UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS BIOLÓGICAS

Ensamblaje del genoma de una levadura cervecera tipo lager (*Saccharomyces cerevisiae* var. *uvarum)*

Por

LBG. RAMIRO ELIZONDO GONZÁLEZ

Como requisito parcial para obtener el Grado de
MAESTRÍA EN CIENCIAS CON ACENTUACIÓN
EN MICROBIOLOGÍA

Cd. Universitaria Noviembre 2014

Ensamblaje del Genoma de una Levadura Cervecera Tipo Lager (*Saccharomyces cerevisiae* var. *uvarum)*

Comité de Tesis

Dr. Benito Pereyra Alférez
Director

Dr. Carlos F. Sandoval Coronado
Secretario

Dr. Luis J. Galán Wong
Vocal

Dr. Hugo A. Luna Olvera
Vocal

Dra. Lilia H. Morales Ramos
Vocal

Dr. Luis C. Damas Buenrostro
Asesor Externo

Cd. Universitaria Noviembre 2014

El presente trabajo se llevó a cabo en el Laboratorio L4 del Instituto de Biotecnología. FCB/UANL, en la Unidad de Genómica del CIDICS/UANL y en el Departamento de Investigación y Desarrollo de Cervecería Cuauhtémoc Moctezuma bajo la asesoría del Dr. Benito Pereyra Alférez, Dra. Rocío Ortiz López y Dr. Luis Cástulo Damas Buenrostro, respectivamente.

Proyecto apoyado por la Convocatoria de Proyectos de Investigación Tecnológica del CONACYT

AGRADECIMIENTOS

A **Dios**, por darme la oportunidad de hacer lo que amo y poner las oportunidades en mi camino.

Al **CONACYT**, por otorgarme la beca para realizar mis estudios de maestría

Al **Dr. Benito Pereyra Alférez**. Muchas gracias por todo su apoyo, consejos y tiempo para la realización de este proyecto.

Al **Dr. Carlos F. Sandoval Coronado** por su apoyo a lo largo del proyecto.

Al **Dr. Luis J. Galán Wong** por su apoyo a lo largo del proyecto.

Al **Dr. Hugo A. Luna Olvera** por su apoyo a lo largo del proyecto.

A la **Dra. Lilia H. Morales Ramos** por su apoyo a lo largo del proyecto.

Al **Dr. Luis C. Damas Buenrostro**. Muchas gracias por su confianza y por tener fe en mí; por la atención al detalle, el estímulo, guía y consejos durante todo el proyecto e incluso en lo personal.

A la **Dra. Rocío Ortiz López**. Muchas gracias por su confianza y apoyo durante todo el proyecto y fuera del él. No podría terminar de agradecerle no solo por sus enseñanzas en el área de las ciencias, sino por, literalmente, cambiar el porvenir de mi vida.

A la **Universidad Autónoma de Nuevo León**, y la **Facultad de Ciencias Biológicas**, por impulsarme como alumno y proveer los medios para mi superación profesional

A **Cervecería Cuauhtémoc Moctezuma**, por todo el apoyo para la realización de este proyecto.

A mi esposa **Dalia**. Gracias no solo por tu apoyo para realizar la maestra, sino por impulsarme a no dejar de superarme. Por ser un ejemplo de esfuerzo y dedicación. Por estar conmigo en las buenas y en las malas (en especial en las tres cirugías de hace unos meses) Gracias por ~~aguantarme~~ amarme, eres lo mejor que me ha pasado.

A **mis padres**, Julián y Susana, por todo su apoyo, amor y comprensión a lo largo de mi vida.

A **mis hermanos** Susy, Julián, Betty y Gaby por todo su apoyo durante toda la vida; mis **cuñados** Kary y Fer; mis **sobrinos** Rodrigo, Eugenia y (en unos meses) Bruno por alegrarme los días con solo verlos (y saber de ellos); a todos mis **tíos**, en especial a la Güera y More

A mis **abuelos** que ya se adelantaron, por ser un ejemplo de honradez y rectitud, por su cariño siempre. Los extraño

A mis **suegros** Astor y Yolanda: Por ser siempre tan amables y haberme aceptado en su familia. No me pudieron tocar mejores suegros; mis **cuñados** Astor y Fer y las **abuelas**.

A mis amigos, compañeros y colegas de **UG** (Arcadio, Ram, Laura, Adriana, Yadira, Ally, Geo, Liz, Ernesto, Pepe, Carmen y Carlos), del **L4** (Clau, Hugo, Zanya, Esme, Laura, Fátima, Will, Jessy, Chuy, Saúl y Astrid) y de **CCM / Heineken** (Esmeralda, Marce, Jan-Maarten Geertman y Marcel van den Broek – heel erg bedankt!) por tantos momentos divertidos, estresantes y de aprendizaje. Muchas gracias por todo su apoyo.

A todos mis **amigos** (en especial a Ricardo, Rafa, Oscar, Prado y R. Arturo), gracias por darme ánimos, escribir miles de mensajes sin parar y por tan buenos momentos durante tantos años.

A la **música** (en especial a Bach, Beethoven, Mozart, Vivaldi, Nirvana, PJ, RHCP, Tool, FF, R.E.M., 311 y muchas otras), por acompañarme durante tantas horas frente a la computadora.

Finalmente, a ti **Yorch**, por hacer que me diera cuenta de lo valiosa que es la vida. Creo que ningún amigo ha tenido (ni tendrá) un impacto tan trascendente en mi vida como lo tuviste tú. Gracias por tu invaluable herencia. Te extrañamos mucho.

DEDICATORIAS

Para Dalia,
Yo contigo y tú conmigo.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTA DE SÍMBOLOS Y ABREVIATURAS

790: Cepa cervecera tipo lager en estudio.

790cer: Subgenoma *Saccharomyces cerevisiae* en 790.

790eub: Subgenoma *Saccharomyces eubayanus* en 790.

CCM: Cervecería Cuauhtémoc Moctezuma.

Contig: Conjunto de dos o más lecturas ensambladas.

DBG: Grafo De Bruijn.

Gap: Secuencia nucleotídica no conocida que une dos *contigs* para formar *scaffolds*, representada por 'Ns'

Mpb: Mega pares de bases ($1x10^6$ pb).

OLC: Superposición-disposición-consenso.

pb: Pares de bases.

PCR: Reacción en Cadena de la Polimerasa.

PFGE: Gel de electroforesis de campos pulsantes.

Q30: Valor de calidad de nucleótidos secuenciados; un posible error cada $1x10^3$ bases.

Q40: Valor de calidad de nucleótidos secuenciados; un posible error cada $1x10^4$ bases.

Scaffold: Conjunto de dos o más contigs ensamblados, con ubicación y dirección.

Scer: *Saccharomyces cerevisiae*

Seub: *Saccharomyces eubayanus*

Software: Programa de cómputo.

RESUMEN

Las levaduras son de gran importancia en la industria cervecera ya que estas llevan a cabo la fermentación de cervezas tipo ale y lager. Éstas últimas representan el mayor mercado en ventas. El genoma nuclear de dichas levaduras suele ser más complejo que el de *Saccharomyces cerevisiae*. Esta complejidad se debe a que el genoma de las tipo lager está conformado por la combinación de dos o tres tipos de levaduras, principalmente *S. cerevisiae* y *Saccharomyces eubayanus*. Los datos del genoma a ensamblar y analizar fueron obtenidos por la secuenciación masiva en las plataformas FLX 454 Titanium (Roche), Ion Torrent e Illumina, los cuales, permiten determinar la secuencia nucleotídica del material genético con precisión y gran cobertura. Dichas secuencias fueron ensambladas por medio de los diversos programas computacionales especializados en bioinformática. Los resultados del ensamblaje mostraron que el genoma nuclear de la levadura se compone los subgenomas de *S. cerevisiae* y *S. eubayanus*, con al menos 32 cromosomas y un tamaño total de 22.7 Mpb comprendidos en 133 *scaffolds*, cuyo tamaño fue, en promedio > 10 kpb. Solo 65 rebasaron esta talla molecular. La versión final del ensamblaje contiene alrededor de 400,000 Ns (*gaps*) distribuidas en sets a lo largo del genoma, 5 de los cuales fueron ubicados, amplificados y validados para dilucidar la secuencia correspondiente. Así mismo, se observaron translocaciones en seis cromosomas diferentes, así como diferente número de copias de cada uno.

SUMMARY

Yeasts are of great relevance in the brewing industry as these perform the fermentation of ale and lager beer styles. The latter represent the largest market in sales. The nuclear genome of said yeasts is usually more complex than that of *Saccharomyces cerevisiae*. This complexity is due to the fact that lager style genomes is form by combining two or three types of yeast, especially *S. cerevisiae* and *Saccharomyces eubayanus.* The genome data to assemble and analyze were obtained by massive sequencing in 454 FLX Titanium (Roche), Ion Torrent and Illumina platforms, which allow determining the nucleotide sequence of the genetic material with precision and high coverage. These sequences were assembled using various computer programs specialized in bioinformatics. Assembly results showed that the nuclear genome of the lager yeast is composed by *S. cerevisiae* and *S. eubayanus* subgenomes, with at least 32 chromosomes and a total size of 22.7 Mbp included in 133 *scaffolds*, with an average size > 10 kbp. Only 65 exceeded that molecular size. The final version of the assembly contains about 400,000 Ns (*gaps*) distributed along the genome, five of which were located, amplified and validated to elucidate the corresponding sequence. Furthermore, translocations are observed in six different chromosomes, as well as different copy number of each.

1. INTRODUCCIÓN

En la mayoría de las sociedades, las bebidas fermentadas tienen una gran importancia debido a su impacto económico y cultural. El desarrollo de tecnologías de fermentación data del año 7000 A.C. en China, donde se descubrió evidencia de una bebida fermentada producida en aquel año (McGovern *et al.*, 2004). A partir de entonces, los procesos se diversificaron y se piensa que la tecnología para la creación de la cerveza es tan antigua como el vino (Legras *et al.*, 2007).

Distintas especies del género *Saccharomyces* han sido aisladas y probadas para la fermentación de cerveza. El complejo de *Saccharomyces* sensu stricto comprende seis especies hermanas: *S. cerevisiae, S. bayanus, S. cariocanus, S. kudriavzevii, S. mikatae y S. paradoxus* (Naumova *et al.*, 2003). Se ha observado que las levaduras cerveceras tipo lager consisten en una especie híbrida con *S. bayanus* y *S. cerevisiae* como sub-genomas. Esto les confiere un rearreglo importante tanto en la posición, número de genes y tamaño del genoma completo (Nakao *et al.*, 2009).

Gracias al advenimiento de nuevas tecnologías de secuenciación masiva como la plataforma Illumina, a la fecha se tienen reportados 34,636 genomas de distintas especies, 12 de las cuales corresponden al género *Saccharomyces* (http://www.ncbi.nlm.nih.gov/genome/browse/ - revisado el día 5 de noviembre de 2014). Esa gran cantidad de información ha permitido el desarrollo de la *Genómica Comparativa*, la cual toma como base genomas de organismos cercanos evolutivamente para un más rápido y preciso procesamiento de datos de secuencias nucleotídicas de nuevos organismos, asignándoles una probable organización cromosómica y permitiendo encontrar genes nuevos que confieren propiedades características a los distintos organismos.

En el presente trabajo ensamblamos el genoma nuclear de una levadura tipo lager siendo que cada cepa cervecera es considerada única en su estructura y eficiencia para la

fermentación, se ha dilucidado el tamaño, secuencia completa y organización de sus cromosomas, así como el número de copias de secuencias de interés comercial.

2. ANTECEDENTES

2.1. Secuenciación de genomas

La publicación de la estructura de doble hélice del ADN en 1953 (Watson y Crick, 1953) llevó a subsecuentes estudios para conocer su secuencia. La secuenciación de los primeros fragmentos de ácidos nucleicos data de 1972, cuando el equipo de trabajo de Fiers y cols. lograron secuenciar el primer gen de ARN completo del bacteriófago MS2 (Jou *et al.*, 1972) y su genoma completo cuatro años después (Fiers *et al.*, 1976). En 1977, Sanger desarrolló la técnica de terminadores de la replicación, en la que se utilizan nucleótidos 2'3'-dideoxi (Sanger *et al.*, 1977 a). Con dicha técnica se logró la secuenciación completa del primer genoma de ADN, el bacteriófago φ X174 de 5,375 pb (Sanger *et al.*, 1977 b). No fue sino hasta 1986 cuando el laboratorio de Leroy Hood y cols. con financiamiento de Applied Biosystems, desarrolló el primer secuenciador automatizado para el análisis de ADN, el cual estaba basado en los principios de la técnica de Sanger, con la adición de fluoróforos para la detección de los nucleótidos (Smith *et al.*, 1986). Ese mismo año, Applied Biosystems sacó al mercado el modelo 370A DNA Sequencing System, revolucionando los campos de descubrimiento genético, genética comparativa y permitiendo la digitalización y análisis computacional de los datos de secuenciación.

En 1995, Craig Venter y cols. publicaron el primer genoma completo de un organismo de vida libre, la bacteria *Haemophilus influenzae*, cuyo genoma consta de 1,830,137 pb (Fleischmann *et al.*, 1995) y siendo el primer uso publicado de la secuenciación tipo *shotgun* de un genoma completo. En 2001, utilizando también la técnica shotgun, se publicó el borrador del genoma humano completo (Lander *et al.*, 2001; Venter *et al.*, 2001).

Los nuevos métodos para determinar la secuencia de ADN desarrollados de mediados a finales de 1990s fueron el inicio de la entonces llamada secuenciación de nueva generación (Next Generation Sequencing), hoy denominada secuenciación de segunda generación. En 1996, la metodología de la pirosecuenciación (en inglés, Pyrosequencing) fue publicada, reportando un ensayo de detección de pirofosfato inorgánico por una enzima luminométrica (luciferasa de luciérnaga) (Ronaghi *et al.*, 1996). Posteriormente, en el año 2000, Lynx Therapeutics publicó un sistema mediado por ligación a adaptadores en perlas denominado Secuenciación Masivamente Paralela (MPSS), sin embargo, el MPSS no salió al mercado (Brenner *et al.*, 2000). Dicha tecnología sirvió como base para el primer secuenciador de segunda generación, el pirosecuenciador 454 de Life Sciences, cuyas ventas comenzaron en 2004 (Margulies *et al.*, 2005). El secuenciador 454 aumentó 50 veces los rendimientos y disminuyó hasta seis veces los costos de cada corrida comparado con la secuenciación automatizada tipo Sanger. Las siguientes plataformas en llegar al mercado fueron Solexa, de Illumina y SOLiD de Applied Biosystems (Schuster, 2008), teniendo cada una de las plataformas sus ventajas comparativamente (Liu *et al.*, 2012).

La Figura 1 muestra una línea de tiempo de los primeros genes y organismos secuenciados, así como la aparición de las tecnologías de secuenciación.

Tabla 1: Comparación de las plataformas de secuenciación. (Adaptado de Liu *et al*., 2012)

Secuenciador	454 GS FLX	HiSeq 2500	SOLiDv4	Sanger 3730xl
Mecanismo	Pirosecuenciación	Por síntesis	Ligación	Terminación por dideoxi
Longitud de lecturas	700pb	2x250pb	50+50pb	400-900pb
Precisión	99.9%	99.9%	99.94%	99.999%
Lecturas	1 M	Hasta 1 T	1.2-1.4 M	-
Tiempo de corrida	24 h	3 – 10 días	7 - 14 días	20 min – 3 h
Ventaja	Longitud de lecturas	Alto rendimiento	Precisión	Alta calidad, lectura larga
Desventaja	Alto costo, bajo rendimiento	Ensamblaje de lecturas cortas	Ensamblaje de lecturas cortas	Alto costo, bajo rendimiento
Precio del instrumento / corrida (USD)	$500,000 / $7,000	$690,000 / $6,000	$495,000 / $15,000	$95,000 / $4 (800 pb)
Memoria	48 GB	48 GB	16 GB	1 GB
Disco duro	1.1 TB	3 TB	10 TB	280 GB
Preparación de librería automatizada	✓	✓	✓	✓
Otros dispositivos necesarios	REM e System	cBOT System	EZ Beads System	Ninguno
Costo por millón de bases (USD)	$10	□0.07	$□.13	$2,400
Resecuenciar	✓	✓	✓	✗
De novo	✓	✓	✗	✓
Cáncer	✓	✓	✓	✗
Arreglos	✓	✓	✓	✓
Bacterias	✓	✓	✓	✗
Detección de mutaciones	✓	✓	✓	✓

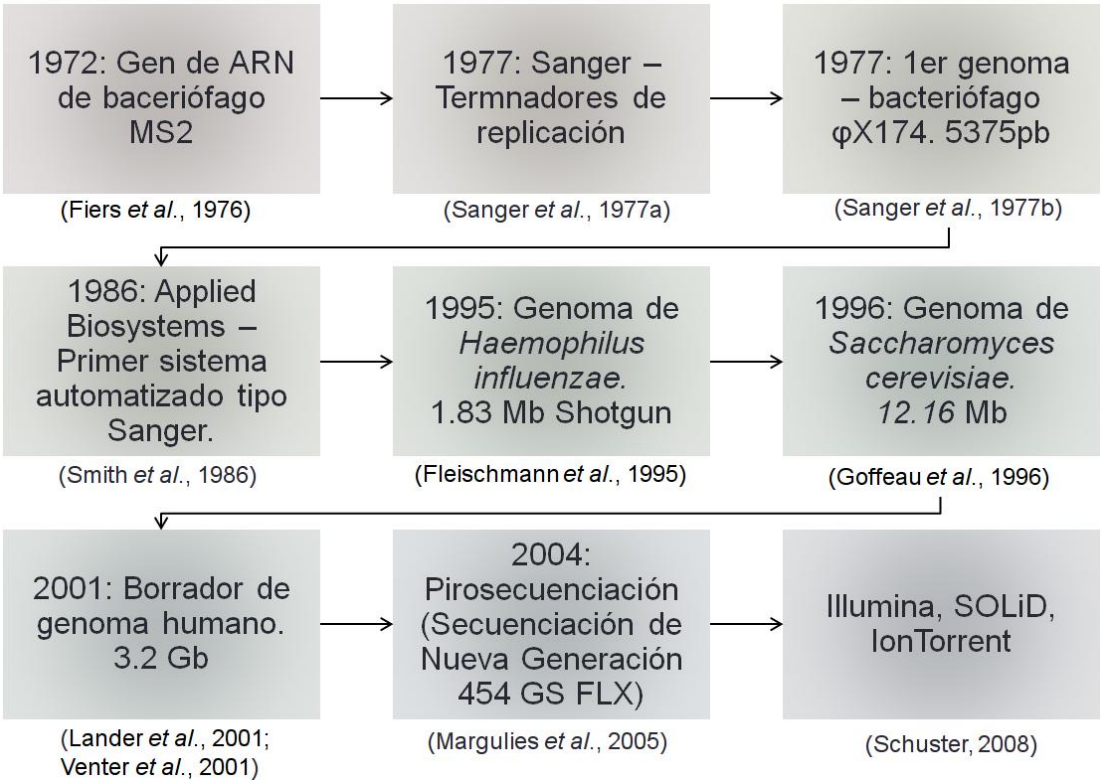


Figura 1: Línea de tiempo - Secuenciación y tecnologías emergentes

2.2. Genoma de levaduras

Saccharomyces cerevisiae, el primer eucariote cuyo genoma fue secuenciado por completo, tiene un genoma nuclear de 12,156,677 pb organizados en un total de 16 cromosomas, los cuales presentan una cantidad considerable de redundancia genética (Goffeau *et al.*, 1996). El mapa físico y genético de los cromosomas, resultado de más de 55 años de análisis genético es accesible gracias a la creación de bases de datos tales como la *Saccharomyces Genome Database* (yeastgenome.org), la cual es constantemente actualizada con los datos de investigaciones recientes (Cherry *et al.*, 1997).

Estudios recientes revelaron que la levadura utilizada en la industria cervecera, específicamente en la producción de cerveza tipo lager, se trata de *Saccharomyces pastorianus:* un híbrido cuyo genoma de 25 Mb de longitud comprende dos sub-genomas nucleares correspondientes a *S. cerevisiae* y *S. bayanus* (formalmente *S. eubayanus*) con un genoma mitocondrial circular de *S. bayanus*. En total, se han encontrado hasta 36 cromosomas distintos incluyendo ocho cromosomas con translocaciones entre los dos genomas (Figura 2). Así mismo, debido a los rearreglos cromosómicos, muchos loci responsables de características típicas de las levaduras cerveceras como la asimilación de maltotriosa y producción de sulfitos se encuentran incrementados en número (Nakao *et al.*, 2009).

