ENSAMBLAJE, ANOTACIÓN Y ANÁLISIS INICIAL DE LOS GENOMAS DE Emmonsia parva Y Emmonsia crescens (Onygenales, Ascomycota)

ELIZABETH MISAS RIVAS

DIRECTOR Y ASESOR:

Oliver Clay, PhD

CIB, Universidad del Rosario

CODIRECTOR:

Juan Guillermo McEwen, M.D, PhD

CIB, Universidad de Antioquia

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
INSTITUTO DE BIOLOGÍA
MEDELLÍN

2011

AGRADECIMIENTOS

Oliver Clay, PhD
CIB, Universidad del Rosario
(Director y asesor de este proyecto)

Juan Guillermo McEwen, M.D, PhD
CIB, Universidad de Antioquia
(Co-director de este proyecto)

José Fernando Muñoz, Estudiante de maestría
CIB, Universidad de Antioquia

Juan Esteban Gallo, Estudiante de doctorado
CIB, Universidad del Rosario

A todos los integrantes de grupo de biología celular y molecular, CIB.

Doctor John W. Taylor y Emily A. Whiston
Universidad de Berkeley, California.

Instituto de Biología
Facultada de ciencias exactas y naturales
Universidad de Antioquia

Corporación para investigaciones biológicas

Colciencias

A mis familiares y amigos

ENSAMBLAJE Y ANOTACIÓN PRELIMINAR DE LOS GENOMAS DE *Emmonsia parva* Y *Emmonsia crescens* (Onygenales, Ascomycota)

ELIZABETH MISAS RIVAS

DIRECTOR Y ASESOR:
Oliver Clay, PhD
CIB, Universidad del Rosario

CODIRECTOR:
Juan Guillermo McEwen, M.D, PhD
CIB, Universidad de Antioquia

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
INSTITUTO DE BIOLOGÍA
MEDELLÍN
2011

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ANEXOS

ABREVIATURAS y TECNICISMOS USADOS

ABREVIATURAS	
pb	pares de bases
kb	Kilo bases
Mb	Mega bases
TECNICISMOS	
reads	Lecturas de secuencia
paired-end reads	Lecturas apareadas de secuencia
k-mer	Subsecuencia de tamaño k
Contig	Secuencias contiguas
Scaffold	Ensamblaje de contigs
Insert size	Tamaño de la librería para secuenciación

RESUMEN

Tres especies de hongos patógenos para los humanos y endémicos de algunas regiones de América del Sur y del Norte son los hongos dimórficos *Paracoccidioides brasiliensis, Histoplasma capsulatum and Blastomyces dermatitidis.* Estos tres hongos dimórficos están estrechamente relacionados con un género de hongos que muy rara vez infecta a los seres humanos, *Emmonsia*, el cual a diferencia de muchos otros hongos en el orden Onygenales todavía no está representado por una secuencia completa de genoma. Para entender mejor la patogenicidad y virulencia de este grupo de hongos a través de estudios de genómica comparativa, se decidió secuenciar los genomas de dos especies *E. parva* y *E, crescens*, utilizando ilumina, una tecnología de secuenciación de nueva generación (NGS, short paired-end reads), para ensamblaje de novo y anotar los dos genomas. Este trabajo representa la primera parte del proyecto, que consistió en el ensamblaje preliminar de los dos genomas, el análisis de los ensamblajes, la obtención de la predicción de los genes, conjuntos de proteínas y sus anotaciones iniciales.

En los genomas de *Emmonsia*, especialmente *E. parva*, se encontró que contienen un alto porcentaje de regiones de ADN pobres en GC, que consiste en su mayor parte en ADN repetitivo o de baja complejidad. En presencia de ADN repetitivo no es posible obtener sin ambigüedades el ensamblaje usando solamente secuencias cortas, pero se logró cubrir las partes del de los dos genomas que contienen casi todos los genes (es decir, “el espacio de genes”), lo que permite obtener una anotación inicial confiable, como se comprobó con diferentes controles de calidad, tales como los análisis de GC y la búsqueda de similitud.

En los siguientes pasos (análisis comparativo de genes), debería ser posible llegar una mejor comprensión de los mecanismos patogénicos y factores de virulencia de hongos patógenos dimórficos del orden Onygenales

ABSTRACT

Three fungal species that are pathogenic to human, and endemic in parts of South and North America, are the dimorphic fungi *Paracoccidioides brasiliensis*, *Histoplasma capsulatum* and *Blastomyces dermatitidis*. These three dimorphic fungi are closely related to a fungal genus that only very rarely infects humans, *Emmonsia*, but unlike many other fungi in the Onygenales order it is not yet represented by a full genome sequence. To better understand pathogenicity and virulence in this group of fungi via comparative genomics studies, we therefore decided to sequence using Illumina's next generation sequencing technology (NGS, short paired-end reads), assemble *de novo*, and annotate strains of the two species *E. parva* and *E. crescens* The work of this thesis represents the first part of the project, in which I was contributed by assembling the two genomes, analyzing the assemblies, obtaining the predicted gene and protein sets and performing their initial annotations. The *Emmonsia* genomes, especially *E. parva*, were found to contain large percentages of GC-poor DNA consisting of much repetitive or low-complexity DNA. Repetitive DNA is not possible to assemble without ambiguities using only short read sequences, but we were able to cover the part(s) of the two genomes containing almost all of the genes (i.e., the 'gene space'), allowing a reliable initial gene annotation, as we could verify by different quality checks such as GC analyses and similarity searches. In the following steps (comparative gene analyses) it should therefore be possible to reach a better understanding of pathogenic mechanisms and virulence factors of dimorphic pathogenic fungi in the Onygenales order.

1. INTRODUCCIÓN

Este proyecto hace parte de un macro proyecto de 3 años (2010-2013), aprobado por COLCIENCIAS, código 2213-48925460, “Comparative Genomics and Virulence in the pathogenic fungus *Paracoccidioides brasiliensis*”, que se está desarrollando en la unidad de biología celular y molecular en la CIB.

Entre los hongos del phylum Ascomycota se encuentran *Emmonsia parva* y *Emmonsia crescens* (anteriormente *Chrysosporium parvum* var. *parvum* y var. *crescens*), agentes causantes de la adiaspiromycosis, enfermedad respiratoria propia de roedores y otros pequeños mamíferos, que ocasionalmente puede afectar a los humanos, aunque se conocen pocos casos de infección por estos hongos (Peterson y Sigler, 1998).

E. crescens y *E. parva* son hongos dimórficos termo dependientes, que a temperaturas inferiores a 30°C crecen como micelio y comparten la misma morfología, caracterizada por la presencia de conidias (2-4µm × 2.5-4.5µm) y numerosos núcleos en las hifas; mientras que a 37°C para *E. crescens* y 40°C para *E. parva* entran en estado adisapórico, que corresponde al estado patógeno para estos hongos, caracterizado por la presencia de adiasporas que para *E. crescens* alcanzan un tamaño de 120µm y son multinucleadas, en *E. parva* las adiasporas tienen un tamaño de 25 µm y generalmente son mono nucleadas (Hejtmánek 1985).

Varias hipótesis filogenéticas basadas en alineamientos de secuencias de ADN señalan a *Emmonsia* como un taxón muy cercano a otros hongos patógenos humanos del mismo orden Onygenales como *Paracoccidioides brasiliensis*, *Histoplasma capsulatum* y *Blastomyces dermatitidis* (Peterson y Sigler, 1998)., por lo cual se hace útil conocer su secuencia genómica completa para establecer comparaciones que podrían elucidar procesos evolutivos de genes (u otras propiedades genómicas) relacionados con la virulencia de los taxa patógenos.

Los miembros del phylum Ascomycota (división *eumicetes*, reino *fungí*, dominio *eucarya*) son hongos con micelio tabicado que producen ascosporas endógenas, pudiendo ser talófitos y unicelulares en el caso de las levaduras. Algunos de ellos son dimórficos: se encuentran en una de dos (o, a veces, tres) formas o fases, de las cuales una puede ser mucho más infecciosa o virulenta que otra (Carlile, 2004). Por ejemplo, *P. brasiliensis* es capaz de crecer como micelio en el ambiente a 19°C, o como levadura a 37°C en el tejido pulmonar humano o de otros mamíferos y causar los daños característicos de la enfermedad paracoccidioidomicosis (PCM) (Klein & Tebbets, 2007). Varios de los hongos dimórficos patógenos que presentan una correspondencia entre una transición de fase y la aparición de la virulencia, se encuentran en el orden Onygenales (Carlile, 2004). Recientemente se ha dado un mayor enfoque sobre la secuenciación de hongos patógenos en los proyectos de genomas, de los cuales una gran parte han sido desarrollados por el Instituto BROAD (http://www.Broadinstitute.org), en la tabla 1 se muestran algunos de ellos.

Tabla 1. Lista de proyectos en curso o ya terminados para hongos del orden Onygenales

Nombre	Tamaño	Cepa/aislamiento
Histoplasma capsulatum	28 Mb	G217,H143,H88,NAm1,G186AR
Blastomyces dermatitidis	65-75 Mb	ATCC 26199+18188,ER-3,SLH14081
Arthroderma benhamiae / Trichophyton mentagrophytes	22 Mb	CBS 112371
Arthroderma gypseum	23 Mb	CBS 118893
Ascosphaera apis	24Mb	USDA-ARSEF 7405
Coccidioides immitis	29 Mb	RMSCC 2394+3703,RS,H538.4
Coccidioides posadasii	27Mb	CPA 0001+0020+0066,RMSCC 1037+1038+1040, RMSCC 2133+3488+3700,str. Silveira,C735 delta SOWgp
Lacazia loboi		EDM7
Microsporum canis/ Arthroderma otae	23 Mb	CBS 113480
Paracoccidioides brasiliensis	30 Mb	Pb01, Pb03, Pb18
Trichophyton equinum	24 Mb	CBS 127.97
Trichophyton rubrum	22 Mb	CBS 118892
Trichophyton tonsurans	22 Mb	CBS 112818

Trichophyton verrucosum	23Mb	HKI 0517
Uncinocarpus reesii *	*22Mb*	*1704*
Emmonsia parva *	30? Mb	UAMH 139 (este proyecto)
Emmonsia crescens *	*30? Mb*	UAMH 3008 (este proyecto)

* Generalmente no patógeno

(Fuentes: Instituto BROAD, NCBI Taxonomy Browser/Entrez Genome, publicaciónes, sitios web de proyectos.)

La cercanía evolutiva de *Emmonsia* a los hongos patógenos de interés se muestra en la filogenia ilustrada en la figura 1, de las cuales hay tres especies patógenas que se considerarán en este proyecto por su cercanía a *Emmonsia* spp son:

(1) *P. brasiliensis* agente causante de la paracoccidioidomicosis, enfermedad sistémica que afecta la población de América latina, área a la cual se restringe su distribución *P*. *brasiliensis* responde al cambio térmico encontrándose como micelio en el ambiente a 19ºC y como levadura a 37ºC en, la primera forma es productora de conidias que es la principal partícula infecciosa y cuya transición a levadura es indispensable para el proceso de infección (García *et al*., 2009).
(2) *B. dermatitidis*, causante de blastomicosis, tiene una amplia distribución y puede permanecer en estado de latencia por largos periodos de tiempo y es de reactivación frecuente en pacientes inmunocomprometidos (Carlile, 2004). Finalmente, (3) *H. capsulatum* cuya manifestación patógena es la histoplasmosis, la infección respiratoria más común causada por hongos. Cerca del 10% de los casos presentan complicaciones como la inflamación del pericardio y fibrosis de los vasos sanguíneos (Carlile, 2004).
Todos ellos son hongos dimórficos, y se ha considerado que dicha capacidad es uno de los principales mecanismos requeridos para la infección, por ello se han desarrollado numerosas investigaciones relacionadas con esta transición (García *et al.,* 2009, 2010).

