies de la familia de las Poáceas

Las plantas, a diferencia de los animales, son organismos sésiles que carecen de la posibilidad de escapar ante situaciones adversas como el ataque de patógenos o el cambio de las condiciones ambientales. Sin embargo, poseen mecanismos de defensa que van desde barreras físicas (películas de cera en la superficie de sus órganos, paredes celulares rígidas, etc.), hasta potentes mecanismos moleculares de resistencia en cada célula y señales sistémicas provenientes del sitio de la infección que tienen marcadas similitudes con la inmunidad innata de los animales. El ataque por patógenos genera una condición desfavorable en la planta que activa una serie de mecanismos de defensa cuyo fin es detener, aminorar o contrarrestar la infección. A lo largo de las generaciones, patógenos y plantas se encuentran enfrentados en una lucha sin fin, en la que las plantas evolucionan para escapar de la infecciones de los patógenos y los patógenos evolucionan para escapar de las defensas de la planta, generando la antes mencionada "carrara armamentista". Aquellos genes involucrados directamente en la interacción, probablemente evolucionan de una manera adaptativa, por medio de la acción de la selección positiva.

Anteriormente (**Capítulo II**), se identificaron 224 genes en *C. graminicola* bajo la acción de la selección positiva, indicando que podrían estar involucrados en la interacción planta-patógeno. De la misma manera, es de esperar que genes de su huésped (maíz), involucrados en la defensa contra el hongo *C. graminicola* y otros patógenos, evolucionen por medio de la acción de la selección positiva. En este capítulo se pretendió determinar si genes involucrados en la defensa del maíz frente al ataque de *C. graminicola*, muestran evidencias de evolución adaptativa a lo largo del linaje de las Poáceas.

El genoma del maíz (*Zea mays*) contiene alrededor de 32.500 genes, y más de 600 de ellos codifican proteínas con dominios relacionados en la defensa ante el ataque de patógenos. La identificación y caracterización de genes de la planta involucrados en la activación de la respuesta de defensa vegetal es una tarea difícil, pero resulta fundamental para el desarrollo de nuevas estrategias de aplicación biotecnológica para el control de enfermedades vegetales.

En un estudio anterior, se identificaron alrededor de 200 genes de maíz expresados durante las primeras etapas de desarrollo de la antracnosis. Entre ellos, se encontraban 36 genes que codifican proteínas con dominios asociados a defensa (*domains associated with defense related genes*, DRGs) y 34 genes que codifican proteínas hipotéticas o desconocidas (*hypothetical or unknown proteins*, HUPGs). Con el objetivo de analizar evidencias de selección positiva en estos genes, se utilizó la base de datos Phytozome para obtener las secuencias ortólogas de cada gen, presentes en todas las especies disponibles dentro de la familia de las poáceas (*Sorghum bicolor*, *Setaria italica, Oryza sativa* and *Brachypodium distachyon*). Luego se aplicó una serie de filtros para evitar el análisis de secuencias demasiado divergentes que saturaran el test de selección y se procedió a realizar dos tipos de test de selección positiva, ambos basados en la estimación de ω (dN/dS) (**Figura 33**). En primer lugar se aplicaron los modelos de Markov (*site models*) implementados en PAML para determinar grupos de ortólogos que se ajustaban mejor a modelos evolutivos que asumían selección positiva. Aquellos grupos de ortólogos que mostraban evidencia de selección positiva de acuerdo a los modelos de PAML (14 grupos), fueron "expandidos". La expansión consistió en asignar secuencias ortólogas adicionales al grupo originario a partir de búsquedas de BLAST en NCBI en otras especies dentro de las Poáceas por medio. El nuevo set de ortólogos fue analizado nuevamente mediante los "*site models"* de PAML y, adicionalmente, fueron analizados en busca de eventos episódicos de selección a través de MEME (Mixed Effects Model of Evolution). La principal característica de MEME con respecto a los "*site models*", es que MEME presenta mayor poder debido a que puede detectar incluso eventos episódicos de selección positiva, es decir sitios que muestran valores elevados de ω a lo largo de algunos linajes, mientras que en el resto de linajes el sitio puede encontrarse bajo neutralidad o selección negativa. Mientras que los "*site models*" sólo detectan selección en un sitio si es evidente en todas las especies analizadas, MEME es capaz de identificar sitios en los que la selección ocurre a lo largo de unos cuantos miembros del grupo.

Como resultado, se identificaron seis genes con evidencia de selección positiva y otros ocho con los sitios que muestran selección positiva episódica (**Tabla 28**). Algunos de ellos, ya habían sido descriptos anteriormente como bajo la acción

de la selección positiva, como en el caso del gene codificante de la quitinasa clase III involucrada en la degradación de la pared celular fúngica, o en miembros de la familia PR5, conformado por proteínas del tipo "*thaumatin-like*" con conocidas funciones antifúngicas e insecticida. Por otra parte, también se identificaron genes bajo selección positiva que no habían sido descriptos anteriormente. Entre ellos una isocitrato liasa que en plantas se encuentra exclusivamente asociada al ciclo del glioxilato. El rol específico del metabolismo del glioxilato en relación a la defensa de las plantas no se ha determinado completamente, pero se ha sugerido que tendría un papel fundamental en la activación de los mecanismos de defensa durante la interacción con patógenos.

Entre los genes con evidencia de selección episódica, se encontraron dos proteínas con funciones bien conocidas en patogenicidad, como lo son la PR1 y PR5 de maíz. Siete de los 14 genes con evidencia de selección positiva codifican proteínas hipotéticas o de función desconocida. Estos genes resultan particularmente interesantes ya que podrían representar secuencias involucradas en la interacción planta-patógeno que aún no han sido identificadas como tales. Un análisis bioinformático más profundo, basado en búsquedas de similitud estructural y de dominios proteicos conservados, permitió determinar funciones putativas para algunas de las proteínas codificadas por estos genes. Entre los dominios y funciones putativas identificadas para estas proteínas se encuentran una clavaminato sintasa involucradas en la biosíntesis de ácido clavulánico, un transportador de metabolitos secundarios, una proteína serina/treonina kinasa, una proteína con repeticiones ricas en leucina (LRR por sus siglas en inglés) y un transportador de amonio. En general, prácticamente todas las secuencias de función desconocida bajo selección positiva mostraron funciones putativas relacionadas con la respuesta defensiva ante el ataque de patógenos.

La reconstrucción tridimensional de algunas de las proteínas bajo selección positiva y la posterior identificación de los aminoácidos identificados como los más afectados por la acción de la selección (**Figura 34**), mostró que los residuos más afectados suelen estar sobre la superficie proteica, indicando que esos sitios podrían estar involucrados en las interacciones proteína-proteína.

En general, en este capítulo se presentan un total de 14 genes con evidencias de selección positiva en el linaje de las Poáceas. Estos genes, están sobre expresados en maíz durante el desarrollo de la antracnosis causada por el hongo *C. graminicola*. Algunos de ellos, ya han sido identificados anteriormente como evolucionando rápidamente, mientras que otros se describen en este trabajo por primera vez. Las evidencias sugieren que estos genes han estado involucrados en la interacción planta-patógeno durante millones de años, participando en una coevolución antagonista como resultado de la carrera armamentista, sugiriendo que los productos de estos genes podrían estar interactuando con efectores producidos por los patógenos o estar implicados en vías metabólicas importantes para la defensa. Esta información, además de contribuir a una mejor comprensión de los mecanismos moleculares implicados en la defensa de las Poáceas, provee un conjunto de candidatos importantes para validación funcional a través de estudios bioquímicos y genéticos con el fin de identificar dianas que podrían tenerse en cuenta en programas de mejoramiento vegetal en la búsqueda de nuevos compuestos fitosanitarios.

1. INTRODUCTION

1.1. Bioinformatics, genomics and molecular evolution applied to agriculture

The global population is estimated to grow to 9.1 billion by the year 2050, and one of the most important challenges that we face is feeding the ever growing population (Godfray et al. 2010; Hubert et al. 2010). Population growth is already driving demand for food and other agricultural products such as animal feed and fuel (Gupta et al. 2010). Furthermore, an increasingly demanding society for more nutritious foods and the need to expand cultivation to new geographical areas, confronting new biotic and abiotic hazards makes classical agriculture insufficient to supply the worldwide demand for food and plant commodities (Altman & Hasegawa 2012). Agricultural and plant biotechnologies, stimulated by the new omics technologies, the improvement of genetic transformation, the increasing understanding of plant biological processes and the accelerated growing of molecular techniques appear as an important, and maybe the only, alternative for supplying the current and future demand for agricultural production (Altman & Hasegawa 2012).

The era of omics including genomics, transcriptomics, epigenomics, and proteomics has led to a massive explosion in the amount of biological information. Such a large quantity of data has resulted in the birth of a new interdisciplinary scientific area composed by biology, mathematics and computer science, known as Bioinformatics. Main areas in bioinformatics consist in the application of computational tools to the processing and managing of data generated by biological experiments, which have had a profound impact on different fields of biology like human health, agriculture, environment, energy and biotechnology (Singh et al. 2011). A complementary field of science, Computational Biology, is mainly focused on the development and application of analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems (McCarthy 2010).

There are currently almost as many subject areas in bioinformatics as kind of biological data, experiments or hypotheses to be tested. Major research areas include sequence analysis, genome annotation (functional genomics), computational evolutionary biology, gene expression analysis, comparative

genomics, network and systems biology, high-throughput image analysis and structural bioinformatics. However, in agriculture, bioinformatics is still in its infancy, since the genomes of most agriculturally important species have only been sequenced recently and have a poor functional annotation, plus the fact that agricultural research communities are much smaller compared with the communities working in model organisms (McCarthy et al. 2006). Next Generation Sequencing (NGS) technologies have revolutionized the fields of sequence analysis, genome annotation and comparative genomics in the recent years, even for agronomically important species, since NGS provides the opportunity to small research groups to obtain quickly and at low cost genomic sequences (**Figure 1**).

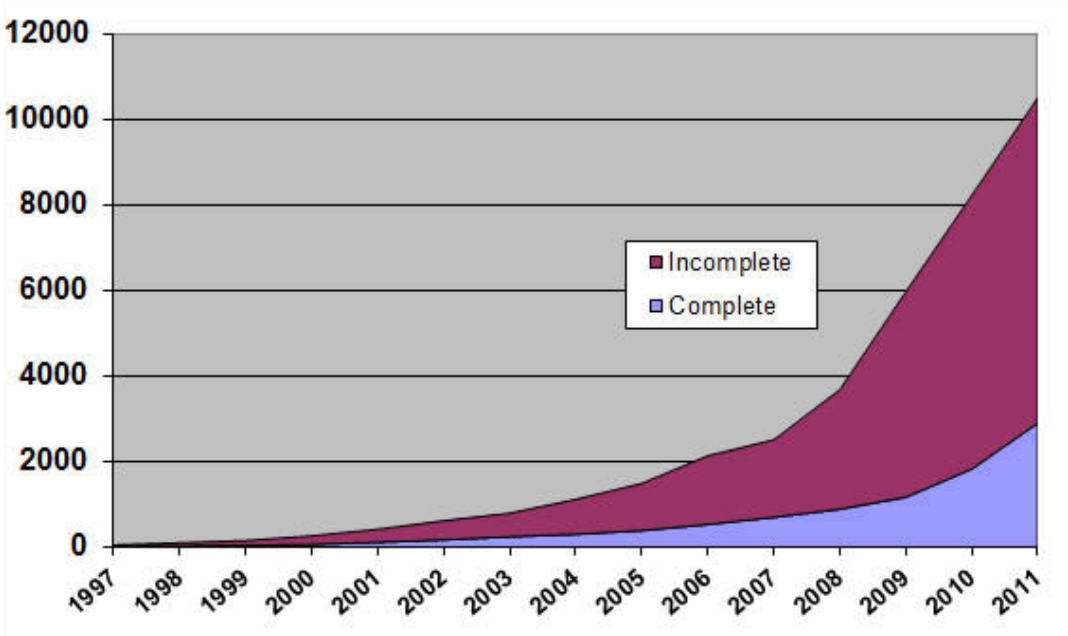

Figure 1. Total number of genome sequencing projects in GOLD (Pagani et al. 2012). A total of 11,472 projects were being monitored by this database at October 2011.

In September 1997, the International Rice Genome Sequencing Project (IRGSP) started an international collaboration to sequence the rice genome (Eckardt 2000), one of the most important crops in the world. Since then, hundreds of genomes of agronomically important plants, animals and pathogens species have been sequenced, providing enormous benefits for the agricultural community (McCarthy et al. 2011). Bioinformatic tools have been developed and used to analyze the genome sequence in order to search for important genes and

genomic traits, facilitating the development of healthier, more productive and disease resistant crops (Singh et al. 2011).

The first plant genome to be completely sequenced was from the model organism *Arabidopsis thaliana* (Initiative TAG 2000), which also actually represent the best quality sequenced and assembled plant genome. Despite its lack of economic, agronomic or environmental importance, *A. thaliana* has become an essential tool for understanding the molecular biology of many plant traits, including development and response to biotic and abiotic stresses. Since release of its genome sequence, a huge amount of data has started to be produced for this organism including gene predictions, gene ontology, protein function prediction and whole genome microarrays. In addition, a wide range of bioinformatics resources and databases have subsequently been developed to visualize and analyze these data (see Graham & May 2011), providing a valuable resource for all researchers in the field. In addition, the Arabidopsis community launched "The 1001 Genomes Project" in 2008 (www.1001genomes.org), which aims to sequence and compare the genomes of 1,000 accessions of *Arabidopsis thaliana*, and will provide an unprecedented resource for the analysis of the whole-genome sequence variation of this reference plant.

Recently, complete genomic sequences for many crop plants have also started to be available such as Rice (IRGSP 2005), Papaya (Ming et al. 2008), Sorghum (Paterson et al. 2009), Maize (Schnable et al. 2009), Soybean (Schmutz et al. 2010), and many others are in the process of being sequenced. The availability of complete genome sequences for many plants has allowed first comparative approaches aimed to identify important genomic traits in the extant cultivars species. The estimation of rates of evolution of genes and gene families, the identification of differential gene loss or retention following duplications and the implication of chromosomal rearrangements shaping genome organization, have collectively contributed to improve the understanding of taxonomic, morphological and physiological variations between cultivated plants (Mehboob-ur-Rahman & Paterson 2009).

One of the most important findings in comparative genomics of plants was that the organization of genes generally remain conserved over evolutionary time, providing an useful framework for inferring correspondence among even distantly-related species (Tang et al. 2008). Despite this, the complexity of plant genomes (usually large, polyploid and full of repetitive DNA) still represent a great challenge to sequencing technologies and bioinformatics. In addition, genome sequences alone may not tell us much about where within the genome to focus our attention. Bioinformatics and computational biology play a fundamental roles during the analysis of such amount of data, helping to the scientific community in the hard task of identify functionally important sequences. There are currently hundreds of on-line resources with applications in biology and agriculture; helping to make sense to the wealth of data that has been produced. Some of the most popular tools are listed in **Table 1**.

Of particular concern to agriculture are emerging infectious diseases (EIDs). EIDs are pathogens that are increasing in incidence, geographic range or host range, and virulence (Daszak et al. 2000; Jones et al. 2008). EIDs caused by fungi are increasingly recognized as presenting a worldwide threat to food security (Fisher et al. 2012; Jones 2013). EIDs are an important threat because our agricultural infrastructure is usually not prepared to control them. All disease control measures need time to be developed. New fungicides may need to be developed or existing fungicides need to be evaluates and approved for control of new diseases. New plant varieties carrying resistance to new diseases need years of breeding and evaluation before they are ready to be introduced for commercial production. Continuing research and development of EIDs are needed to maintain a high level of food security.

Table 1. Characteristics of some of the main on-line resources providing bioinformatics tools for the analysis of biological data in general and for agronomically important species.

Resource	Description	Tools	Website/Reference
General			
NCBI	Databases from The National Center for Biotechnology Information	Similarity searches against GeneBank, PubMed and others.	www.ncbi.nlm.nih.gov (Mizrachi 2007)
UniProt	Database of protein sequences and functional information.	Similarity searches, domains identification against Swiss-Prot and TrEMBL.	www.uniprot.org (The UniProt Consortium 2012)
Gene Ontology	Controlled vocabulary of terms for describing gene product characteristics.	Annotation, terms enrichment, proteins interaction.	www.geneontology.org (Ashburner et al. 2000)
GOLD	Genomes Online Database	Access to all genomes and metagenomes sequencing projects around the world.	www.genomesonline.org (Pagani et al. 2012)
Agriculture			
AgBase	Functional modeling resources for agricultural species.	Analysis of proteomics and other experimental data using the GO.	www.agbase.msstate.edu (McCarthy et al. 2011)
Gramene	Resource for Comparative Grass Genomics	Analysis of QTLs, metabolic pathways, genetic diversity, GO and markers.	www.gramene.org (Youens-Clark et al. 2011)
Phytozome	Comparative platform for green plant genomics.	Evolutionary history of plant genes. Functional annotations of complete genomes.	www.phytozome.net (Goodstein et al. 2011)
PlantGDB	Resource for comparative plant genomics.	Sequence data for >70,000 plant species, custom EST assemblies.	www.plantgdb.org (Duvick et al. 2008)
FungiDB	An integrated functional genomics database for fungi.	Comparative and functional genomic for fungi.	http://fungidb.org (Stajich et al. 2011)
Phi-Base	Pathogen-Host interaction Database	Sequence search against expertly curated database of genes proven to affect the pathogen-host interactions.	www.phi-base.org (McCarthy et al. 2011)
DFVF	Database of Virulence Factors in Fungal Pathogens.	Search for virulence factors and pathogen-host interaction mechanisms in fungi.	sysbio.unl.edu/DFVF/ (Lu et al. 2012)

The availability of whole genome sequences for pathogenic fungi has revolutionized the field of molecular plant pathology and the past few years has seen a great improvement in our ability to leverage comparative genomics to learn more about the function and evolution of fungal genomes (Galagan et al. 2005). In 2011, an international research team in collaboration with the Joint Genome Institute of the Department of Energy of the EEUU has embarked on a five-year project to sequence 1,000 fungal genomes from across the fungal tree of life (http://1000.fungalgenomes.org/). This resource will provide reference information for fungal comparative genomics, facilitating research on fundamental features of the plant-microbe interaction. The value of comparative genomics has already been demonstrated in the identification of important virulence genes with host specific functions (Schirawski et al. 2010; Spanu et al. 2010; O'Connell et al. 2012). However, the low taxonomic coverage of the fungal phylogenetic tree places limits on the types and scope of comparative genomic studies that can be performed. Studies of genome organization and tests of selection pressures in particular, are hampered by the large evolutionary gaps between sequenced taxa. To maximize the effectiveness of these types of comparative genomics, comparative analyses need to be within different strains of the same fungal species or between closely related species (Besnard & Christin 2010; Stukenbrock et al. 2011; Gibbons et al. 2012; Manning et al. 2013).

Initial analyses of fungal genomes have shown that fungi are highly divergent compared with higher eukaryotes, even among species that look morphologically and phylogenetically very similar. For example the genomes of three *Aspergillus* species are on about 68% similar (Galagan, et al. 2005). Significant sequence differences can be found even among different strains of the same species (Cuomo et al. 2007; Fedorova et al. 2008). Among plant pathogenic fungi, a pattern is emerging in which genes within specific regions of the genome tend to evolve more rapidly than others and genes with roles in plant disease, such as effectors (see below for a definition of effector) tend to be found within these regions (Cuomo et al. 2007; Thon et al. 2004). Increasing interest in the study of effectors, coupled with the increasing availability of genome sequences has caused the creation of a new term, effectoromics, which

is used to describe the use of comparative genomics to study this class of genes (Oh et al. 2009).

In addition to effectors, many plant pathogenic fungi also produce a wide diversity of secondary metabolites (SMs), organic compounds that do not have a role in primary metabolism. SMs have been shown to be important in virulence and host specificity. Fungi produce four major groups of SM: polyketides produced by polyketide synthases (PKS); peptides produced by nonribosomal peptide synthases (NRPS); alkaloids produced by dimethylallyl tryptophan synthases (DMATS); and terpenes produced by terpene synthases (TS) (Keller et al. 2005). These gene families are large, and most fungal genomes contain multiple copies of each. The genes encoding secondary metabolites show great interspecific and intraspecific variability in their presence and expression patterns, probably reflecting their importance in determining host specificity and virulence (O'Connell et al. 2012; Ohm et al. 2012).

Comparative genomics between pathogenic and non-pathogenic fungi may also provide evidence regarding genes or gene families unique to pathogens (Dean et al. 2005). An understanding of how pathogens and non-pathogens have evolved to fill their ecological niches will help us to identify which physiological biochemical and developmental process are truly unique to pathogens. This knowledge can be used to help identify targets (physiological processes and individual genes or gene products) for developing environmentally friendly control strategies that are effective against a broad array of pathogens and can leave non-pathogens unharmed.

Plant-pathogen genomics is one of the most promising areas in bioinformatics, comparative genomics and molecular evolution applied to agricultural systems (Atallah & Subbarao 2012; de Jonge 2012; Stukenbrock & Dutheil 2012). Studies of pathogen population genetics, including the effects of natural selection and recombination, are essential to describe the disease epidemiology and to face management strategies. The most important aspects regarding plant-pathogen interaction and population genetics/genomics of fungal pathogens are described below and in the "Introduction" sections of later chapters.

1.2. The plant-pathogen interaction (PPI)

At the moment, between 80,000 and 120,000 species of fungi have been identified, but it is estimated that there may be as many 5.1 million (O'Brien et al. 2005; Blackwell 2011). There are many different kinds of fungi with many different forms and are widespread all around. Based on their ecological role, fungi can be classified as saprotrophs, mutualistic symbionts, parasites, or hyperparasites. Most of the known fungal species are strictly saprophytic, i.e., they live on dead organic matter, which they help decompose. More than 10,000 species of fungi, however, can cause disease in plants (Agrios 2005). Pathogenic fungi can lead to economic losses estimated at more than 200 billion $US per year due to infected crops (Cuomo et al. 2007). Fungal diseases in humans have also become significant in number in the last few decades, infecting billions of people every year, and being deadly in many of the immunocompromised individuals (Brown et al. 2012).

One of the biggest problems with the development of effective strategies against fungal pathogens is that fungi adapt to a variety of lifestyles and therefore use different strategies to infect and colonize their hosts. For instance, plant pathogenic fungi can be obligate biotrophs (cannot survive in a dead host and therefore does not kill the host), necrotroph (obtains its nutrients from dead cells and tissues of its host organism) or hemibiotroph (parasites in living tissues for a period and continues its life cycle on dead tissues) (Parbery 1996). Additionally, fungi can be divided into two basic morphological forms, filamentous and yeast-like. Some fungi can exhibit either the yeast form or the filamentous form depending on environmental conditions or the stage of their life cycle and are called dimorphic (Soanes et al. 2008). These differences in lifestyles and morphological traits make finding targets for disease control difficult.

Why do pathogens attack plants? In the large number of filamentous fungi, only a few members are able to infect plants. These parasites are successful because they can invade, take food and proliferate in a host plant. The most parsimonious answer to this question is because during their evolutionary history they have acquired the ability to live off the nutrients produced by the host plants or, in some cases depend of them for survival. To access these

substances, pathogens must first penetrate the outer barriers formed by the cuticle and/or cell walls and finally arrive to the cell interior. In addition, the plant cell contents are not always readily usable by the fungus and must be digested to be utilizable. On the other hand, the plant detects the presence of the pathogen and drives its defense mechanism producing structures and chemical substances that interfere with the progress or existence of the pathogen. Therefore, for a pathogen to infect a plant, it must be able of overcome the structural barriers of the plant, obtain nutrients, and meanwhile neutralize antipathogen substance produced by the plant. In order to this, pathogens need to secrete a variety of chemical compounds which affect specific metabolic pathways of their hosts (Agrios 2005).

At the genomic level, there are three compatible processes that may explain the emergence of the many adaptations to the pathogenic lifestyle in fungi. First, parasitism is associated with the presence of novel genes. Such genes may have a specific role during host infection and could be acquired by some of the many mechanisms for the generation of new genes (for a review see Chandrasekaran & Betrán, 2008) such as gene duplication followed by functional divergence or horizontal gene transfer (e.g. Armijos Jaramillo et al. 2013). Second, adaptations to pathogenic habits may result from differences in the regulation of gene expression (King & Wilson 1975; Gasch et al. 2004; Aguileta et al. 2009). Third, the pathogenic life style is associated with gene loss and deletions (Dean et al. 2005). For instance, genes involved in a free-living lifestyle or those that allow the host to detect the pathogen are expected to be lost in pathogenic fungi (Tunlid & Talbot 2002; Aguileta et al. 2009).

When a parasitic relationship is established between a fungus and its host plant, a competition for survival is generated. At the population level, two evolutionary scenarios have been proposed to describe the genetic population dynamics: the “Red Queen” model and the “Arms Race” model. In the Red Queen model, two or more alleles of a pathogenicity gene exist in the pathogen population, and the allele with the largest contribution to pathogen fitness increases in frequency. At the same time, the host allele that can defend its host from this pathogenic gene also increases its frequency in the population (Terauchi & Yoshida 2010). This

processes lead to a non-directional evolution, in which both populations are polymorphic at all stages (**Figure 2**).

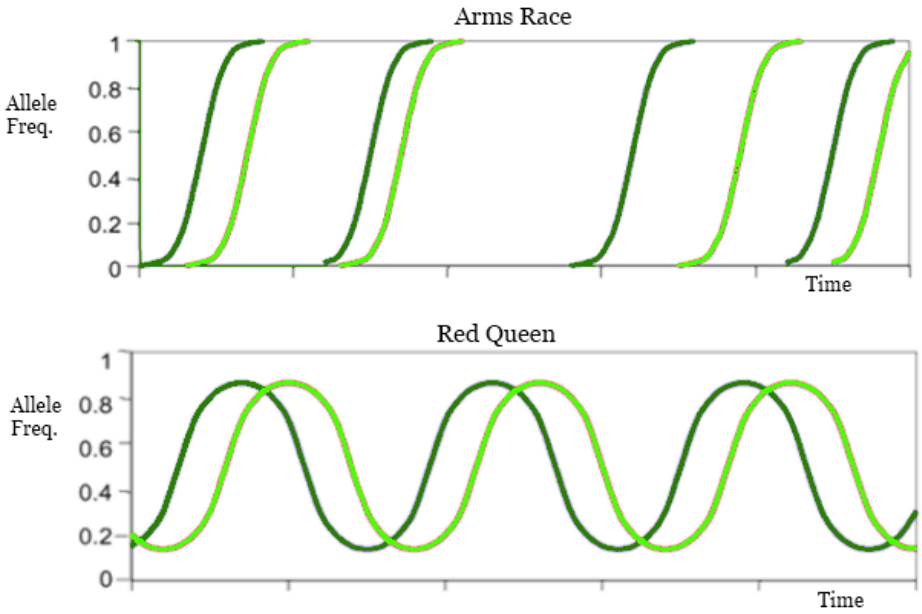

Figure 2. Allele frequency changes in the Arms Race (top) and Red Queen (bottom) models of host–pathogen interactions. Dark green line, the allele frequency of the pathogenicity related gene in the pathogen. Ligth green line, the allele frequency of the host defence gene. Modified from Woolhouse et al. (2002).

In the Arms Race model, host-pathogen interaction constantly selects for genetic changes in both populations. Genetic variations that improve fitness in both, pathogens (as the capacity to avoid host detection) and hosts (as the ability to recognize pathogens), will be maintained in the population (Stahl & Bishop 2000; Maor & Shirasu 2005). Selection is directional, since genetic change accumulates in both populations. At any given stage of the process there may be polymorphisms in either, both or neither of the two populations (Woolhouse et al. 2002). These phenomena require rapid adaptations and a continual evolutionary progression in those protein-coding genes directly involved in the interaction, which may leave footprints in the genome in the shape of highly variable coding sequences, reflecting rapid evolution (Takken & Rep 2010). In many cases, rapid divergence is driven by adaptive evolution (positive Darwinian selection) owing this, the detection and characterization of natural selection involved in plant–pathogen coevolution stands as a very important and exciting research topic.

The Red Queen and the Arms Race scenarios may leave contrasting patterns of DNA polymorphisms and divergence in the genomes of both hosts and pathogens, which can be detected using statistical tests developed for molecular population genetics studies (Terauchi & Yoshida 2010). Particularly interesting to researchers of molecular plant–microbe interactions is the coevolution of pathogen genes encoding for effectors vs. host genes effector target proteins and R proteins (Terauchi & Yoshida 2010). Effector proteins under positive selection have been identified in viruses, pathogenic bacteria, oomycetes and fungal species. In bacteria and viruses, a significant number of studies have identified proteins exposed at the surface under positive selection, which could be indicating that they are probably involved in the evolutionary arms race (Aguileta et al. 2009). In oomycetes and fungi, a number of toxins, Avr products, degradative enzymes and reproductive related proteins were also identified as being under positive selection.

Coevolution is defined as the change of a biological object triggered by the change of a related object (Yip et al. 2008). At the genetic level, coevolution is a dynamic process, which implies reciprocal changes in the genetic composition of the species that are involved. During the antagonistic coevolution between a plant and its pathogen, at the population level pathogens may reduce plant fitness forcing the selection of novel defense strategies that, in case of being effective, will be spread through the plant population. On the other hand, effectively defended plants will decrease pathogen fitness, thus selecting for a genotype that can overcome the defense, which then spreads through the pathogen population (Stahl & Bishop 2000). This struggle for survival between populations causes genes encoding proteins involved in the interaction to exhibit rapid evolution resulting in a larger amount of amino acid replacements between the species, meanwhile a selective sweep reduces the genetic variation (intra-species polymorphism) in the regions tightly linked to the selected sites (Terauchi & Yoshida 2010).

During evolution, plant pathogens have developed several strategies for the localization and invasion of their hosts. The recognition is usualy conducted by the identification of the host topography or their chemical compounds. The entrance is usually achieved through natural openings and wounds, but in many

cases pathogens can directly penetrate the cell wal by mean of the secretion of enzymes or by the generation of high turgor pressure in specialized structures (Agrios 2005). Against this pathogen attack, it is not surprising that plants have developed mechanisms to defend themselves. Some defenses are constitutive while others are induced by the challenging organism. The induced responses are often triggered by recognition of the invader and are usually followed by a hypersensitive response (HR) that causes localized cell death, keeping the pathogen localized to the primary infection site. The responsible genetic elements for the plant and the pathogen are being termed resistance (R) and avirulence (Avr) genes, respectively (Flor 1942; Van Der Biezen & Jones 1998; Strange 2003; Rouxel & Balesdent 2010).

Many Avr genes have been cloned from different fungal species. They usually encode small secreted proteins (63–314 aa), are often rich in cysteines and show no sequence similarity among each other (Rouxel & Balesdent 2010). Most characterized R genes encode proteins having a leucine-rich-repeat (LRR) domain and one or many nucleotide-binding sites (NBSs) (Xiao et al. 2008). These plant proteins, belonging to the nucleotide-binding site-leucine-rich repeat (NBS-LRR) family, are usually involved in the pathogen detection. In these proteins, the NBS is typically highly conserved and the LRR region is highly variable and it is involved in recognition and specificity (DeYoung & Innes 2006).

There are multiple means by which the gene products involved in the PPI may be brought together and interact. The “gene-for-gene” (GFG) model, which was first described by Flor (Flor 1955), provides a good scheme for understanding the coevolution between plants and their pathogens. Flor claimed that for every gene determining resistance in the host, there was a corresponding and complementary gene for avirulence in the pathogen. A plant variety can carry the gene for resistance (R) or lack it, carrying the gene for susceptibility (r) to a pathogen. In similar way, a pathogenic strain can carry the gene for avirulence (A) or the gene for virulence (a) against the resistance gene (R). Thus, depending on the kind of genes in the interaction, there are four possible gene combinations and reaction types. A host will be resistant only if the R gene-encoded host receptor recognizes the pathogen’s Avr gene product, which will

trigger the defense reactions (**Figure 3.a** - upper left interaction). When an incompatible interaction occurs, it is because the host lacks the specific receptor for the Avr product. The R gene-encoded by the host finds no elicitor to recognize, the pathogen's Avr product is not produced or no host receptor is present (**Figure 3.a** - other interactions), no defense reactions are triggered, so the host lacks resistance to this pathogen (Agrios 2005; Stukenbrock & McDonald 2009).

Although the GFG model provides a good scheme for understanding the genetic basis of the PPI, usually a more complex range of interaction mechanisms involving a number of genes in both partners is found. In addition, since direct interaction between R and Avr proteins could often not be demonstrated experimentally, it was recognized that R proteins may also monitor the state of host components targeted by pathogen Avr molecules to establish disease (de Jonge et al. 2011). On the other hand, cells of the innate immune system of plants recognize broadly conserved microbial "elicitors" associated with groups of pathogens. These molecules have been named pathogen-associated molecular patterns (PAMPs) and microbe-associated molecular patterns (MAMPs) (Janeway 1989; Boller 1995; Janeway & Medzhitov 2002). The PAMPs recognition by the host results in PAMP-triggered immunity (PTI) and restricts pathogen development or kills it (Rouxel & Balesdent 2010).

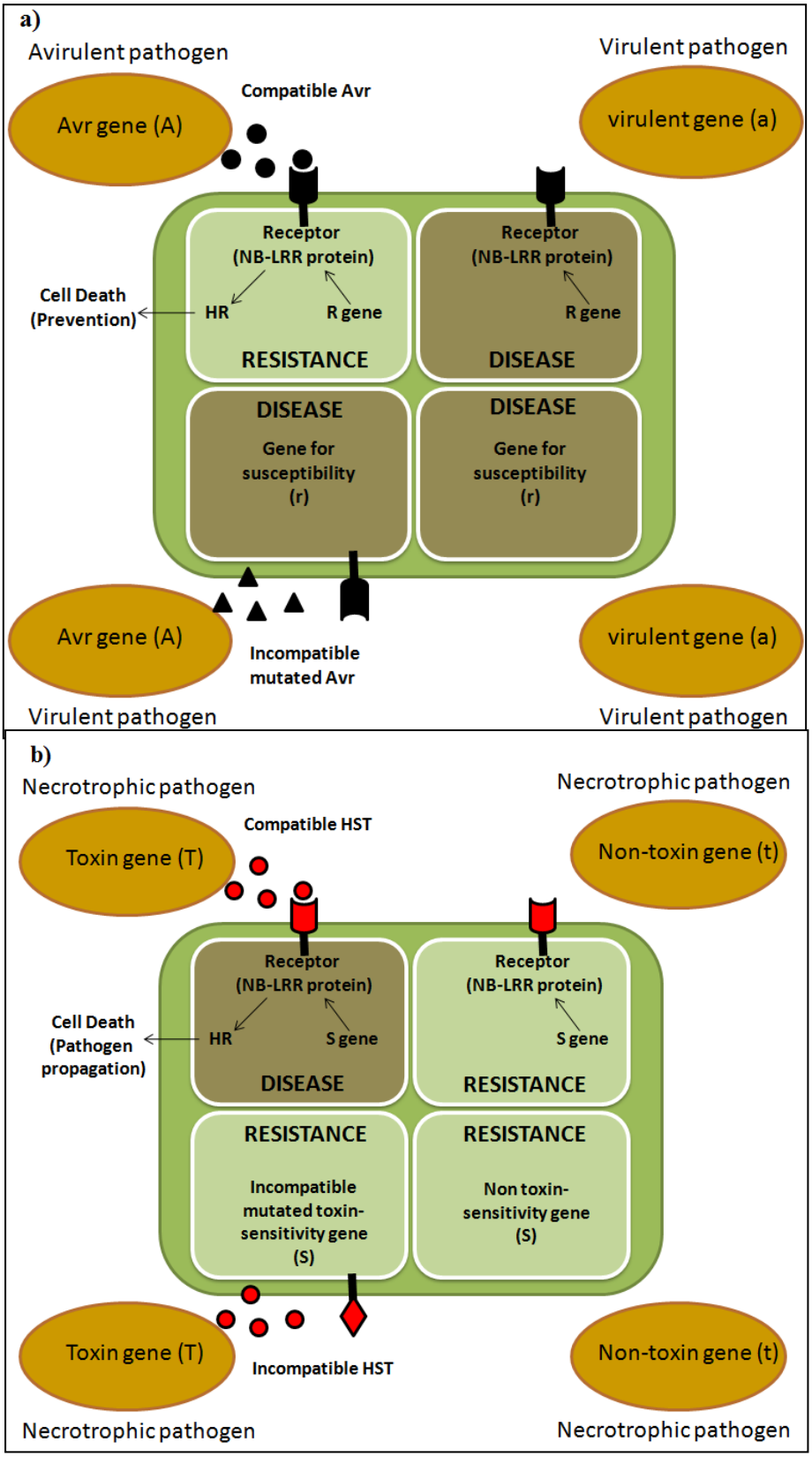

Figure 3. Gene-for-gene interactions based on a) an avirulence protein and b) a host-specific toxin. Figure adapted from Agrios 2005 and Stukenbrock & McDonald 2009.

Under the GFG model of interaction, the fungus would produce an elicitor, encoded by an avirulence gene, which directly or indirectly triggers a defense reaction in the host, characterized by a HR. Because PAMPs are essential for the normal function of the pathogen, the conventional wisdom was that they cannot be lost or significantly altered without affecting pathogen fitness. However, some recent publications have demonstrated that in some flagellated pathogenic bacteria, even PAMPs may be able to evolve to escape host recognition, making it easy to think that in fungal pathogens something similar could happen (Stukenbrock & McDonald 2009). So, in order to avoid recognition by the host R locus, avirulence genes are expected to evolve towards loss of function through either mutations leading to stop codons, insertions of transposable elements within the gene, gene deletion, or protein-coding genes modifications (Aguileta et al. 2009). Under this scenario, is expected that positive selection acts to foster mutations in the Avr genes to enable evasion of R gene recognition and plant R genes co-evolve to recognize new Avr products also through positive selection, R gene duplication or recombination (Stukenbrock & McDonald 2009).

Recently, a new consensus has emerged about the origin, evolution and biological role of a special class of pathogen effectors, the Host-Specific Toxins (HSTs), which are critical for virulence in necrotrophic fungal pathogens. As their name suggests, these are molecules that are toxic only to specific genotypes of the host and are innocuous to the great majority of other plants.. The GFG interaction between HST and their corresponding toxin-sensitivity genes (S) is similar to that of Avr and R genes, but in this case the compatible interaction between the HST and the toxin receptor is required for the pathogen propagation (**Figure 3.b** - upper left interaction). However, when an incompatible interaction occurs between the HST and its host receptor (when the host lacks the specific receptor, no pathogen HST is produced or no host receptor is present) the host will be resistant to the necrotrophic fungus (**Figure 3.b** – other interactions). In this case, positive selection acts favoring mutations in genes encoding host toxin-sensitivity (S) to avoid binding the pathogen's toxin and, as a result, pathogen HST genes coevolve though positive selection to recognize new S genes.

Differences between the outcome of Avr and HST in the GFG interactions are probably linked to the different strategies pathogens use to obtain nutrients. An Avr-induced resistance leads to a HR and local cell death through apoptosis restricts access of biotrophic pathogens to nutrients and water. For necrotrophic pathogens, HR and apoptosis provide easily acquired nutrients for further proliferation in the host. Thus, an HST produced by a necrotroph can operate by activating an R gene (and the corresponding HR pathway) that protects plants from biotrophs but allow the further proliferation of necrotrophic pathogens (Stukenbrock & McDonald 2009).

The usage of terms as “effectors”, “elicitors” and “virulence factors” has become popular in the field of PPI, but there is no clear definition of these terms yet. For example, various scientific communities define effectors differently. According to Hogenhout et al. (2009), who are for a broader, inclusive definition of the term, effectors are all proteins and small molecules produced by the pathogen that alter host-cell structure and function. These alterations either facilitate infection (virulence factors and toxins) or trigger defense responses (avirulence factors and elicitors) or both. This broader definition of effectors includes many molecules, such as PAMPs, toxins, and degradative enzymes. However, Dodds & Rathjen (2010) use the term “elicitor” to refer to both PAMPs and effectors, because these are molecules that induce (“elicit”) an immune defense response. Finally, Hogenhout et al. (2009) propose that in the absence of enough information, it would be suitable to call all of them “effectors” until the exact activities of the pathogen molecule is revealed, after which it may be renamed to reflect their specific activities.

A recent review (de Jonge et al. 2011) summarizes the current knowledge of effectors in filamentous fungi and the role of effectors in the interactions between fungi and their host plants (**Figure 4**). Under their model, fungi secrete effectors in the interface between pathogen and host after host penetration (1). Some effectors protect the hyphae against the hydrolytic enzymes secreted by the host (2) or inactivating them (3). Some effectors act as scavenger molecules searching for potential PAMP molecules (4) that may alarm host defense (5). Many effectors are translocated inside of the host’s cell

(6), where they may affect cytoplasmic processes related to host defense (7) or may go to the nucleus where they may regulate the transcription of target genes (8). The host's cell has various kinds of receptors on the cell surface (9) and in the cytoplasm (10) which, in case of detecting the presence of the pathogen, will trigger the defense response (**Figure 4**).

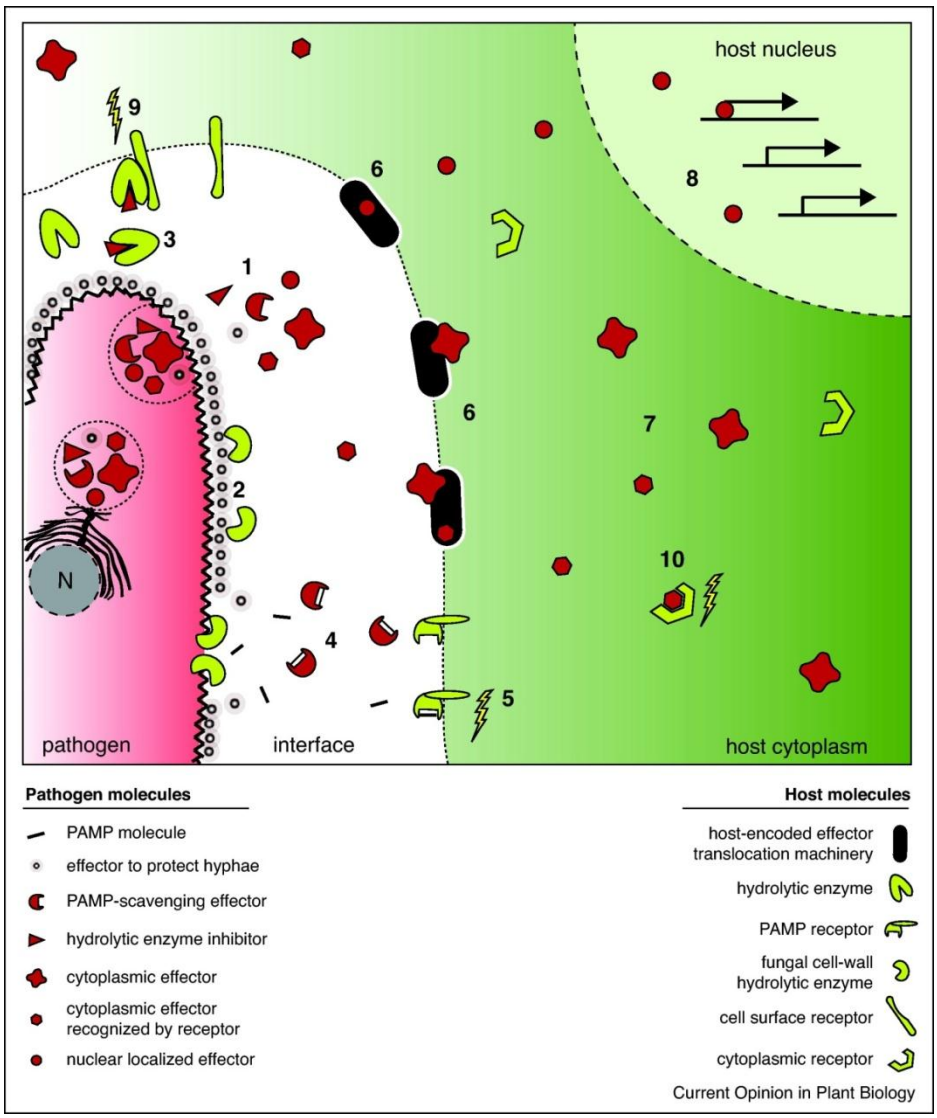

Figure 4. The role of multiple effectors during the interaction between fungi and host plants. Taken from de Jonge et al., 2011.

As predicted by the arms race model, several phytopathogen effector genes and their plant targets have been found to be under positive selection (Bishop et al. 2000; Rech et al. 2012; Win et al. 2007; Stergiopoulos et al. 2007; Van der Merwe et al. 2009; Van de Wouw et al. 2010; Pedersen et al. 2012). To date, most representative categories of detected genes under positive selection are

those involved in adaptation to contrasting environments, arms race adaptation, sexuality, hybrid inviability and sensorial perception, but positive selection has also been detected in housekeeping genes (Viscidi & Demma 2003; Lefebure & Stanhope 2009). More recent studies have taken advantage of the availability of whole genome sequences to identify genes under positive selection without *a priori* gene candidates (Aguileta et al. 2010; Stukenbrock et al. 2011; Aguileta et al. 2012; Gibbons et al. 2012). This approach becomes increasingly feasible as more complete genome sequences are available. In the case of PPI, this approach could help us to detect the most rapidly evolving genes in host and pathogen genomes. These genes will potentially be involved in the plant–pathogen struggle or in pathogen specialization on new hosts (**Table 2**). In pathogens, such genes may be involved in evading host defenses or generating novel mechanisms of infection and its detection can be important for a better understanding of the PPI (Aguileta et al. 2009).

Closer links between population genomics and experimental studies will provide us with the essential tools to describe and predict the emergence, establishment, and adaptation of plant pathogens to agro-ecosystems. A combination of genome-wide evolutionary analyses in crops and pathogens and functional studies, will provide fundamental resources to design novel and environmentally friendly strategies to slow down the emergence and spread of pathogens (Stukenbrock & Bataillon 2012).

Table 2. A selected set of genes showing evidence of positive selection in pathogenic fungi, divided by functional categories.

Gene	Function	Organism	References	Observations
Avirulence Genes				
Avr4	Triggers defense response.	*Cladosporium fulvum*	(Stergiopoulos et al. 2007)	-
AvrLm1 **AvrLm6**	Triggers defense response.	*Leptosphaeria maculans*	(Van de Wouw et al. 2010)	90% worldwide population lacked AvrLm1 gene.
Avr-Pita1	Dispensable for virulence on rice.	*Magnaporthe oryzae*	(Dai et al. 2010)	Presence of indels in virulent isolates.
AvrP4 **AvrL567** **AvrM**	Induce cell death and necrotic responses.	*Melampsora spp.*	(Van der Merwe et al. 2009; Catanzariti et al. 2006; Dodds et al. 2006)	Small secreted protein. Expressed in haustoria.
Avr_{k1}	Establishment of haustoria.	*Blumeria graminis*	(Sacristán et al. 2009)	Lacked secretion signal peptides.
Mycotoxin-related Genes				
Trichothecene cluster	Inhibit protein synthesis.	*Fusarium spp.*	(Ward et al. 2002)	Act as Avr genes and toxin.
NEP-like genes.	Induce necrosis and ethylene production.	*Botrytis spp.*	(Staats et al. 2007)	Act as host-specific toxin.
NIP1	Induce necrosis.	*Rhynchosporium secalis*	(Schürch et al. 2004)	Strains lacking NIP1 or with a specific point mutations overcome barley resistance.
SnToxA	Induce necrosis.	*Pyrenophora tritici-repentis* *Phaeosphaeria nodorum*	(Stukenbrock & McDonald 2007)	Selection imposed by the host.
Self-recognition and Sexual Compatibility genes				
HET Genes	Heterokaryon incompatibility.	*Neurospora crassa* *Podospora anserina*	(Paoletti et al. 2007)	Selection acting at the WD-repeat domain.
MAT Genes	Control of sexual reproduction.	*Fusamium spp.* *Neurospora spp.*	(S. H. Martin et al. 2011) (Wik et al. 2008)	Positive selection discovered in heterothallic species.

Other pathogenicity-related genes				
β-xylosidase	Plant cell wall degrading enzyme.	*Mycosphaerella graminicola*	(Brunner et al. 2009)	Positive selection in domesticated wheat pathogens
Poly-galacturonase (Bcpg1 and Bcpg2)	Plant cell wall degrading enzyme. Tissue maceration,	*Botrytis cinerea*	(Cettul et al. 2008)	Both genes are required for full virulence.
Hydrophobins	Confer hydrophobicity to fungal surfaces.	*Paxillus involutus* *Trichoderma spp*.	(Rajashekar et al. 2007; Kubicek et al. 2008)	Evolution according to the birth-and-death model[1]
Laccases	Biodegradation and alteration of lignified substrates.	*Basidiomycets group*	(Levasseur et al. 2010)	-

[1] Birth-and-Death Model of Evolution: In this model of evolution, duplicate genes are produced by tandem or block gene duplication, and some of the duplicate genes diverge functionally but others become pseudogenes owing to deleterious mutations or are deleted from the genome. The end result of this mode of evolution is a multigene family with a mixture of divergent groups of genes and highly homologous genes within groups plus a substantial number of pseudogenes (Nei et al. 1997).

1.3. *Colletotrichum graminicola* as a model organism

Among plant diseases caused by fungi, Anthracnose is one of the most destructive, causing significant crop losses on all five continents. Anthracnose disease symptoms consist mainly of necrotic lesions affecting leaves, stems, flowers and fruit, but may also have an impact on stem rots and seedlings (Agrios 2005; Cannon et al. 2012) (**Figure 5**).

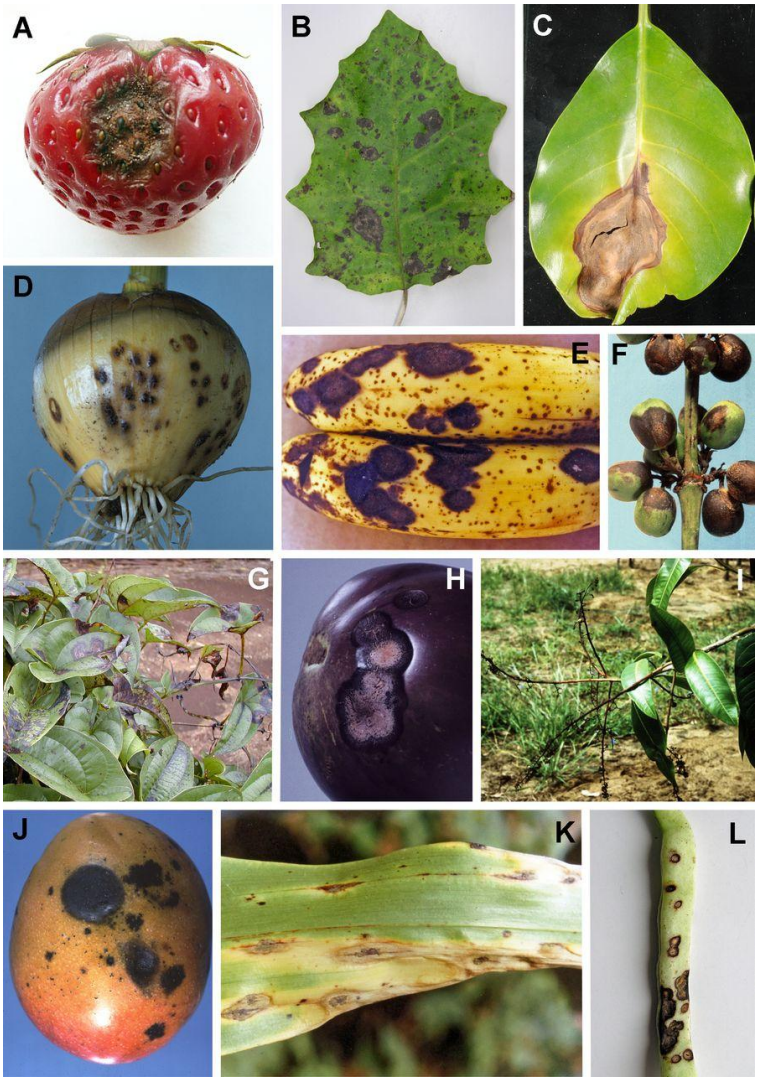

Figure 5. Disease symptoms caused by *Colletotrichum spp.* on different plant species. A) Strawberry fruit (*C. nymphaeae*). B) Leaf of Rangiora (*C. beeveri*). C). Leaf of *Tecomanthe speciosa* (*C. boninense*). D) Onion bulb (*C. circinans*). E) Banana (*C. musae*). F) Coffee berry (*C. kahawae subsp. kahawae*). G) Leaf of yam (*C. gloeosporioides*). H) Aubergine - Eggplant (*C. gloeosporioides*). I) Leaf of Mango (undetermined *Colletotrichum sp*). J) Mango (*C. gloeosporioides*). K) Leaf of maize (*C. graminicola*). L) Bean pod (*C. lindemuthianum*). Taken from Cannon et al. (2012).

The most well-known etiologic agents for anthracnose are species of the filamentous fungus *Colletotrichum*. *Colletotrichum* is the anamorphic stage of the genus *Glomerella* and belongs to the Ascomycota, Pezizomycotina, Sordariomycetes, Hypocreomycetidae, Glomerellales, Glomerellaceae (Reblova et al. 2011; Zhang et al. 2006) (**Figure 6**).

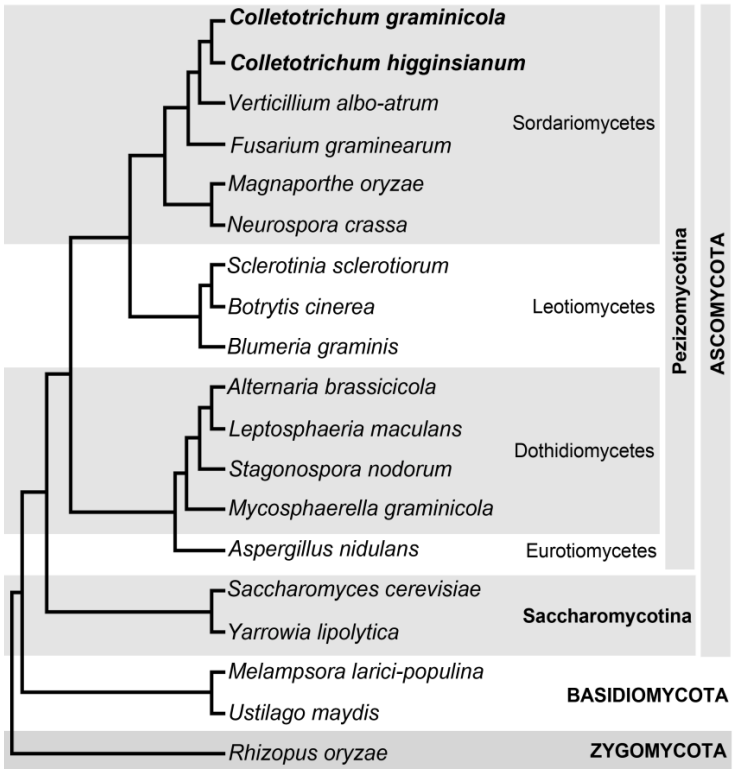

Figure 6. Whole-genome cladogram showing the phylogenetic relationships of *Colletotrichum spp.* and 17 other sequenced fungi. Taken from O'Connell et al. (2012).

Colletotrichum spp. infect virtually every plant family that is agronomically or horticulturally important (Hyde et al. 2009; Prusky et al. 2000), causing significant crop losses on all five continents. Numerous species of this pathogen are attributed to the development of disease in several crops of valuable economic importance in the Mediterranean region including cucurbits, pepper, strawberries, beans and olives (Wasilwa 1993; Freeman & Katan 1997; Méndez-Vigo et al. 2005; Talhinhas et al. 2005).

Colletotrichum fungi are important as experimental models in studies of many aspects of plant disease, including fungal development, carbohydrate degrading enzymes, infection processes, host resistance, and the molecular biology of

plant-pathogen interactions (Thon et al. 2002; Sukno et al. 2008; Wattad et al. 1994; Wattad et al. 1995; Perfect et al. 1999; Latunde-Dada 2001; Vargas et al. 2012). Some of the earliest research of host-pathogen interactions and fungal elicitors of pathogenicity was conducted using *Colletotrichum* pathosystems (Dixon & Lamb 1990; Wijesundera et al. 1989; Lawton & Lamb 1987) and today dozens of laboratories around the world are studying the biology and pathology of various species of *Colletotrichum*. In addition, *Colletotrichum spp.* are model organisms for the study of hemibiotrophic pathogens, those that begin their infection as biotrophs (keeping the host cell alive) but later switch to a necrotrophic lifestyle, killing their hosts and feeding on dead cells (Bergstrom & Nicholson 1999; Vargas et al. 2012; O'Connell et al. 2012) (**Figure 7**).

One of the most important species of the genus *Colletotrichum* is *C. graminicola*, which causes anthracnose stalk rot and leaf blight of maize (*Zea mays*) (LeBeau 1950; Jamil & Nicholson 1991), producing annual yield losses of more than 1 billion dollars in the U.S.A. alone (Frey et al. 2011) and has a great potential to damage agricultural and natural ecosystems (Kamenidou et al. 2013). Maize - *C. graminicola* is one of the most intriguing pathosystems. *C. graminicola* is among the best characterized and most tractable *Colletotrichum* species. It is one of very few in which sexual crosses can be made and is easily cultured and stored and experiments like transformations, gene disruptions and pathogenicity assays are relatively easy to perform. Maize, on the other hand, is a classical genetic model as well as an important crop plant. In 2006, the USDA/National Science Foundation Microbial Genome Sequencing Project National Research Initiative funded a project to sequence the genome of *C. graminicola* strain M1.001, which was the first member of the Colletotrichum genus to be fully sequenced, providing a high-quality reference genome assembly for this species. As result, a 57.4-Mb genome was obtained, distributed among 13 chromosomes, including three minichromosomes with less than 1 Mb in size and the annotation process resulted in the identification of 12,006 gene models (O'Connell et al. 2012)[2].