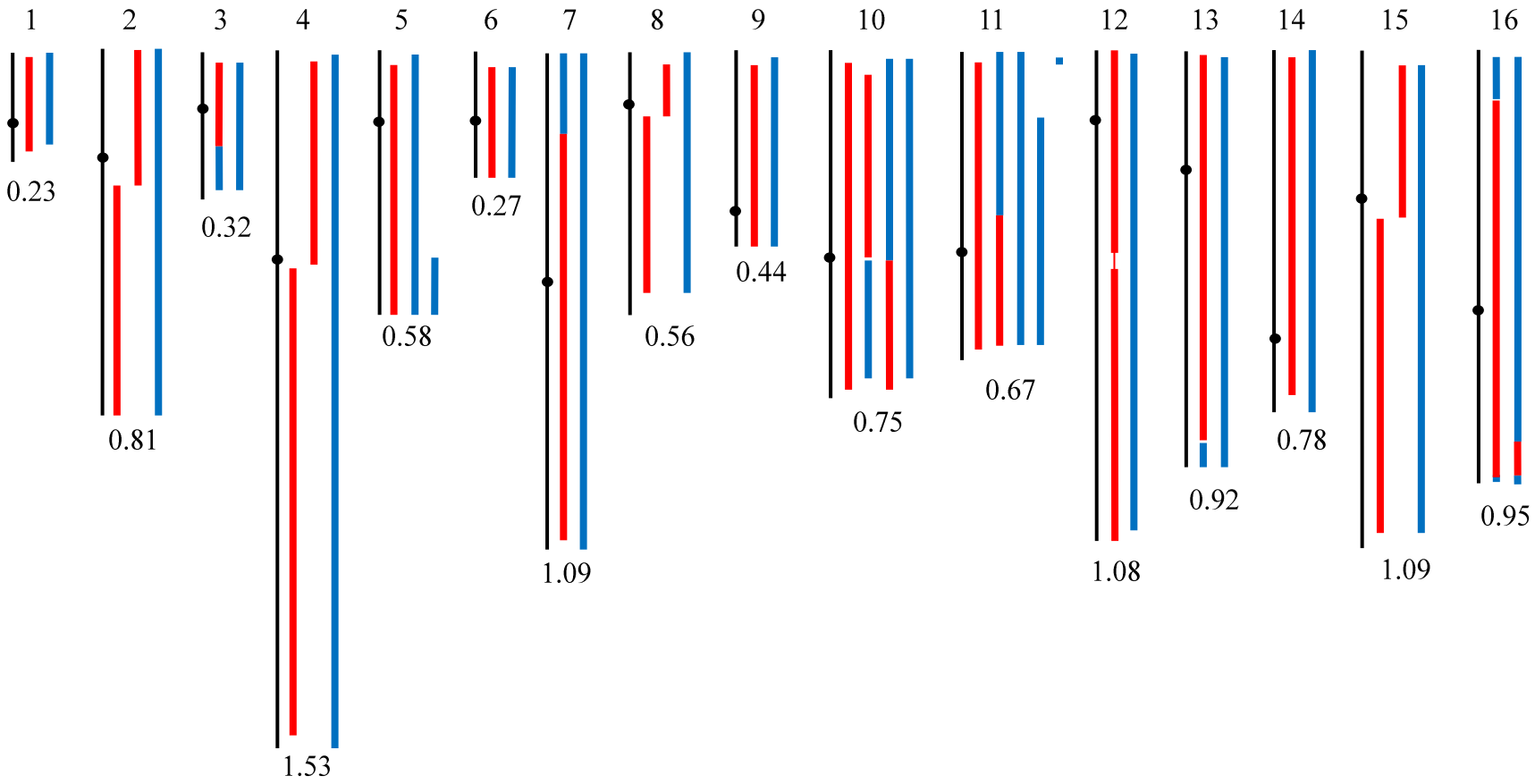


Figura 2: Genoma de *S. pastorianus*; Líneas Negro - cromosomas del genoma de referencia *S. cerevisiae* S288C; Rojo – Subgenoma *S. eubayanus*; Azul Subgenoma *S. cerevisiae*. Se pueden observar los rearreglos, translocaciones, y diferente número de copias de los cromosomas. En la parte inferior se muestra el tamaño de cada cromosoma de referencia en Mpb (Adaptada de Nakao *et al*., 2009).

2.3. Herramientas bioinformáticas

Con el advenimiento de las nuevas tecnologías para determinar la secuencia nucleotídica de los ácido nucleicos (secuenciación), los costos disminuyeron y los datos generados se incrementaron considerablemente; al mismo tiempo, se incrementó la necesidad de nuevos algoritmos y programas bioinformáticos para su análisis, el cual ha sido un "cuello de botella" en las investigaciones de ésta índole.

Las dos clases de algoritmos mayormente utilizados para el ensamblaje de las secuencias son: OLC, superposición-disposición-consenso (*Overlap-Layout-Consensus*) y DBG, grafo de-brujn (D*e-Bruijn-Graph*).

OLC generalmente funciona en tres pasos: i) encontrar los sobrelapes de todas las lecturas; ii) llevar a cabo una disposición de todas las lecturas y sobrelapes en un grafo y iii) inferir el consenso de la secuencia. Es un algoritmo de ensamblaje tipo intuitivo desarrollado en 1980 y mejorado, posteriormente, por muchos grupos de investigación. Es utilizado ampliamente en programas como Arachne, Celera Assembler, CAP3, PCAP, Phrap, Phusion, MIRA3 y Newbler (Li *et al.*, 2011).

DBG es un algoritmo anti-intuitivo: primero corta las lecturas en *k*-mers (k-mer: todas las posibles subsecuencias de longitud *k* de una lectura) mucho más pequeños y utiliza todos los *k*-mers para formar un DGB, finalmente infiriendo la secuencia genómica (Idury y Waterman, 1995). Este algoritmo fue originalmente introducido en 1995, pero no fue ampliamente utilizado sino hasta la introducción de la plataforma Illumina; el primer ensamblador DBG, EULER, fue publicado en 2001 y desde entonces se han desarrollado muchos otros ensambladores DBG para lecturas cortas, tales como Euler-USR, Velvet, AbySS, AllPath-LG, SOAPdenovo, IDBA y más (Peng *et al., 2012;* Li *et al.*, 2011).

Independientemente del programa y algoritmo utilizado, el ensamblaje de secuencias de plataformas de segunda generación se realiza en pasos: las lecturas se sobrelapan para formar *contigs* los cuales se ordenan y orientan para construir *scaffolds*. Dichos *scaffolds*, que son rellenados con N's para dar continuidad a los *contigs*, deben ser validados para corroborar el ordenamiento y rellenar los *gaps* faltantes.

3. HIPÓTESIS

El ensamblaje del genoma nuclear de la levadura permite asignar una posición de las lecturas de secuenciación en los cromosomas, así como conocer el número de copias de regiones de interés.

4. OBJETIVOS

5.1. OBJETIVO GENERAL

Realizar el ensamblaje del genoma de una levadura cervecera tipo lager por medio del sobrelape de lecturas de secuenciación.

5.2. OBJETIVOS PARTICULARES

1. Analizar las lecturas de secuenciación para buscar y filtrar posible contaminación.
2. Dilucidar el número de cromosomas de la levadura por medio de PFGE
3. Ensamblar las lecturas en *contigs*.
4. Mapear las secuencias ensambladas para formar *scaffolds*.
5. Ordenar los *scaffolds* para construir los cromosomas.
6. Validar el ensamblaje por medio de PCRs de los segmentos dudosos

5. JUSTIFICACIÓN

El conocimiento básico y aplicado de las levaduras de interés industrial se encuentra circunscrito a pruebas fisiológicas y amplificaciones de genes en particular. El conocimiento de la secuencia nucleotídica del genoma nuclear de levaduras cerveceras, desde el punto de vista básico como el tamaño, organización por cromosomas, hasta el número de copias que posee de regiones de interés, nos coloca en la posición de desarrollar, de manera más racional, no solo tecnologías de fermentación sino también herramientas para la detección de viabilidad, rutas metabólicas, etc.

6. MATERIALES Y MÉTODOS

El presente trabajo fue realizado con la secuencia genómica de una levadura cervecera tipo lager, la cual pertenece a la colección de la Cervecería Cuauhtémoc Moctezuma (CCM). La Tabla 2 muestra las distintas plataformas y cantidad de lecturas obtenidas por cada una:

Tabla 2: Datos crudos de secuenciación

Plataforma (tipo de secuenciación – tamaño de inserto)	Numero de lecturas (en millones)	Tamaño promedio de lecturas (pb)
FLX 454 Titanium	0.8	400
Illumina (*Pair end*)	6.0	101-192
Illumina (*Mate pair* – 350pb)	5.0	101 x 2
Illumina (*Mate pair* – 8 kb)	11.7	51 x 2

Así mismo, se obtuvieron los datos de la secuenciación de los ocho cromosomas de menor tamaño por medio de la plataforma Ion Torrent. Dichos cromosomas fueron secuenciados por la Empresa de manera individual por medio de Gel de Electroforesis de Campo Pulsante (PFGE). El número de lecturas obtenidas de cada uno de los cromosomas (denominados "Bandas" y ordenados del 1 al 8 de forma ascendente en tamaño) son mostrados en la Tabla 3.

Tabla 3: Secuenciación de cromosomas individuales

Banda	Número de lecturas	Banda	Número de lecturas
B1	76,299	**B5**	758,392
B2	78,987	**B6**	1,309,324
B3	93,633	**B7**	199,020
B4	91,364	**B8**	147,158

El análisis bioinformático fue realizado en una computadora y dos servidores con las características mostradas en la Tabla 4:

Tabla 4: Equipo computacional para el análisis de datos.

Computadora / Servidor	Memoria RAM	CPU	Sistema operativo
HP xw4600 Workstation (Computadora)	4 GB	Intel® Core™2 Quad CPU Q9300 2.50GHz × 4	Linux / Ubuntu 12.10 32-bit
HP xw4600 Workstation (Servidor)	4 GB	Intel® Core™2 Quad CPU Q9300 2.50GHz x 4	Linux / Ubuntu Server 64-bit
GS FLX Titanium Cluster	32 GB	2.33 GHz x 20	Linux / Fedora 64-bit

El ensamblaje del genoma consiste en sobrelapar las lecturas arrojadas por el secuenciador para formar segmentos mayores del genoma denominados *contigs*. Éstos son numerados de mayor a menor tamaño pero no llevan un orden real con respecto a su posición en los cromosomas. Para ser considerado un *contig*, se requiere de al menos dos lecturas sobrelapadas y una longitud mínima de 500 pb. No es necesario un análisis estadístico debido a la naturaleza de la técnica y a la profundidad del genoma (cuantas veces está representado, en promedio, cada nucleótido).

Inicialmente, por medio del programa computacional (software) FastQC 0.10.1 (Andrews, 2010) se verificó la calidad de las lecturas de Illumina, asegurando que éstas tuviesen un valor de calidad ≥ Q30 (un posible error de interpretación cada 1000 bases). Habiendo pasado dicho filtro, las lecturas *Pair end* de Illumina y las lecturas obtenidas por pirosecuenciación (FLX 454 Titanium) fueron ensambladas por medio del software Newbler 2.6 (Roche). Los *contigs* resultantes fueron entonces procesados por medio del programa SSPACE 1.0 (Boetzer *et al.*, 2011) el cual enfoca el ensamblaje de los *gaps* y forma así *scaffolds* (al menos dos *contigs* con una posición y dirección específica en los cromosomas (Figura 3). Se utilizaron los parámetros estándar para los programas mencionados.

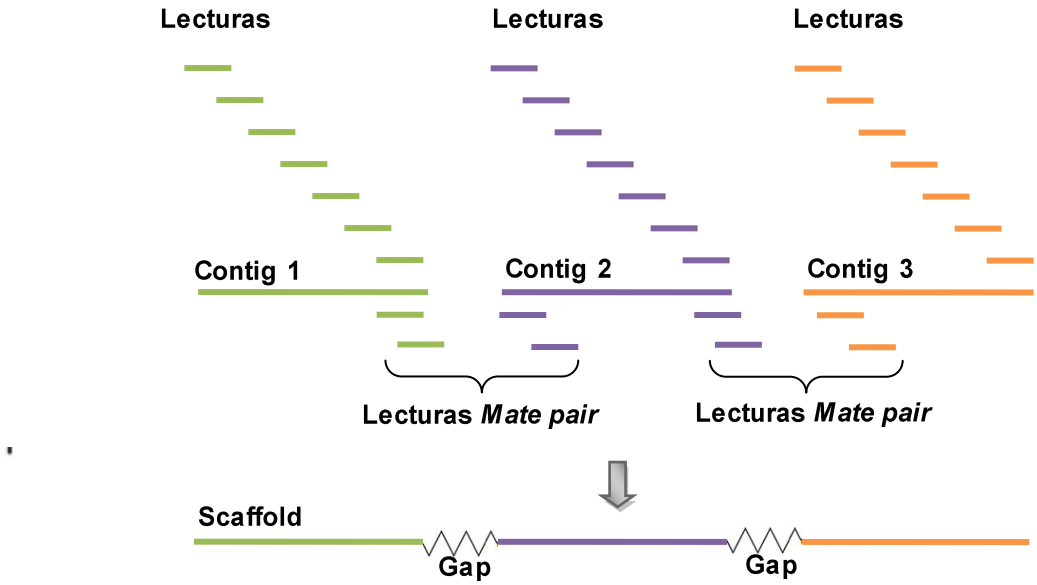


Figura 3: Formación de *Scaffolds*

Posteriormente, y teniendo ya los datos de secuenciación de los cromosomas por separado, éstos fueron ensamblados con el software Newbler 2.6 y alineados contra los *scaffolds* del genoma completo. Esto delimitó cuáles son los *scaffolds* correspondientes a cada uno de los ocho cromosomas de menor tamaño.

Todos los *scaffolds* fueron entonces alineados contra los cromosomas de referencia reportadas en bases de datos como *S. cerevisiae* (obtenida de http://www.ncbi.nlm.nih.gov/genome/15 - revisado el día 5 de noviembre de 2014) y *S. eubayanus* (secuencia no reportada), con ello se obtuvo el orden y asignación de los *scaffolds* por cromosoma. Dichos alineamientos se realizaron con el programa MUMmer 3.23 (Delcher, *et al.*, 2003)

El número total de cromosomas, así como su tamaño estimado fue evaluado por medio de los PFGE mencionados anteriormente; en ellos se puede llevar a cabo la

separación e identificación de los cromosomas individualmente. El protocolo para la preparación de las muestras de ADN, así como los geles fue proporcionado por la empresa y realizado en el Centro de Investigación y Desarrollo en Ciencias de la Salud (CIDICS) de la Universidad Autónoma de Nuevo León.

El análisis de la ploidía fue realizado con el *software* Magnolya 0.14 (Nijkamp *et al*., 2012). Mientras que los *scaffolds* fueron visualizados con el programa Ugene 1.11.3 (Okonechnikov *et al*., 2012). Este programa permite encontrar *gaps* (secuencias cuyo nucleótido se desconoce y representados por la letra "N"). El diseño de oligonucleótidos (primers) para PCR flanqueando 5 de estos *gaps* fue realizado con el programa Primer3 0.4.0 (Rozen y Skaletsky. 1999) y su especificidad verificada con BLAST (Johnson *et al*., 2008) (Tabla 5). La PCR y secuenciación de los amplicones fue realizada por la empresa Vitagénesis S.A. de C.V. quien estandarizó se realizó de acuerdo a las características de los mismos.

Tabla 5: Oligonucleótidos para llenado de *gaps* en la cepa 790.

Nombre	Ubicación	Secuencia nucleotídica (5'-3')	Tamaño (bases)	Tamaño esperado (pb)	Tm (°C)
Gap1 F	Scaf30_8320-9787	TTTACCATGAGCGCAACAGC	20	1750	56.3
Gap1 R	Scaf30_8320-9787	AAAAAGCAGAACGACGCACC	20		56.6
Gap2 F	Scaf30_140561-141053	TCTTCGAATCTGGCTCTGGT	20	0 (posible deleción)	55.7
Gap2 R	Scaf30_140561-141053	GCTGCTATGAATCCTTCGAATACC	24		55.6
Gap3 F	Scaf8_17006-18762	AAGTAGAGTCAATTTGGCACCAG	23	927	55.2
Gap3 R	Scaf8_17006-18762	TCAAGCAACAGCTCATCCAC	20		55.5
Gap4 F	Scaf13_519000-520800	GCTCAGAAGCAGATGTCTTCAA	22	327	55.5
Gap4 R	Scaf13_519000-520800	TGCGGTGTAAGAAGAGGACA	20		55.9
Gap5 F	Scaf13_523000-524500	TCCCAGCAGAAAAGATGGAC	20	681	54.8
Gap5 R	Scaf13_523000-524500	GGAGCCGCAGTGAAAGTTAAT	21		55.6

El análisis de las secuencias amplificadas se llevó a cabo mediante el software Geneious 7.1.6 (Kearse *et al.*, 2012) para la visualización de los electroferogramas, así como Ugene 1.11.3 para la reconstrucción y llenado de los *gaps*.

La Figura 4 muestra el diagrama de flujo en el análisis de las secuencias para su ensamblaje completo.

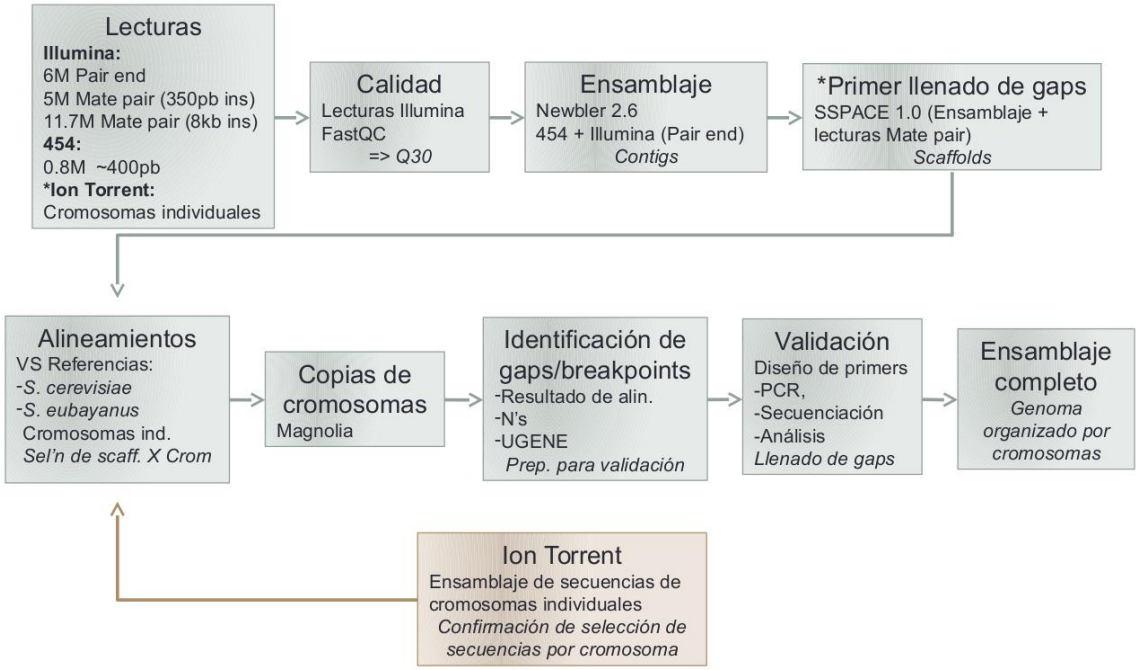


Figura 4: Diagrama de flujo del ensamblaje del genoma.

7. RESULTADOS

7.1. Calidad de las lecturas

7.1.1. Librería FLX 454 Titanium: Las lecturas del FLX 454 Titanium con valor de calidad <Q40 son filtradas automáticamente, por lo cual se tiene asegurado el uso de lecturas de muy alta calidad (un error cada 10,000 bases).

7.1.2. Librería Illumina *Pair end*: Las lecturas de Illumina analizadas con FastQC mostraron tener una calidad >Q30 (un error cada 1,000 bases). La Figura 5 muestra los valores de calidad Q (eje Y) con respecto a la posición de las bases (eje X) de la librería *Pair end* (ver Tabla 2). Así mismo, se muestra la distribución del tamaño de las lecturas obtenidas en dicha librería (Figura 6)

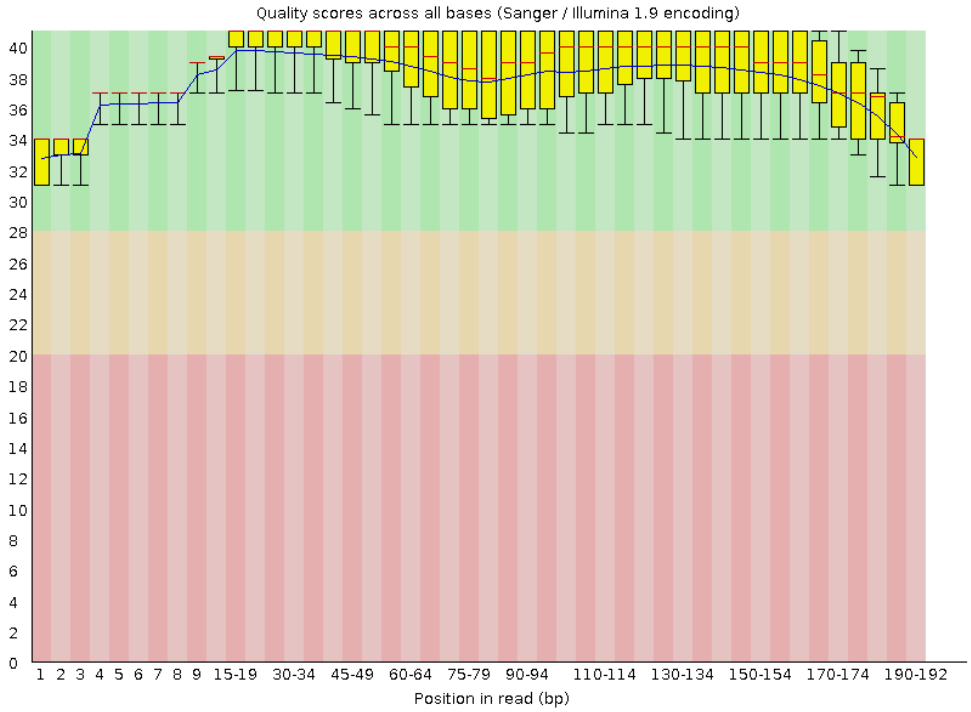


Figura 5: Valor de calidad de librería *Pair end* (Illumina).