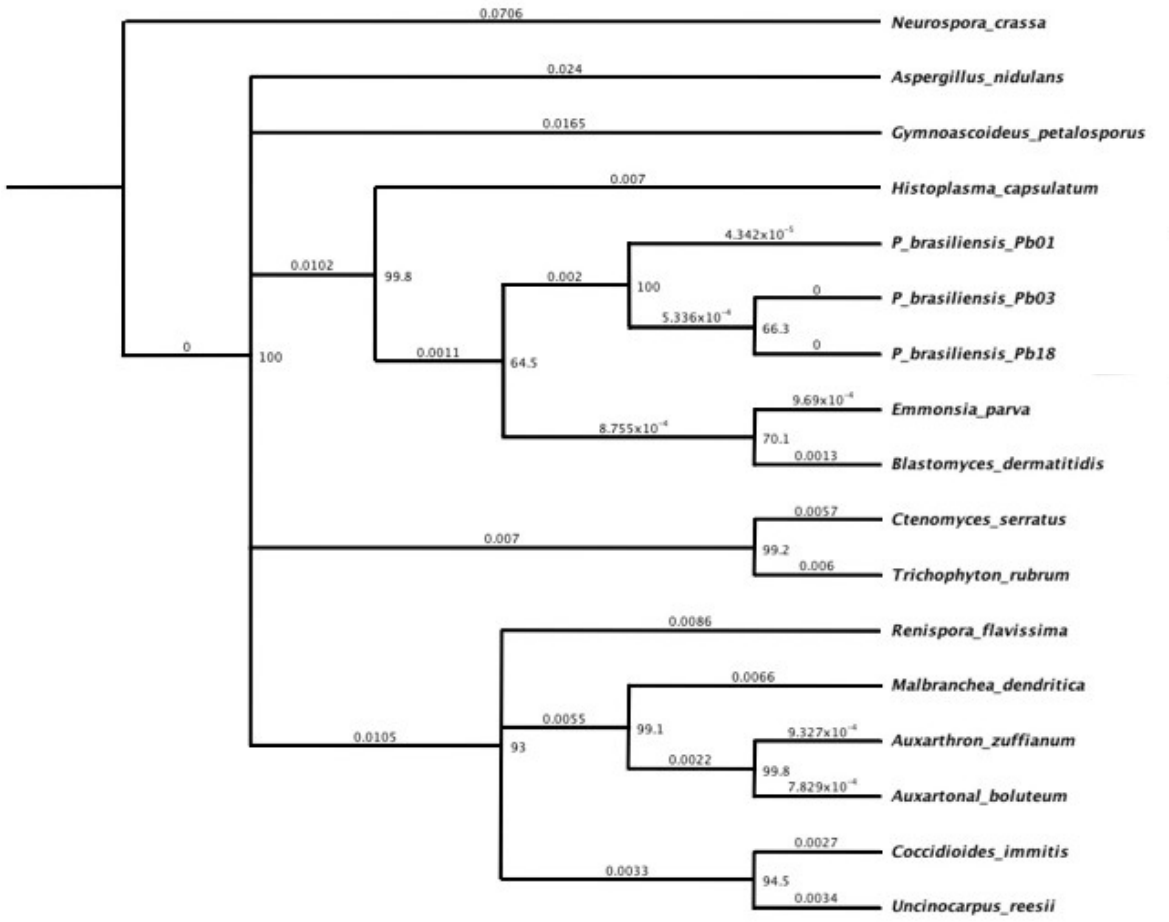


Figura 1: Árbol filogenético de varios miembros del phylum Ascomycota, construida usando secuencias de ADN 18S ribosomal y el método “neighbor-joining” y mostrando relaciones de los hongos patogénicos (Pb01, Pb03, Pb18: tres grupos representativos de la especie *P. brasiliensis)*.y no patogénicos (*Uncinocarpus reesii*, *E. parva*; cf. Bowman, White & Taylor, 1996).

Algunas investigaciones previas han reportado genes asociados a virulencia en *P. brasiliensis*, *H. capsulatum* y *B. dermatitidis* porque participan en o regulan el proceso de transición o se relacionan con algún otro cambio morfológico necesario para la adaptación del hongo a las condiciones del hospedero (Nemecek *et. al.,* 2006). Sin embargo, para las tres especies patógenas mencionadas, que están enteramente secuenciadas, no se dispone de estudios comparativos donde estos hongos patógenos de interés médico sean enfrentados a una especie no patógena pero cercana en el árbol filogenético. Es precisamente mejorar el nivel de comprensión de la virulencia en estos hongos dimórficos de interés médico, la motivación principal del proyecto que se propone.

En un trabajo del laboratorio asociado al nuestro, en Berkeley, California, se ha realizado, junto con el instituto BROAD, la secuenciación, ensamblaje y anotación de genomas de cepas del género Coccidioides, haciendo también un análisis comparativo de las cepas de Coccidioides entre ellas y con una especie no patógena *Uncinocarpus reesii,* cercana a *Coccidioides* spp. según los árboles filogenéticos disponibles (Sharpton *et al.,* 2009; Neafsey *et al.,* 2010). El proyecto actual utilizará algunos de los métodos exitosos de dicho trabajo, para el ensamblaje y la anotación inicial de los dos genomas de *Emmonsia* spp

En la comparación propuesta entre *Emmonsia* spp y los hongos de interés médico *B. dermatitidis*, *P. brasiliensis* e *H. capsulatum*, el proyecto actual tiene como enfoque la secuenciación, ensamblaje, anotación de los genomas de *E. parva* y *E. crescens* y el análisis inicial de sus genes.

2. MATERIALES Y MÉTODOS

2.1 Cepas

- *E. parva:* cepa UAMH 139, aislada de comadreja, en Ravelli County, Montana.
- *E. crescens:* cepa UAMH 3008, aislada de pulmón de roedor (*Arvicola terrestris*) en Noruega. (Peterson y Sigler, 1998).

2.2 Extracción de ADN

El ADN fue extraído en la Universidad de California, usando el método de fenol-cloroformo, con las modificaciones previas al tratamiento con los solventes orgánicos que reporta Diez et al.,1999, como las siguientes: una lisis mecánica de la suspensión de hongos que se logra mediante el uso de perlas de vidrio, buffer de lisis y pasos sucesivos de vortex-hielo y posterior incubación en hielo por aproximadamente 20 minutos para separar las fases, el sobrenadante (sin las perlas de vidrio), se incuba por una hora a 65°C. Posteriormente se realizan los tratamientos con fenol-cloformo-alcohol isoamilico (25:24:1), y un tratamiento solo con cloformo-alcohol isoamilico (24:1) usando un volumen igual a la fase acuosa recuperada. Luego el ADN se precipita con alcohol y acetato de amonio, se resuspende en buffer TE y se trata con RNAasa a 37°C durante una hora, para eliminar los posibles residuos celulares, se repite el proceso de extracción con solventes orgánicos, la precipitación con alcohol y acetato de amonio y se resuspende en buffer TE.

2.3 Preparación de librerías

La preparación de librerías se realizó siguiendo los protocolos propuestos por illumina en “Preparing Samples for Sequencing Genomic ADN”(illumina 2008). El ADN es fragmentado por nebulización para producir fragmentos de 200-800pb, se reparan los extremos y se adiciona una cola de poli A, después se ligan los adaptadores para paired end reads, mediante un gel se seleccionan los fragmentos que ligaron el adaptador y tienen un tamaño adecuado. Se realiza enriquecimiento por PCR y por último se usa un bioanalizador para revisar

tamaño/cantidad, este genera electroferogramas que permiten elegir el tamaño de librería para secuenciación (illumina 2008).

2.4 Secuenciación

Tecnología de secuenciación basada en terminadores fluorescentes reversibles, los fragmentos ADN son amplificados usando primers con estos terminadores fluorescentes reversibles, se amplifica de modo que se forman colonias de clones locales (ampliación en puente), Se adicionan los cuatro tipos de ddNTPs, el ADN puede ser extendido un nucleótido a la vez; en el Genome Analyzed, una cámara registra imágenes de los niveles de fluorescencia. La secuenciación Paired-end genera reads de ambos extremos de las librerías (illumina 2009).

Los genomas de *E. parva* (cepa UAMH 139) y *E. crescens* (cepa UAMH 3008) fueron secuenciados usando tecnología Solexa/Illumina en una colaboración entre el laboratorio de acogida (CIB/UCB) y un laboratorio de la Universidad de California. Los dos genomas, secuenciados mediante dicha tecnología, se obtienen como millones de fragmentos (‘reads’) apareados de longitudes cortas de 101 pb que corresponden a los extremos de la librería (figura 2) (Pop, 2009; http://www.illumina.com).

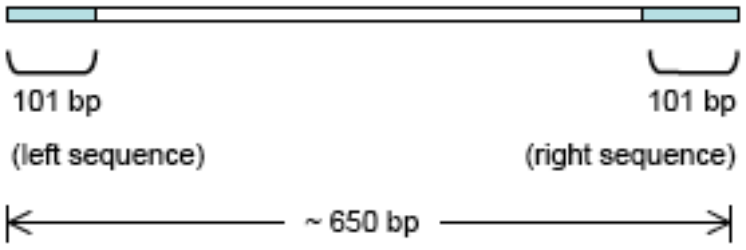


Figura 2: Esquema del insert size

Esquema de la librería, el tamaño de la librería es de 650pb y se obtienen la secuencia de 101 pb ambos extremos

2.5 Ensamblaje

El ensamblaje de los fragmentos secuenciados de un genoma (especie o cepa) se puede realizar de dos maneras:

(1) *de novo*, es decir sin la necesidad de poseer la secuencia de otro genoma como referencia (Nowrousian *et al.,* 2010),

(2) por alineación con genomas ya ensamblados y cercanos filogenéticamente.

El ensamblaje se realizó usando la primera opción, mediante softwares SOAPdenovo (Li *et al.,* 2010), ABySS (Simpson *et al.,* 2009) y Velvet (Zerbino & Birney, 2008; Zerbino *et al.,* 2009;http://www.ebi.ac.uk/~zerbino/velvet/Manual.pdf) Para SOAPdenovo y ABySS se prueban diferentes valores de K-mer (tamaño de los segmentos cortos de los reads que se usa para efectuar el ensamblaje). En el caso de velvet se elige un solo k-mer para cada especie, debido a que el corrido de este programa es computacionalmente muy costoso (Conway y Bromage, 2011).

La mayoría de los ensambladores de paired-end reads (short reads) se basan en algoritmos usando grafo de Bruijn, algoritmos que pueden procesar datos de secuencias genómicas o transcriptomas. Primero, todos las posibles subcadenas de longitud k (k-mer determinado) son generadas de las secuencias o reads, después se construye un grafo de Bruijn en el cual los nodos (vértices) son las secuencias cortas, y los enlaces (edges) van de una de las secuencias cortas a las `siguientes', por ejemplo la secuencia ACGT puede tener enlaces dirigidas a CGTA, CGTC, CGTG y/o CGTT. Aunque los programas tienen un fondo teórico en común, se distinguen entre ellos en detalles como el tratamiento de repeticiones, de errores, etc.

Basado en experiencias del laboratorio asociado al nuestro en Berkeley, en las cuales compararon varios programas, en sus versiones actuales, para ensamblar otras especies de hongos, en este proyecto se optó por SOAPdenovo como ensamblador principal. Sin embargo, se realizaron algunas pruebas y comparaciones usando otros dos programas Velvet y ABySS.

2.5.1 SOAPdenovo

El ensamblaje de los datos de Illumina / Solexa se realizó usando SOAPdenovo (Short Oligonucleotide Analysis Package) y los programas accesorios requeridos: corrector, k-mer freq, SOAPdenovo y gapcloser. Se probaron varios valores de k-mer 17,19, 21, 23, 25, 27, 29, 31, 33, 43, 53, 63(pares de bases o nucleótidos).

2.5.2 ABySS

ABySS (Assembly By Short Sequences), procede en dos estados. Primero, todos las posibles subcadenas de longitud k (k-mer determinado) son generadas de la secuencia de reads, el conjunto de datos obtenido es procesado para remover los errores e iniciar la construcción de los contigs. En el segundo estado la información pareada (mate-pair) es usada para extender los contigs resolviendo ambigüedades de sobrelapamiento de los contigs (Simpson et al. 2009). Se probaron los siguientes valores de k-mer 21, 43 y 63 para *E. parva* y 23, 43,6 para *E. crescens*

2.5.2 Velvet

Programa o paquete algorítmico que es expresamente diseñado para tratar con secuencias en fragmentos cortos; Velvet tiene dos tareas: eliminar los errores en los datos (secuencias cortas o 'reads') y luego desenredar las regiones repetidas del genoma (Zerbino & Birney 2008; Zerbino *et al.,* 2009). Velvet consiste en dos programas, velveth y, posteriormente, velvetg que tiene exigencias de memoria (RAM) muy grandes (> 32 GB, ver Conway y Bromage, 2011).