[2] Broad Institute Colletotrichum Genome Database
http://www.broadinstitute.org/annotation/genome/colletotrichum_group.

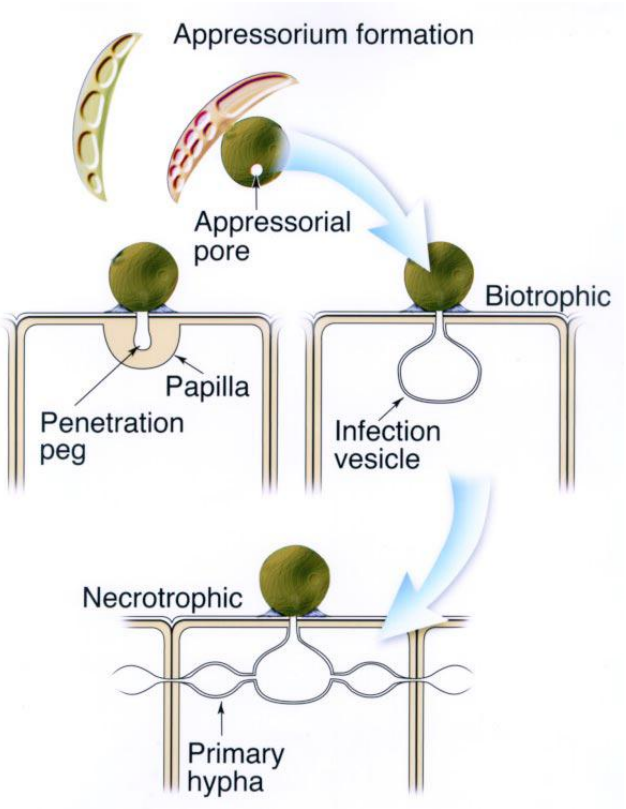

Figure 7. In hemibiotrophic species of *Colletotrichum*, spores (conidium) germinate and form dome-shaped appressoria that enter the plant cells using a combination of mechanical force and degradative enzymes. Generally, the host cell attempts to form a lignified papilla, which may prevent the entry of the penetration peg into the cell. If penetration is successful, the fungus forms an infection vesicle that invaginates the host plasma membrane. Upon entering the plant tissue, the fungus lives as a biotroph for a short time within living host cells. The fungus switches to a necrotrophic lifestyle in which the fungus kills plant cells, causing expanding necrotic lesions. Taken from Bergstrom & Nicholson (1999).

Maize is widely cultivated throughout the world and represents one of the most important commodities worldwide with an estimated harvested area of more than 176 million hectares (FAOSTAT 2012). In Spain, the largest maize producing region is Castilla y León, representing approximately 30% of the arable land (Subdirección General de Análisis, Prospectiva y Coordinación 2012). While significant crop losses due to anthracnose in Europe are not reported in the literature, the presence of the pathogen has already been described in Croatia (Palaversic et al. 2009), Germany (Böning & Wallner 1936), France (Messiaen et al. 1960 and Syngenta, *personal comunication*) and Switzerland (Sukno et al. 2013, *in press*). To date, the fungus has not been

isolated from diseased plants in Spain, but considering that maize anthracnose is found in southern France, very close to the border of Spain, we suspect that maize production, especially in Northern Spain is at risk of anthracnose. Traditionally, anthracnose has not been a serious concern for maize producers because our dry climate inhibits the spread of the disease. However, climate change brings the risk of wetter, more humid conditions during the growing season, especially during the early part of the growing season when *C. graminicola* can cause a disease known as seedling blight. Additionally, the increasing use of irrigation may also lead to conditions in which anthracnose can spread quickly. Basic research into the nature of this emerging disease and pathogen will provide us necessary tools to control it.

2. Hypotheses and Objectives

This thesis has been divided into four main chapters. Each one of them is written in scientific paper format, including introduction, materials and methods, results and discussion sections. General conclusions for all of them are presented together at the end of the thesis. As well as most scientific projects, several hypotheses were developed and tested, and most of them were evaluated at different sections of this thesis. Main hypotheses and objectives are described below.

Overall Objective

To understand evolutionary aspects of the plant-pathogen interaction using the maize-*Colletotrichum graminicola* pathosystem.

Hypothesis I

Variation in virulence in C. graminicola is correlated with changes in the genome, including rapidly evolving sequences, gene gain/loss and genomic structural variations.

Objective I.A

To characterize phenotypic variability in a sample of eight strains of *C. graminicola*.

Objective I.B

To resequence, assemble and annotate whole genome sequences of seven field isolates of *C. graminicola*.

Objective I.C

To determine whether there is evidence of recombination between isolates of *C. graminicola*.

Objective I.D

To determine the phylogenetic relationships between isolates of *C. graminicola*.

Objective I.E

To characterize genomic structural variations in the genomes of *C. graminicola* and their possible implications in pathogenicity and virulence.

Objective I.F

To determine patterns of gene gain and loss between the isolates of *C. graminicola*.

Objective I.G

To search for evidence of natural selection acting on coding and non-coding DNA sequences at the whole-genome level.

Hypothesis II

Fungi and oomycetes produce several non-coding RNAs (ncRNAs) involved in multiple functions, including pathogenicity.

Objective II

To search for and annotate putative ncRNAs in fungal and oomycete genomes using publicly available ESTs.

Hypothesis III

Plant defense related genes are under the action of positive selection as a consequence of the arms race between plant and pathogens.

Objective III

To search for evidence of positive selection in disease response genes within members of the Poaceae.

2.1. Thesis outline

Chapter I is aimed at describing phenotypic and genomic characteristics of the eight isolates of *C. graminicola* widely used along the thesis. For phenotypic characterization (**Objective I.A**) I implemented *in vitro* and *in planta* assays to quantify and classify isolates according to their virulence against maize. Genomic characterization was performed by sequencing and analyzing the whole genomes of the isolates. I initially describe the genome assembly strategies carried out and the statistics derived from the assembled genomes (**Objective I.B**). I further analyzed four aspects of the genomes of the isolates: recombination (**Objective I.C**), phylogenetic relationships (**Objective I.D**), genomic structural variations (**Objective I.E**) and the gene content (**Objective I.F**) and I attempt to correlate genomic traits with the phenotypic characteristics of the isolates.

In **Chapter II**, I analyzed genome-wide patterns of selection acting on different classes of sequences between the eight isolates of *C. graminicola* based on the mapping assemblies built in the previous chapter (**Objective I.G**). I used both frequency spectrum and maximum likelihood based methods to analyze selective pressures acting on coding and non-coding DNA sequences. One of the most striking results from this study was the detection of natural selection acting on non-coding sequences in *C. graminicola* isolates, implying that those regions are functionally important as previously described for higher eukaryotes. Two main hypotheses were formulated regarding the function of such putatively functional non-coding sequences. The first one has to do with the presence of regulatory elements driving the transcriptional process. This issue was addressed in **Chapter II**, with results suggesting that genes which are upregulated during infection tend to show different patterns of selection in their neighboring non-coding regions than other genes.

The functionality of non-coding DNA regions could also be linked to the presence of non-protein-coding RNAs (ncRNAs) implicated in some of the multiple functions already discovered for this sequences such as RNA processing, modification, transcriptional regulation, mRNA stability and translation, and protein secretion. I address this issue at **Chapter III**, in which I successfully adapted and applied a pipeline previously developed for the

identification and annotation of bovine ncRNAs using ESTs (**Objective II**). Unfortunately, few ESTs for *C. graminicola* are available at the moment, which is why I decided to use publicly available ESTs for 42 species of filamentous fungi and 4 species of oomycetes to annotate and classify putative ncRNAs at all of them. At the moment, there are no genome-wide studies of ncRNA in filamentous fungi and this first report will potentially open a poorly explored field fungal genomics.

Finally, since coevolution between plants and pathogens can leave footprints on their genomes and genes involved on the interaction are expected to show patterns of positive selection, in **Chapter IV** I used information about genes that are expressed in maize during infection with *C. graminicola* to search for evidence of positive selection acting on these defense-related genes between members of the Poaceae lineage (**Objective III**).

Overall, this thesis provides a valuable resource for plant pathologist and for the development of environmentally friendly strategies to control fungal diseases at the time that highlights the importance of bioinformatics analysis applied to the study of agricultural pathogens, contributing to our understanding of aspects regarding molecular and evolutionary process taking place during the plant-pathogen interaction.

Note: Supplementary Material is included in the digital version of this Thesis.

3. CHAPTER I

Characterization, sequencing and comparative genomics of *Colletotrichum graminicola* isolates

3.1. Introduction

Understanding the diversity of genome organization, the processes that create it, and its consequences is particularly important for understanding the genetic basis of the huge diversity of life (Budd 2012). When dealing with pathogens, these studies may additionally help us to learn more about genomic traits shaping its ability to attack other organisms. To determine the extent of genetic variation among isolates of *Colletotrichum graminicola*, I sequenced seven field isolates and jointly analyzed them with the high-quality reference genome of *C. graminicola* strain M1.001 (O'Connell et al. 2012). In the present chapter, I introduce the phenotypic characterization of these isolates, the sequencing, assembly and annotation process carried out on each one of them and the evaluation of three important aspects usually addressed in comparative and population genomics of pathogens: recombination, genomic structural variations and gene content of the isolates.

New sequencing technologies have enabled many biological research labs to sequence and assemble small-sized eukaryotic genomes using their own resources (Haridas et al. 2011). These next generation sequencing (NGS) technologies such as SOLiD™ Sequencing (www.appliedbiosystems.com) or Illumina Sequencing (www.illumina.com) can produce massive amounts of data in short time and at a low cost, but producing tens of millions of short reads (usually between 50 and 400 nt long), which poses a major challenge for genome assemblies. Depending on whether a high quality reference genome is available, current genome assembly strategies generally fall into one of two categories: mapping assemblies or *de novo* assemblies (**Figure 8**).

The development of *de novo* genome assembly algorithms is an extremely active area of research, with many assemblers published and many more being produced (Tritt et al. 2012). Among the most popular and freely available *de novo* assemblers are Abyss (Simpson et al. 2009), SOAPdenovo (Luo et al. 2012), Velvet (Zerbino & Birney 2008) and IDBA (Peng et al. 2010). The main goal of all of them is piece together individual sequence reads to form the longest contigs possible. Of course, this task is much more computationally and methodologically costly than mapping sequence reads to a reference genome, but allow us to identify new genomic regions that were not previously described

in the reference genome. On the other hand, mapping assembly is usually performed in order to identify variants present in one individual's genetic profile compared to a reference sequence. A number of programs to map short sequence reads against a reference genome have been developed. Examples of such applications are Bowtie (Langmead et al. 2009), BWA (Li & Durbin 2009), MAQ (H. Li et al. 2008), and SOAP (R. Li et al. 2008).

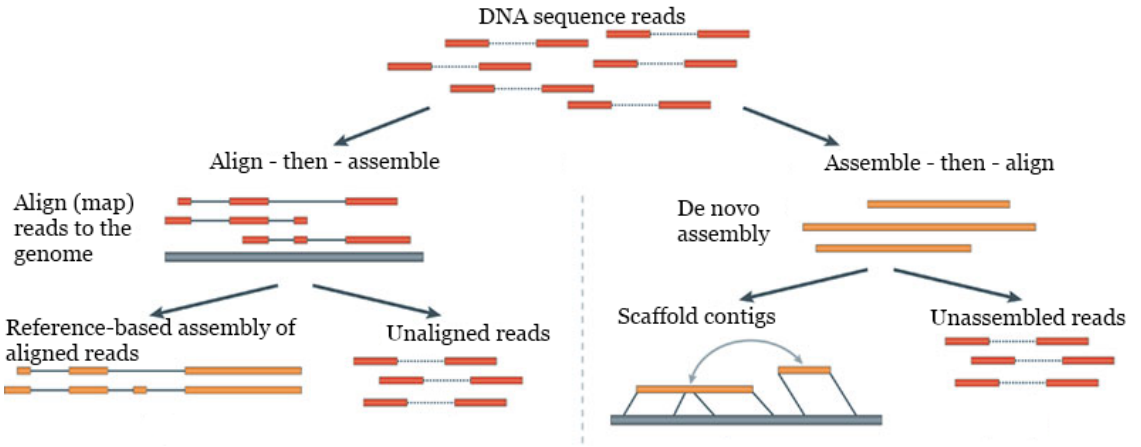

Figure 8. Alternative approaches for genome assembly. Pipeline describing the mapping assembly (left) and the *de novo* assembly (right) strategies. Modified from Martin & Wang (2011).

At the present work, I carried out both kinds of assemblies aimed to address different goals. Alignments of mapped reads to M1.001 reference genome were used to analyze structural variations (chromosomal rearrangements). Assembled consensus sequences resulting from mapping/aligning were used for the recombination and phylogenetic analyses presented at this chapter and for the selection analysis introduced at **Chapter II**. Finally, *de novo* assemblies were used to analyze the gene content at each isolate.

One of the first questions that should be addressed when study genetic diversity is whether or not sexual reproduction is shuffling existing genetic material in the population (Williams 1975). Sexual recombination contributes significantly to genetic diversity in eukaryotic genomes and increases the rate of adaptation of the population (Heitman 2006; Michod et al. 2008; de Jonge et al. 2013). This process is particularly important for pathogenic species, which usually face new hosts (like resistant cultivars or new host species), environments or agricultural practices (Saleh et al. 2012), which might even result in the creation of new infective strains (Michod et al. 2008). In addition, the study of

recombination is important to study because of the potentially disruptive influences that it can have on evolutionary analyses of sequence alignments, since evolutionary histories of recombinant sequences cannot be accurately described by standard bifurcating phylogenetic trees, but instead along a set of correlated trees (Arenas & Posada 2010) and also because patterns of genetic variability created by recombination can closely reassemble the effects of molecular adaptation (Anisimova et al. 2003).

In fungi, sexual reproduction is very difficult to detect since it could be cryptic or facultative (Agrios 2005). In addition, sexual organs may be difficult to observe when sexual reproduction is restricted to limited spatial areas or periods of time (Saleh et al. 2012). Sexual reproduction is rare in most *Colletotrichum* species (Wharton & Uribeondo 2004; O'Connell et al. 2012). The teleomorphic phase has been characterized in three species in this genus: *C. falcatum* (Caravajal & Edgerton 1944), *C. graminicola* (Politis & Wheeler 1972) and *C. sublineola* (Vaillancourt & Hanau 1992), but only *C. falcatum* has been found in nature (outside of the laboratory) (Crouch & Beirn 2009). *C. graminicola* was reported to be homothallic (Politis 1975), but many strains, including M1.001, are self-sterile and cross-fertile (Vaillancourt 1991). In an attempt to clarify whether or not sexual reproduction had taken place between isolates of *C. graminicola*, I analyzed patterns of genetic recombination at the whole-genome level between the eight isolates. Various bioinformatic strategies can be explored to address this question, depending on the objectives. In this case, I focused mainly in two goals: to detect evidence of recombination in our dataset and to identify mosaic sequences.

A simple and robust statistical test for detecting the presence of recombination is by means of the PHI, Pairwise Homoplasy Index (Φ_w) described by Bruen et al., 2006. This test (called PhiTest) is able to detect recombination within any given set of aligned sequences, regardless of population history, demographic history or recombination and mutation rates. PhiTest examines "incompatibilities" in phylogenetic signals (so called phylogenetically incompatible site-pairs), and is based on the "four gamete test" (Hudson & Kaplan 1985). If two isolates never recombine, adjacent polymorphisms will likely be "compatible", that is both polymorphic sites will have the same

phylogenetic signal (they will support the same tree topology). On the other hand, "incompatible" sites could have two possible histories, one in which there was recurrent (convergent) mutation in different isolates and one in which there was recombination between them (**Figure 9**). Under an infinite-sites model (Kimura 1969) of sequence evolution, the possibility of a recurrent mutation does not exist, and so incompatibility for a pair of sites implies that at least one recombination event must have occurred. The Pairwise Homoplasy Index (Φ_w) estimate the mean refined incompatibility score from nearby sites, which is interpreted as the minimum number of convergent or recurrent mutations necessarily present on any tree describing the history of any two sites. Significance of the observed Φw statistic is later obtained by using a permutation test.

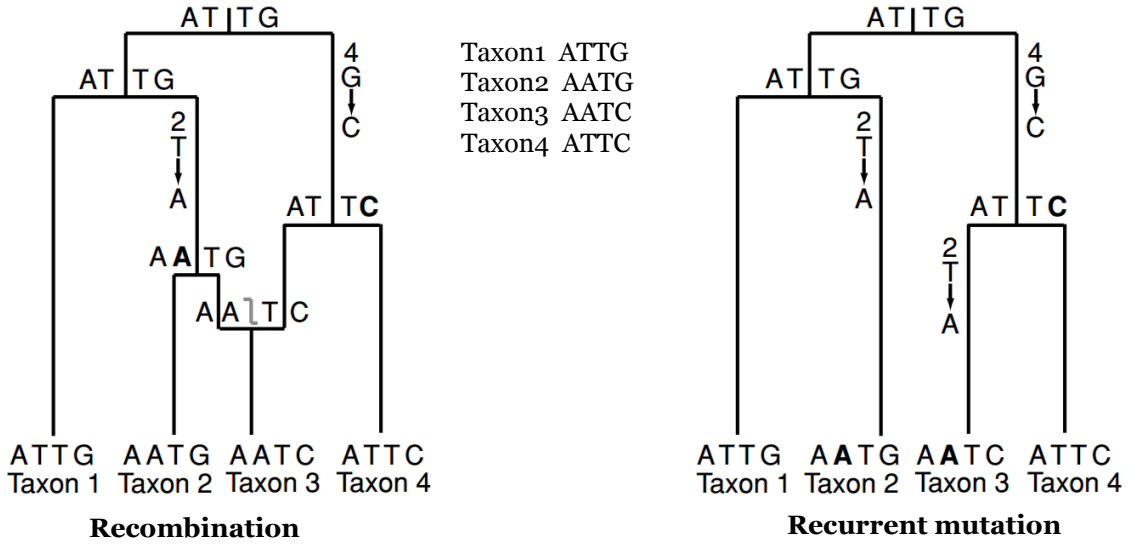

Figure 9. In this example, sites 2 and 4 are informative; however these two sites are "incompatibles" because there is no single tree that can represent the evolutionary history of these taxa. Two different histories can explain this situation: Recurrent mutation or Recombination. At the left, recombination event is represented by the joining of two branches into the lineage of taxon 3, which carries a recombinant sequence. At the right, recurrent mutation (T→A) at site 2 in taxon 2 and taxon 3. Taken from Lemey & Posada (2009).

Recombination breakpoints (locations where the crossing-over have occurred) in a given set of genomes from individuals in a population divide the genome into haplotype blocks, resulting in a mosaic structure of the genome (Zhang et al. 2009). Mosaic structures can be identified by considering three sequences at a time in the multiple sequence alignment to determine the existence of one

sequence that is a mosaic of a second and a third. In this work, I tested for the existence of such mosaics by applying the program 3SEQ (Boni et al. 2007) to different aligned regions in the genome, as well as to each gene and its neighboring regions.

While homologous recombination is an important mechanism for genome rearrangements (Bishop & Schiestl 2000), in fungi it is not strictly needed for the generation of new structural genome variations (Kistler & Miao 1992). Even though specific mechanisms were not always recognized, widespread chromosomal rearrangements have been identified at sexual and asexual filamentous fungi and many of them seem to be involved in the generation of new variants to mediate virulence (Hane et al. 2011; de Jonge et al. 2013). A second question addressed in this chapter is regarding whether or not structural variations and chromosomal rearrangements are playing some role in the evolution of *C. graminicola*.

Genomic structural variations are considered to be any DNA sequence alteration other than a single nucleotide polymorphism (SNP) (Feuk et al. 2006). These include insertions, deletions, inversions, inter and intra chromosomal translocations (**Figure 10**). The genome-wide detection of structural variants has recently improved enormously due to the information provided by next generation paired-end sequencing reads. Since paired reads are generated at an approximately known distance in the donor genome, paired reads that map to the reference genome with a distance between them substantially different from the expected length or with anomalous orientation, provide evidence for the occurrence of a structural variants between the genomes (Medvedev et al. 2009). I evaluated the existence of structural variations by using the derived alignment of mapped paired reads from each isolate against the reference genome of *C. graminicola* strain M1.001.

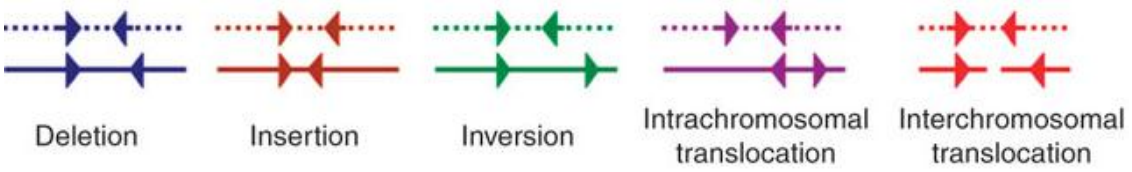

Figure 10. Different classes of structural variations. Arrows pairs represent the location and the orientation of the structural variation. Dotted line represents a chromosome in the analyzed genome. Solid line represents a chromosome in the reference genome. **Deletion**: A piece of a chromosome is missing in the analyzed genome. **Insertion**: A segment of DNA is added in the analyzed genome. **Inversion**: A segment of DNA is reversed orientated regarding the same place in the reference chromosome. **Translocation**: A change in position of a chromosomal segment within a genome that involves no change to the total DNA content. Translocations can be intra or inter chromosomal. Modified from Chen et al. (2009).

Patterns of gene gain and loss have been demonstrated to be an important aspect during evolution of pathogenic fungi (Dean et al. 2005; Powell et al. 2008; Soanes et al. 2008; Spanu et al. 2010). In addition, changes in the gene pool between different strains of the same fungal species may help to identify those genes being part of the "core gene set" and also those genes that are unique for each isolate. This analysis may result even more significant when isolates in the sample present differences in virulence (the degree of aggressiveness) and pathogenicity (the capacity to cause a disease), as in our case. In order to analyze the gene content at each isolate, I used the *de novo* assembled genomes to annotate most genes as possible using *ab initio* and evidence-based methods. Finally, I attempted to correlate phenotypic variations in terms of virulence with the genomic characteristics of each isolate.

3.2. Materials and Methods

3.2.1. Isolates phenotypic characterization[3] and sequencing

Infection assays

C. graminicola isolates obtained from culture collections (**Table 3**) were maintained at 23°C on PDA medium with continuous illumination under white fluorescent light. Cultures used for maize infection assays were grown for 15 to 20 days on PDA as described previously (Sukno et al. 2008; Vargas et al. 2012). Spores were harvested from the culture surface of the Petri dishes as described previously (Thon et al. 2000), filtered through three layers of cheesecloth and then washed with two rounds of re-suspension in sterile distilled water (dH_2O) followed by centrifugation (3,000 rpm, 3 min, room temperature) (Sukno et al. 2008). Conidia from each *C. graminicola* fungal isolate were inoculated on the highly susceptible maize inbred line M0940 (Warren et al. 1975) by placing 10mL droplets containing approximately 650 conidia on the adaxial side of the third leaf of the plant. The position of each infection site was marked for future reference. Infected plants, growing in Cone-Tainers (Stuewe & Sons, Corvallis, OR), were then sealed with plastic to preserve humidity and were incubated for 18 hours at 23°C. After this period, the plants were left undisturbed for several hours until the drops were dry and were then transferred to a growth chamber for four days at 25°C and 50% humidity. The assay was repeated three times, with 2-3 plants inoculated with each fungal strain. Average lesion size for each strain was measured by scanning leaves and using an image processing software (Paint.NET v3.5: www.getpaint.net) to quantify the lesion size. Differences among average lesion sizes were statistically tested using a Tukey HSD pairwise comparisons test.

[3] Obtaining isolates and maintaining them, as well as the infection and *in vitro* assays, were performed by José M. Sanz-Martín, Walter A. Vargas and Serenella A. Sukno, members of Laboratory 1 of CIALE.

In vitro assays

Growth rate assay: *C. graminicola* isolates were grown on three Petri dishes (90mm Ø) using three different solid media and maintained at 23°C with continuous illumination under white fluorescent light. Different culture media were used to measure growth rate: PDA medium (Potato Dextrose Agar, Difco Laboratories), representing the complete medium, covering all nutritional requirements, Minimum Medium (MM), providing just basal essential nutrients (Politis 1975; Thon et al. 2002) and Oatmeal Agar medium (OA) to encourage sporulation. Previously to the assay, each isolate was grown for 5 days on PDA to obtain mycelia as source of inoculum. We extracted a piece of actively growing vegetative mycelia using a sterile cork borer (0.5cm Ø, Carl Roth, Karlsruhe, Germany; cat number Catalog number 0581.1. http://www.carl-roth.de/) and placed it at the center of each Petri plate. The diameter of the colony was measured at 2, 4 and 6 days. The experiment was repeated two times.

Spores per plate assay: Each isolate was grown on PDA medium for 20 days at 23°C with continuous illumination under white fluorescent light. Conidia were picked up from plates, filtered through sterile gauze and washed three times with distilled water. A hematocytometer was used to count spores in order to determine the total number per milliliter which was later extrapolated to the total plate volume. For analyze data we transformed the total number of spores per plate to a logarithmic (Log_{10}) scale in order to avoid the influence of big differences between the samples.

In vitro percentage of germination assay: 100uL drops of the spore suspensions (104 spores/mL) were deposited in the bottom of empty Petri dishes and were transferred into closed humidity chambers and placed in an incubator (23 °C, continuous light). The percentage of germination was assessed at 24 hours by counting the number of germinated spores within a single field observed in the center of the drop at 10x magnification. The experiment was performed twice, analyzing three samples from each isolate in each experiment.

Culture conditions and genomic DNA (gDNA) purification

We incubated mycelia from monosporic cultures with orbital shaking in Fries' medium (Vaillancourt & Hanau 1992) for four days, at 25°C and 150 rpm with continuous illumination and then we used the Maxi DNA extraction protocol adapted from Baek & Kenerley (1998) to extract genomic DNA (gDNA).

Sequencing the Internal Transcribed Spacer (ITS) 1

Prior to genome sequencing, we verified the identity of the isolates by sequencing the ribosomal RNA gene Internal Transcribed Spacer 1 (ITS1). Primers ITS5 (GGAAGTAAAAGTCGTAACAAGG) and ITS4 (TCCTCCGCTTATTGATATGC) were used to amplify ITS region 1 from all isolates. gDNA was used as template for PCR with the following conditions: 94°C (2'), 35 cycles of 94°C (30"), 55°C (30") and 72°C (45") and a final extension step of 72°C (5'). We performed all PCR reactions with Biotools DNA polymerase. PCR amplifications were visualized by agarose (0.8%) gel electrophoresis at 90V, confirming the amplicon size. The amplified ITS fragments were purified using NucleoSpin Gel and PCR Clean up (Macherey Nagel) kits and sequenced at the Sequencing Service of the University of Salamanca.

Genome Sequencing and Illumina sequence reads processing

Genomic DNA (1µg/100µL) was sent to the Keck Center for Comparative and Functional Genomics (University of Illinois) where shotgun DNAseq libraries were prepared with Illumina's TruSeq DNAseq Sample Prep kit. Libraries were then pooled, quantified by qPCR and sequenced on one lane for 100 cycles from each end of the fragments on an Illumina HiSeq2000 system, producing over 400 million 100 bp reads with insert sizes between 400 and 500 bp. I assessed the quality of the reads using the FastQC software (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Most isolates passed all quality controls. Two isolates (i51134 and iJAB2) failed in the "per base quality score", so these libraries were re-sequenced, but FastQC still failed for these isolates after re-sequencing. Warn/Fail results in FastQC are common for very small genomes according to the manufacturer. In fact, the total number of

sequenced reads for these two isolates was much lower than for the others. I presume that FastQC results as well as the lower number of sequenced reads could be due to problems with gDNA quality for isolates i51134 and iJAB2.

3.2.2. Genome assembly

Mapping assembly

For this strategy, *C. graminicola* genomes were assembled by mapping reads to the reference genome of *C. graminicola* strain M1.001 (BioProject: PRJNA37879) (O'Connell et al. 2012) and calling the consensus sequence using MAQ v0.7.1 (H. Li et al. 2008). Due to reduced quality of sequence reads from isolates i51134 and iJAB2 I was extremely strict for mapping reads to the reference genome, to call the consensus sequence and to identify SNPs, which was reflected at both the lower read depth and the greater percentage of ambiguously called based for isolates i51134 and iJAB2. For mapping reads to the reference genome I allowed a maximum of two mismatches per read and all nucleotides in the consensus sequence were required to present a minimum mapping quality of 40, a minimum neighboring quality of 20 and read depth of 3X. All regions where these requirements were not met were masked with "Ns" at all genomes and were not used for the SNPs calling. Due to the highly strict filtering of reads, many of them were discarded from the analysis (between 7% of isolate i63127 and 60% for isolate iJAB2) (**Table 5**). Pairwise identities between consensus sequences from each isolate were calculated using EMBOSS: infoalign (Rice et al. 2000).

De novo assembly[4]

Using FastQ sequence reads from isolate i13649, I first tested accuracy and speed for four different assemblers: Abyss (Simpson et al. 2009), SOAPdenovo (Luo et al. 2012), Velvet v.1.1 (Zerbino & Birney 2008) and IDBA (Peng et al. 2010). Accuracy was measured in terms of assembly size, number of contigs >

[4] De novo assemblies were performed with the help of Vinicio Armijos, from Lab. 1 from CIALE.

200nt and the N50/N90[5] values. I decided to use Velvet since it showed the best performance of the four tested assemblers (data not shown). For each isolate, I evaluated a broad range of k-mer[6] values (from 21 to 95) until I obtained the k-mer value providing the greatest N50 value for the assembly.

3.2.3. Recombination analyses

Exploratory recombination analysis using the Phi statistic

Using chromosomal consensus sequences previously obtained by the *Mapping Assembly*, I analyzed all 13 entire chromosomes by mean of the PhiPack software (Bruen et al. 2006). A p-value for each chromosome under the null hypothesis of no recombination was obtained. Incompatibilities between all pairs of parsimony informative sites were represented by an incompatibility matrix for each chromosome. To determine which regions exhibit the strongest evidence of recombination, I calculated the Phi statistic along a sliding-window using the Profile package (Bruen et al. 2006). I estimated Phi values at 100nt windows for each 1,000nt genomic region with jumps of 500nt. P-values of recombination within each genomic region were corrected for multiple comparisons using False Discovery Rate (FDR) controlling procedure by the method of Benjamini & Hochberg (1995).

Analysis of mosaic structures using 3SEQ

Mosaic structure analysis was performed in two different ways using the 3SEQ (-xrun) v.1.1 software (Boni et al. 2007). First, I computed a multiple genome alignment of the *de novo* assembled genomes and the *C. graminicola* M1.001 contigs using progressiveMauve (Darling et al. 2004; Darling et al. 2010). I then extracted all Locally Collinear Blocks (LCBs)[7] containing sequences for all the

[5] N50/N90: Represent the size of the smallest contig such that 50%/90% of the genome is contained in contigs of size N50/N90 or greater. The larger N50/N90 statistics, the better quality of the assemblies.
[6] K-mer: In genome assembly, this refers to the number of perfectly matching adjacent nucleotides among reads that are needed to make contigs.
[7] Locally Collinear Blocks (LCBs) are conserved genomic regions initially identified by Mauve to perform the multiple genome alignment. Once a set of alignment anchors that define each LCB

eight isolates and 3SEQ was run over each LCB (duplicated -identical- sequences were removed). 3SEQ reports a p-value for each LCB corrected for multiple comparisons by the method of Dunn-Šidák. In a second experiment, mosaic structures of each individual gene in the genome including 500nt of flanking sequence were identified. In order to this, I extracted sequences between positions 500nt upstream and 500nt downstream of the coding sequences (CDS±500nt) of each gene from the genomic consensus sequences of each isolated created by the *Mapping Assembly*. Each gene locus was extracted based on coordinates described for the *C. graminicola* strain M1.001 annotation (O'Connell et al. 2012; Broad Fungal Genome Initiative).

3.2.4. Phylogenetic Tree

I used the genomic coordinates of the 12,006 gene models predicted by the Broad Institute in *C. graminicola* M1.001 (O'Connell et al. 2012) to extract transcript sequences from each consensus genome constructed by the mapping assembly. I excluded from the analysis any gene showing evidence of recombination according to 3SEQ results. I then selected orthologous clusters containing transcript sequences with less than 20% of unambiguously called bases (Ns) for all the 8 isolates and I clustered them together to create multiple sequence alignments. In addition, in order to be used as an outgroup in the phylogenetic tree, I identified orthologous transcript sequences between *C. graminicola* and the closely related species *C. higginsianum* by computing reciprocal best blast hits. Clusters of sequences containing one transcript per isolate and the orthologous transcript sequence from *C. higginsianum* were aligned using MUSCLE v3.8 (Edgar 2004). A total of 8,288 multiple sequence alignments were then concatenated to create a multigene multiple sequence alignment, which was used to build a Maximum Likelihood tree using PhyML v.3.0 (Guindon & Gascuel 2003; Guindon et al. 2010).

---

have been determined, Mauve performs a gapped global alignment of each LCB. Mauve applies the ClustalW progressive global alignment algorithm to each LCB.

3.2.5. Structural variants analyses

Structural variants including insertions, deletions, inversions, and translocations were identified using BreakDancer-Max v.1.3 (Chen et al. 2009) over each alignment derived from mapping paired reads to the reference genome of *C. graminicola* strain M1.001 (see *Mapping Assembly*). Output *MAP* alignments from MAQ were converted to *BAM* using SAMtools (H. Li et al. 2009). All variants were required to have a minimum PhredQ score of 60 and at least 10 read pairs supporting the variant. In addition, I filtered out inter-chromosomal translocations having an anchoring region shorter than 100nt, which highly improved the specificity at cost of sensitivity in the results. Such strict filtering did not allow us to analyze isolates iJAB2 nor i51134, which had in most of cases, a low number of reads supporting the structural variation.

3.2.6. Gene prediction and annotation

Gene annotations for each *de novo* assembled genome were obtained using the annotation pipeline MAKER v.2.28 (Cantarel et al. 2008; Holt & Yandell 2011). MAKER leverages existing software tools and integrates their output to produce the best possible gene model for a given location. I used three *Ab initio* gene prediction models: Augustus v.2.5 (Stanke & Waack 2003; Stanke et al. 2008), GeneMark-ES (Lomsadze et al. 2005) and SNAP (Korf 2004). Augustus was trained using *C. graminicola* ESTs previously downloaded from GenBank. Proteins, ESTs and transcripts sequences from *C. graminicola* were submitted to MAKER as evidence. Illumina RNA-Seqs reads sequenced by O'Connell et al. (2012) were assembled using Cufflinks v.2.1 (Trapnell et al. 2010) and then mapped to each genome using TopHat v.2 (Kim et al. 2013) to provide MAKER such alignments as evidence. Transcript sequences from *C. graminicola* were mapped to each genome using GMAP v. 2012-06-25 (Wu & Watanabe 2005) and also submitted as evidence to MAKER. I additionally used a specific *C. graminicola* repeats library previously created (O'Connell et al. 2012) as evidence for RepeatMasker (Smit et al. 1996) to identify and mask out repeat elements at each genome.

3.2.7. Gene content analysis

In order to determinate unique genes for each isolate, I identify homologous sequences between the isolates based on both global and local alignments. I initially identified gene models predicted at genomic regions showing high similarity with M1.001 transcripts. This task was performed by analyzing those gene models at each genome that presented evidence provided by GMAP (Wu & Watanabe 2005), which was used to map M1.001 transcripts to each genome by mean of performing global Needleman & Wunsch (1970) alignments. These genes were assumed to be homologous sequences of the gene mapping at such region. Remaining gene models (those predicted at regions where none M1.001 transcript was mapped) were then clustered based on local alignment similarities. In order to this, I performed an all vs. all BLASTP between remaining proteins plus all M1.001 proteins. I clustered proteins using the Markov clustering program MCL (Enright et al. 2002) based on BLASTP results. Unclustered genes were considered to be unique for each genome. The core genome was identified by merging results from both methods and looking for genes with evidence of being present in all isolates according to any of the strategies. The functional categories analysis was performed as describe at the Material and Methods of **Chapter II**.

3.3. Results

3.3.1. Phenotypic characterization of the C. graminicola isolates

Twelve strains of *Colletotrichum graminicola* obtained from different sources (mainly from culture collections), were initially evaluated for their virulence on maize. Seven strains were selected to be sequenced and analyzed in more detail, representing the broadest diversity in geographic origin and virulence possible. These seven isolates including the recently sequenced reference strain M1.001 (O'Connell et al. 2012) represent a worldwide sample of this pathogen with a variable range of virulence against maize (**Table 3** and **Figure 11**).

Table 3. Information about *C. graminicola* isolates used in the present study.

Isolate	Other name/code	Origen	Source
M1.001	CgM2 CBS-130836	Missouri, USA	Dr. Lisa Vaillancourt, University of Kentucky
i318	LARS 318	Nigeria	Warwick (HRI) Genetic Resources Unit, UK
i113173	CBS113173 IMI 84302	Zimbabwe	Common Access to Biological Resources and Information (CABRI)
i47511	NRRL47511	Michigan, USA	USDA-ARS Culture Collection (NRRL), USA
iJAB2		Brazil	Warwick (HRI) Genetics Resources Unit, UK
i13649	NRRL13649 ATCC 34167	Alabama, USA	USDA-ARS Culture Collection (NRRL), USA
i63127	DMS-63127	Germany	Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ)
i51134	MAFF51134	Nagano, Japan	National Institute of Agrobiological Sciences, Japan

In order to characterize the isolates, I evaluated four phenotypic characters: virulence measured as lesion size on maize leaves, growth rate *in vitro*, spores production measured as spores per plate and percentage of spore germination (see Materials and Methods). All four experiments showed statistical differences between the isolates (**Table 4** and **Figure 12**). The infection assay, one of the most representative experiments used to characterize the virulence of the isolates, divided isolates into four overlapping groups (ANOVA, $p=2.8e^{-70}$): A: M1.001, i318, i113173; B: i318, i113173, i47511; C: iJAB2, i13649 and D: i63127, i51134 (**Table 4**). The most virulent isolate was M1.001, while isolates i63127 and i51134 did not show symptoms at all four days after infection on maize (**Figure 11**). It is important to highlight that isolate i63127 did not produce appresoria, specialized infection structures that are required to penetrate leaf

surfaces, under any of the conditions tested (W.A. Vargas and S.A. Sukno, *unpublished results*). This isolate did not show symptoms even 15 days after infection. Infection assays of wounded leaves using a GFP (green fluorescent protein) labeled strain of i63127, demonstrated that this isolate is able to grow inside wounded leaves, but has a significantly lower growth rate than isolate M1.001 and also has a much lower success rate of achieving a successful infection (W.A. Vargas, *personal communication*). In contrast, isolate i51134, which produces melanized appressoria, started showing symptoms around 6-7 days after infection (data not shown). However, at these time points, borders for lesions caused by the most virulent isolates (such as M1.001 or i318) are extremely difficult to discern, reason by why I decided to use four days to measure it. This time point additionally gives us the best quantitative differences between isolates (S.A. Sukno and M.R. Thon, *unpublished results*). Further studies using wound-inoculated maize leaves showed that isolate i51134 is almost as virulent on wounded plants as isolate M1.001 (W.A. Vargas, *personal communication*).

Growth rate assays showed statistical differences between the isolates on all three tested media. For PDA (complete) medium (ANOVA, $p=5.7e^{-18}$), isolate i47511 showed the higher growth rate with an average of 11.9mm per day, while isolate i63127 showed the lower rate (4.6mm per day). Moreover, on Minimal Medium (MM) (ANOVA, $p=5.584e^{-14}$) isolates M1.001 and i318 showed the highest rate of growing, while i63127 was again the most delayed. Statistical differences were lower, but still significant between isolates at the Oatmeal Agar (OA) medium (ANOVA, $p=6.562e^{-06}$), and once again isolate i63127 showed the lower growth rate.

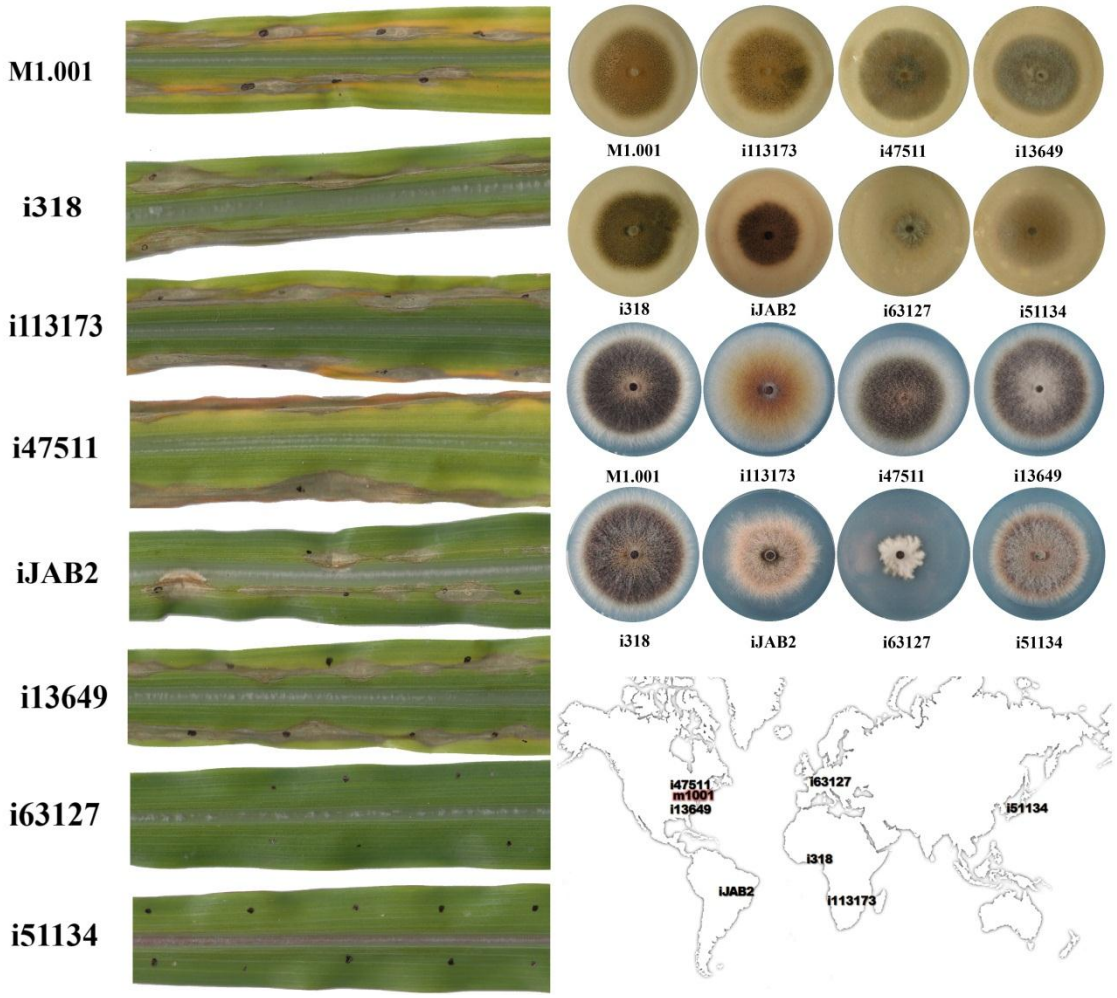

Figure 11. *C. graminicola* isolates used in this study. At the left, the symptoms showed by maize leaf six days after inoculation with each isolate. Black dot indicate inoculation points. At the right-top, isolates grown during four days on Oatmeal Agar (top) and PDA (bottom) cultures. At the right-bottom, the geographical distribution of isolates analyzed in this work.

Table 4. Phenotypic characteristics of each isolate.

Isolate	Lesion Size (mm² ± std. err.)	Growth rate (mm/day)			% Germination	Spore/Plate (LOG_{10})
		PDA	MM	OA		
M1.001	6.27 (± 0.39)$_{A}$	9.8$_{AB}$	10.5$_{A}$	9.8$_{A}$	83%$_{A}$	8.5$_{AB}$
i318	5.48 (± 0.37)$_{AB}$	11.1$_{AB}$	10.3$_{A}$	9.1$_{A}$	68%$_{B}$	8.4$_{AB}$
i113173	5.14 (± 0.31)$_{AB}$	9.9$_{AB}$	7.6$_{B}$	9.8$_{A}$	13%$_{D}$	7.5$_{B}$
i47511	5.01 (± 0.43)$_{B}$	11.9$_{A}$	8.5$_{AB}$	8.9$_{A}$	80%$_{A}$	8.7$_{A}$
iJAB2	3.26 (± 0.33)$_{C}$	7.2$_{C}$	6.4$_{B}$	8.4$_{AB}$	0%$_{E}$	8.2$_{AB}$
i13649	3.12 (± 0.32)$_{C}$	9.3$_{B}$	6.8$_{B}$	8.5$_{AB}$	7%$_{DE}$	6.9$_{C}$
i63127	0.00 (± 0.00)$_{D}$	4.6$_{D}$	3.6$_{C}$	6.9$_{B}$	0%$_{E}$	6.9$_{C}$
i51134	0.00 (± 0.00)$_{D}$	10.2$_{AB}$	9.4$_{AB}$	8.4$_{AB}$	37%$_{C}$	7.4$_{BC}$

Note: Lesion Size: Average lesion size (mm²) four days afeter infection. Growth rate: Average geowth rate (mm/day) measured at 2, 4 and 6 days on different media (PDA: Potato Dextrose Agar, MM: Minimum Medium and OA: Oatmeal Agar). % Germination: Average percentage of germination. Spore/Plate (LOG10): Average of total number of spores per plate in logarithmic scale. Letters indicate homogeneous groups according to Tukey HSD pairwise comparisons test for each experiment.

The *in vitro* percentage of spore germination was extremely variable between isolates (ANOVA, $p=3.387e^{-17}$), showing five statistically significant homogeneous groups with almost no overlapping (**Table 4**). Isolates M1.001 and i47511 had the higher percentage of germination, while isolates i63127 and iJAB2 showed none germinated spores at any of the assays. The total number of spores per plate also varied significantly between isolates (ANOVA, $p=1.216e^{-05}$). Isolate i47511 showed the largest number (approximate $7.55e^{8}$ spores per plate), while isolate i63127 presented the smaller (approximate $8.43e^{6}$ spores per plate) (**Table 4**).

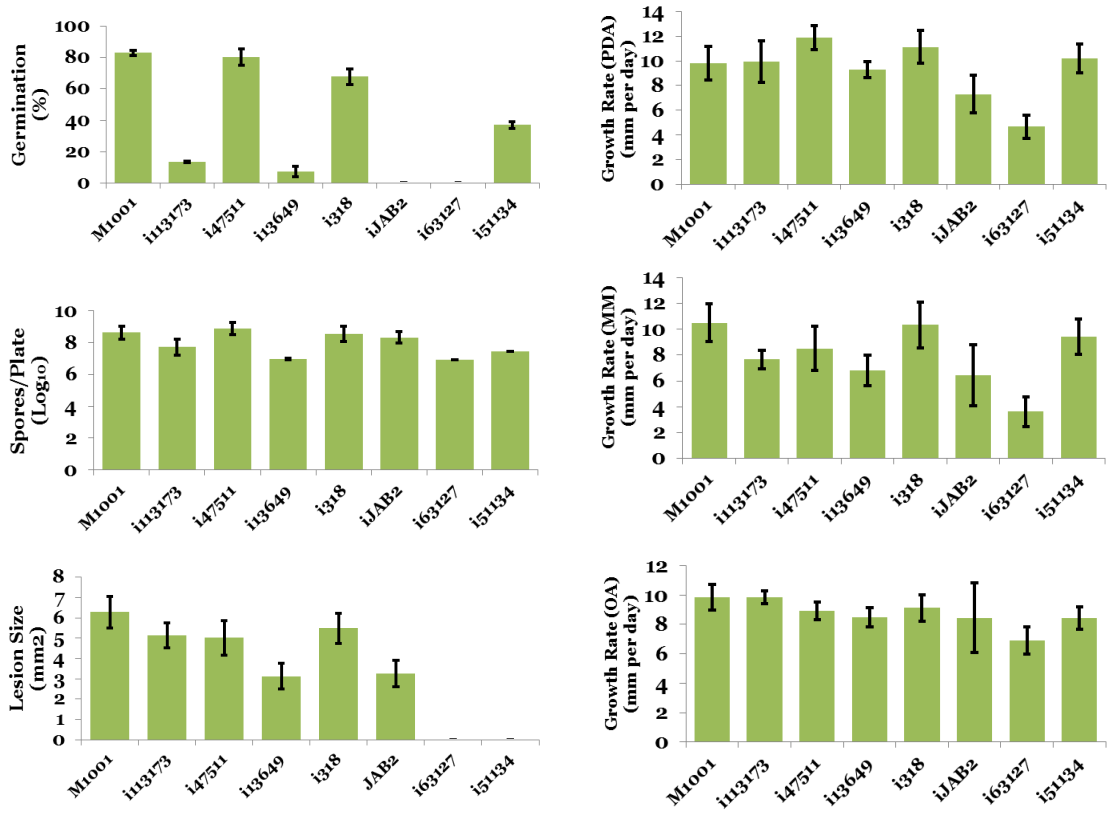

Figure 12. *In vitro* and *in planta* assays carried out in order to characterize phenotypic variability between the *C. graminicola* isolates. Bars indicate average values and standard deviation for each isolate in each experiment.

In order to determine whether or not any *in vitro* phenotypic characteristic were correlated with the average lesion size caused by the isolate, I estimated the Pearson's correlation coefficient (r) between the average lesion size and each of the average values obtained at the *in vitro* experiments (**Figure 13**). I found a positive correlation at all comparisons, indicating that such phenotypic features are probably associated with ability of the pathogen to cause disease. The

greater correlation with the lesion size was observed with the growth rate at Oatmeal Agar medium (r=0.84). However, since isolates i63127 and i51134 presented an average lesion size of cero, also estimated the Pearson's correlation without taking into account this values. This new dataset revealed a high correlation between the lesion size and both the growth rate in Minimum Medium (MM) and the percentage of germination.

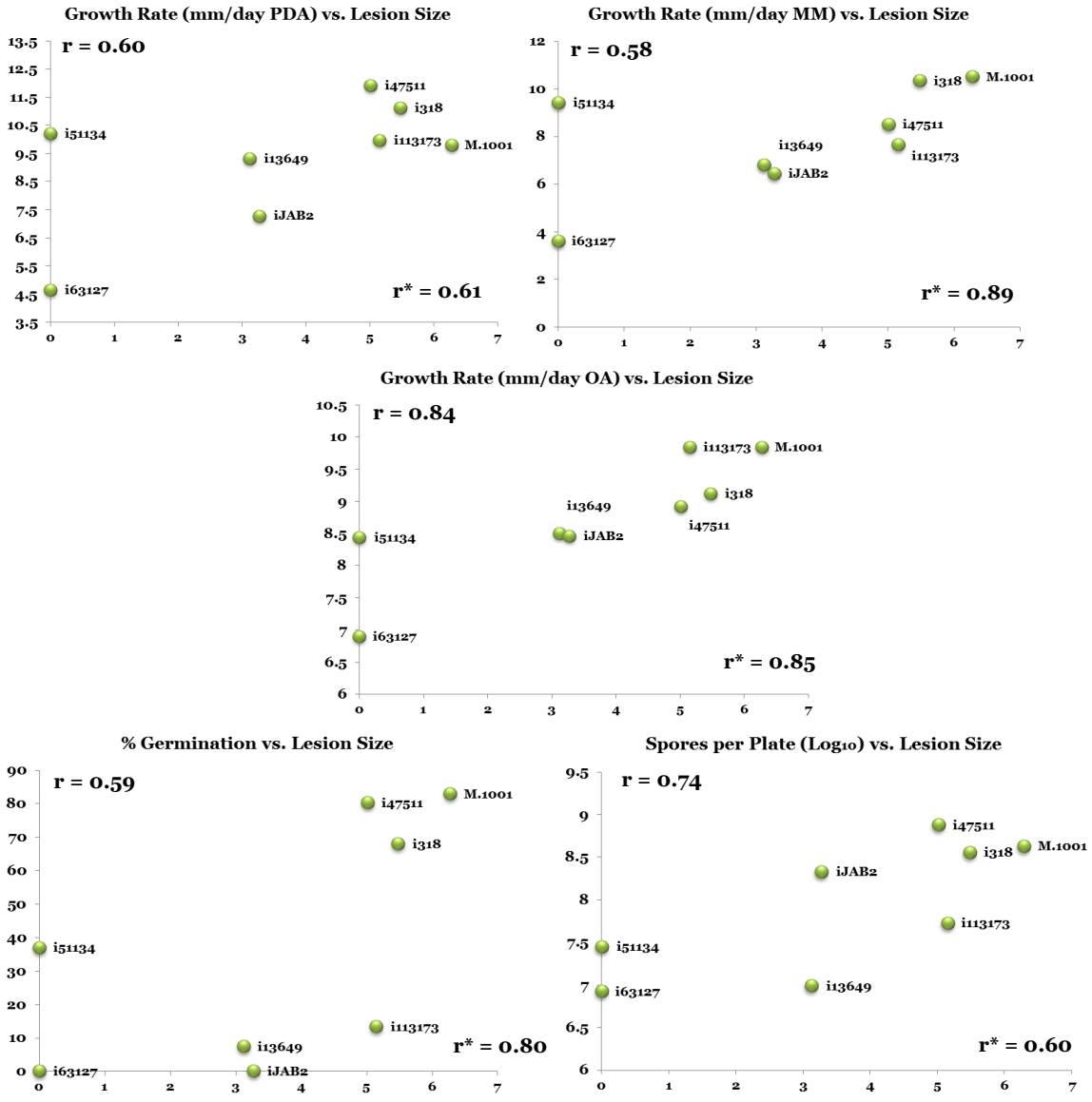

Figure 13. Correlation between phenotypic characteristics and average lesion size (mm^2). (r) Pearson's correlation coefficient using all isolates. (r*) Pearson's correlation coefficient removing isolates i63127 and i51134.

3.3.2. Genome assembly

Mapping Assembly

Mapping and assembling genomic reads from each isolate to the *C. graminicola* M1.001 reference genome resulted in high quality draft genome sequences for each isolate (**Table 5**), with average read depth ranging from 24X to 132X, and covering between 84% and 99% of the reference genome with at least 3X coverage. The average sequence identity between isolates was 96%. The most dissimilar pair is iJAB2-i63127 with an overall similarity of 93% and most similar pair is M1.001-i318 with an overall similarity of 99%. The average occurrence of SNPs per isolate was 2,126 SNPs/Mb, similar to the SNP density previously described between isolates of the wild phytopathogenic filamentous fungus *Aspergillus flavus* (1,854 SNPs/Mb), and much larger than that found between isolates of the atoxigenic, domesticated *Aspergillus oryzae* (1,042 SNPs/Mb) (Gibbons et al. 2012), but less than half of that discovered between isolates of the dimorphic human pathogen *Coccidioides immitis* (5,242 SNPs/Mb) (Neafsey et al. 2010). A more detailed analysis of SNPs and nucleotide diversity based on this assembly is presented in **Chapter II**.

Table 5. Summary statistics of mapping assemblies of the *C. graminicola* isolates.

Isolate	Sequenced Reads	Mapped Reads	%Used Reads	Read Depth	%Ns	SNPs
M1.001	Reference	-	-	-	9.21	-
i318	70,427,250	60,957,326	86.6	121X	9.83	9,170
i113173	83,773,114	58,434,572	69.8	120X	14.43	160,983
i47511	72,005,328	52,486,812	72.9	108X	13.9	141,118
iJAB2	27,686,062	11,251,096	40.6	24X	25.24	155,561
i13649	51,358,762	46,081,744	89.7	93X	14.72	82,206
i63127	67,186,960	62,416,798	92.9	132X	19.79	115,695
i51134	34,414,680	14,884,038	43.2	31X	19.53	139,134

Note: Sequenced Reads: Total number of raw sequence reads received from the sequencer. Mapped reads: Total number of effectively mapped reads from each isolates to the M1.001 genome. %Used Reads: Percentage of the number of sequenced reads effectively used for the assembly. Read depth: The average per-base depth for each genome, taking into account just unambiguous sites. SNPs: Number of SNPs identified against the M1.001 reference genome. %Ns: For M1.001, the value represents the percentage of ambiguously called bases in the reference genome (non A, T, C or G). For the sequenced isolates, the values represent the percentage of the genome with less than 3 reads coverage and therefore where SPNs were not called.

***De novo* assembly**

I initially analyzed several k-mer values (ranging from 21 to 95) for each isolate, looking for the k-mer value that resulted in the highest N50 for the assembly (**Figure 14**).