Figura 6: Distribución del tamaño de lecturas *Pair end* (Illumina).

7.1.3. Librería Illumina *Mate pair* (inserto 8 Kpb): Al igual que las lecturas *Pair end*, las lecturas *Mate pair* con un inserto de 8 Kb mostraron tener una calidad por arriba de Q30 y un tamaño específico de 51pb en cada extremo (51 x 2). Dichos resultados se muestran en las Figura 7 y 8

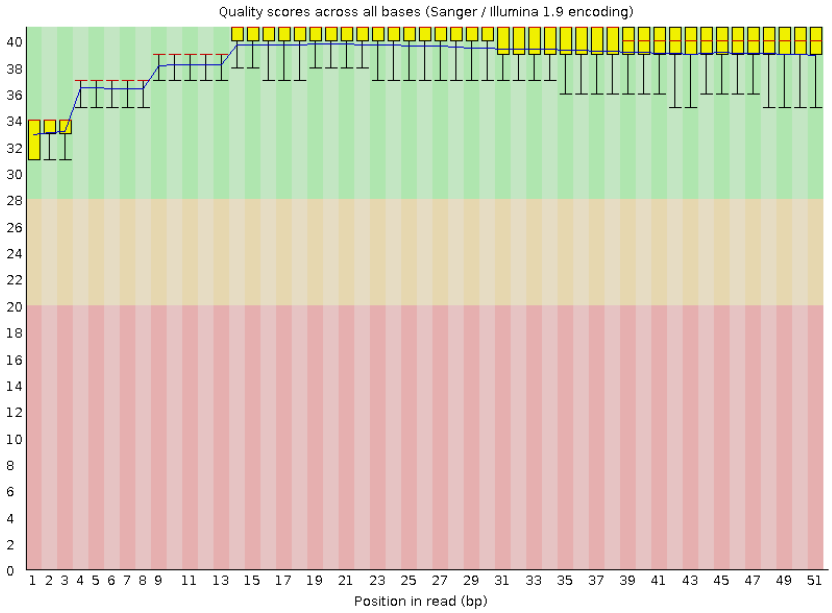


Figura 7: Valor de calidad de libreria *Mate pair*, inserto: 8 Kb (Illumina).

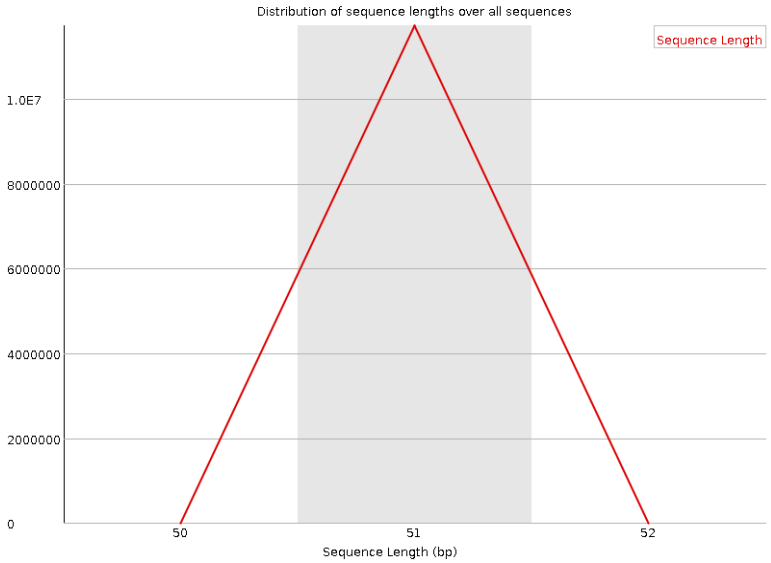


Figura 8: Distribución del tamaño las lecturas *Mate pair*, inserto: 8 Kpb (Illumina).

7.1.4. Librería Illumina *Mate pair* (inserto 350 pb): Así mismo, se realizó la comprobación de calidad de las lecturas *Mate pair* con inserto de 350 pb. Se observó también una calidad ≥ Q30 con un tamaño de lectura de 101pb en cada extremo (101 x 2). Dichos resultados se muestran en las Figura 9 y 10.

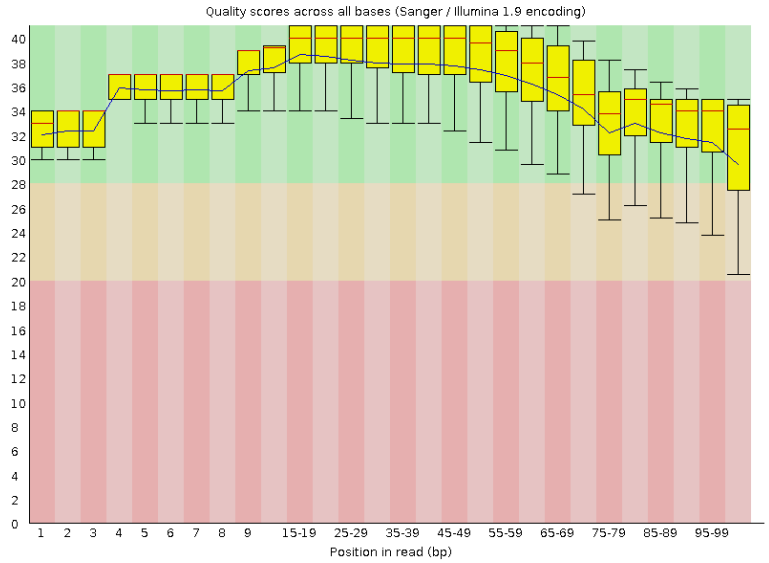


Figura 9: Valor de calidad de libreria *Mate pair*, inserto: 350 pb (Illumina).

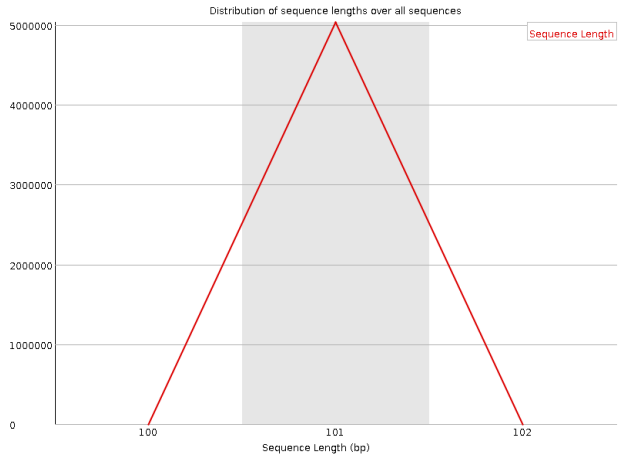


Figura 10: Distribución del tamaño las lecturas *Mate pair*, inserto: 350 pb (Illumina).

7.2. Ensamblaje *de novo* y obtención de *Scaffolds:* Durante el transcurso de la realización de esta tesis, se tuvieron tres versiones distintas del ensamblaje (Tabla 6):

V1- Lecturas FLX 454 Titanium

V2- Lecturas FLX 454 Titanium + Illumina *Mate pair* inserto de 350 pb.

V3- Lecturas FLX 454 Titanium + Illumina *Mate pair* inserto de 350 pb. + Illumina *Mate pair* inserto de 8 kb + Illumina *Pair end* (versión final).

Tabla 6: Versiones del ensamblaje de la levadura cervecera.

	V1	V2	V3
Lecturas ensambladas	769113	10715485	17034361
Profundidad	~8x	~41x	~70x
Tamaño estimado del Genoma (Mb)	33.8	31	22.7
Número de *contigs/scaffolds*	2,850	1,103	133
Tamaño promedio de *contigs/scaffolds* (pb)	7,699	20,130	170,987
Contig/scaffold de mayor tamaño (pb)	154,254	233,917	1,404,408
N50	13,964	42,237	568,800
Ns	0	0	399,699

Como se puede observar en la tabla anterior, la última versión del ensamblaje asignó el orden y orientación a muchos de los *contigs*, disminuyendo de 1103 *contigs* a solo 133 *scaffolds* cuya distribución de tamaños se muestran en la

Tabla *7*.

Tabla 7: Distribución del tamaño de *Scaffolds*

Nivel de tamaño	Número de *scaffolds*	Pb
0 : 999	47	31,168
1k : 9,999	21	44,610
10k : 99,999	18	1,063,816
100k : 199,999	11	1,608,982
200k : 299,999	6	1,528,906
300k : 399,999	6	2,092,805
400k : 499:999	4	1,700,637
500k : 599,999	7	3,759,004
600k : 699,999	5	3,213,208
700k : 799,999	1	712,480
800k : 899,999	1	806,158
900k : 999,999	5	4,775,094
1;404,408	1	1,404,408
Total	**133**	**22,741,276**

7.3. Alineamientos

7.3.1. 790 vs *S. cerevisiae* S288C: Los *scaffolds* obtenidos (133) fueron alineados contra cada uno de los 16 cromosomas de la cepa de referencia *S. cerevisiae* S288C (Fig. 11). Donde observamos dos niveles de identidad: ~100% (rojo) y ~80% (verde). Los últimos corresponden al subgenoma de *S. eubayanus.*

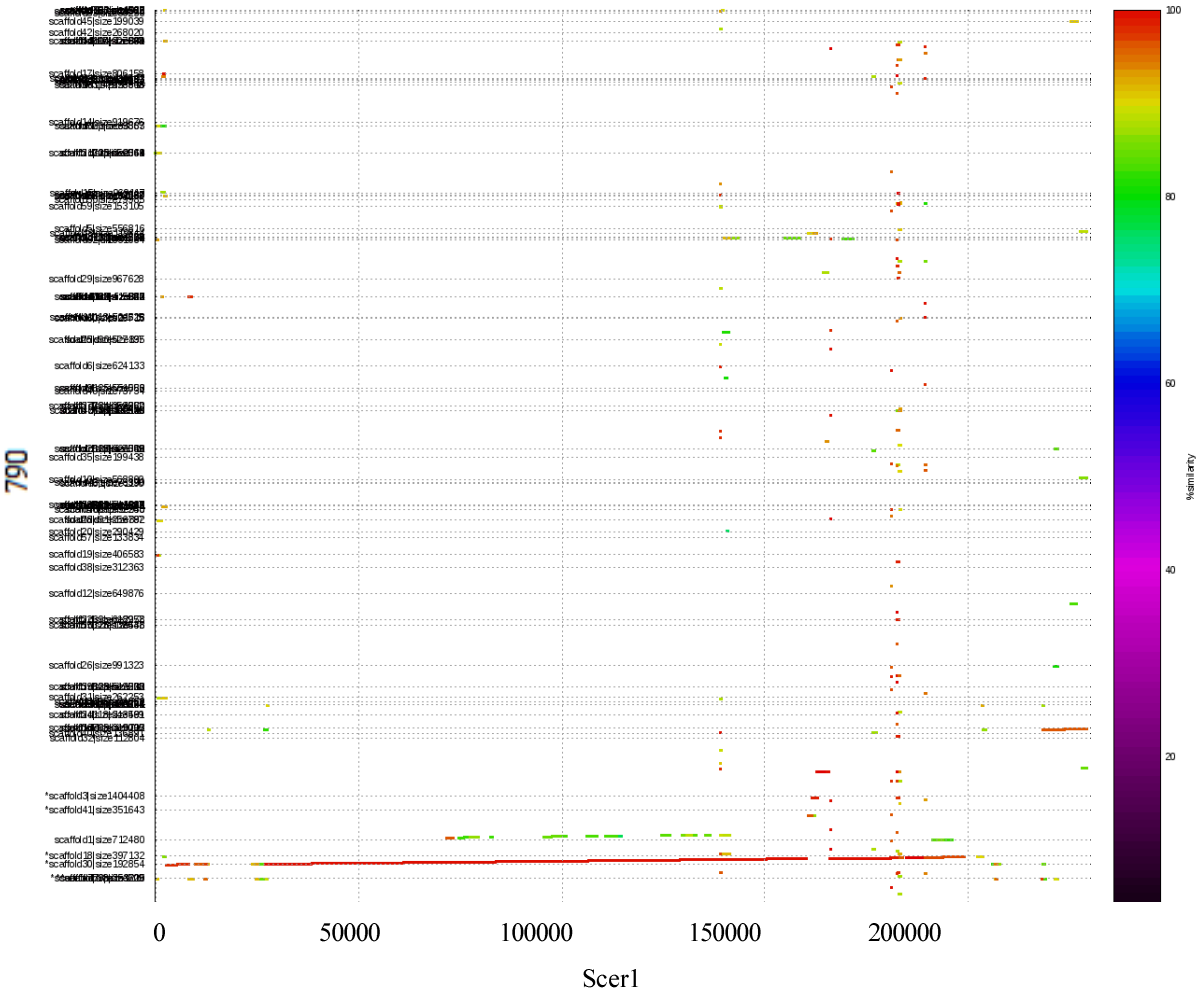


Figura 11: Alineamiento de los *scaffolds* de 790 (eje Y) vs *S. cerevisiae* S288C (eje X).

7.3.2. 790 vs *S. eubayanus*: El mismo método se utilizó para el alineamiento de *S. eubayanus* (secuencia no publicada en bases de datos) contra 790 (Fig. 12). Donde los datos de identidad ~100% (rojo) corresponden a *S. eubayanus*, mientras que ~80% (verde) al subgenoma de *S. cerevisiae* (ver Capitulo 12: Anexos).

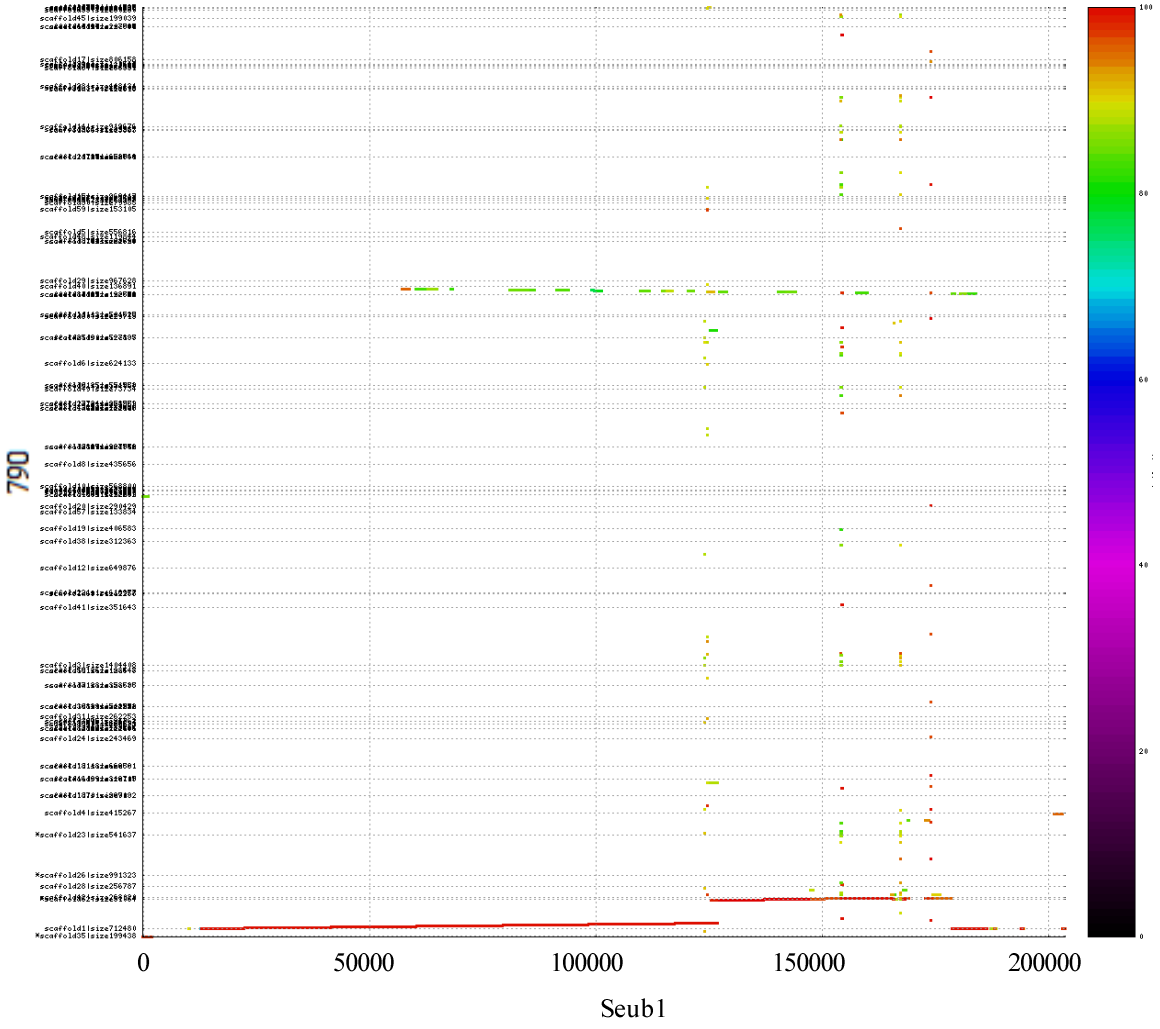


Figura 12: Alineamiento de los *scaffolds* de 790 (eje Y) vs *S. eubayanus* (eje X).

7.4. Asignación de *scaffolds* a cromosomas

A partir de las tablas de alineamientos obtenidas con MUMmer 3.23 (Tabla 8), se asignaron los *scaffolds* correspondientes para cada uno de los cromosomas del subgenoma de *S. cerevisiae* (Scer) y *S. eubayanus* (Seub), asi como el porcentaje de cobertura de cada cromosoma de referencia (Tabla 9).

Tabla 8: Ejemplo de las tablas de alineamiento

[S1]	[E1]	[S2]	[E2]	[CovR]	[CovQ]	[Ref]	[Qry]
2709	8320	187227	192854	2.44	2.92	S288C-Ch1	Scf30
9787	12346	182869	185428	1.11	1.33	S288C-Ch1	Scf30
27136	140561	67495	180918	49.27	58.81	S288C-Ch1	Scf30
141053	160238	48297	67493	8.33	9.95	S288C-Ch1	Scf30
166163	181044	33407	48302	6.46	7.72	S288C-Ch1	Scf30
184871	188778	16267	20174	1.70	2.03	S288C-Ch1	Scf30
188780	199064	4666	15051	4.47	5.39	S288C-Ch1	Scf30

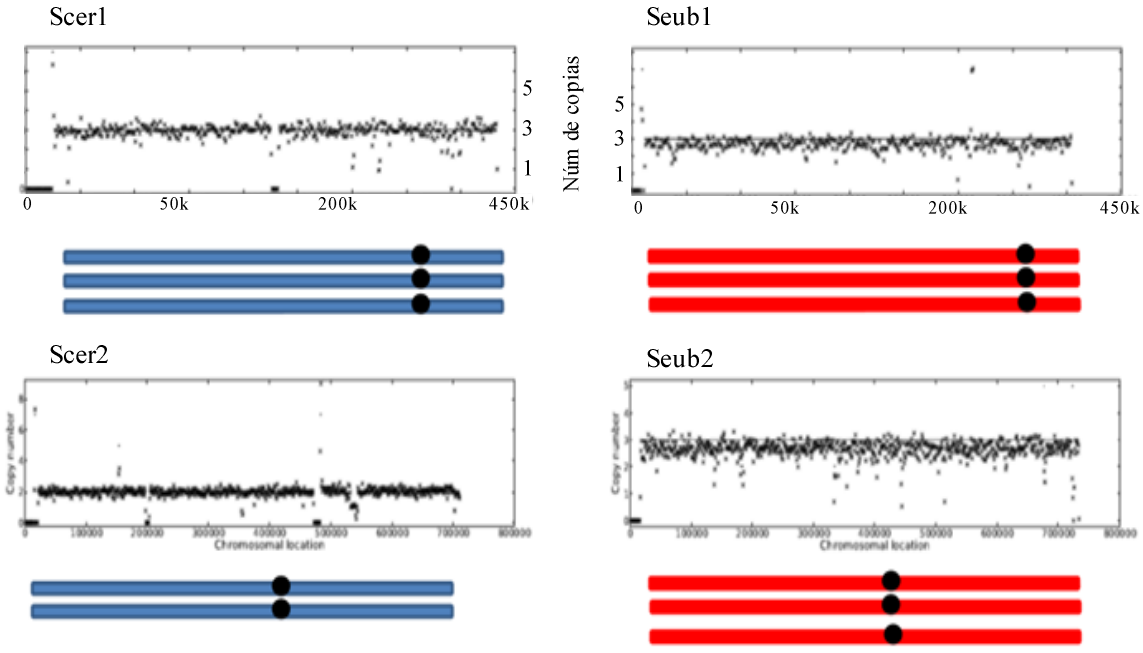
Tabla 9: Asignacion de *scaffolds* y cobertura de los cromosomas