2.6 Anotación de genes

Se usaron programas para predicción *ab initio* como AUGUSTUS y FGENESH (en servidor web) y se usan predicciones basadas en homología como Blast.

2.6.1 BlastN

Para los ensamblajes obtenidos por SOAPdenovo para cada valor de k-mer, se

realizó un blastn con las secuencias disponibles de tres especies cercanas filogenéticamente; *B. dermatiditis, H. capsulatum y P. brasiliensis (Anexo 3).*

2.6.2 Augustus

Es un programa para la predicción ab initio de los genes codificantes de proteínas en genomas eucariotas basado en un “Hidden Markov Model” (HMM) y se integran una serie de métodos para modelar la longitud de los intrones. Usa un nuevo modelo de sitio de splice (corte y empalme), de un donador, en este caso particular se usó el de *H. capsulatum*, donde una región corta directamente upstream del sitio modelo de splice elige el mejor marco de lectura dependiendo del contenido de GC (Stanke 2003). Este programa se instala en el servidor y se usa para los ensamblajes obtenidos de SOAPdenovo para cada valor de k-mer.

2.6.2.1 Blast2GO (B2G)

Es una herramienta de investigación diseñada con el objetivo principal de obtener la anotación Gene Ontology (GO) para secuencias que aún no tiene disponible anotación GO, es una aplicación basada en búsqueda de similaridad con análisis estadístico y visualización (Conesa et al 2005). Blast2GO se usó para realizar blast a los genes predichos por Augustus para los ensamblajes de *E. parva* k-mer 21 y *E.crescens* usando SOAPdenovo en el presente trabajo se analizan únicamente los resultados de la parte Blast, ya que las columnas de anotación GO generadas por el programa no estaban completas para todos los genes

2.6.2 FGENESH

Programa de predicción basado en “Hidden Markov Model”, de Softberry, que no es de descarga libre, por lo cual se usa el servidor web disponible en: http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind, FGENESH es un programa sensible para detectar genes, pero ha sido criticado por su tendencia a reportar cantidades importantes de ‘falsos positivos’ (Cruveiller et al., 2003), sobre todo en genomas que contienen muchas secuencias repetidas o contrastes en GC como es el caso, por ejemplo, en el genoma del genero `vecino’ de *Emmonsia*, es decir *Blastomyces*. Por lo tanto, usamos el programa FGENESH únicamente a fines de comparación para los scaffolds más

grandes de dos ensamblajes, uno de *E. parva* usando k-mers de tamaño 21 y uno de *E. crescens* usando k-mers de 23. Los scaffold más grandes se eligen usando Geneious (http://www.geneious.com).

E. crescens	*E. parva*
scaffold18	scaffold1062
scaffold133	scaffold1060
scaffold174	scaffold292
scaffold431	scaffold47
scaffold462	scaffold54

3 RESULTADOS

3.1 Secuenciación

Para los genomas de *E. parva* (cepa UAMH 139) y *E. crescens* (cepa UAMH3008), se elige un tamaño de librería de aproximadamente 650pb. A continuación se muestran los electroferogramas (Figura 3). Para detalles de la electroforesis, imagen del gel y especificaciones del bioanalizador que genero los electroferogramas ver anexo 1.

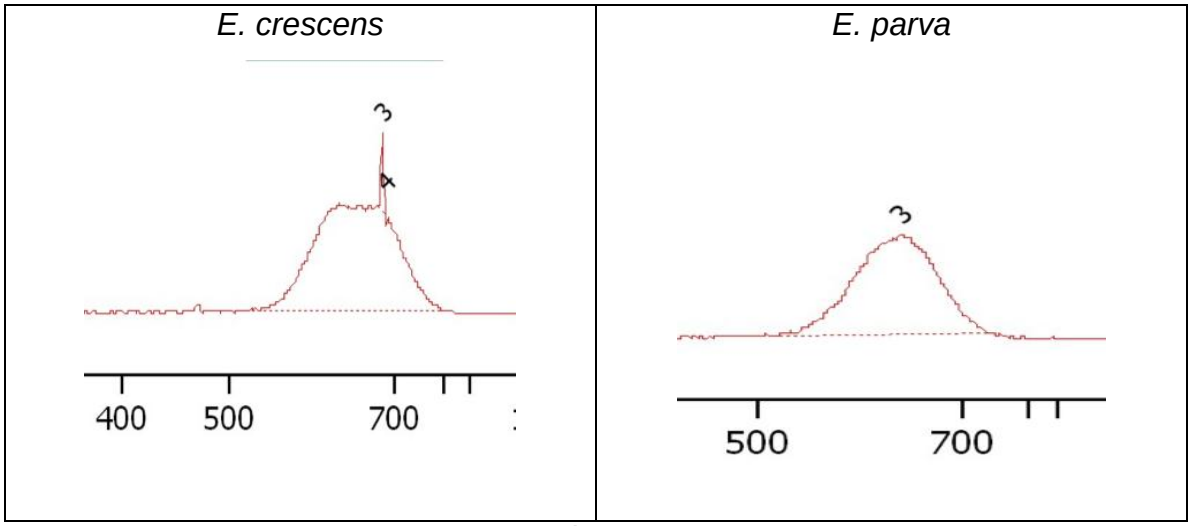


Figura 3. Electroferogramas (escala logarítmica) de los tamaños de los insert de *E. parva* y *E. crescens.* Observe que su tamaño promedio es aproximadamente 650pb. Tomado del reporte de resultados del Genome Analyzer II (Anexo 1).

Se generaron 25940870 pair-end reads para *E. parva* y 31709328 pair-end reads para *E. crescens,* cada read con una longitud de 101pb

3,2 Ensamblaje

3.2.1 SOAPdenovo

Se obtuvo un ensamblaje para cada uno de los tamaños de *k*-mer probados. Por ejemplo, el parámetro *k*-mer 21 para *E. parva* resultó en 32,82 Mb de secuencia en 21’490 contigs, 2787 scaffolds y un N50 con un tamaño de 30,06 kb (Tabla 2a),

para *E. crescens* con *k*-mers de tamaño 23 se obtuvo 33,47 Mb de secuencia en 20’140 contigs, 1473 scaffolds y un N50 de 87,74 kb (Tabla 2 b).

Tabla 2 a. Resumen de ensamblaje de *E. parva* usando SOAPdenovo.

k-mer size	Scaffolds	Ensamblaje (Mb)	N50 (kb)	Longitud promedio (kb)
17	3003	41,7	31,4	13,9
19	2616	34,2	32,2	13,1
21	2787	32,8	30,1	11,8
23	3026	33	29,3	10,9
25	3420	33,8	27,2	9,9
27	3706	35	27,3	9,4
29	4079	35,8	25,6	8,8
31	4331	36,7	25,5	8,5
33	4347	37,6	25,2	8,6
43	2870	40,8	32,6	14,2
53	1680	39,7	40,7	23,6
63	1433	37	42	25,8

Tabla 2b. Resumen de ensamblaje de *E. crescens* usando SOAPdenovo

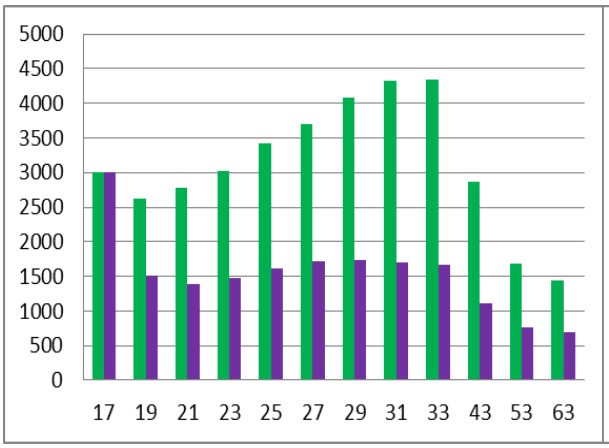
k-mer size	Scaffolds	Ensamblaje (Mb)	N50 (kb)	Longitud promedio (kb)
17	3008	43,4	35,6	14,4
19	1506	36,4	72	24,2
21	1396	34,3	82,2	24,5
23	1473	33,5	87,7	22,7
25	1610	33,3	86,6	20,7
27	1719	33,4	87,9	19,4
29	1734	33,4	92,9	19,2
31	1706	33,3	89,6	19,5
33	1659	33,4	92,7	20,1
43	1108	34,3	93,2	30,9
53	758	33,9	100,4	44,7
63	702	33,1	101	47,2

El parámetro N50 es uno de los más comúnmente reportados en la literatura para evaluar la calidad del ensamblaje y está definido como la longitud para la cual el 50% de todas las bases en el ensamble están en un contig o scaffold de al menos esa longitud, es decir que el 50% del ensamblaje este contenido en contigs o scaffolds de longitud mínima N50. Para los valores de *k*-mer evaluados se encontraron valores más altos de N50 para *E. crescens*, y en ambas especies este

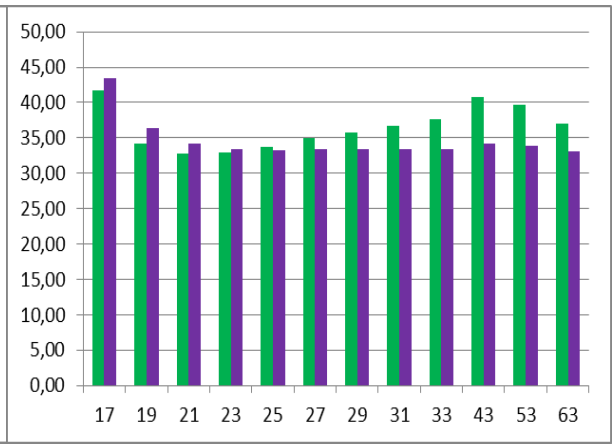
parámetro parece tener una tendencia a incrementar su valor a medida que aumenta el valor del *k*-mer.

En general para todos los valores de *k*-mer de *E. parva*, se obtuvo un número de scaffolds más alto con respeto al que se obtiene para *E. crescens,* sin embargo la longitud promedio de los scaffolds es mayor en *E. crescens* (figura 4 y 6), lo que sugiere que las partes de los genomas que no contienen cantidades importantes de ADN repetido (y por eso se pueden ensamblar sin ambigüedad a partir de reads cortos) tienen tamaños parecidos, siendo el genoma de *E. crescens* el que tiene una contribución menor de secuencias repetidas.

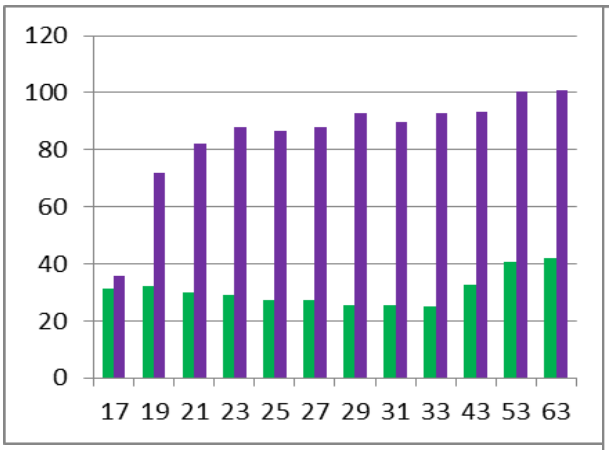
a) Numero de scaffolds



b) Tamaño del ensamblaje (Mb)



c) N50 (kb)



d) Longitud promedio(kb)

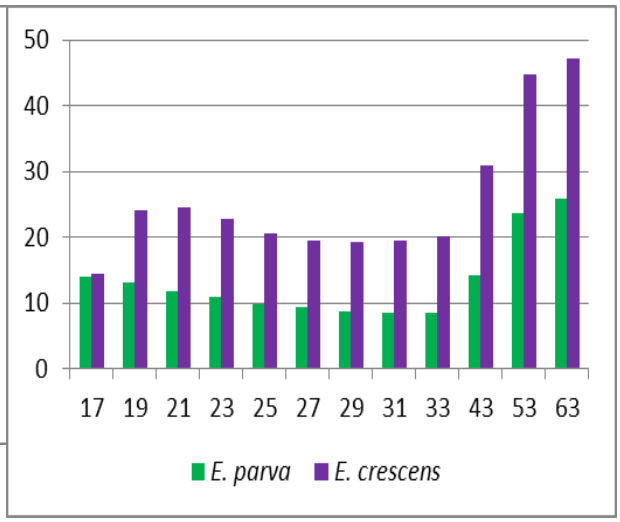


Figura 4. Comparación de varios parámetros de los ensamblajes obtenidos mediante SOAPdenovo a partir de los reads de *E. parva* (verde claro) y *E. crescens* (morado oscuro), usando valores del parámetro *k* entre 17 y 63 (eje horizontal).