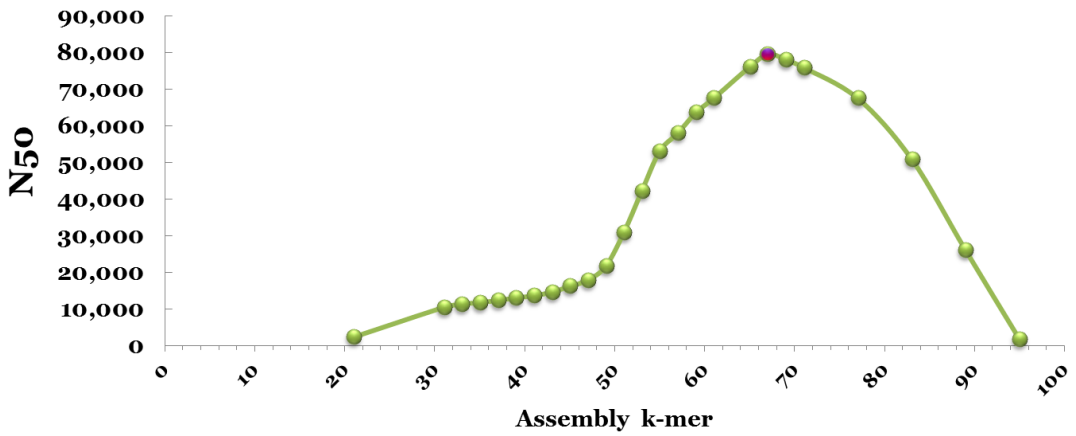

Figure 14. Genome assembly N50 values for each k-mer analyzed for the genome of i13649. In this case, k-mer = 67 had the higher N50 value (79,638) among all of the assemblies.

With the exception of genomes from isolates iJAB2 and i51134, the *de novo* assembled genomes resulted in high quality assemblies, with more than 30X coverage and comprised of relatively few large contigs (**Table 6**).

Table 6. *De novo* assembly statistics for the genomes of the seven sequenced isolates of *C. graminicola*.

Isolate	Best K-mer	% Used Reads	Assembly Size (Mb)	Number Contigs*	N50	N90	Coverage (Fold)
i318	29	64.4	43.3	4,412	60,655	10,896	65.0
i113173	29	60.3	43.9	4,857	59,544	9,424	68.6
i47511	27	61.1	43.2	4,472	55,697	9,164	63.1
iJAB2	35	24.8	38.3	65,795	757	279	3.0
i13649	67	84.3	41.5	1,727	79,638	19,217	31.2
i63127	67	90.9	42.7	1,074	99,065	24,237	46.7
i51134	35	51.8	39.1	66,883	784	271	2.1

Note: *Just contigs > 200nt were taken into account.

The sequence assemblies from isolates iJAB2 and i51134 had a much larger number of contigs, most of them short with a very low percentage of reads used for the assembly, which resulted in considerably lower coverage compared to the other genomes (**Table 6**). Because of the low quality of the genome

assemblies obtained for isolates iJAB2 and i51134, I searched for evidence of bacterial contamination in those genomes. I first identified all contigs containing at least one gene of *C. graminicola* or with high similarity to any gene from any Colletotrichum species. The remaining contigs were analyzed using BLASTn against the "nr" database from NCBI. BLAST hits results were then parsed in order to identify contigs showing a high percentage of hits with sequences from bacterial genomes. I identified 18 contigs from i51134 having >90% hits with bacterial sequences. Such a small number of contigs showing putative contamination does not seem to completely explain the results obtained for the assembly of genomes iJAB2 and i51134. Anyway, I discarded these 18 contigs for further experiments. On the other hand, the total number of sequenced reads for these two isolates was much lower than for the others (**Table 5**), which make us suspect that the lower quality assemblies for isolates i51134 and iJAB2 was mainly consequence of a poor genomic DNA quality. Without taking into account these two isolates, most genomes showed a similar genome size, ranging from 41.5 Mb to 43.9 Mb, a smaller size than the M1.001 reference genome (50.9 Mb). However, since the presence of interspersed repeats larger than sequencing reads can confound assembly algorithms (Miller et al. 2010) and repetitive DNA comprises around 12% of the *C. graminicola* genome (O'Connell et al. 2012) it is very likely that the smaller genome sizes obtained is indeed linked to the presence of repetitive DNA sequences. In addition, contigs smaller than 200 nt were not included in this analysis, therefore it remains possible that the complete genomes of the isolates presented in this work vary in abundance of repetitive sequences and actually have greater sizes.

3.3.3. Recombination analyses

Exploratory recombination analysis

I applied the Phi test (Bruen et al. 2006) to each individual chromosome using the consensus genomic sequence for each isolate previously obtained by the *Mapping Assembly*. A Phi score with a p-value<0.01 shows that recombination occurs in the data set. I found evidence of recombination within the entire alignment for all chromosomes with the exception of chromosome 12, which also showed highest diversity (**Table 7**).

Table 7. Statistics derived from the Phi test of recombination applied to each entire chromosome.

Chromosome	Length	Diversity	Informative Sites	p-value*
Chr1	6,787,984	0.2%	20,312	0.0 (< 10^{-99})
Chr2	6,748,533	0.2%	21,187	0.0 (< 10^{-99})
Chr3	6,027,927	0.2%	15,690	0.0 (< 10^{-99})
Chr4	5,882,731	0.2%	13,533	0.0 (< 10^{-99})
Chr5	5,174,327	0.2%	13,875	0.0 (< 10^{-99})
Chr6	4,553,816	0.2%	12,358	0.0 (< 10^{-99})
Chr7	4,372,745	0.2%	10,922	0.0 (< 10^{-99})
Chr8	2,496,009	0.3%	9,168	0.0 (< 10^{-99})
Chr9	3,032,982	0.2%	9,752	0.0 (< 10^{-99})
Chr10	3,325,116	0.2%	8,911	0.0 (< 10^{-99})
Chr11	289,653	0.7%	1,787	7.44^{-05}
Chr12	169,361	1.2%	957	3.51^{-01}
Chr13	138,786	0.5%	770	0.0 (< 10^{-99})

Note: Nucleotide diversity represents the percentage of variable sites at each chromosome and was estimated by pairwise-deletion (ignoring missing/ambiguity sites). *p-values under the null hypothesis of no recombination using the Phi statistic.

I additionally constructed incompatibility matrices for phylogenetically informative sites in each chromosome (**Figure 15**). I found evidence of incompatible pairs of sites along all chromosomes. Incompatible pairs can be explained by either recurrent mutation or recombination. Under an infinite-sites model (Kimura 1969) of sequence evolution, the possibility of a recurrent mutation does not exist, so incompatibility for a pair of sites implies that at least one recombination event must have occurred. However, if the mutation rate is higher than the recombination, the infinite-sites assumption is violated. For that reason it is difficult to determine whether the great amount of incompatible pair sites showed in **Figure 15** are due to recombination or recurrent mutation.

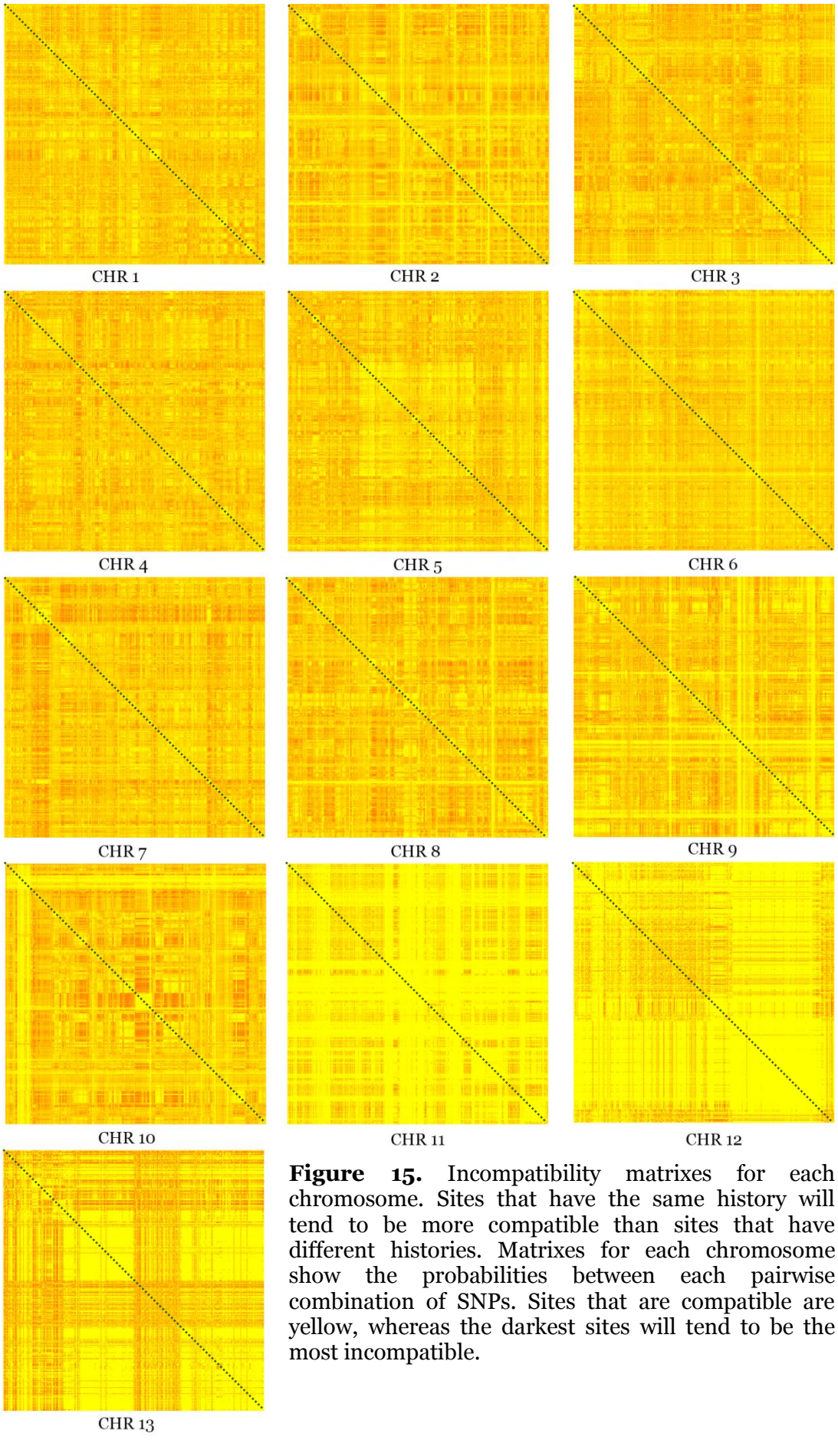

Figure 15. Incompatibility matrixes for each chromosome. Sites that have the same history will tend to be more compatible than sites that have different histories. Matrixes for each chromosome show the probabilities between each pairwise combination of SNPs. Sites that are compatible are yellow, whereas the darkest sites will tend to be the most incompatible.

Divergence between the sequences was low (**Table 7**), making it unlikely that recurrent mutation is the main cause of incompatibility between sites. To further analyze individual local regions at each chromosome, I estimated the Phi statistic using a sliding-window approach. P-values for Phi were individually estimated within local regions (1,000 nt) with step sizes of 500 nt (**Figure 16**). This analysis demarcated 134 regions (out of the 65,824 tested, representing 0.2%) showing a p-value < 0.01 (after correction for multiple comparisons). Neither specific chromosome nor region inside each chromosome showed an increased recombinant signal.

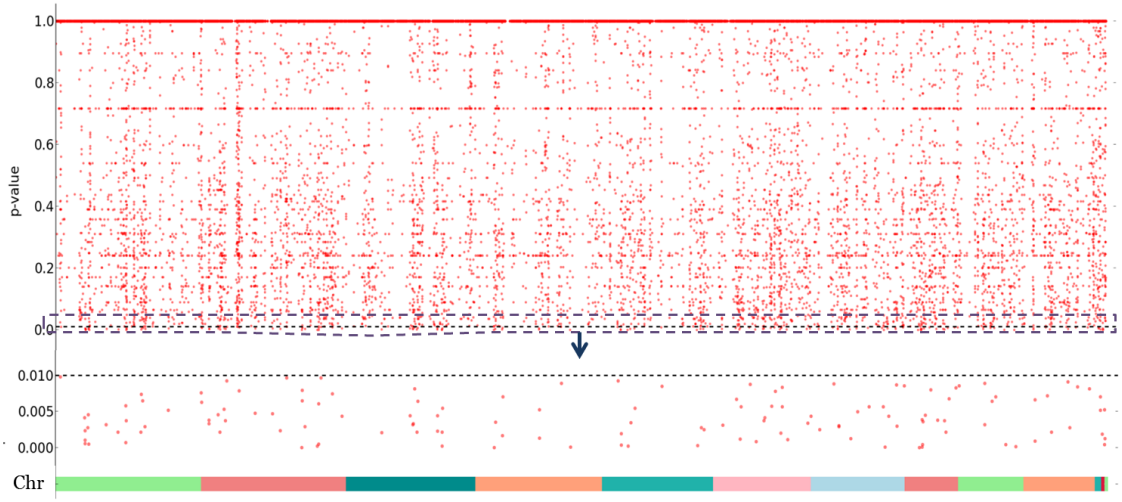

Figure 16. Sliding-windows analysis for recombination using the Phi test across whole genome sequence. Each chromosome was individually tested at each 1,000nt genomic region with step sizes of 500nt. Dots indicate p-values for each of the 65,824 windows after correction for multiple comparisons using the Benjamini & Hochberg (1995) procedure. At the bottom, zoomed in region showing the 134 statistically significant regions (p<0.01). Color bar at the bottom indicates each of the 13 chromosomes.

In order to further analyze recombination events at the genomic level, I looked for evidence of genome-wide mosaic structures between the eight isolates at the multiple sequence alignment of different genomic regions. I first performed a multiple genome alignment between the *de novo* assembled contigs and M1.001 contigs using Mauve (Darling et al. 2004; Darling et al. 2010). I then extracted Locally Collinear Blocks (LCBs) from the alignment and analyzed them looking for mosaic structures using 3SEQ (Boni et al. 2007) (**Table 8**). I examined a total of 37,645 LCBs, of which 37,167 were tested with 3SEQ since they presented one sequence for each genome. I detected 1,889 (5.1%) LCBs with

evidence of mosaicism ($p < 0.01$ after Bonferroni correction provided by 3SEQ) (**Table 8**). Large chromosomes (from 1 to 10) showed a similar percentage of LBCs with evidence of mosaicism (between 4.6% and 5.4%), while results were more diverse for minichromosmes (11, 12 and 13). The higher percentage was observed for chromosome 11 (9.6%), while chromosome 13 did not show any LCB with signatures of mosaic structure. On the other hand, Mauve aligner identified just one LCB from chromosome 12, which also showed evidence of recombination in the 3SEQ analysis. The lower number of LCBs identified on chromosome 12 is probably related with its high diversity (**Table 7**), which makes it difficult to find homologous genomic regions to build LCBs. An additional outcome from this analysis was regarding the length of the LCBs showing evidence of mosaicism. The average sequence length for LCBs showing evidence of mosaicism was 715 nt, much higher than the average length of all LCBs examined (579 nt). A plausible explanation for such results has to do with the algorithm, since 3SEQ only examines variable nucleotide positions from the alignment, the chance to find polymorphic sites increases with the alignment sequence length.

Table 8. Statistics from the mosaic structure analysis applied to each LBC derived from multiple genome alignment.

Chromosome	LCBs	Mean length LCB	Tested (3SEQ)	Positive (3SEQ)	Mean length LCB Positive
Chr1	5,470	583	5,398	294 (5.4%)	745
Chr2	4,924	584	4,839	262 (5.4%)	730
Chr3	4,631	574	4,592	233 (5.1%)	714
Chr4	4,748	577	4,693	226 (4.8%)	698
Chr5	3,971	575	3,922	190 (4.8%)	715
Chr6	3,139	576	3,095	143 (4.6%)	734
Chr7	3,288	583	3,248	168 (5.2%)	717
Chr8	2,047	573	2,026	97 (4.8%)	733
Chr9	2,302	574	2,272	120 (5.3%)	724
Chr10	2,310	565	2,283	106 (4.6%)	703
Chr11	56	606	52	5 (9.6%)	717
Chr12	1	673	1	1 (-)	-
Chr13	48	503	48	0 (0%)	-
Unknown	710	571	698	44 (6.3%)	697
Total	**37,645**	**579.7**	**37,167**	**1,889 (5.1%)**	**715.4**

Note: LCBs indicate the total number of LCBs identified by Mauve aligner. Positive (3SEQ) represent the number of LCBs showing a p-value<0.01. Unknown chromosome corresponds to unanchored contigs.

In order to determinate whether assembly and/or alignment strategies were biasing our results in 3SEQ, I estimated p-values for mosaic structures based on the consensus sequences derived from the *Mapping Assembly*. 3SEQ program was run over each gene locus (CDS±500nt, see Material and Methods) containing sequences for all the eight isolates (**Table 9**).

Table 9. Statistics from mosaic structure analysis applied to each gene locus.

Chromosome	Loci	Tested (3SEQ)	Positive (3SEQ)
Chr1	1,761	1,706	41 (2.4%)
Chr2	1,522	1,473	45 (3.0%)
Chr3	1,435	1,415	28 (1.9%)
Chr4	1,456	1,444	23 (1.6%)
Chr5	1,220	1,199	18 (1.5%)
Chr6	990	970	13 (1.3%)
Chr7	1,010	992	16 (1.6%)
Chr8	662	634	19 (3.0%)
Chr9	791	761	28 (3.7%)
Chr10	704	697	19 (2.7%)
Chr11	24	16	1 (6.3%)
Chr12	10	0	0 (0%)
Chr13	21	21	0 (0%)
Unknown	400	312	6 (1.9%)
Total	**12,006**	**11,640**	**257 (2.2%)**

Note: Loci indicate the total number of loci (CDS±500nt) at each chromosome. Unknown indicate the number of gene predicted at unanchored supercontigs. Positive (3SEQ) represent the number of gene loci showing a p-value<0.01.

This strategy revealed a lower percentage of genomic regions with evidence of mosaicism (2.2%) than the analysis performed over LCBs. At least two main causes may account for such a result: 1) Genome Assembly: While LCBs were based on the *de novo* assembly, gene loci were extracted from consensus sequence estimated from the mapping assembly. The last approach, does not allow for indels or highly polymorphic regions, since such regions were probably masked during the consensus building because they did not have enough depth of coverage. 2) Genomic Regions Tested: While LCBs were randomly extracted from different regions in the genome (including non-coding, intergenic and repetitive regions), for the second approach I just used coding and neighboring regions for each gene. It is expected that intergenic regions show a higher

polymorphism than coding regions, which can increase the change of find mosaic structures regarding more conserved regions. Overall, I still find some evidence of recombination between the isolates. Interestingly, similar to at the LCBs approach, chromosome 13 does not show evidence of mosaicism, suggesting that positive results from the Phi test (**Table 7** and **Figure 17**) are due to recurrent mutation rather than recombination. Additionally, chromosome 12 does not show any gene loci with evidence of recombination, suggesting that recombination is not accounting for its higher diversity. As well as in the LCB approach, chromosome 11 showed the higher percentage of gene loci with evidence of mosaicism, although in this case such percentage was represented just by one gene locus. For large chromosomes, the higher percentage was for chromosome 9 (3.7%), which also showed one of the higher percentage of LCBs (5.3%) with mosaic structures. Overall, despite the reduced number of events detected at the gene locus approach, both methodologies using 3SEQ showed similar patterns, indicating that neither assembly nor alignment strategies significantly biased the results.

3.3.4. Phylogenetic relationships between the isolates

A phylogenetic tree showing the relationships between the isolates was built using a concatenated multiple sequence alignment of coding sequences from 8,288 genes (see Materials and Methods). I used genes presenting less than 20% of unambiguously called bases in all isolates, showing no evidence of recombination and having the corresponding orthologous sequence in the closely related species *C. higginsianum* (**Figure 17** and **Figure 18**).

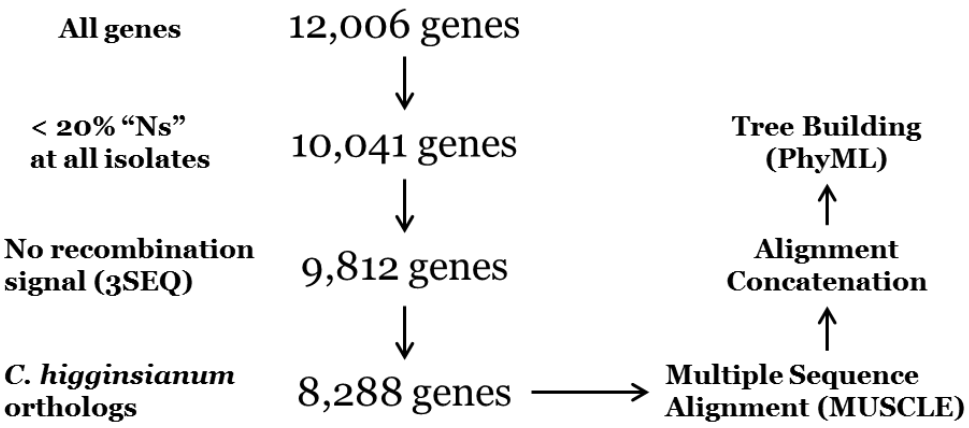

Figure 17. Pipeline designed to build the phylogenetic tree.

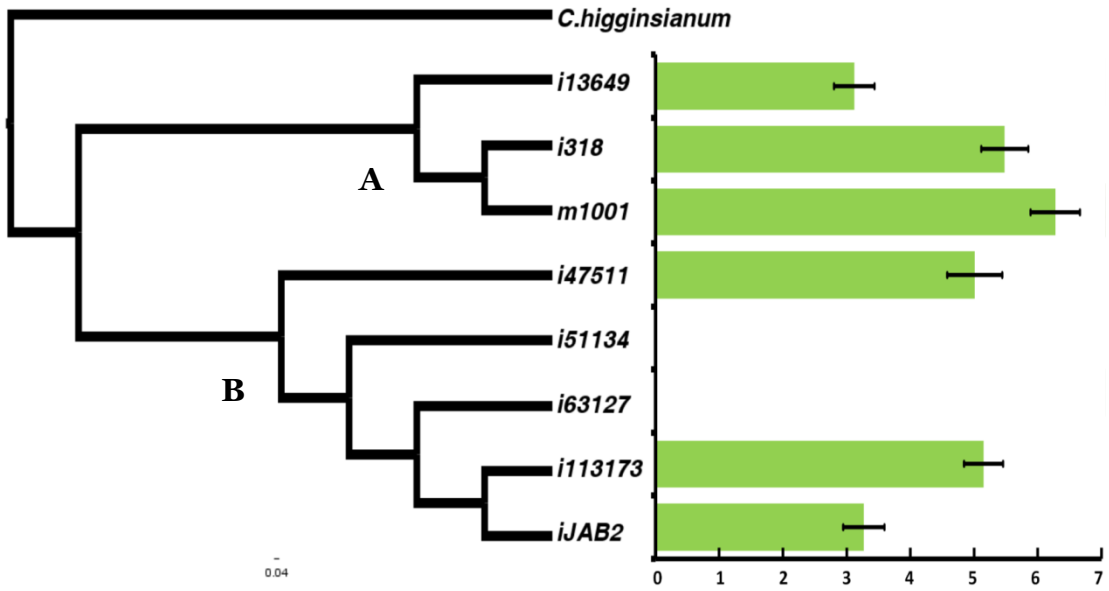

Figure 18. Maximum Likelihood tree showing genetic distances between the eight isolates and the close related species *C. higginsianum* based on multigene sequence alignment of 8,288 genes. At the right, average lesion size (mm^2) caused by each isolate on maize leaves.

Due to the conservative method used to select genes for the phylogenetic reconstruction, I expect these genes to be among the most functionally constrained between the *C. graminicola* isolates and *C. higginsianum*. In fact, just 14% out of 13,800,919 sites were polymorphic in the concatenated alignment. I did not find evidence of a relationship between virulence and genetic distance between the isolates. However, the phylogenetic tree was split into two main clades, one including isolates i318, M1.001 and i13649 (**Figure 18 A**) and the other containing isolates i47511, i51134, i63127, i113173 and iJAB2 (**Figure 18 B**). Clade A includes the two most virulent isolates (i318 and M1.001), grouped together in a sub-clade, and showing the greatest pairwise identity (99%). On the other hand, clade B includes two highly virulent isolate (i47511 and i113173), one moderately virulent isolate (iJAB2), the reduced virulence isolate (i51134) and the non pathogenic isolate i63127. The most parsimonious explanation for clade B topology required multiple acquisitions and losses in the ability of cause disease (e.g. the loss of the ability to cause disease in the common ancestor of isolate i51134, i63127, i113173 and iJAB2 and the gain of this ability in the ancestor of isolates i113173 and iJAB2). Even though these kinds of events have been previously described between members of the Ascomycota (Berbee 2001; Stajich et al. 2009), considering that only the most conserved genes were used for the phylogenetic reconstruction, I hypothesize that mayor changes at the genomic level (like gene gain and loss, rapidly evolving sequences and genomic structural variations) might represent the main factors accounting for the variability in virulence showed by *C. graminicola* isolates.

3.3.5. Analysis of genomic structural variations

Alignments of genomic sequences reads to the M1.001 genome were used to identify intra and inter-chromosomal variations using Breakdancer-Max (Chen et al. 2009). The low depth of coverage obtained for isolates iJAB2 and i51134 did not allow me to identify structural variations with confidence, so they were excluded. In total, 2,377 intra-chromosomal variations (inversions, deletions, insertions and translocations) were found between the five genomes analyzed (**Table 10**).

Table 10. Total number of each kind of intra-chromosomal structural variation detected in each genome.

Intra-chromosomal variations	i63127	i318	i47511	i113173	i13649
Inversions	26	27	51	56	231
Deletions	439	154	420	433	287
Insertions	47	34	14	26	27
Translocations	9	20	37	36	3
Total	**521**	**235**	**522**	**551**	**548**

Most genomes showed a similar number of intra-chromosomal structural variations (between 521 and 551), excluding the genome of isolate i318, which showed considerably fewer events (235). Most intra-chromosomal structural variations were represented by deletions, except for isolate i13649 which also showed a significant number of inversions. Many breakpoints were shared between the genomes, suggesting that these may represent genomic rearrangements in the M1.001 genome (**Figure 19**).

The total number of inter-chromosomal translocations for each isolate looked more variable than intra-chromosomal variations (**Table 11**). Genomes from isolates i63127 and i13649 showed the lowest number of events detected (20 and 13, respectively). An in depth analysis of breakpoint distribution showed that the largest number of inter-chromosomal translocations were in chromosome 6. This chromosome mainly interchanges DNA with chromosome 1 in isolate i47511 and with chromosome 4 in isolate i113173 (**Table 11** and **Figure 20**).

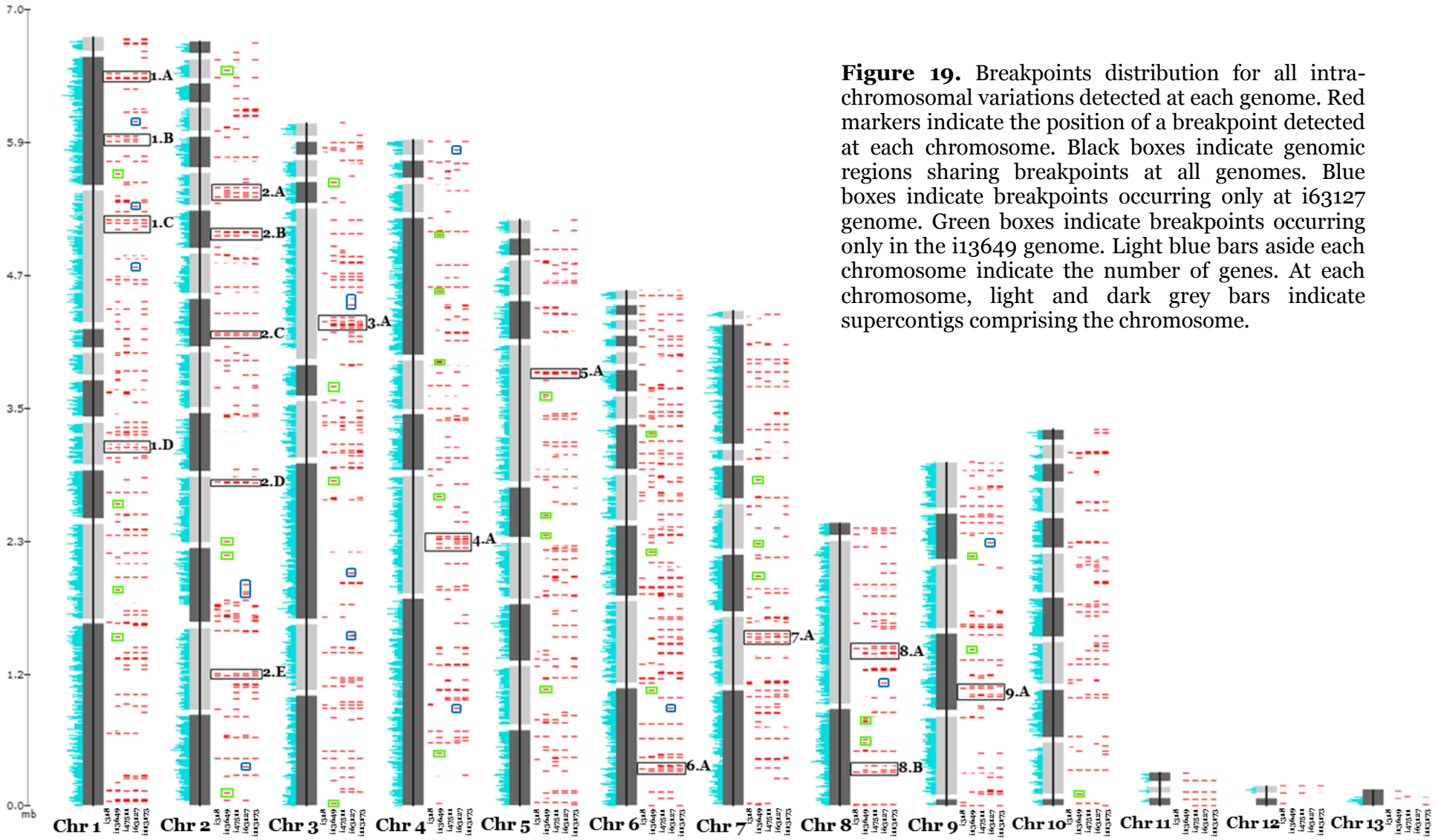

Figure 19. Breakpoints distribution for all intra-chromosomal variations detected at each genome. Red markers indicate the position of a breakpoint detected at each chromosome. Black boxes indicate genomic regions sharing breakpoints at all genomes. Blue boxes indicate breakpoints occurring only at i63127 genome. Green boxes indicate breakpoints occurring only in the i13649 genome. Light blue bars aside each chromosome indicate the number of genes. At each chromosome, light and dark grey bars indicate supercontigs comprising the chromosome.

Table 11. Number of inter-chromosomal translocations detected in each genome when compare to M1.001 genome.

Inter-chromosomal Translocations	i63127	i318	i47511	i113173	i13649
Total	**20**	**152**	**201**	**221**	**13**
Chr1	4	37	46	51	4
Chr2	8	41	45	53	5
Chr3	6	18	40	39	2
Chr4	4	29	38	46	4
Chr5	2	32	45	35	3
Chr6	2	**47**	**71**	**78**	5
Chr7	2	28	21	39	3
Chr8	1	12	7	9	0
Chr9	3	16	25	18	0
Chr10	5	28	42	40	0
Chr11	3	1	6	11	0
Chr12	0	10	6	11	0
Chr13	0	5	10	12	0

Note: Total indicates the number of events identified at each genome. Numbers per chromosome indicate the breakpoints detected at each chromosome (each translocation consist in two breakpoints at two different chromosomes).

Interestingly, despite the relatively low number of intra-chromosomal variations (**Table 10**), i318's genome showed a high number of inter-chromosomal translocations. On the other hand, genomes from isolates i63127 and i13649, which had a high number of intra-chromosomal variations, showed considerably fewer inter-chromosomal translocation events. Overall, the genome of i113173 showed the greatest number of structural variations (intra and inter-chromosomal).

By analyzing all structural variations together, I found that chromosome 6 was the most affected of the large chromosomes, showing 139 intra-chromosomal breakpoints per Mb and 45 inter-chromosomal translocation breakpoints per Mb. In addition, I found that minichromosomes (11, 12 and 13) show a considerably higher occurrence of both types of structural variations (intra and inter-chromosomal) (**Table 12**). One of the main characteristics of minichromosomes is that they are enriched with repetitive DNA (O'Connell et al. 2012). I observed a positive correlation between the percentage of repetitive DNA in the chromosome and the occurrence of chromosomal rearrangements (**Figure 21**).

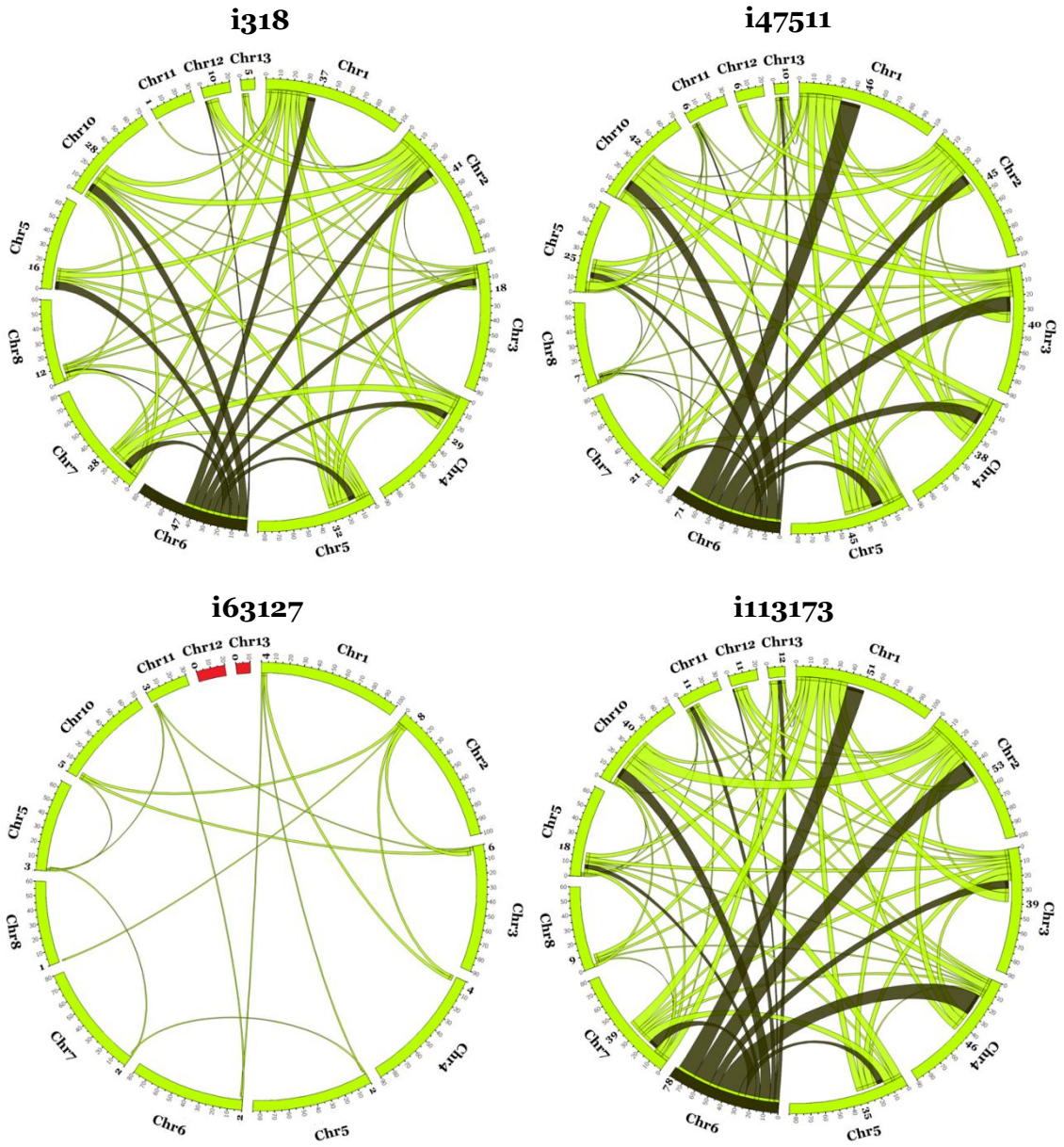

Figure 20. Circos plots depicting inter-chromosomal translocations among the four genomes showing the largest number of events. Translocations involving chromosome 6 appear in black, while the rest in green. Bold number aside each chromosome represent the total number of breakpoints. Line widths represent the number of translocations between two chromosomes.

Table 12. Intra and inter chromosomal variations per chromosome.

Chr.	Chr. Size (bp)	Intra SV	Intra SV/Mb	Inter SV	Inter SV/Mb	Repetitive DNA (%)
Chr1	6,787,984	659	97	142	21	5.12
Chr2	6,748,533	551	82	152	23	6.12
Chr3	6,027,927	567	94	105	17	5.06
Chr4	5,882,731	487	83	121	21	4.49
Chr5	5,174,327	452	87	117	23	5.86
Chr6	4,553,816	634	**139**	203	**45**	**8.55**
Chr7	4,372,745	438	100	93	21	5.87
Chr8	2,496,009	244	98	29	12	4.58
Chr9	3,032,982	318	105	62	20	6.68
Chr10	3,325,116	286	86	115	35	6.81
Chr11	289,653	64	**221**	21	**73**	**26.89**
Chr12	169,361	16	94	27	**159**	6.92
Chr13	138,786	38	**274**	27	**195**	**37.16**

Note: For each chromosome, the total number of intra-chromosomal variations (Intra SV) and the occurrence per Megabase (Intra SV/Mb); the total number of inter-chromosomal variations (Inter SV) and the occurrence per Megabase (Inter SV/Mb) and the percentage of repetitive DNA.

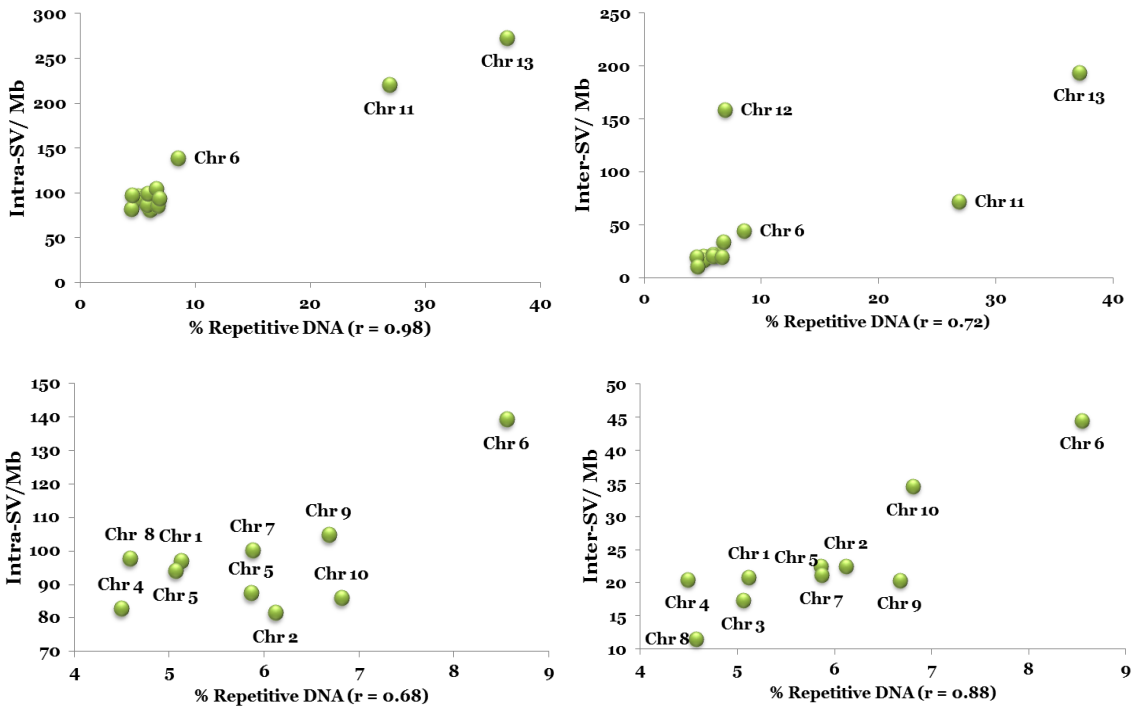

Figure 21. Correlation between the percentage of repetitive DNA and the occurrence of structural variations in each chromosome. At the top, intra and inter structural variations occurrence vs. % of repetitive DNA including all 13 chromosomes. At the bottom, same plots but including just large chromosomes. For each case, the Pearson' (r) correlation coefficient is showed.

For instance, a strong positive correlation was found between the percentage of repetitive DNA and the occurrence of intrachromosomal variations (**Figure 21**, upper-left, r=0.98). Most affected chromosomes were two of the minichromosomes (11 and 13) and chromosome 6. A similar pattern was observed for inter-chromosomal translocations (**Figure 21**, upper-right, r=0.72), although in this case, one of the minichromosomes (chromosome 12), which have a lower percentage of repetitive DNA, showed a high occurrence of translocations. By removing minichromosomes from the analysis, the correlation between the occurrence of intra-chromosomal variations and the percentage of repetitive DNA decreased (**Figure 21**, bottom-left, r=0.68). Conversely, this procedure increased the correlation between the occurrence of inter-chromosomal translocations and the percentage of repetitive DNA (**Figure 21**, bottom-right, r = 0.88). These results suggest that the amount of repetitive DNA, is affecting the occurrence of intra-chromosomal structural variations, mainly for chromosome 6, 11 and 13.

To explore the nature of genes affected by intra-chromosomal variations, I analyzed the function of genes located in genomic regions affected by three different kinds of breakpoints. First, I explored genomic regions showing breakpoints shared by all five genomes, likely to be a consequence of rearrangements only in in the M1.001 genome (the most virulent isolate). Such regions are indicated by black boxes in **Figure 19**. The analysis of all 17 regions together yielded a total of 374 genes putatively affected by the occurrence of neighboring intra-chromosomal variations (**Supplementary Table S1**). This set of genes was slightly but significantly enriched with *C. graminicola* specific effectors (Chi-square test, Yates corrected: $X^2 = 8.6$, p = 0.0016) (**Table 13**).

Table 13. 2x2 contingency table for genes annotated as putative species specific effectors.

	Other genes	***C. graminicola* specific effectors**
All chromosomal genes	11,198	34 (0.30%)
Genes at breakpoints	369	5 (1.33%)

Four out the five *C. graminicola* specific effectors are located in box 8.A (**Figure 19**, Chromosome 8: positions 1,303,016 to 1,425,522. Genes: GLRG_01031, GLRG_01041, GLRG_01042 and GLRG_01043). All of them encode small (<187aa) secreted proteins with unknown function and represent excellent candidates for functional validation. In addition, another 40 genes located in these regions encode secreted proteins and nine of them have sequence similarity with previously described virulence factors (**Supplementary Table S1**).

I analyzed genomic regions showing breakpoints unique to the i63127 genome (the non pathogenic isolate), represented by blue boxes in **Figure 19**. I identified 72 genes likely to be affected by the occurrence of neighboring intra-chromosomal variations (**Supplementary Table S2**). I found no statistically significant enrichments in this set, however I detected the presence of many genes coding essential proteins for a pathogenic lifestyle, such as one cellulose (GLRG_06263), two glycoside hydrolases (GLRG_03145 and GLRG_08886), and a tannase and feruloyl esterase (GLRG_01824). In addition, five of the 72 genes are upregulated in isolate M1.001 during the necrotrophic phase of infection (O'Connell et al. 2012). These genes, showing structural variations exclusively in the genome of the endophytic isolate i63127, also represent excellent candidates for functional validation since disruptions in any of them could be the cause of the absence of phatogenic phase in this isolate.

Finally, I analyzed genomic regions affected by the occurrence of intra-chromosomal variations unique to the genome of isolate i13649 (green boxes, **Figure 19**). Isolate i13649 was collected in Alabama (USA), a geographic region relatively close to the region where isolate M.1001 was originally collected (**Figure 11**). Isolate i13649 has high genetic similarity with two of the most virulent isolates (M1.001 and i318) (**Figure 18**), but a clear difference in terms of percentage of spore germination, growth rate in minimal medium and sporulation as well as virulence against maize (**Table 4** and **Figure 12**), therefore genomic structural variations uniquely found in the genome of i13649 could shed light into the genetic mechanism involved in such phenotypic characteristics. I identified 46 genes affected by the occurrence of neighboring intra-chromosomal variations (**Supplementary Table S3**) in this genome.

Interestingly, this set of genes was significantly enriched (Z Test, Z = -3.23, p = 0.0012) with a class of transcription factors containing a DNA-binding domain that consists of six cysteine residues bound to two zinc atoms (Zn2Cys6) (Genes: GLRG_01663, GLRG_03014, GLRG_07227 and GLRG_09181). This class of zinc finger transcription factors is unique to fungi and have been described as involved in different regulatory functions such as the production of toxins, secondary metabolites, meiosis, growth and conidiation (Burger et al. 1991; Woloshuk et al. 1994; Anderson et al. 1995; Brown et al. 2007; Zhao et al. 2011). In addition, seven genes affected by the occurrence of genomic variations at this isolate are upregulated during infection of M1.001 in maize. All these genes also represent excellent candidates for functional validation since may prove important in looking for genes affecting phenotypic characters directly or indirectly involved in the virulence of *C. graminicola*.

3.3.6. Gene content

In order to analyze the gene content of each isolate, I first annotated the genome sequences of the *de novo* assemblies using MAKER (Cantarel et al. 2008; Holt & Yandell 2011). Total number of gene models predicted for each isolate ranked from 9,635 to 11,484 (**Figure 22 A**). Because of the sequencing strategy, which was based only on short reads, the resulting genome sequences were quite fragmented. This fragmentation was much higher for genomes iJAB2 and i51134 (**Table 6**). The use of fragmented genomes implied that some genes were probably split into two or more predicted gene models whereas others could be truncated versions of the complete gene, which makes it extremely difficult to determine whether two gene models correspond to the same gene or to a duplicated version of the gene in the same genome. Because of this, in this first approach I did not attempt to analyze gene copy number variations between the genomes. Instead, I focused on two main goals: to identify genes shared between all isolates and to identify genes that are unique for each isolate.

In order to identify unique genes for each isolates based on sequence similarity, I used a strategy based on both global and local alignment. I initially identified gene models predicted in genomic regions showing significant global alignments with M1.001 transcripts. The remaining genes (**Figure 22 B**) were then clustered using MCL (Enright et al. 2002) based on local alignments resulting from an all vs. all BLASTP. The remaining sequences that were not placed into clusters were considered to be unique for each isolate (**Figure 22 C**). As expected, genomes for isolates iJAB2 and i51134 were the most affected by the fragmentation, and many gene models predicted for them did not show similarities with any other gene, even after correcting for fragmentation (**Figure 22 C**). Additionally, I identified 83 genes from the M1.001 genome that neither mapped to some genome nor were clustered by local alignment similarity. Based on both analyses, I partially solve the putative overestimation of gene models due to the genome fragmentation, at the cost of considering that each gene model represents either, a fragment of the same gene or a truncated version of a complete gene (**Table 14**).

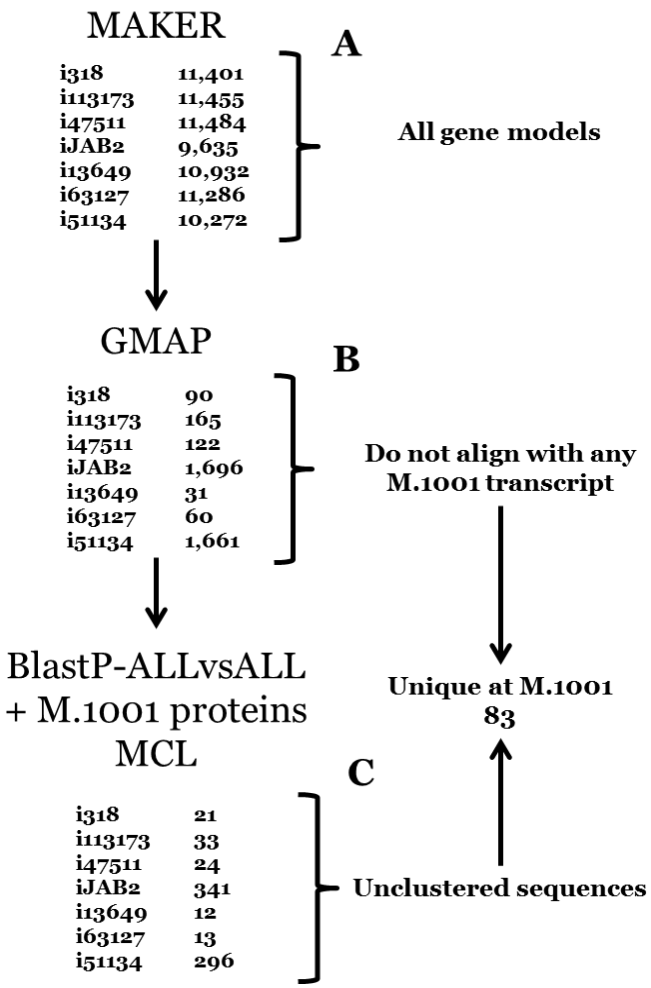

Figure 22. Pipeline and results for the gene content analysis. A) Gene models predicted by MAKER. B) Gene models located at genomic regions where none M1.001 transcript could be mapped. C) Unique gene after clustering with MCL based on all vs. all BLASTP results using proteins from genes identified at B.

Table 14. Total number of genes predicted after correcting for genome fragmentation.

Isolate	M1.001	i318	i113173	i47511	iJAB2	i13649	i63127	i51134
Genes	12,006*	11,219	11,182	11,208	8,607	10,853	11,148	8,940

Note: *Genes predicted by the Broad Institute annotation pipeline.

Due to the odd results retrieved for isolates iJAB2 and i51134 (considerably lower number of unique gene models: **Figure 22 C** and total number of genes: **Table 14**), and given that such results are likely consequence of the high level of genome fragmentation (**Table 6**), I did not include them for future analysis regarding gene content.

Based on the analyses of global alignments (GMAP) and local alignments (all vs. all BLASTP + MCL) and without taking into account genes predicted for genomes iJAB2 and i51134, I estimate a core gene set (genes present at all isolates) of 10,379 genes (**Figure 23**).

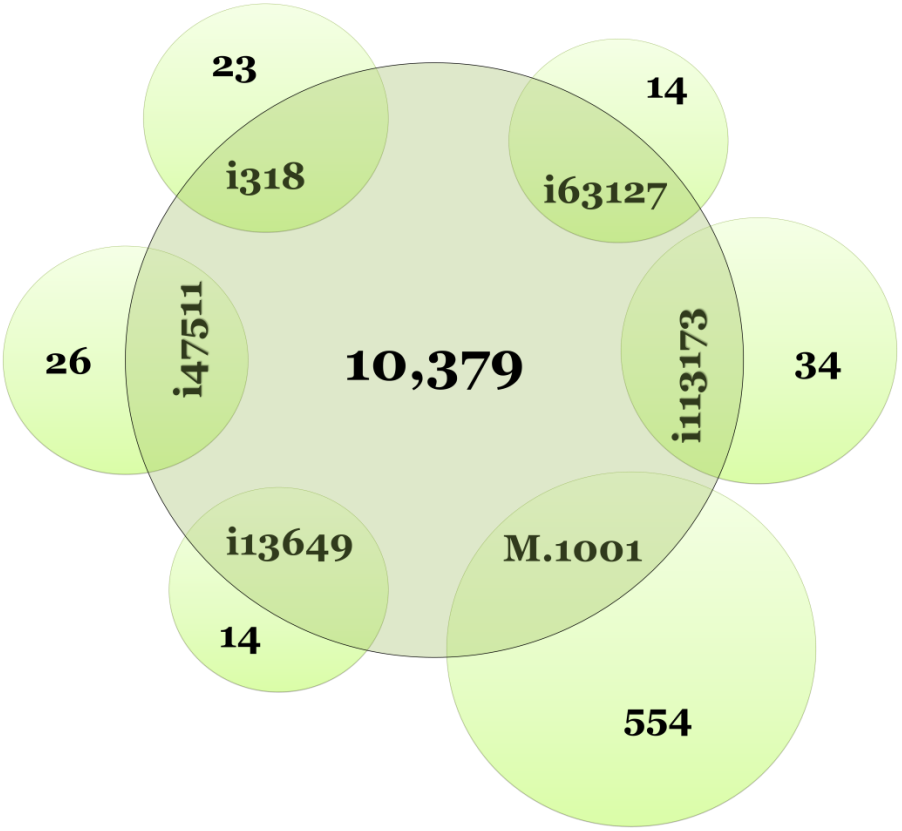

Figure 23. The number of genes shared by all genomes (core gene set) and unique genes in each genome.

Functional analysis of gene ontologies in the core gene set revealed enrichment for categories essential for the basic metabolism of the cell, including genes involved in primary metabolism, regulation of cellular and biological processes or genes coding for structural component of the cell (**Figure 24**). In addition, an in depth analysis of genes categories involved in pathogenicity (see Materials and Methods, **Chapter II**, for a description of functional categories), showed that genes coding for putative virulence factors, transporters and transcription factors are also overrepresented in the core gene set (**Table 15**).

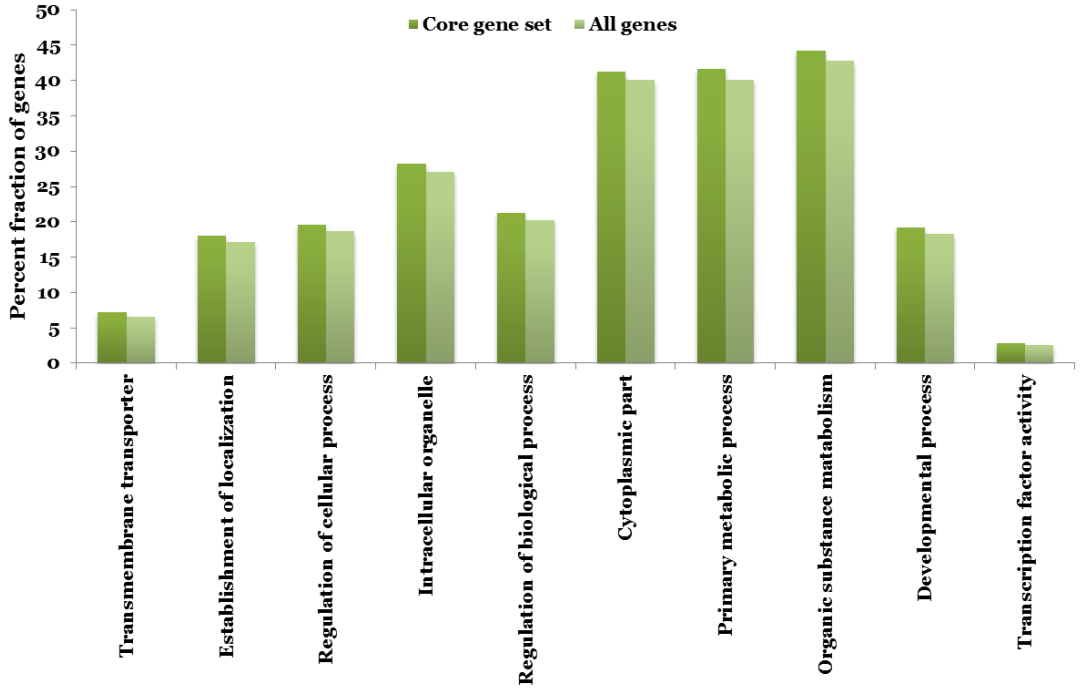

Figure 24. Representative gene ontology categories enriched in the core gene set of the *C. graminicola* isolates. All categories showed a significant p-value (< 0.01) for independence of rows and columns at the Fisher's exact test, after correction for multiple comparisons using Bonferroni method.

Table 15. Enrichment test for gene categories involved in pathogenicity and/or virulence.

Gene category	Proportion	Proportion in core set	p-value
Virulence Factor	1,445/12,006	1,356/10,379	**1.03E-20**
Transporters	662/12,006	625/10,379	**7.03E-11**
Transcription Factors	551/12,006	517/10,379	**9.62E-08**
Cytochrome P450	147/12,006	137/10,379	6.53E-02
Secreted Proteases	110/12,006	101/10,379	5.26E-01
CAZymes	494/12,006	437/10,379	9.06E-01
Secondary Metabolism	300/12,006	264/10,379	1.00E+00
Secreted	1,347/12,006	1,165/10,379	1.00E+00
Genus Specific Effectors	177/12,006	136/10,379	1.00E+00

Note: P-values estimated for each 2x2 contingency table were corrected for multiple comparisons using Bonferroni method. See Materials and Methods, **Chapter II**, for a description of functional categories.

I also identified unique genes in each genome (**Figure 23**). The largest set of unique genes was found in the reference genome M1.001. This set of 554 genes represents genes that were not identified in any of the *de novo* assembled

genomes and therefore could represent evolutionary innovations in M1.001, the most virulent isolate among all the isolates analyzed in the present study. Among the 554, I identified 87 genes coding for proteins involved in the infection process, including 10 secondary metabolites, 20 carbohydrate-active enzymes (CAZymes), 3 *C. graminicola* specific effectors, 1 genus specific effector, 47 genes upregulated during infection (O'Connell et al. 2012) and 33 secreted proteins (**Table 16**). Although it is difficult to determine whether genome fragmentation or gene annotation processes are accounting for the absence of these genes in the resequenced isolates, finding none of them in any of the genomes provides strong support for the hypothesis that most of these genes represent evolutionary innovations in the genome of isolate M1.001.