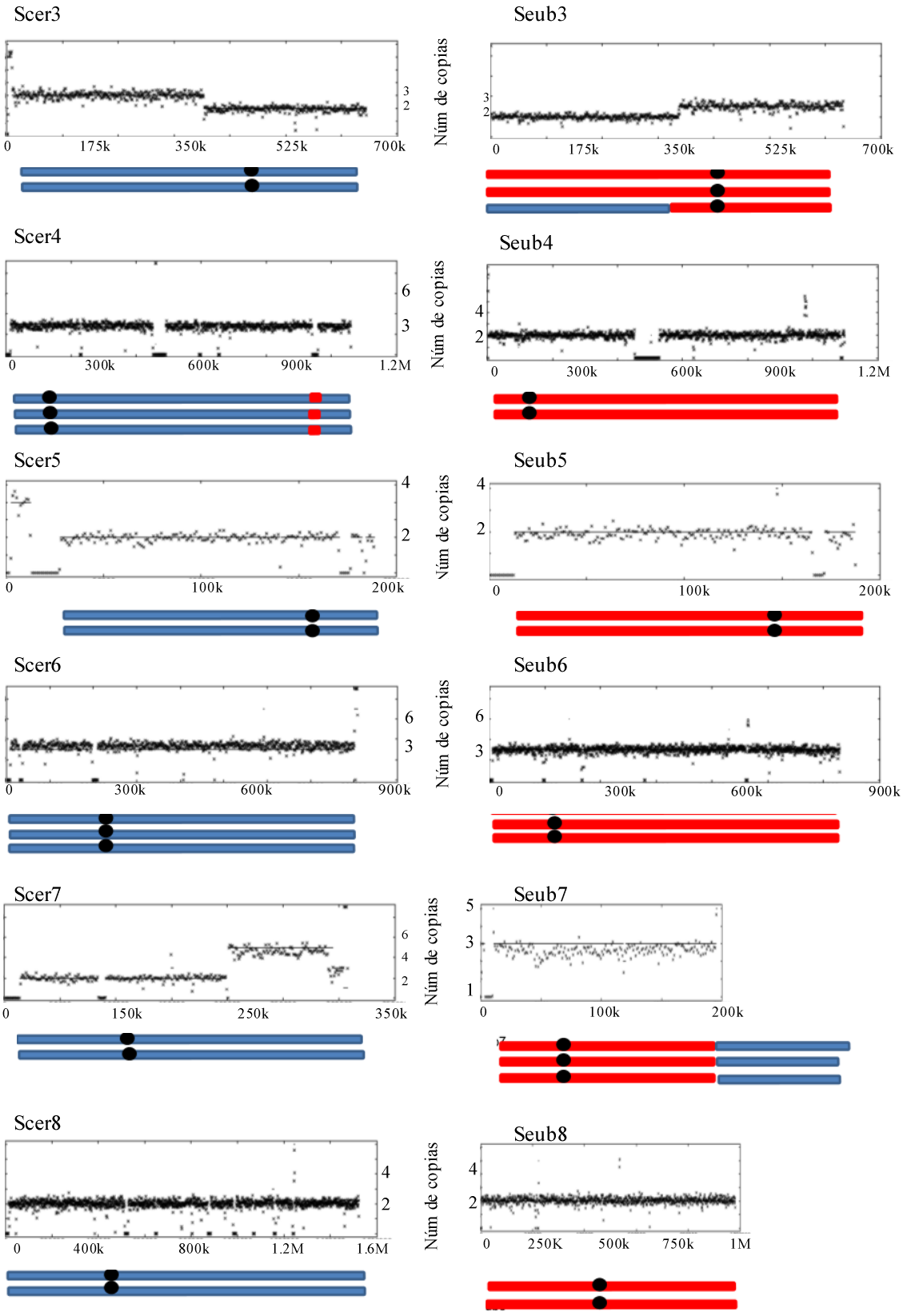
790cer scaff	Cobertura	Scer	Seub	Cobertura	790eub scaff
30	73.78	1	1	75.57	1, 62
61, 22, 11	87.18	2	2	94.40	14, 16, 26
46, 63, 32, 37, 65	85.92	3	3	60.47	45
29, 17	87.53	4	4	94.83	14, 16, 26
34, 17, 22, 54, 32	90.86	5	5	93.77	25
40, 47	86.05	6	6	91.58	42
31, 39, 41	86.38	7	7	67.55	31, 50, 28, 23
59, 7	82.27	8	8	97.10	4, 6, 55, 3
8	86.19	9	9	94.18	19
33, 24	87.48	10	10	93.96	27, 49, 20
13	93.97	11	11	97.77	12
56, 18, 60, 10, 48	85.88	12	12	90.91	1, 36, 10, 6, 1
43, 3	86.75	13	13	91.67	15
9, 44	89.08	14	14	97.84	5, 35
2, 53, 58	87.75	15	15	97.10	4, 6, 55, 3
52, 11, 57, 64, 38	75.88	16	16	72.06	21

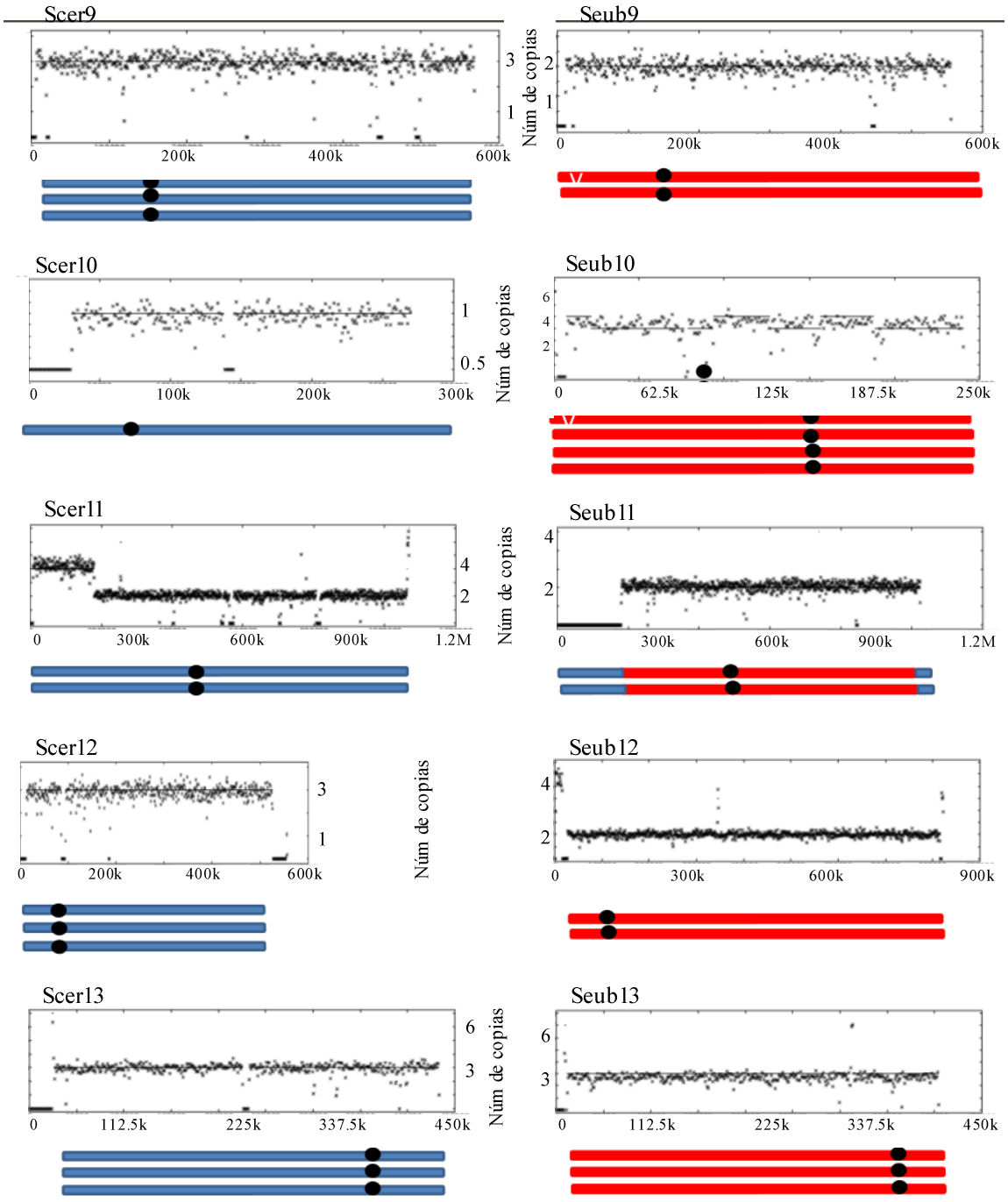
Los *scaffolds* correspondientes a cada cromosoma fueron seleccionados en base al porcentaje de identidad de cada *scaffold* con los cromosomas. P. ej.: el *scaffold* 46 tiene un 98.15% de identidad con un fragmento del cromosoma 3 del subgenoma de *S. cerevisiae*, y no tiene identidad con ningún otro cromosoma, por ende, se asigna el *scaffold* 46 a dicho cromosoma. Es importante señalar que algunos de los *scaffolds* se repiten tanto para cromosomas del mismo subgenoma, como entre los subgenomas, de lo cual se hace referencia en la discusión.

7.5. Ploidía y rearreglos cromosómicos

El número de copias de los cromosomas, así como sus rearreglos cromosómicos fue evaluado por medio del software Magnolya 0.14, el cual mapea las lecturas crudas contra los cromosomas de referencia y en base a la profundidad encontrada en cada fragmento de los cromosomas deduce al número de copias del mismo. La Figura 13 muestra el número de copas de cada uno de los cromosomas de 790 con respecto a su subgenoma *S. cerevisiae* (Scer – azul) y *S. eubayanus* (Seub – rojo).




Scer3
Seub3
Núm de copias
0
175k
350k
525k
700k
Scer4
Seub4
300k
600k
900k
1.2M
Scer5
Seub5
100k
200k
Scer6
Seub6
Scer7
Seub7
150k
250k
350k
Scer8
Seub8
400k
800k
1.6M
250K
500k
750k
1M


Scer9
Seub9
Núm de copias
0
200k
400k
600k
Scer10
Seub10
100k
200k
300k
62.5k
125k
187.5k
250k
Scer11
Seub11
300k
600k
900k
1.2M
Scer12
Seub12
Scer13
Seub13
112.5k
225k
337.5k
450k

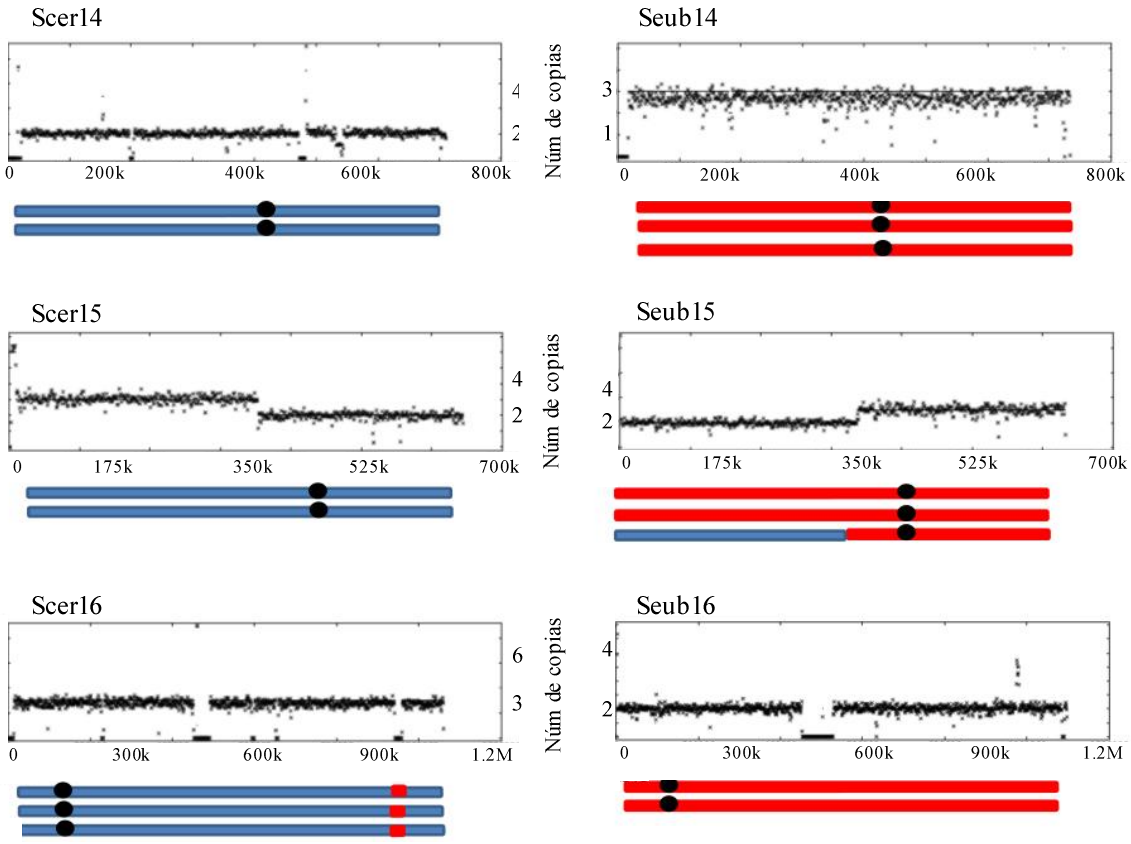


Figura 13: Número de copias de los cromosomas y rearreglos. Eje Y: Número de copias de cada cromosoma; Eje X: Posición en cada cromosoma.

Estos resultados muestran algunos rearreglos importantes en el genoma de 790, tales como translocaciones en los cromosomas 3, 4, 7, 11, 15 y 16, en los que algunas regiones del subgenoma de *S. cerevisiae* se encuentran insertadas en el subgenoma de *S. eubayanus*, y viceversa.

7.6. Identificación de *gaps*

El ensamblaje final (Versión 3) incluyó un total de 399699 'Ns' (ver Tabla 6) o secuencias nucleotídicas desconocidas, denominados "*gaps*", los cuales visualizados con el software Ugene 1.11.3 como se muestra en el ejemplo de la Figura 14 marcados en amarillo a lo largo de la secuencia del *scaffold* 30.

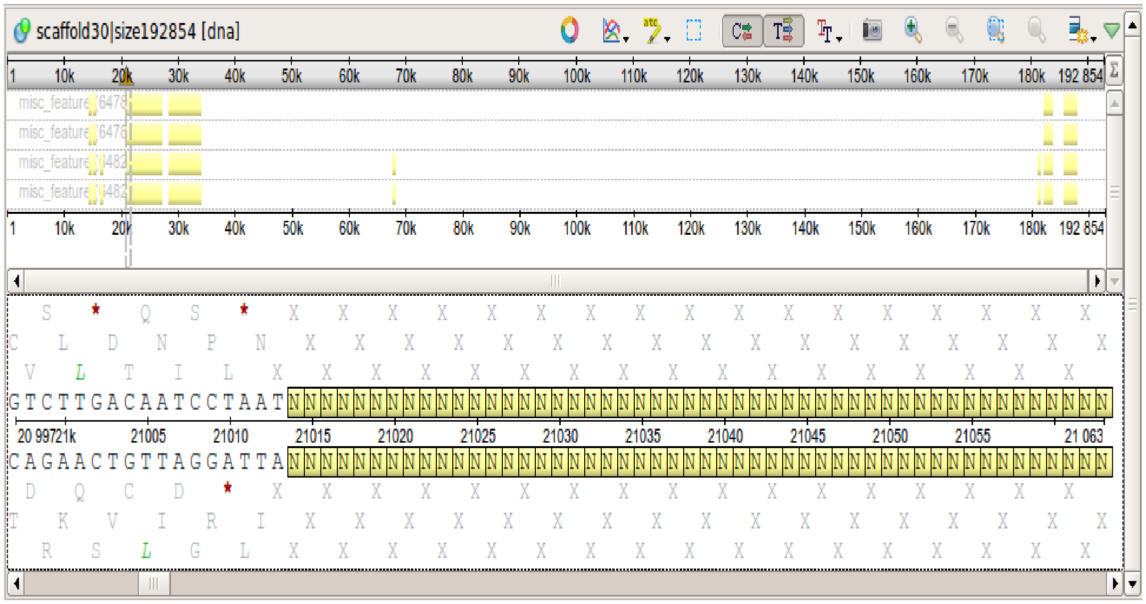


Figura 14: *Gaps* en *scaffold* 30.

7.7. Amplificación, secuenciación y llenado de *gaps*

7.7.1. *Gap* 1

Por medio de las tablas de alineamiento generadas con MUMmer 3.23 (ver Tabla 8) se identificó una región no resuelta en la posición 8,320 a 9,787 del *scaffold* 30 con respecto a la referencia *S. cerevisiae* S288C. El gel del producto amplificado se muestra en la Figura 16; en ella se observa el fragmento de ~1,750 pb en las tres cepas en estudio (*S. cerevisiae*, 790 y *S. eubayanus*).

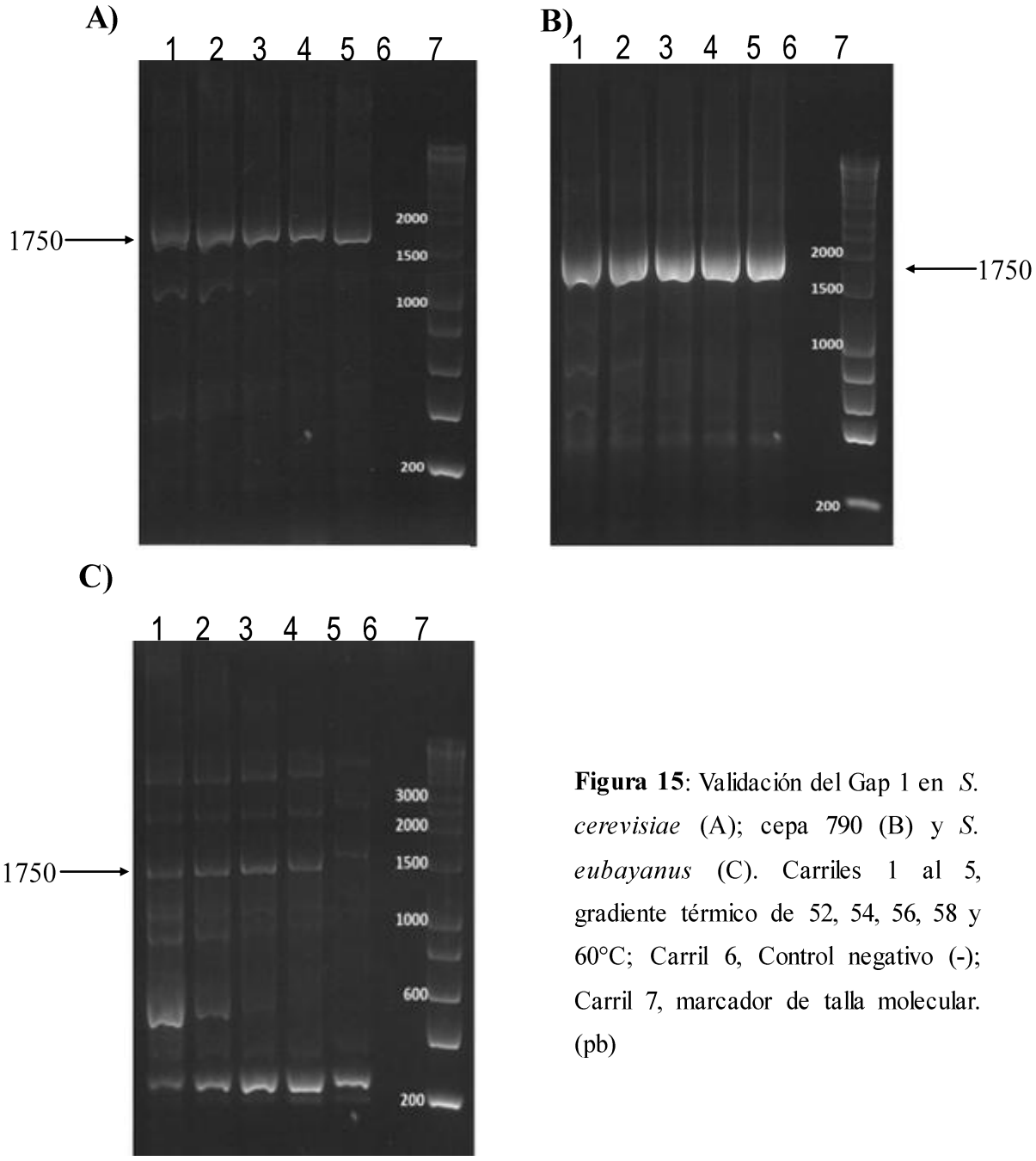


Figura 15: Validación del Gap 1 en *S. cerevisiae* (A); cepa 790 (B) y *S. eubayanus* (C). Carriles 1 al 5, gradiente térmico de 52, 54, 56, 58 y 60°C; Carril 6, Control negativo (-); Carril 7, marcador de talla molecular. (pb)

Al día de hoy no se han recibido los datos de secuenciación por capilar de los fragmentos.

7.7.2. *Gap* 2

El gel del producto amplificado para la región faltante en la posición 140,561 a 141,053 del *scaffold* 30 (cromosoma 1 del subgenoma *cerevisiae*) se muestra en la Figura 16; en ella se observa un fragmento de ~734pb en las tres cepas en estudio (*S. cerevisiae*, 790 y *S. eubayanus*), dicho fragmento coincide con el tamaño amplificable por los *primers* en el genoma de referencia de *S. cerevisiae*.