Para los valores de *k*-mer entre 25 y 63, *E. parva* alcanza un tamaño de ensamblaje superior al de *E. crescens,* lo que se consideró un artilugio (valores no confiables), producto de las regiones repetidas que son detectadas en el ensamblaje de *E. parva* a valores de *k*-mer altos (Figura 5a) y que no suceden en *E. crescens* (Figura 5b).

Para *E. parva* y *E. crescens* se evaluó la distribución de frecuencia del porcentaje de GC para algunos *k*-mer (figura 5a). Al aumentar el valor *k* aparecen contigs con menor contenido de GC, sugiriendo la presencia de regiones repetitivas en el genoma, sobre todo en el caso de *E. parva*.

De hecho, se puede ya acceder a ensamblajes de varias cepas de *B dermatitidis*, obtenidos por el Broad Institute y no todavía publicados. Un análisis de estos ensamblajes que se realizaron, teniendo en cuenta las anotaciones preliminares del mismo instituto, muestran que *B. dermatitidis*, especie muy cercano a los *Emmonsia* spp. y sobre todo a *E. parva* (Peterson y Sigler, 1998, y nuestros análisis de varios genes seleccionados), contiene una cantidad muy elevada de ADN repetida y de contenido en GC bajo. Alineamientos de varias de nuestros scaffolds de *E. parva* con los scaffolds de *B. dermatitidis* usando el programa Geneious Pro mostraron que nuestros scaffolds corresponden a segmentos grandes de los scaffolds de *B. dermatitidis*, interrumpidos por bloques grandes de GC mucho más bajo, conteniendo secuencias repetidas o de baja complejidad que nuestro ensamblaje no alcanzó a recorrer (Figura 5b). Es decir, tales bloques largos de contenido bajo en GC apenas se encuentran en los scaffolds de nuestras ensamblajes de *E. parva*, pero sus subsecuencias cortas se encuentran en los reads, es decir en nuestros datos primarios (Figura 5c). Las observaciones sobre *B. dermatitidis* concuerdan con otras observaciones reportadas por otros grupos: los perfiles bimodales de GC de *B. dermatitidis* obtenidos por ultra centrifugación del ADN en el trabajo de Bowman, Garrison y Fina (1972); los análisis recientes de Clutterbuck (2011); y el hecho que del tamaño total de este genoma es aproximadamente el doble de él que observamos en otros genomas de hongos cercanos en la filogenia, como *P. brasiliensis* y *H. capsulatum* (ver Tabla 1). Es decir, un ensamblaje *de novo* y automatizado de reads cortos de 101 pb no

alcanzaría, por si mismo, a cubrir el genoma de *B. dermatitidis* y, como lo vemos, no logró el cubrimiento de la especie más estrechamente relacionada, *E. parva*. Como muestran las figuras 4 y 5, el caso de *E. crescens* es más favorable en el sentido que nuestros ensamblajes cubren una proporción más alta del genoma; por ejemplo, los tamaños N50 de los scaffolds son más elevados porque hay menos secuencias repetidas que impiden el ensamblaje.

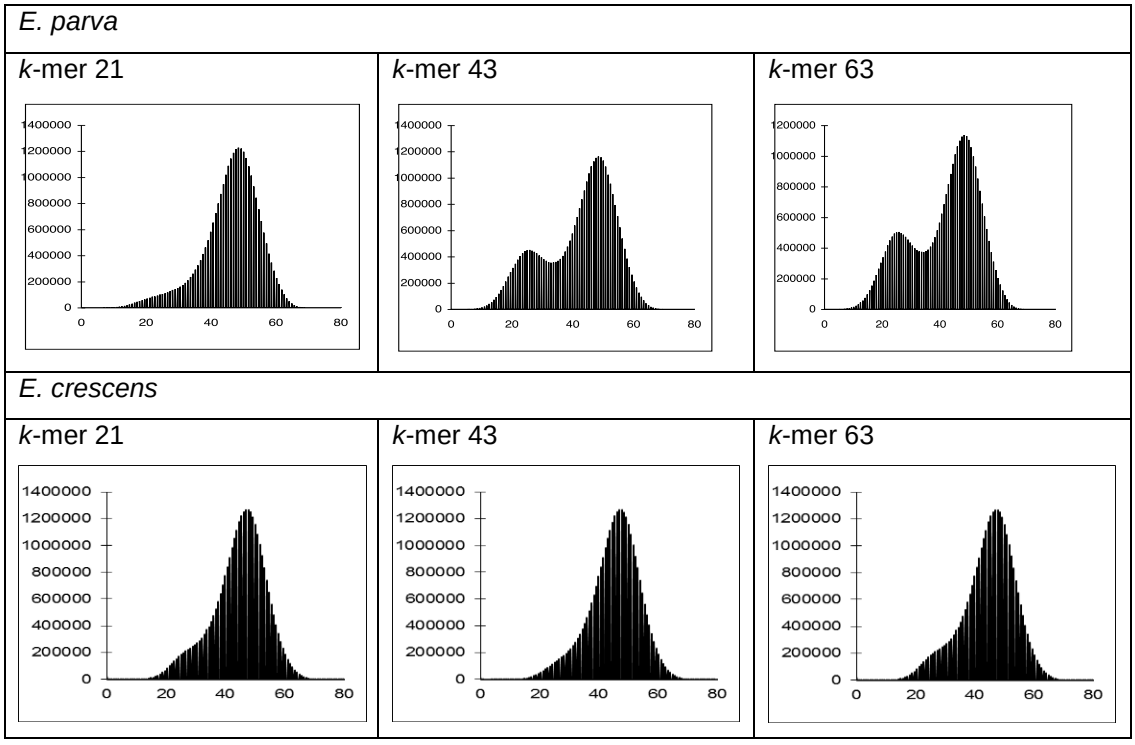


Figura 5 a. Histogramas (distribuciones de frecuencias) del porcentaje de GC en los ensamblajes de *E. parva* y *E. crescens*. El eje vertical muestra el número de segmentos (teniendo aproximadamente el mismo tamaño que los reads, 128 pb) contenidos en los scaffolds de los ensamblajes por SOAPdenovo para los valores de *k*-mer mostrados.

b) Scaffold interrumpido

c) Histograma de GC de los reads

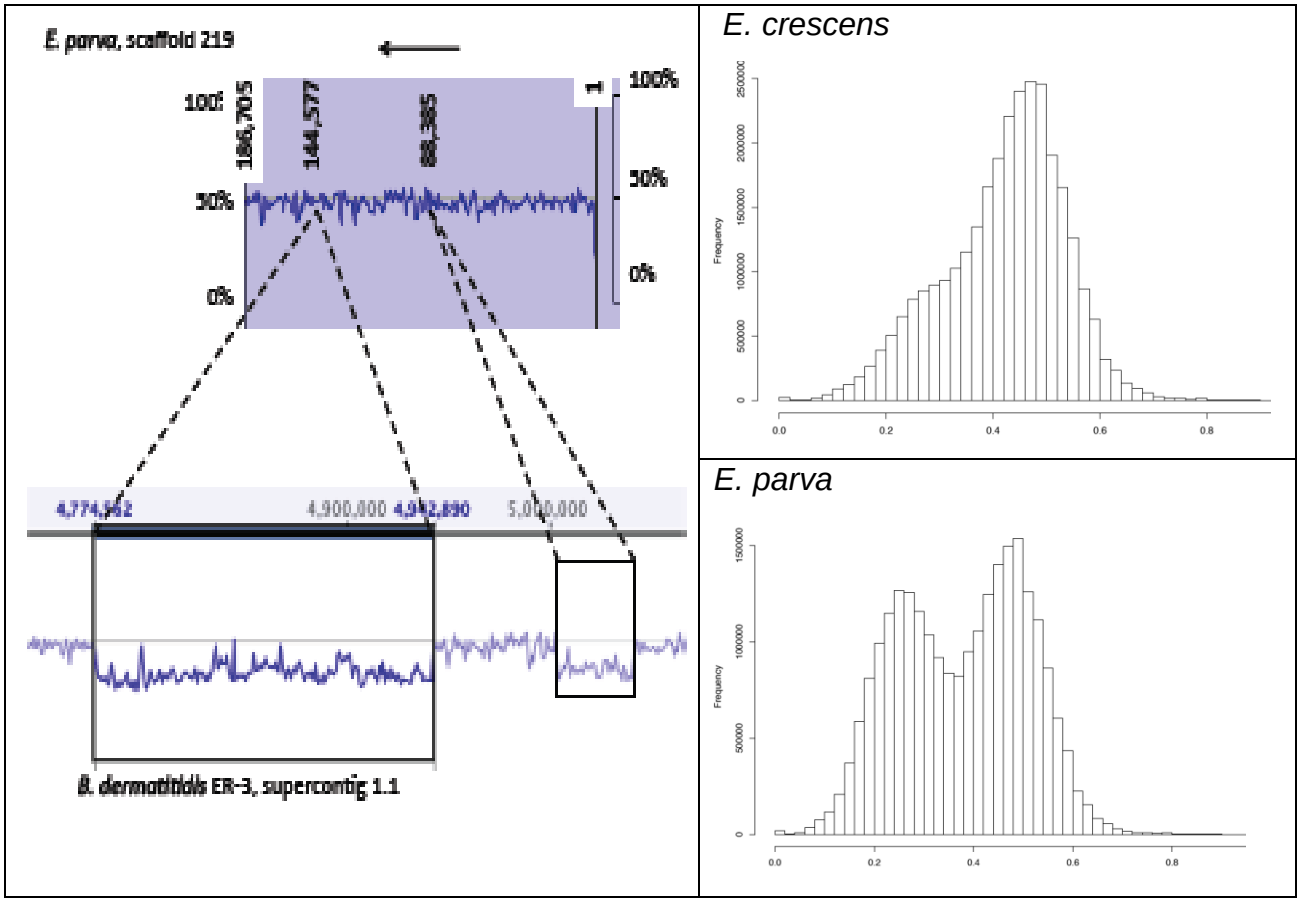


Figura 5b y 5c. Análisis de regiones pobres en GC. En 5b se observan los sitios de interrupción por largas regiones repetitivas pobres en GC en un supercontig de *B. dermatitidis* con respecto al scaffold correspondiente en *E. parva.* En 5c se observa la distribución de los paired-end reads de *E. parva* y *E. crescens*, se muestra un componente predominante de contenido pobre en GC en *E. parva* que corresponde a las regiones pobres en GC que no son ensambladas.

A partir del análisis de estos parámetros, número de scaffolds, longitud promedio de los scaffolds, tamaño del ensamblaje, N50 y el análisis del contenido de GC junto con otros parámetros adicionales N90 y longitud del scaffold más grande (*Anexo 2*), se eligen como mejores ensamblajes iniciales: *E. parva*, *k*-mer 21 y *E. crescens*, *k*-mer 23. En este contexto, es importante señalar que una estrategia de ensamblaje usando solamente las pequeños reads (‘next generation sequencing’ o ‘NGS’) pueda alcanzar a cubrir casi todos los genes de un genoma como *E. parva* o *E. crescens*, pero sin cubrir las partes altamente repetidas del mismo genoma.