Moreover, I analyzed genes predicted to be unique in each of the *de novo* assembled genomes (**Figure 23**), also representing putative innovations in each of the isolates. None of the proteins encoded by the unique genes showed significant similarities to other proteins in the NCBI nr BLAST database and most of them are less than 200 aa long (**Table 17**). Additionally, a high percentage of the unique proteins are predicted to be secreted, especially for the most virulent isolates i318 and i113173. When comparing with the proportion of genes coding for secreted proteins in M1.001 (11%), I found that the set of new genes was enriched with secreted proteins in isolates i318 (35%, p=0.00036), i113173 (26%, p=0.0049) and i13649 (29%, p=0.04), but not in isolates i47511 nor in the non pathogenic isolate i63127 (**Table 17**). I searched for conserved protein domains in these proteins using InterProScan (Quevillon et al. 2005). Very few proteins showed sequence similarity to domain or family models in the InterPro database. Among them, I found a beta-D-glucan hydrolase (IPR002772) and a peptidase C48 in isolate i318, a phosphoribosylglycinamide synthetase (IPR020560) in isolate i13649, a clavaminate synthase (IPR003819) in isolate i47511, an aspartic peptidase (IPR021109) in isolate i63127 and a cysteine proteinase (IPR003653) and chaperone DnaJ-like protein (IPR024586) at isolate i113173 (**Table 17**). These sequences, together with all the others without assigned functions could represent evolutionary novelties specific for each isolates and represent target genes for functional validation experiments. Of particular interest are unique small secreted proteins, since they could be part of the isolate-specific effectors repertory.

Table 16. List of the 87 genes from M1.001 that are not present in any of the re-sequenced isolates and with putative functions associated with pathogenicity.

Gene ID	Len. (aa)	Annotation	Chr	Sec. Met.	CAZymes	Eff.	Up.	Sec.
GLRG_00111	408	hypothetical protein	Chr4			Cg		
GLRG_00173	1,081	Oxidoreductase	Chr4					X
GLRG_00509	1,116	hypothetical protein	Chr1					X
GLRG_00602	295	hypothetical protein	Chr1				X	
GLRG_00663	654	hypothetical protein	Chr1				X	
GLRG_00788	1,347	endo-beta-1	Chr1		X			X
GLRG_00995	447	hypothetical protein	Chr8					X
GLRG_01082	760	hypothetical protein	Chr8					X
GLRG_01103	1,663	glycosyl hydrolase	Chr8		X			X
GLRG_01263	920	malate dehydrogenase	Chr8				X	X
GLRG_01347	681	WSC domain	Chr3					X
GLRG_01358	609	hypothetical protein	Chr3				X	
GLRG_01660	551	hypothetical protein	Chr3				X	
GLRG_01848	527	hypothetical protein	Chr3			Cg		
GLRG_01981	1,023	aldose 1-epimerase	Chr3				X	
GLRG_02019	817	hypothetical protein	Chr3					X
GLRG_02026	527	hypothetical protein	Chr3				X	X
GLRG_02113	368	hypothetical protein	Chr4			Cg		X
GLRG_02271	1,591	hypothetical protein	Chr4				X	X
GLRG_02340	261	hypothetical protein	Chr4				X	
GLRG_02403	775	hypothetical protein	Chr5		X			
GLRG_02423	1,089	hypothetical protein	Chr5					X
GLRG_02521	913	hypothetical protein	Chr5					X
GLRG_02584	463	hypothetical protein	Chr5	X				
GLRG_02648	399	hypothetical protein	Chr5				X	
GLRG_02883	532	hypothetical protein	Chr1				X	
GLRG_02890	442	hypothetical protein	Chr1		X			
GLRG_03485	198	hypothetical protein	Chr7				X	
GLRG_03535	1,657	beta-galactosidase	Chr7		X			X
GLRG_03547	1,261	hypothetical protein	Chr7				X	
GLRG_03690	1,603	glycosyl hydrolase	Chr4		X		X	X
GLRG_04693	1,726	hypothetical protein	Chr8				X	
GLRG_04712	623	hypothetical protein	Chr8				X	
GLRG_05280	882	hypothetical protein	Chr2				X	
GLRG_05281	1,511	hypothetical protein	Chr2				X	
GLRG_05287	234	hypothetical protein	Chr2				X	
GLRG_05421	372	hypothetical protein	Chr2					X
GLRG_05496	369	hypothetical protein	Chr6			C		
GLRG_05507	851	glycosyl hydrolase	Chr6		X			X
GLRG_05869	2,001	rhamnogalacturonate lyase	Chr9		X		X	X
GLRG_05874	618	hypothetical protein	Chr9				X	
GLRG_06056	976	carbonic anhydrase	Chr5				X	
GLRG_06108	687	hypothetical protein	Chr5				X	

GLRG_06158	615	LSM domain	Chr5	X				
GLRG_06161	1,932	MFS transporter	Chr5	X				
GLRG_06314	404	hypothetical protein	Chr2				X	
GLRG_06465	1,061	hypothetical protein	Chr6					X
GLRG_06735	1,159	hypothetical protein	Chr3				X	
GLRG_06745	1,718	alkaline phosphatase	Chr3				X	
GLRG_06881	1,262	beta-glucosidase	Chr2		X			
GLRG_06934	1,276	aldose 1-epimerase	Chr2					X
GLRG_06968	360	LysM domain	Chr2		X			
GLRG_07009	1,366	Mannosyltransferase	Chr9		X			
GLRG_07191	747	hypothetical protein	Chr5				X	
GLRG_07704	999	GDSL-like	Chr3		X			X
GLRG_07727	522	tat pathway signal	Chr3					X
GLRG_07728	372	hypothetical protein	Chr3				X	
GLRG_07957	536	hypothetical protein	Chr4	X				
GLRG_08137	811	glycosyl hydrolase	Chr2		X			
GLRG_08320	1,406	glutathione S-transferase	Chr5				X	
GLRG_08472	1,433	hypothetical protein	Chr4					X
GLRG_08509	939	phytanoyl-CoA dioxygenase	Chr1				X	
GLRG_08913	932	PHB depolymerase	Chr9		X		X	X
GLRG_08950	543	hypothetical protein	Chr9	X				
GLRG_09062	1,353	hypothetical protein	Chr9				X	
GLRG_09070	797	glycosyl hydrolase	Chr9	X	X		X	X
GLRG_09270	1,162	pyoverdine/dityrosine	Chr7				X	
GLRG_09398	2,210	caspase domain	Chr1				X	
GLRG_09440	612	SCP-like	Chr1				X	X
GLRG_09506	633	hypothetical protein	Chr10				X	X
GLRG_09507	621	hypothetical protein	Chr10		X		X	
GLRG_09620	1,065	hypothetical protein	Chr2				X	
GLRG_09709	516	hypothetical protein	Chr10	X			X	
GLRG_09710	861	hypothetical protein	Chr10	X			X	
GLRG_10218	351	hypothetical protein	Chr7				X	
GLRG_10224	369	hypothetical protein	Chr7					X
GLRG_10770	1,035	Pectinesterase	Chr6		X			X
GLRG_10919	684	hypothetical protein	Chr2				X	X
GLRG_11025	1,219	hypothetical protein	super_73				X	
GLRG_11052	1,659	O-acetylhomoserine	Chr10				X	
GLRG_11267	951	hypothetical protein	Chr4		X		X	X
GLRG_11503	534	Acetyltransferase	Chr7	X				
GLRG_11643	463	hypothetical protein	super_99		X			
GLRG_11796	612	hypothetical protein	super_125		X			
GLRG_11860	1,502	FAD binding	super_150	X				
GLRG_11861	1,017	kelch domain	super_150					X
GLRG_11918	1,069	pyoverdine/dityrosine	super_205				X	

Note (**Table 16**): For each gene, Gene ID indicates the gene name according to the Broad Institute annotation, Len. (aa) indicate the length of the encoded protein, Annotation according to the Broad Institute and chromosome (Chr). Genes showing an "X" were annotated as: secondary metabolisms gene (Sec. Met.), carbohydrate-active enzymes (CAZymes), gene upregulated during infection on maize (Up.) and secreted protein according to SignalP program (Sec.). Eff. column shows *Colletotrichum graminicola* specific effector (Cg) or Genus specific effector (C). Annotation and classification of each gene is described at the section Material and Methods of **Chapter II**.

Table 17. Characteristics of genes uniquely identified in each isolate.

Isolate	Unique Seqs.	Average seq. length (aa)	SignalP (%)	InterPro
i318	23	182	8 (35%)**	IPR002772 IPR003653
i13649	14	121	4 (29%)*	IPR020560
i47511	26	132	6 (23%) n.s.	IPR003819
i63127	14	161	2 (14%) n.s.	IPR021109
i113173	34	136	9 (26%)**	IPR003653 IPR024586

Note: SignalP (%) Indicates the number and percentage of unique genes coding for proteins putatively secreted. A Z test for proportions was performed between each set of genes and all genes in the *C. graminicola* M1.001 genome. **, * and n.s. indicate significantly differences at 0.01, 0.05 and not significantly different respectively, when comparing the proportion of secreted proteins in the new genes set with the proportion of secreted proteins in the M1.001 genome. InterPro column indicates InterPro IDs matching with some of the unique proteins.

I analyzed genes present in all isolates but lost in the asymptomatic isolate i63127. I found 35 genes (**Table 18**), including two genus specific effectors (GLRG_04199 and GLRG_07719), two putative virulence factors (GLRG_06773 and GLRG_08164) one cutinase (GLRG_00892) and six genes upregulated during infection of isolate M1.001 on maize (GLRG_01034, GLRG_03065, GLRG_04402, GLRG_08155 and GLRG_08167). These genes, likely essential for the pathogenicity of *C. graminicola*, also represent very interesting genes for functional validation since they are uniquely lost in this asymptomatic isolate.

Table 18. List of the 35 genes present in all isolates but lost in the non pathogenic isolate i63127.

Gene ID	Len. (aa)	Annotation	Chr	CAZymes	Eff.	Up.	Vir. Fac.	Sec.
GLRG_00046	2501	hypothetical protein	chr4					
GLRG_00446	333	hypothetical protein	chr4					
GLRG_00883	733	hypothetical protein	chr1					
GLRG_00892	894	cutinase	chr1	X			X	
GLRG_01034	258	hypothetical protein	chr8			X		
GLRG_01212	263	hypothetical protein	chr8					
GLRG_01219	833	SAP domain protein	chr8					
GLRG_01400	1176	hsp20-like protein	chr3					
GLRG_02202	1428	hypothetical protein	chr4					
GLRG_02351	1779	WD domain-containing	chr4					
GLRG_02531	934	cenp-O kinetochore	chr5					
GLRG_03065	917	DJ-1/PfpI family protein	chr1			X		X
GLRG_04199	626	hypothetical protein	chr7		Cg			
GLRG_04402	1063	endo alpha-1,4	chr3	X		X		
GLRG_04623	1757	ubiquitin carboxyl hydrolase	chr8					
GLRG_04722	940	hypothetical protein	chr8					
GLRG_04733	538	hypothetical protein	chr8					
GLRG_05129	1569	hypothetical protein	chr2					
GLRG_05573	1068	hypothetical protein	chr6					
GLRG_05723	1018	hypothetical protein	chr9					
GLRG_05745	883	ubiquitin carboxyl hydrolase	chr9					
GLRG_05943	434	hypothetical protein	chr9					
GLRG_06262	775	phosphate transporter	chr2					
GLRG_06772	1113	hypothetical protein	chr3					
GLRG_06773	1260	hypothetical protein	chr3				X	
GLRG_06875	2213	hypothetical protein	chr2					
GLRG_06954	578	ribosomal protein L24	chr2					
GLRG_07719	510	hypothetical protein	chr3		C			X
GLRG_08155	279	hypothetical protein	chr2			X		
GLRG_08164	4364	GMC oxidoreductase	chr2				X	
GLRG_08167	975	hypothetical protein	chr2			X		X
GLRG_08168	1341	hypothetical protein	chr2					
GLRG_08169	1137	hypothetical protein	chr2					
GLRG_10065	1134	hypothetical protein	chr1					
GLRG_11581	438	hypothetical protein	super_94					

Note: For each gene, Gene ID indicates the gene name according to the Broad Institute annotation, Len. (aa) indicate the length of the encoded protein, Annotation according to the Broad Institute and chromosome (Chr). Genes showing an "X" were annotated as: carbohydrate-active enzymes (CAZymes), gene upregulated during infection on maize (Up.) and secreted protein according to SignalP program (Sec.). Eff. column shows *Colletotrichum graminicola* specific effector (Cg) or Genus specific effector (C). Vir. Fac. Indicates genes showing similarities with known virulence factors. Annotation and classification of each gene is described at the section Material and Methods of **Chapter II**.

3.4. Discussion

In the present chapter I introduce some phenotypic and genomic characteristics of a worldwide sample of eight isolates of the filamentous fungus *C. graminicola* with a wide range of virulence against maize (**Figure 11**). I used *in planta* assays to quantify virulence in terms of average lesion size produced by the isolates four days after infection. This experiment divided isolates into four overlapping categories (**Table 4**), ranging from asymptomatic isolates (i63127 and i51134) to highly virulent isolates (M1.001 and i318). In addition, *in vitro* assays also showed statistically significant differences between isolates in all experiments (**Table 4**). Most characteristics were positively correlated with the degree of virulence on maize (**Figure 13**). Among them, the growth rate on oatmeal agar (OA) medium was the most strongly correlated with the average lesion size. This medium induces sporulation in *Colleotrichum spp*., which may be showing that the ability to cause disease in *C. graminicola* could be related with faster sporulation. In addition, I also found a strong positive correlation between average lesion size and growth rate at minimum medium (MM), suggestive of a potential relationship between the virulence and the ability of the isolate to grow under extreme nutritional conditions. On the other hand, the correlation between virulence and *in vitro* growth rate significantly decreases when all nutritional requirements are supplied (complete medium, PDA). The ability of pathogens to access nutrients is a determinant for its success. When a parasitic relationship is established, the primary goal of pathogens is to obtain nutrients from their hosts for growth and reproduction. Meanwhile the host plant limits the pathogen's access to nutrients and initiates immune responses. Under these conditions, the pathogen's ability to survive using limited resources could make the difference between success and failure.

By removing asymptomatic isolates (i63127 and i51134) from the study, I found that the degree of virulence was more strongly correlated with the *in vitro* percentage of germination (appressoria formation) than with the total number of spores produced (**Figure 13**). Sporulation is essential to the organism not only for its persistence in the environment but also for increasing the chance to infect its hosts. However, it is well known that the ability of the organism to sporulate is highly dependent of environmental factors like humidity,

temperature, light and host health (Agrios 2005) and therefore, the amount of spores produced by the isolates under laboratory conditions just give us a vague idea about what actually happens in nature. Instead, the formation of the appresoria is more likely to be associated with the virulence of *C. graminicola* isolates, since penetration of the host cells only occurs after maturation and melanization of this structure, which creates exceptionally high hydrostatic pressures to facilitate the entrance (Bergstrom & Nicholson 1999). In order to further analyze possible causes of phenotypic variability between the isolates, I looked for genes affected by genomic structural variations occurring uniquely in the genome of isolate i13649. Although isolate i13649 show high genetic similarity with isolate M1.001 (**Figure 18**), they exhibit a clear difference in terms of percentage of spore germination, growth rate in minimum medium and sporulation (**Table 4** and **Figure 12**). I found 46 genes (**Supplementary Table S3**) affected by intra-chromosomal variations, some of which could explain the phenotypic difference observed. Interestingly, this set of genes was enriched with genes coding for Zn2Cys6 transcription factors, which are involved in multiple functions such as the production of toxins, secondary metabolites, meiosis, growth and conidiation (Burger et al. 1991; Woloshuk et al. 1994; Anderson et al. 1995; Brown et al. 2007; Zhao et al. 2011). For instance, in the filamentous fungus *Fusarium graminearum*, the absence of one Zn2Cys6 transcription factor (named EBR1) causes a significant reduction in the growth rate, virurelence, spore germination and conidiation (Zhao et al. 2011). Structural variations affecting functionality in some of these transcription factors in *C. graminicola* could potentially explain variations observed at the phenotypic level; therefore they represent excellent candidates for functional validation.

The complete absence of symptoms in isolate i63127 could be related with either the incapacity to form appresoria (W.A. Vargas and S.A. Sukno, *personal communication*) or the inability to secrete sufficient quantities of degradative enzymes like cutinases, cellulases, pectinases, and polygalacturonases which are responsible for degradation of the host cell wall that characterizes the destructive necrotrophic phase of the infection (O'Connell et al. 2012). However, isolate i63127 is still able to grow inside of maize leaves without apparent damage to the plant (W.A. Vargas and S.A. Sukno, *unpublished results*), which

suggests that this isolate could represent an endophytic variant of *C. graminicola*. Symptomless endophytes have been described in *Colletotrichum* spp. (Lu et al. 2004; Joshee et al. 2009; Rojas et al. 2010; Yuan et al. 2011). In fact, a single disruption event of a pathogenicity gene (called CPR1) transformed the pathogenic *C. graminicola* strain M1.001 into an endophyte-like strain, fully capable of penetrating and colonizing host cells during the initial biotrophic phase of the disease, but unable to switch to a necrotrophic lifestyle, and showing no symptoms of disease on maize leaves (Thon et al. 2002). On the other hand, recent studies showed that *C. graminicola* is unable to suppress many of the known plant defense mechanisms during the biotrophic phase, and the switch to a necrotrophic phase may enable it to avoid direct contact with host defenses by means of killing the plant cells (Vargas et al. 2012). Loss, disruption or mutation in genes involved in the change of lifestyle (from biotroph to necrotroph) may also be responsible of the incapacity of isolate i63127 to develop symptoms on maize leaves, since the switch is essential for the differentiation of secondary hyphae, which allow it to kill the host cells and proliferate as a necrotroph. In addition, lost of one or more genes coding for effector proteins, could affect the ability of the fungus to complete the entrance into the host, as demonstrated by our Group by knocking gene GLRG_04079 (a nuclear effector), which make the fungus able to develop functional appressoria, but unable to penetrate into the host cell (W. A. Vargas and S. A. Sukno, *personal communication*).

In the present study I identifed a set of genes partially affected or completely lost in the genome of isolate i63127. First, by looking at genes located at regions near breakpoints of intra-chromosomal structural variations (blue boxes, **Figure 19**) I found four genes coding for enzymes directly involved in the degradation of plant cell walls: one cellulose, two glycoside hydrolases and one a tannase and feruloyl esterase. One of them (GLRG_08886), together with another four (see **Supplementary Table S2**) are differentially upregulated during the necrotrophic phase of infection of isolate M1.001 (O'Connell et al. 2012). Furthermore, I identified 35 genes present in all virulent isolates, but lost in the genome of i63127 (**Table 18**). Among these genes, I found five that are upregulated during the necrotrophic phase in M1.001, four of which have unknown function and another encodes for a glycosyl hydrolase

(GLRG_04402). In addition, this set includes one *C. graminicola* specific and one genus specific effectors and two putative virulence factor. One of them, gene GLRG_06773, is highly conserved between all other isolates of *C. graminicola* and its ortholog in *Colletotrichum lagenarium* (gene CMK1) encodes a mitogen-activated protein (MAP) kinase involved in conidial germination and appresorium formation (Takano et al. 2000). The loss of this gene in i63127 may also explain the lack of spore germination in this strain, similar to the observed for gene CMK1 (Takano et al. 2000). Overall, genes located in regions close to structural breakpoints unique to genome i63127 and genes identified in all isolates but not present in i63127, represent excellent candidates for further studies of genetic basis for the loss of pathogenicity in this isolate.

One of the first questions raised in this chapter was whether there is evidence of recombination in the *C. graminicola* field isolates. To address this question, I applied two different methods using data obtained through the two genome assembly strategies. Initially, I performed a whole-chromosome analysis of incompatible sites using consensus chromosomal sequences from each isolate obtained by the mapping assembly. I applied the Phi test to this data in order to detect recombination by comparing the frequency and distribution of phylogenetically incompatible site-pairs with the frequency of such site-pairs expected in the absence of recombination (Bruen et al. 2006). Overwhelming evidence of recombination was found in all chromosomes with the exception of chromosome 12 (**Table 7** and **Figure 15**). I additionally estimated p-values for Phi in sliding windows of 1,000 nt along each chromosome and I found that approximately 2% of the analyzed windows show evidence of recombination (**Figure 16**). The Phi test is a powerful test for finding evidence of recombination between very closely related sequences (D. P. Martin et al. 2011), however, this test is based on the infinite sites model (Kimura 1969) and therefore assumes the absence of recurrent mutations (Liu & Fu 2008), which are relatively common in genomic regions with high mutation rates. Mutation hot-spots have been widely documented in fungal genomes and they are usually associated with the presence of highly repetitive regions containing transposons, which may promote the occurrence of mutations in several ways, for instance by means of repeat induced point-mutation (RIP), a fungal genomic defense mechanism that hypermutates repetitive DNA promoting C:G to T:A

transitions (Cambareri et al. 1989; Ikeda et al. 2002; Van de Wouw et al. 2010; Ben-Ami & Kontoyiannis 2012; Ma et al. 2012; Jeon et al. 2013). Given that the *C. graminicola* genome has a high percentage of repetitive DNA, the possibility of recurrent mutations cannot be excluded and therefore it is difficult to determine whether the great amount of incompatible pair sites detected is actually due to recombination or to recurrent mutation, especially for minichromosomes 11 and 13, which have a higher proportion of repetitive DNA (**Table 12**).

In order to further analyze recombination events at the genomic level, I looked for evidence of mosaic structures using the 3SEQ program (Boni et al. 2007), which infers mosaicism by analyzing sequence triplets in multiple sequence alignments. Mosaic structure exists in a nucleotide sequence if different segments of the sequence descend from different ancestors[8]. 3SEQ assesses all possible combinations of three sequences and evaluates the hypothesis that one of them is a recombinant of the other two. I initially applied this method to a set of Locally Colinear Blocks (LCBs) previously obtained by a multiple genome alignment of *de novo* assembled contigs. I found around 5% of the LCBs with statistical support for the presence of mosaicism (**Table 8**). Almost the same proportions of positive LCBs were found for each of the large chromosomes, but minichromosomes showed more variable results. For instance, chromosomes 11 and 13 showed the highest and lowest percentages of LCBs with signatures of mosaicism (9.6% and 0%, respectively), while just one LCB could be aligned from chromosome 12, which also presented evidence of mosaic structure (**Table 8**). Similarly, by using 3SEQ to analyze each gene locus extracted from the consensus sequence created by the mapping assembly, I found that the highest percentage of loci with signatures of mosaicism were also on chromosome 11 (6.3%), while none of them were significant on chromosome 13 (**Table 9**). These results suggest that the incompatible site pairs detected by the Phi test on chromosome 13 are likely to be consequence of recurrent mutation rather than recombination. Moreover, chromosome 11 showed evidence of recombination in

[8] During meiosis, chromosome by chromosome homologous recombination produces a new haploid gamete containing half of the genetic information from each parental DNA. This mixing of genomes leads to mosaic chromosome (haplotype) structure composed of segments of DNA from each parent.

all tests, while the high diversity of chromosome 12 makes extremely difficult to make conclusions about whether or not recombination is acting on it.

For large chromosomes, I detected evidence of recombination in all tests. The likely consequence for this result is the occurrence of sexual reproduction at some point during the evolution of the field isolates of *C. graminicola* (Awadalla 2003; D. P. Martin et al. 2011). Even though the sexual stage of *C. graminicola* has been described in laboratory conditions and it has never been found in nature (Crouch & Beirn 2009), the possibility of sexual recombination in field isolates of this pathogen have been previously proposed (Vaillancourt & Hanau 1992; Crouch et al. 2006). Evidence of a sexual cycle in natural strains of *C. graminicola* is not trivial, since it would impact the population biology of the species, perhaps enabling more rapid evolution of antifungal resistance or virulence factors. This fundamental aspect of the organism's biology must be taken into account when developing control strategies for emerging infectious diseases (EIDs) (Heitman 2006; Zeyl 2009).

However, several asexual fungal pathogens are known to recombine at least occasionally, for example *Magnaporthe oryzae* (Zeigler et al. 1997), *Candida albicans* (Alby et al. 2009), *Batrachochytrium dendrobatidis* (Farrer et al. 2013), *Aspergillus fumigatus* (Samson et al. 2009), *Aspergillus nidulans* (Schoustra et al. 2007), *Alternaria alternata* (Stewart et al. 2013), *Fusarium oxysporum* (Ma et al. 2010) and *Verticillium dahliae* (de Jonge et al. 2013), which implies that meiotic recombination is not the only mechanism able to explain the observed evidence of recombination in *C. graminicola*. Even though mechanisms of recombination are extremely varied across pathogenic species (Awadalla 2003), parasexual recombination[9] has been largely proposed as one of the most likely instruments accounting for the occurrence of recombination in asexual fungal species (Hastie 1964; Glass et al. 2000; Schoustra et al. 2007; Ma et al. 2010; de Jonge et al. 2013). Thus, the recombination that I detected may be due to either sexual recombination, or parasexual recombination. A

[9] Parasexual recombination: Results from the parasexual cycle in some fungi. In this case, two haploid hyphae fuse together, resulting in the formation of a diploid nucleus. This nucleus, usually unstable, can later return to a haploid state involving a mitotic crossing-over, which may result in the exchange of genetic material.

more in depth analysis, combining biological and population genetics data would enable us to more accurately determine whether or not sexual reproduction is accounting for the recombination detected in *C. graminicola* strains. For instance, a recent study carried out by Saleh et al. (2012) combines reproductive biology, population genetics and computer simulations to explore sexuality in supposedly clonal populations of *Magnaporthe oryzae*. By mean of sampling 456 strains from 55 countries and using microsatellite markers, the authors found that this pathogen actually reproduces sexually in regions near to it center of origin, but clonally outside of it. Such approaches, applied to *C. graminicola* populations, are likely to shed light on its reproductive behaviors.

Another issue addressed in this chapter was about the phylogenic relationships between the isolates of *C. graminicola*. A maximum likelihood tree based on the concatenated alignment of 8,288 genes (**Figure 18**) showed that the two most virulent isolates (M1.001 and i318) also represent the two most related isolates in terms of genetic similarity. This finding is even more interesting given the geographical regions from which these isolates were collected. *C. graminicola* strain M1.001 was collected in Missouri (USA) in the year 1978 from infected maize (Forgey et al. 1978) and isolate i318 was collected in Nigeria at the ends of the 1980s (Cardwell 1989). Nigeria is the largest maize producer in Africa (IITA 2013), and many of the varieties cultivated nowadays are the product of the breeding of maize materials introduced from all over the world (Iken & Amusa 2005). Due to the high phenotypic and genotypic similarities between these isolates, I suggest that isolate i318 might represent a clonal strain of isolate M1.001, which probably arrived in Nigeria by means of infected seeds (or any other plant material) from the United States. Furthermore, the phylogenetic tree did not reveal any tendency of clustering isolates according to their degree of virulence against maize (**Figure 18**). This implied that the great majority of genes used to reconstruct the phylogenetic tree (represented by the most conserved genes between the isolates) are apparently neutral in relation to virulence, which leads us to suggest that mayor changes at the genomic level, such as rapidly evolving sequences, genomic structural variations or the gain and loss of genes, could be mainly accounting for the differences in virulence showed by *C. graminicola* isolates.

I analyzed genomic alignments of paired-end sequence reads against M1.001 reference genome to identify genomic structural variations in each of the sequenced isolates as compared to the M1.001 genome. Among the five highest quality resequenced genomes 2,377 intra-chromosomal and 607 inter-chromosomal variations were detected. Chromosome 6 was the most affected of the large chromosomes by both kinds of variations (**Table 11**, **Figure 20** and **Table 12**), suggesting that this chromosome may represent a source for the generation of new genetic variants. In addition, chromosome 6 is considerably enriched with repetitive DNA, which is known to promote genome instability through recurrent excisions and insertions (Chen et al. 2010). A similar pattern was found for minichromosomes 11 and 13, which also have a high amount of repetitive DNA (**Table 12** and **Figure 21**). By analyzing genes located in the neighboring regions of intra-chromosomal breakpoints shared by all the isolates, I found five *C. graminicola* specific effectors (**Table 13**), expected to be involved in the pathogenicity of *C. graminicola*, indicating that chromosomal rearrangements are likely to be genetically non-neutral and that they may be increasing the adaptive capability of this pathogen by means of promoting variability in genes coding for proteins that interact with the host.

Finally, I analyzed the core gene set and unique genes in each isolate. This analysis revealed that the core gene set was enriched with genes essential for the basic metabolism of the cell (**Figure 24**), but also with putative virulence factors, transporters and transcription factors (**Table 15**), suggesting that even less virulent isolates have the basic machinery to be a pathogen. On the other hand, unique genes in each isolate were usually represented by small secreted proteins with unknown function and enzymes involved in the degradation of the cell wall (**Table 16** and **Table 17**). These genes could potentially represent evolutionary innovations in each isolate, directly involved in the host or environment specificity and represent excellent candidates for functional validation.

Overall in this chapter, in addition to presenting phenotypic and genomic characteristics of eight isolates of *C. graminicola*, I found striking evidence of recombination and I identified genomic translocations and variations in the gene content likely to be involved in the pathogenicity and specificity of the

isolates, providing a valuable resource for further functional and population genetic analyses in this pathogen.

4. CHAPTER II

Natural selection in coding and non-coding DNA in *Colletotrichum graminicola* isolates

4.1. Introduction

The rapid accumulation of genome sequences and the development of powerful statistical methods to detect signatures of adaptation provide us an unprecedented opportunity to increase our understanding of functionally important genomic regions. Even though main sources of adaptive characters causing phenotypic differences between organisms remain under debate (Hughes 2012), it is generally accepted that positive selection (selection in favor of advantageous mutations) plays an important role in the origin of new phenotypes (Anisimova & Liberles 2012). In fact, the evidence of selection acting on protein-coding sequences has increased enormously in the last twenty years (Fitch et al. 1991; McDonald & Kreitman 1991; Bishop et al. 2000; Bustamante et al. 2005; Aguileta et al. 2010; Rech et al. 2012). However, the high level of similarity between proteins (in number and function) from phenotypically very different organisms and the fact that a large proportion of the non-protein coding DNA of eukaryotic genomes is actually functional (Kondrasov 2005; Taft et al. 2007; Raffaele and Kamoun 2012), has led to many researchers to ask whether phenotypic diversity is mainly determined by changes in protein-coding sequences or in the non-coding regulatory sequences (King & Wilson 1975; Oleksiak et al. 2002; Gasch et al. 2004; Whitehead & Crawford 2006; Wray 2007). For that reason, much attention has recently been given to the molecular evolution of non-coding DNA sequences (for an updated review see Zhen and Andolfatto 2012). Nevertheless, studies of adaptive evolution in non-coding DNA are currently restricted to model organisms like yeast (Fay & Benavides 2005; Borneman et al. 2007; Ronald & Akey 2007; Emerson et al. 2010), *Arabidopsis* (Kim et al. 2007), *Drosophila* (Andolfatto 2005; Haddrill et al. 2008), mice (Kousathanas et al. 2011) and humans (Keightley et al. 2005; Haygood et al. 2007; Haygood et al. 2010). These surveys of non-coding sequences were performed using different methods, and patterns of natural selection acting at this level are sometimes difficult to interpret but it is becoming clear that natural selection acts on large portions of the non-coding genome. As stated by Zhen and Andolfatto (2012), "the emerging picture, in many eukaryotic organisms, is that a much larger fraction of non-coding DNA is functional and subject to both positive and negative natural selection than previously believed".

In this chapter, I have investigated patterns of selection operating on both protein-coding DNA sequences (CDSs) and non-coding intergenic and intronic sequences in a worldwide sample of eight strains of the filamentous fungus *Colletotrichum graminicola*. The genus *Colletotrichum* represents one of the 10 most economically devastating groups of plant pathogens, causing post-harvest rots and anthracnose spots and blights of aerial parts of the plant in a vast range of agronomic and horticultural crops throughout the world (**Figure 5**) (Cannon et al. 2012; Dean et al. 2012). *C. graminicola* only infects maize (*Zea mays*) (LeBeau 1950; Jamil & Nicholson 1991), producing annual yield losses of more than 1 billion dollars in the U.S.A. alone (Frey et al. 2011) and having a great potential to damage agricultural and natural ecosystems (Kamenidou et al. 2013). In addition, *C. graminicola* is a model organism for the study of hemibiotrophic pathogens, those that begin their infection as biotrophs (keeping alive the host cell) but later switch to a necrotrophic lifestyle, killing their hosts and feeding on dead cells (**Figure 7**) (Bergstrom & Nicholson 1999; Vargas et al. 2012; O'Connell et al. 2012).

The search for signatures of natural selection at the molecular level has become an increasingly popular strategy to identify genes involved in host-pathogen interactions, especially for phytopathogens, which are constantly subject to the strong selection pressure imposed by the host plant population (Stukenbrock & McDonald 2009; Van de Wouw et al. 2010). Of particular interest are genes coding for effector proteins or secreted pathogenicity factors that modulate plant immunity and facilitate infection (Kamoun 2007; Ellis et al. 2009). Two different hypotheses have been proposed to explain the population dynamics of pathogenicity genes during the co-evolution between plants and pathogens: the red queen hypothesis and the arms race hypothesis (**Figure 2**) (Woolhouse et al. 2002; Terauchi and Yoshida 2010). The red queen hypothesis, also known as the trench warfare hypothesis, is characterized by the persistence of dynamic polymorphisms caused by the action of balancing selection, which maintain multiple alleles in the population. In the arms race hypothesis, alleles offering a greater ability to both survive and reproduce, will rapidly increase in the population and will be eventually fixed. It is expected that such genes evolve faster than others by the action of positive selection, resulting in a large amount of substitutions at the genomic regions being selected. (Win et al. 2007;

Stergiopoulos et al. 2007; Van der Merwe et al. 2009; Van de Wouw et al. 2010; Pedersen et al. 2012). The arms race hypothesis is the one most widely used to describe the evolution of genes that encode proteins in the plant host and in the pathogen, that control the interaction between the plant and the pathogen although plant avirulence genes have been described as evolving in a manner consistent with the red queen model (Stahl et al. 1999; Okuyama et al. 2011).

Large scale scans for genes under positive selection in plant pathogenic fungi have revealed that genes coding for effectors show patterns of adaptive evolution, however these studies have also shown that genes involved in the transport of molecules across the plasma membrane, synthesis of secondary metabolites, protein and peptide degradation, carbohydrate metabolism, signaling transduction, hyphal development and differentiation and regulation of expression of other genes are also subject to adaptive evolution (Aguileta et al. 2010; Stukenbrock et al. 2011; Aguileta et al. 2012; Gibbons et al. 2012) (**Table 2**). These studies suggest that the ability of plant pathogens to infect and cause disease in their hosts (or to adapt to a new environment) rely, at least partially, on a wide range of rapidly evolving genes involved in multiple functions. Such genes could provide pathogens with a broad array of genetics variants to either avoid host defenses or generate new mechanisms of infection (Stukenbrock & McDonald 2009). However, as stated above, an increasing amount of evidence suggest that many phenotypic differences between eukaryotic organisms could also be due to changes on gene expression (Hahn 2007) and consequently it has been proposed that many traits involved in the host-pathogen interaction could potentially be determined by mutations in non-coding regulatory regions of the genome (Aguileta et al. 2009).

In order to investigate the selective pressures acting on different regions of the genome, I analyzed genome sequences of eight phenotypically and geographically diverse isolates of *C. graminicola* (**Table 3** and **Figure 11**). I first performed a genome-wide analysis of the empirically derived sliding-window distribution of Tajima's D values (Tajima 1989; Carlson et al. 2005; Kelley et al. 2006) and later I examined Tajima's D outliers in each class of coding and non-coding sequences looking for genomic regions showing unusual patterns of polymorphism. I looked for evidence of positive selection acting on

both coding and non-coding sequences, by comparing rates of substitution with the neutral expectation in a maximum likelihood (ML) framework. By analyzing coding and non-coding sequences showing extreme Tajima's D values and/or evidence of positive selection, I identified genes upregulated during infection and other genes involved in pathogenicity significantly enriched with 5'UTR and intronic regions with unusual patterns of polymorphisms relative to neutral expectations and 3'UTR sequences under positive selection. I found that positive selection in the CDS regions mainly affects genes coding for putative effector proteins and genes involved in the production of secondary metabolites. This study is the first report of selective pressures at the genomic level acting on both coding and non-coding DNA for an agronomically important phytopathogenic filamentous fungus. Our results are expected to contribute not only to our understanding of the evolutionary processes at the molecular level, but also to the development of environmentally friendly strategies to control fungal diseases.

4.2. Materials and Methods

4.2.1. Whole genome polymorphism analysis

I used the genomic coordinates of the 12,006 gene models predicted in *C. graminicola* M1.001 (O'Connell et al. 2012) to define the following classes of sequences: *coding* (CDSs), *introns* (all introns of the gene were concatenated), *5'upstream* (500 bp upstream of the transcription start codon), *3'downstream* (500 bp downstream of the transcription stop codon), *5'UTR* (120 bp upstream the transcription start codon) and *3'UTR* (200 bp downstream the transcription stop codon) (**Figure 26A**). For all classes, I extracted genomic sequences from each consensus genome and I clustered them together to create the multiple sequence alignments. 5'upstream and 3'downstream lengths were selected for be the region expected to be enriched with regulatory elements implicated in the control of transcription and translation (Xie et al. 2005; Kousathanas et al. 2011). The length of the UTRs were defined according to the average UTR length for fungi (Mazumder et al. 2003). In all cases sequences with more than 50% ambiguously called bases (Ns) were discarded. I included only the intergenic regions where the start and stop codons were annotated in all isolates. In addition, to preserve the identity of the non-coding sequences, I removed intergenic regions that, according the fixed lengths used in this study, overlapped with another intergenic region from a neighboring gene.

I concatenated all of the nuclear chromosomes together and analyzed them using Variscan *v2.0.3* (Hutter et al. 2006). I calculated Tajima's D values in 5 kb windows, with a 500 bp step size, comprising a total of 98,624 windows. All sites with more than three ambiguities were excluded from the analysis. I classified outlier windows according to the empirical distribution of Tajima's D values as Extreme Windows (EWs) when $D<5^{th}$ percentile (-1.33) or $D>95^{th}$ percentile (1.51). Additionally, I calculated the percentage of coding sequences at each window based on the *C. graminicola* M1.001 genome annotation, and the percentage of repetitive DNA at each window according to the annotation of transposable elements (O'Connell et al. 2012) in order to perform a point-biserial correlation test (r_{pbi}) between the membership to a EW (1) or not (0) and the percentage of coding or repetitive DNA in the window. Genes in EWs

were selected when at least 50% of genes sequences fitted in three consecutive EWs.

4.2.2. Site frequency spectrum analysis of different sequence classes

I pooled site classes across loci and used the Perl library Polymorphorama (Andolfatto 2007; Haddrill et al. 2008) to estimate Tajima's D for every individual sequence for each class (introns, 5'upstream, 3'downstream, 5'UTR, 3'UTR and coding). For coding sequences I separately estimated Tajima's D values for synonymous and non-synonymous sites as defined by the method of Nei and Gojobori (1986). Based on the empirical distribution of Tajima's D values for each class, I classified outlier values as D*: D*<0 (D< 5^{th} percentile) and D*>0 (D> 95^{th} percentile). To identify enriched functional or gene expression categories for outlier sequences, I performed Fisher's exact tests, correcting p-values for multiple comparisons using false discovery rate (FDR < 0.05) separately for each class.

4.2.3. Positive selection tests on coding regions

I analyzed all orthologous sets of CDSs with at least 3 sequences (11,995). Maximum likelihood phylogenies for each orthologous set were inferred using CodonPhyML (Gil et al. 2013). Positive selection was measured by the *dN*/*dS* ratio (ω), where *dN* represents the rate of non-synonymous substitutions per non-synonymous site and *dS* is the rate of synonymous substitutions per synonymous site. When a coding sequence is under negative selection, non-synonymous substitutions are constrained with respect to the neutral evolution due to their deleterious effect, and therefore ω < 1. Under neutrality, the rate of synonymous substitutions is equal to the rate of non-synonymous substitutions (ω = 1). Alternatively, if the sequence is evolving under positive diversifying selection, *dN* > *dS* and ω > 1. I estimated ω using Markov codon models using the ML approach as implemented in the CODEML program from PAML *v4* (Yang 2007) software package. I fitted six site models of codon evolution to each sets of orthologous sequences and obtained the optimized log likelihood (lnL) values for each model. Three likelihood ratio tests (LRTs) were performed. The significance of the tests was evaluated using the LRT statistic 2*(lnL1-lnL0) =

2ΔL, which was compared with a χ^2 distribution (see Anisimova et al. (2001) for details) to test whether there were statistical differences between the null (0) and the alternative (1) models. The LRTs compared the following models: M0 vs M3 to test for heterogeneity in ω among sites in a sequence, and M1a vs M2a and M7 vs M8 both to test for positive selection ($\omega > 1$). I considered a sequence as evolving under positive selection when the LRT for the ω-heterogeneity and at least one of the LRTs for positive selection were significant, all with a p-value < 0.05.

4.2.4. Positive selection tests of non-coding regions

Positive selection tests of non-coding sequences were performed according to Wong and Nielsen (2004) using the HyPhy (Kosakovsky Pond et al. 2005) batch file written by Dr. Oliver Fedrigo (Haygood et al. 2007). In this analysis, the rate of nucleotide substitution in the non-coding region (*dNC*) is compared with the rate of an *a priori assumed* neutral rate of substitutions (*dS*) by $\zeta = dNC/dS$. The parameter ζ represents the nucleotide substitution rate in the non-coding region, normalized by the rate of neutral substitutions (e.g. synonymous substitutions in the adjacent coding regions). Therefore, under neutrality $\zeta = 1$, under negative selection $\zeta < 1$ and under positive selection $\zeta > 1$. I analyzed each non-coding sequence using as neutral substitution rate the pooled synonymous substitution rate of the adjacent gene as well as the upstream and downstream genes. In addition, I only analyzed coding and non-coding sequences present in all isolates, as previously described. The ζ values were also estimated in an ML framework, which allows us to test hypotheses concerning this parameter using LRTs. I fitted three different models to the data according to Wong and Nielsen (2004): the neutral model (NM), the two-category model (2CM) and the three-category model (3CM), assuming no positive selection ($\zeta \leq 1$), allowing for $\zeta < 1$ or $\zeta \geq 1$ and allowing for $\zeta < 1$, $\zeta = 1$, or $\zeta > 1$, respectively. Two different LRTs were then performed: NM vs 2CM and NM vs 3CM for each non-coding orthologs sequence for all classes analyzed. I considered a sequence as evolving under positive selection when at least one LRT showed a p-value < 0.05.

4.2.5. Enrichment analysis of functional categories

To investigate if selective pressures act preferentially on specific types of sequences, I analyzed nine functional categories relevant to pathogenicity. Seven categories (Carbohydrate-Active Enzymes (CAZymes), Cytochrome P450, Genus specific Effectors, Secondary Metabolism, Secreted Proteases, Transcription Factors and Transporters) were previously described by (O'Connell et al. 2012). Due to the extremely conservative definition of the Genus-specific Effector proteins as "predicted extracellular proteins without any homology to proteins outside the genus *Colletotrichum*" (O'Connell et al. 2012), I also analyzed two additional categories likely to be involved in pathogenicity: all putative secreted proteins (potentially also effectors) and putative virulence factors. I identified secreted proteins using SignalP *v4.0* (Petersen et al. 2011). Putative virulence factors were annotated by performing whole-proteome BLASTp searches against the Pathogen Host Interaction Database (PhiBase v3.2) (Winnenburg et al. 2006) and against the Database of Fungal Virulence Factors (DFVF) (Lu et al. 2012) and I classified as putative virulence factors those genes showing at least one hit (e-value ≤1e-10) in both databases. In addition, I assigned gene ontology (GO) terms to *C. graminicola* genes using Goanna *v.2* (McCarthy et al. 2006) based on sequence similarity using BLASTp. I used the UniProt and AgBase_community (Fungi) databases filtering out sequences and annotations with automatically assigned GO terms (GO evidence code: IEA). An e-value ≤1e-5 and at least three BLAST hits with the same GO term was required to transfer annotations. At least one GO category was identified for 8,176 (68%) genes and each gene was also considered to belong to all parent categories of the directly assigned GOs (Kosiol et al. 2008). I analyzed only GO terms with at least 5 genes. Finally, I also analyzed differentially expressed genes in *C. graminicola* during different phases of infection in maize. I analyzed upregulated genes during infection at three different categories according to experimental RNA-seq data (O'Connell et al. 2012): Biotrophic/PA (significally upregulated genes in biotrophy regarding *in planta* appressoria), Necrotrophic/PA (significally upregulated genes in necrotrophy regarding *in planta* appressoria) and Necrotrophic/Biotrophic (significally upregulated genes in necrotrophy regarding biotrophic phase). A detailed description of whole genome gene annotation and gene categories analyzed at the present

study is showed in **Supplementary Table S4**. All enrichment tests were performed by creating 2×2 contingency table for the number of genes assigned or not assigned to the category and by estimating the p-value for independence of rows and columns by the Fisher's exact test, corrected for multiple comparisons (false discovery rate (FDR) or Bonferroni).

4.3. Results

4.3.1. Whole-genome nucleotide polymorphism analysis

I analyzed whole-genome sequences of eight isolates of *C. graminicola* from different regions of the world with a variable range of virulence against maize (**Table 3**, **Table 5** and **Figure 11**). I examined whole-genome polymorphisms by analyzing the empirically derived sliding-window distribution of Tajima's D values in order to identify regions in the genome showing unusual patterns of nucleotide polymorphism (**Figure 25**). Tajima's D reflects the difference between two estimators of the population mutation rate: the Watterson's estimator of nucleotide diversity per site (Θ_w) and the average pairwise nucleotide differences per site (Π). Θ_w is influenced by the number of segregating sites (it assigns the same importance to all polymorphisms), while Π is more affected by frequent polymorphisms. Under neutrality, the average of Θ_w and Π are expected to be equal, so Tajima's D = 0. Significant deviations from zero (e.g. an excess of low or high frequency polymorphisms) represent a signal of non-neutrally evolving sequence (e.g. D > 0: balancing selection or D < 0: positive or purifying selection). A potential problem with neutrality tests is that they can be difficult to interpret because of the confounding effects of demographic processes and/or sequencing errors (Achaz 2008) (e.g. D > 0 as a consequence of population subdivision, D<0 due to population expansion or D<0 due to large Θ_w values derived from randomly called SNPs from sequencing errors). However, since such events are expected to affect all regions in the genome equally (Bickel et al. 2013), looking for unusual values of the empirical distribution of Tajima's D values is a valid approach to identify genomic regions under selection (Przeworski et al. 2000; Carlson et al. 2005; Biswas & Akey 2006; Kelley et al. 2006). I classified windows with outlier Tajima's D values as Extreme Windows (EWs) if Tajima's D values were greater than the 95th percentile and lower than the 5th percentile of their genomic distribution. I identified 2,040 genes present in EWs (see Materials and Methods for a description on how genes were selected). In addition, I divided EWs into positive EWs (D>0) and negative EWs (D<0). Negative EWs are expected to contain the most selectively constrained genes since they represent windows with excess of low frequencies polymorphisms, likely to be a consequence of

positive or negative selection, whereas positive EWs are expected to contain sequences with an excess of intermediate frequency polymorphisms mainly as a consequence of balancing selection or relaxation of selective constraints. A gene ontology (GO) enrichment test did not reveal an enrichment of functional categories in either set of genes after correction for multiple comparisons. However, I found a weak positive correlation between negative EWs and the percentage of coding sequences in the window (r_{pbi} = 0.096, p = 6e-204) and a negative but stronger correlation with the percentage of repetitive DNA in the window (r_{pbi} = -0.118, p = 1e-304) (**Figure 25**). A negative correlation with the percentage of repetitive DNA is expected since most repetitive DNA is likely to evolve neutrally and not to be under strong selective constrains (Gaffney & Keightley 2006). However, the weak positive correlation of most constrained windows with the percentage of coding sequences and the lack of enrichment of GO functional categories may be suggesting that different class of polymorphisms at the whole window are under differential selective pressures. While functional polymorphisms (i.e. polymorphisms that alter the gene function or its regulation) are likely to be under positive or negative selection, non-functional polymorphisms (i.e. polymorphisms at repetitive DNA or at synonymous sites) are more likely to be neutral. I analyzed this scenario by examining the distribution of polymorphisms at different class of sites.

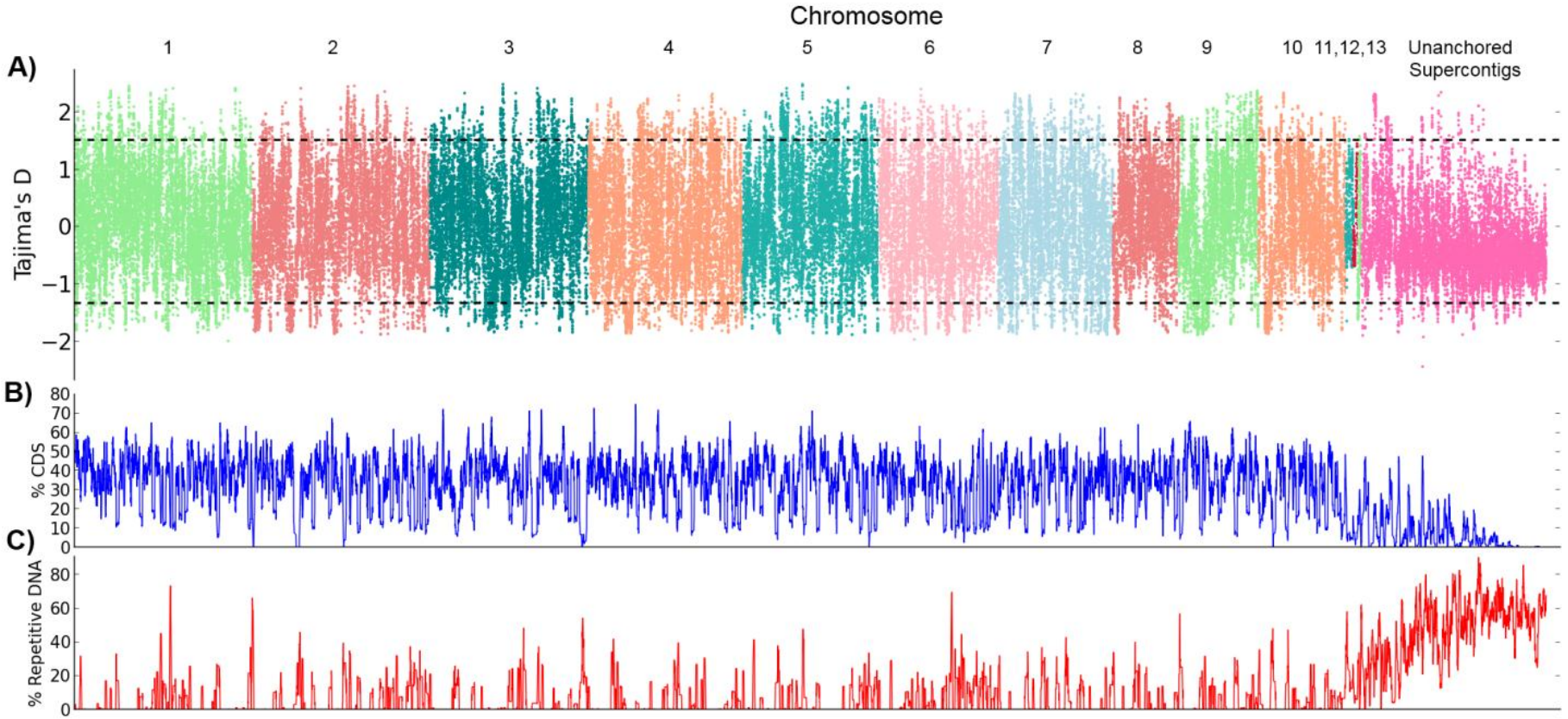

Figure 25. Sliding-windows analysis of Tajima's D values across the genome. **A)** Tajima's D values for 5kb windows with 500bp slides (jumps), comprising 98,624 windows covering the 13 chromosomes plus unanchored supercontigs. Dashed lines represent 5th and 95th percentiles. **B)** Percentage of coding sequence (CDS) and **C)** percentage of repetitive DNA at each window.

4.3.2. Different site frequency spectra between classes of sequences

In order to thoroughly investigate the distribution of Tajima's D values across different classes of sequences, I classified the entire genome into seven different classes of sequences (**Figure 26A**): introns, intergenic (5'upstream, 3'downstream, 5'UTR and 3'UTR) and coding (synonymous and non-synonymous). Since I discarded overlapping sequences and sequences with many ambiguously called bases, I did not include all of the genes in the genome (see Materials and Methods). The mean Tajima's D values were positive and very close to zero for all of the sequence classes except for synonymous sites for which the distribution was more positively skewed compared to the other classes (**Table 19**).

Table 19. Summary statistics for coding and non-coding sequences.

	Tajima's D					Positive Selection	
	Num. Seq.	**D***	**D*<0 (5th P)**	**D*>0 (95th P)**	**Mean Tajima's D**	**Num. Seq.**	**ω or ζ > 1**
Syn	11,860	872	331 (-1.31)	541 (1.44)	0.102		
Non-Syn	11,860	812	309 (-1.31)	503 (1.44)	0.019 (**)	11,995	224
3' Down	5,706	476	204 (-1.44)	272 (1.60)	0.063 (ns)	5,693	668
3' UTR	9,652	537	221 (-1.31)	316 (1.44)	0.065 (*)	9,648	613
5' Up	7,949	715	370 (-1.31)	345 (1.60)	0.080 (ns)	7,944	728
5'UTR	10,733	611	329 (-1.05)	282 (1.44)	0.052 (**)	10,724	456
Introns	8,893	741	388 (-1.05)	353 (1.44)	0.059 (*)	8,742	457

Note: Num. Seq.: Total number sequences analyzed in each class. D*: Number of sequences with Tajima's D < 5th P (percentile) or D > 95th P (percentile). D*<0: Number of sequences with Tajima's D < 5th percentile. D*>0: Number of sequences with Tajima's D > 95th percentile. Mean Tajima's D: Symbols between brackets indicate significantly differences at Wilcoxon rank-sum test versus *synonymous* (**: p < 1e-10, *: p < 1e-3, ns: not significant). Num. Seq. (Positive Selection): Total number of sequences analyzed. ω or ζ > 1: Sequences under positive selection (PS) at each class: Coding sequences were classified under PS when $p < 0.05$ at LRT (M0vsM3) and $p < 0.05$ at LRT (M1avsM2a) or LRT (M7vsM8). Non-coding sequences were classified under PS when any of the LRTs (NMvs2CM or NMvs3CM) showed a $p < 0.05$.

Reduced levels of polymorphism in non-coding and non-synonymous sites compared with the putatively neutrally evolving sites (synonymous) indicates that many of these sequences could be functionally constrained (Andolfatto 2005). The distributions of polymorphism frequencies at non-synonymous and 5'UTR sites are the most skewed towards rare frequencies relative to synonymous polymorphisms (Wilcoxon rank-sum test versus synonymous sites: Z = 7.98, p = 1.5e-15 and Z = 7.28, p = 3.2e-13, respectively). The distribution of

polymorphism frequencies in intron and 3'UTR classes showed lower but still significant differences relative to the distribution of Tajima's D values at synonymous sites (Z= 5.72, p = 1e-8 and Z = 3.78, p = 1.5e-4, respectively). However, neither 3'downstream nor 5'upstream regions showed differences in the synonymous distribution of Tajima's D values (Z = 1.66, p = 0.09 and Z = 0.73, p = 0.46, respectively). An excess of negative Tajima's D values in non-synonymous sites compared to synonymous sites is expected since most non-synonymous polymorphisms will be kept at lower frequencies due to functional constrains. Additionally, our results indicate that polymorphisms in introns and non-coding sequences in the immediate neighborhood of CDSs (UTRs) are more constrained on average, compared to polymorphisms at synonymous sites and at non-coding sequences further away from CDSs. By analyzing the distribution of Tajima's D values for each class of sequences (**Figure 26B**), I found that 3'downstream and 5'upstream regions show an extremely similar pattern of distribution, with values spread uniformly across the range, symmetrically distributed (Fisher's skewness coefficient, g_1 = 0.06 in both cases) and without extreme outliers, indicating that there are almost the same number sequences with excesses of both low and high frequency polymorphisms, and a lack of extreme values that deviate from the expected under neutrality. In contrast, 3'UTR and Intron sequences showed slightly positively skewed distributions of D values (Fisher's skewness coefficient, g_1 = 0.15 and g_1 = 0.23, respectively) with a concentration of values located near to zero or slightly positive (median = Q1), a greater dispersion of positive values and most extreme values corresponding to an excess of high frequency polymorphisms. 5'UTR values showed a peaked, but also a positively skewed distribution (Fisher's skewness coefficient, g_1 = 0.26), with at least 50% of the middle data concentrated around zero (median = Q1 = Q3), which indicate that most 5'UTR sequences presented none segregating sites. Overall, with the exception of 3'downstream and 5'upstream, most distributions showed many extreme positive and negative outliers. I analyzed functions and transcriptional profiles of genes found in the regions with extreme Tajima's D values at each distribution by selecting outliers in the lowest and/or the highest 5% of the distribution.