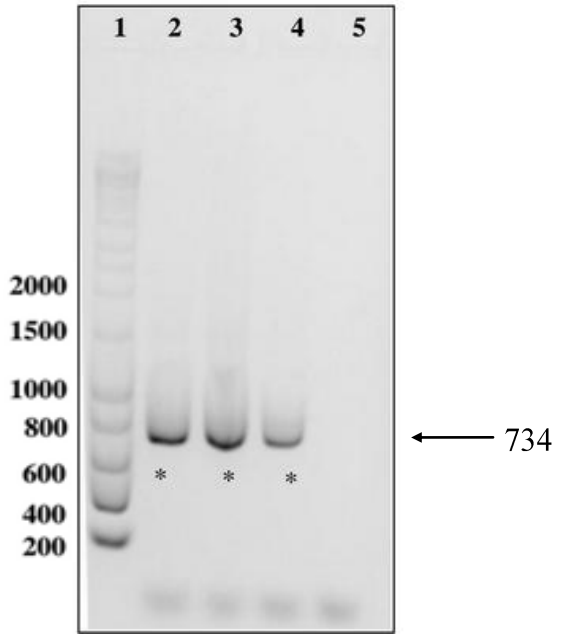


Figura 16: Validación del *Gap* 2. Carriles: 1, Marcador de talla molecular; 2, cepa S288C (Control +); 3, cepa 790; 4, cepa *S. eubayanus*; 5, Control negativo (-) (pb)

Dicha amplificación sugirió una anomalía en el ensamblaje de ese fragmento del genoma. Para verificar que en efecto se tratara del fragmento en el *scaffold* 30, se analizaron los electroferogramas (Figura 17) y se extrajo el fragmento con secuencia de alta calidad. Posteriormente, se buscó dicho fragmento en la secuencia del *scaffold* 30 y la referencia *S. cerevisiae* S288C, hallando un sobrelape en la secuencia conocida del *scaffold* 30 (Figura 18 barra celeste). Un fragmento de 23 nucleótidos (Figura 18 barra

roja) no pudo ser rellenado debido a el decaimiento en la calidad de la secuencia Rv (Figura 18 barra amarilla)

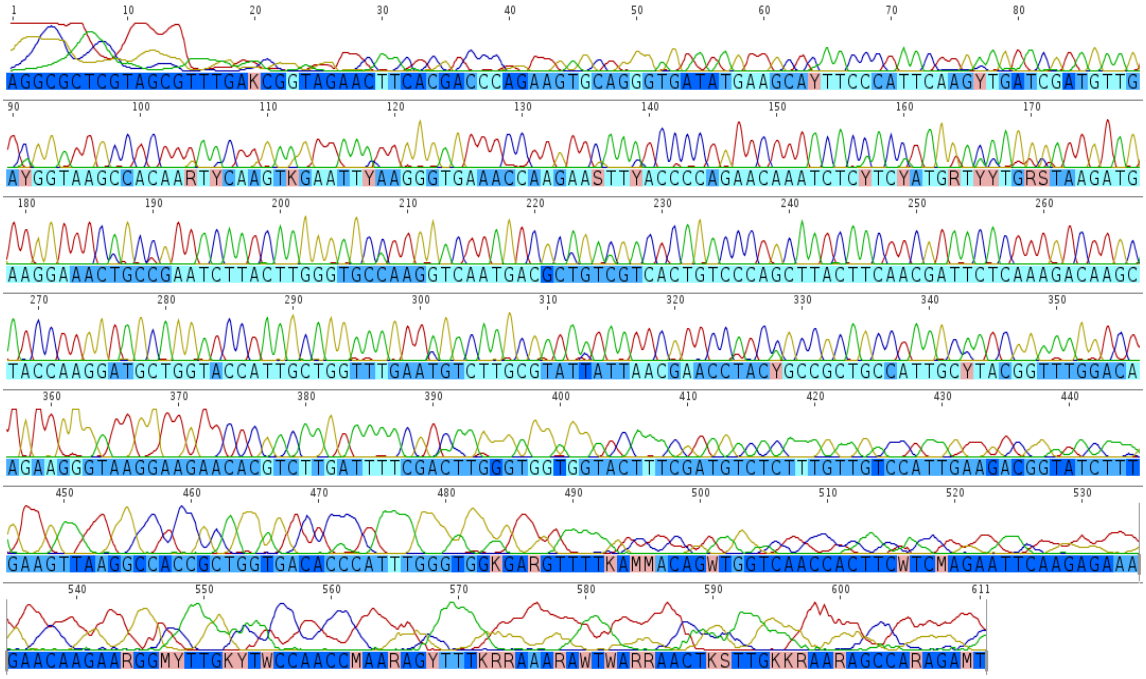


Figura 17: Electroferograma *Gap* 2 en 790

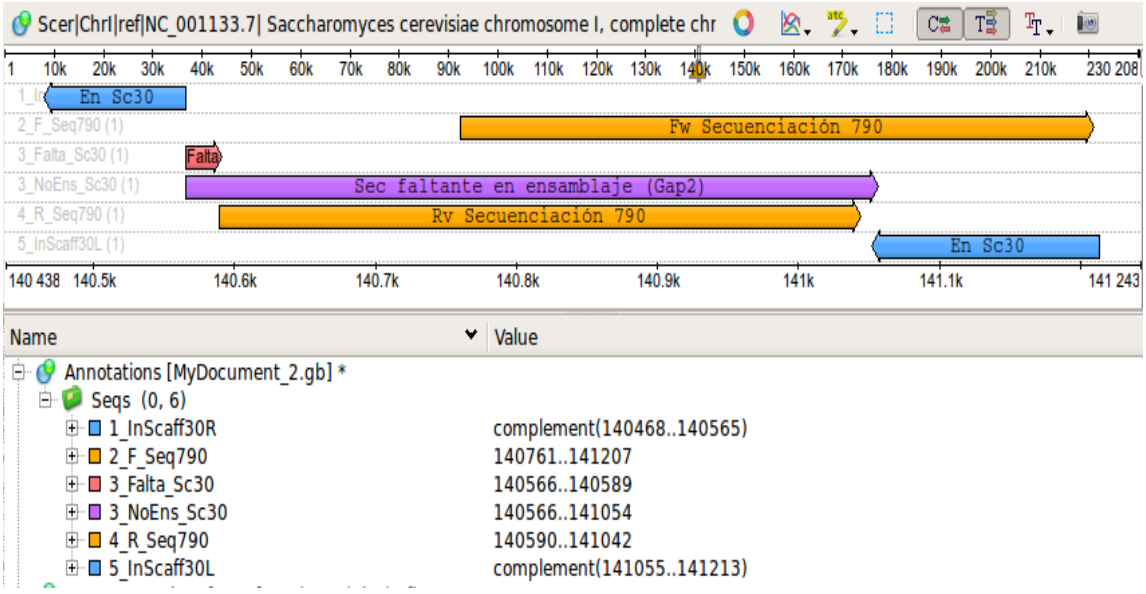


Figura 18: Llenado de *Gap* 2

7.7.3. *Gaps* 3, 4 y 5

La Figura 19 muestra el gel del producto amplificado para los *gaps* 3, 4 y 5, cuyos tamaños esperados eran de 927, 327 y 681 pb respectivamente.

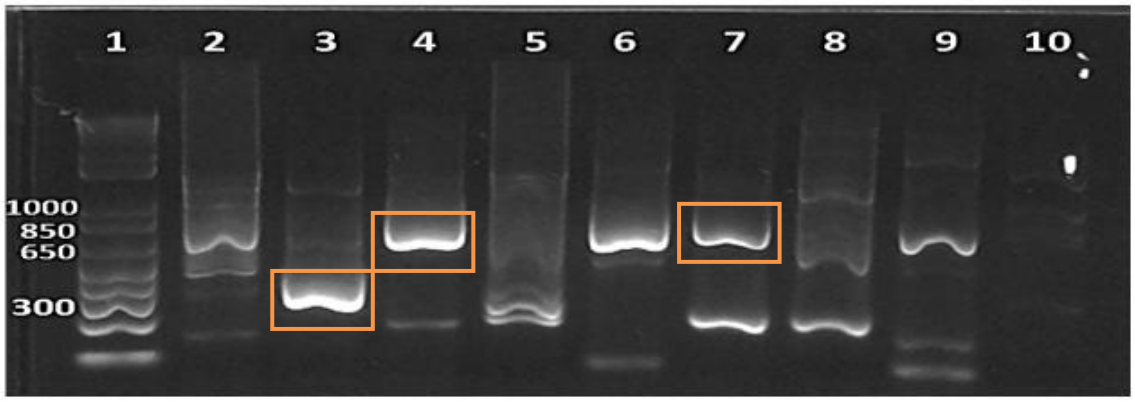


Figura 19: Validación de los *Gaps* 3, 4 y 5. Carriles: 1, Marcador de talla molecular; 2 a 4, *S. cerevisiae gap* 3, 4 y 5 respectivamente; 5 a 7, cepa 790 *gap* 3, 4 y 5 respectivamente; 8 a 10, *S. eubayanus gap* 3, 4 y 5 respectivamente (ver Tabla 5). Los cuadros naranjas indican los tamaños esperados.

Se puede observar que se obtuvieron bandas inespecíficas tanto en *S. cerevisiae* como en 790. Era de esperarse que no apareciera la banda esperada en *S. eubayanus,* debido a que los *primers* fueron diseñados de manera específica para el subgenoma de *S. cerevisiae*, sin embargo, también se encontraron bandas inespecíficas para los gaps 4 y 5.

Debido a estas inconsistencias, se decidió no continuar con la secuenciación de estos *gaps*.

8. DISCUSIÓN

Gracias al advenimiento de nuevas tecnologías de secuenciación masiva como la plataforma Illumina, a la fecha se tienen reportados 34,636 genomas: 1,741 eucariotes, 28,601 procariotes (incluyendo cepas y variedades) y 4,294 virus, además de 5,072 plásmidos (http://www.ncbi.nlm.nih.gov/genome/browse/ - revisado el día 5 de noviembre de 2014). Esa gran cantidad de información ha permitido el desarrollo de la *Genómica Comparativa*, la cual toma como base genomas de organismos cercanos evolutivamente para un más rápido y preciso procesamiento de datos de secuencias nucleotídicas de nuevos organismos, asignándoles una probable organización cromosómica y permitiendo encontrar genes nuevos que confieren propiedades características a los distintos organismos.

Al día de hoy, conocemos la secuencia nucleotídica del genoma de 12 especies del género *Saccharomyces* las cuales son: *S. cerevisiae, S. kudriavzevi , S. paradoxus, S. mikatae, S. bayanus,, S. boulardii, S. arboricola, S. uvarum, S. carlsbergensis, S. pastorianus* y los híbridos *S. cerevisiae - S. kudriavzevii y S. pastorianus - S. weihenstephan* 34/70. De las cuales solo las cuatro últimas son utilizadas con frecuencia en el proceso cervecero.

El tamaño del genoma de las no cerveceras ronda los 12 Mpb (1.2×10^7 pb), mientras que las cerveceras, poseen una talla sustancialmente mayor de ~20 Mpb (2×10^7 pb). Es importante resaltar que el genoma de las cerveceras se compone, en su mayoría, de híbridos (http://www.ncbi.nlm.nih.gov/genome/?term=saccharomyces - revisado el día 5 de noviembre de 2014).

Según los resultados obtenidos, se conoce ahora que la levadura cervecera en estudio (clave 790) posee, tentativamente, 32 cromosomas y un tamaño calculado en 22.7 Mpb (2.27×10^7 pb). El principal obstáculo para conocer de manera exacta el número de cromosomas, se debió a una interferencia en la parte superior de los geles tipo PFGE,

donde, posiblemente, se pudiesen encontrar cromosomas grandes (Dr. Luis Damas. Comunicación personal). Aunque existe esta posibilidad, nuestros resultados parecen ser los correctos, ya que la talla molecular global es muy semejante a lo reportado para otras levaduras cerveceras tipo lager. Por lo que existe una probabilidad muy baja de encontrar cromosomas extras (Dunn y Sherlock, 2008; Nakao *et al*., 2009; Borneman *et al*., 2012; Walter *et al*., 2014).

El tamaño calculado coincide con la información previamente conocida, ya que consta de 16 cromosomas del subgenoma de *S. cerevisiae* y 16 cromosomas de *S. eubayanus.* Así mismo, el tamaño es cercano a la suma de estos dos genomas (~12 Mpb cada uno). Esto es consistente tanto con los datos previamente reportados por Nakao y cols. (2009), Borneman y cols. (2012), así como por el grupo de Walther y cols. (2014), estos últimos secuenciaron y ensamblaron el genoma de las levaduras cerveceras *S. carlsbergensis* (78 *scaffolds,* 29 cromosomas con 19.5 Mpb de longitud) y *S. weihenstephan* (985 scaffolds, ~29 cromosomas y 22.9 Mpb).

Existen distintos niveles de ensamblaje de los genomas secuenciados y reportados. Muchas secuencias nucleotídicas de los genomas han sido publicadas como "borrador" de *contigs* (*draft genome*), sin darles una orientación y ubicación cromosómica; quedando a consideración del investigador el tamaño minino de cada contig para ser considerado para su publicación. En otras ocasiones, se hacen públicos los *scaffolds* obtenidos a partir de una asignación, dirección y orden a los contigs y tomando parámetros distintos en cuanto al número de *contigs* necesarios para conformar los *scaffolds.* En el presente trabajo se consideraron al menos dos *contigs* de tamaño mínimo de 500 pb para formar los *scaffolds*

El mejor ensamblaje del genoma de la 790, arrojó un total de 133 *scaffolds*, 65 de los cuales tienen un tamaño > 10 kpb, similar o mejor a los reportados anteriormente. La Tabla 10 muestra una comparación del nivel de ensamblaje de cada genoma reportado

del género *Saccharomyces*. (http://www.ncbi.nlm.nih.gov/genome/?term=saccharomyces - revisado el día 5 de noviembre de 2014).

Tabla 10: Nivel de ensamblaje de genomas secuenciados de especies de *Saccharomyces*

Organismo	Genoma Terminado	Número de Cromosomas	Número de *Scaffolds*	Número de *Contigs*	Tamaño (Mpb)
S. cerevisiae	✓	16	17	-	12.16
S. kudriavzevii	✗	16	2,054	-	11.19
S. pastorianus	✗	-	-	2,425	24.21
S. paradoxus	✗	16	-	832	11.87
S. mikatae	✗	16	-	1,648	11.47
S. bayanus	✗	16	-	586	11.87
S. boulardii	✗	16	48	-	11.64
S. arboricola	✓	16	35	-	11.62
S. uvarum	✗	-	-	3,985	11.60
S. carlsbergensis	✗	29	77	-	19.37
S. cerevisiae - *S. kudriavzevii*	✗	-	60	419	23.37
S. pastorianus – *S. weihenstephan* 34/70	✗	-	-	1,358	22.96
790	✗	**~32**	**133**	-	**22.74**

La cobertura del subgenoma de *S. cerevisiae* varió entre un 73.78% (cromosoma 1) y un 93.79% (cromosoma 11). Por otro lado, la cobertura del subgenoma de *S. eubayanus* varió de 60.47% (cromosoma 3) a un 97.84% (cromosoma 14). Dicha variación puede deberse a la presencia de secuencias de mayor complejidad que por la naturaleza del proceso de secuenciación, o las limitaciones del algoritmo del software utilizado, no pudo resolverse para obtener el 100% de la cobertura.

Gracias a los alineamientos realizados pudimos constatar que la identidad de ambos subgenomas (*S. cerevisiae* y *S. eubayanus*) es de ~80%, lo cual favoreció a un mejor ensamblaje y distinción entre los mismos, lo cual también es consistente a lo encontrado por Nakao y cols. (2009) y Walther y cols. (2014).

La variación en el número de copias de cada cromosoma correspondiente a los subgenomas, así como la presencia de translocaciones entre los mismos coincide con lo previamente observado por Dunn y Sherlock (2008), Nakao y cols. (2009), Borneman y cols. (2012) y Walther y cols. (2014). La Tabla 11 muestra el número de copias para cada cromosoma de los subgenomas de *S. cerevisiae* (790cer) y *S. eubayanus* (790eub):

Tabla 11: Número de copias de los cromosomas de 790

Cromosoma	790cer	790eub	Cromosoma	790cer	790eub
1	3	3	**9**	3	2
2	2	3	**10**	1	4
3	2.5	2.5	**11**	2	2
4	3	2	**12**	3	2
5	2	2	**13**	3	3
6	3	3	**14**	2	3
7	2	3	**15**	2.5	2.5
8	2	2	**16**	3	2

Con respecto a la intensidad de la hibridación y translocaciones cromosómicas, encontramos, básicamente dos tipos de arreglos: i) cromosomas híbridos pero con continuidad en la secuencia (translocación) y ii) híbridos con regiones cortas de otro cromosoma (recombinación homóloga). Estas características fueron encontradas en los

cromosomas 3, 4, 7, 11, 15 y 16, en los que algunas regiones del subgenoma de *S. cerevisiae* se encuentran insertadas en el subgenoma de *S. eubayanus*, y viceversa. Los puntos de translocación posiblemente contengan secuencias conservadas con mayor identidad entre ambos subgenomas.

Las secuencias nucleotídicas obtenidas por medio de Ion Torrent no fueron utilizadas para el ensamblaje del genoma debido a la baja calidad de las mismas.

La presencia y abundancia de 'Ns' (nucleótidos no asignados aun) en la Versión 3 del ensamblaje se debe al proceso de secuenciación de las librerías Illumina *Mate pair,* las cuales tienen un espaciador de 350 pb y 8 kpb, con el fin de unir y dar dirección a los *contigs*. Como se demostró en el llenado de *gaps*, se requiere de la validación de cada uno para lograr una mayor cobertura de cada cromosoma.

Gap 1: Se requirió optimizar la PCR con el fin de eliminar fragmentos inespecíficos. Con ello se obtuvieron las bandas esperadas tanto en *S. cerevisiae* S288C como en 790. Era de esperarse que en *S. eubayanus* no se encontrara dicho fragmento ya que los *primers* fueron diseñados específicamente para el subgenoma *cerevisiae*. Al día de hoy no se han recibido los datos de secuencaicion por capilar de los fragmentos.

Gap 2: Como se mencionó en los resultados, el fragmento amplificado en *S. eubayanus* fue inesperado debido a que los *primers* fueron diseñados específicamente para el subgenoma *cerevisiae*. En la secuencia ensamblada y correspondiente a *eubayanus* no se ha encontrado el fragmento amplificado en esa referencia. Esto puede deberse a que en la cepa cervecera este fragmento no exista o que tiene una identidad tan alta con el subgenoma de *cerevisiae* que el ensamblador no pudo discernir entre ambos, asignando ese fragmento únicamente en dicho subgenoma. El sobrelape entre las secuencias ya conocidas en el *scaffold* 30 y el fragmento rellenado nos hacen pensar que, en efecto, corresponde al cromosoma 1 del subgenoma *cerevisiae*.

Gaps 3, 4 y 5: La inespecificidad en los productos amplificados pudo deberse a un error en el diseño de los *primers*, aun y cuando estos fueron cuidadosamente estudiados para que amplificasen solo la región de interés. De los tres *gaps*, solo el 5 se pudiera llegar a mejorar con un gradiente de temperatura, para eliminar la banda inespecífica, sin embargo, se realizará como parte de las perspectivas y trabajos a futuro en este proyecto.

9. CONCLUSIONES

A partir de lo encontrado en el presente trabajo de investigación se puede concluir que:

La levadura 790 es un hibrido entre *Saccharomyces cerevisiae y S. eubayanus*

Su genoma nuclear está formado por 32 cromosomas

16 cromosomas corresponden al subgenoma de *S. cerevisiae* y 16 al subgenoma *S. eubayanus*.

En la versión final del ensamblaje se conformaron 133 *scaffolds.*

65 de los *scaffolds* son de un tamaño mayor a 10 kpb. Esto se puede interpretar como si cada cromosoma estuviese dividido en 2 *scaffolds*.

En cada subgenoma, encontramos 10 cromosomas con identidad al 100% para cada especie.

4 cromosomas presentaron translocaciones continuas (cromosomas 3, 7, 11 y 15, todos en su subgenoma de *S. eubayanus*).

2 cromosomas presentaron regiones no continuas (cromosomas 4 y 16 del subgenoma de *S. cerevisiae*). Lo cual nos indica eventos de recombinación homóloga

Los datos obtenidos, tanto de número de cromosomas como del tamaño del genoma y cantidad de *scaffolds* es consistente con respecto a genomas del género *Saccharomyces*.

10. PERSPECTIVAS

1. Se requiere continuar trabajando en el proyecto para completar las secuencias nucleotídicas no resueltas (*gaps*), así como la asignación y reducción de *scaffolds*.
2. Con los datos obtenidos se puede continuar con la anotación del genoma. Esto puede ser de utilidad para corroborar la información ya conocida por métodos bioquímicos así como encontrar genes anteriormente no considerados con importancia industrial
3. Es necesario mejorar el ensamblaje. Para ello se requiere diseñar oligonucleótidos que flanqueen las regiones no resueltas ('Ns' en *gaps*) con el fin de "rellenarlos" y optimizar los ya diseñados para los gaps 3, 4 y 5.
4. Aunque faltan regiones del genoma por resolver, los datos obtenidos hasta el momento son de enorme utilidad para la comprensión de la estructura general del mismo y sirve como base para la búsqueda de mayor información con respecto a las características intrínsecas de la levadura cervecera.

11. LITERATURA CITADA.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Ref. Source.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., y Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.

Borneman, A.R., Desany, B.A., Riches, D., Affourtit, J.P., Forgan, A.H., Pretorius, I.S., Egholm, M., y Chambers, P.J. (2012). The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii* origins. FEMS Yeast Res. *12*, 88–96.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., *et al.* (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. *18*, 630–634.