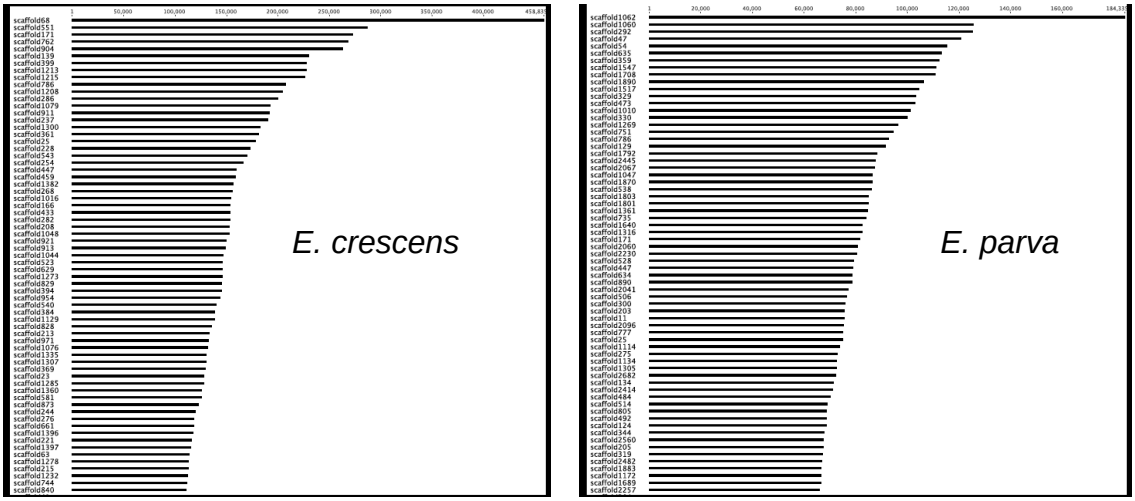


Figura 6. Longitud de los scaffolds en el ensamblaje para valores del parámetro *k*-mer 21 en *E. parva* y *k*-mer 23 en *E. crescens* (solo se muestran los scaffolds más grandes). Imágenes generadas por el software Geneious.

3.2.2 ABySS

Para este ensamblador se eligen tres valores de k-mer 21, 43 y 63. Para el k-mer 63 en *E. parva* el ensamblaje resultó en 38.29 Mb de secuencia en 30405 scaffolds y un N50 con un tamaño de 29359 pb (Tabla 3), para *E. crescens* con el mismo valor de k-mer ABySS generó 33,93 Mb de secuencia en 5022 scaffolds y un N50 de 66.53 kb.

Tabla 3: Resumen de ensamblaje de *E. crescens* y *E. parva* usando ABySS

	k-mer	Scaffolds (pb)	Ensamblaje (Mb)	N50(pb)	Longitud promedio(pb)
E. crescens	21	451973	30.79	5438	1766
	43	34554	32.88	56466	337391
	63	5022	33.93	66533	11875
E. parva	21	6144	28.79	5445	1783
	43	175804	38.62	24306	1530
	63	30405	38.29	29359	3707

3.2.3 Velvet

Este ensamblador no arrojó resultados, Velveth generaba los resultados para Velvetg, este último tras varios días de corrido fallaba, los requerimientos computacionales del programa no permitían su funcionamiento óptimo.

3.3 Anotación de genes

3.3.1 BlastN

Para *Emmonsia* spp*. s*e encuentra un mayor grado de similaridad con *B. dermatiditis* ER3. Para *E. parva* k-mer 21 y para *E. crescens* k-mer 23 con 87,6 % y 73,829% respectivamente (Tabla 4 y Figura 7).

Tabla 4. Porcentaje de genes en común entre *Emmonsia* spp. y especies cercanas filogenéticamente

Especie	Genes totales	% de genes *E. parv*a 21	% de genes *E. crescens* 23
B. dermatitidis ER3	9522	87,6	73,8
B. dermatitidis SLH	9555	87,2	68,7
H. capsulatum G186AR	8741	63,1	64
P. brasilinesis 03	7876	34,3	40,8
P. brasilinesis 01	9132	30,7	46
P.brasilinesis18	9233	31,0	41,6

De los 9522 genes de *B. dermatitidis ER3,* 8342 genes se encuentran presentes en el ensamblaje de *E. parva* k-mer 21(ANEXO 3), *B. dermatitidis* es de las especies consideradas en este estudio la más cercana a *E. parva,* seguida por *H. capsulatum* y por último *P. brasilinesis* quien más cercana *E. crescens,* para la cepa *P. brasilinesis 01* se encontró 46% de similaridad entre ellas (Tabla 4).

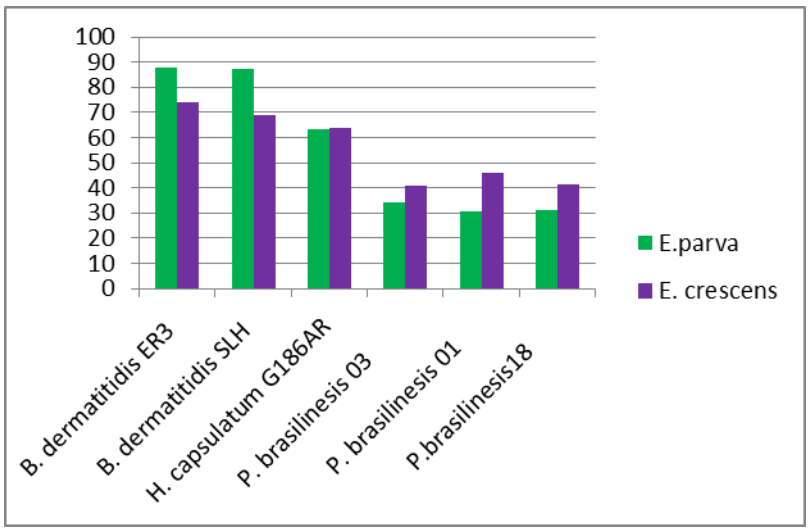


Figura 7. Porcentaje de similaridad de genes entre *Emmonsia* spp. y especies relacionadas filogenéticamente, obtenidos mediante alineamientos con BlastN (usando los parámetros por defecto).

3.3.2 AUGUSTUS v2.5.2

Se obtienen 10286 genes en *E. parva* para el ensamblaje K-mer 21 y 10640 genes para *E. crescens* k-mer 23. El número de genes obtenidos para cada k-mer se muestra en la Tabla 5 y Figura 8.

Tabla 5. Resultados de Augustus, número de genes según el k-mer.

	E. parva	*E. crescens*
kmer	**Número de Genes**	
17	10264	10581
19	10343	10682
21	10286	10681
23	10286	10640
25	10177	10698
27	10083	10700
29	9915	10711
31	9860	10726
33	9835	10739
43	9610	10684
53	9512	10709
63	9368	10677

Para todos los valores de k-mer probados Augustus predice un mayor número de genes para *E. crescens,* y la cantidad de genes predichos no cambia significativamente entre k-mer mientras que para *E. parva* el número de genes predichos tiene una tendencia a disminuir a medida que aumenta el valor del k-mer (figura 8).

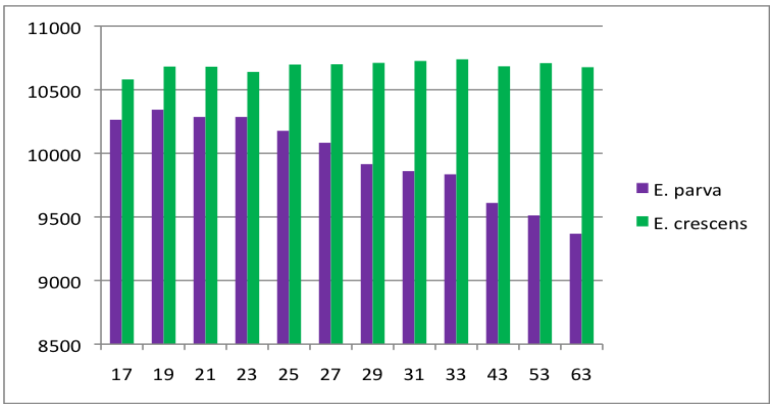


Figura 8: Histograma de número de genes predichos por Augustus para cada valor de k-mer.

3.3.2.1 Blast2Go

El programa realizó blast para los genes predichos por Augustus para los ensamblajes de *E. parva* k-mer 21 y *E.crescens* usando SOAPdenovo, para *E. parva* se encontró que de los 10286 genes, 2503 no recuperan resultados esto es el 24,3% del total de genes predichos (Tabla 6a Figura 9a). Para *E.crescens* se encontró que el 23,06% de los 10340 genes predichos no recuperan ningún registro (Tabla 7a Figura 9b).

Los hallazgos reportados como proteínas predichas o proteínas hipotéticas conservadas, para *E. parva* en su mayoría fueron encontradas en *A. dermatitidis SLH14081* con 959 proteínas predichas o proteínas hipotéticas conservadas, los resultados para otras especies relacionadas se muestran en la tabla 6b (Figura 9a). Mientras que para *E.crescens* la mayoría de los hallazgos reportados como proteínas predichas o proteínas hipotéticas conservadas fueron encontrados en *A. dermatitidis SLH14081* y *A. dermatitidis ER-3* como se muestra en la tabla 7b (Figura9b)

Tabla 6 a. Resumen de los resultados de la parte blast de Blast2Go para *E. parva*

TIPO	GENES	PORCENTAJE
Sin anotación (NA)	2503	24,33
Proteínas no especificadas (Proteínas)	485	4,72
Proteínas hipotéticas conservadas	776	7,54
Proteínas predichas	183	1,78
Proteínas especificadas (Otros)	6339	61,63
Total	10286	100

Tabla 6 b. Cantidad de proteínas predichas o proteínas hipotéticas conservadas para especies relacionadas en *E. parva*.

Especie	Porcentaje
Ajellomyces dermatitidis ER-3	34,2
Ajellomyces dermatitidis SLH14081	47,4
Ajellomyces dermatitidis ATCC	0,0
Ajellomyces capsulatum NAm1	3,2
Ajellomyces capsulatum G186AR	4,2
Ajellomyces capsulatum H88	1,6
Ajellomyces capsulatum H143	4,0
Paracoccidioides brasiliensis Pb03	0,6
Paracoccidioides brasiliensis Pb01	0,6
Paracoccidioides brasiliensis Pb18	0,2
Coccidioides posadasi	0,3
Coccidioides immitis	0,3
Uncinocarpus reesii 1704]	0,9

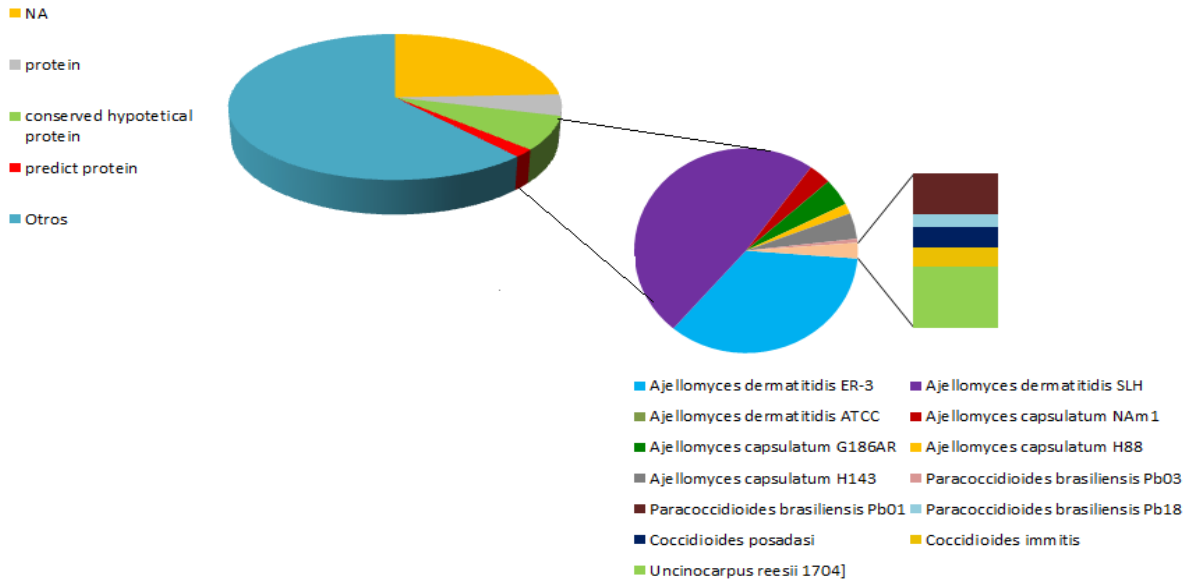


Figura 9 a. Diagrama de distribución de porcentajes según el tipo de anotación preliminar en Blast2GO para *E. parva*.