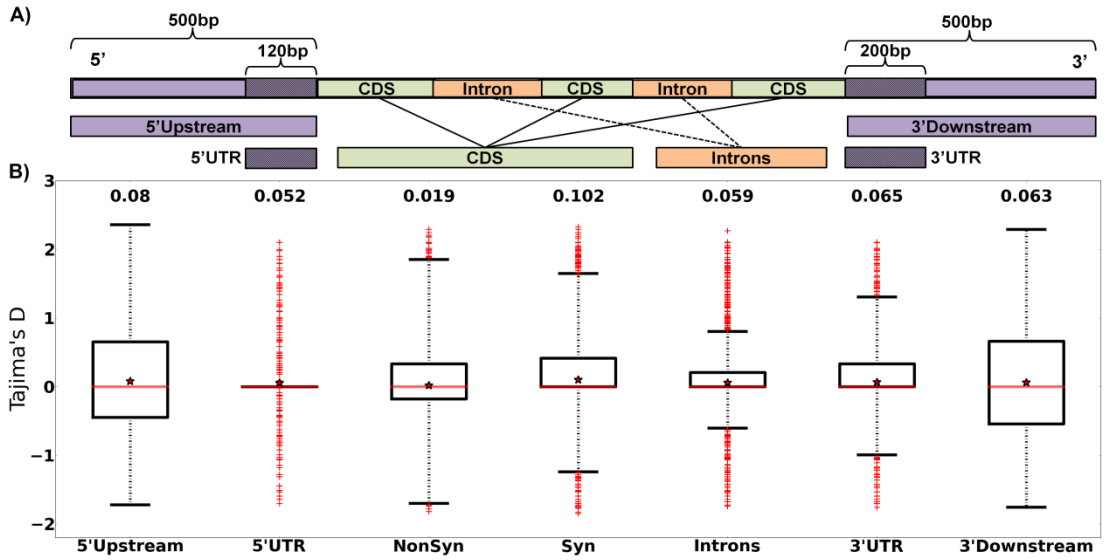

Figure 26. Distribution of Tajima's D values for each class of sequences. **A)** Typical eukaryotic protein-coding gene and sequence classes analyzed in the present study. **B)** Boxplots showing the distribution of Tajima's D values in each class of sequences. Values inside each box correspond to the middle 50% of the data (between the 25th (Q1) and 75th (Q3) percentiles) and the red line within the box represents the median. The ends of the vertical dotted lines (whiskers) at the top and bottom of each box indicate the maximum and minimum limits to consider outliers according to the Inter Quartile Range (IQR = Q3-Q1). Whiskers lengths were calculated as Q3+3*IQR (upper) and Q1-3*IQR (lower). Values behind the lines (red crosses) are extreme outliers. Red stars and values at the top of the boxplots indicate mean Tajima's D value for each class of sequence.

I separately analyzed two kinds of Tajima's D outliers representing sequences with unusual patterns of polymorphisms relative to the rest of the sequences at the same class: those showing extreme negative or extreme positive D (**Table 19**). The functional category enrichment analysis revealed that many non-coding sequences showing extreme D values belonged to genes related to pathogenicity (**Figure 27**, **Supplementary Table S5** and **Supplementary Table S6**). Such is the case of genes coding for secreted proteins, which were significantly enriched with both 5'UTR and 3'UTR sequences showing an excess of high frequency polymorphisms, or for genes upregulated during infection, significantly enriched with 5'UTR and intronic regions showing different patterns of polymorphisms relative to the expected for the class. I did not conduct an evaluation of putative false positives, and hence I don't expect all outliers to have experienced selective pressures, it is not possible to conclude about the strength of selection acting at each class. In contrast, the goal of this test was to explore whether or not non-coding sequences in proximity to genes involved in pathogenicity showed different patterns of polymorphisms

compared to other genes. Given that I expect the same rate of false positives in both kind of genes, our results are unlikely to be consequence of an excess of false positive outliers identified as extreme observations (Biswas & Akey 2006; Kelley et al. 2006).

4.3.3. Positive selection in coding and non-coding sequences

The use of maximum likelihood (ML) to estimate parameters under different evolutionary models provides a powerful and flexible framework for the hypotheses testing concerning the evolutionary process (Anisimova et al. 2001; Yang 2006). In order to investigate patterns of positive selection acting at both CDSs and non-coding sequences I applied different models of evolution aimed to compare the nucleotide substitutions rate at the region of interest with neutral expectation.

For CDSs, I fitted six Markov codon models of substitution implemented in PAML *v4* (Yang 2007) to test different hypothesis regarding the estimation of the nonsynonymous to synonymous rate ratio dN/dS (also known as ω). I used likelihood ratio tests (LRTs) for positive selection on the protein level, which compare log likelihood values obtained from each model. I performed three LRTs (M0vsM3, M1avsM2a and M7vsM8) and classified coding sequences as evolving under positive selection (PS) when the LRT comparing M0 vs M3 showed heterogeneity of ω among sites and one of the other LRTs (M1a vsM2a or M7 vs M8) showed evidence for PS (all with $p < 0.05$; **Table 19**). I identified 1.86% (224 out of 11,995) CDSs under positive selection. For most of them (205) all three LRTs were significant. The functional categories enrichment tests showed that many classes of genes previously described as evolving under PS in pathogenic fungi were also significantly enriched in our set of CDSs under PS. Such genes mainly encode for secondary metabolites, secreted proteins that likely interact with host molecules acting as pathogen effectors and putative virulence factors (**Figure 27** and **Supplementary Table S5**). Additionally, an in-depth analysis of GO categories enriched with CDSs under PS, showed genes involved in the binding of vitamins and amino acids, in the biosynthesis of polyketides and fatty acids and genes controlling methylation (**Supplementary Table S6**).

p (FDR) < 0.01
p (FDR) < 0.05
Not Significant

Functional Categories	TEST	Coding Syn	Coding NoS	3' Downstream	3' UTR	5' Upstream	5' UTR	Introns
CAZymes	D*	39	35	31	32	41	35	40
	D*<0	16	12	16	11	23	17	20
	D*>0	23	23	15	21	18	18	20
	PS	9		37	31	30	22	21
Cytochrome P450	D*	17	17	9	3	11	4	16
	D*<0	6	6	4	3	5	2	8
	D*>0	11	11	5	0	6	2	8
	PS	4		8	9	8	5	6
Genus spec. Effectors	D*	8	14	8	8	13	17	10
	D*<0	2	5	3	2	4	7	5
	D*>0	6	9	5	6	9	10	5
	PS	6		11	7	11	7	3
Secondary Metabolism	D*	30	22	15	19	18	22	29
	D*<0	15	10	7	9	11	11	12
	D*>0	15	12	8	10	7	11	17
	PS	21		23	25	18	9	18
Secreted Proteins	D*	108	116	64	79	121	114	95
	D*<0	41	42	28	25	64	55	54
	D*>0	67	74	36	54	57	59	41
	PS	44		97	83	113	56	51
Secreted Proteases	D*	7	10	2	6	8	4	5
	D*<0	1	1	0	1	3	2	3
	D*>0	6	9	2	5	5	2	2
	PS	3		11	11	9	3	4
Transcription Factors	D*	33	51	24	33	34	22	44
	D*<0	7	19	9	14	26	16	26
	D*>0	26	32	15	19	8	6	18
	PS	16		44	34	33	20	37
Transporters	D*	65	35	20	24	43	42	70
	D*<0	20	13	8	8	22	19	34
	D*>0	45	22	12	16	21	23	36
	PS	20		54	41	46	32	37
Virulence Factors	D*	135	101	62	68	95	80	129
	D*<0	63	46	33	37	60	36	68
	D*>0	72	55	29	31	35	44	61
	PS	39		95	102	100	54	65
Expresion Categories								
Biotrophic/PA	D*	32	32	17	20	46	31	44
	D*<0	11	10	7	11	23	14	24
	D*>0	21	22	10	9	23	17	20
	PS	9		33	34	28	12	16
Necrotrophic/Biotrophic	D*	50	52	26	34	43	41	52
	D*<0	17	15	13	16	19	18	25
	D*>0	33	37	13	18	24	23	27
	PS	16		46	38	45	25	35
Necrotrophic/PA	D*	68	76	43	47	71	65	93
	D*<0	22	26	19	23	34	29	49
	D*>0	46	50	24	24	37	36	44
	PS	26		69	70	68	34	52

Figure 27. Enrichment of putative non-neutrally evolving sequences at different functional gene categories related with pathogenicity. Numbers indicate sequences for each class and gene category resulted for each test. Tests: D* (D<5th percentile or D>95th percentile), D*<0 (D<5th percentile), D*>0 (D > 95th percentile) and PS (sequences under positive selection according to LRTs results). Background colors indicate significance of the Fisher's exact test for enrichment after correction for multiple comparisons by the false discovery rate (FDR).

For non-coding sequences, I applied combined DNA and codon models developed by (Wong & Nielsen 2004) to compare the rate of nucleotide substitutions in the non-coding regions with the rate of synonymous substitutions in the nearby coding regions, measured by parameter ζ. However, instead just using the synonymous substitutions rate for the nearest coding sequence, I also used the synonymous substitutions rate of the pooled CDSs from the upstream and downstream. With this approach I expected to avoid, at least partially, the bias in synonymous substitutions rate with respect to individual gene history. Non-coding sequences were classified under positive selection when any of the two LRTs (comparing models NM vs 2CM or NM vs 3CM) showed statistical differences supporting the model that allow for $\zeta > 1$ (**Table 19**). The 3'downstream sequences showed the largest amount of sequences putatively under PS (11.73%), while 5'UTR were the least affected by PS (4.25%). Despite our conservative procedure (from calling SNPs to the assignation of neutral sites, passing through the annotation of non-coding sequences), it is difficult to estimate how much of non-coding sequences showing PS are actually true or false positives since selection at synonymous sites, a higher mutation rate or a relaxation of selective constrains may also contribute to the signal detected by the ML test. In order to further investigate whether selection at synonymous sites were influencing our results I analyzed the intersection of genes showing PS at the five non-coding classes. If selection at synonymous sites were biasing the detection of PS at the non-coding regions, I would expect that most non-coding sequence from the same gene show PS. I found only two genes with PS in all five classes, suggesting that selection in synonymous sites is not biasing our results (**Figure 28**). An additional outcome from this analysis is the high overlap between 3'UTR and 3'downstream (183) and between 5'UTR and 5'upstream (135), which indicates that many PS signals at the more distant intergenic regions are actually coming from its contained UTR region.

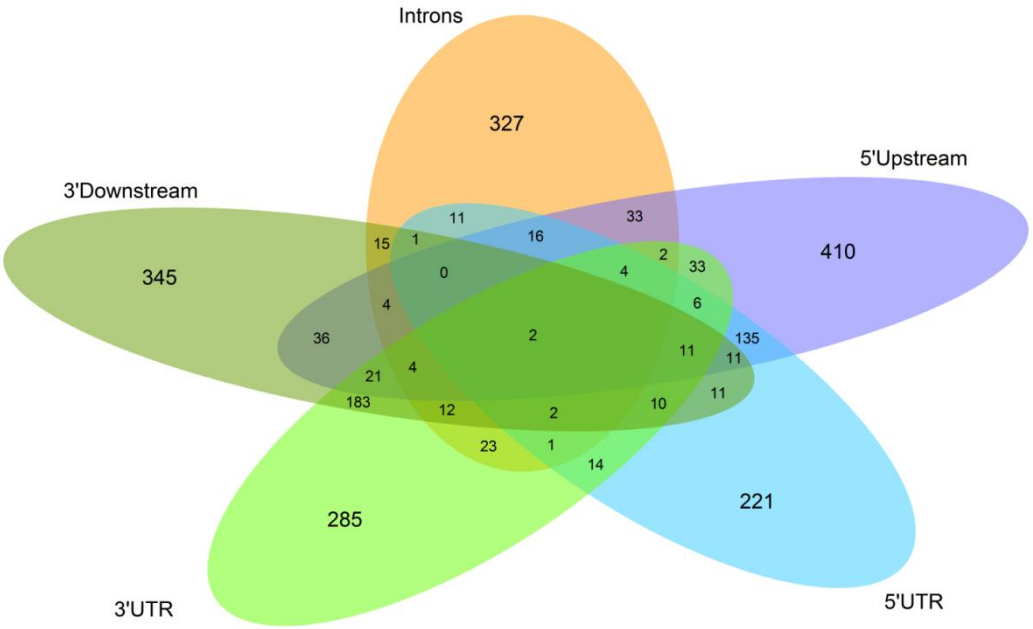

Figure 28. Five sets Venn's diagram showing gene intersections between different classes of non-coding sequences under positive selection.

4.4. Discussion

By analyzing whole genomes of eight field isolates of *C. graminicola* collected from different regions of the world, I found that selection differentially affects patterns of polymorphism in coding and non-coding sequences of pathogenicity-related genes. Our first approach to study the frequency and distribution of polymorphisms was to perform a sliding-windows analysis of Tajima's D values at the whole-genome level (**Figure 25**). This method is based on the principle that natural selection alters the site frequency spectrum (but see Zhu and Bustamante (2005), which can be detected based on deviations from the null model. Although Tajima's D test is not aimed to distinguish between different forms of natural selection, it will tend to be negative under selective sweeps and positive under balancing selection (excluding demographic effects). For instance, genes coding for products involved in basic metabolic processes (e.g. respiration, DNA synthesis, etc.) or belonging to cellular components are likely to show more negative D values than others due to purifying selection. When I studied genes located in windows with extreme Tajima's D (EWs) I did not find enrichments for any particular gene category. I hypothesize that such results might be consequence of considering that all polymorphisms in the window are equally affected by natural selection. In fact, functional polymorphisms (Albert 2011) (i.e. polymorphisms that alter the gene function or its regulation) are likely to show a different frequency spectrum than non-functional polymorphism (i.e. synonymous polymorphisms in coding regions) since they evolve differently in response to the force of selection (Andolfatto 2005).

I divided the whole genome into different classes of sequences that I expected to vary in strength and type of selection and individually estimated Tajima's D for all sequences of each class. Usually, synonymous sites are considered to evolve neutrally, therefore Tajima's D = 0 is expected. I obtained a positive value in our study (0.102), probably as consequence of sampling from many populations. However, such a demographic effect will equally affect all regions in the genome (Tajima 1989) and consequently differences between regions are expected to reflect the effects of selection (Bickel et al. 2013). I found that the distributions of polymorphisms in UTRs, introns and non-synonymous sites are skewed

toward rare frequencies relative to synonymous polymorphisms showing less positive Tajima's D values, whereas 3'downstream and 5'upstream did not show statistical differences relative to D values at synonymous sites (**Table 19** and **Figure 26B**). Reduced levels of polymorphism in these regions relative to synonymous sites suggests that, on average, they are functionally constrained and thus under purifying selection (Andolfatto 2005). This is expected for non-synonymous polymorphisms, since most of them will be deleterious and rapidly removed from the population. Functional constrains at UTRs and intronic sequences have been previously reported in higher eukaryotes such as *Drosophila* (Andolfatto 2005), murids (Gaffney & Keightley 2006) and *Acyrthosiphon pisum* (pea aphid) (Bickel et al. 2013). In fungi, some evidence for functional constrains in non-coding sequences were found in yeast (Gasch et al. 2004; Doniger et al. 2008) and in the filamentous pathogenic fungus *Pyrenophora* (Ellwood et al. 2012). However, even though non-coding DNA near to coding sequences are expected to be involved in the control of transcription and translation, most of them have no known function and selective constraints acting on these regions can sometimes be difficult to interpret (Zhen & Andolfatto 2012). On the other hand, previous work found a correlation between non-coding sites under selection and the function of the adjacent gene (Gaffney & Keightley 2006). By analyzing the outlying Tajima's D values of each sequence class I found that many non-coding sequences showing unusual patterns of polymorphism belong to genes related with pathogenicity. For instance, 5'UTR sequences showing an excess of high frequency polymorphisms (D*>0) were significantly associated with genes encoding for secreted proteins (putative effectors) and to genes upregulated during all of the three most important phases of infection of *Cg* in maize (**Figure 27**, **Supplementary Table S5** and **Supplementary Table S6**). Balancing selection has been largely proposed as one of the main mechanisms responsible for the maintenance of polymorphisms in populations (Kojima 1971). However these patterns, consistent with the red queen model of evolution, have been mainly described for sequences encoding effector proteins in the pathogen and the corresponding resistance genes in the plant (Dodds et al. 2006; Win et al. 2007). Our results suggest that evolution of 5'UTR regions, likely to be involved in the regulation of the adjacent gene, could also be driven by a red queen-like

model, in which variability at the regulatory sequence would confer flexibility for gene expression. Organisms carrying a monomorphic pathogenicity gene, but with polymorphic regulatory sequences could show very different patterns of virulence, depending on its ability to control the amount and timing of transcription. In contrast, 5'UTR from genes involved in the most basic functions such as the production of cellular components or organelle organization were underrepresented in the set with D*>0 (**Supplementary Table S6**), probably as consequence of purifying selection acting on the regulatory regions from these genes.

Genes upregulated during infection also showed enrichment of intronic sequences with an excess of both high and low frequency polymorphisms. There is ample evidence about the influence of introns on eukaryotic gene expression (MacKenzie & Quinn 1999; Le Hir et al. 2003; Gaffney & Keightley 2006; Albert 2011). Our results reflect the fact that some introns are expected to have a greater proportion of regulatory sequences than others, resulting in different strength and/or types of selection acting on them (Gazave et al. 2007). In addition, introns with D*>0 were specially enriched in transporter genes (**Supplementary Table S6**) which have been previously reported to change its transcriptional regulation due to variations in the intronic regions (Greenwood & Kelsoe 2003; Hranilovic et al. 2004). Our results show that putative regulatory non-coding sequences from many genes involved in pathogenicity are subject to different selective pressures compared to non-coding sequences from other genes.

Whereas neutrality tests (like Tajima's D) have been demonstrated to be powerful enough to identify unusual patterns of polymorphism at the genome level, sometimes they do not easily allow conclusions about the adaptive processes acting at the molecular level (Yang & Bielawski 2000). I applied LRTs to investigate patterns of positive selection acting on both, coding and non-coding sequences. One of the main concerns when looking for non-coding sequences under PS is the choice of reliable "neutrally evolving" references sites to compare with (Zhen & Andolfatto 2012). Synonymous sites and intronic DNA (Haygood et al. 2007) are the two most widely used reference sites. However, an increasing amount of evidence suggests that intronic sequences are enriched

with regulatory sequences or encoded for non-coding RNAs (Qu and Adelson 2012a; Qu and Adelson 2012b), which would lead these sequences to be under purifying selection, resulting in an overestimation of positive selection in the non-coding region. As I previously showed, a large number of intronic sequences from genes that are upregulated during infection appear to be evolving differently to other introns (**Figure 27**). For that reason, I used synonymous sites as the neutral reference. However, synonymous substitutions are not exempt from selection either. Both positive and negative selection at the synonymous sites have been previously described in higher eukaryotes (Resch et al. 2007; Kosiol & Anisimova 2012; Lawrie et al. 2013) and selection on codon usage has been demonstrated to take place at the population level in the fungal model organism *Neurospora crassa* (Whittle et al. 2012). In addition, adaptation of synonymous codon usage seems to be related to the gene's function (Liu et al. 2005; Basak & Ghosh 2006; Hudson et al. 2011). I identified a significant enrichment of positive D values at synonymous sites in genes coding for transporters and enrichment of negative D values in genes coding for putative virulence factors (**Figure 27**). An in-depth analysis of genes with negative D at synonymous sites showed that most of them are ribosome-related (**Supplementary Table S6**), indicating a putative codon usage bias in that class of genes, in line with the previously described results for *Saccharomyces cerevisiae* (Hudson et al. 2011). In an attempt to eliminate the bias regarding possible selection at the synonymous sites, I compared non-coding substitution rates for each class of sequence with the pooled synonymous substitutions rate at the coding regions of the three neighboring genes. Even through such conservative approach, I identified many non-coding regions in all classes showing evidences of PS (**Table 19**). However, for 5'upstream, 5'UTR and intronic sequences under PS, I did not find enrichment of any GO or pathogenicity related gene category. It is possible that the highly conservative use of putative neutral synonymous substitution rates did not allow us to identify non-coding regions under weak PS or those with just a few sites under selection. Additionally, by removing sequences with many ambiguously called bases or without well annotated start and stop codons in all of the isolates, I was actually analyzing just the most conserved sequences, since the discarded ones are likely to be the consequence of a low number of sequence reads mapping to

them due to too many mismatches. 3'downstream and 3'UTR sequences under PS were significantly enriched for gene categories related with pathogenicity. For instance, genes coding for transporters showed a significant enrichment of the 3'downstream region under PS, while putative virulence factors and two out of the three categories of genes upregulated during infection showed enrichment of 3'UTR under PS. These findings take relevance in light of the increasing evidence of the active contribution of 3'UTRs in the regulation of gene expression (Thon et al. 2002b; Mazumder et al. 2003; Merritt et al. 2008), suggesting that such regions could be also playing an important role in adaptations of virulence forced by the constantly changing host and environment.

At the CDS level, I found that 1.86% of the genes have evidence of PS. The PS genes are enriched with the functional categories secondary metabolism, secreted proteins and putative virulence factors. Interestingly, putative genus-specific effectors were not enriched for CDSs under PS. Previous studies have also shown that PS genes are not always enriched with putative effectors (Stukenbrock et al. 2010; Pedersen et al. 2012) although it should be noted that the definition of effector proteins varies considerably among authors. Given the difficulty associated with the definition of effector proteins, our results could be consequence of a misclassification of effectors. This claim is partially supported by many studies that show genes of unknown function are often under PS (Y.-D. Li et al. 2009; Stukenbrock et al. 2010; Neafsey et al. 2010; Aguileta et al. 2012). Several authors employ a broader definition and consider that any secreted protein is a potential effector (see Doehlemann and Hemetsberger, 2013). Secreted proteins are enriched among the PS genes, which is consistent with our original expectations.

I identified 52 coding sequences under PS with no known function, of which just five were identified as genus specific putative effectors, while another 17 are predicted to be secreted. Such proteins are very likely to be part of the effector repertory and therefore are excellent candidates for functional validation (**Table 20**). One of these genes (*GLRG_04079*) has been functionally analyzed and is involved in pathogenicity (WA Vargas et al. unpublished observations). The, virulence factors category includes genes that are homologous to genes that

have been shown to have some role in virulence in other fungi. However, this category may include genes involved in multiple functions. Out of 39 virulence factors identified in the present study under PS, 17 were related to the production of secondary metabolites, 8 encode transporters, 3 carbohydrate active enzymes, 2 transcription factors, 1 secreted protease and 2 that participate in mycelium development. Secondary metabolism genes deserve special consideration in our study since *Colletotrichum* species, like most necrotrophs and hemibiotrophs, produce a broad range of secondary metabolites with important functions such as antibiosis, protection from stress and pathogenicity (Keller et al. 2005; Spanu & Kämper 2010; O'Connell et al. 2012). I found that the set of genes under PS was enriched with members of this category (**Figure 27**), and especially with polyketide biosynthesis (PKS) related genes (**Supplementary Table S6**). Secondary metabolism gene clusters are suspected to have undergone expansions in *Colletotrichum* species, which could shed light into the reason for finding many of them under positive selection. In fact, I found a new case of PS after gene duplication (Zhang et al. 1998; Emes & Yang 2008) for gene *GLRG_03511* (under PS) and its paralog *GLRG_05714*. This gene, as well other secondary metabolism related genes under PS, may represent a source for the production of new bioactive molecules, with implications to both phytopathology and biochemistry. Overall, genes identified under PS within the coding region belong to categories previously identified in other plant pathogenic fungi as evolving adaptively (Aguileta et al. 2010; Stukenbrock et al. 2011; Aguileta et al. 2012; Gibbons et al. 2012), supporting the hypothesis that they are involved in the evolutionary arms-race or in the adaptation to new environments. Some of the most interesting genes showing PS and its characteristics are listed in **Table 20** providing a valuable resource for future functional characterization.

Table 20. Interesting genes under positive selection at the coding sequence.

Gene ID	Annotation	Exp.Cg	Secr.	PhiBase	Phenotype of mutant	Pathogen species	Host
GLRG_00054	ABC transporter	S/S/S	NO	PHI:202	Reduced virulence	*Botrytis cinerea*	Grape
GLRG_00247	ABC-2 type transporter	D/D/S	NO	PHI:258	Reduced virulence	*Gibberella pulicaris*	Potato
GLRG_00264	Hypothetical protein	S/D/S	YES	-	-	-	-
GLRG_00468	ABC transporter	S/S/S	NO	PHI:1018	Loss of pathogenicity	*Magnaporthe oryzae*	Rice
GLRG_00469	AMP-binding enzyme	S/S/S	NO	PHI:160	Loss of pathogenicity	*Alternaria alternata*	Apple
GLRG_00513	ABC-2 type transporter	S/S/S	NO	PHI:258	Reduced virulence	*Gibberella pulicaris*	Potato
GLRG_00514	ABC transporter	S/S/S	NO	PHI:267	Reduced virulence	*Candida albicans*	Mouse
GLRG_00920	Amino acid adenylation	S/S/S	NO	PHI:12	Loss of pathogenicity	*Cochliobolus carbonum*	Maize
GLRG_01586	Hypothetical protein	S/U/S	YES	-	-	-	-
GLRG_01804	Hypothetical protein	U/U/S	YES	-	-	-	-
GLRG_01845	Transcription factor	S/S/S	NO	PHI:169	Loss of pathogenicity	*Colletotrichum lindemuthianum*	Bean
GLRG_01860	Beta-ketoacyl synthase	S/S/S	NO	PHI:55	Reduced virulence	*Cochliobolus heterostrophus*	Maize
GLRG_02650	Adhesin protein Mad1	S/S/S	YES	-	-	-	-
GLRG_02963	Hypothetical protein	S/S/S	NO	PHI:404	Reduced virulence	*Magnaporthe oryzae*	Rice
GLRG_03150	Hypothetical protein	S/S/S	NO	PHI:211	Reduced virulence	*Candida albicans*	Mouse
GLRG_03507	Beta-ketoacyl synthase	S/S/S	NO	PHI:101	Reduced virulence	*Aspergillus fumigatus*	Mouse
GLRG_03511	Beta-ketoacyl synthase	S/S/S	NO	PHI:55	Reduced virulence	*Cochliobolus heterostrophus*	Maize
GLRG_04079	Hypothetical protein	S/D/D	YES	-	-	-	-
GLRG_04142	Hypothetical protein	U/U/S	YES	-	-	-	-
GLRG_05009	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_05053	Hypothetical protein	S/S/S	NO	PHI:864	Loss of pathogenicity	*Aspergillus fumigatus*	Mouse
GLRG_05268	Methyltransferase	U/U/S	NO	PHI:482	Reduced virulence	*Aspergillus fumigatus*	Mouse
GLRG_05919	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_05958	Glycosyl hydrolase	S/S/S	YES	PHI:1071	Loss of pathogenicity	*Ustilago maydis*	Maize
GLRG_06092	Mannosyltransferase	S/D/D	NO	PHI:455	Loss of pathogenicity	*Cryptococcus neoformans*	Mouse
GLRG_06331	WD domain	S/S/S	NO	PHI:211	Reduced virulence	*Candida albicans*	Mouse

GLRG_06371	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_06485	Hypothetical protein	S/D/S	YES	-	-	-	-
GLRG_06732	Cytochrome P450	S/S/S	NO	PHI:438	Reduced virulence	*Botrytis cinerea*	Bean
GLRG_06861	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_07140	Hypothetical protein	D/D/S	YES	-	-	-	-
GLRG_07145	Cytochrome P450	S/S/S	NO	PHI:438	Reduced virulence	*Botrytis cinerea*	Bean
GLRG_07254	Alpha-L-rhamnosidase	S/S/S	YES	-	-	-	-
GLRG_07434	Beta-ketoacyl synthase	S/U/U	NO	PHI:325	Effector	*Magnaporthe oryzae*	Rice
GLRG_07527	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_07678	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_07748	Hypothetical protein	S/D/S	YES	-	-	-	-
GLRG_07825	Sugar lactone oxidase	S/S/S	NO	PHI:197	Reduced virulence	*Candida albicans*	Mouse
GLRG_08161	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_08505	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_08566	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_08615	MFS Transporter	S/S/S	NO	PHI:141	Reduced virulence	*Cercospora kikuchii*	Soybean
GLRG_08620	Beta-ketoacyl synthase	S/S/S	NO	PHI:433	Reduced virulence	*Cercospora nicotianae*	Tobacco
GLRG_08878	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_08901	Hypothetical protein	S/S/S	YES	-	-	-	-
GLRG_09110	Hypothetical protein	S/D/D	YES	PHI:256	Reduced virulence	*Magnaporthe oryzae*	Rice
GLRG_09221	DEAD/DEAH box helicase	S/S/S	NO	PHI:423	Loss of pathogenicity	*Claviceps purpurea*	Ergot
GLRG_09382	Tannase/feruloyl esterase	S/S/S	YES	-	-	-	-
GLRG_09394	Hypothetical protein	U/U/S	YES	-	-	-	-
GLRG_09842	AMP-binding	S/S/S	NO	PHI:325	Effector	*Magnaporthe oryzae*	Rice
GLRG_10257	MFS Transporter	S/S/S	NO	PHI:511	Reduced virulence	*Candida albicans*	Mouse
GLRG_10317	Beta-ketoacyl synthase	S/S/S	NO	PHI:55	Reduced virulence	*Cochliobolus heterostrophus*	maize
GLRG_10367	AMP-binding	S/S/S	NO	PHI:325	Effector	*Magnaporthe oryzae*	Rice
GLRG_10457	Hypothetical protein	S/D/D	YES	-	-	-	-

GLRG_11626	Beta-ketoacyl synthase	S/S/S	NO	PHI:325	Effector	*Magnaporthe oryzae*	Rice
GLRG_11821	Linoleate diol synthase	D/D/S	NO	PHI:496	Hypervirulence	*Aspergillus fumigatus*	Mouse
GLRG_11938	Hypothetical protein	S/D/D	YES	-	-	-	-

Note: Gene Id and Annotation according to the Broad Institute annotation. Expression in *C graminicola* indicate whether the gene is UP (U) regulated DOWN (D) regulated or STABLE (S) at each time point during infection on maize (Biotrophic/PA / Necrotrophic/PA / Necrotrophic/Biotrophic) according to O'Connell et al. (2012). PhiBase ID: Best match on PhiBase (Winnenburg et al. 2006)

Transcriptional profiling of pathogens during the infection process has been useful to identify genes involved in host invasion and nutrient acquisition. However, similar to previous studies (Aguileta et al. 2012), I did not find an enrichment of coding sequences under PS that are also upregulated during infection. A similar pattern was also found in mammals, in which coding sequences under PS showed a reduced expression level at different analyzed tissues (Kosiol et al. 2008) suggesting, according to the authors, a relationship between expression patterns and the likelihood of positive selection at the coding region. Alternatively, I discovered that many upregulated genes during infection show evidence for non-neutral evolution at the non-coding regions that are likely to have roles in the transcriptional regulation. King and Wilson's hypothesis concerning the predominant role of regulatory mutations in organismal evolution (King & Wilson 1975) has lately received great support, revealing, for example, the leading role of gene expression on the local adaptations in humans (Fraser 2013). Following this line, I hypothesize that even though adaptations in the coding sequences are important for proteins expected to interact directly with the host's molecules, changes in regulatory sequences may drive the evolution of many other characters involved in the virulence of *Cg*. However, the role of regulatory sequence evolution remains unclear until information about genome-wide variation in gene expression, sequence polymorphisms and phenotypic variability become available.

The present study represents the first report of selective pressure acting on both coding and non-coding DNA at the whole genome level, for an agronomically important phytopathogenic filamentous fungus. I found evidence that both protein-coding and non-coding DNA sequences of pathogenicity-related genes are under differential selective pressures compared to other genes. Furthermore, I found that genes coding for proteins expected to interact directly with the host's molecules (such as effector proteins and secondary metabolites) show evidence of positive selection acting on the coding sequence, whereas genes upregulated during infection are enriched with UTRs and intronic DNA sequences under selective sweeps, balancing and positive selection. Our findings contribute to our understanding of the evolutionary process at the molecular level and provide a valuable resource for the development of environmentally friendly strategies to control fungal diseases.

5. CHAPTER III

Annotation of fungal non-coding RNAs (ncRNAs)

5.1. Introduction

In recent years, the sequencing and annotation of hundreds of genomes have revealed that most organisms contain a lower number of protein coding genes than initially anticipated and that protein coding genes in fact only account for a small fraction of the genome of higher eukaryotes (Mattick 2001). However, for many organisms it has been discovered that a large proportion of the genome is actually transcribed (Carninci et al. 2005; Frith et al. 2005; Jochl et al. 2008; Arrial et al. 2009; Jacquier 2009; Mercer et al. 2009). These new data have led some researchers to propose that the complexity of eukaryote genomes cannot be exclusively explained by the classical view of the central dogma of molecular biology "DNA-RNA-protein", since much of the DNA that does not encode proteins is now known to encode various types of functional RNAs, known as non-protein-coding RNAs (ncRNAs) (Barry 2007; Gerstein et al. 2007; ENCODE Project Consortium et al. 2007).

Whether most ncRNAs are functional or not is still under debate (Carninci et al. 2005; Hüttenhofer et al. 2005; van Bakel et al. 2010), however multiple ncRNAs, either on their own or in complex with proteins, have already been described as fundamental for some important cellular functions (Hüttenhofer et al. 2002; Gottesman 2002; Eddy 2002; Storz et al. 2004; Mattick 2004; Costa 2005; Costa 2008; Taft et al. 2009; Huarte et al. 2010; Ørom et al. 2010). Some of the most important functions of ncRNAs are RNA processing and modification, transcriptional regulation, DNA replication, mRNA stability and translation, and protein secretion (**Figure 29**). Therefore it is not surprising that ncRNAs come in a huge variety of flavors. Some of the best-studied classes of short (<200 nt) ncRNAs include small nucleolar RNAs (snoRNAs), which are primarily involved in the chemical modification of other RNAs (Lafontaine & Tollervey 1998), microRNAs (miRNAs) which participates in the transcriptional and post-transcriptional regulation of gene expression (Chen & Rajewsky 2007), small interfering RNA (siRNA) involved in the RNA interference (RNAi) pathway among many other roles (Hamilton & Baulcombe 1999), small nuclear RNAs (snRNA) whose primary function is the processing of pre-mRNA in the nucleus (Kretzner et al. 1990) and the piwi-interacting RNAs (piRNAs) which form a RNA-protein complex through interactions with piwi proteins which

participates in epigenetic and post-transcriptional gene silencing of retrotransposons and other "genomic parasites" (Girard et al. 2006). On the other hand, long ncRNAs (>200 nt) have been largely proposed as crucial players in the transcriptional, post-transcriptional and epigenetic regulation (Mercer et al. 2009; Wilusz et al. 2009; Kung et al. 2013).

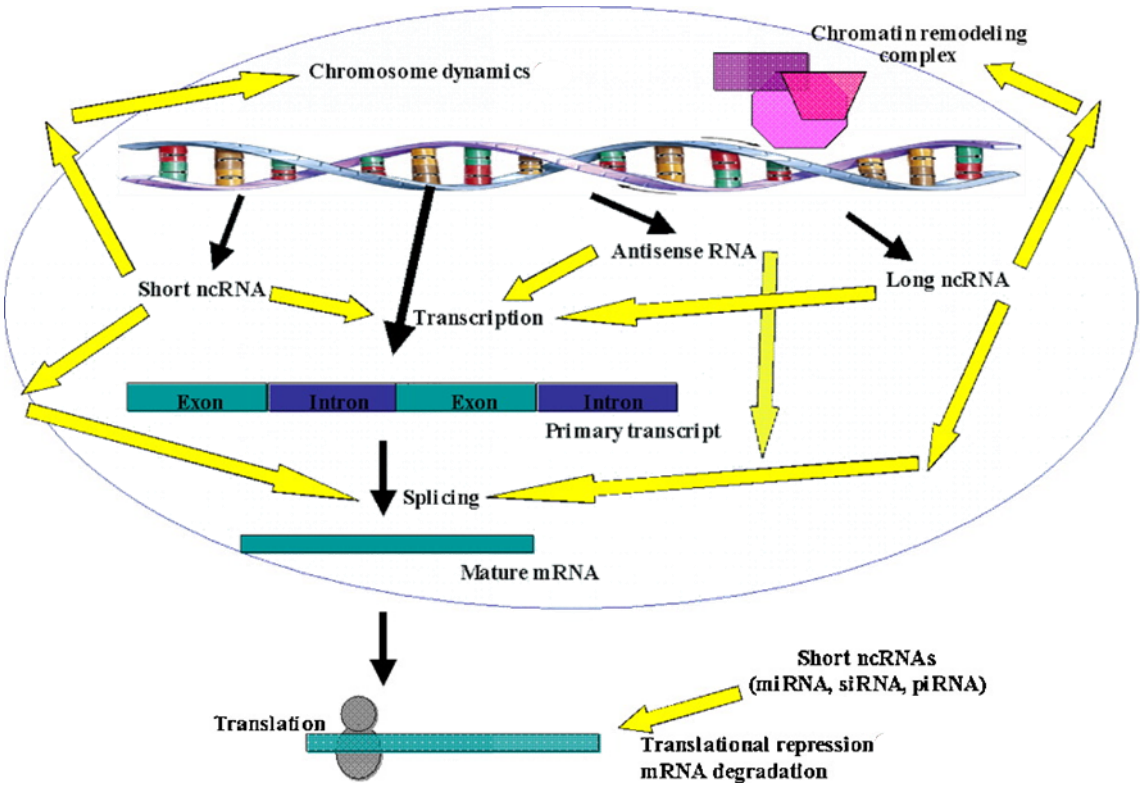

Figure 29. Main function of ncRNAs in eukaryotes. Taken from (Matrajt 2010).

Even though thousands of pervasively transcribed ncRNAs have been described in higher eukaryotes (Qu & Adelson 2012b), the information for simple eukaryotes such as fungi is very limited (Gowda et al. 2010). One of the first studies was performed by analyzing microarray expression data of *Saccharomyces cerevisiae*, revealing some possible regulatory mechanisms based on non-coding transcripts (Havilio et al. 2005). Comparative genomics of seven yeast species, providing stronger evidence for the presence of a large number of small ncRNA genes and structural motifs that overlap with known features such as coding sequences and UTRs, suggesting their roles in post-transcriptional regulation (Steigele et al. 2007). In *Aspergillus fumigatus*, an experimental screening of ncRNAs expressed under various growth conditions

and during specific developmental stages revealed 30 novel candidate ncRNAs from known ncRNA classes such as snRNAs and snoRNAs and other 15 candidates that could not be assigned to any known ncRNA class (Jochl et al. 2008). A new type of small interfering RNA (named qiRNAs) was firstly described in *Neurospora crassa* (Lee et al. 2009). The synthesis of qiRNAs is mainly induced by DNA damage, suggesting a role in DNA repair by inhibition of the protein translation. Additionally, the discovery of several pathways for the production of Dicer-dependent miRNAs and the Dicer-independent Piwi-interacting RNAs in *N. crassa* have also provided evidence of the existence of such kind of sequences in filamentous fungi (Lee et al. 2010). Small RNAs mapping to numerous nuclear and mitochondrial genomic features including repetitive elements were also found in *Magnaporthe oryzae*. Most of them accumulate in vegetative and specialized-infection tissues, suggesting a key role in the regulation of growth and development (Gowda et al. 2010; Nunes et al. 2011). Overall, these works provide evidence of the presence and the importance of ncRNAs in filamentous fungi; however it is likely that the vast majority of ncRNAs in fungi remain unknown.

Previously (**Chapter II**), I found evidence of natural selection acting on non-coding DNA sequences of the plant pathogenic, filamentous fungus *C. graminicola*. These sequences, located in the immediate neighborhood of coding regions, likely represent functionally important sequences involved in the organismal phenotype. I have proposed two mutually non-exclusive hypotheses regarding the nature of such sequences: 1) they represent *cis* and *trans* regulatory elements, controlling the transcription of genes and 2) they are transcribed as functional ncRNAs implicated in some of the multiple functions detailed above. The first hypothesis was addressed in **Chapter II**, with results suggesting that many traits involved in the host-pathogen interaction could potentially be determined by mutations in non-coding regulatory regions of the genome. Unfortunately, at the time of addressing the second hypothesis, there were not sufficient transcriptome sequences available for *C. graminicola* to conduct a whole genome analysis of ncRNAs. For that reason, I extended the exploration of putative ncRNAs to many other fungal species containing a much larger amount of publicly available transcript sequences.

To characterize ncRNAs at a genome scale in filamentous fungi, I adapted a computational pipeline initially applied to identify non-protein coding transcripts in bovines and developed by Dr. Zhipeng Qu and Prof. David Adelson at the Centre for Bioinformatics and Computational Genetics from The University of Adelaide (Qu & Adelson 2012a). I analyzed 2,127,338 public Expressed Sequence Tags (ESTs) from 42 fungal and 4 oomycetes species, representing most important fungal taxonomic groups. I identified a total of 8,251 putative ncRNAs, most of them (93%) previously un-annotated. Fungal ncRNAs are poorly conserved at the sequence level between analyzed species, and many of them seem to be involved in cell differentiation. The functional importance of putative fungal ncRNAs remains unknown, but their characterization could potentially open the possibility of exploring new strategies for the development of novel anti-fungal drugs by, for instance, targeting regulatory ncRNAs that control different aspects of fungal growth and virulence, which could have multiple implications in agriculture and human health.

5.2. Materials and Methods

5.2.1. Pipeline description

Fungal and oomycete ESTs were obtained from the dbEST of NCBI (Boguski et al. 1993). I extracted all available ESTs (as of September, 2011) for each of the 46 species. I used SeqClean[10] to identify contaminating sequences using the UniVec[11] database. I scanned EST sequences for repetitive sequences using RepeatMasker (Smit et al. 1996) with the Repbase v.16.08 library (Jurka et al. 2005) and by performing a WU-BLAST[12] search against a repetitive protein database (GB-TE) (Smith et al. 2007). Once repetitive and low quality sequences were masked, the remaining ESTs were clustered and assembled using the TGICL[13] software. Unigenes longer than 100 nt were then mapped to the fungal genomes using GMAP (Wu & Watanabe 2005). Sequences mapping to the genome (>95% identity and >90% coverage) were used as query sequences in both BLASTN searches of a database built with the protein-coding transcript sequences and BLASTX against the SwissProt database (Bairoch & Boeckmann 1994). Unigenes showing an e-value lower than 1e-5 in any of the searches were removed. The remaining sequences were evaluated with the Getorf program from the EMBOSS package (Rice et al. 2000). Sequences showing an open reading frame longer than 300nt were discarded. Finally, I evaluated the coding potential of the unigenes by means of the a support vector machine algorithm PORTRAIT (Arrial et al. 2009). This pipeline was individually run for each species by means of a Python script linking all of the programs.

5.2.2. Annotation of ncRNAs

Several ncRNA databases were used to annotate the ncRNAs. The miRNA database, miRBase release 17, which includes 16,772 entries representing hairpin precursor miRNAs, expressing 19,724 mature miRNA products, in 153

[10] SeqClean: http://compbio.dfci.harvard.edu/tgi/software/
[11] UniVec: http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/
[12] WU_BLAST has been renamed to AB_BLAST, which is no longer freely accessible. We used version 2 of WU_BLAST in the pipeline. (http://blast.wustl.edu/)
[13] TGICL: http://compbio.dfci.harvard.edu/tgi/software/

species was obtained from miRBase (http://www.mirbase.org/) (Kozomara & Griffiths-Jones 2011). Rfam10.1, which contains non-coding RNA genes, structured cis-regulatory elements and self-splicing RNAs was obtained from http://rfam.janelia.org/ (Griffiths-Jones 2005). The Stand-alone Rfam search was performed by a Perl script Rfam_scan.pl provided with Rfam. NONCODE3.0 was obtained from http://www.noncode.org/ (Bu et al. 2012). FRNAdb was obtained from http://www.ncrna.org/frnadb/. I built a database using the 46 ncRNAs identified in *Aspergillus fumigatus* (Jochl et al. 2008). I also used tRNAscan-SE to identify transfer RNA genes (Lowe & Eddy 1997). All searches against different databases were performed using BLASTN requiring 80% identity and 80% coverage of the query sequence.

5.2.3. Analysis of putative ncRNAs

I reconstruct a phylogenetic tree for the 46 species based on the single copy gene MS456, which corresponds to the gene Mcm7, a DNA replication licensing factor required for DNA replication initiation and cell proliferation (Moir et al. 1982; Kearsey & Labib 1998) and previously described as one of the best single-copy genes for fungal phylogenetic reconstructions (Aguileta et al. 2008). I identified orthologs of MS456 in each of the 46 species, aligned them using MUSCLE (Edgar 2004) and built the phylogeny using PhyML (Guindon et al. 2010).

Sequence conservation of the putative ncRNAs was assessed by mapping each of the 8,251 ncRNAs to each of the 46 genomes under study using GMAP (Wu & Watanabe 2005) with the option "cross-species". I considered an ncRNA as "conserved" when mapped to a genome of a different species with at least 20% coverage and 90% sequence identity.

In order to determine whether any EST library was enriched with transcript sequences classified as putative ncRNAs, I categorized all unique transcript sequences according to the EST library of origin. Unigenes consisting of just one EST, were assigned to the EST library from which it was obtained. For unigenes consisting of more than one EST, I determined whether the ESTs originated from different EST libraries, in which case the contig sequence was assigned to the library sourcing at least the 80% of the ESTs. When this

criterion was impossible to determine (e.g. contigs assembled with similar percentage of ESTs from different libraries), the transcript was excluded from the analysis. Once transcript sequences were assigned to each library, I performed a Chi-Squared test to examine differences between the expected number of transcript sequences for each library and the observed number of transcripts classified as putative ncRNAs. For those species showing a significant p-value ($p<0.01$) at the Chi-Squared test, I applied the Z-test for proportions to determine which libraries were enriched with transcripts classified as ncRNAs. P-values were corrected for each species using the Bonferroni method taking into account all the EST libraries analyzed for the species.

5.3. Results

5.3.1. Exploration of putative ncRNAs

I used a previously described pipeline (Qu & Adelson 2012a) to identify putative ncRNA transcripts from all publically available ESTs of 46 fungal and oomycete species. The criteria to select the species were: 1) Genome completely sequenced, 2) Availability of whole-genome annotation of the genome and 3) More than 10,000 ESTs in the NCBI - dbEST database (as of September, 2011). In addition, I also included some species that didn't meet criterion 3, but represented economical, agricultural or biologically important fungal species such as *Aspergillus oryzae*, *Colletotrichum graminicola*, *Colletotrichum higginsianum*, *Fusarium oxysporum* and *Paxillus involutus* (**Table 21**).

Overall, 2,127,338 ESTs were used for the exploration of putative ncRNAs. The pipeline was individually applied to each species (**Figure 30** and **Table 21**). After quality control, repeat filtering and EST assembly, I identified 553,510 unique transcripts. Transcripts for each species were mapped to the corresponding genome; resulting in 376,923 transcripts sequences effectively mapped overall the species. Unmapped transcripts were discarded. Mapped transcripts were analyzed looking for those that likely encode for proteins. In this way, sequences showing similarities with protein coding transcripts of the species or with any protein in the SwissProt database were discarded (e-value < 1e-5). In addition I also remove ORF containing transcripts (>100 codons) and transcripts with a probability >50% of being protein-coding according to Portrait (Arrial et al. 2009). The large number of transcripts removed by the two last filtering, raises the possibility that there are a significant number of protein-coding genes at the fungal genomes that remain un-annotated. The final data set contained 8,251 putative fungal ncRNAs, ~37% of them represented by long (>200nt) ncRNAs (**Table 22**).

Table 21. Fungal and Oomycete species used for the exploration of putative ncRNAs.

Species	Genes	Size	Source
Aspergillus flavus	12,604	36.8	www.broadinstitute.org
Aspergillus nidulans	10,560	30.1	www.broadinstitute.org
Aspergillus niger	11,200	37.2	www.broadinstitute.org
Aspergillus oryzae	12,063	37.1	www.broadinstitute.org
Aspergillus terreus	10,406	29.3	www.broadinstitute.org
Botryotinia fuckeliana	16,448	42.7	www.broadinstitute.org
Coccidioides immitis RS	9,910	28.9	www.broadinstitute.org
Coccidioides posadasii	10,228	27.6	www.broadinstitute.org
Cochliobolus heterostrophus	9,633	34.9	www.jgi.doe.gov
Colletotrichum graminicola	12,006	51.6	www.broadinstitute.org
Colletotrichum higginsianum	16,172	49.1	www.broadinstitute.org
Cryphonectria parasitica	11,609	43.9	www.jgi.doe.gov
Fusarium graminearum	13,322	36.4	www.broadinstitute.org
Fusarium oxysporum	17,708	61.4	www.broadinstitute.org
Fusarium solani	15,707	51.3	www.jgi.doe.gov
Fusarium verticillioides	14,188	41.8	www.broadinstitute.org
Heterobasidion annosum	13,405	33.6	www.jgi.doe.gov
Laccaria bicolor	23,130	60.7	www.jgi.doe.gov
Leptosphaeria maculans	12,469	44.9	www.jgi.doe.gov
Magnaporthe grisea	12,991	41.0	www.broadinstitute.org
Melampsora larici	16,831	101.1	www.jgi.doe.gov
Mucor circinelloides	11,719	36.6	www.jgi.doe.gov
Mycosphaerella fijiensis	13,107	74.1	www.jgi.doe.gov
Mycosphaerella graminicola	10,952	39.7	www.jgi.doe.gov
Neurospora crassa	9,907	41.0	www.broadinstitute.org
Paracoccidioides brasiliensis	9,137	32.9	www.broadinstitute.org
Paxillus involutus	17,968	58.3	www.jgi.doe.gov
Phanerochaete chrysosporium	10,048	35.1	www.jgi.doe.gov
Phycomyces blakesleeanus	16,528	53.9	www.jgi.doe.gov
Phytophthora capsici	19,805	64.0	www.jgi.doe.gov
Phytophthora infestans	18,181	228.5	www.broadinstitute.org
Phytophthora sojae	26,584	82.6	www.jgi.doe.gov
Pleurotus ostreatus	12,330	34.3	www.jgi.doe.gov
Podospora anserina	10,635	35.7	podospora.igmors.u-psud.fr
Postia placenta	17,173	90.9	www.jgi.doe.gov
Puccinia triticina	11,638	162.9	www.broadinstitute.org
Pythium ultimum	15,323	44.9	pythium.plantbiology.msu
Rhizopus oryzae	17,459	46.1	www.broadinstitute.org
Schizophyllum commune	13,181	38.5	www.jgi.doe.gov
Thielavia terrestris	9,813	36.9	www.jgi.doe.gov
Tremella mesenterica	8,313	28.6	www.jgi.doe.gov
Trichoderma atroviride	11,863	36.1	www.jgi.doe.gov
Trichoderma reesei	9,129	33.5	www.jgi.doe.gov
Trichoderma virens	12,427	39.0	www.jgi.doe.gov
Tuber melanosporum	7,496	124.9	www.genoscope.cns.fr
Ustilago maydis	6,522	19.7	www.broadinstitute.org

Note: Genes column indicates the total number of protein coding genes annotated at each genome. Size column indicates the genome size in Mb. Source column indicates the URL from where the genomes were downloaded.

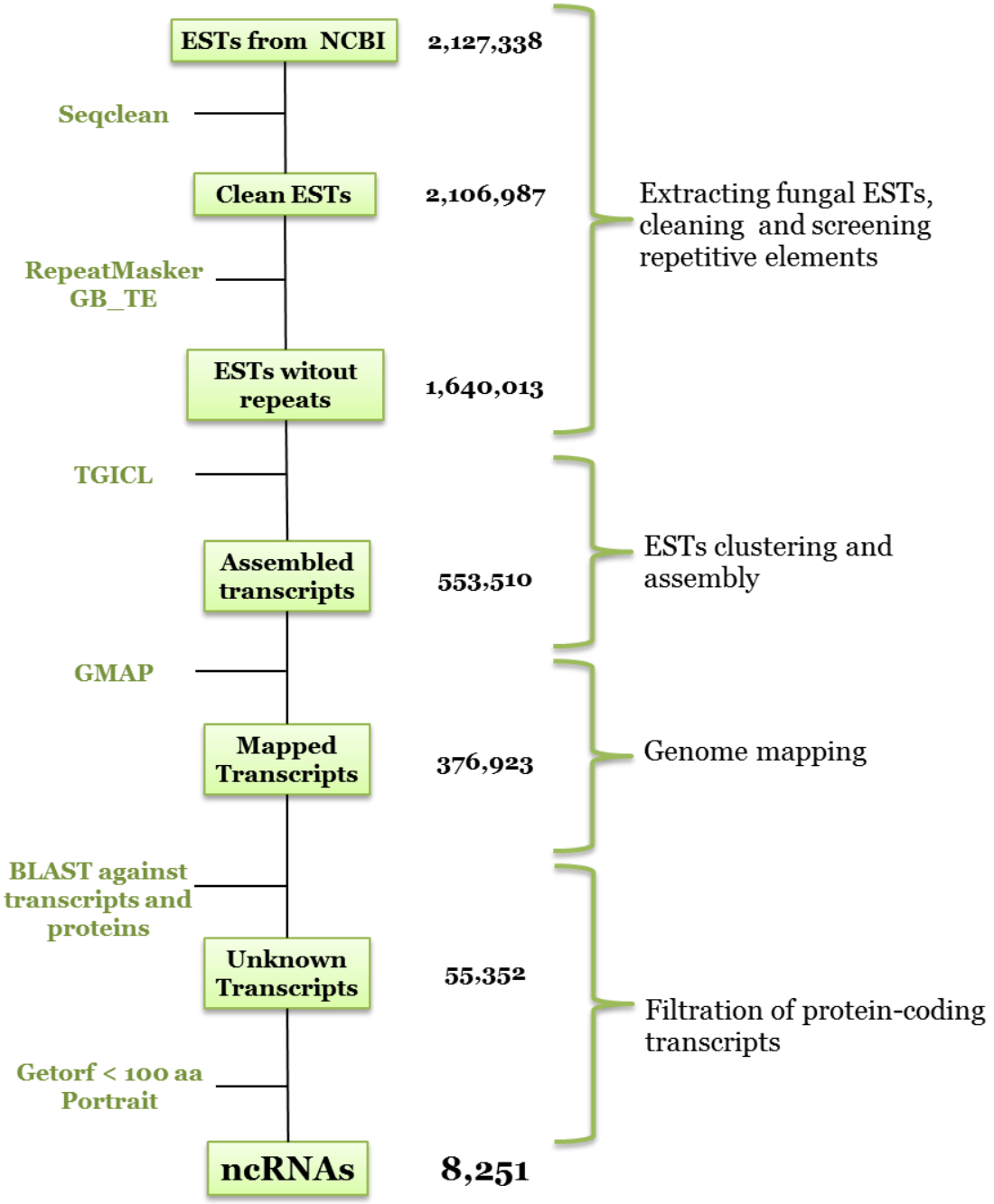

Figure 30. Flowchart describing the pipeline and results of the ncRNAs exploration. Numbers indicate totals for all species, but the pipeline was applied individually for each species.

The phytopathogenic ascomycete *Leptosphaeria maculans* showed the largest percentage of uniquely mapped transcripts classified as putative ncRNAs (11.1%), while the oomycete *Phytophthora capsici* had the lowest (0.1%) (**Table 22**). On average, 2.2% of the transcripts were classified as putative ncRNAs, a much lower percentage than the previously found by applying the same pipeline over bovine ESTs (15.5%) (Qu & Adelson 2012a).

Table 22. nsRNA content of each of the 46 fungal and oomycete species.