Cherry, J.M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R.K., *et al.* (1997). Genetic and physical maps of *Saccharomyces cerevisiae*. Nature *387*, 67–73.

Delcher, A.L., Salzberg, S.L., y Phillippy, A.M. (2003). Using MUMmer to identify similar regions in large sequence sets. Curr. Protoc. Bioinforma. 10–13.

Dunn, B., y Sherlock, G. (2008). Reconstruction of the genome origins and evolution of the hybrid lager yeast *Saccharomyces pastorianus*. Genome Res. *18*, 1610–1623.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., *et al.* (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature *260*, 500–507.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., *et al.* (1995). Whole-Genome Random Sequencing and Assembly of *Haemophilus Influenzae* Rd. Science *269*, 496–498.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., *et al.* (1996). Life with 6000 genes. Science *274*, 546, 563–567.

Idury, R.M., y Waterman, M.S. (1995). A New Algorithm for DNA Sequence Assembly. J. Comput. Biol. *2*, 291–306.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., y Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. *36*, W5–W9.

Jou, W.M., Haegeman, G., Ysebaert, M., y Fiers, W. (1972). Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. Nature *237*, 82–88.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., *et al.* (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., y FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Legras, J.-L., Merdinoglu, D., Cornuet, J.-M., y Karst, F. (2007). Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. Mol. Ecol. *16*, 2091–2102.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., *et al.* (2011). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. Brief. Funct. Genomics.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., y Law, M. (2012). Comparison of Next-Generation Sequencing Systems. J. Biomed. Biotechnol. *2012*, 1–11.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

McGovern, P.E., Zhang, J., Tang, J., Zhang, Z., Hall, G.R., Moreau, R.A., Nuñez, A., Butrym, E.D., Richards, M.P., Wang, C.-S., *et al.* (2004). Fermented beverages of pre- and proto-historic China. Proc. Natl. Acad. Sci. U. S. A. *101*, 17593–17598.

Nakao, Y., Kanamori, T., Itoh, T., Kodama, Y., Rainieri, S., Nakamura, N., Shimonaga, T., Hattori, M., y Ashikari, T. (2009). Genome Sequence of the Lager Brewing Yeast, an Interspecies Hybrid. DNA Res. *16*, 115–129.

Naumova, E.S., Korshunova, I.V., Jespersen, L., y Naumov, G.I. (2003). Molecular genetic identification of *Saccharomyces sensu stricto* strains from African sorghum beer. FEMS Yeast Res. *3*, 177–184.

Nijkamp, J.F., van den Broek, M.A., Geertman, J.-M.A., Reinders, M.J., Daran, J.-M.G., y de Ridder, D. (2012). De novo detection of copy number variation by co-assembly. Bioinformatics *28*, 3195–3202.

Okonechnikov, K., Golosova, O., Fursov, M. (2012). Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics *28*, 1166–1167.

Peng, Y., Leung, H.C.M., Yiu, S.M., y Chin, F.Y. L. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinforma. Oxf. Engl. *28*, 1420–1428.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., y Nyrén, P. (1996). Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. Anal. Biochem. *242*, 84–89.

Rozen, S., y Skaletsky, H. (1999). Primer3 on the WWW for General Users y for Biologist Programmers. In Bioinformatics Methods and Protocols, Humana Press, pp. 365–386.

Sanger, F., Nicklen, S., y Coulson, A.R. (1977a). DNA Sequencing with Chain-Terminating Inhibitors. Proc. Natl. Acad. Sci. *74*, 5463–5467.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M., y Smith, M. (1977b). Nucleotide sequence of bacteriophage φX174 DNA. Nature *265*, 687–695.

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. Nat. Methods *5*, 16–18.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., y Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature *321*, 674–679.

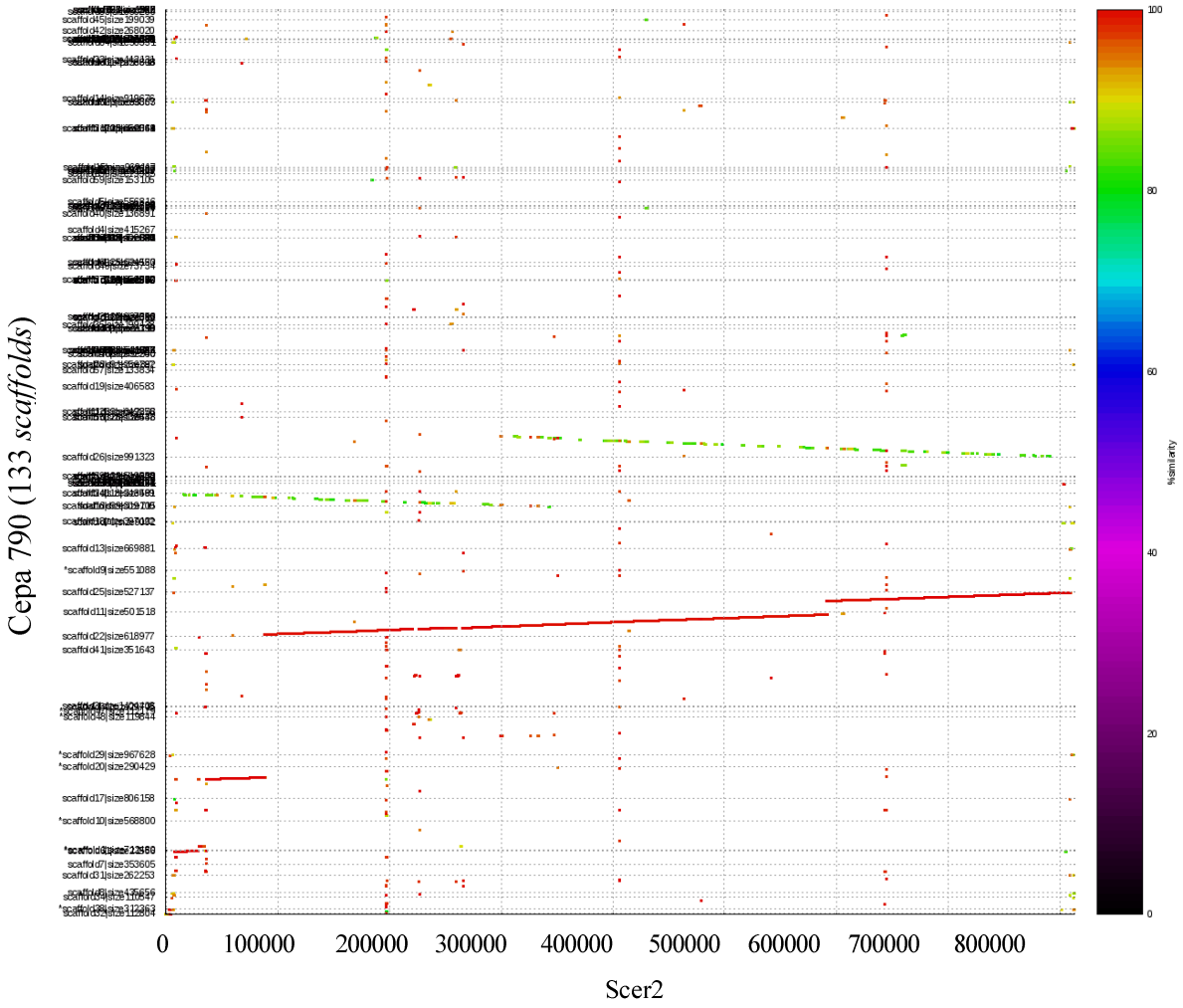
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Walther, A., Hesselbart, A., y Wendland, J. (2014). Genome sequence of *Saccharomyces carlsbergensis*, the world's first pure culture lager yeast. G3 Bethesda Md *4*, 783–793.

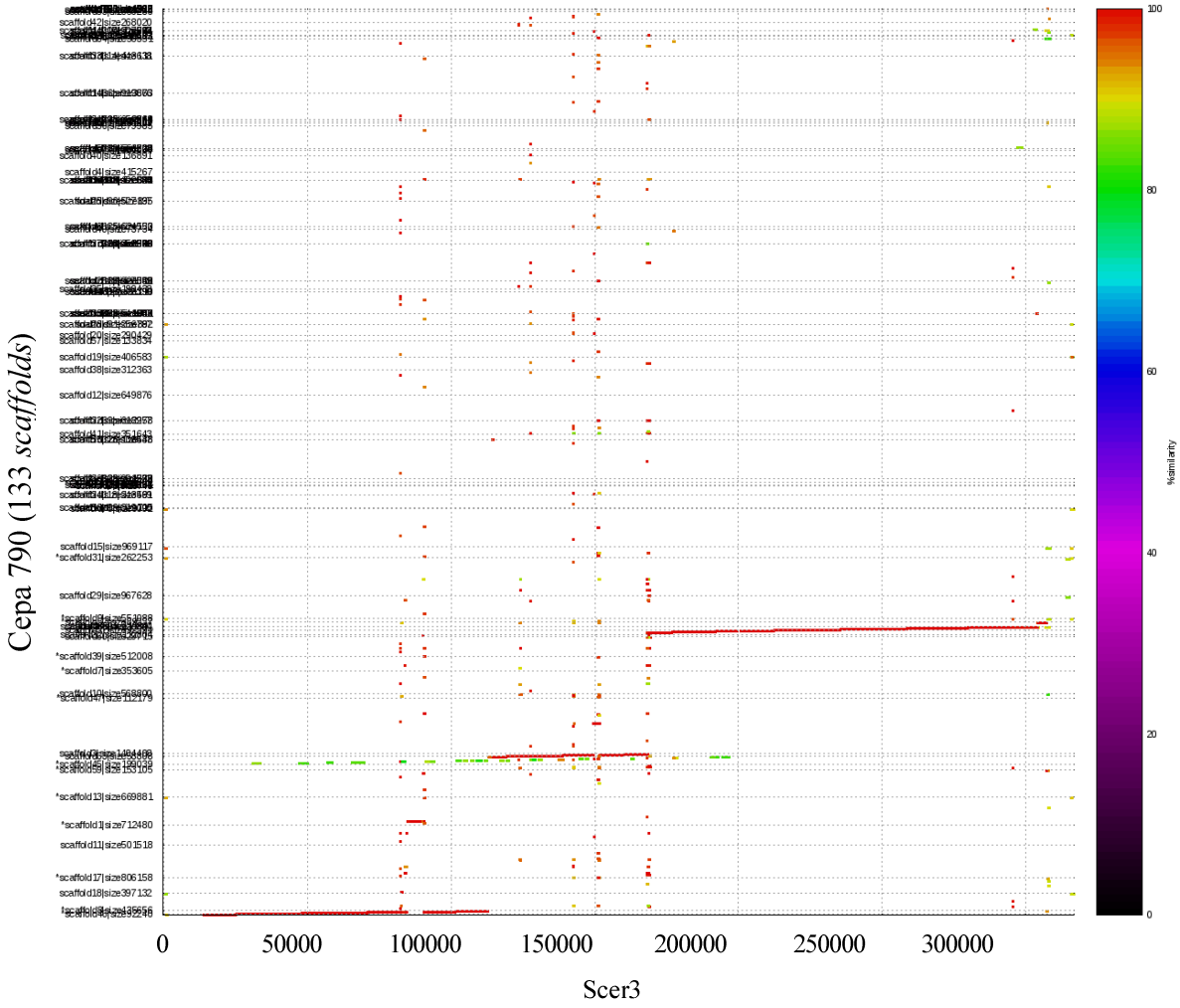
Watson, J.D., y Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

12. ANEXOS

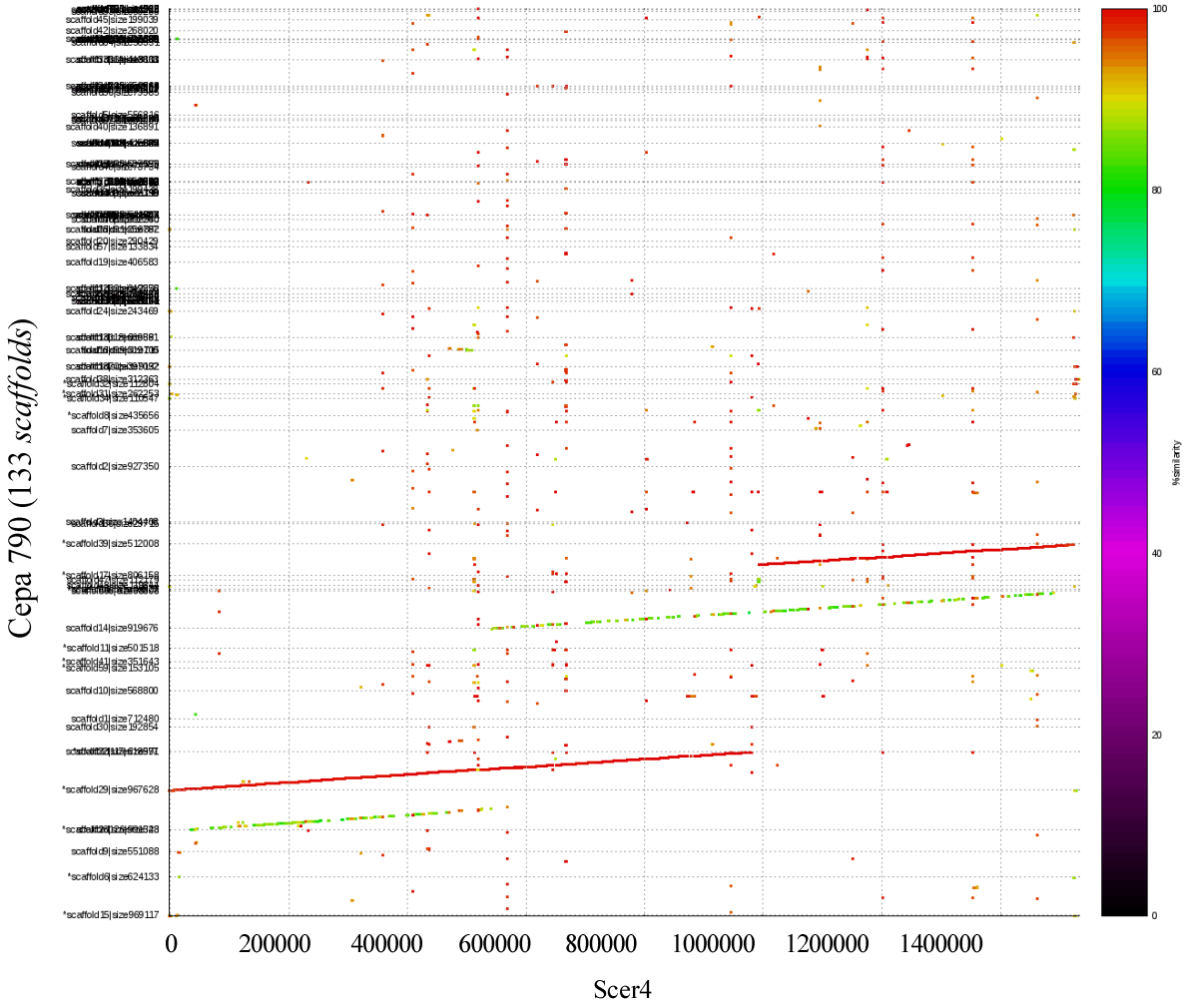
De la sección **7.3.1. 790 vs *S. cerevisiae* S288C:**



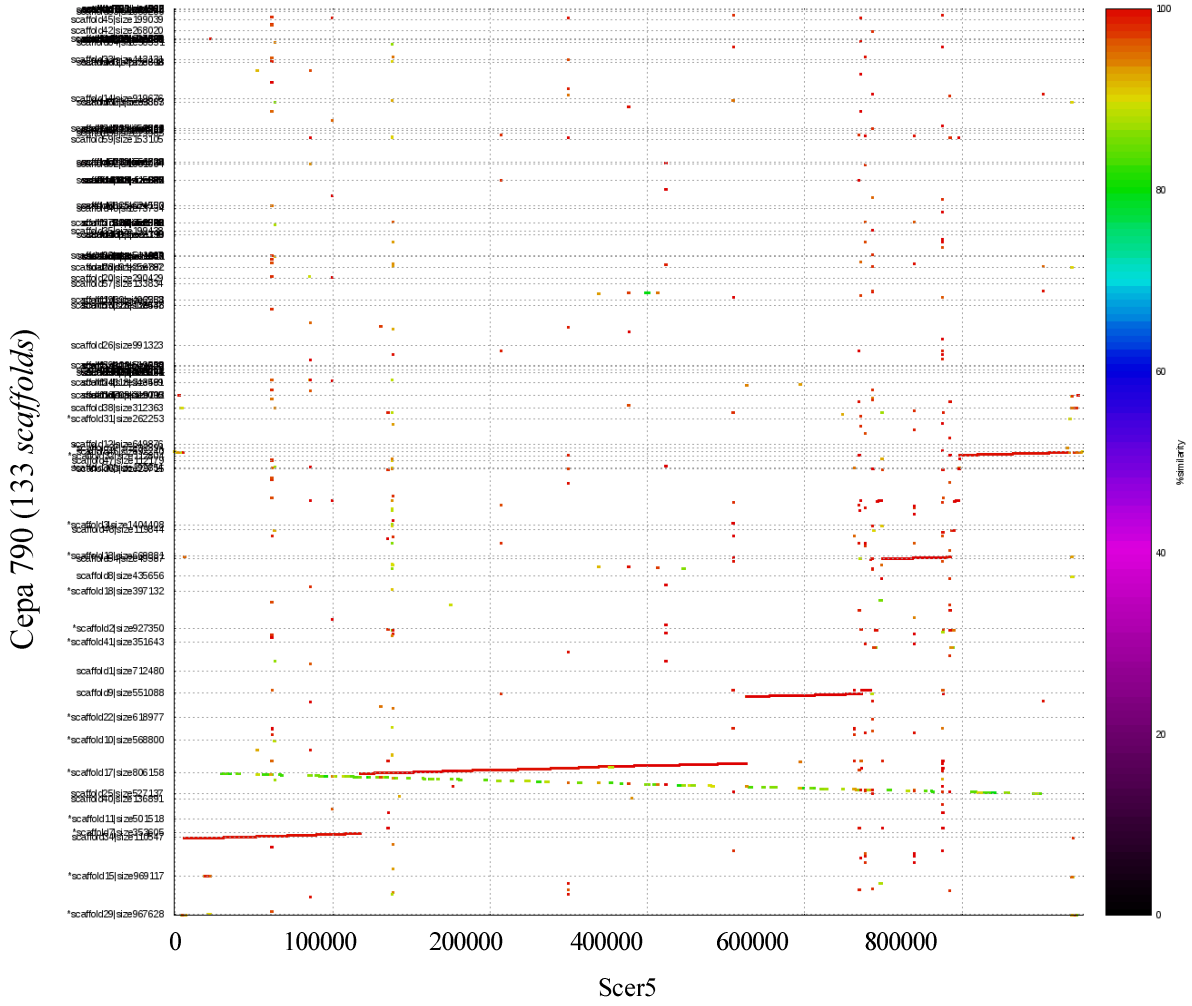
1. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 2 (referencia)



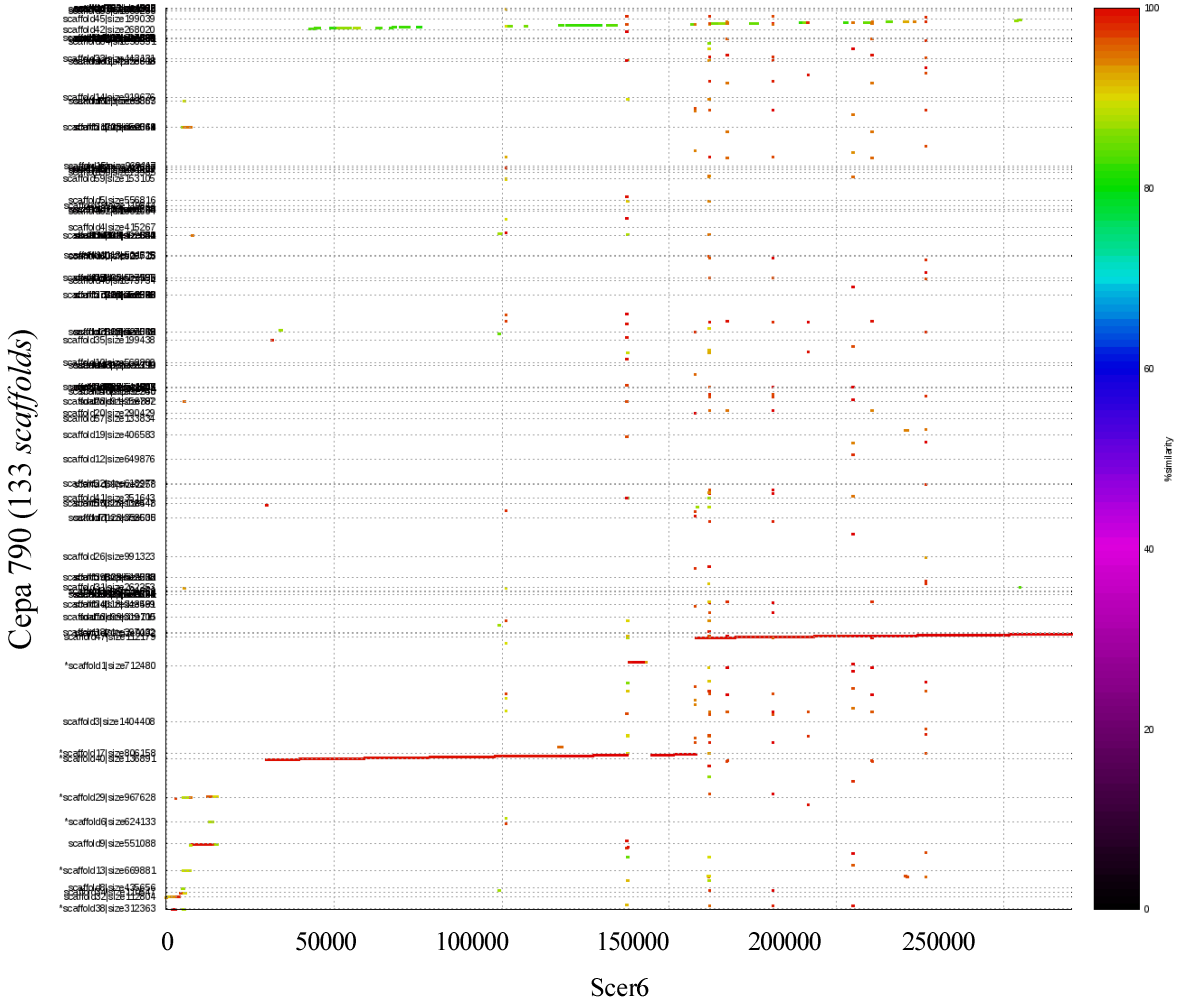
2. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 3 (referencia)