Tabla 7 a. Resumen de los resultados de la parte blast de Blast2Go para *E. crescens*

TIPO	GENES	PORCENTAJE
Sin anotación (NA)	2384	23,06
Proteínas no especificadas (Proteínas)	688	6,65
Proteínas hipotéticas conservadas	852	8,24
Proteínas predichas	243	2,35
Proteínas especificadas (Otros)	6173	59,70
Total	10340	100

Tabla 7 b. Cantidad de proteínas predichas o proteínas hipotéticas conservadas para especies relacionadas en *E. crescens*.

Especie	Porcentaje
Ajellomyces dermatitidis ER-3	20,2
Ajellomyces dermatitidis SLH14081	29,0
Ajellomyces dermatitidis ATCC	0,0
Ajellomyces capsulatum NAm1	7,9
Ajellomyces capsulatum G186AR	15,1
Ajellomyces capsulatum H88	5,7
Ajellomyces capsulatum H143	12,3
Paracoccidioides brasiliensis Pb03	2,2
Paracoccidioides brasiliensis Pb01	2,5
Paracoccidioides brasiliensis Pb18	1,7
Coccidioides posadasi	1,2
Coccidioides immitis	0,8
Uncinocarpus reesii 1704	1,5

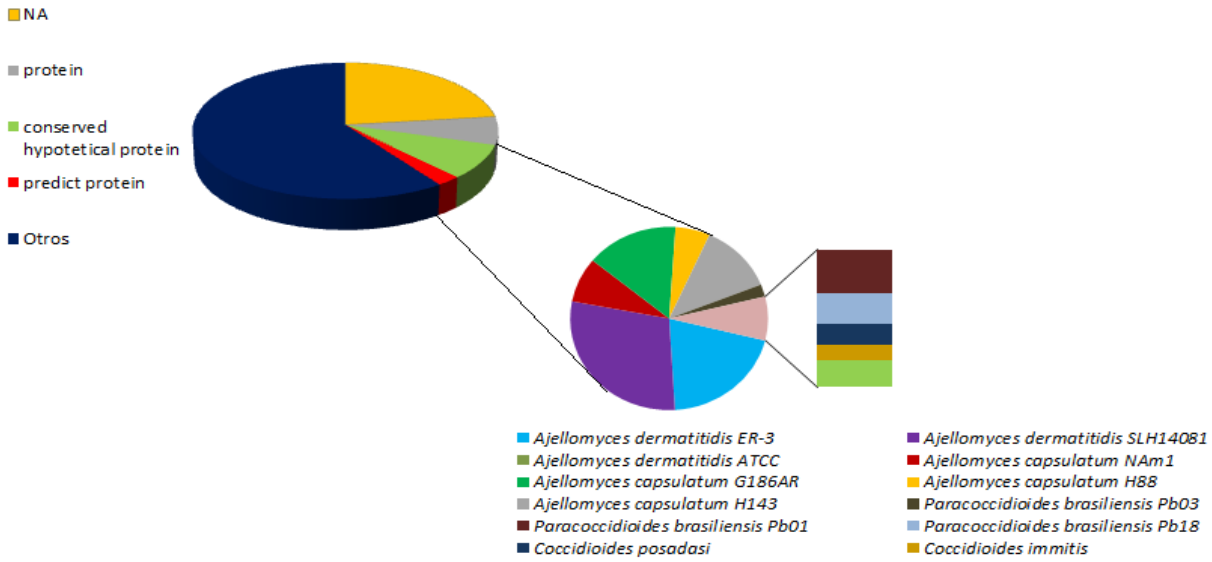


Figura 9 b. Diagrama de distribución de porcentajes según el tipo de anotación preliminar en Blast2GO para *E. crescens*.

3.3.3 FGENESH

Únicamente para los scaffolds más grandes de los ensamblajes k-mer 21 de *E. parva* y el K-mer 23 de *E. crescens* (Tabla 8).

Tabla 8: Scaffolds usados para la predicción de genes por FGENESH y genes predichos.

	Scaffold	**Longitud (Mb)**	**Genes predichos**
			FGENESH
E. crescens	scaffold18	458.835	154
	scaffold133	268.621	98
	scaffold174	286.791	91
	scaffold431	263.301	83
	scaffold462	272.324	80
E. parva	scaffold1062	184.335	57
	scaffold1060	125.772	41
	scaffold292	125.347	37
	scaffold47	120.848	34
	scaffold54	115.534	35

4 DISCUSIÓN

Las nuevas tecnologías de secuenciación como 454 y sobre todo Solexa son ahora capaces de generar secuencias por un precio mucho más razonable, pero al mismo tiempo con fragmentos mucho más cortos con respecto a los métodos anteriores, por lo cual, el ensamblaje de aquellos datos es considerablemente más complejo, ya que fragmentos tan cortos implican que el ensamblador debe ser capaz de tratar numerosos casos ambiguos que se superponen (Zerbino *et al.,* 2009).

El tiempo de computación y memoria son restricciones que limitan el uso práctico de estos algoritmos para genomas cuyo tamaño es del orden de mega bases (Simpson 2009). Particularmente por esta razón Velvet se corrió para un solo valor de k-mer para cada especie (*E. parva* 21 y *E. crescens* 23) que fue elegido tomando como referencia los resultados obtenidos en SOAPdenovo.

La principal innovación de ABySS es una distribución representativa del Bruijn graph, que permite un cálculo paralelo del algoritmo en una red de computadores (Simpson 2009), lo que permitió el uso de varios procesadores simultáneamente (los 24 procesadores disponibles). Con este programa se probaron tres valores de k-mer por especie (para *E. parva* 21, 43 y 63 y para *E. crescens* 23, 43 y 63)

Para ABySS dentro de los parámetros de optimización que se reportan en illumina 2009, se recomienda usar valores de k-mer no inferiores a la mitad de la longitud de los reads, por lo cual no es sorprendente que para ABySS los ensamblajes con mejor N50 (otros parámetros) se obtuvieron al usar los k-mer 43 y 63 que superan la mitad de la longitud de los reads de 101pb.

Sin embargo el programa más accesible (computacionalmente menos costoso) y que arrojó mejores resultados fue SOAPdenovo. Este programa acepta valores de k-mer entre 13 y 127, k-mer más grandes tienen mayor probabilidad de ser únicos en el genoma lo cual haría el gráfico más simple, sin embargo se requiere de mucha profundidad en la secuenciación para garantizar la superposición de k-

mers grandes en todos las localizaciones del genoma (SOAPdenovo 2010), entendiéndose profundidad como la superposición de muchos fragmentos en la misma localidad del genoma. En este proyecto se eligieron valores de k-mer de longitudes relativamente cortas entre 17 y 63, ya que como lo reporta (Illumina 2009), el uso de diferentes tamaños de librería genera mayor profundidad y en este caso se dispone de un solo tamaño de librería 650pb.

Por ejemplo en los ensamblajes de *E. parva* k-mer 21 y *E. crescens* k-mer 23, comparados con las estadísticas publicadas por el Broad Institute para las especies filogenéticamente cercanas, SOAPdenovo proporcionó resultados coherentes en cuanto al tamaño el genoma y un número similar de genes después de la anotación, usando el programa Augustus los ensamblajes obtenidos contienen un número de genes cercano al esperado (Figura 10, Figura 11, Anexo 4).

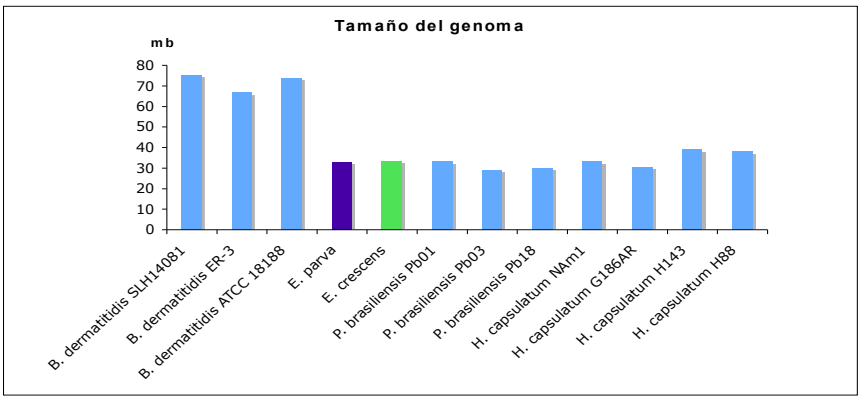


Figura 10. Estadísticas publicada por el Broad Institute para el tamaño de los genomas de especies relacionadas (azul) a *Emmonsia* spp. y los resultados obtenidos para los ensamblajes de *E. parva* (k-mer 21; morado) y *E. crescens* (k-mer 23; verde).

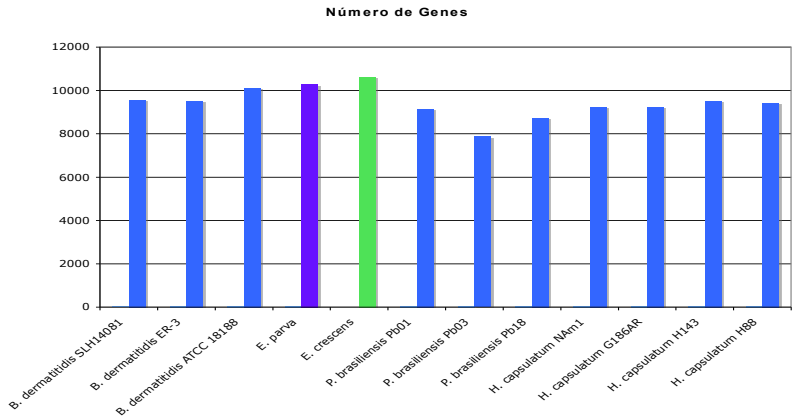


Figura 11. Estadística publicada por el Broad Institute del número de genes para las especies relacionadas (azul) a *Emmonsia* spp. y los resultados obtenidos usando SOAPdenovo y Augustus para *E. parva* (k-mer 21; morado) y *E. crescens* (k-mer 23; verde).

En general los hongos unicelulares y filamentosos contienen pocas regiones repetitivas en sus genomas (30-90Mb) y por lo tanto son candidatos adecuados para probar ensamblajes *de novo* a partir de short reads (Nowrousian 2010). Sin embargo, la especie *B. dermatitidis* relacionada con *Emmonsia* spp. tiene una cantidad importante de regiones repetitivas o de baja complejidad a lo cual puede atribuirse la diferencia en el tamaño de genoma respecto a las otras especies (Figura 10).

Esta es una de las principales razones por las que se realizó un ensamblaje *de novo*, ya que la especie más cercana filogenéticamente *era B. dermatitidis* y su genoma no es lo suficientemente similar al de *Emmonsia* spp. para generar un buen ensamblaje de referencia.

A pesar de que los genomas de *B. dermatitidis* tienen un tamaño que duplica el de los ensamblajes de *Emmonsia* spp. estos tienen más genes predichos que los reportados para los genomas de *B. dermatitidis,* lo cual sugiere que los ensamblajes tienen completo el "gene space", parte o partes del genoma que contienen los genes y generalmente se caracteriza por un intervalo limitado del contenido GC, que excluye gran parte del ADN repetitivo del genoma (Carels et al., 1995).

5 REFERENCIAS BIBLIOGRÁFICAS

Bawdona Roger E., Garrisonb Robert G., Finac Louis R (1972) “Deoxyribonucleic Acid Base Composition of the Yeastlike and Mycelial Phases of Histoplasma capsulatum and Blastomyces dermatitidis” *J. Bacteriol. August 1972 vol. 111 no. 2 593-596*

Blanco, E., Parra, G., Guigó, R. (2002). "Using geneid to identify genes." En: Current Protocols in Bioinformatics. Vol. 1, Unit 4.3. (eds. Baxevanis, A.D., Davison, D.B.) John Wiley & Sons, New York.

Bowman, B.H., White, T.J., Taylor, J.W. (1996). “Human pathogeneic fungi and their close nonpathogenic relatives.” *Mol. Phylogenet. Evol*. 6:89-96.