Species	NCBI ESTs	Transcripts Mapped	ncRNAs	Ratio (%)	% Long ncRNA
Aspergillus flavus	20,371	6,008	128	2.1	61.72
Aspergillus nidulans	16,848	5,563	371	**6.7**	39.62
Aspergillus niger	46,938	9,565	146	1.5	59.59
Aspergillus oryzae	9,051	7,205	188	2.6	60.11
Aspergillus terreus	12,776	2,007	18	0.9	27.78
Botryotinia fuckeliana	60,925	7,569	138	1.8	35.51
Coccidioides immitis RS	62,729	12,259	55	0.4	**89.09**
Coccidioides posadasii	110,163	14,903	105	0.7	46.67
Cochliobolus heterostrophus	28,747	8,269	87	1.1	27.59
Colletotrichum graminicola	2,380	988	22	2.2	**72.73**
Colletotrichum higginsianum	5,469	1,290	65	**5.0**	21.54
Cryphonectria parasitica	41,858	6,667	80	1.2	36.25
Fusarium graminearum	21,355	6,304	232	**3.7**	31.47
Fusarium oxysporum	9,304	3,090	277	**9.0**	43.32
Fusarium solani	33,120	5,339	27	0.5	59.26
Fusarium verticillioides	87,086	10,214	151	1.5	14.57
Heterobasidion annosum	41,232	7,914	24	0.3	16.67
Laccaria bicolor	34,348	6,015	30	0.5	**83.33**
Leptosphaeria maculans	42,319	6,069	675	**11.1**	22.07
Magnaporthe grisea	88,292	22,065	1,053	**4.8**	48.91
Melampsora larici	54,445	7,684	204	2.7	50.00
Mucor circinelloides	27,614	5,315	87	1.6	33.33
Mycosphaerella fijiensis	36,233	8,595	19	0.2	68.42
Mycosphaerella graminicola	32,194	11,008	10	0.1	50.00
Neurospora crassa	277,147	24,919	179	0.7	58.10
Paracoccidioides brasiliensis	41,487	7,602	78	1.0	**70.51**
Paxillus involutus	9,951	1,810	42	2.3	42.86
Phanerochaete chrysosporium	18,106	3,949	39	1.0	30.77
Phycomyces blakesleeanus	47,847	5,906	18	0.3	**83.33**
Phytophthora capsici	56,457	8,930	6	0.1	**83.33**
Phytophthora infestans	111,106	21,118	118	0.6	38.14
Phytophthora sojae	28,467	6,559	49	0.7	12.24
Pleurotus ostreatus	29,211	6,230	49	0.8	34.69
Podospora anserina	51,862	7,607	87	1.1	68.97
Postia placenta	38,114	7,364	26	0.4	65.38
Puccinia triticina	44,407	2,911	231	**7.9**	43.29
Pythium ultimum	100,391	23,797	2,249	**9.5**	18.90
Rhizopus oryzae	13,313	3,221	34	1.1	26.47
Schizophyllum commune	31,874	5,787	38	0.7	10.53
Thielavia terrestris	27,991	7,545	46	0.6	21.74
Tremella mesenterica	20,586	2,770	21	0.8	61.90
Trichoderma atroviride	35,125	9,452	255	2.7	56.47
Trichoderma reesei	44,966	6,239	46	0.7	30.43
Trichoderma virens	35,475	6,586	53	0.8	**77.36**
Tuber melanosporum	92,371	14,664	284	1.9	61.27
Ustilago maydis	45,287	10,052	111	1.1	20.72
Total	**2,127,338**	**376,923**	**8,251**	**2.2%**	**36.9**

Note: NCBI ESTs: Total number of ESTs downloaded from dbEST as at September, 2011. Transcripts Mapped: Number of transcripts mapped to the genome (including assembled contigs). ncRNAs: Total number of predicted ncRNAs. Ratio (%): Indicates the percentage of unique mapped transcripts classified as ncRNAs. %Long ncRNA: Indicates the percentage of ncRNAs larger than 200nt.

5.3.2. Annotation of putative ncRNAs

I used several methods to annotate and classify the 8,251 putative ncRNAs. As a result, I identified 590 (7.1%) ncRNAs with similarities to known ncRNAs (**Table 23**).

Table 23. Annotation of putative ncRNAs.

Class	Number	Database/Program
miRNAs	547	Mature miRNA (v.17)
snoRNAs/snRNAs	7	fRNAdb, Rfam10.0, *A. fumigatus*
tRNAs	7	tRNAscan-SE, fRNAdb
rRNAs	4	Rfam10.0
piRNA	6	fRNAdb
Conserved ncRNA	12	fRNAdb
Other	7	Rfam10.0, *A. fumigatus*
Total annotated	**590**	

Even though the vast majority of the ncRNAs did show similarity with well-annotated ncRNAs, considering that ncRNAs are usually less conserved than protein-coding genes and the fact that most databases used for the annotation were built with sequences from higher eukaryotes, these results show that the pipeline applied to fungal transcripts effectively identified sequences that encode for ncRNAs, but I was not able to classify most of them based on sequence similarity, in agreement with previous results obtained by Qu & Adelson (2012a).

5.3.3. Phylogenetic distribution of ncRNAs

I determined whether the proportion of transcripts classified as putative ncRNAs was correlated with the phylogenetic relationships of the fungal species (**Figure 31**). I found no evident pattern in the distribution of the ratio of putative ncRNAs neither along the phylogenetic tree of the 46 species nor significant differences between different phyla (ANOVA, $F=0.39$, $p=0.81$) (**Table 24**).

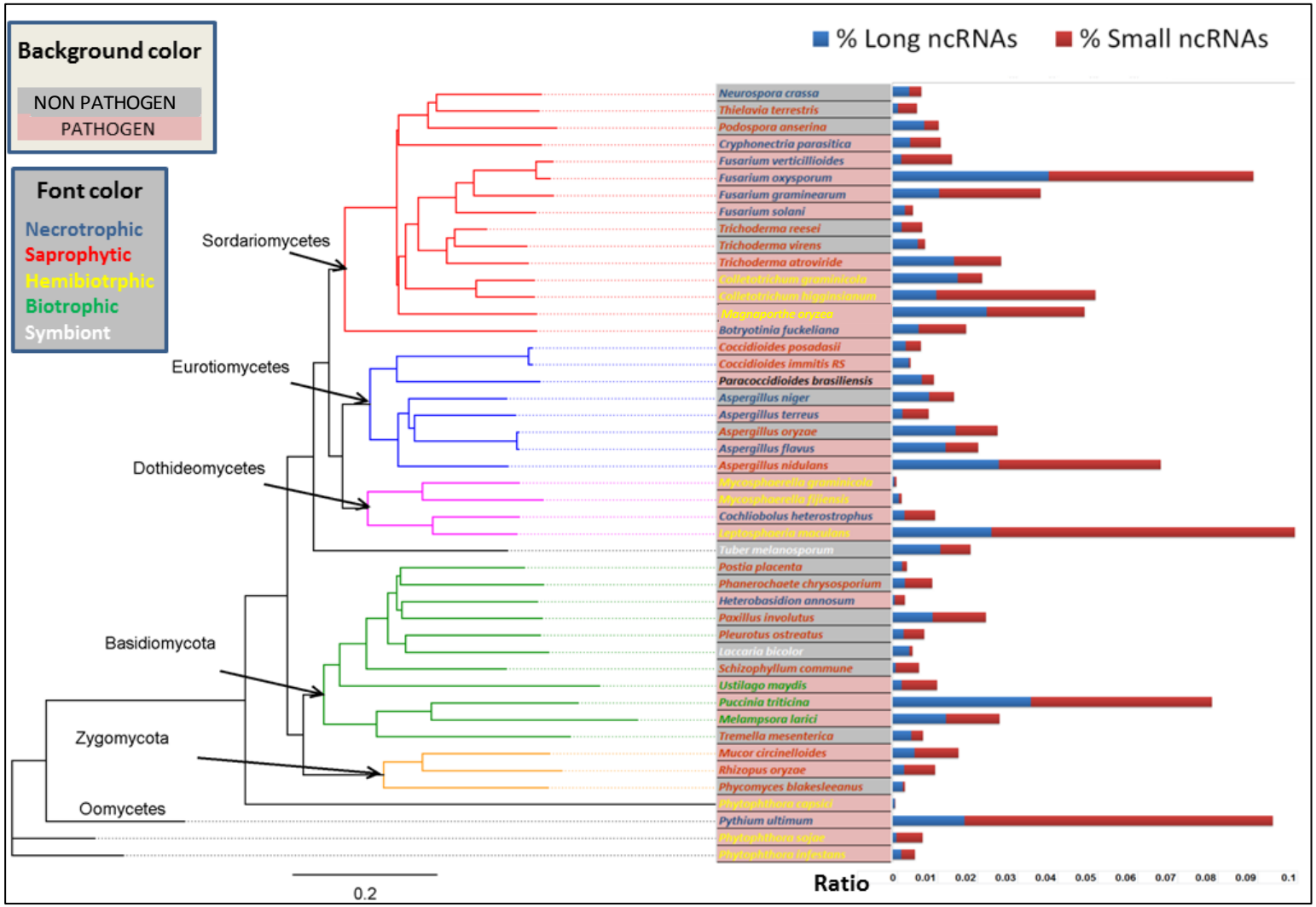

Figure 31. Phylogenetic reconstruction based on the single copy gene MS456. Background colors indicate whether the species is pathogenic or non-pathogenic. Font colors indicate the lifestyle of the species. Bars next to species name indicate the ratio (proportion) of transcripts classified as putative ncRNAs (number of ncRNAs/number of unique mapped transcripts). Blue and red bars colors indicate the percentage of long (>200nt) and short (<200nt) putative ncRNAs respectively.

However, by dividing species according to their ability to cause disease I found a slight but significant difference between the average ratios of annotated ncRNAs. Overall, pathogenic species have a larger ratio of ncRNAs than non-pathogenic species (ANOVA, F=4.63, p=0.036) (**Table 24**). On the other hand, lifestyle of fungal species is not related to the ratio of ncRNAs (ANOVA, F=0.91, p=0.46).

Table 24. Average ratio of transcript sequences classified as ncRNAs for different biological groups.

Pathogenicity (F=4.63, p=0.036)	**Species**	**Average Ratio (Std.Dev.)**
Pathogen	30	0.0274 (0.030)
Non-pathogen	16	0.0105 (0.006)
Lifestyle (F=0.91, p=0.46)	**Species**	**Average Ratio (Std.Dev.)**
Necrotrophic	13	0.0259 (0.029)
Saprophytic	18	0.0140 (0.015)
Hemibiotrophic	9	0.0276 (0.036)
Biotrophic	3	0.0390 (0.035)
Symbiotic	2	0.0122 (0.010)
Phylum (F=0.39, p=0.81)	**Species**	**Average Ratio (Std.Dev.)**
Sordariomycetes	15	0.0242 (0.023)
Eurotiomycetes	8	0.0200 (0.020)
Dothideomycetes	4	0.0312 (0.053)
Basidiomycota	11	0.0166 (0.022)
Zygomycota	3	0.0100 (0.006)
Oomycetes	4	0.0270 (0.045)

It is important to highlight that both the correlation or the absence of correlation between the ratio of ncRNAs and phylogeny, pathogenicity or lifestyle of the species included in the present analysis could be highly biased by the accuracy of the genome annotation of each species. Since most annotation projects usually applied different methods and levels of stringency to call gene models, and because one of the steps in the pipeline consists of discarding putative coding sequences by similarity to protein coding genes annotated in the genome, the annotation processes carried out in different species could lead to an over or under estimation in the total number of putative ncRNAs. The improvement of whole genome annotation as well as larger sets of ESTs could help to finally elucidate whether there is a correlation of any of these characteristics and the amount of ncRNAs transcribed.

5.3.4. Sequence conservation of predicted ncRNAs

In order to determine the sequence conservation of the predicted ncRNAs among the 46 species, I attempted to map each ncRNA to each of the 46 genomes using GMAP. I considered an ncRNA as "conserved" when it maps to a genome of a species other than its originating species with at least 20% coverage and 90% sequence identity. Using this criterion, 7,021 ncRNAs (~85%) did not map to any genome, with the exception of the genome from the species to which they belong (**Figure 32**). These results, in addition to the short number of ncRNAs that were annotated based on sequences similarity (**Table 23**), illustrates the low sequence conservation of ncRNAs in fungi.

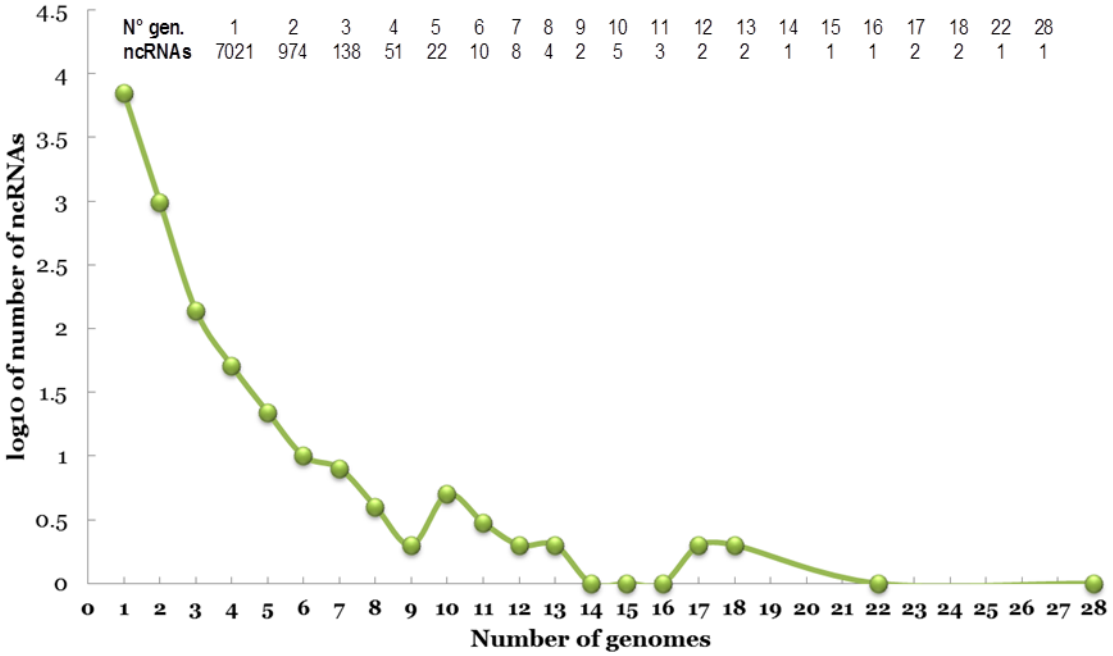

Figure 32. Conservation of putative ncRNAs among the 46 genomes. Axis Y indicates the log10 of the number of ncRNAs vs. the number of genomes on which the ncRNAs map at axis X. At the top, total number of ncRNAs mapping to different number of genomes.

5.3.5. EST library analysis

All 2,127,338 ESTs initially downloaded for the 46 species come from 423 different libraries in the NCBI-dbEST database. After cleaning and identifying uniquely mapped sequences, I obtained 376,923 uniquely mapped transcripts (**Figure 30** and **Table 22**), representing sequences from 407 libraries (**Table 25**). Most of these libraries were created under different conditions or from different vegetative or reproductive structures. Therefore, I determined whether nsRNAs were differentially expressed under different conditions.

For each species, I classified transcript sequences according to the library from which they were derived (see Materials and Methods). Two species (*Pythium ultimum* and *Colletotrichum graminicola*) contain ESTs obtained for just one library, so I could not analyze them. I performed a Chi-Squared test to examine differences between the expected number of transcript sequences for each library, and the observed number of transcripts classified as putative ncRNAs. For 35 species there were statistically significant differences between the two sets (**Table 25**), suggesting that ncRNAs were predominantly predicted from some of the libraries at these species. I compared the proportion of uniquely mapped transcripts with the proportion of predicted ncRNAs at each library by mean of a Z-test, correcting p-values for multiple comparisons (**Table 26**). I found that most libraries showing statistically significant differences were built from reproductive structures (cleistothecia, apothecia, spherules, perithecia and conidia) or during the formation of specialized vegetative structures (germ tube, haustoria), suggesting a relationship between the production of ncRNAs and the differentiation of specialized cell types.

Table 25. Total number of EST libraries analyzed in each species and p-values for the Chi-Squared (X^2) test for differences in the expected number of transcripts classified as putative ncRNAs.

Species	Libraries	p (X^2)
Aspergillus flavus	2	1.10E-01
Aspergillus nidulans	4	**3.17E-92**
Aspergillus niger	9	**1.02E-09**
Aspergillus oryzae	2	1.00E+00
Aspergillus terreus	5	**1.07E-03**
Botryotinia fuckeliana	9	**8.40E-36**
Coccidioides immitis RS	3	8.54E-02
Coccidioides posadasii	10	**1.00E-32**
Cochliobolus heterostrophus	4	**1.53E-18**
Colletotrichum graminicola	1	NA
Colletotrichum higginsianum	3	**1.29E-09**
Cryphonectria parasitica	3	**7.41E-13**
Fusarium graminearum	22	**4.53E-22**
Fusarium oxysporum	8	**5.76E-47**
Fusarium solani	4	**3.43E-03**
Fusarium verticillioides	12	6.01E-02
Heterobasidion annosum	6	**1.04E-08**
Laccaria bicolor	13	**3.01E-17**
Leptosphaeria maculans	15	**2.56E-03**
Magnaporthe grisea	20	**9.47E-10**
Melampsora larici	4	**1.42E-32**
Mucor circinelloides	2	4.32E-02
Mycosphaerella fijiensis	10	**2.32E-18**
Mycosphaerella graminicola	14	**1.72E-07**
Neurospora crassa	21	**3.34E-19**
Paracoccidioides brasiliensis	8	**8.52E-12**
Paxillus involutus	12	**9.74E-03**
Phanerochaete chrysosporium	5	**9.39E-26**
Phycomyces blakesleeanus	4	**1.06E-08**
Phytophthora capsici	3	**3.54E-04**
Phytophthora infestans	33	**1.10E-03**
Phytophthora sojae	16	**1.69E-19**
Pleurotus ostreatus	8	**7.72E-05**
Podospora anserina	7	**7.97E-06**
Postia placenta	12	**1.57E-05**
Puccinia triticina	20	2.57E-02
Pythium ultimum	1	NA
Rhizopus oryzae	3	**1.66E-06**
Schizophyllum commune	6	5.13E-01
Thielavia terrestris	2	1.44E-01
Tremella mesenterica	2	5.90E-01
Trichoderma atroviride	19	**6.57E-28**
Trichoderma reesei	7	**1.65E-09**
Trichoderma virens	12	**5.00E-21**
Tuber melanosporum	2	**1.88E-03**
Ustilago maydis	19	**3.04E-23**

Table 26. Results for the Z-test for proportions of unique mapped transcripts classified as ncRNAs.

Library (NCBI ID)	ALL (%)	ncRNA (%)	Z-test (p-value)	Library Description	ESTs
Aspergillus nidulans					
LIBEST_000419	1.20	7.00	6.28E-17	Sexual phase - cleistothecium	**272**
LIBEST_001614	2.00	7.60	9.94E-12	Early sexual phase	**264**
LIBEST_001083	76.00	84.00	2.60E-03	Vegetative mycelia - asexual structures	**12,457**
Aspergillus niger					
LIBEST_001923	0.40	3.30	1.21E-07	unknown	**250**
LIBEST_000268	5.00	16.00	9.37E-10	unknown	**2283**
LIBEST_017622	4.20	26.00	1.41E-40	Mycelial growth	**2,768**
Botryotinia fuckeliana					
LIBEST_027483	2.40	16.00	3.04E-25	Apothecia	**3,450**
LIBEST_027480	13.20	27.00	1.23E-05	pH3 to pH8 shift	**7,284**
Coccidioides posadasii					
LIBEST_014533	11.00	29.00	1.43E-08	Saprobic phase (mycelia)	**17,154**
LIBEST_007112	2.00	12.00	8.38E-12	Spherule	**2,452**
Cryphonectria parasitica					
LIBEST_021137	4.60	38.00	5.78E-42	Mycelia, conidium infected	**3,384**
Fusarium graminearum					
LIBEST_015778	11.00	31.00	1.11E-20	Sexual phase - perithecia	**1,843**
LIBEST_013803	2.00	5.80	1.96E-03	Mycelia on wheat heads	**547**
Fusarium oxysporum					
LIBEST_019886	72.00	96.00	1.45E-17	Conidium, germ tube and mycelium	**6,448**
Fusarium solani					
LIBEST_023777	0.80	11.00	3.05E-07	Mycelia, 27 hr culture	**563**
Heterobasidion annosum					
LIBEST_011935	0.90	20.00	5.67E-20	Germinating spores	**1,274**
Laccaria bicolor					
LIBEST_012444	3.60	54.00	1.71E-49	Three-weeks-old	**896**
LIBEST_020925	2.00	12.00	5.97E-03	30 to 40 mm mature fruitbody	**730**

Leptosphaeria maculans					
LIBEST_026972	18.00	34.00	3.61E-22	Mycelium in V8 for 5 days	**6,624**
LIBEST_026970	5.00	9.00	3.05E-04	Conidia germinating in liquid Fries for 48 h	**3,528**
LIBEST_026971	0.20	1.10	6.76E-03	Conidia germinating in V8 juice for 28h	**212**
LIBEST_026969	3.00	6.00	7.36E-04	Conidia germinating in liquid Fries for 48 h	**2,934**
Magnaporthe oryzae					
LIBEST_015475	6.00	22.00	3.25E-90	Biotrophic phase	**7,729**
LIBEST_015481	14.00	35.00	6.53E-78	Biotrophic phase	**13,798**
Melampsora larici					
LIBEST_026131	7.00	56.00	1.41E-152	Haustoria from infected poplar	**3,028**
Neurospora crassa					
LIBEST_001553	2.00	9.50	7.55E-14	Mycelia 22hr growth in dark	**10,850**
LIBEST_001554	1.50	16.00	2.20E-34	Mycelia 22hr growth in dark	**9,075**
LIBEST_008672	1.60	4.70	6.79E-03	Perithecia (fruiting bodies)	**2,682**
LIBEST_024059	13.00	25.00	1.74E-05	7 DayssPost-Cross Sexual Growth	**35,139**
LIBEST_001625	0.30	2.60	4.07E-07	Perithecia (fruiting bodies)	**496**
Paracoccidioides brasiliensis					
LIBEST_022143	0.13	6.32	9.82E-38	Differentially expressed sequences	**17**
LIBEST_020653	4.58	21.05	2.56E-12	Transition library	**1,107**
Phanerochaete chrysosporium					
LIBEST_015053	2.48	41.46	3.49E-48	unknown	**253**
LIBEST_015054	0.81	12.20	8.86E-13	unknown	**66**
Phytophthora infestans					
LIBEST_021205	0.60	4.08	7.79E-06	Interaction with Potato	**6,230**
LIBEST_016734	0.02	0.68	1.72E-05	unknown	**23**
LIBEST_018798	0.23	8.16	2.62E-69	Germinated cysts	**139**
Phytophthora sojae					
LIBEST_014539	12.67	36.54	1.36E-05	Infection and propagation	**4,031**
LIBEST_014544	0.46	5.77	2.93E-06	Free swimming (Zoospores)	**1,019**
LIBEST_018786	0.41	5.77	5.16E-07	unknown	**74**
Podospora anserine					
LIBEST_023239	13.37	32.00	1.50E-06	Perithecia older than 48h	**8,007**

Rhizopus oryzae					
LIBEST_020022	1.52	15.00	6.35E-10	unknown	**350**
LIBEST_020021	1.06	10.00	1.75E-06	unknown	**254**
Trichoderma atroviride					
LIBEST_020345	5.81	11.58	1.10E-03	Mycelium	**990**
LIBEST_023781	22.41	54.66	9.86E-38	Conidiation	**5,352**
Trichoderma reesei					
LIBEST_018903	4.58	28.00	3.19E-13	unknown	**2,047**
LIBEST_014605	4.46	8.00	8.82E-07	Asexual Mycelium	**3,825**
Trichoderma virens					
LIBEST_019599	3.69	19.72	1.43E-10	Mycelia	**1,613**
LIBEST_02009	3.14	22.54	9.96E-18	Mycelia Grown in Crab Chitin	**468**
LIBEST_020322	0.62	4.23	5.01E-03	Mycelia Grown insN-acetylglucosamine	**196**
Ustilago maydis					
LIBEST_019704	1.65	7.87	4.96E-06	unknown	**544**

Note: Just tests showing a significant p-value ($p<0.01$) are displayed. For each library, ALL (%) represent the percentage of the unique mapped transcripts originated from that library and ncRNA (%) shows the percentage of ncRNAs originated from that library. Z-test (p^*) shows the p-value for the Z-test of proportions after correction for multiple comparison using the Bonferroni method. ESTs indicate the initial number of ESTs for the library.

5.3.6. Natural selection in C. graminicola ncRNAs

Previously, I have identified hundreds of non-protein coding sequences showing evidences of natural selection between isolates of *C. graminicola* (see **Chapter II**). I analyzed the 22 ncRNAs predicted for *C. graminicola* to determinate whether some of them show evidence of natural selection (**Table 27**). Six of them were not tested since they are located at intergenic regions too far away from the adjacent protein coding gene (see Materials and Methods in **Chapter II** for a detailed explanation of why some regions were not tested). Four of the remaining 18, represent ncRNAs transcribed at regions previously described as under the action of natural selection. ncRNA 170324397, which is transcribed from the 3'UTR of gene GLRG_06645 shows evidence of selective sweeps (D<0), ncRNAs 170325157 and CL108Contig1, transcribed from the 3'UTR of genes GLRG_01081 and GLRG_08685 respectively show evidence of balancing selection (D>0) and ncRNA 170325164, transcribed from the 3'UTR of gene GLRG_02993 shows signals of positive selection (**Table 27**). These sequences, representing putative ncRNAs that show evidence of be under the action of different selective constraints, represent very interesting candidates for functional validation since they could be part of the RNA regulatory network involved in the control of gene expression determining different aspects in the biology of this fungus like the virulent behavior, adaptation to new environments of switches in the lifestyle.

Table 27. Description for the 22 ncRNAs identified at *C. graminicola*.

ncRNA	Region	Gene	Chr.	Selection
170324008 (-)	3'UTR	GLRG_10507 (-)	4	No Selection
170324031 (+)	3'UTR	GLRG_05155 (+)	2	No Selection
170324226 (+)	INTER	GLRG_01327 (-) GLRG_01328 (+)	3	Not tested
170324260 (-)	3'UTR	GLRG_09727 (-)	10	No Selection
170324397 (+)	3'UTR	GLRG_06645 (+)	7	**D*< 0**
170324463 (+)	3'UTR	GLRG_00437 (+)	4	No Selection
170324478 (-)	3'UTR	GLRG_00013 (-)	4	No Selection
170324491 (-)	INTER	GLRG_07779 (-) GLRG_07780 (-)	3	Not tested
170324577 (-)	3'UTR	GLRG_00941 (-)	8	No Selection
170324851 (+)	5'UP	GLRG_05371 (+)	2	Not tested
170324872 (+)	3'UTR	GLRG_04871 (+)	1	No Selection
170325094 (-)	3'UTR	GLRG_09823 (-)	10	No Selection
170325157 (-)	3'UTR	GLRG_01081 (+)	8	**D* > 0**
170325164 (+)	3'UTR	GLRG_02993 (-)	1	**Positive Selection**
170325317 (-)	Mean position: 3,428,684		2	Not tested
170325373 (-)	INTER	GLRG_08631 (+) GLRG_08632 (-)	1	Not tested
170325585 (-)	3'UTR	GLRG_11836 (-)	S*.141	No Selection
170325749 (-)	3'UTR	GLRG_04275 (-)	7	No Selection
170325793 (-)	3'UTR	GLRG_05486 (-)	6	No Selection
170325834 (+)	INTER	GLRG_02236 (+) GLRG_02237 (-)	4	No Selection
170325885 (-)	3'UTR	GLRG_08737 (+) GLRG_08738 (-)	2	Not tested
CL108Contig1 (+)	3'UTR	GLRG_08685 (+)	2	**D* > 0**

Note: Each ncRNAs appears with its corresponding GenBank ID (gi). Contig CL108Contig1 was built using ESTs 170325334 and 170325621. Region indicates the relative position of the ncRNAs regarding the nearest gene. 3'UTR implies that the ncRNAs is located at less than 200nt from the stop codon of the coding sequence of gene indicated at the Gene column. (+) and (-) Indicate transcriptional orientation (sense) of putative ncRNAs (ncRNA column) and mRNA (Gene column). ITER indicate that the ncRNA is located at the intergenic region between two very close genes. Chr. indicates the chromosome at which the ncRNAs is located (S*.141 = unanchored supercontig 141). Selection indicates results obtained for the genomic region at **Chapter II** regarding test of selection. D*>0: Tajima's D value significantly positive, D*<0: Tajima's D value significantly negative and Positive Selection: significant p-value ($p < 0.05$) at the Maximun Likelihood Ratio test according to Wong & Nielsen (2004) (see **Chapter II**).

5.4. Discussion

Recent, large transcriptomic analyses have shown that almost the entire genome of complex eukaryotes is transcribed (Carninci et al. 2005; Frith et al. 2005; Jochl et al. 2008; Arrial et al. 2009; Jacquier 2009; Mercer et al. 2009). Further functional and *in silico* analyses have revealed that a large fraction of such transcript sequences is comprised of ncRNAs that are likely to be functional (Tupy et al. 2005; Havilio et al. 2005; Ravasi et al. 2006; Araki et al. 2006; Qu & Adelson 2012a; Qu & Adelson 2012b). Even though the abundance and functionality of ncRNAs in eukaryotic genomes have long been supported, large-scale scans for ncRNAs are restricted to higher or model organisms. In the present study I attempted to explore, annotate and characterize ncRNAs along the fungal and oomycete phylogeny. I analyzed publicly available ESTs for 42 fungal and 4 oomycetes species using a computational pipeline to identify non protein coding transcript sequences. After curation, clustering, assembly, mapping and removing protein coding transcripts, I found 8,251 putative ncRNA sequences, representing the 2.2% of the unique transcript sequences.

I identified 590 (7.1%) putative ncRNAs with similarities to previously described ncRNAs (**Table 23**). The vast majority of annotated ncRNAs correspond to microRNAs (miRNAs), which are involved in the transcriptional and post-transcriptional regulation of gene expression and are widespread among eukaryotes (Cerutti & Casas-Mollano 2006). However, most ncRNAs did not have detectable sequence similarity with well-annotated ncRNAs. Three main reasons could be accounting for this observation. First, in contrast to other methods aimed at detecting putative ncRNAs, the pipeline applied in the present study is not restricted to the identification of conserved ncRNAs, which enables us to detect putative ncRNAs that were not previously described (Qu & Adelson 2012a). Second, most RNA databases are largely comprised of sequences from mammals and other higher eukaryotes, which make homology search methods difficult to applied for fungal ncRNAs. And third, in contrast to protein coding sequences, many functional ncRNAs have been shown to be poorly conserved at the sequence level, even between closely related species (Pang et al. 2006), suggesting that some classes of ncRNAs might be under rapid sequence evolution, making the identification and annotation ncRNAs

based on sequence similarity useless. In fact, I evaluated sequence conservation of the 8,251 fungal ncRNAs and I found that the 85% of them were not present in any of the other 46 species and just 21 ncRNAs were conserved among 10 or more species (**Figure 32**). These results, together with the low percentage of ncRNAs that were effectively classified (**Table 23**), suggest that fungal ncRNAs are poorly conserved at the sequence level. However, functions of many types of ncRNAs are determined not only by their sequences but also by their secondary structures, which leaves open the possibility that many of the ncRNAs identified in the present study actually play similar functions because their structural similarities (Sun & Buhler 2008; Achawanantakun et al. 2011). A more detailed analysis using RNA secondary structure modeling and alignment methods could be implemented in order to address the structural conservation of these fungal ncRNAs.

I analyzed the proportion (ratio) of transcript sequences classified as ncRNAs between different phylogenetic and ecological groups among the 46 fungal and oomycetes species (**Figure 31** and **Table 24**). I did not find any pattern regarding the amount of ncRNA and the fungal phylogeny or lifestyle. Nevertheless, I found a small but significantly larger amount of ncRNAs in pathogenic species than in non-pathogenic species. Pathogenic organisms are expected to be more complex than their free-living relatives since they usually cause disease by modulating the metabolism of the host through enzymes, toxins, growth regulators, and other substances they secrete and they have to absorb foodstuffs from the host cells for their own use, meanwhile avoid host defenses. In fact, several (but not all) genomes of filamentous plant pathogens are larger than the genomes of non-pathogenic species, but this difference has been largely associated with the expansion and contraction of specific gene families or the proliferation of repetitive DNA, which should confer genomic plasticity, contributing to the emergence of new virulence traits (Raffaele & Kamoun 2012). However, as proposed for higher eukaryotes, it remains possible that the greater amount of non-coding DNA in pathogenic organisms could also be related to an increased regulatory sophistication by ncRNAs. Whole genome transcriptomic experiments under different conditions, comparing pathogenic and non-pathogenic species or strains from the same species could help to clarify whether or not ncRNAs play fundamental roles in pathogenicity.

In order to determine the nature of putative fungal and oomycete ncRNAs I analyzed the source EST libraries from which transcript sequences were generated. I found that a significant proportion of ncRNAs were predicted from transcript sequences expressed in reproductive structures or during the formation of specialized vegetative structures (**Table 26**). For instance, for the plant pathogenic fungus *Melampsora larici*, which causes poplar leaf rust and represent the most devastating and widespread pathogen of poplars, I identified 7,684 unique mapped transcripts (**Table 22**) from the analysis of four EST libraries (**Table 25**). Three of the four libraries (LIBEST_026130, LIBEST_024927 and LIBEST_024926) contain a mixture of ESTs generated under diverse conditions. However, the fourth (LIBEST_026131) contains ESTs generated uniquely from haustoria during infection on poplar. While only 7% of total transcript sequences are represented by ESTs from library LIBEST_026131, ncRNAs of *Melampsora larici* contains 56% of transcript sequences represented by this library, suggesting that ncRNAs were significantly overrepresented in this library (**Table 26**). Such a clear result is not so evident for all some EST libraries do not have a sufficiently detailed description of the tissue from which EST libraries were made. Also, some of the EST libraries contain a mixture of transcripts from different conditions or from different structures. However, as general rule, many EST libraries significantly overrepresented in the ncRNAs were generated during advanced stages of the mycelial growth, during the germination of reproductive structures or at specialized reproductive or vegetative structures (**Table 26**). These results suggest that ncRNAs could be involved in the differentiation of specialized cell types in fungal species. The implication of ncRNAs in cellular differentiation and development has been largely proposed (Mattick 2007; Amaral & Mattick 2008; Amaral et al. 2008). As in higher eukaryotes, it is possible that ncRNAs in fungi are involved in complex gene regulatory networks responsible for the modulation of mRNAs at particular growing stages and tissues. Determining the number and functions of ncRNAs controlling cellular differentiation is a fundamental issue for pathogenic organisms, mainly for those that exhibit distinct phases in their life cycle like hemibiotrophic (e.g. Colletotrichum spp., *Magnaporthe oryzae* or *Leptosphaeria maculans*) or dimorphic (e.g. *Coccidioides immitis*, *Paracoccidioides brasiliensis* or *Ustilago maydis*) fungi,

since they could potentially be targets for the development of novel anti-fungal drugs (Jochl et al. 2008).

One of the key questions raised in **Chapter II**, was whether evidence of natural selection acting on non-coding DNA sequences in *C. graminicola* was influenced by the presence of ncRNA sequences. Even for the low number of putative ncRNAs identified in this species, I found that four of them have evidence of some kind of selection (**Table 27**), suggesting their relative importance in the adaptive process of this pathogenic fungus. An additional outcome from the analysis of the genomic localization of the 22 putative ncRNAs of *C. graminicola* was the prevalence of ncRNAs mapping very close to 3' ends of protein-coding genes, most of them transcribed in the same sense than the closest genes (**Table 27**). The possibility exists that some of these ncRNAs are extensions of the 3'UTRs produced by alternative transcription start or termination sites of protein-coding genes (Qu & Adelson 2012a). On the other hand, ncRNAs transcribed from the 3'UTRs could also correspond to a recently described class of 3'UTR-related ncRNAs found in humans, mouse and fly called 3'UTR-derived RNAs (uaRNAs), which can be expressed separately from the associated protein-coding gene and act as *cis* or *trans* regulatory sequences (Mercer et al. 2011). Interestingly, the only two "antisense" ncRNAs (regarding transcriptional orientation of the closest gene) in *C. graminicola* (170325157 and 170325164) show evidence of balancing and positive selection (**Table 27**), in agreement with the results of Qu and Adelson (2012a), who found, in bovines, that antisense ncRNAs tend to show a lower sequence conservation than ncRNAs transcribed in the same orientation than protein coding genes. These antisense ncRNAs, showing patterns of rapid evolution could potentially be involved in central regulatory roles controlling aspects of the plant-pathogen interaction like the progression through the lifecycle or the expression of virulence factors as have been demonstrated in other microbial pathogens (Svenningsen et al. 2009; Bordi et al. 2010; Chinni et al. 2010; Matrajt 2010).

In conclusion, by applying a pipeline based on the analysis of public EST, I identified a set of putative fungal and oomycete ncRNAs, most of them previously unknown. Understanding and determining functional importance of

these ncRNAs could potentially open the possibility of explore new strategies for the development of novel anti-fungal drugs whit multiple implications in agriculture and human health.

6. CHAPTER IV

Positive selection in disease response genes within members of the Poaceae

6.1. Introduction

Antagonistic coevolution between plants and pathogens triggers an evolutionary arms race in which the host evolves to escape pathogen infections, and pathogens evolve to escape host defenses. Genes involved in this interaction are expected to show signatures of adaptive molecular evolution through positive selection (Aguileta et al. 2009) where advantageous mutations are retained in the population. For instance, several plant genes related to defense mechanisms have been reported to evolve under positive selection (Bishop et al. 2000; Roth & Liberles 2006; Zamora et al. 2009).

The maize genome contains more than 32,500 protein coding genes. 22,874 of the proteins were assigned to a least one InterPro term (Schnable et al. 2009). Domains associated with defense related genes (DRGs) are well represented, including 240 glycoside hydrolases, 213 pathogenesis-related transcriptional factors, and 211 with N-terminal leucine-rich repeats. Identifying those genes that play a fundamental role during pathogen attack is a difficult but important task. In a previous study, we carried out suppression subtractive hybridization experiments in the maize-*Colletotrichum graminicola* pathosystem during early stages of anthracnose leaf blight development (Vargas et al. 2012). We found more than 200 differentially expressed genes from maize, including 36 DRGs and 34 genes encoding hypothetical or unknown proteins (HUPGs) identified as up-regulated during infection. Moreover, the recent release of Phytozome v8.0 (Goodstein et al. 2011) offers the opportunity for fast and accurate access to families of orthologous and paralogous genes in the Poaceae lineage. Using the information available in Phytozome, I further analyzed the previously identified genes (Vargas et al. 2012) to investigate genomic patterns of positive selection on HUPGs and DRGs across the Poaceae lineage.

6.2. Material and Methods

Figure 33 shows the workflow designed to analyze positive selection in the 36 DRGs and the 34 HUPGs. It consists of three main steps: 1) Identification of orthologs and paralogs, 2) Remove highly divergent gene clusters and 3) Positive selection tests.

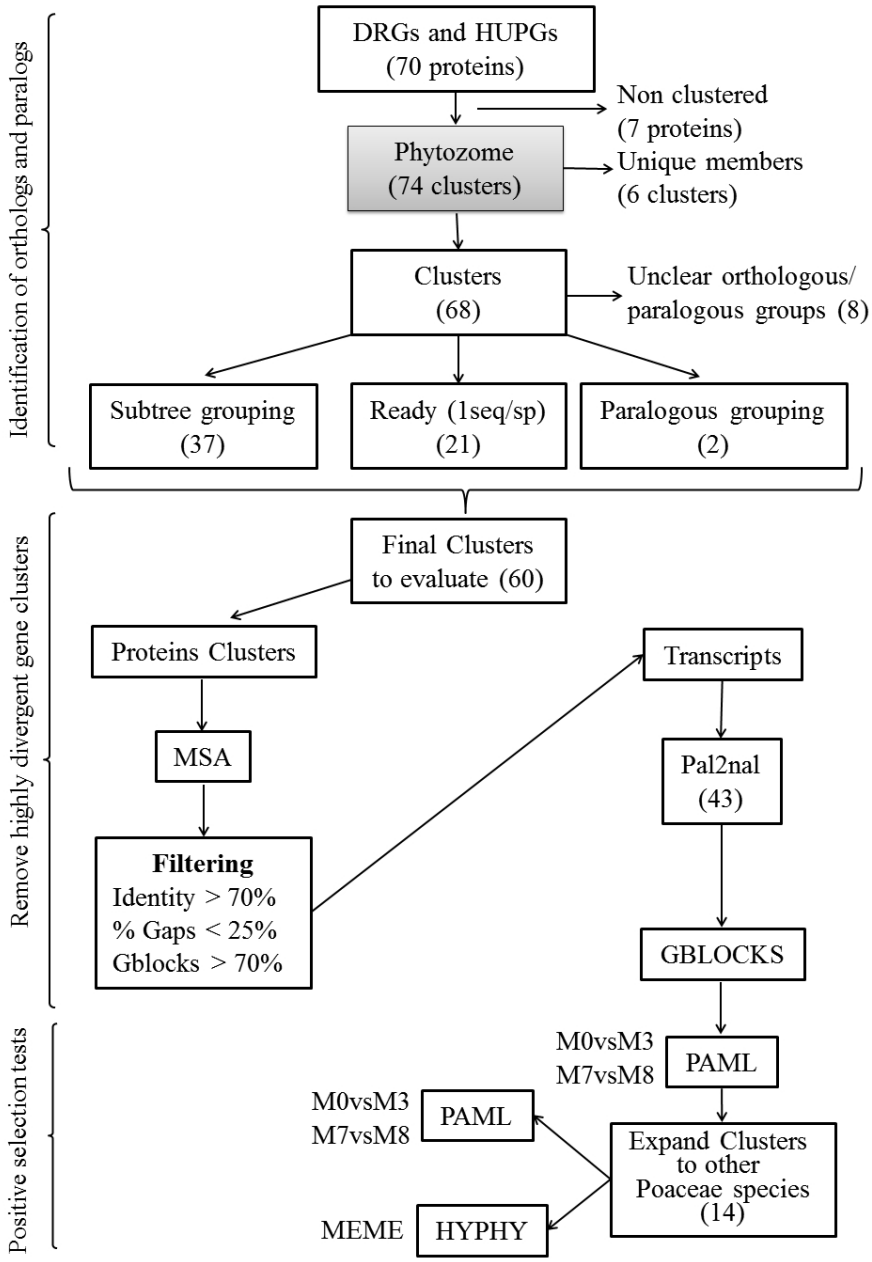

Figure 33. Flowchart of methods used to identify families of defense-related genes (DRGs) and hypothetical or unknown protein-coding genes (HUPGs) under positive selection in the Poaceae lineage.

6.2.1. Identification of orthologs and paralogs

Using a custom Python script I systematically identified Phytozome gene clusters in the Poaceae (Grass) node looking for all clusters containing at least one of the 70 genes IDs under study. In some cases, maize genes appear in more than one cluster, so all of them were taken into account for the analysis. Grass gene clusters in Phytozome v8.0 contain sequences from *Zea mays* (Zm), *Sorghum bicolor* (Sb), *Setaria italica* (Si*), Oryza sativa* (Os) and *Brachypodium distachyon* (Bd). Gene families with more than 1 sequence for each species were manually curated by analyzing the phylogenetic tree and selecting those sequences clustered together in a subtree resembling the species phylogeny, an indication of orthology (Zamora et al. 2009). I additionally analyzed gene clusters containing more than 4 sequences of maize in order to look patterns of positive selection in groups of paralogs.

6.2.2. Remove highly divergent gene clusters

I first aligned protein clusters using MUSCLE v3.8 (Edgar 2004) with the default options and then filtered the multiple sequence alignments (MSA) to discard clusters with highly divergent sequences. MSAs were retained for further analysis if the average pairwise identity within the cluster was greater than 70%, the percentage of gapped residues was less than 25% (Mondragón-Palomino et al. 2002) and the percentage of sites conserved by Gblocks (using the default parameters) (Castresana 2000) was greater than 70%. Clusters that did not meet these requirements were discarded (as in the case of both paralogous sets) and the ones that did it were then tested for positive selection.

6.2.3. Positive selection tests

CODEML implemented in the PAML v4 software package (Yang 2007) was used to fit two kinds of models to the data, models that allow positive selection (M3 and M8) and models that do not (M0 and M7). For each model, the log likelihood (lnL) values were obtained and two likelihood ratio tests (LRTs) were performed (M0vsM3 and M7vsM8) as $2*(\ln L1-\ln L0) = 2\Delta L$, which was compared with a χ^2 distribution to test whether there was statistical difference between the two models (critical value 5.99 at 5% significance level).

Clusters showing statistically significant differences at any of the two LRTs were "expanded" by adding more orthologous sequences from other Poaceae species and then conducting the positive selection tests again. For each of the 14 clusters, I used the protein sequence from maize as query in TBLASTN searches against the "nr" and "est" DNA databases and filtering results for Poaceae species. An e-value $<10^{-20}$, and a coverage > 90% of the complete ORF were required to be considered as putative orthologs. The new "expanded" orthologous clusters were then tested again using CODEML. When the LRT indicated positive selection, the Bayes empirical Bayes (BEB) method was used to calculate the posterior probabilities that each codon is from the site class of positive selection under model M8 (Yang et al. 2005).

I used the MEME algorithm (Murrell et al. 2012), part of the HyPhy package (Kosakovsky Pond et al. 2005) implemented in the Datamonkey webserver (Delport et al. 2010), to test for evidence of episodic selection. The LRT is reliable even if only a few similar sequences are analyzed (Anisimova et al. 2001b), but the power of prediction of positive selection sites by M8+BEB is low when few closely related sequences are used (Anisimova et al. 2002). Sizes of clusters analyzed here were between 5 and 12 members and the relative sequence divergence was between 0.11 (31452004-e) and 0.56 (31456723-e) nucleotide changes per codon per branch (estimated as S/(2T-3) according to Anisimova et al. (2001), so the low numbers of taxa as well as the low sequence divergence may prevent the identification of sites under positive selection by M8+BEB. In addition, the M8+BEB method may fail to recognize sites where selection is episodic. The MEME algorithm pools information over branches in the phylogenetic tree to gain power to detect episodic selection at a site, reducing the stringency of the analysis. Sites indicated as evolving under positive selection by MEME and M8+BEB were visually inspected, and only the ones that do not appear at low-quality alignment regions were reported.

6.2.4. Three-D modeling

When possible, protein 3-D structure models were built based on fold recognition using the Protein Homology/analogY Recognition Engine v.2.0 (Phyre2) (Kelley & Sternberg 2009) and the 3-D representation of molecular

structures was obtained using Geneious v5.4 (Drummond et al. 2011). Amino acids predicted as evolving under positive selection where then mapped into the protein structure.

6.3. Results and Discussion

A total of 74 clusters were identified in Phytozome across the Poaceae (grass) node, which contain the 70 up-regulated maize genes (36 DRGs and 34 HUPGs). Seven sequences were classified in at least two different clusters and six out the 74 clusters contain the maize sequence as the only member. Three of them GRMZM2G040493, GRMZM2G080466 and GRMZM2G078124, encode proteins with unknown function and without any known functional domains. The remaining are genes GRMZM2G016922 (a putative kaurene synthase), GRMZM2G001084 (putative ATP-dependent Clp protease) and GRMZM2G088088 (a putative importin subunit alpha protein). It is interesting to highlight that sequence GRMZM2G001084 (cluster 32722009) shows high similarity (99.7%) with GRMZM2G123922 (cluster 31448528), however they are grouped into different clusters. While cluster 31448528 contains one member for each grass species, GRMZM2G001084 is the only member of cluster 32722009. All sequences in cluster 31448528 show high similarity along the full length of the protein and GRMZM2G001084 differs only in the N-terminal region, outside the highly conserved functional domain, which may be responsible for the separation into two different clusters. There is no supporting EST or protein evidence for the N-terminal region of the GRMZM2G001084 gene model so I cannot exclude the possibility that the gene is incorrectly annotated. Subcellular localization analysis using TargetP v1.1 (Emanuelsson et al. 2000) suggests that the GRMZM2G123922 protein is targeted to the chloroplasts while that signal is absent in GRMZM2G001084. These results suggest that either GRMZM2G001084 is not correctly annotated, or that GRMZM2G001084 represents an ancient gene copy of GRMZM2G123922 that has gained a novel function after the gene duplication event.

In addition, genes GRMZM2G080466 and GRMZM2G016922, unique members of specific clusters, are up-regulated during infection with both, *C. graminicola* and *Ustilago maydis* (Doehlemann et al. 2008; Horst et al. 2010; Vargas et al. 2012). These genes induced during infection with both fungal species may represent evolutionary innovations in the maize genome directly related with defense mechanisms and are very interesting candidates for future functional characterization.

The remaining 68 clusters were analyzed in order to obtain a set of orthologous genes and filtered to avoid highly divergent sequences that may cause false positive results in the positive selection tests. A total of 43 clusters outperformed the filtering and were tested for signatures of positive selection. CODEML implemented in the PAML v4 software package (Yang 2007) was used to fit two kinds of models to the data: models that allow positive selection (M3 and M8) and models that do not (M0 and M7). For each model, the log likelihood (lnL) value were obtained and two Likelihood Ratio Test (LRTs) were performed (M0vsM3 and M7vsM8) as 2*(lnL1-lnL0) = 2ΔL, which were compared with a χ^2 distribution to test whether there was statistical difference between the two models. Fourteen clusters showed statistically significant differences at any of the two LRTs, indicating that either there was heterogeneity in ω values among sites, or evidence of positive selection in the clusters. Next, I used BLAST to search for additional Poaceae orthologous sequences in public databases for each cluster and the positive selection tests were performed again.

Results and characteristics for the 14 clusters are summarized in **Table 28**. The comparison M0vsM3 shows that all 14 clusters have variations for *dN*/*dS* ratio (ω) among their codons, indicating that selective constraints are heterogeneous between sites. A total of 6 clusters (31469106-se, 31456723-e, 31466309-se, 31477240-se, 31452004-e and 31443992-se) showed evidence for positive selection through M7vsM8 comparison. Using the Bayes empirical Bayes method implemented under M8 (M8+BEB) (Yang et al. 2005), I identified several sites in the alignment for each cluster in which the approximate mean of the posterior distribution for ω was >1.

Since the M7vsM8 LRT is a conservative test and in some cases can fail to detect positive selection due to lack of power of the LRT (Anisimova et al. 2001b), as in the case of sites where episodes of positive selection are confined to a small subset of branches in a phylogenetic tree, I also used the Mixed Effects Model of Evolution (MEME) (Murrell et al. 2012), which is able to identify lineage-specific events of positive selection (episodic selection) even though the same site is neutrally or negatively evolving in the rest of the lineages. As expected,

MEME showed a greater power to detect sites under positive selection, finding episodic selection in all 14 clusters.

Two of six clusters identified by LRT (M7vsM8) have already been described as evolving under positive selection in plants. One of them is cluster 31466309-se, in which four sites showed ω>1 in the approximate mean of the posterior distribution using M8+BEB, two of them with posterior probabilities (Pr ω>1) > 0.95 (sites 4A and 285I). Additionally, 10 sites were identified by MEME showing episodic selection. This cluster contains members of the chitinase class III, a well-known defense related protein involved in fungal cell-wall degradation, and described by Bishop et al. (2000) as evolving under positive selection. Furthermore, cluster 31443992-se showed 4 sites with ω>1 using M8+BEB, two of them with (Pr ω>1) > 0.8 (148V and 166M) and 10 sites under episodic selection identified by MEME. Three-D structure of maize protein (GRMZM2G402631_P01) was predicted based on template PDB ID: 1AUN with confidence=100% and coverage=86%. Sites predicted as evolving under positive selection were then mapped onto the tertiary structure (**Figure 34D**), and sites under positive selection are situated on the surface of the protein, which is consistent with the results already described by Zamora et al. (2009). This cluster contains members of the PR5 family, thaumatin-like proteins with known antifungal and anti-insect activity.

Table 28. Results for positive selection tests and characteristics for the fourteen clusters.

Maize Gene	Genome Annotation	Cluster	Species	M0 vs M3		M7 vs M8			MEME
				ω	Δ2L	ω	Δ2L	PSS (BEB)	PSS
GRMZM2G149800	Hypothetical protein LOC100277913	31469106-se	Zm, Si, Bd, Os, Sb, Ta, Hv, Pe	24.09	**245.65**	2.59	**21.2360**	None	71T, 227R, 232S, 342Y
GRMZM2G324297	Unknown (similar to arogenate dehydrogenase)	31456723-e	Zm, Si, Bd, Os, Sb, Ta, Hv	1.10	**470.95**	1.30	**17.4570**	**151Q****, **158R**, 240K, **253D**, **257A**, 306S, **323Q**, **349R**, 368F	147A, 182S, 211L, 301S, 358R
GRMZM2G453805	Chitinase class III	31466309-se	Zm, Si, Bd, Os, Sb, Ta, Pe	6.83	**216.60**	7.31	**14.2870**	**4A****, 9A, **260P**, **285I***	87G, 138G, 160S, 169A, 184L, 222G, 224A, 269I, 282I, 293V
GRMZM2G168502	Hypothetical protein LOC100217285	31477240-se	Zm, Si, Bd, Os, Sb, Ta, Hv, Pe	0.83	**222.73**	3.85	**10.2076**	130V, **347S***, 356T	32V, 99N, 120N, 144A, 184V, 203C, 206A, 235E, 293S, 296C
GRMZM2G056369	Putative Isocitrate lyase	31452004-e	Zm, Si, Bd, Os, Sb, Ta, Hv	0.40	**145.16**	1.10	**8.5392**	30G, 76G, **334K**, **345R**, **371T**, **560P**, 561R, 564T, 572M	142L, 293S, 320C, 330G, 331V, 344D, 345R
GRMZM2G402631	Pathogenesis-related protein 5	31443992-se	Zm, Si, Bd, Os, Sb, Ta, Hv, As, Pe, Sc, Or, Tm	1.12	**201.12**	1.31	**8.3213**	57P, 91Q, **148V, 166M**	39G, 53T, 56N, 58G, 84G, 98A, 109L, 146R, 182T, 185P
GRMZM2G465226	Pathogenesis-related protein 1	31458297-se	Zm, Si, Bd, Os, Sb, Ta, Hv, Pe	0.79	**219.62**	7.57	1.8310	None	99S, 108D, 128A, 134V
GRMZM2G338809	Hypothetical protein LOC100382111	31445034-se	Zm, Si, Bd, Os, Sb, Ta, Hv, Pe	0.33	**272.13**	1.63	0.5130	None	35T, 37A, 121F, 131A, 266V, 271S, 278C, 342I, 393K, 413A
GRMZM2G011085	Uncharacterized protein	31447249-e	Zm, Si, Bd, Os, Sb, Hv, Pe	0.10	**20.92**	1.00	0.0032	None	95L, 160T, 195V

GRMZM2G039639	Protein P21	31461924-se	Zm, Si, Bd, Os, Sb, Ta, Hv, Pe	0.80	**179.11**	1.00	0.0002	None	23L, 34V, 98V, 107A, 145G, 161S, 165A, 202K, 209P
GRMZM2G080499	Hypothetical protein	31455526-e	Zm, Si, Bd, Os, Sb, Hv	0.52	**82.64**	1.00	0.0001	None	67T, 90T, 93T, 161P, 400S, 452T, 803V
GRMZM2G117971	Hypothetical protein LOC100191593	31462333-se	Zm, Si, Bd, Os, Sb, Ta, Hv	0.45	**133.46**	2.57	0.0001	None	33Q, 36G, 50N, 103T, 108S, 118S, 131K, 138Q
GRMZM2G322129	Putative uncharacterized protein	31457048	Zm, Si, Bd, Os, Sb	0.85	**189.35**	1.00	1.6598	None	28V, 37P, 38H, 119R, 135L, 172G, 173S, 203S, 228R, 250K, 255D, 263M, 284H, 312D, 434L, 595S
GRMZM2G057093	Chitinase	31480783-s	Zm, Si, Bd, Os, Sb	0.74	**80.97**	4.38	0.0614	None	177T, 277S

Note: Cluster: clusters are named according to Phytozome v8.0 (Grass node), adding "*e*" when the cluster was expanded to other Poaceae species and "*s*" when a subtree was extracted from original Phytozome cluster. Species: one sequence for each of the following species was included in the cluster: Zm: *Zea mays*, Sb: *Sorghum bicolor*, Si: *Setaria italica*, Os: *Oryza sativa*, Bd: *Brachypodium distachyon*, Ta: *Triticum aestivum*, Hv: *Hordeum vulgare*, Pe: *Phyllostachys edulis*, As: *Aegilops speltoides*, Sc: *Secale cereale*, Or: *Oryza rufipogon*, Tm: *Triticum monococcum*. ω=dN/dS estimated under M3 and M8. Δ2L: likelihood ratio estimated as 2*(lnL1-lnL0) between M0vsM3 and M7vsM8, in bold those are statically significant at 0.05. PSS: Positively selected sites indentified by BEB when the approximate mean of the posterior distribution for w is > 1. Clusters with posterior probability > 0.8, *>0.95 and ** > 0.99 are in bold type. PSS (MEME) show sites identified by MEME as under episodic selection (p-value <0.1). In both cases amino acids refer to maize sequence.

Among clusters without previous evidence of positive selection I found 31452004-e. This cluster contains members of isocitrate lyase (ICL) enzyme, which in plants is an enzyme exclusively found in the glyoxylate cycle. Even though its specific role in defense mechanisms has not been elucidated yet, glyoxylate metabolism seems to be an important part of the defense mechanisms activated by plants during infection by pathogens (Scheideler et al. 2002). Sequence analysis revealed that nine sites showed ω>1 using M8+BEB, four of them with (Pr ω>1) > 0.8 (334K, 345R, 371T and 560P), and seven sites displayed episodic selection after analysis using MEME. The 3-D structure for protein GRMZM2G056369_P01 was modeled based on PDB template 1DQU (confidence=100% and coverage=90%), the crystal structure of the tetrameric ICL from *Aspergillus nidulans*. Each subunit defines two domains. Domain I is associated with the centre of the tetramer and shows high similarity to triose phosphate isomerase (TIM) barrel. Domain II forms a peripheral head to the subunit (Britton et al. 2000). Most sites identified as evolving under positive selection were located between sites 293 and 371 in the protein alignment (**Figure 35**), which match with residues belonging to domain II (**Figure 34C**). This domain appears to be unique to eukaryotic ICLs and has been proposed to be important for the association of ICL with peroxisomes (Dunn et al. 2009). Further functional studies will contribute to a better understanding of domain II and its functional relevance during plant-pathogen interactions.