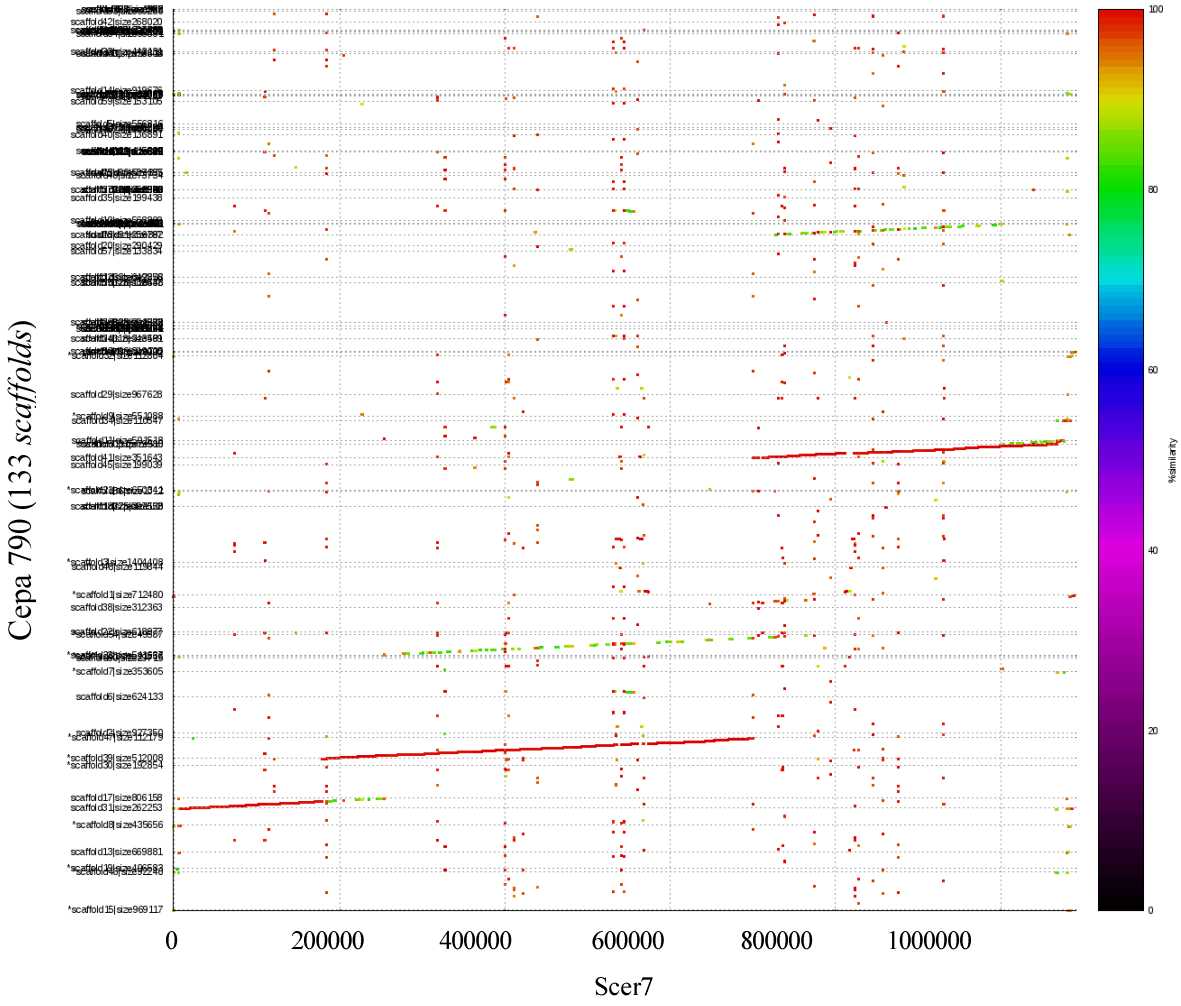
3. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 4 (referencia)



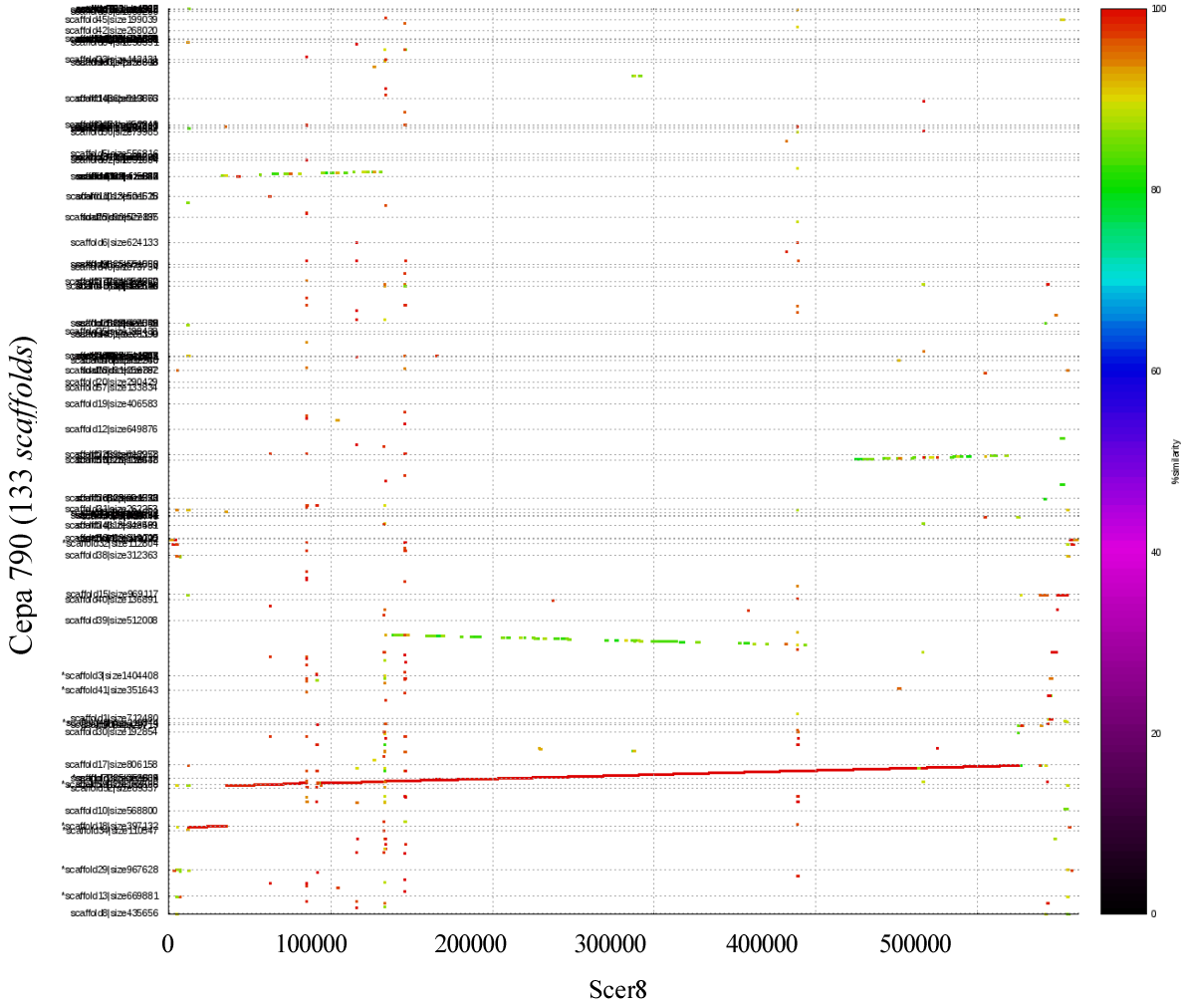
4. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 5 (referencia)



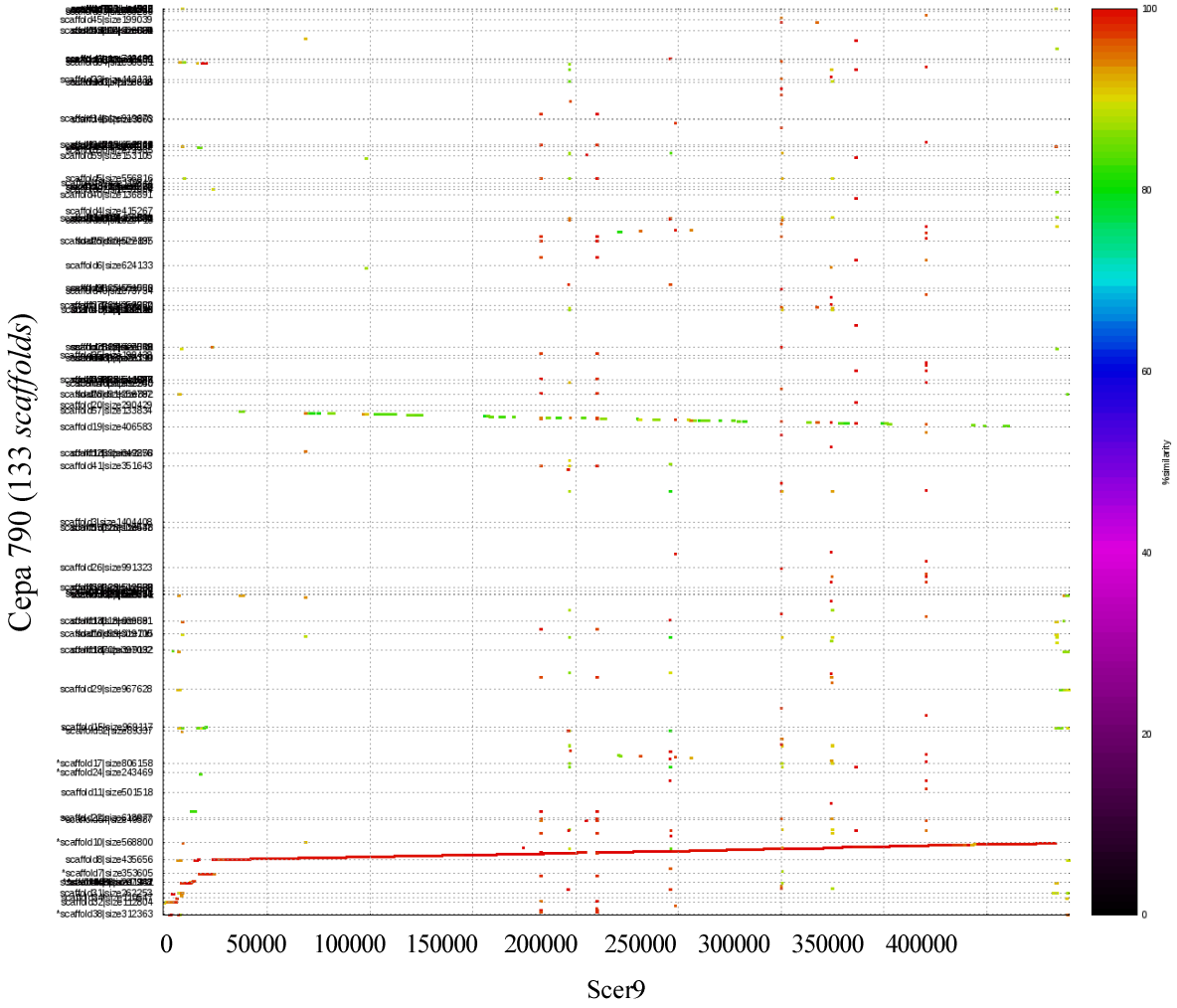
5. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 6 (referencia)



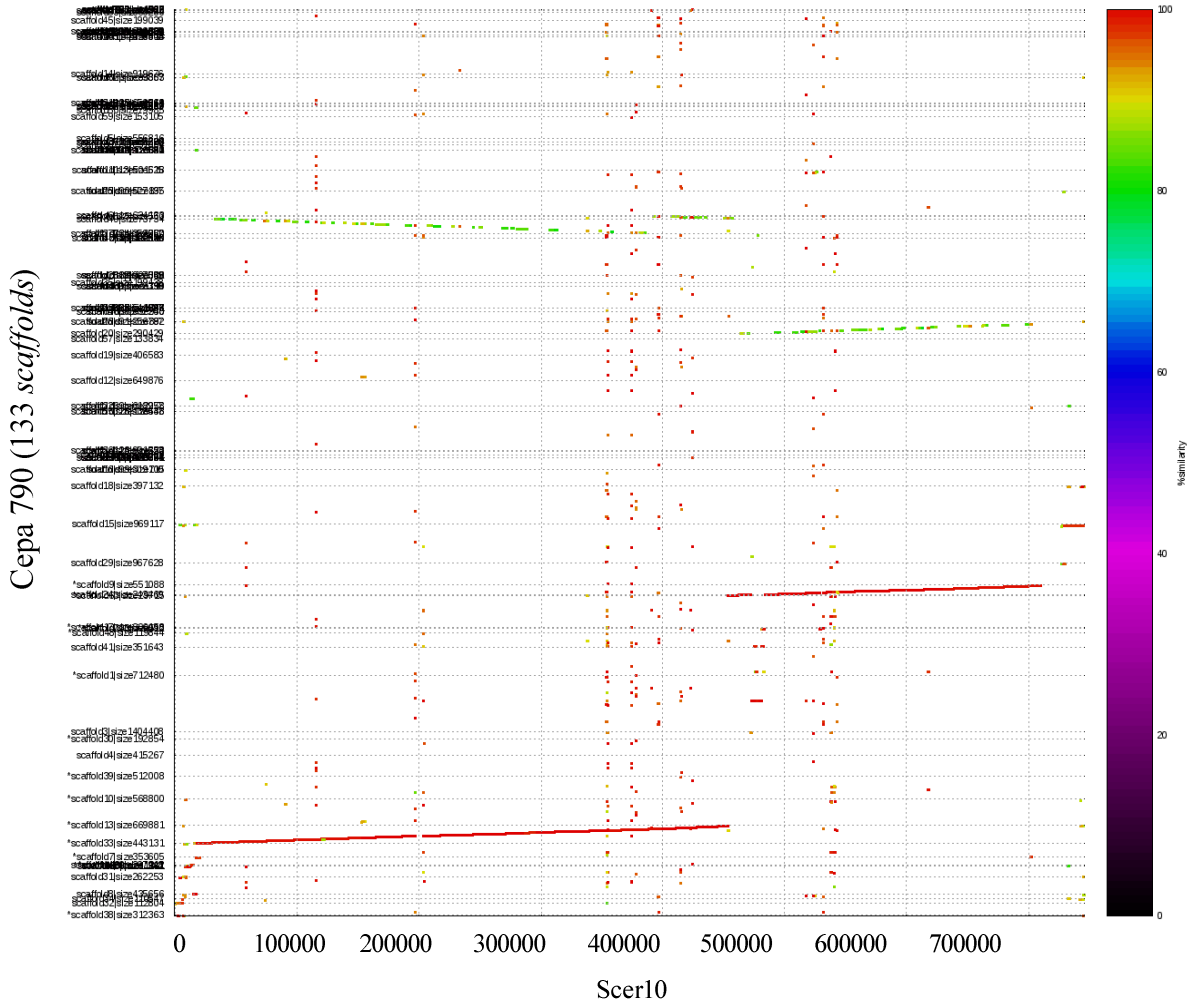
6. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 7 (referencia)



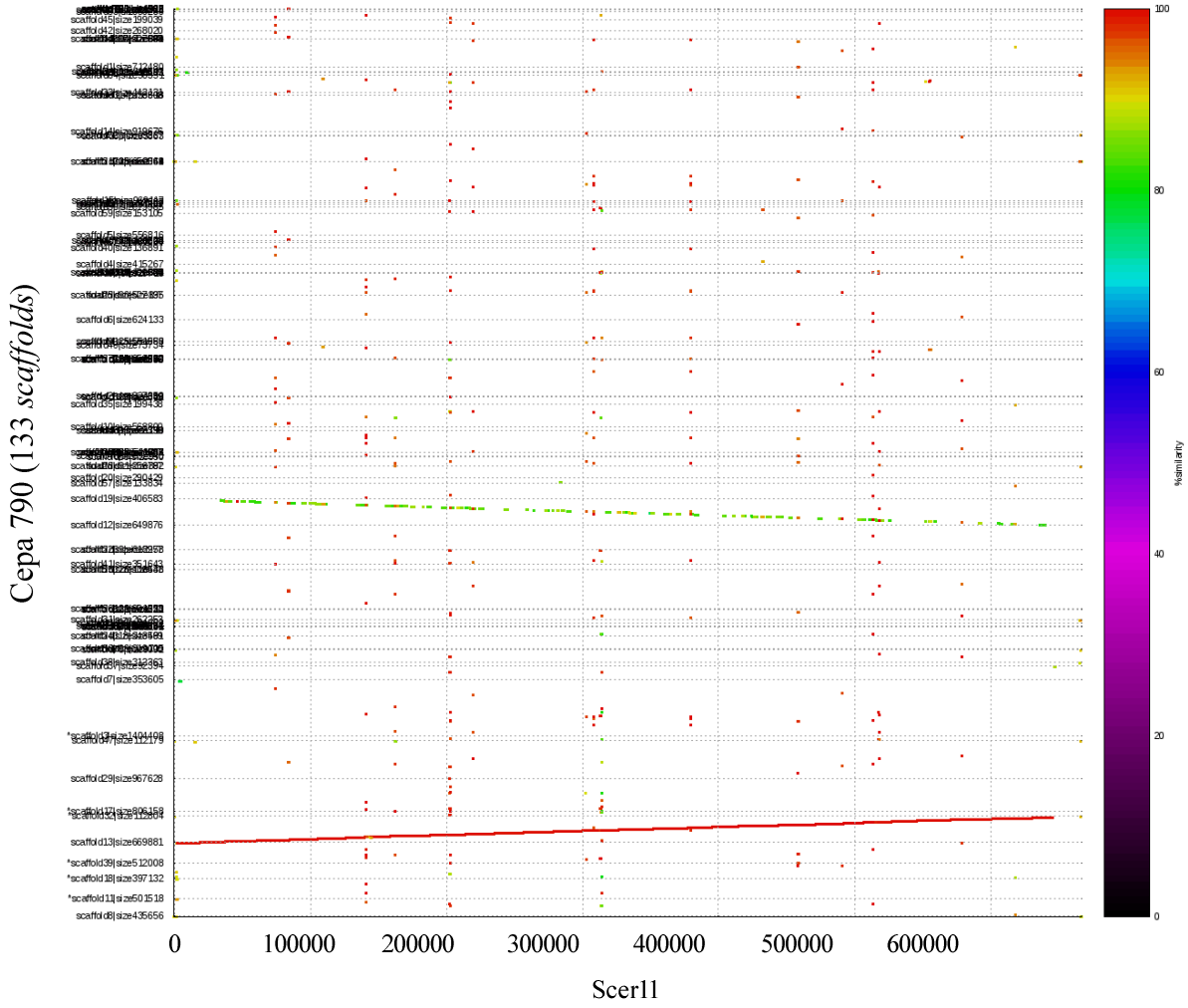
7. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 8 (referencia)