BROAD institute, http://www.broadinstitute.org

Carels, N., Barakat, A. & Bernardi, G. (1995). The gene distribution of the maize genome.Proc. Natl. Acad. Sci. USA 92, 11057-11060.

Carlile, M., *et al*. (2004). ”The Fungi”, Second edition, Elsevier Academic press, Chapter 7: Parasities and mutualistic symbionts, p. 435.

Chaisson, M.J., Brinza, D., Pevzner, P.A. (2009). “De novo fragment assembly with short mate-paired reads: Does the read length matter?” *Genome Res*. 19:336-346.

Conesa Ana, Götz Stefan, *et al*. (2005). “Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research”. *Bioinformatics 21 (18): 3674-3676.*

Conway, TC. Bromage, AJ. (2011). Succinct data structures for assembling large genomes. *Bioinformatics* 27:479-486.

Cruveiller, S., Jabbari, K., Clay, O., Bernardi, G. (2003). “Compositional features of eukaryotic genomes for checking predicted genes.” *Brief. Bioinf.* 4:43-52.

D., Pevzner, P.A. (2009). “De novo fragment assembly with short mate-paired reads: Does the read length matter?” *Genome Res*. 19:336-346.

Diez, S., *et al*. (1999). “PCR with *Paracoccidioides brasiliensis* specific primers: potential use in ecological studies”. *Rev. Inst. Med. Trop*. *Sao Paulo* 41, 351–358.

García, A.M., Hernández, O., Aristizábal, B.H. *et al*. (2009). “Identification of genes associated with germination of conidia to form mycelia in the fungus *Paracoccidioides brasiliensis*.” *Biomedica* 29:403-412.

García, A.M., Hernández, O., Aristizábal, B.H. *et al*. (2010). “Gene expression analysis of *Paracoccidioides brasiliensis* transition from conidium to yeast cell.” *Med. Mycol*. 48:147-154.

Drummond, A.J., Ashton, B., Buxton, S., *et al*. (2011). Geneious v5.4, http://www.geneious.com .

Illumina (2009). “Sequencing User Guide for single-read and pair-end reads Sequencing”.
http://sfgf.stanford.edu/documents/solexa_docs/User_Manuals/SequencingUserGuide_1006747_RevA.pdf

Illumina (2008). “Preparing Samples for Sequencing Genomic ADN”. en: http://www.ucl.ac.uk/wibr/services/solexa/Sample_Prep.pdf

Haas, B.J., Kamoun, S., Zody, M.C. *et al.* (2009). “Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*.” *Nature* 461:393-398.

Hejtmánek, M. (1985). “Dimorphism in *Chrysosporium parvum*.” En: *Fungal Dimorphism, with Emphasis on Fungi Pathogenic for Humans*, ed. Szanislo, P.J., Plenum Press, New York & London, pp. 237-261.

Klein, B.S., Tebbets, B. (2007). “Dimorphism and virulence in fungi.” *Curr. Opin. Microbiol.* 10:314-319.

Korf, I. (2004). “Gene finding in novel genomes.” *BMC Bioinformatics* 5:59.

Langmead, B., Trapnell, C., Pop, M., Salzberg, M.L. (2009). “Ultrafast and memory-efficient alignment of short ADN sequences to the human genome.” *Genome Biol.* 10:R25.

Li, R., Zhu, H., Ruan, J., *et al.* (2010). “De novo assembly of human genomes with massively parallel short read sequencing.” *Genome Res*. 20:265-272.

Marini, M.M., Zanforlin, T., Santos, P.C. *et al.* (2010). “Identification and characterization of Tc1/mariner-like ADN transposons in genomes of the pathogenic fungi of the Paracoccidioides species complex.” *BMC Genomics* 11:130.

Morgan, M., Anders, S., Lawrence, M., *et al.* (2009). “ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data.” *Bioinformatics* 25:2607-2608.

NCBI Taxonomy, www.**ncbi**.nlm.nih.gov

Neafsey, D.E., Barker, B.M., Sharpton, T.J. (2010). “Population genomic sequencing of Coccidioides fungi reveals recent hybridization and transposon control.” *Genome Res*. 20:938-946.

Nowrousian, M., Stajich, J.E., Chu, M., et al. (2010). “*De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis.” *PLoS Genet.*, 6(4):e1000891.

Peterson, S.W., Sigler, L. (1998). “Molecular genetic variation in *Emmonsia crescens* and *Emmonsia parva*, etiologic agents of adiaspiromycosis, and their phylogenetic relationship to *Blastomyces dermatitidis* (*Ajellomyces dermatitidis*) and other systemic fungal pathogens.” *J. Clin. Microbiol.* 36:2918–2925.

Pop, M. (2009). “Genome assembly reborn: recent computational challenges.” *Brief. Bioinform.* 10:354-366.

Richard, G.F., Kerrest, A., Dujon, B. (2008). “Comparative genomics and molecular dynamics of ADN repeats in eukaryotes.” *Microbiol. Mol. Rev.* 72:686-727.

Sharpton, T.J., Stajich, J.E., Rounsley, S.D., *et al.* (2009). “Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives.” *Genome Res.* 19:1722-1731.

Simpson, J.T., Wong, K., Jackman, S.D., *et al.* (2009). “ABySS: a parallel assembler for short read sequence data.” *Genome Res.* 19:1117-1123.

Stanke, M., Tzvetkova, A., Morgenstern, B. (2006). “AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome.” *Genome Biol.* 7:S11.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., Borodovsky, M. (2008). “Gene prediction in novel fungal genomes using *ab initio* algorithm with unsupervised training.” *Genome Res*. 18: 1979-1990.

Zaragoza, O., García-Rodas, R., Nosanchuk, J.D. *et al.* (2010). “Fungal cell gigantism during mammalian infection.” *PloS Pathog*. 6(6):e1000945.

Zerbino, D.R., Birney, E. (2008). “Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs.” *Genome Res.*, 18:821-829.

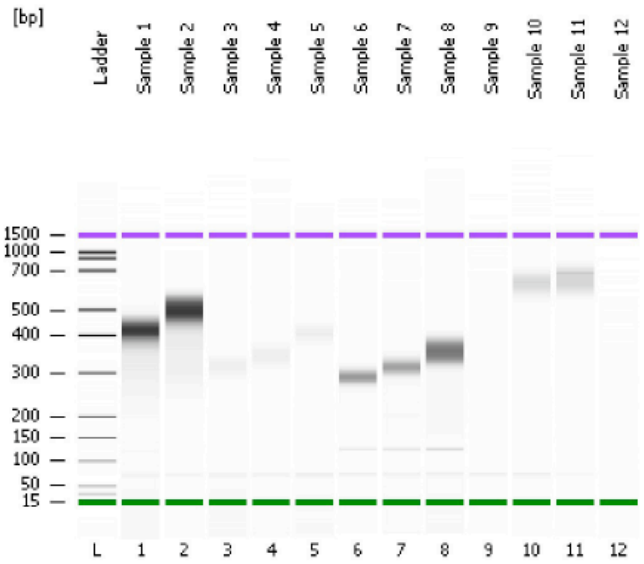
Zerbino, D.R., McEwen, G.K., Margulies, E.H., Birney, E. (2009). “Pebble and Rock Band: heuristic resolution of repeats and scaffolding in the Velvet short-read *de novo* assembler.” *PLoS One*, 4(12):e8407.

ANEXOS

Assay Class: DNA 1000
Data Path: C:\...-29\2100 expert_DNA 1000_DE02000247_2010-07-29_12-50-45.xad

Created: 7/29/2010 12:50:45 PM
Modified: 7/29/2010 1:44:41 PM

Electrophoresis File Run Summary



Instrument Information:

Instrument Name: DE02000247 Firmware: C.01.069
Serial#: DE02000247 Type: G2938A

Assay Information:

Assay Origin Path: C:\Program Files\Agilent\2100 bioanalyzer\2100 expert\assays\dsDNA\DNA 1000 Series II.xsy

Title: DNA 1000 Series II

Version: 2.0

Assay Comments:

Chip Information:

Chip Lot:

Reagent Kit Lot:

Chip Comments:

Assay Class: DNA 1000
Data Path: C:\...-29\2100 expert_DNA 1000_DE02000247_2010-07-29_12-50-45.xad

Created: 7/29/2010 12:50:45 PM
Modified: 7/29/2010 1:44:41 PM

Electrophoresis Assay Details

General Analysis Settings

Number of available sample and ladder wells (max.) : 13
Minimum visible range [s] : 30
Maximum visible range [s] : 129
Start analysis time range [s] : 30
End analysis time range [s] : 128.95
Ladder Concentration [ng/µl] : 4
Uses standard area for ladder fragments
Lower Marker Concentration [ng/µl] : 4.2
Upper Marker Concentration [ng/µl] : 2.1
Used upper marker for quantitation
Standard curve fit is Point to Point
Show data aligned to lower and upper marker

Integrator Settings

Integration start time [s] : 30
Integration end time [s] : 128.95
Slope Threshold : 0.5
Height Threshold [FU] : 20
Area Threshold : 0.1
Width Threshold [s] : 1
Baseline Plateau [s] : 0.5

Filter Settings

Filter width [s] : 0.5
Polynomial order : 4

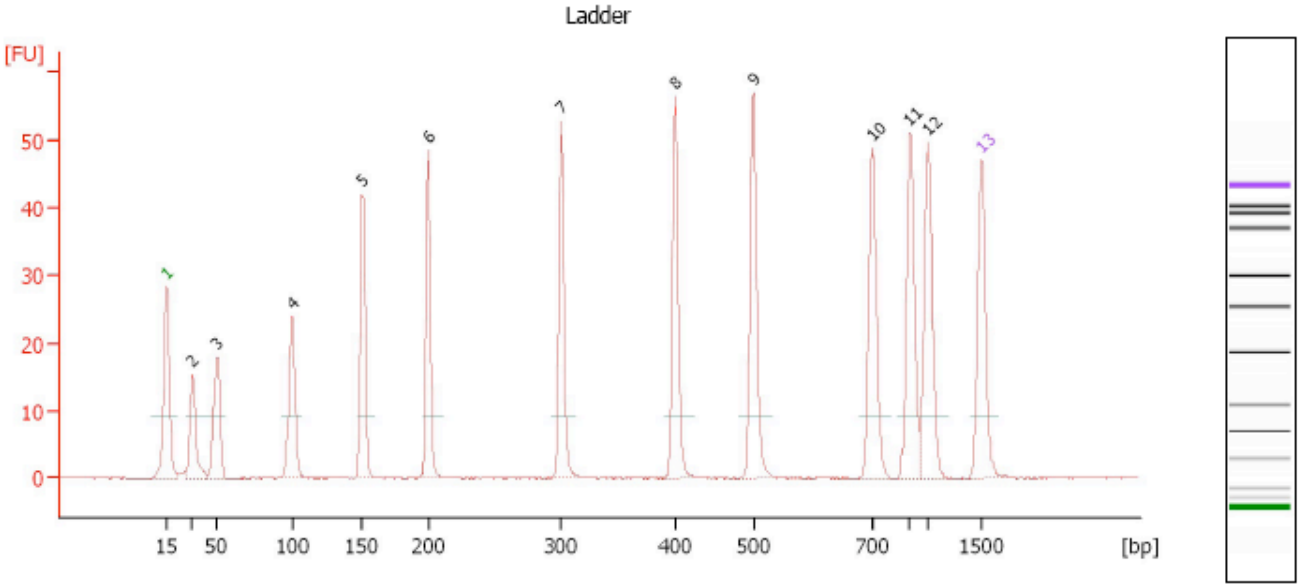
Ladder

Ladder Peak	Size	Area
1	15	25
2	25	26
3	50	34
4	100	41
5	150	45
6	200	52
7	300	63
8	400	76
9	500	83
10	700	88
11	850	86
12	1000	90
13	1500	52

Assay Class: DNA 1000
Data Path: C:\...-29\2100 expert_DNA 1000_DE02000247_2010-07-29_12-50-45.xad