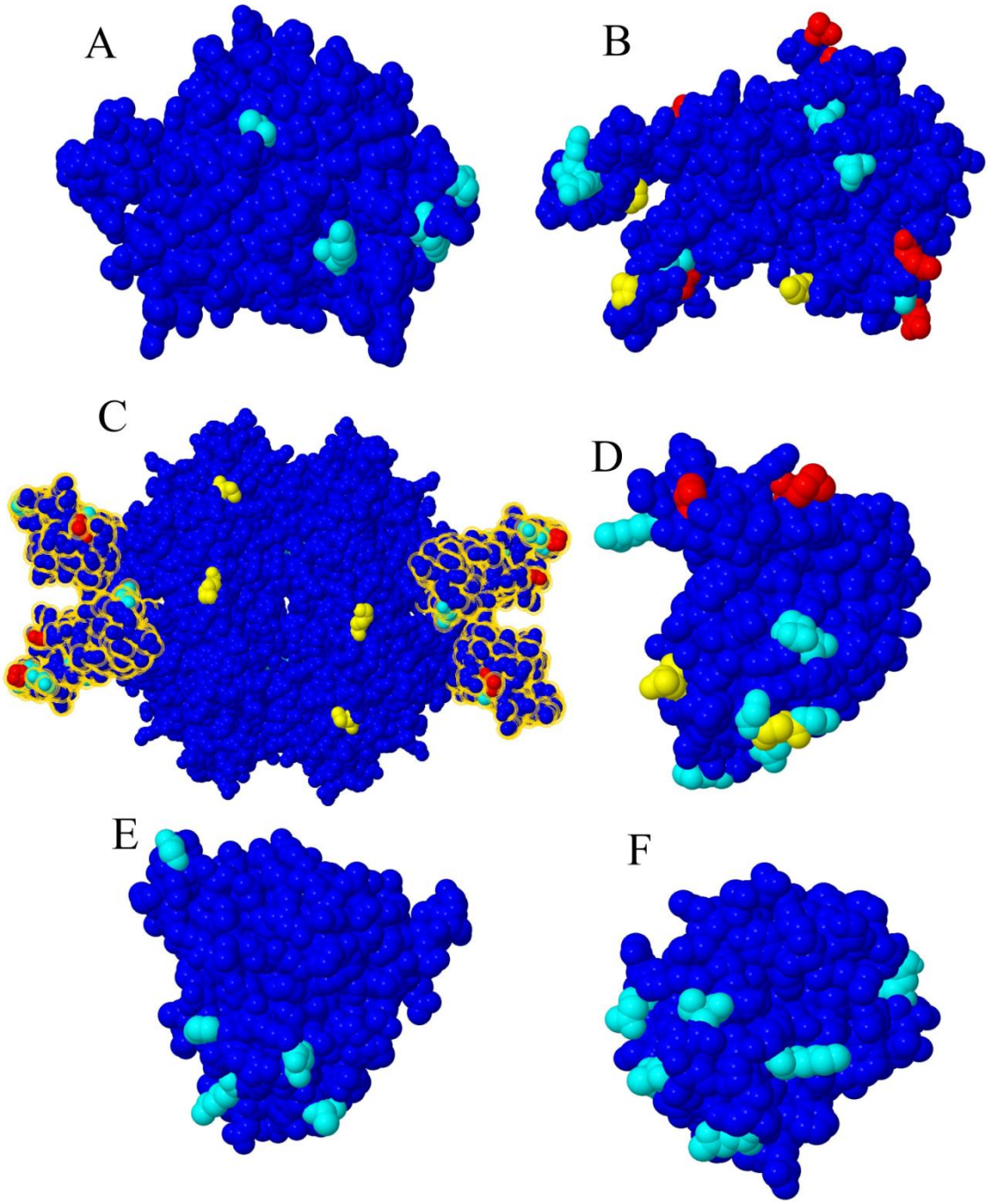

Figure 34. Modeled 3-D structure for six maize proteins based on fold recognition using Phyre2. **A)** GRMZM2G149800_P02 (cluster: 31469106-se, PDB ID: 3OOX). **B)** GRMZM2G324297_P02 (cluster: 31456723-e, PDB ID: 3KTD). **C)** Tetrameric structure of isocitrate lyase from *A. nidulans* (PDB ID: 1DQU). Domain II highlighted in yellow. **D)** GRMZM2G402631_P01 (cluster: 31443992-se, PDB ID: 1AUN). **E)** GRMZM2G039639_P01 (cluster: 31461924-se, PDB ID: 1DU5). **F)** GRMZM2G117971_P01 (cluster: 31462333-se, PDB ID: 1BW3). Amino acids inferred as evolving under positive selection are colored. PSS (M8+BEB) when the approximate mean of the posterior distribution for ω is > 1 and posterior probability > 0.5 (yellow) and > 0.8 (red). PSS (MEME), p-value < 0.1 (cyan).

Clusters 31469106-se, 31456723-e and 31477240-se contain proteins annotated as hypothetical or with unknown function. In the case of 31469106-se, sites under positive selection could not be identified by M8+BEB, but MEME identified four sites with evidence of episodic selection. Even though the protein GRMZM2G149800_P02 (cluster 31469106-se) has been annotated as a hypothetical protein, its 3-D structure was modeled using as template the crystal structure of a putative 2OG-Fe(II) oxygenase (PDB ID: 3OOX, confidence=100.0% and coverage=84%) (**Figure 34A**). In addition, using the Superfamily database (Gough et al. 2001) I found that this maize protein belongs to a superfamily of clavaminate synthase-like proteins, an oxidoreductase involved in the biosynthesis of clavulanic acid and other 5S clavams (Tahlan et al. 2004). Clavulanic acid is a well-known antibiotic, with metabolites of the clavam family having shown antibacterial and antifungal activities (Paradkar A S et al. 1997). In the case of cluster 31456723-e, nine sites showed ω>1 using M8+BEB, six of them with (Pr ω>1) > 0.8. In addition, MEME identified five sites under episodic positive selection. The 3-D structure of maize protein belonging to this cluster (GRMZM2G324297_P02) was predicted based on template PDB ID: 3KTD (confidence=100% and coverage=65%) and sites identified as evolving under positive selection were located throughout the protein surface (**Figure 34B**). Proteins in this cluster show similarities with arogenate/prephenate dehydrogenases. Members of this group of enzymes catalyze a step during tyrosine biosynthesis in the shikimate pathway, which is involved in the biosynthesis of aromatic amino acids and a wide range of secondary metabolites. Although the regulation and coordination of the synthesis of these amino acids are not well understood, many secondary metabolites have shown to be important during plant defense against herbivores, pests and pathogens (Bennett et al. 1994; Tzin & Galili 2010). In fact, the accumulation of phenolic compounds and increased levels of hydroxycinnamic acid derivatives have been detected in maize leaves during *C. graminicola* infection (Vargas et al. 2012). Finally, cluster 31477240-se showed three sites with ω>1 using M8+BEB, one of them with (Pr ω>1) > 0.95 and 10 sites with evidences of episodic selection. While annotated as hypothetical proteins, all sequences in this cluster contain two copies of a domain associated with drug/metabolite transport. Simmons et al. (2003) have also described a

putative multidrug transporter (*Zm-mfs1*) up-regulated in maize during *Cochliobolus heterostrophus* and *Cochliobolus carbonum* infection. Multidrug transporters play critical roles in plant defense, such as decreasing the accumulation of toxins secreted by the pathogen and exporting secondary metabolites out of the cell (Peng et al. 2011). Ten transmembrane regions were predicted using TMHMM v.0.92b (Krogh et al. 2001) for all sequences in this cluster and eight out of ten sites identified by MEME were found in the transmembrane regions (**Figure 36**). Interestingly, substitutions located at transmembrane segments cause modifications in the activity and substrate specificity of some transporters belonging to this family (Egner et al. 2002; Loo & Clarke 1994). Unlike most transporters, the relationship between multidrug transporters and their substrates is not highly specific, and they have the ability to recognize and transport a wide variety of structurally different organic compounds (Conte & Lloyd 2011). However, it is possible that the sites under positive selection might be crucial to determine the kinetic capabilities of the transporter. In this sense, rapid evolution would be acting on the selection of a more efficient transporter protein to grant a better metabolic fitness of the host cells during the pathogenic attack.

The clusters of proteins annotated as hypothetical proteins deserve attention since they may represent genes that have not been previously described as important during plant-pathogen interactions. One of them is cluster 31462333-se, which showed eight sites under episodic selection. Maize protein (GRMZM2G117971_P01) shows high similarity with Barwin proteins and the 3-D structure was predicted based on template PDB ID: 1BW3 (confidence=100% and coverage=83%, **Figure 34F**). Barwin domains are common to pathogenicity related proteins IV (PR4), which have been described with antifungal activity against a wide variety of pathogens, such as *LrPR4* (Li et al. 2010). In the case of cluster 31455526-e, seven sites showed signatures of episodic selection. Maize protein in this cluster (GRMZM2G080499_P01) is annotated as unknown function protein, but presents a serine/threonine protein kinases domain. Several serine/threonine protein kinases have shown to be involved in the recognition of pathogen-derived signal molecules (Martin et al. 1993; Romeis 2001; Song et al. 1995) and some have already been identified as evolving under positive selection in plants (Roth & Liberles 2006).

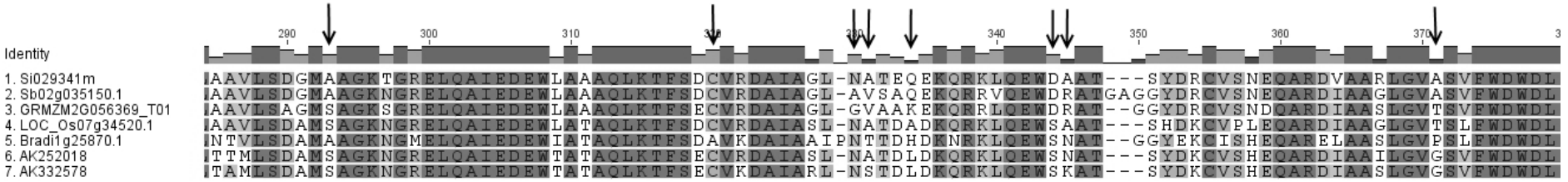

Figure 35. Multiple sequence alignment of protein cluster 31452004-e. Arrows indicate sites identified under positive selection by both methods, M8+BEB and MEME.

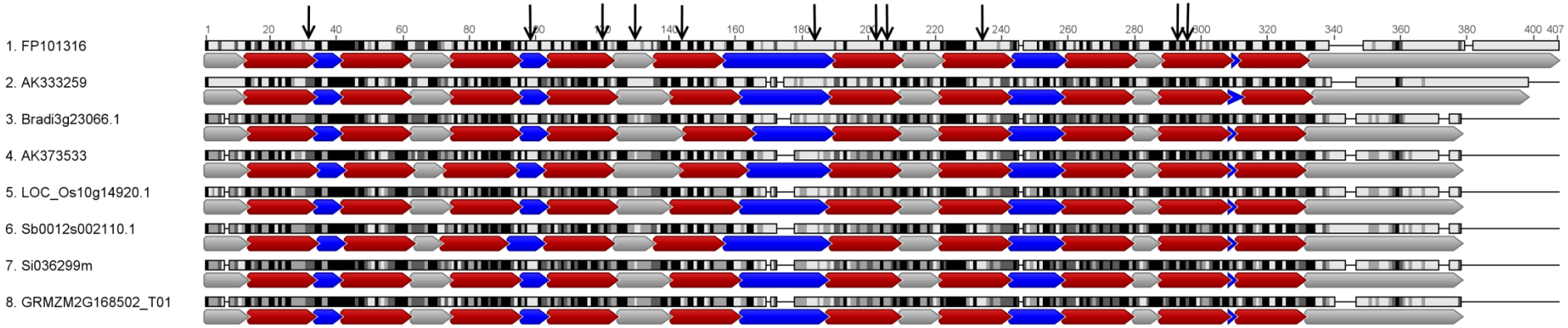

Figure 36. Multiple sequence alignment of protein cluster 31477240-se. Vertical black arrows indicate sites identified under positive selection, horizontal arrows indicate: potential transmembrane regions (red), potential cytoplasmic regions (grey) and potential extracellular regions (blue). Most sites under positive selection appear at the potential transmembrane regions.

Another cluster with several sites presenting episodic selection is cluster 31457048. A total of 16 sites were identified by MEME, the highest number of sites between all the clusters analyzed in this study. The maize protein in this cluster (GRMZM2G322129_P01) shows similarity to a transducin-like protein from sorghum. Ontologies associated with this cluster in Phytozome are mainly related to nucleotide/protein binding activities. In addition, the C-terminal region shows similarity to the armadillo (Arm) repeat, which is present in proteins acting in intracellular signaling events and in proteins with other essential functions such as cytoskeletal regulation (Coates 2003). Interestingly, 15 out of 16 sites under episodic selection appear outside the armadillo domain, at the N-terminal region of the protein. Four motifs with the conserved pattern LxxLxL in the C-terminal region were also identified in all sequences in the cluster with the exemption of GRMZM2G322129_P01 (containing only three motifs). The LxxLxL pattern has been proposed by molecular modeling as being sufficient to provide the characteristic horseshoe curvature present in leucine-rich repeat (LRRs) proteins (Kajava & Kobe 2002). Among the most important LRR proteins involved in plant defense are the nucleotide-binding site-leucine-rich repeat (NBS-LRR), pivotal players for pathogen detection. However, no NBSs were detected in any of the proteins of the cluster. Furthermore, most sites under positive selection were detected outside of the putative LRR region, where the recognition of the pathogen would be expected to occur. However, Mondragón-Palomino et al. (2002) have already described sites under positive selection outside of the LRR region in this family, concluding that these sites could also be important for the detection and signaling transduction during plant pathogen interactions.

I also found evidence of episodic selection in two additional clusters containing hypothetical or uncharacterized proteins. The first is cluster 31445034–se that includes proteins similar to ammonium transporters. These transporters may play important roles in plant defense against alkalinizing pathogens as *Colletotrichum spp.*, which secrete ammonia to increase the pH of the host tissue and infect the host (Prusky et al. 2001; Miyara et al. 2012). The second cluster is 31447249-e, which contains proteins with similarity to the PRELI/MSF1 domain, and has three sites that have undergone episodic selection. The function of the PRELI/MSF1 domain is unknown, although it is

conserved in lipid-binding proteins and proteins involved in vesicle transport and secretion mechanisms (Anantharaman & Aravind 2002).

Finally, I obtained three clusters of proteins with known function where MEME detected sites under episodic selection. Two of them, 31458297-se and 31480783-s contain well-known pathogenicity related proteins as PR1 and chitinases displaying just four and two sites evolving rapidly, respectively. The third cluster is 31461924-se and showed nine sites under episodic selection. The maize protein from this cluster (GRMZM2G039639_P01) is annotated as protein P21, a family that belongs to the thaumatin-like proteins and with highly similarity to osmotins PR5. The three-D structure for GRMZM2G039639_P01 was modeled based on template PDB ID: 1DU5 (confidence=100% and coverage=89%), the crystal structure of zeamatin, an antifungal protein from maize with membrane-permeabilizing activity (Batalia et al. 1996; Roberts & Selitrennikoff 1990). Sites predicted as evolving under episodic selection were mapped along the whole surface of the protein (**Figure 34E**), showing a similar pattern to the PR5 previously described in this work and by Zamora et al. (2009). Although the mechanism of action of osmotins has not been elucidated yet, it has been proposed that these proteins may target cell-wall and membranes of the fungal cell (Anžlovar & Dermastia 2003).

I present here a total of 14 gene clusters with evidence for positive selection in the Poaceae lineage. Maize genes belonging to these clusters have been shown to be up-regulated in maize during development of anthracnose caused by the hemibiotrophic fungus *C. graminicola*. Some of them have already been identified as evolving under positive selection while others are reported here for the first time. In general, most of the residues under positive selection detected in this study are exposed to the surface of the proteins (**Figure 34**). These results are expected for proteins with important functions such as ligand-protein interactions, allosteric regulations and signal perception. Six clusters showed positive selection even using one of the most stringent tests, the LRT between models M7 and M8 implemented in CODEML. According to Phytozome v8.0, five out the 14 clusters contain a single-copy gene at each of the analyzed Poaceae species (31456723-e, 31452004-e, 31447249-e, 31455526-

e and 31457048), therefore these genes are excellent candidates for further functional analysis.

In the present work, I identified a set of genes that probably have been playing important roles during plant-microbe interactions for millions of years of antagonist coevolution between grasses and their pathogens. This information may aid the understanding of molecular mechanisms involved in plant defense since products of these genes may be interacting with effectors produced by pathogens or involved in metabolic pathways that are important for defense. Accordingly, these genes represent a set of important candidates for functional validation through biochemical and genetic studies in order to identify targets to take into account in plant breeding programs as well as in pursuit of environmentally friendly plant protection compounds.

7. CONCLUSIONS / CONCLUSIONES

1. Phenotypic aspects of eight strains of the filamentous fungus *Colletotrichum graminicola* were successfully determined, with results suggesting a correlation between virulence and the ability of the isolates to grow under extreme conditions, to sporulate and to produce appresoria.

2. A non pathogenic isolate was described in the present work, providing a useful resource for the identification of genomic traits involved in the development of anthracnose.

3. Whole genomes of seven field isolates of *C. graminicola* showing variability in virulence to maize and collected from different regions of the world were successfully resequenced, assembled and annotated.

4. Overwhelming evidence of recombination was identified in the genomes of *C. graminicola*, mainly in large chromosomes, suggesting the possibility of sexual reproduction in field isolates of this fungus, but other processes could also account for these results.

5. Phylogenetic relationships between isolates of *C. graminicola* did not reveal any correlation between relatedness and their degree of virulence to maize.

6. Genomic structural variations in each chromosome are correlated with the amount of repetitive DNA and affect mainly chromosome 6 and minichromosomes, suggesting that they may represent sources for the generation of new genetic variants.

7. Genes coding for *C. graminicola* specific effectors are surrounded by breakpoints of genomic structural variations, suggesting that this mechanisms could be involved in promoting variability on these pathogenicity related proteins.

8. Four genes coding for enzymes directly involved in the degradation of plant cell walls are affected by genomic structural variations occurring only in the genome of the non pathogenic isolate, suggesting that the disruption of some of these genes could be the consequence of the absence of pathogenicity in this isolate.

9. The unique genes in each isolate are mainly pathogenicity related genes, with a high proportion of small secreted proteins (putative effectors), which could potentially represent evolutionary innovations directly involved in the host or environment specificity.

10. Thirty five genes are present in all of the pathogenic isolates, but are missing in the non pathogenic isolate, including some effectors, virulence factors, a cutinase and six genes upregulated during infection of isolate M1.001 on maize, representing excellent candidates for functional validation since they are uniquely lost in this asymptomatic isolate.

11. Natural selection differentially affects patterns of polymorphism in coding and non-coding sequences of pathogenicity-related genes. Genes coding for effector proteins and secondary metabolites show evidence of positive selection acting on the coding sequence, in agreement with an Arms Race model of evolution. Genes upregulated during different phases of infection show unusual patterns of polymorphisms relative to neutral expectations in the 3' UTR, 5' UTR and intronic regions.

12. 5' UTRs of genes coding for effector proteins and genes upregulated during infection show an excess of high frequency polymorphisms likely the consequence of balancing selection and consistent with the Red Queen hypothesis of evolution between host and pathogen acting on these putative regulatory sequences.

13. Results from the analysis of natural selection, suggest that even though adaptive substitutions on coding sequences are important for proteins that interact directly with the host, polymorphisms in the regulatory sequences may confer flexibility in the virulence processes.

14. A computational pipeline for the identification of ncRNAs based on the analysis of publicly available sequences was successfully applied over 2,127,338 ESTs from 46 fungal and oomycetes species, resulting in the identification of 8,251 putative ncRNAs, most of them previously unknown.

15. Fungal ncRNAs are poorly conserved at the sequence level between the analyzed species and most of them seem to be involved in cell differentiation. The functional importance of putative fungal ncRNAs remains unknown, but the characterization of ncRNAs could potentially open the possibility of exploring new strategies for the development of novel anti-fungal drugs with multiple implications in agriculture and human health.

16. Fourteen genes differentially expressed in maize during *Colletotrichum graminicola* infection show evidence of positive selection in the Poaceae lineage, suggesting that they have a role in the arms race between plants and pathogens. These proteins may be targets for pathogen produced effectors.

17. Some of the fourteen genes have already been described as evolving under positive selection, but others are reported here for the first time, including genes encoding an isocitrate lyase, dehydrogenases, a multidrug transporter, a protein containing a putative leucine-rich repeat and other proteins with unknown functions.

18. Mapping positively selected residues onto the predicted 3-D structure of proteins showed that most of them are located on the surface, where proteins are in contact with other molecules.

1. Se determinaron y cuantificaron diferentes aspectos fenotípicos de ocho aislados del hongo filamentoso *Colleototrichum graminicola*. Los resultados muestran una correlación positiva entre virulencia y la tasa de crecimiento *in vitro* bajo condiciones de nutrientes limitados, el número de esporas por placa y el porcentaje de germinación.

2. La caracterización fenotípica y posteriores análisis llevados a cabo por nuestro Grupo, revelaron la existencia de un aislado no patogénico de *C. graminicola*, el cual proporciona un recurso útil para la identificación de características genéticas involucradas en el desarrollo de la antracnosis.

3. Se re-secuenciaron, ensamblaron y anotaron los genomas completos de siete aislados de *C. graminicola* fenotípicamente diferentes, con un grado variable de virulencia frente al maíz y procedentes de diferentes regiones del mundo.

4. Se detectaron evidencias de recombinación genética entre los genomas de *C. graminicola*, principalmente para los cromosomas grandes, lo que sugiere la posibilidad de que *C. graminicola* se reproduzca sexualmente en campo. Sin embargo, otros procesos como la recombinación parasexual también podrían explicar los resultados obtenidos.

5. El análisis filogenético no reveló correlación entre distancia genética y grado de virulencia de los aislados.

6. El cromosoma 6 y los minicromosomas se encuentran significativamente más afectados por variaciones estructurales, las cuales a su vez están correlacionadas con la cantidad de ADN repetitivo. Estos cromosomas podrían estar involucrados en la generación de nuevas variantes genéticas en *C. graminicola*.

7. El análisis de genes afectados por variaciones estructurales reveló un enriquecimiento de genes codificantes de efectores específicos de *C. graminicola*, por lo que estos mecanismos genéticos podrían promover la variabilidad en genes involucrados en patogenicidad.

8. Entre los genes afectados por variaciones estructurales únicas al genoma del aislado no patogénico, se identificaron cuatro genes codificantes de enzimas degradadoras de pared celular, los cuales podrían ser los responsables de la ausencia de patogenicidad en este aislado.

9. El análisis de genes únicos en cada genoma reveló enriquecimiento de genes codificantes de pequeñas proteínas secretadas, las cuales podrían actuar como efectores específicos en determinados cultivos huéspedes (variedades) o representar innovaciones evolutivas involucradas en la adaptación a nuevos ambientes.

10. El análisis del contenido génico reveló que el aislado no patógenico carecía de 35 genes presentes en todos los otros aislados patógenos. Entre ellos, dos

efectores, dos factores de virulencia, una cutinasa y otros seis genes sobreexpresados durante la infección de la cepa silvestre M1.001 en hojas de maíz. La pérdida de estos genes, probablemente fundamentales para el completo desarrollo de la antracnosis, representan excelentes candidatos para el diseño de experimentos funcionales, ya que podrían estar directamente involucrados en el desarrollo de los característicos síntomas de la antracnosis.

11. Se encontraron evidencias de selección natural a nivel molecular en regiones codificantes y no codificantes de genes involucrados en patogenicidad. Como es de esperar bajo el modelo de la Carrera Armamentista, la selección positiva parece ser predominante en genes que codifican proteínas que interactúan con el huésped (efectores, factores de virulencia y metabolitos secundarios). Por otra parte, genes sobreexpresados durante la infección muestran evidencias de selección en las regiones reguladoras, sugiriendo un rol adaptativo de estas secuencias.

12. Las regiones 5'UTR de genes sobreexpresados durante la infección y probablemente implicadas en la regulación de dichos genes, muestran evidencias de selección balanceadora, en concordancia con el modelo de la Reina Roja.

13. Los análisis de selección natural a nivel molecular sugieren que, mientras que las substituciones adaptativas a nivel de secuencias codificantes podrían ser importantes en proteínas que interactúan con el huésped, los polimorfismos en las regiones regulatorias podrían conferir mayor flexibilidad en el proceso de infección mediante la regulación diferencial (en tiempo y lugar) de genes importantes para la patogénesis.

14. Se logró aplicar satisfactoriamente una metodología computacional para la anotación de ARNs no codificantes (ncRNAs) en hongos filamentosos. El análisis de 2.127.338 ESTs, pertenecientes a 46 especies de hongos y Oomicetes, resultó en la identificación de 8.251 ncRNAs, la mayoría de ellos previamente desconocidos.

15. El análisis de los ncRNAs, reveló una pobre conservación a nivel de secuencias y una probable implicación en procesos de diferenciación celular. Si bien se desconoce la importancia funcional de los ncRNAs identificados, la caracterización de estas nuevas secuencias podría abrir vías en el desarrollo de nuevos antifúngicos.

16. Catorce genes expresados diferencialmente en maíz durante la infección con *C. graminicola*, muestran evidencias de selección positiva a lo largo del linaje de las Poáceas. Estos genes podrían estar implicados en la Carrera Armamentista, a través de la interacción con efectores del patógeno.

17. Algunos de los catorce genes bajo selección positiva ya se han descripto anteriormente, pero otros se describen en este trabajo por primera vez. Entre ellos se encuentran genes codificantes de una isocitrato liasa, una deshidrogenasa, un transportador, una proteína con repeticiones ricas en leucina y otros genes con función desconocida.

18. El análisis tri-dimensional de las proteínas bajo selección positiva, mostró que la mayoría de los sitios afectados se encuentran en la superficie protéica, donde es de esperar que interactúen con otras moléculas.

8. BIBLIOGRAPHY

Achawanantakun, R., Sun, Y. & Takyar, S.S., 2011. NcRNA consensus secondary structure derivation using grammar strings. *Journal of bioinformatics and computational biology*, 9(2), pp.317-337.

Achaz, G., 2008. Testing for neutrality in samples with sequencing errors. *Genetics*, 179(3), pp.1409-1424.

Agrios, G.N., 2005. Plant Pathology, Fifth Edition 5.a ed., *Academic Press*.

Aguileta, G., Lengelle, J., Chiapello, H., Giraud, T., Viaud, M., Fournier, E., Rodolphe, F., Marthey, S., Ducasse, A., Gendrault, A., Poulain, J., Wincker, P. & Gout, L., 2012. Genes under positive selection in a model plant pathogenic fungus, Botrytis. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 12(5), pp.987-996.

Aguileta, G., Lengelle, J., Marthey, S., Chiapello, H., Rodolphe, F., Gendrault, A., Yockteng, R., Vercken, E., Devier, B., Fontaine, M.C., Wincker P., Dossat C., Cruaud C., Couloux A. & Giraud T., 2010. Finding candidate genes under positive selection in Non-model species: examples of genes involved in host specialization in pathogens. *Molecular ecology*, 19(2), pp.292–306.

Aguileta, G., Marthey, S., Chiapello, H., Lebrun, M.-H., Rodolphe, F., Fournier, E., Gendrault-Jacquemard, A. & Giraud, T., 2008. Assessing the performance of single-copy genes for recovering robust phylogenies. *Systematic Biology*, 57(4), pp.613 - 627.

Aguileta, G., Refrégier, G., Yockteng, R., Fournier, E. & Giraud, T., 2009. Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infection, Genetics and Evolution*, 9(4), pp.656-670.

Albert, P.R., 2011. What is a functional genetic polymorphism? Defining classes of functionality. *Journal of Psychiatry & Neuroscience*, 36(6), pp.363-365.

Alby, K., Schaefer, D. & Bennett, R.J., 2009. Homothallic and heterothallic mating in the opportunistic pathogen *Candida albicans*. *Nature*, 460(7257), pp.890-893.

Altman, A. & Hasegawa, P.M., 2012. Introduction to plant biotechnology 2011: Basic aspects and agricultural implications. En Arie Altman & Paul Michael Hasegawa, eds. Plant Biotechnology and Agriculture. San Diego: *Academic Press*, pp. xxix-xxxviii.

Amaral, P.P., Dinger, M.E., Mercer, T.R. & Mattick, J.S., 2008. The eukaryotic genome as an RNA machine. *Science*, 319(5871), pp.1787-1789.

Amaral, P.P. & Mattick, J.S., 2008. Noncoding RNA in development. *Mammalian genome*, 19(7-8), pp.454-492.

Ben-Ami, R. & Kontoyiannis, D.P., 2012. Resistance to echinocandins comes at a cost: the impact of FKS1 hotspot mutations on *Candida albicans* fitness and virulence. *Virulence*, 3(1), pp.95-97.

Anantharaman, V. & Aravind, L., 2002. The GOLD domain, a novel protein module involved in Golgi function and secretion. *Genome Biology*, 3(5), pp.1–7.

Anderson, S.F., Steber, C.M., Esposito, R.E. & Coleman, J.E., 1995. UME6, a negative regulator of meiosis in *Saccharomyces cerevisiae*, contains a C-terminal Zn2Cys6 binuclear cluster that binds the URS1 DNA sequence in a zinc-dependent manner. *Protein Science*, 4(9), pp.1832-1843.

Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature*, 437(7062), pp.1149-1152.

Andolfatto, P., 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Research*, 17(12), pp.1755-1762.

Anisimova, M., Bielawski, J.P. & Yang, Z., 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular biology and evolution*, 19(6), pp.950-958.

Anisimova, M., Bielawski, J.P. & Yang, Z., 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular biology and evolution*, 18(8), pp.1585-1592.

Anisimova, M. & Liberles, D., 2012. Detecting and understanding natural selection. En G. Cannarozzi & A. Schneider, eds. Codon Evolution: mechanisms and models. *Oxford University Press*.

Anisimova, M., Nielsen, R. & Yang, Z., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, 164(3), pp.1229-1236.

Anžlovar, S. & Dermastia, M., 2003. The comparative analysis of osmotins and osmotin-like PR-5 proteins. *Plant Biology*, 5(2), pp.116-124.

Araki, R., Fukumura, R., Sasaki, N., Kasama, Y., Suzuki, N., Takahashi, H., Tabata, Y., Saito, T. & Abe, M., 2006. More than 40,000 transcripts, including novel and noncoding transcripts, in mouse embryonic stem cells. *Stem cells*, 24(11), pp.2522-2528.

Arenas, M. & Posada, D., 2010. The effect of recombination on the reconstruction of ancestral sequences. *Genetics*, 184(4), pp.1133-1139.

Armijos Jaramillo, V.D., Vargas, W.A., Sukno, S.A. & Thon, M.R., 2013. Horizontal transfer of a subtilisin gene from plants into an ancestor of the plant pathogenic fungal genus Colletotrichum. *PLoS ONE*, 8(3), p.e59078.

Arrial, R.T., Togawa, R.C. & Brigido, M.M., 2009. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus *Paracoccidioides brasiliensis*. *BMC bioinformatics*, 10(1), p.239.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), pp.25-29.

Atallah, Z.K. & Subbarao, K.V., 2012. Population biology of fungal plant pathogens. En M. D. Bolton & B. P. H. J. Thomma, eds. Plant Fungal Pathogens. *Methods in Molecular Biology*. *Humana Press*, pp. 333-363.

Awadalla, P., 2003. The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics*, 4(1), pp.50-60.

Baek, J.-M. & Kenerley, C.M., 1998. The arg2 gene of *Trichoderma virens*: Cloning and development of a homologous transformation system. *Fungal Genetics and Biology*, 23(1), pp.34-44.

Bairoch, A. & Boeckmann, B., 1994. The SWISS-PROT protein sequence data bank: current status. *Nucleic acids research*, 22(17), pp.3578-3580.

Van Bakel, H., Nislow, C., Blencowe, B.J. & Hughes, T.R., 2010. Most «Dark Matter» transcripts are associated with known genes. *PLoS Biology*, 8(5), p.e1000371.

Barry, P., 2007. Genome 2.0: Mountains of new data are challenging old views. *Science News*, 172(10), pp.154–156.

Basak, S. & Ghosh, T.C., 2006. Temperature adaptation of synonymous codon usage in different functional categories of genes: A comparative study between homologous genes of *Methanococcus jannaschii* and *Methanococcus maripaludis*. *FEBS Letters*, 580(16), pp.3895-3899.

Batalia, M.A., Monzingo, A.F., Ernst, S., Roberts, W. & Robertus, J.D., 1996. The crystal structure of the antifungal protein zeamatin, a member of the thaumatin-like, PR-5 protein family. *Nature Structural Biology*, 3(1), pp.19-23.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), pp.289-300.

Bennett, R.N., Wallsgrove, R.M., Bennett, R.N. & Wallsgrove, R.M., 1994. Secondary metabolites in plant defence mechanisms. *New Phytologist*, 127(4,), pp.617-633.

Berbee, M.L., 2001. The phylogeny of plant and animal pathogens in the Ascomycota. *Physiological and Molecular Plant Pathology*, 59(4), pp.165-187.

Bergstrom, G.C. & Nicholson, R.L., 1999. The biology of corn anthracnose: Knowledge to exploit for improved management. *Plant Disease*, 83(7), pp.596-608.

Besnard, G. & Christin, P.A., 2010. Evolutionary genomics of C4 photosynthesis in grasses requires a large species sampling. *Comptes rendus biologies*, 333(8), pp.577-581.

Bickel, R.D., Dunham, J.P. & Brisson, J.A., 2013. Widespread selection across coding and noncoding DNA in the Pea Aphid genome. *G3: Genes|Genomes|Genetics*, 3(6), pp.993-1001.

Bishop, A.J.R. & Schiestl, R.H., 2000. Homologous recombination as a mechanism for genome rearrangements: environmental and genetic effects. *Human Molecular Genetics*, 9(16), pp.2427-2334.

Bishop, J.G., Dean, A.M. & Mitchell-Olds, T., 2000. Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution. *Proceedings of the National Academy of Sciences*, 97(10), pp.5322-5327.

Biswas, S. & Akey, J., 2006. Genomic insights into positive selection. *Trends in Genetics*, 22(8), pp.437-446.

Blackwell, M., 2011. The Fungi: 1, 2, 3 ... 5.1 million species? *American Journal of Botany*, 98(3), pp.426-438.

Boguski, M.S., Lowe, T.M. & Tolstoshev, C.M., 1993. dbEST--database for «expressed sequence tags». *Nature genetics*, 4(4), pp.332-333.

Boller, T., 1995. Chemoperception of microbial signals in plant cells. *Annual Review of Plant Physiology and Plant Molecular Biology*, 46(1), pp.189-214.

Boni, M.F., Posada, D. & Feldman, M.W., 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*, 176(2), pp.1035-1047.

Böning, K. & Wallner, F., 1936. Fußkrankheit und andere Schädigungen an Mais durch *Colletotrichum graminicolum (Ces.) Wilson*. *Phytopath Ztschr*, 9, pp.99-110.

Bordi, C., Lamy, M.-C., Ventre, I., Termine, E., Hachani, A., Fillet, S., Roche, B., Bleves, S., Méjean, V., Lazdunski, A. & Filloux, A., 2010. Regulatory RNAs and the HptB/RetS signalling pathways fine-tune *Pseudomonas aeruginosa* pathogenesis. *Molecular microbiology*, 76(6), pp.1427-1443.

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M. & Snyder, M., 2007. Divergence of transcription factor binding sites across related yeast species. *Science*, 317(5839), pp.815-819.

Britton, K., Langridge, S., Baker, P.J., Weeradechapon, K., Sedelnikova, S.E., De Lucas, J.R., Rice, D.W. & Turner, G., 2000. The crystal structure and active site location of isocitrate lyase from the fungus *Aspergillus nidulans*. *Structure*, 8(4), pp.349-362.

Broad Fungal Genome Initiative, **Colletotrichum Database**. Available at: http://www.broadinstitute.org/annotation/genome/colletotrichum_group/MultiHome.html.

Brown, D.W., Butchko, R.A.E., Busman, M. & Proctor, R.H., 2007. The *Fusarium verticillioides* FUM gene cluster encodes a Zn(II)2Cys6 protein that affects FUM gene expression and fumonisin production. *Eukaryotic Cell*, 6(7), pp.1210-1218.

Brown, G.D., Denning, D.W. & Levitz, S.M., 2012. Tackling human fungal infections. *Science*, 336(6082), pp.647-647.

Bruen, T.C., Philippe, H. & Bryant, D., 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172(4), pp.2665-2681.

Brunner, P.C., Keller, N. & McDonald, B.A., 2009. Wheat domestication accelerated evolution and triggered positive selection in the β-Xylosidase enzyme of *Mycosphaerella graminicola P. Michalak*. *PLoS ONE*, 4(11), p.e7884.

Bu, D., Yu, K., Sun, S., Xie, C., Skogerbø, G., Miao, R., Xiao, H., Liao, Q., Luo, H., Zhao, G., Zhao, H., Liu, Z., Liu, C., Chen, R. & Zhao, Y., 2012. NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic acids research*, 40(Database issue), pp.D210-215.

Budd, A., 2012. Diversity of Genome Organisation. En M. Anisimova, ed. Evolutionary Genomics. Methods in Molecular Biology. *Humana Press*, pp. 51-76.

Burger, G., Strauss, J., Scazzocchio, C. & Lang, B.F., 1991. nirA, the pathway-specific regulatory gene of nitrate assimilation in *Aspergillus nidulans*, encodes a putative GAL4-type zinc finger protein and contains four introns in highly conserved regions. *Molecular and cellular biology*, 11(11), pp.5746-5755.

Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Todd Hubisz, M., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., Civello, D., Adams, M.D., Cargill, M. & Clark, A.G., 2005. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062), pp.1153-1157.

Cambareri, E.B., Jensen, B.C., Schabtach, E. & Selker, E.U., 1989. Repeat-induced G-C to A-T mutations in Neurospora. *Science*, 244(4912), pp.1571-1575.

Cannon, P.F., Damm, U., Johnston, P.R. & Weir, B.S., 2012. Colletotrichum - current status and future directions. *Studies in Mycology*, 73(1), pp.181-213.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A. & Yandell, M., 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1), pp.188-196.

Caravajal, F. & Edgerton, C.W., 1944. The perfect stage of *Colletotrichum falcatum*. *Phytopathology*, 34(2), pp.206-213 pp.

Cardwell, K.F., 1989. Variation in virulence of *Colletotrichum graminicola (Ces.) Wils.* and competition between monoconidial isolates. *Texas A & M University*.

Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E., Livingston, R.J., Rieder, M.J. & Nickerson, D.A., 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11), pp.1553-1565.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., Allen, J.E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R.N., Bailey, T.L., Bansal, M., Baxter, L., Beisel, K.W., Bersano, T., Bono, H., Chalk, A.M., Chiu, K.P., Choudhary, V., Christoffels, A., Clutterbuck, D.R., Crowe, M.L., Dalla, E., Dalrymple, B.P., Bono, B. de, Gatta, G.D., Bernardo, D. di, Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C.F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T.R., Gojobori, T., Green, R.E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T.K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S.P.T., Kruger, A., Kummerfeld, S.K., Kurochkin, I.V., Lareau, L.F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Babu, M.M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K.C., Pavan, W.J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J.F., Ring, B.Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S.L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. a. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S.L., Tang, S., Taylor, M.S., Tegner, J., Teichmann, S.A., Ueda, H.R., Nimwegen, E. van, Verardo, R., Wei, C.L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S.M., Teasdale, R.D., Liu, E.T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J.S., Hume, D.A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J. & Hayashizaki, Y., 2005. The transcriptional landscape of the mammalian genome. *Science*, 309(5740), pp.1559-1563.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4), pp.540-552.

Catanzariti, A.-M., Dodds, P.N., Lawrence, G.J., Ayliffe, M.A. & Ellis, J.G., 2006. Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *The Plant Cell Online*, 18(1), pp.243 -256.

Cerutti, H. & Casas-Mollano, J.A., 2006. On the origin and functions of RNA-mediated silencing: from protists to man. *Current genetics*, 50(2), pp.81-99.

Cettul, E., Rekab, D., Locci, R. & Firrao, G., 2008. Evolutionary analysis of endopolygalacturonase-encoding genes of *Botrytis cinerea*. *Molecular Plant Pathology*, 9(5), pp.675-685.

Chandrasekaran, C. & Betrán, E., 2008. Origins of new genes and pseudogenes | Learn Science at Scitable. *Nature Education*, 1(1).

Chen, J.-M., Cooper, D.N., Férec, C., Kehrer-Sawatzki, H. & Patrinos, G.P., 2010. Genomic rearrangements in inherited disease and cancer. *Seminars in Cancer Biology*, 20(4), pp.222-233.

Chen, K. & Rajewsky, N., 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nature reviews*. Genetics, 8(2), pp.93-103.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., Shi, X., Fulton, R.S., Ley, T.J., Wilson, R.K., Ding, L. & Mardis, E.R., 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, 6(9), pp.677-681.

Chinni, S.V., Raabe, C.A., Zakaria, R., Randau, G., Hoe, C.H., Zemann, A., Brosius, J., Tang, T.-H. & Rozhdestvensky, T.S., 2010. Experimental identification and characterization of 97 novel npcRNA candidates in *Salmonella enterica serovar Typhi*. *Nucleic acids research*, 38(17), pp.5893-5908.

Coates, J.C., 2003. Armadillo repeat proteins: beyond the animal kingdom. *Trends in Cell Biology*, 13(9), pp.463-471.

Conte, S.S. & Lloyd, A.M., 2011. Exploring multiple drug and herbicide resistance in plants—spotlight on transporter proteins. *Plant Science*, 180(2), pp.196-203.

Costa, F.F., 2008. Non-coding RNAs, epigenetics and complexity. *Gene*, 410(1), pp.9–17.

Costa, F.F., 2005. Non-coding RNAs: new players in eukaryotic biology. *Gene*, 357(2), pp.83–94.

Crouch, J.A. & Beirn, L.A., 2009. Anthracnose of cereals and grasses. *Fungal Diversity*, 39, pp.19-44.

Crouch, J.A., Clarke, B.B. & Hillman, B.I., 2006. Unraveling evolutionary relationships among the divergent lineages of Colletotrichum causing anthracnose disease in turfgrass and corn. *Phytopathology*, 96(1), pp.46-60.

Cuomo, C.A., Güldener, U., Xu, J.-R., Trail, F., Turgeon, B.G., Di Pietro, A., Walton, J.D., Ma, L.-J., Baker, S.E., Rep, M., Adam, G., Antoniw, J., Baldwin, T., Calvo, S., Chang, Y.-L., DeCaprio, D., Gale, L.R., Gnerre, S., Goswami, R.S., Hammond-Kosack, K., Harris, L.J., Hilburn, K., Kennell, J.C., Kroken, S., Magnuson, J.K., Mannhaupt, G., Mauceli, E., Mewes, H.-W., Mitterbauer, R., Muehlbauer, G., Münsterkötter, M., Nelson, D., O'Donnell, K., Ouellet, T., Qi, W., Quesneville, H., Roncero, M.I.G., Seong, K.-Y., Tetko, I.V., Urban, M., Waalwijk, C., Ward, T.J., Yao, J., Birren, B.W. & Kistler, H.C., 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, 317(5843), pp.1400-1402.

Dai, Y., Jia, Y., Correll, J., Wang, X. & Wang, Y., 2010. Diversification and evolution of the avirulence gene AVR-Pita1 in field isolates of *Magnaporthe oryzae*. *Fungal Genetics and Biology,* 47(12), pp.973-980.

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), pp.1394-1403.

Darling, A.E., Mau, B. & Perna, N.T., 2010. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, 5(6), p.e11147.

Daszak, P., Cunningham, A.A. & Hyatt, A.D., 2000. Emerging infectious diseases of wildlife-- threats to biodiversity and human health. *Science*, 287(5452), pp.443-449.

Dean, R., Van Kan, J. a. L., Pretorius, Z.A., Hammond-Kosack, K.E., Di Pietro, A., Spanu, P.D., Rudd, J.J., Dickman, M., Kahmann, R., Ellis, J. & Foster, G.D., 2012. The top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology*, 13(4), pp.414–430.

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.-R., Pan, H., Read, N.D., Lee, Y.-H., Carbone, I., Brown, D., Oh, Y.Y., Donofrio, N., Jeong, J.S., Soanes, D.M., Djonovic, S., Kolomiets, E., Rehmeyer, C., Li, W., Harding, M., Kim, S., Lebrun, M.-H.,

Bohnert, H., Coughlan, S., Butler, J., Calvo, S., Ma, L.-J., Nicol, R., Purcell, S., Nusbaum, C., Galagan, J.E. & Birren, B.W., 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, 434(7036), pp.980-986.

Delport, W., Poon, A.F.Y., Frost, S.D.W. & Kosakovsky Pond, S.L., 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, 26(19), pp.2455-2457.

Der Biezen Van, E.A. & Jones, J.D.G., 1998. Plant disease-resistance proteins and the gene-for-gene concept. *Trends in Biochemical Sciences*, 23(12), pp.454-456.

DeYoung, B.J. & Innes, R.W., 2006. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nature Immunology*, 7(12), pp.1243-1249.

Dixon, R.A. & Lamb, C.J., 1990. Molecular communication in interactions between plants and microbial pathogens. *Annual Review of Plant Physiology and Plant Molecular Biology*, 41(1), pp.339-367.

Dodds, P.N., Lawrence, G.J., Catanzariti, A.-M., Teh, T., Wang, C.-I.A., Ayliffe, M.A., Kobe, B. & Ellis, J.G., 2006. Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proceedings of the National Academy of Sciences*, 103(23), pp.8888 -8893.

Dodds, P.N. & Rathjen, J.P., 2010. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nature Review Genetics*, 11(8), pp.539-548.

Doehlemann, G. & Hemetsberger, C., 2013. Apoplastic immunity and its suppression by filamentous plant pathogens. *New Phytologist*, 198(4), pp.1001–1016.

Doehlemann, G., Wahl, R., Horst, R.J., Voll, L.M., Usadel, B., Poree, F., Stitt, M., Pons-Kühnemann, J., Sonnewald, U., Kahmann, R. & Kämper, J., 2008. Reprogramming a maize plant: transcriptional and metabolic changes induced by the fungal biotroph *Ustilago maydis*. *The Plant journal: for cell and molecular biology*, 56(2), pp.181-195.

Doniger, S.W., Kim, H.S., Swain, D., Corcuera, D., Williams, M., Yang, S.-P. & Fay, J.C., 2008. A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genetics*, 4(8), p.e1000183.

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T. & Wilson, A., 2011. Geneious v5.4. Available at: http://www.geneious.com/.

Dunn, M.F., Ramírez-Trujillo, J.A. & Hernández-Lucas, I., 2009. Major roles of isocitrate lyase and malate synthase in bacterial and fungal pathogenesis. *Microbiology*, 155(10), pp.3166-3175.

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M.D., Lawrence, C.J., Lushbough, C. & Brendel, V., 2008. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Research*, 36(1), pp.D959-D965.

Eckardt, N.A., 2000. Sequencing the rice genome. *The Plant Cell*, 12(11), pp.2011-2018.

Eddy, S.R., 2002. Computational genomics of noncoding RNA genes. *Cell*, 109(2), pp.137-140.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5(1), p.113.

Egner, R., Bauer, B.E. & Kuchler, K., 2002. The transmembrane domain 10 of the yeast Pdr5p ABC antifungal efflux pump determines both substrate specificity and inhibitor susceptibility. *Molecular Microbiology*, 35(5), pp.1255-1263.

Ellis, J.G., Rafiqi, M., Gan, P., Chakrabarti, A. & Dodds, P.N., 2009. Recent progress in discovery and functional analysis of effector proteins of fungal and oomycete plant pathogens. *Current Opinion in Plant Biology*, 12(4), pp.399-405.

Ellwood, S.R., Syme, R.A., Moffat, C.S. & Oliver, R.P., 2012. Evolution of three Pyrenophora cereal pathogens: Recent divergence, speciation and evolution of non-coding DNA. *Fungal Genetics and Biology*, 49(10), pp.825-829.

Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G., 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology*, 300(4), pp.1005-1016.

Emerson, J.J., Hsieh, L.-C., Sung, H.-M., Wang, T.-Y., Huang, C.-J., Lu, H.H.-S., Lu, M.-Y.J., Wu, S.-H. & Li, W.-H., 2010. Natural selection on *cis* and *trans* regulation in yeasts. *Genome Research*, 20(6), pp.826-836.

Emes, R.D. & Yang, Z., 2008. Duplicated paralogous genes subject to positive selection in the genome of *Trypanosoma brucei*. *PLoS ONE*, 3(5), p.e2295.

ENCODE Project Consortium, 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799-816.

Enright, A.J., Van Dongen, S. & Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7), pp.1575 -1584.

FAOSTAT, 2012. Food and Agriculture Organization of the United Nations. Available at: http://faostat.fao.org/.

Farrer, R.A., Henk, D.A., Garner, T.W.J., Balloux, F., Woodhams, D.C. & Fisher, M.C., 2013. Chromosomal copy number variation, selection and uneven rates of recombination reveal cryptic genome diversity linked to pathogenicity. *PLoS Genetics*, 9(8), p.e1003703.

Fay, J.C. & Benavides, J.A., 2005. Hypervariable noncoding sequences in *Saccharomyces cerevisiae*. Genetics, 170(4), pp.1575-1587.

Fedorova, N.D., Khaldi, N., Joardar, V.S., Maiti, R., Amedeo, P., Anderson, M.J., Crabtree, J., Silva, J.C., Badger, J.H., Albarraq, A., Angiuoli, S., Bussey, H., Bowyer, P., Cotty, P.J., Dyer, P.S., Egan, A., Galens, K., Fraser-Liggett, C.M., Haas, B.J., Inman, J.M., Kent, R., Lemieux, S., Malavazi, I., Orvis, J., Roemer, T., Ronning, C.M., Sundaram, J.P., Sutton, G., Turner, G., Venter, J.C., White, O.R., Whitty, B.R., Youngman, P., Wolfe, K.H., Goldman, G.H., Wortman, J.R., Jiang, B., Denning, D.W. & Nierman, W.C., 2008. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genetics*, 4(4), p.e1000046.

Feuk, L., Carson, A.R. & Scherer, S.W., 2006. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2), pp.85-97.

Fisher, M.C., Henk, D.A., Briggs, C.J., Brownstein, J.S., Madoff, L.C., McCraw, S.L. & Gurr, S.J., 2012. Emerging fungal threats to animal, plant and ecosystem health. *Nature*, 484(7393), pp.186-194.

Fitch, W.M., Leiter, J.M., Li, X.Q. & Palese, P., 1991. Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences*, 88(10), pp.4270-4274.

Flor, H., 1942. Inheritance of pathogenicity in *Melmapsora lini*. *Phytopathology*, 32, pp.653-669.

Flor, H., 1955. Host-parasite interaction in flax rust - its genetics and other implications. *Phytopathology*, 45(12), pp.680-685.

Forgey, W.M., Blanco, M.H. & Loegering, W.Q., 1978. Differences in pathological capabilities and host specificity of *Colletotrichum graminicola* on *Zea mays* [maize]. *Plant Disease Reporter*, v. 62(7) p. 573-576.

Fraser, H., 2013. Gene expression drives local adaptation in humans. *Genome Research*, 23(7), pp.1089-1096.

Freeman, S. & Katan, T., 1997. Identification of Colletotrichum species responsible for anthracnose and root necrosis of strawberry in Israel. *Phytopathology*, 87(5), pp.516-521.

Frey, T.J., Weldekidan, T., Colbert, T., Wolters, P.J.C.C. & Hawk, J.A., 2011. Fitness evaluation of Rcg1, a locus that confers Resistance to *Colletotrichum graminicola* (Ces.) G.W. Wils. using near-isogenic maize hybrids. *Crop Science*, 51(4), pp.1551-1563.

Frith, M.C., Pheasant, M. & Mattick, J.S., 2005. The amazing complexity of the human transcriptome. *European journal of human genetics*, 13(8), pp.894-897.

Gaffney, D.J. & Keightley, P.D., 2006. Genomic selective constraints in murid noncoding DNA. *PLoS Genetics*, 2(11), p.e204.

Galagan, J.E., Calvo, S.E., Cuomo, C., Ma, L.-J., Wortman, J.R., Batzoglou, S., Lee, S.-I., Baştürkmen, M., Spevak, C.C., Clutterbuck, J., Kapitonov, V., Jurka, J., Scazzocchio, C., Farman, M., Butler, J., Purcell, S., Harris, S., Braus, G.H., Draht, O., Busch, S., D'Enfert, C., Bouchier, C., Goldman, G.H., Bell-Pedersen, D., Griffiths-Jones, S., Doonan, J.H., Yu, J., Vienken, K., Pain, A., Freitag, M., Selker, E.U., Archer, D.B., Peñalva, M.Á., Oakley, B.R., Momany, M., Tanaka, T., Kumagai, T., Asai, K., Machida, M., Nierman, W.C., Denning, D.W., Caddick, M., Hynes, M., Paoletti, M., Fischer, R., Miller, B., Dyer, P., Sachs, M.S., Osmani, S.A. & Birren, B.W., 2005. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature*, 438(7071), pp.1105-1115.

Galagan, J.E., Henn, M.R., Ma, L.-J., Cuomo, C.A. & Birren, B., 2005. Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Research*, 15(12), pp.1620-1631.

Gasch, A.P., Moses, A.M., Chiang, D.Y., Fraser, H.B., Berardini, M. & Eisen, M.B., 2004. Conservation and evolution of cis-regulatory systems in Ascomycete fungi. *PLoS Biol*ogy, 2(12), p.e398.

Gazave, E., Marqués-Bonet, T., Fernando, O., Charlesworth, B. & Navarro, A., 2007. Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biology*, 8(2), p.R21.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S. & Snyder, M., 2007. What is a gene post-ENCODE? History and updated definition. *Genome Research*, 17(6), pp.669-681.

Gibbons, J.G., Salichos, L., Slot, J.C., Rinker, D.C., McGary, K.L., King, J.G., Klich, M.A., Tabb, D.L., McDonald, W.H. & Rokas, A., 2012. The evolutionary imprint of domestication on genome variation and function of the filamentous fungus *Aspergillus oryzae*. *Current Biology*, 22(15), pp.1403-1409.

Gil, M., Zanetti, M.S., Zoller, S. & Anisimova, M., 2013. CodonPhyML: Fast maximum likelihood phylogeny estimation under codon substitution models. *Molecular Biology and Evolution*, 30(6), pp.:1270-1280.

Girard, A., Sachidanandam, R., Hannon, G.J. & Carmell, M.A., 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 442(7099), pp.199-202.

Glass, N.L., Jacobson, D.J. & Shiu, P.K., 2000. The genetics of hyphal fusion and vegetative incompatibility in filamentous ascomycete fungi. *Annual Review of Genetics*, 34, pp.165-186.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M. & Toulmin, C., 2010. Food Security: The challenge of feeding 9 billion people. *Science*, 327(5967), pp.812-818.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. & Rokhsar, D.S., 2011. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1), pp.D1178-D1186.

Gottesman, S., 2002. Stealth regulation: biological circuits with small RNA switches. *Genes & development*, 16(22), pp.2829-2842.

Gough, J., Karplus, K., Hughey, R. & Chothia, C., 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4), pp.903-919.

Gowda, M., Nunes, C.C., Sailsbery, J., Xue, M., Chen, F., Nelson, C.A., Brown, D.E., Oh, Y., Meng, S., Mitchell, T., Hagedorn, C.H. & Dean, R.A., 2010. Genome-wide characterization of methylguanosine-capped and polyadenylated small RNAs in the rice blast fungus *Magnaporthe oryzae*. *Nucleic Acids Research*, 38(21), pp.7558-7569.

Graham, N. & May, S., 2011. Bioinformatics resources for *Arabidopsis thaliana*. En R. Schmidt & I. Bancroft, eds. Genetics and Genomics of the Brassicaceae. *Plant Genetics and Genomics*: Crops and Models. Springer New York, pp. 585-596.

Greenwood, T.A. & Kelsoe, J.R., 2003. Promoter and intronic variants affect the transcriptional regulation of the human dopamine transporter gene. *Genomics*, 82(5), pp.511-520.

Griffiths-Jones, S., 2005. Annotating non-coding RNAs with Rfam. *Current Protocols in Bioinformatics*, 9, pp.12.5.1–12.5.12.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3), pp.307-321.

Guindon, S. & Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5), pp.696-704.

Gupta, P.K., Balyan, H.S. & Varshney, R.K., 2010. Quantitative genetics and plant genomics: an overview. *Molecular Breeding*, 26(2), pp.133-134.

Haddrill, P.R., Bachtrog, D. & Andolfatto, P., 2008. Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Molecular Biology and Evolution*, 25(9), pp.1825-1834.

Hahn, M.W., 2007. Detecting natural selection on cis-regulatory DNA. *Genetica*, 129(1), pp.7-18.

Hamilton, A.J. & Baulcombe, D.C., 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286(5441), pp.950-952.

Hane, J.K., Rouxel, T., Howlett, B.J., Kema, G.H., Goodwin, S.B. & Oliver, R.P., 2011. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biology*, 12(5), p.R45.

Haridas, S., Breuill, C., Bohlmann, J. & Hsiang, T., 2011. A biologist's guide to de novo genome assembly using next-generation sequence data: A test with fungal genomes. *Journal of microbiological methods*, 86(3), pp.368-375.

Hastie, A.C., 1964. The parasexual cycle in *Verticillium albo-atrum*. *Genetics Research*, 5(02), pp.305-315.