Created: 7/29/2010 12:50:45 PM
Modified: 7/29/2010 1:44:41 PM

Electropherogram Summary



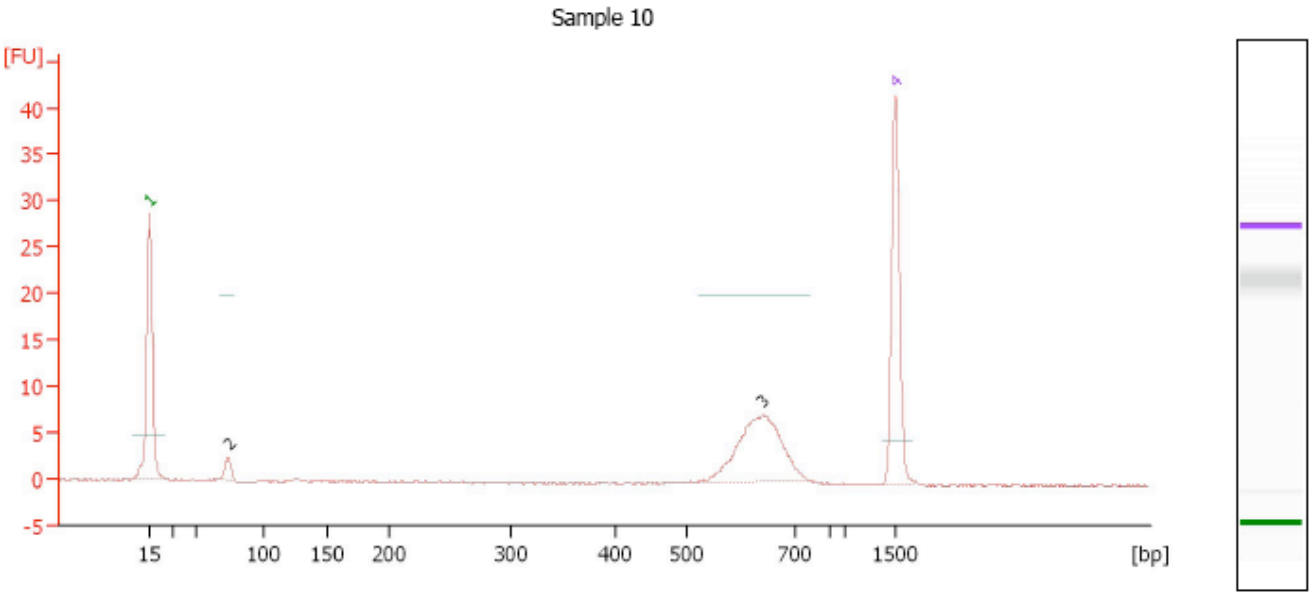
Peak table for Ladder

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	25	4.00	242.4	Ladder Peak
3	50	4.00	121.2	Ladder Peak
4	100	4.00	60.6	Ladder Peak
5	150	4.00	40.4	Ladder Peak
6	200	4.00	30.3	Ladder Peak
7	300	4.00	20.2	Ladder Peak
8	400	4.00	15.2	Ladder Peak
9	500	4.00	12.1	Ladder Peak
10	700	4.00	8.7	Ladder Peak
11	850	4.00	7.1	Ladder Peak
12	1,000	4.00	6.1	Ladder Peak
13	1,500	2.10	2.1	Upper Marker

Assay Class: DNA 1000
Data Path: C:\...-29\2100 expert_DNA 1000_DE02000247_2010-07-29_12-50-45.xad

Created: 7/29/2010 12:50:45 PM
Modified: 7/29/2010 1:44:41 PM

Electropherogram Summary Continued ...



Overall Results for sample 10 : **Sample 10**

Peak table for sample 10 : **Sample 10**

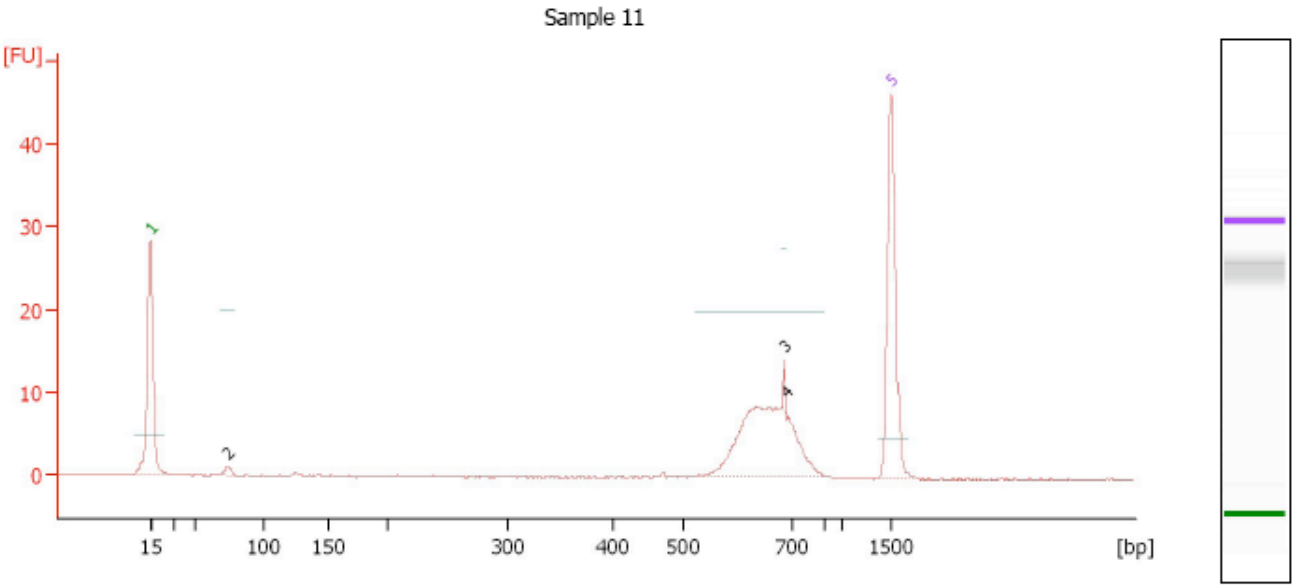
Peak		Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	◀	15	4.20	424.2	Lower Marker
2		74	0.24	5.0	
3		639	2.23	5.3	
4	▶	1,500	2.10	2.1	Upper Marker

Number of peaks fou

Assay Class: DNA 1000
Data Path: C:\...-29\2100 expert_DNA 1000_DE02000247_2010-07-29_12-50-45.xad

Created: 7/29/2010 12:50:45 PM
Modified: 7/29/2010 1:44:41 PM

Electropherogram Summary Continued ...



Overall Results for sample 11 : Sample 11

Peak table for sample 11 : Sample 11

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	73	0.10	2.2	
3	686	3.12	6.9	Number of peaks fou
4	689	0.07	0.2	
5	1,500	2.10	2.1	Upper Marker

ANEXO 2

SOAPdenovo	Kmer size	Scaffolds	Contigs	Sum up (bp)	Average length	Scadffolds& Singleton	Sum up (bp)	Average length	Longest (bp)	Scaffold N50	Scaffold N90
Emmonsia parva	17	3003	66307	41685198	13881	8737	42524360	4867	190021	31429	5937
	19	2616	30397	34233989	13086	4822	34565193	7168	203463	32230	5740
	21	2787	21490	32825334	11778	5576	33371694	5984	193220	30060	4590
	23	3026	19531	32969869	10895	6671	33836463	5072	284823	29272	3954
	25	3420	20176	33800392	9883	8230	34986166	4251	190521	27239	3387
	27	3706	22229	35005021	9445	9726	36508840	3753	282079	27293	3160
	29	4079	25199	35798817	8776	11535	37649125	3263	280641	25640	2769
	31	4331	28773	36744126	8483	13252	38834751	2930	196086	25463	2623
	33	4347	32495	37580345	8645	14702	39891972	2713	188599	25169	2618
	43	2870	42843	40832464	14227	12672	43567760	3438	269541	32563	5391
	53	1680	121219	39671804	23614	93116	46847845	503	266249	40734	54
	63	1433	96223	36981126	25806	78503	44323841	564	357305	41960	64
Emmonsia crescens	17	3008	71381	43446278	14443	14998	45215925	3014	227079	35600	4879
	19	1506	34309	36394611	24166	4862	36881464	7585	484033	72036	13610
	21	1396	24101	34265149	24545	4102	34657491	8448	451397	82194	13698
	23	1473	20140	33470667	22722	4434	33903273	7646	488252	87741	12098
	25	1610	18870	33295251	20680	4865	33746468	6936	410195	86624	10953
	27	1719	18622	33356499	19404	5602	33875328	6047	405321	87920	9927
	29	1734	18800	33359270	19238	6293	33918762	5389	403812	92864	9047
	31	1706	19260	33345061	19545	6970	33969526	4873	400797	89630	8928
	33	1659	20013	33427350	20149	7663	34159473	4457	399482	92729	8016
	43	1108	19247	34261799	30922	5949	34898968	5866	467643	93225	15894
	53	758	28734	33867390	44679	16897	35060896	2074	332992	100449	20013
	63	702	28127	33112535	47168	19435	34565180	1778	332761	101039	19573

ANEXO 3

***E. parva* k=21**

Número de genes

Kmer size	BlastoER3	BlastoSLH	Pb18	Pb03	Pb01	HistoG186AR
19	8341	8333	2750	2719	2813	5899
21	8342	8335	2737	2704	2806	5867
23	8320	8313	2713	2677	2766	5824
25	8299	8287	2690	2661	2768	5784
27	8276	8242	2661	2628	2727	5751
29	8231	8198	2639	2606	2701	5704
31	8209	8177	2627	2596	2707	5685
33	8203	8163	2611	2581	2691	5656
43	8078	8038	2554	2527	2633	5534
53	7987	7957	2517	2490	2594	5442
63	7888	7855	2475	2440	2543	5362

***E. crescens* k=23**

Número de genes

k-mer	BlastoER3	BlastoSLH	PB18	PB01	PB03	BlastoSLH
19	6480	6546	3624	3716	3610	6546
21	6513	6579	3627	3721	3615	6579
23	7030	6569	3634	3725	3616	6569
25	7094	6600	3859	3728	3617	6600
27	7215	6605	3873	3723	3614	6605
29	7294	6624	3890	3730	3618	6624
31	7265	6616	3895	3729	3618	6616
33	7338	6621	3896	3730	3616	6621
43	7555	6626	3904	3726	3615	6626
53	7634	6619	3925	3723	3612	6619
63	7599	6599	3897	3705	3597	6599

Porcentaje

Kmer size	BlastoER3 (9522 genes)	BlastoSLH (9555 genes)	Pb18 (8741 genes)	Pb03 (7876 genes)	Pb01 (9132 genes)	Histo G186AR (9233 genes)
19	87,597	87,211	31,461	34,523	30,804	63,890
21	87,608	87,232	31,312	34,332	30,727	63,544
23	87,377	87,002	31,038	33,989	30,289	63,078
25	87,156	86,729	30,775	33,786	30,311	62,645
27	86,915	86,259	30,443	33,367	29,862	62,287
29	86,442	85,798	30,191	33,088	29,577	61,778
31	86,211	85,578	30,054	32,961	29,643	61,573
33	86,148	85,432	29,871	32,770	29,468	61,259
43	84,835	84,123	29,219	32,085	28,833	59,937
53	83,879	83,276	28,795	31,615	28,406	58,941
63	82,840	82,208	28,315	30,980	27,847	58,074

Porcentaje

K-mer	BlastoER3 (9522 genes)	BlastoSLH (9555 genes)	Pb18 (8741 genes)	Pb01 (9132 genes)	Pb03 (7876 genes)	BlastoSLH (9555 genes)
19	68,053	68,509	41,460	40,692	45,835	68,509
21	68,399	68,854	41,494	40,747	45,899	68,854
23	73,829	68,749	41,574	40,791	45,912	68,749
25	74,501	69,074	44,148	40,823	45,924	69,074
27	75,772	69,126	44,308	40,769	45,886	69,126
29	76,602	69,325	44,503	40,845	45,937	69,325
31	76,297	69,241	44,560	40,834	45,937	69,241
33	77,064	69,294	44,572	40,845	45,912	69,294
43	79,343	69,346	44,663	40,802	45,899	69,346
53	80,172	69,273	44,903	40,769	45,861	69,273
63	79,805	69,063	44,583	40,572	45,670	69,063