Havilio, M., Levanon, E.Y., Lerman, G., Kupiec, M. & Eisenberg, E., 2005. Evidence for abundant transcription of non-coding regions in the *Saccharomyces cerevisiae* genome. *BMC Genomics*, 6(1), p.93.

Haygood, R., Babbitt, C.C., Fedrigo, O. & Wray, G.A., 2010. Contrasts between adaptive coding and noncoding changes during human evolution. *Proceedings of the National Academy of Sciences*, 107(17), pp.7853-7857.

Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D. & Wray, G.A., 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genetics*, 39(9), pp.1140-1144.

Heitman, J., 2006. Sexual reproduction and the evolution of microbial pathogens. *Current Biology*, 16(17), pp.711-725.

Le Hir, H., Nott, A. & Moore, M.J., 2003. How introns influence and enhance eukaryotic gene expression. *Trends in Biochemical Sciences*, 28(4), pp.215-220.

Hogenhout, S.A., Van der Hoorn, R.A.L., Terauchi, R. & Kamoun, S., 2009. Emerging concepts in effector biology of plant-associated organisms. *Molecular Plant-Microbe Interactions*, 22(2), pp.115-122.

Holt, C. & Yandell, M., 2011. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, 12(1), p.491.

Horst, R.J., Doehlemann, G., Wahl, R., Hofmann, J., Schmiedl, A., Kahmann, R., Kämper, J., Sonnewald, U. & Voll, L.M., 2010. *Ustilago maydis* infection strongly alters organic nitrogen allocation in maize and stimulates productivity of systemic source leaves. *Plant Physiology*, 152(1), pp.293-308.

Hranilovic, D., Stefulj, J., Schwab, S., Borrmann-Hassenbach, M., Albus, M., Jernej, B. & Wildenauer, D., 2004. Serotonin transporter promoter and intron 2 polymorphisms: relationship between allelic variants and gene expression. *Biological Psychiatry*, 55(11), pp.1090-1094.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., Attardi, L.D., Regev, A., Lander, E.S., Jacks, T. & Rinn, J.L., 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142(3), pp.409-419.

Hubert, B., Rosegrant, M., van Boekel, M.A.J.S. & Ortiz, R., 2010. The future of food: scenarios for 2050. *Crop Science*, 50(1), pp.S-33-S-50.

Hudson, N.J., Gu, Q., Nagaraj, S.H., Ding, Y.-S., Dalrymple, B.P. & Reverter, A., 2011. Eukaryotic evolutionary transitions are associated with extreme codon bias in functionally-related proteins. *PLoS ONE*, 6(9), p.e25457.

Hudson, R.R. & Kaplan, N.L., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1), pp.147-164.

Hughes, A.L., 2012. Evolution of adaptive phenotypic traits without positive Darwinian selection. *Heredity*, 108(4), pp.347-353.

Hüttenhofer, A., Brosius, J. & Bachellerie, J.P., 2002. RNomics: identification and function of small, non-messenger RNAs. *Current Opinion in Chemical Biology*, 6(6), pp.835-843.

Hüttenhofer, A., Schattner, P. & Polacek, N., 2005. Non-coding RNAs: hope or hype? *Trends in Genetics*, 21(5), pp.289-297.

Hutter, S., Vilella, A.J. & Rozas, J., 2006. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, 7(1), p.409.

Hyde, K.D., Cai, L., Cannon, P.F., Crouch, J.A., Crous, P.W., Damm, U., Goodwin, P.H., Chen, H., Johnston, P.R., Jones, E.B.G., Lui, Z.Y., McKenzie, E.H.C., Moriwaki, J., Noireung, P., Pennycook, S.R., Pfenning, L.H., Prihastuti, H., Sato, H., Shivas, R.G., Tan, Y.P., Taylor, P.J.W., Weir, B.S., Yang, Y.L. & Zhang, J.Z., 2009. Colletotrichum – names in current use. *Fungal Diversity*, 39, pp.147-182.

IITA, 2013. International Institute of Tropical Agriculture Report 2013. Available at: http://www.iita.org/maize.

Ikeda, K., Nakayashiki, H., Kataoka, T., Tamba, H., Hashimoto, Y., Tosa, Y. & Mayama, S., 2002. Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context. *Molecular Microbiology*, 45(5), pp.1355-1364.

Iken, J.E. & Amusa, N.A., 2005. Maize research and production in Nigeria. *African Journal of Biotechnology*, 3(6), pp.302-307.

Initiative TAG, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814), pp.796-815.

IRGSP (International Rice Genome Sequencing Project), 2005. The map-based sequence of the rice genome. *Nature*, 436(7052), pp.793-800.

Jacquier, A., 2009. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Reviews Genetics*, 10(12), pp.833-844.

Jamil, F.F. & Nicholson, R.L., 1991. Response of sorghum lines of different ages to *Colletotrichum graminicola* isolates from shattercane, sorghum and corn. *Pakistan Journal of Phytopathology*, 3(1-2), pp.12-18.

Janeway, C.A., 1989. Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harbor Symposia on Quantitative Biology*, 54, pp.1-13.

Janeway, C.A., Jr & Medzhitov, R., 2002. Innate immune recognition. *Annual Review of Immunology*, 20, pp.197-216.

Jeon, J., Choi, J., Lee, G.-W., Dean, R.A. & Lee, Y.-H., 2013. Experimental evolution reveals genome-wide spectrum and dynamics of mutations in the rice blast fungus, *Magnaporthe oryzae*. *PLoS ONE*, 8(5), p.e65416.

Jochl, C., Rederstorff, M., Hertel, J., Stadler, P.F., Hofacker, I.L., Schrettl, M., Haas, H. & Huttenhofer, A., 2008. Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein synthesis. *Nucleic Acids Research*, 36(8), pp.2677-2689.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L. & Daszak, P., 2008. Global trends in emerging infectious diseases. *Nature*, 451(7181), pp.990-993.

Jones, N., 2013. Planetary disasters: It could happen one night. *Nature*, 493(7431), pp.154-156.

De Jonge, R., 2012. In silico identification and characterization of effector catalogs. En M. D. Bolton & B. P. H. J. Thomma, eds. Plant Fungal Pathogens. Methods in Molecular Biology. *Humana Press*, pp. 415-425.

De Jonge, R., Bolton, M.D., Kombrink, A., Berg, G.C.M. van den, Yadeta, K.A. & Thomma, B.P.H.J., 2013. Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Research*, 23(8), pp.1271-1282.

De Jonge, R., Bolton, M.D. & Thomma, B.P., 2011. How filamentous pathogens co-opt plants: the ins and outs of fungal effectors. *Current Opinion in Plant Biology*, 14(4), pp.400-406.

Joshee, S., Paulus, B.C., Park, D. & Johnston, P.R., 2009. Diversity and distribution of fungal foliar endophytes in New Zealand Podocarpaceae. *Mycological Research*, 113(9), pp.1003-1015.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. & Walichiewicz, J., 2005. Repbaseu, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110(1-4), pp.462-467.

Kajava, A.V. & Kobe, B., 2002. Assessment of the ability to model proteins with leucine-rich repeats in light of the latest structural information. *Protein Science*, 11(5), pp.1082-1090.

Kamenidou, S., Jain, R., Hari, K., Robertson, J.M. & Fletcher, J., 2013. The Microbial Rosetta Stone Central Agricultural Database: An information resource on high-consequence plant pathogens. *Plant Disease*, 97(8), pp.1097-1102.

Kamoun, S., 2007. Groovy times: filamentous pathogen effectors revealed. *Current Opinion in Plant Biology*, 10(4), pp.358-365.

Kearsey, S.E. & Labib, K., 1998. MCM proteins: evolution, properties, and role in DNA replication. *Biochimica et biophysica acta*, 1398(2), pp.113-136.

Keightley, P.D., Lercher, M.J. & Eyre-Walker, A., 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol*, 3(2), p.e42.

Keller, N.P., Turner, G. & Bennett, J.W., 2005. Fungal secondary metabolism - from biochemistry to genomics. *Nature Reviews Microbiology*, 3(12), pp.937-947.

Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W. & Akey, J.M., 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, 16(8), pp.980-989.

Kelley, L.A. & Sternberg, M.J.E., 2009. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols*, 4(3), pp.363-371.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4), p.R36.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D. & Nordborg, M., 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, 39(9), pp.1151-1155.

Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4), pp.893-903.

King, M.C. & Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184), pp.107-116.

Kistler, H.C. & Miao, V.P.W., 1992. New modes of genetic change in filamentous fungi. *Annual Review of Phytopathology*, 30(1), pp.131-153.

Korf, I., 2004. Gene finding in novel genomes. *BMC Bioinformatics*, 5(1), p.59.

Kosakovsky Pond, S.L., Frost, S.D.W. & Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5), pp.676 -679.

Kosiol, C. & Anisimova, M., 2012. Selection on the Protein-Coding Genome. En M. Anisimova, ed. Evolutionary Genomics. Methods in Molecular Biology. *Humana Press*, pp. 113-140.

Kosiol, C., Vinař, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R. & Siepel, A., 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genetics,* 4(8), p.e1000144.

Kousathanas, A., Oliver, F., Halligan, D.L. & Keightley, P.D., 2011. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Molecular Biology and Evolution*, 28(3), pp.1183-1191.

Kozomara, A. & Griffiths-Jones, S., 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(1), pp.D152-D157.

Kretzner, L., Krol, A. & Rosbash, M., 1990. *Saccharomyces cerevisiae* U1 small nuclear RNA secondary structure contains both universal and yeast-specific domains. *Proceedings of the National Academy of Sciences*, 87(2), pp.851-855.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3), pp.567-580.

Kubicek, C., Baker, S., Gamauf, C., Kenerley, C. & Druzhinina, I., 2008. Purifying selection and birth-and-death evolution in the class II hydrophobin gene families of the ascomycete Trichoderma/Hypocrea. *BMC Evolutionary Biology*, 8(1), p.4.

Kung, J.T.Y., Colognori, D. & Lee, J.T., 2013. Long noncoding RNAs: past, present, and future. *Genetics*, 193(3), pp.651-669.

Lafontaine, D.L. & Tollervey, D., 1998. Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. *Trends in Biochemical Sciences*, 23(10), pp.383-388.

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.

Latunde-Dada, A.O., 2001. Colletotrichum: tales of forcible entry, stealth, transient confinement and breakout. *Molecular Plant Pathology*, 2(4), pp.187-198.

Lawrie, D.S., Messer, P.W., Hershberg, R. & Petrov, D.A., 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet*, 9(5), p.e1003527.

Lawton, M.A. & Lamb, C.J., 1987. Transcriptional activation of plant defense genes by fungal elicitor, wounding, and infection. *Molecular and Cellular Biology*, 7(1), pp.335-341.

LeBeau, F.J., 1950. Pathogenicity studies with Colletotrichum from different hosts on sorghum and sugar cane. *Phytopathology*, 40(5), pp.430-438 pp.

Lee, H.-C., Chang, S.-S., Choudhary, S., Aalto, A.P., Maiti, M., Bamford, D.H. & Liu, Y., 2009. qiRNA is a new type of small interfering RNA induced by DNA damage. *Nature*, 459(7244), pp.274-277.

Lee, H.-C., Li, L., Gu, W., Xue, Z., Crosthwaite, S.K., Pertsemlidis, A., Lewis, Z.A., Freitag, M., Selker, E.U., Mello, C.C. & Liu, Y., 2010. Diverse pathways generate microRNA-like RNAs and Dicer-independent small interfering RNAs in fungi. *Molecular Cell*, 38(6), pp.803-814.

Lefebure, T. & Stanhope, M.J., 2009. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus Campylobacter. *Genome Research*, 19(7), pp.1224-1232.

Lemey, P. & Posada, D., 2009. Introduction to recombination detection. En A. M. Vandamme, M. Salemi, & P. Lemey, eds. The Phylogenetic Handbook. *Cambridge University Press*, pp. 493-518.

Levasseur, A., Saloheimo, M., Navarro, D., Andberg, M., Pontarotti, P., Kruus, K. & Record, E., 2010. Exploring laccase-like multicopper oxidase genes from the ascomycete *Trichoderma reesei*: a functional, phylogenetic and evolutionary study. *BMC Biochemistry*, 11, p.32.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078-2079.

Li, H., Ruan, J. & Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), pp.1851-1858.

Li, R., Li, Y., Kristiansen, K. & Wang, J., 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5), pp.713-714.

Li, X., Xia, B., Jiang, Y., Wu, Q., Wang, C., He, L., Peng, F. & Wang, R., 2010. A new pathogenesis-related protein, LrPR4, from *Lycoris radiata*, and its antifungal activity against *Magnaporthe grisea*. *Molecular Biology Reports*, 37(2), pp.995-1001.

Li, Y.-D., Liang, H., Gu, Z., Lin, Z., Guan, W., Zhou, L., Li, Y.-Q. & Li, W.-H., 2009. Detecting positive selection in the budding yeast genome. *Journal of Evolutionary Biology*, 22(12), pp.2430–2437.

Liu, Q., Dou, S., Ji, Z. & Xue, Q., 2005. Synonymous codon usage and gene function are strongly related in *Oryza sativa*. *Biosystems*, 80(2), pp.123-131.

Liu, X. & Fu, Y.-X., 2008. Algorithms to estimate the lower bounds of recombination with or without recurrent mutations. *BMC Genomics*, 9(1), p.S24.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. & Borodovsky, M., 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33(20), p.6494.

Loo, T.W. & Clarke, D.M., 1994. Mutations to amino acids located in predicted transmembrane segment 6 (TM6) modulate the activity and substrate specificity of human P-glycoprotein. *Biochemistry*, 33(47), pp.14049-14057.

Lowe, T.M. & Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5), pp.955-964.

Lu, G., Cannon, P.F., Reid, A. & Simmons, C.M., 2004. Diversity and molecular relationships of endophytic Colletotrichum isolates from the Iwokrama Forest Reserve, Guyana. *Mycological Research*, 108(1), pp.53-63.

Lu, T., Yao, B. & Zhang, C., 2012. DFVF: database of fungal virulence factors. *Database: The Journal of Biological Databases and Curation* 2012 (bas032).

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Yunjie, Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Yong, Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, Jian, Lam, T.-W. & Wang, J., 2012. SOAPdenovo2:

an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1), p.18.

Ma, L.-J., van der Does, H.C., Borkovich, K.A., Coleman, J.J., Daboussi, M.-J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., Houterman, P.M., Kang, S., Shim, W.-B., Woloshuk, C., Xie, X., Xu, J.-R., Antoniw, J., Baker, S.E., Bluhm, B.H., Breakspear, A., Brown, D.W., Butchko, R.A.E., Chapman, S., Coulson, R., Coutinho, P.M., Danchin, E.G.J., Diener, A., Gale, L.R., Gardiner, D.M., Goff, S., Hammond-Kosack, K.E., Hilburn, K., Hua-Van, A., Jonkers, W., Kazan, K., Kodira, C.D., Koehrsen, M., Kumar, L., Lee, Y.-H., Li, L., Manners, J.M., Miranda-Saavedra, D., Mukherjee, M., Park, G., Park, J., Park, S.-Y., Proctor, R.H., Regev, A., Ruiz-Roldan, M.C., Sain, D., Sakthikumar, S., Sykes, S., Schwartz, D.C., Turgeon, B.G., Wapinski, I., Yoder, O., Young, S., Zeng, Q., Zhou, S., Galagan, J., Cuomo, C.A., Kistler, H.C. & Rep, M., 2010. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature*, 464(7287), pp.367-373.

Ma, X., Rogacheva, M.V., Nishant, K.T., Zanders, S., Bustamante, C.D. & Alani, E., 2012. Mutation hot spots in yeast caused by long-range clustering of homopolymeric sequences. *Cell Reports*, 1(1), pp.36-42.

MacKenzie, A. & Quinn, J., 1999. A serotonin transporter gene intron 2 polymorphic region, correlated with affective disorders, has allele-dependent differential enhancer-like properties in the mouse embryo. *Proceedings of the National Academy of Sciences*, 96(26), pp.15251-15255.

Manning, V.A., Pandelova, I., Dhillon, B., Wilhelm, L.J., Goodwin, S.B., Berlin, A.M., Figueroa, M., Freitag, M., Hane, J.K., Henrissat, B., Holman, W.H., Kodira, C.D., Martin, J., Oliver, R.P., Robbertse, B., Schackwitz, W., Schwartz, D.C., Spatafora, J.W., Turgeon, B.G., Yandava, C., Young, S., Zhou, S., Zeng, Q., Grigoriev, I.V., Ma, L.-J. & Ciuffetti, L.M., 2013. Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. *G3: Genes|Genomes|Genetics*, 3(1), pp.41-63.

Maor, R. & Shirasu, K., 2005. The arms race continues: battle strategies between plants and fungal pathogens. *Current Opinion in Microbiology*, 8(4), pp.399-404.

Martin, D.P., Lemey, P. & Posada, D., 2011. Analysing recombination in nucleotide sequences. *Molecular Ecology Resources*, 11(6), pp.943–955.

Martin, G.B., Brommonschenkel, S.H., Chunwongse, J., Frary, A., Ganal, M.W., Spivey, R., Wu, T., Earle, E.D. & Tanksley, S.D., 1993. Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science*, 262(5138), pp.1432-1436.

Martin, J.A. & Wang, Z., 2011. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10), pp.671-682.

Martin, S.H., Wingfield, B.D., Wingfield, M.J. & Steenkamp, E.T., 2011. Structure and evolution of the Fusarium mating type locus: New insights from the *Gibberella fujikuroi* complex. *Fungal Genetics and Biology*, 8(7), pp.731-740.

Matrajt, M., 2010. Non-coding RNA in apicomplexan parasites. *Molecular and Biochemical Parasitology*, 174(1), pp.1–7.

Mattick, J.S., 2007. A new paradigm for developmental biology. *The Journal of Experimental Biology*, 210(9), pp.1526-1547.

Mattick, J.S., 2001. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Reports*, 2(11), pp.986-991.

Mattick, J.S., 2004. RNA regulation: a new genetics? *Nature Reviews Genetics*, 5(4), pp.316-323.

Mazumder, B., Seshadri, V. & Fox, P.L., 2003. Translational control by the 3′-UTR: the ends specify the means. *Trends in Biochemical Sciences*, 28(2), pp.91-98.

McCarthy, F.M., Gresham, C.R., Buza, T.J., Chouvarine, P., Pillai, L.R., Kumar, R., Ozkan, S., Wang, H., Manda, P., Arick, T., Bridges, S.M. & Burgess, S.C., 2011. AgBase: supporting functional modeling in agricultural organisms. *Nucleic Acids Research*, 39(1), pp.D497-D506.

McCarthy, F.M., Wang, N., Magee, G.B., Nanduri, B., Lawrence, M.L., Camon, E.B., Barrell, D.G., Hill, D.P., Dolan, M.E., Williams, W.P., Luthe, D.S., Bridges, S.M. & Burgess, S.C., 2006. AgBase: a functional genomics resource for agriculture. *BMC Genomics*, 7(1), p.229.

McDonald, J.H. & Kreitman, M., 1991. Adaptive protein evolution at the *Adh* locus in Drosophila. *Nature*, 351(6328), pp.652-654.

Medvedev, P., Stanciu, M. & Brudno, M., 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11s), pp.S13-S20.

Mehboob-ur-Rahman & Paterson, A.H., 2009. Comparative Genomics in Crop Plants. En S. M. Jain & D. S. Brar, eds. Molecular Techniques in Crop Improvement. *Springer Netherlands*, pp. 23-61..

Méndez-Vigo, B., Rodríguez-Suárez, C., Pañeda, A., Ferreira, J.J. & Giraldez, R., 2005. Molecular markers and allelic relationships of anthracnose resistance gene cluster B4 in common bean. *Euphytica*, 141(3), pp.237-245.

Mercer, T.R., Dinger, M.E. & Mattick, J.S., 2009. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, 10(3), pp.155–159.

Mercer, T.R., Wilhelm, D., Dinger, M.E., Soldà, G., Korbie, D.J., Glazov, E.A., Truong, V., Schwenke, M., Simons, C., Matthaei, K.I., Saint, R., Koopman, P. & Mattick, J.S., 2011. Expression of distinct RNAs from 3' untranslated regions. *Nucleic Acids Research*, 39(6), pp.2393-2403.

Merritt, C., Rasoloson, D., Ko, D. & Seydoux, G., 2008. 3' UTRs are the primary regulators of gene expression in the *C. elegans* germline. *Current Biology*, 18(19), pp.1476-1482.

Van der Merwe, M.M., Kinnear, M.W., Barrett, L.G., Dodds, P.N., Ericson, L., Thrall, P.H. & Burdon, J.J., 2009. Positive selection in AvrP4 avirulence gene homologues across the genus Melampsora. *Proceedings. Biological Sciences / The Royal Society*, 276(1669), pp.2913-2922.

Messiaen, C.M., Lapon, R. & Molot, P., 1960. Root necroses, stalle rots, and parasitic lodging of maize. *Annual Epiphyte*, 10(4), pp.441-474.

Michod, R.E., Bernstein, H. & Nedelcu, A.M., 2008. Adaptive value of sex in microbial pathogens. *Infection, Genetics and Evolution*, 8(3), pp.267-285.

Miller, J.R., Koren, S. & Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), pp.315-327.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L.T., Salzberg, S.L., Feng, L., Jones, M.R., Skelton, R.L., Murray, J.E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R.E., Michael, T.P., Wall, K., Rice, D.W., Albert, H., Wang, M.-L., Zhu, Y.J., Schatz, M., Nagarajan, N., Acob, R.A., Guan, P., Blas, A., Wai, C.M., Ackerman, C.M., Ren, Y., Liu, C., Wang, Jianmei, Wang,

Jianping, Na, J.-K., Shakirov, E.V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J.E., Gschwend, A.R., Delcher, A.L., Singh, R., Suzuki, J.Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Pérez, R., Torres, M.J., Feltus, F.A., Porter, B., Li, Y., Burroughs, A.M., Luo, M.-C., Liu, L., Christopher, D.A., Mount, S.M., Moore, P.H., Sugimura, T., Jiang, J., Schuler, M.A., Friedman, V., Mitchell-Olds, T., Shippen, D.E., dePamphilis, C.W., Palmer, J.D., Freeling, M., Paterson, A.H., Gonsalves, D., Wang, L. & Alam, M., 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, 452(7190), pp.991-996.

Miyara, I., Shnaiderman, C., Meng, X., Vargas, W.A., Díaz-Mínguez, J.M., Thon, M., Sherman, A. & Prusky, D.B., 2012. Role of nitrogen-metabolism genes expressed during pathogenicity of the alkalinizing *Colletotrichum gloeosporioides* and their differential expression in acidifying pathogens. *Molecular Plant-Microbe interactions*, 25(9), pp.1251-1263.

Mizrachi, I., 2007. GenBank: The Nucleotide Sequence Database. *The NCBI Handbook*. Available at: http://www.ncbi.nlm.nih.gov/books/NBK21105/#ch1.History.

Moir, D., Stewart, S.E., Osmond, B.C. & Botstein, D., 1982. Cold-sensitive cell-division-cycle mutants of yeast: isolation, properties, and pseudoreversion studies. *Genetics*, 100(4), pp.547-563.

Mondragón-Palomino, M., Meyers, B.C., Michelmore, R.W. & Gaut, B.S., 2002. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Research*, 12(9), pp.1305-1315.

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K. & Kosakovsky Pond, S.L., 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, 8(7), p.e1002764.

Neafsey, D.E., Barker, B.M., Sharpton, T.J., Stajich, J.E., Park, D.J., Whiston, E., Hung, C.-Y., McMahan, C., White, J., Sykes, S., Heiman, D., Young, S., Zeng, Q., Abouelleil, A., Aftuck, L., Bessette, D., Brown, A., FitzGerald, M., Lui, A., Macdonald, J.P., Priest, M., Orbach, M.J., Galgiani, J.N., Kirkland, T.N., Cole, G.T., Birren, B.W., Henn, M.R., Taylor, J.W. & Rounsley, S.D., 2010. Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control. *Genome Research*, 20(7), pp.938-946.

Needleman, S.B. & Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology,* 48(3), pp.443-453.

Nei, M. & Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5), pp.418 -426.

Nei, M., Gu, X. & Sitnikova, T., 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences*, 94(15), pp.7799 -7806.

Nunes, C.C., Gowda, M., Sailsbery, J., Xue, M., Chen, F., Brown, D.E., Oh, Y., Mitchell, T.K. & Dean, R.A., 2011. Diverse and tissue-enriched small RNAs in the plant pathogenic fungus, *Magnaporthe oryzae*. *BMC Genomics*, 12(1), p.288.

O'Brien, H.E., Parrent, J.L., Jackson, J.A., Moncalvo, J.-M. & Vilgalys, R., 2005. Fungal community analysis by large-scale sequencing of environmental samples. *Applied and Environmental Microbiology*, 71(9), pp.5544-5550.

O'Connell, R.J., Thon, M.R., Hacquard, S., Amyotte, S.G., Kleemann, J., Torres, M.F., Damm, U., Buiate, E.A., Epstein, L., Alkan, N., Altmüller, J., Alvarado-Balderrama, L., Bauser, C.A., Becker, C., Birren, B.W., Chen, Z., Choi, J.,

Crouch, J.A., Duvick, J.P., Farman, M.A., Gan, P., Heiman, D., Henrissat, B., Howard, R.J., Kabbage, M., Koch, C., Kracher, B., Kubo, Y., Law, A.D., Lebrun, M.-H., Lee, Y.-H., Miyara, I., Moore, N., Neumann, U., Nordström, K., Panaccione, D.G., Panstruga, R., Place, M., Proctor, R.H., Prusky, D., Rech, G., Reinhardt, R., Rollins, J.A., Rounsley, S., Schardl, C.L., Schwartz, D.C., Shenoy, N., Shirasu, K., Sikhakolli, U.R., Stüber, K., Sukno, S.A., Sweigard, J.A., Takano, Y., Takahara, H., Trail, F., Does, H.C. van der, Voll, L.M., Will, I., Young, S., Zeng, Q., Zhang, J., Zhou, S., Dickman, M.B., Schulze-Lefert, P., Themaat, E.V.L. van, Ma, L.-J. & Vaillancourt, L.J., 2012. Lifestyle transitions in plant pathogenic Colletotrichum fungi deciphered by genome and transcriptome analyses. *Nature Genetics*, 44(9), pp.1060-1065.

Oh, S.-K., Young, C., Lee, M., Oliva, R., Bozkurt, T.O., Cano, L.M., Win, J., Bos, J.I.B., Liu, H.-Y., Damme, M. van, Morgan, W., Choi, D., Vossen, E.A.G.V. der, Vleeshouwers, V.G.A.A. & Kamoun, S., 2009. *In planta* expression screens of *Phytophthora infestans* RXLR effectors reveal diverse phenotypes, including activation of the *Solanum bulbocastanum* disease resistance protein Rpi-blb2. *The Plant Cell*, 21(9), pp.2928-2947.

Ohm, R.A., Feau, N., Henrissat, B., Schoch, C.L., Horwitz, B.A., Barry, K.W., Condon, B.J., Copeland, A.C., Dhillon, B., Glaser, F., Hesse, C.N., Kosti, I., LaButti, K., Lindquist, E.A., Lucas, S., Salamov, A.A., Bradshaw, R.E., Ciuffetti, L., Hamelin, R.C., Kema, G.H.J., Lawrence, C., Scott, J.A., Spatafora, J.W., Turgeon, B.G., de Wit, P.J.G.M., Zhong, S., Goodwin, S.B. & Grigoriev, I.V., 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathogens*, 8(12), p.e1003037.

Oleksiak, M.F., Churchill, G.A. & Crawford, D.L., 2002. Variation in gene expression within and among natural populations. *Nature Genetics*, 32(2), pp.261-266.

Ørom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R. & Shiekhattar, R., 2010. Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143(1), pp.46-58.

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M.A., Smirnova, T., Nosrat, B., Markowitz, V.M. & Kyrpides, N.C., 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 40(Database issue), pp.D571-579.

Palaversic, B., Jukic, M., Buhinicek, I., Vragolovic, A. & Kozic, Z., 2009. Breeding maize for resistance to stalk anthracnose. *Maydica*, 54(2/3), pp.229-232.

Pang, K.C., Frith, M.C. & Mattick, J.S., 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics*, 22(1), pp.1-5.

Paoletti, M., Saupe, S.J. & Clavé, C., 2007. Genesis of a fungal non-self recognition repertoire. *PLoS ONE*, 2(3), p.e283.

Paradkar A.S., Jensen S.E. & Mosher R.H., 1997. Comparative genetics and molecular biology of β-lactam biosynthesis. En Biotechnology of antibiotics, 2nd ed., revised and expanded. New York N.Y.: *Informa Healthcare*, pp. 241–277.

Parbery, D., 1996. Trophism and the ecology of fungi associated with plants. *Bological Reviews of the Cambridge Philosophical Society*, 71(3), p.473.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A.K., Chapman, J., Feltus, F.A., Gowik, U., Grigoriev, I.V., Lyons, E., Maher, C.A., Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y.,

Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., McCann, M.C., Ming, R., Peterson, D.G., Mehboob-ur-Rahman, Ware, D., Westhoff, P., Mayer, K.F.X., Messing, J. & Rokhsar, D.S., 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457(7229), pp.551-556.

Pedersen, C., Themaat, E.V.L. van, McGuffin, L.J., Abbott, J.C., Burgis, T.A., Barton, G., Bindschedler, L.V., Lu, X., Maekawa, T., Weßling, R., Cramer, R., Thordal-Christensen, H., Panstruga, R. & Spanu, P.D., 2012. Structure and evolution of barley powdery mildew effector candidates. *BMC Genomics*, 13(1), p.694.

Peng, H., Han, S., Luo, M., Gao, J., Liu, X. & Zhao, M., 2011. Roles of multidrug transporters of MFS in plant stress responses. *International Journal of Bioscience, Biochemistry and Bioinformatics*, 1(2), pp.109-113.

Peng, Y., Leung, H.C.M., Yiu, S.M. & Chin, F.Y.L., 2010. IDBA – A practical iterative de Bruijn Graph *de novo* assembler. En B. Berger, ed. Research in Computational Molecular Biology. Lecture Notes in Computer Science. *Springer Berlin Heidelberg*, pp. 426-440.

Perfect, S.E., Hughes, H.B., O'Connell, R.J. & Green, J.R., 1999. Colletotrichum: A model genus for studies on pathology and fungal-plant interactions. *Fungal Genetics and Biology*, 27(2-3), pp.186-198.

Petersen, T.N., Brunak, S., von Heijne, G. & Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8(10), pp.785-786.

Politis, D.J., 1975. The identity and perfect state of *Colletotrichum graminicola*. *Mycologia*, 67(1), p.56.

Politis, D.J. & Wheeler, H., 1972. The perfect stage of *Colletotrichum graminicola*. *Plant Disease Reporter*, 56(12), pp.1026-1027.

Powell, A.J., Conant, G.C., Brown, D.E., Carbone, I. & Dean, R.A., 2008. Altered patterns of gene duplication and differential gene gain and loss in fungal pathogens. *BMC Genomics*, 9, p.147.

Prusky, D., Freeman, S. & Dickman, M.B., 2000. Colletotrichum: host specificity, pathology, and host-pathogen interaction, *APS Press*.

Prusky, D., McEvoy, J.L., Leverentz, B. & Conway, W.S., 2001. Local modulation of host pH by Colletotrichum species as a mechanism to increase virulence. *Molecular Plant-Microbe Interactions*, 14(9), pp.1105-1113.

Przeworski, M., Hudson, R.R. & Di Rienzo, A., 2000. Adjusting the focus on human variation. *Trends in Genetics*, 16(7), pp.296-302.

Qu, Z. & Adelson, D.L., 2012a. Bovine ncRNAs are abundant, primarily intergenic, conserved and associated with regulatory genes. *PLoS ONE*, 7(8), p.e42638.

Qu, Z. & Adelson, D.L., 2012b. Evolutionary conservation and functional roles of ncRNA. *Frontiers in Genetics*, 3:205.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. & Lopez, R., 2005. InterProScan: protein domains identifier. *Nucleic Acids Research*, 33(Web Server issue), pp.116-120.

Raffaele, S. & Kamoun, S., 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nature Reviews Microbiology*, 10(6), pp.417-430.

Rajashekar, B., Samson, P., Johansson, T. & Tunlid, A., 2007. Evolution of nucleotide sequences and expression patterns of hydrophobin genes in the ectomycorrhizal fungus *Paxillus involutus*. *The New Phytologist*, 174(2), pp.399-411.

Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., Grimmond, S.M., Hume, D.A., Hayashizaki, Y. & Mattick, J.S., 2006. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Research*, 16(1), pp.11-19.

Reblova, M., Gams, W. & Seifert, K.A., 2011. Monilochaetes and allied genera of the Glomerellales, and a reconsideration of families in the Microascales. *Studies in Mycology*, 68, pp.163-191.

Rech, G.E., Vargas, W.A., Sukno, S.A. & Thon, M.R., 2012. Identification of positive selection in disease response genes within members of the Poaceae. *Plant signaling & behavior*, 7(12), pp.1667-1675.

Resch, A.M., Carmel, L., Mariño-Ramírez, L., Ogurtsov, A.Y., Shabalina, S.A., Rogozin, I.B. & Koonin, E.V., 2007. Widespread positive selection in synonymous sites of mammalian genes. *Molecular Biology and Evolution*, 24(8), pp.1821-1831.

Rice, P., Longden, I. & Bleasby, A., 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6), pp.276-277.

Roberts, W.K. & Selitrennikoff, C.P., 1990. Zeamatin, an antifungal protein from maize with membrane-permeabilizing activity. *Journal of General Microbiology*, 136(9), pp.1771-1778.

Rojas, E.I., Rehner, S.A., Samuels, G.J., Van Bael, S.A., Herre, E.A., Cannon, P., Chen, R., Pang, J., Wang, R., Zhang, Y., Peng, Y.-Q. & Sha, T., 2010. *Colletotrichum gloeosporioides* s.l. associated with *Theobroma cacao* and other plants in Panama: multilocus phylogenies distinguish host-associated pathogens from asymptomatic endophytes. *Mycologia*, 102(6), pp.1318-1338.

Romeis, T., 2001. Protein kinases in the plant defence response. *Current Opinion in Plant Biology*, 4(5), pp.407-414.

Ronald, J. & Akey, J.M., 2007. The evolution of gene expression QTL in *Saccharomyces cerevisiae*. *PLoS ONE*, 2(8), p.e678.

Roth, C. & Liberles, D.A., 2006. A systematic search for positive selection in higher plants (Embryophytes). *BMC Plant Biology*, 6(1), p.12.

Rouxel, T. & Balesdent, M.-H., 2010. Avirulence Genes. En John Wiley & Sons, Ltd, ed. Encyclopedia of Life Sciences. Chichester, UK: John Wiley & Sons, Ltd..

Sacristán, S., Vigouroux, M., Pedersen, C., Skamnioti, P., Thordal-Christensen, H., Micali, C., Brown, J.K.M. & Ridout, C.J., 2009. Coevolution between a family of parasite virulence effectors and a class of LINE-1 retrotransposons. *PLoS ONE*, 4(10), p.e7463.

Saleh, D., Xu, P., Shen, Y., Li, C., Adreit, H., Milazzo, J., Ravigné, V., Bazin, E., Nottéghem, J.-L., Fournier, E. & Tharreau, D., 2012. Sex at the origin: an Asian population of the rice blast fungus *Magnaporthe oryzae* reproduces sexually. *Molecular Ecology*, 21(6), pp.1330–1344.

Samson, R.A., Varga, J. & Dyer, P.S., 2009. Morphology and reproductive mode of *Aspergillus fumigatus*. En Latgé, J. P. &Steinbach, W. J. eds. *Aspergillus fumigatus* and aspergillosis, pp.7-13.

Scheideler, M., Schlaich, N.L., Fellenberg, K., Beissbarth, T., Hauser, N.C., Vingron, M., Slusarenko, A.J. & Hoheisel, J.D., 2002. Monitoring the switch

from housekeeping to pathogen defense metabolism in *Arabidopsis thaliana* using cDNA arrays. *Journal of Biological Chemistry*, 277(12), pp.10555-10561.

Schirawski, J., Mannhaupt, G., Munch, K., Brefort, T., Schipper, K., Doehlemann, G., Di Stasio, M., Rossel, N., Mendoza-Mendoza, A., Pester, D., Muller, O., Winterberg, B., Meyer, E., Ghareeb, H., Wollenberg, T., Munsterkotter, M., Wong, P., Walter, M., Stukenbrock, E., Guldener, U. & Kahmann, R., 2010. Pathogenicity determinants in smut fungi revealed by genome comparison. *Science*, 330(6010), pp.1546-1548.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C. & Jackson, S.A., 2010. Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278), pp.178-183.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.-T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A. & Wilson, R.K., 2009. The B73 maize genome: Complexity, diversity, and dynamics. *Science*, 326(5956), pp.1112-1115.

Schoustra, S.E., Debets, A.J.M., Slakhorst, M. & Hoekstra, R.F., 2007. Mitotic recombination accelerates adaptation in the fungus *Aspergillus nidulans*. *PLoS Genetics*, 3(4), p.e68.

Schürch, S., Linde, C.C., Knogge, W., Jackson, L.F. & McDonald, B.A., 2004. Molecular population genetic analysis differentiates two virulence mechanisms of the fungal avirulence gene NIP1. *Molecular Plant-Microbe Interactions*, 17(10), pp.1114-1125.

Simmons, C.R., Fridlender, M., Navarro, P.A. & Yalpani, N., 2003. A maize defense-inducible gene is a major facilitator superfamily member related to bacterial multidrug resistance efflux antiporters. *Plant Molecular Biology*, 52(2), pp.433-446.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. & Birol, İ., 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6), pp.1117-1123.

Singh, V.K., Singh, A.K., Chand, R. & Kushwaha, C., 2011. Role of bioinformatics in agriculture and sustainable development. *International Journal of Bioinformatics Research*, 3(2), pp.221-226.

Smit, A.F.A., Hubley, R. & Green, P., 1996. RepeatMasker Open-3.0, Available at: http://www.repeatmasker.org.

Smith, C.D., Edgar, R.C., Yandell, M.D., Smith, D.R., Celniker, S.E., Myers, E.W. & Karpen, G.H., 2007. Improved repeat identification and masking in Dipterans. *Gene*, 389(1), pp.1-9.

Soanes, D.M., Alam, I., Cornell, M., Wong, H.M., Hedeler, C., Paton, N.W., Rattray, M., Hubbard, S.J., Oliver, S.G. & Talbot, N.J., 2008. Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *PLoS ONE*, 3(6), p.e2300.

Song, W.-Y., Wang, G.-L., Chen, L.-L., Kim, H.-S., Pi, L.-Y., Holsten, T., Gardner, J., Wang, B., Zhai, W.-X., Zhu, L.-H., Fauquet, C. & Ronald, P., 1995. A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. *Science*, 270(5243), pp.1804-1806.

Spanu, P.D. & Kämper, J., 2010. Genomics of biotrophy in fungi and oomycetes—emerging patterns. *Current Opinion in Plant Biology*, 13(4), pp.409-414.

Spanu, P.D., Abbott, J.C., Amselem, J., Burgis, T.A., Soanes, D.M., Stuber, K., Loren van Themaat, E.V., Brown, J.K.M., Butcher, S.A., Gurr, S.J., Lebrun, M.-H., Ridout, C.J., Schulze-Lefert, P., Talbot, N.J., Ahmadinejad, N., Ametz, C., Barton, G.R., Benjdia, M., Bidzinski, P., Bindschedler, L.V., Both, M., Brewer, M.T., Cadle-Davidson, L., Cadle-Davidson, M.M., Collemare, J., Cramer, R., Frenkel, O., Godfrey, D., Harriman, J., Hoede, C., King, B.C., Klages, S., Kleemann, J., Knoll, D., Koti, P.S., Kreplak, J., Lopez-Ruiz, F.J., Lu, X., Maekawa, T., Mahanil, S., Micali, C., Milgroom, M.G., Montana, G., Noir, S., O'Connell, R.J., Oberhaensli, S., Parlange, F., Pedersen, C., Quesneville, H., Reinhardt, R., Rott, M., Sacristan, S., Schmidt, S.M., Schon, M., Skamnioti, P., Sommer, H., Stephens, A., Takahara, H., Thordal-Christensen, H., Vigouroux, M., Wessling, R., Wicker, T. & Panstruga, R., 2010. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*, 330(6010), pp.1543-1546.

Staats, M., van Baarlen, P., Schouten, A., van Kan, J.A.L. & Bakker, F.T., 2007. Positive selection in phytotoxic protein-encoding genes of Botrytis species. *Fungal Genetics and Biology*, 44(1), pp.52-63.

Stahl, E.A. & Bishop, J.G., 2000. Plant-pathogen arms races at the molecular level. *Current Opinion in Plant Biology*, 3(4), pp.299-304.

Stajich, J.E., Berbee, M.L., Blackwell, M., Hibbett, D.S., James, T.Y., Spatafora, J.W. & Taylor, J.W., 2009. The Fungi. *Current Biology*, 19(18), pp.R840-R845.

Stajich, J.E., Harris, T., Brunk, B.P., Brestelli, J., Fischer, S., Harb, O.S., Kissinger, J.C., Li, W., Nayak, V., Pinney, D.F., Stoeckert, C.J. & Roos, D.S., 2011. FungiDB: an integrated functional genomics database for fungi. *Nucleic Acids Research*, 40(D1), pp.D675-D681.

Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D., 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5), pp.637-644.

Stanke, M. & Waack, S., 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(2), pp.215-225.

Steigele, S., Huber, W., Stocsits, C., Stadler, P.F. & Nieselt, K., 2007. Comparative analysis of structured RNAs in *S. cerevisiae* indicates a multitude of different functions. *BMC Biology*, 5(1), p.25.

Stergiopoulos, I., De Kock, M.J., Lindhout, P. & De Wit, P.J.G., 2007. Allelic variation in the effector genes of the tomato pathogen *Cladosporium fulvum* reveals different modes of adaptive evolution. *Molecular Plant-Microbe Interactions*, 20(10), pp.1271–1283.

Stewart, J.E., Thomas, K.A., Lawrence, C.B., Dang, H., Pryor, B.M., Timmer, L.M.P. & Peever, T.L., 2013. Signatures of recombination in clonal lineages of the citrus brown spot pathogen, *Alternaria alternata* sensu lato. *Phytopathology*, 103(7), pp.741-749.

Storz, G., Opdyke, J.A. & Zhang, A., 2004. Controlling mRNA stability and translation with small, noncoding RNAs. *Current Opinion in Microbiology*, 7(2), pp.140-144.

Strange, R.N., 2003. Introduction to plant pathology, *John Wiley and Sons*.

Stukenbrock, E.H. & Bataillon, T., 2012. A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. *PLoS Pathogens*, 8(9), p.e1002893.

Stukenbrock, E.H., Bataillon, T., Dutheil, J.Y., Hansen, T.T., Li, R., Zala, M., McDonald, B.A., Wang, J. & Schierup, M.H., 2011. The making of a new pathogen: Insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. *Genome Research*, 21(12), pp.2157-2166.

Stukenbrock, E.H. & Dutheil, J.Y., 2012. Comparing Fungal Genomes: Insight into Functional and Evolutionary Processes. En M. D. Bolton & B. P. H. J. Thomma, eds. Plant Fungal Pathogens. Methods in Molecular Biology. *Humana Press*, pp. 531-548.

Stukenbrock, E.H., Jørgensen, F.G., Zala, M., Hansen, T.T., McDonald, B.A. & Schierup, M.H., 2010. Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. *PLoS Genetics*, 6(12), p.e1001189.

Stukenbrock, E.H. & McDonald, B.A., 2007. Geographical variation and positive diversifying selection in the host-specific toxin SnToxA. *Molecular Plant Pathology*, 8(3), pp.321–332.

Stukenbrock, E.H. & McDonald, B.A., 2009. Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. *Molecular Plant-Microbe Interactions*, 22(4), pp.371–380.

Subdirección General de Análisis, Prospectiva y Coordinación, 2012. Dossier Autonómico; Comunidad Autónoma de Castilla y León. Available at: http://www.magrama.gob.es/es/ministerio/servicios/analisis-y-prospectiva/Dossier_Castilla_y_Leon_tcm7-183049.pdf

Sukno, S.A., Sanz-Martín, J.M., González-Fuente, M., Hiltbrunner, J. & Thon, M.R., 2013. First report of anthracnose stalk rot of maize caused by *Colletotrichum graminicola* in Switzerland. *Plant Disease*, in press.

Sukno, S.A., García, V.M., Shaw, B.D. & Thon, M.R., 2008. Root infection and systemic colonization of maize by *Colletotrichum graminicola*. *Applied and Environmental Microbiology*, 74(3), pp.823-832.

Sun, Y. & Buhler, J., 2008. Designing secondary structure profiles for fast ncRNA identification. *Computational Systems Bioinformatics / Life Sciences Society*, 7, pp.145-156.

Svenningsen, S.L., Tu, K.C. & Bassler, B.L., 2009. Gene dosage compensation calibrates four regulatory RNAs to control *Vibrio cholerae* quorum sensing. *The EMBO Journal*, 28(4), pp.429-439.

Taft, R.J., Glazov, E.A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G.J., Lassmann, T., Forrest, A.R.R., Grimmond, S.M., Schroder, K., Irvine, K., Arakawa, T., Nakamura, M., Kubosaki, A., Hayashida, K., Kawazu, C., Murata, M., Nishiyori, H., Fukuda, S., Kawai, J., Daub, C.O., Hume, D.A., Suzuki, H., Orlando, V., Carninci, P., Hayashizaki, Y. & Mattick, J.S., 2009. Tiny RNAs associated with transcription start sites in animals. *Nature Genetics*, 41(5), pp.572-578.

Taft, R.J., Pheasant, M. & Mattick, J.S., 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays*, 29(3), pp.288–299.

Tahlan, K., Park, H.U., Wong, A., Beatty, P.H. & Jensen, S.E., 2004. Two sets of paralogous genes encode the enzymes involved in the early stages of clavulanic acid and clavam metabolite biosynthesis in *Streptomyces clavuligerus*. *Antimicrobial Agents and Chemotherapy*, 48(3), pp.930-939.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), pp.585-595.

Takano, Y., Kikuchi, T., Kubo, Y., Hamer, J.E., Mise, K. & Furusawa, I., 2000. The *Colletotrichum lagenarium* MAP kinase gene CMK1 regulates diverse aspects of fungal pathogenesis. *Molecular Plant-Microbe Interactions*, 13(4), pp.374-383.

Takken, F. & Rep, M., 2010. The arms race between tomato and *Fusarium oxysporum*. *Molecular Plant Pathology*, 11(2), pp.309–314.

Talhinhas, P., Sreenivasaprasad, S., Neves-Martins, J. & Oliveira, H., 2005. Molecular and phenotypic analyses reveal association of diverse *Colletotrichum acutatum* groups and a low level of *C. gloeosporioides* with olive anthracnose. *Applied and Environmental Microbiology*, 71(6), pp.2987-2998.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. & Paterson, A.H., 2008. Synteny and collinearity in plant genomes. *Science*, 320(5875), pp.486-488.

Terauchi, R. & Yoshida, K., 2010. Towards population genomics of effector-effector target interactions. *The New Phytologist*, 187(4), pp.929-939.

The UniProt Consortium, 2012. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, 41(D1), pp.D43-D47.

Thon, M.R., Martin, S.L., Goff, S., Wing, R.A. & Dean, R.A., 2004. BAC end sequences and a physical map reveal transposable element content and clustering patterns in the genome of *Magnaporthe grisea*. *Fungal Genetics and Biology*, 41(7), pp.657-666.

Thon, M.R., Nuckles, E.M., Takach, J.E. & Vaillancourt, L.J., 2002a. CPR1: A gene encoding a putative signal peptidase that functions in pathogenicity of *Colletotrichum graminicola* to maize. *Molecular Plant-Microbe Interactions*, 15(2), pp.120–128.

Thon, M.R., Nuckles, E.M. & Vaillancourt, L.J., 2000. Restriction enzyme-mediated integration used to produce pathogenicity mutants of *Colletotrichum graminicola*. *Molecular Plant-Microbe Interactions*, 13(12), pp.1356-1365.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. & Pachter, L., 2010. Transcript assembly and

quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), pp.511-515.

Tritt, A., Eisen, J.A., Facciotti, M.T. & Darling, A.E., 2012. An integrated pipeline for de novo assembly of microbial genomes. *PLoS ONE*, 7(9), p.e42304.

Tunlid, A. & Talbot, N.J., 2002. Genomics of parasitic and symbiotic fungi. *Current Opinion in Microbiology*, 5(5), pp.513-519.

Tupy, J.L., Bailey, A.M., Dailey, G., Evans-Holm, M., Siebel, C.W., Misra, S., Celniker, S.E. & Rubin, G.M., 2005. Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, 102(15), pp.5495-5500.

Tzin, V. & Galili, G., 2010. New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. *Molecular Plant*, 3(6), pp.956-972.

Vaillancourt, L.J., 1991. A method for genetic analysis of *Glomerella graminicola* (*Colletotrichum graminicola*) from maize. *Phytopathology*, 81(5), p.530.

Vaillancourt, L.J. & Hanau, R.M., 1992. Genetic and morphological comparisons of Glomerella (Colletotrichum) isolates from maize and from sorghum. *Experimental Mycology*, 16(3), pp.219-229.

Vargas, W.A., Sanz-Martín, J.M., Rech, G.E., Rivera, L.P., Benito, E.P., Díaz-Mínguez, J.M., Thon, M.R. & Sukno, S.A., 2012. Plant defense mechanisms are activated during biotrophic and necrotrophic development of *Colletotricum graminicola* in maize. *Plant Physiology*, 158(3), pp.1342-1358.

Viscidi, R.P. & Demma, J.C., 2003. Genetic diversity of *Neisseria gonorrhoeae* housekeeping genes. *Journal of Clinical Microbiology*, 41(1), pp.197-204.

Ward, T.J., Bielawski, J.P., Kistler, H.C., Sullivan, E. & O'Donnell, K., 2002. Ancestral polymorphism and adaptive evolution in the trichothecene mycotoxin gene cluster of phytopathogenic Fusarium. *Proceedings of the National Academy of Sciences*, 99(14), pp.9278-9283.

Warren, H.L., Nicholson, R.L. & Turner, M.T., 1975. Field reaction of corn inbreds to *Colletotrichum graminicola*. *Plant Disease Reporter*, 59(9), pp.767-769.

Wasilwa, L.A., 1993. Reexamination of races of the cucurbit anthracnose pathogen *Colletotrichum orbiculare*. *Phytopathology*, 83(11), p.1190.

Wattad, C., Dinoor, A. & Prusky, D., 1994. Purification of pectate lyase produced by *Colletotrichum gloeosporioides* and its inhibition by epicatechin: a possible factor involved in the resistance of unripe avocado fruits to anthracnose. *Molecular Plant-Microbe Interactions*, 7(2), pp.293-297.

Wattad, C., Freeman, S., Dinoor, A. & Prusky, D., 1995. A nonpathogenic mutant of *Colletotrichum magna* is deficient in extracellular secretion of pectate lyase. *Molecular Plant-Microbe Interactions,* 8(4), pp.621-626.

Wharton, P.S. & Uribeondo, J.D., 2004. The biology of *Colletotrichum acutatum*. *Anales del Jardín Botánico de Madrid*, 61(1), pp.3-22.

Whitehead, A. & Crawford, D.L., 2006. Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences*, 103(14), pp.5425-5430.

Whittle, C.A., Sun, Y. & Johannesson, H., 2012. Genome-wide selection on codon usage at the population level in the fungal model organism *Neurospora crassa*. *Molecular Biology and Evolution*, 29(8), pp.1975-1986.

Wijesundera, R.L.C., Bailey, J.A., Byrde, R.J.W. & Fielding, A.H., 1989. Cell wall degrading enzymes of *Colletotrichum lindemuthianum*: their role in the development of bean anthracnose. *Physiological and Molecular Plant Pathology*, 34(5), pp.403-413.

Wik, L., Karlsson, M. & Johannesson, H., 2008. The evolutionary trajectory of the mating-type (mat) genes in Neurospora relates to reproductive behavior of taxa. *BMC Evolutionary Biology*, 8, pp.109-109.

Williams, G.C., 1975. Sex and Evolution, *Princeton University Press*.

Wilusz, J.E., Sunwoo, H. & Spector, D.L., 2009. Long noncoding RNAs: functional surprises from the RNA world. *Genes & Development*, 23(13), pp.1494-1504.

Win, J., Morgan, W., Bos, J., Krasileva, K.V., Cano, L.M., Chaparro-Garcia, A., Ammar, R., Staskawicz, B.J. & Kamoun, S., 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic Oomycetes. *The Plant Cell*, 19(8), pp.2349-2369.

Winnenburg, R., Baldwin, T.K., Urban, M., Rawlings, C., Köhler, J. & Hammond-Kosack, K.E., 2006. PHI-base: a new database for pathogen host interactions. *Nucleic Acids Research*, 34(1), pp.D459-D464.

Woloshuk, C.P., Foutz, K.R., Brewer, J.F., Bhatnagar, D., Cleveland, T.E. & Payne, G.A., 1994. Molecular characterization of aflR, a regulatory locus for aflatoxin biosynthesis. *Applied and Environmental Microbiology*, 60(7), pp.2408-2414.

Wong, W.S.W. & Nielsen, R., 2004. Detecting selection in noncoding regions of nucleotide sequences. *Genetics*, 167(2), pp.949-958.

Woolhouse, M.E.J., Webster, J.P., Domingo, E., Charlesworth, B. & Levin, B.R., 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, 32(4), pp.569-577.

Van de Wouw, A.P., Cozijnsen, A.J., Hane, J.K., Brunner, P.C., McDonald, B.A., Oliver, R.P. & Howlett, B.J., 2010. Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants. *PLoS Pathogens*, 6(11), p.e1001180.

Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, 8(3), pp.206-216.

Wu, T.D. & Watanabe, C.K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9), pp.1859-1875.

Xiao, S., Wang, W. & Yang, X., 2008. Evolution of Resistance Genes in Plants. En H. Heine, ed. Innate Immunity of Plants, Animals, and Humans. Berlin, Heidelberg: *Springer Berlin Heidelberg*, pp. 1-25.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. & Kellis, M., 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, 434(7031), pp.338-345.

Yang, Z., 2006. Computational molecular evolution, *Oxford University Press*.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8), pp.1586-1591.

Yang, Z. & Bielawski, J., 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12), pp.496-503.

Yang, Z., Wong, W.S.. & Nielsen, R., 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, 22(4), p.1107.

Yip, K.Y., Patel, P., Kim, P.M., Engelman, D.M., McDermott, D. & Gerstein, M., 2008. An integrated system for studying residue coevolution in proteins. *Bioinformatics*, 24(2), pp.290 -292.

Youens-Clark, K., Buckler, E., Casstevens, T., Chen, C., Declerck, G., Derwent, P., Dharmawardhana, P., Jaiswal, P., Kersey, P., Karthikeyan, A.S., Lu, J., McCouch, S.R., Ren, L., Spooner, W., Stein, J.C., Thomason, J., Wei, S. & Ware, D., 2011. Gramene database in 2010: updates and extensions. *Nucleic Acids Research*, 39(Database issue), pp.D1085-1094.

Yuan, Z., Su, Z., Mao, L., Peng, Y., Yang, G., Lin, F. & Zhang, C., 2011. Distinctive endophytic fungal assemblage in stems of wild rice (*Oryza granulata*) in China with special reference to two species of Muscodor (Xylariaceae). *Journal of Microbiology (Seoul, Korea)*, 49(1), pp.15-23.

Zamora, A., Sun, Q., Hamblin, M.T., Aquadro, C.F. & Kresovich, S., 2009. Positively selected disease response orthologous gene sets in the cereals identified using *Sorghum bicolor* L. Moench expression profiles and comparative genomics. *Molecular Biology and Evolution*, 26(9), pp.2015-2030.

Zeigler, R.S., Scott, R.P., Leung, H., Bordeos, A.A., Kumar, J. & Nelson, R.J., 1997. Evidence of parasexual exchange of DNA in the rice blast fungus challenges its exclusive clonality. *Phytopathology*, 87(3), pp.284-294.

Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821-829.

Zeyl, C., 2009. The role of sex in fungal evolution. *Current Opinion in Microbiology*, 12(6), pp.592-598.

Zhang, J., Rosenberg, H.F. & Nei, M., 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences*, 95(7), pp.3708-3713.

Zhang, N., Castlebury, L.A., Miller, A.N., Huhndorf, S.M., Schoch, C.L., Seifert, K.A., Rossman, A.Y., Rogers, J.D., Kohlmeyer, J., Volkmann-Kohlmeyer, B. & Sung, G.H., 2006. An overview of the systematics of the Sordariomycetes based on a four-gene phylogeny. *Mycologia*, 98(6), p.1076.

Zhang, Q., Wang, W., McMillan, L., Pardo-Manuel De Villena, F. & Threadgill, D., 2009. Inferring genome-wide mosaic structure. *Pacific Symposium on Biocomputing*, pp.150-161.

Zhao, C., Waalwijk, C., de Wit, P.J.G.M., van der Lee, T. & Tang, D., 2011. EBR1, a novel Zn2Cys6 transcription factor, affects virulence and apical dominance of the hyphal tip in *Fusarium graminearum*. *Molecular Plant-Microbe Interactions*, 24(12), pp.1407-1418.

Zhen, Y. & Andolfatto, P., 2012. Methods to detect selection on noncoding DNA. *Methods in Molecular Biology*, 856, pp.141-159.

Zhu, L. & Bustamante, C.D., 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics*, 170(3), pp.1411-1421.