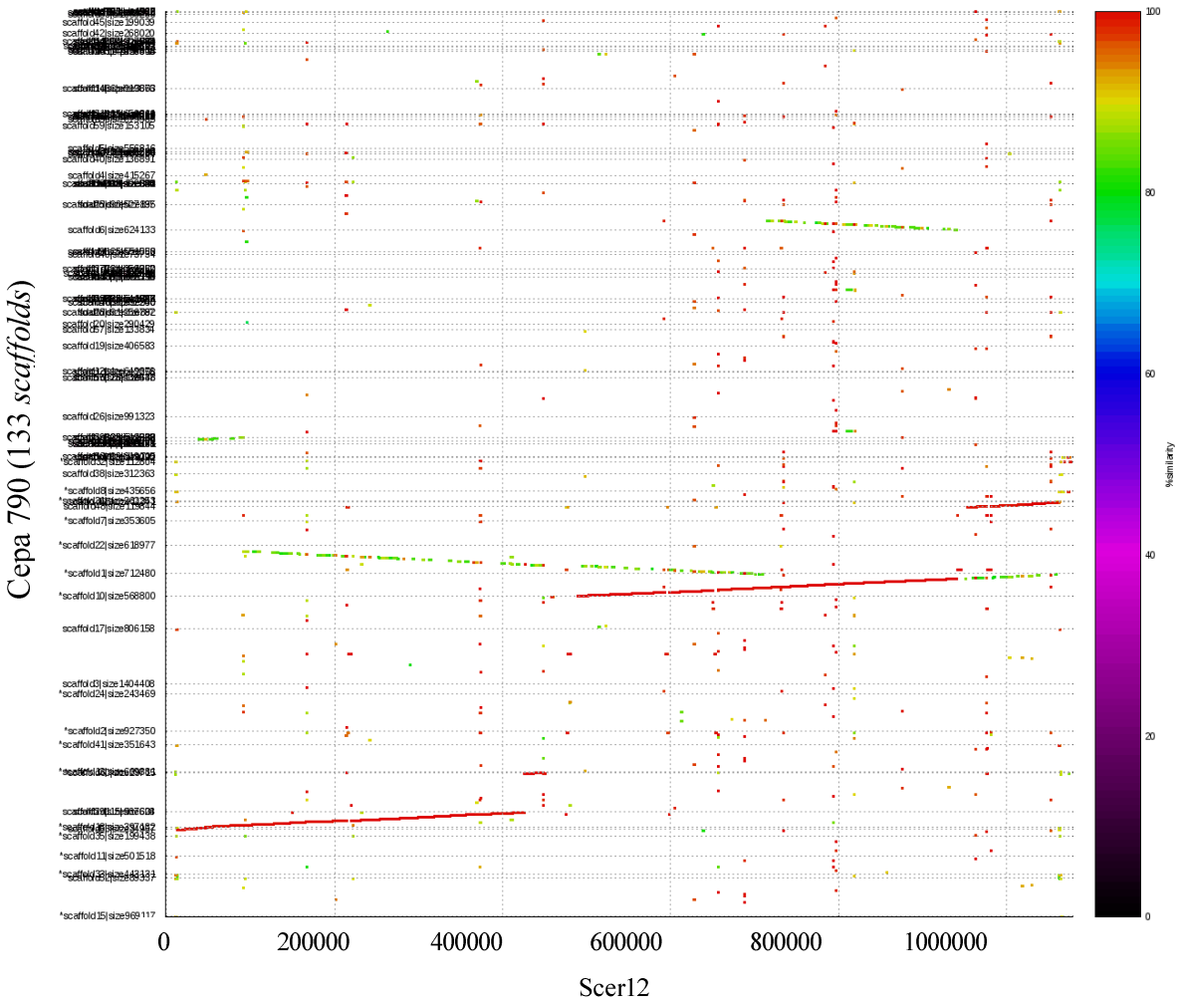
8. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 9 (referencia)



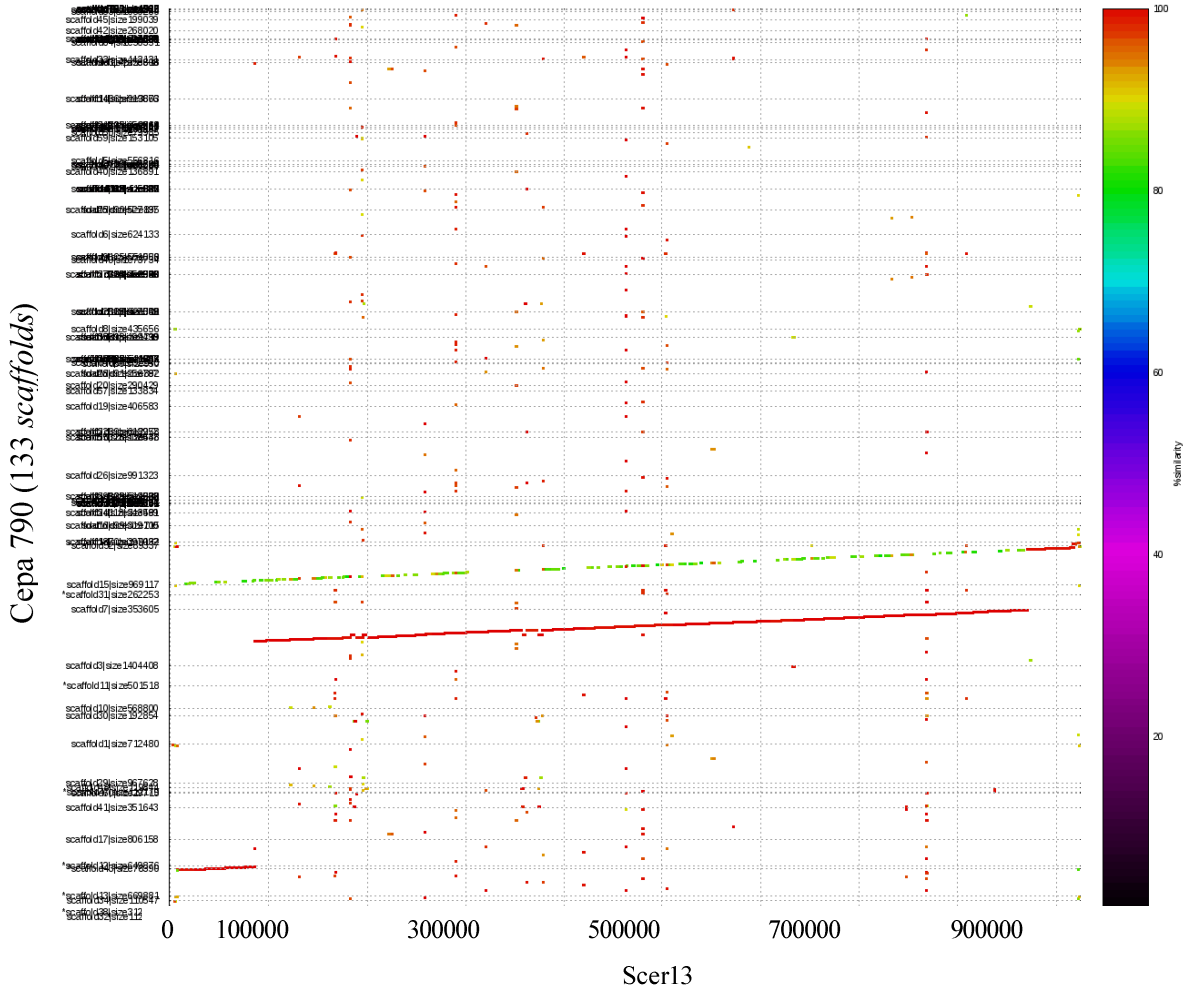
9. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 10 (referencia)



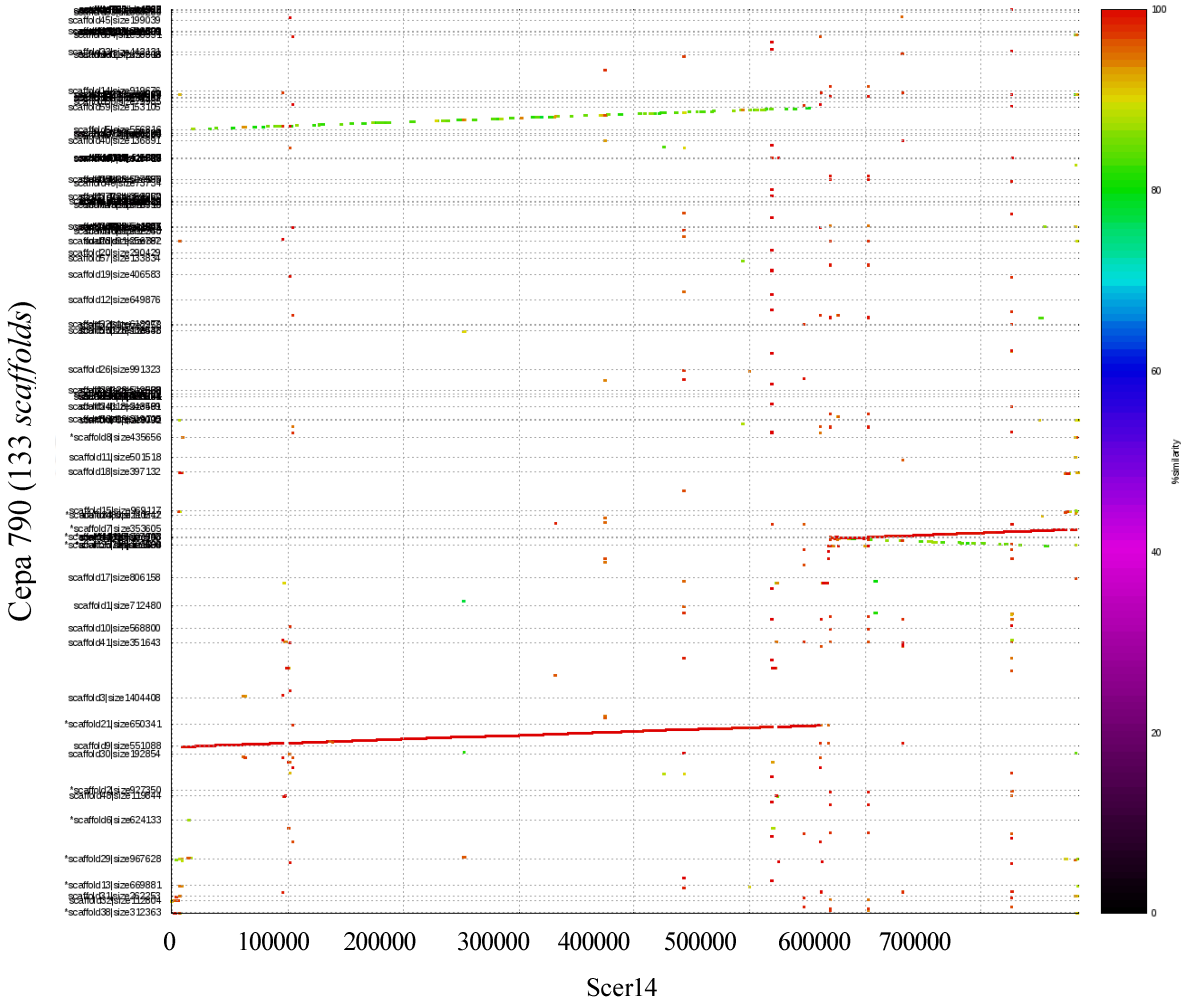
10. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 11 (referencia)



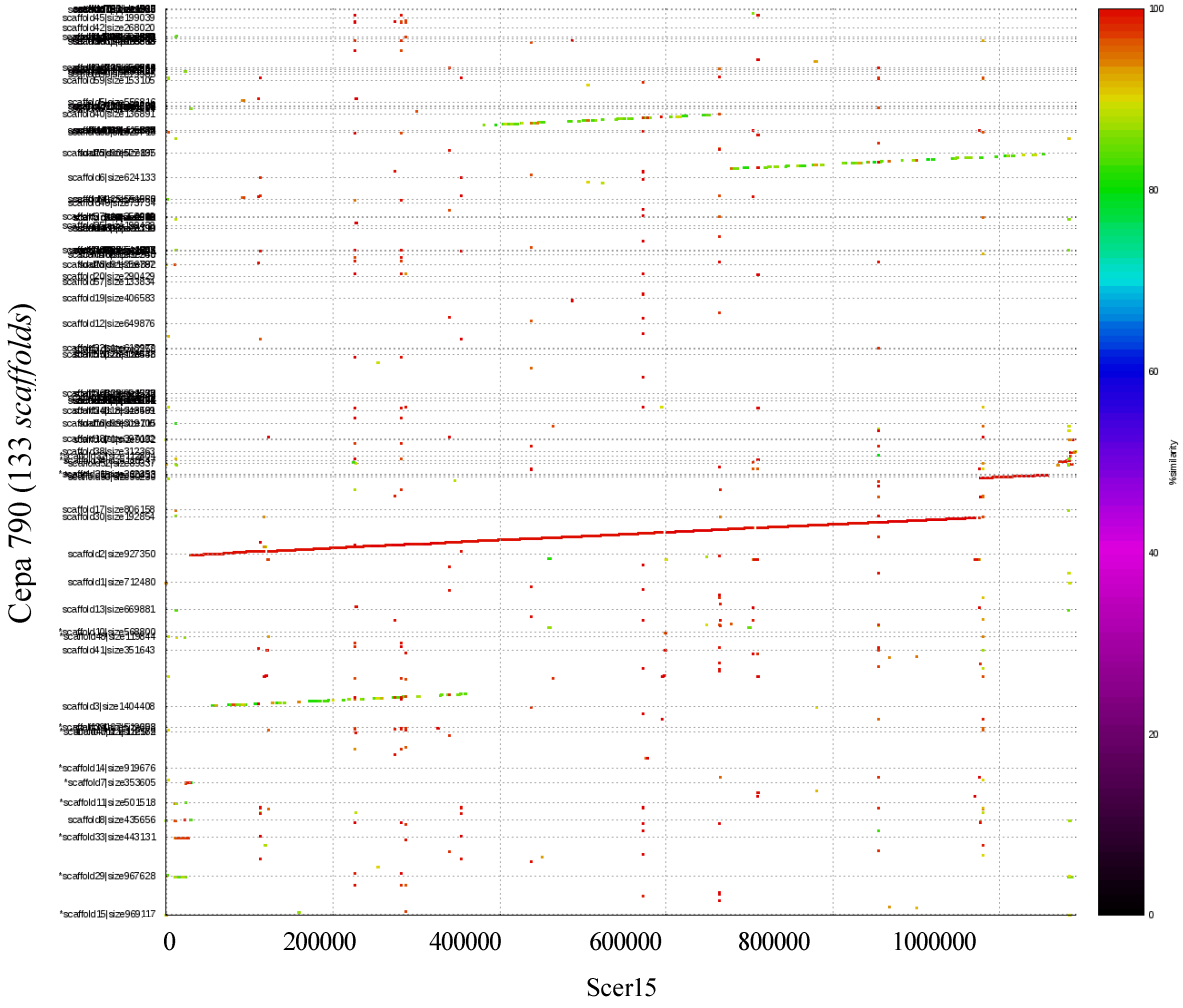
11. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 12 (referencia)



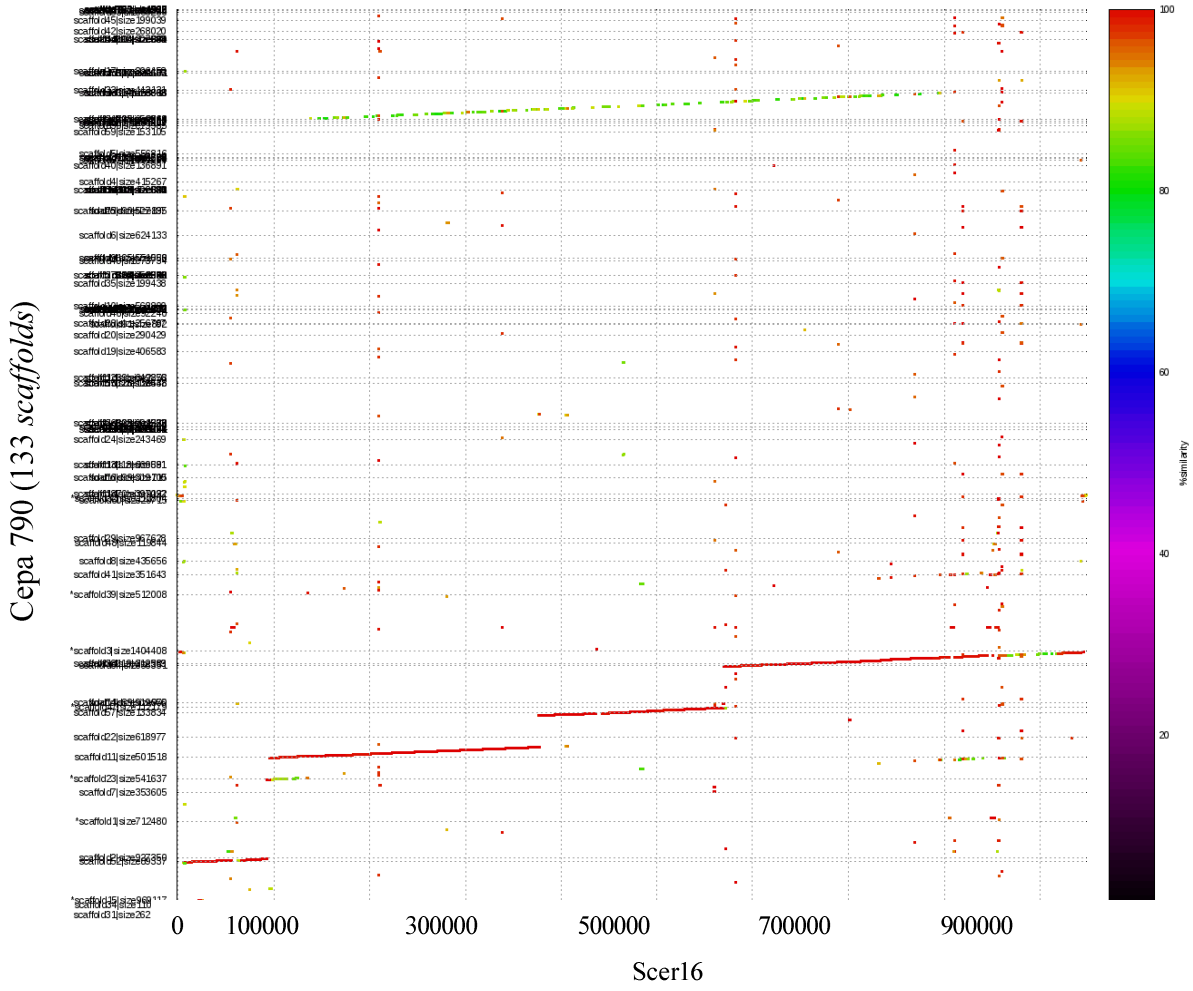
12. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 13 (referencia)



13. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 14 (referencia)



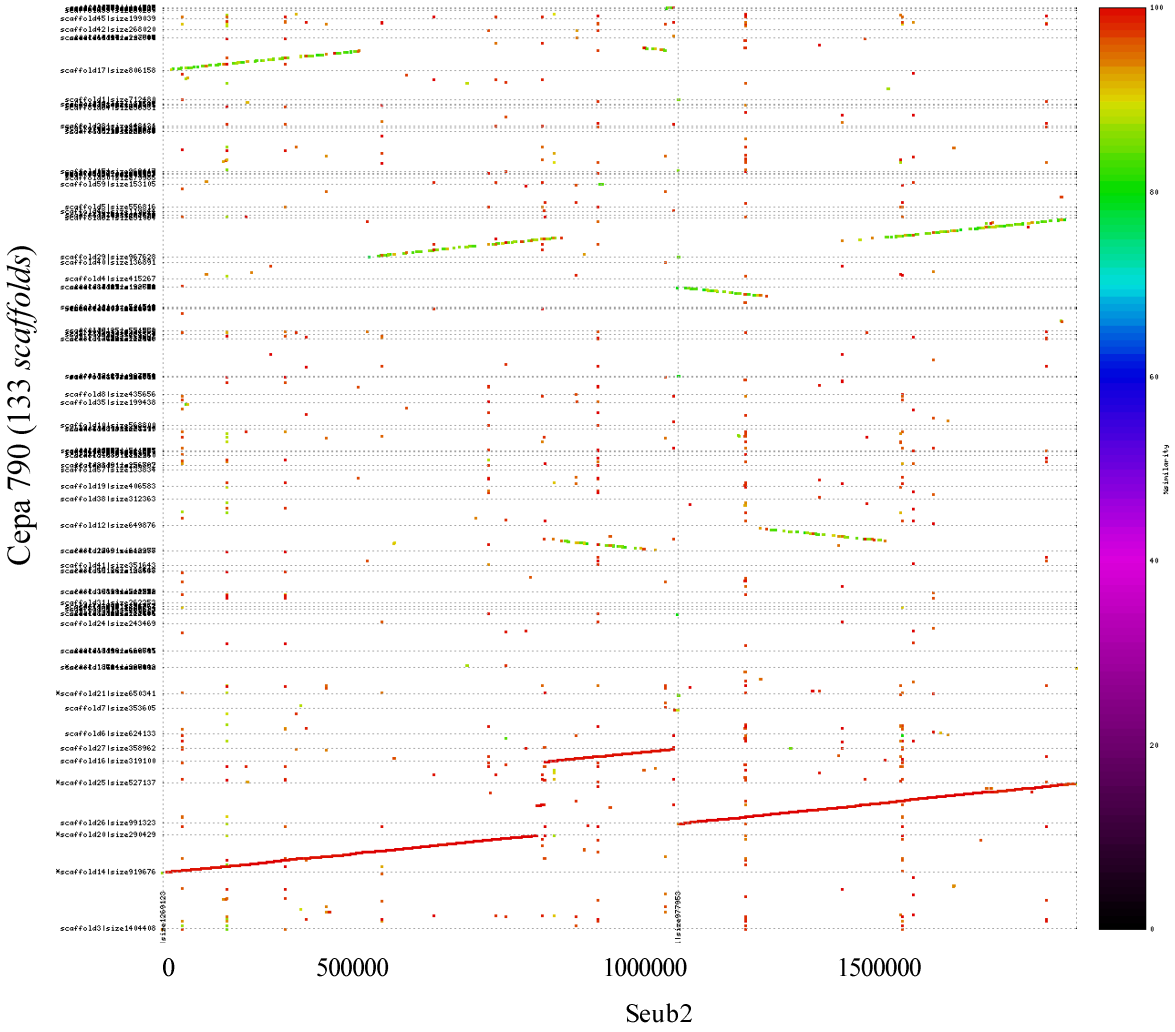
14. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 15 (referencia)



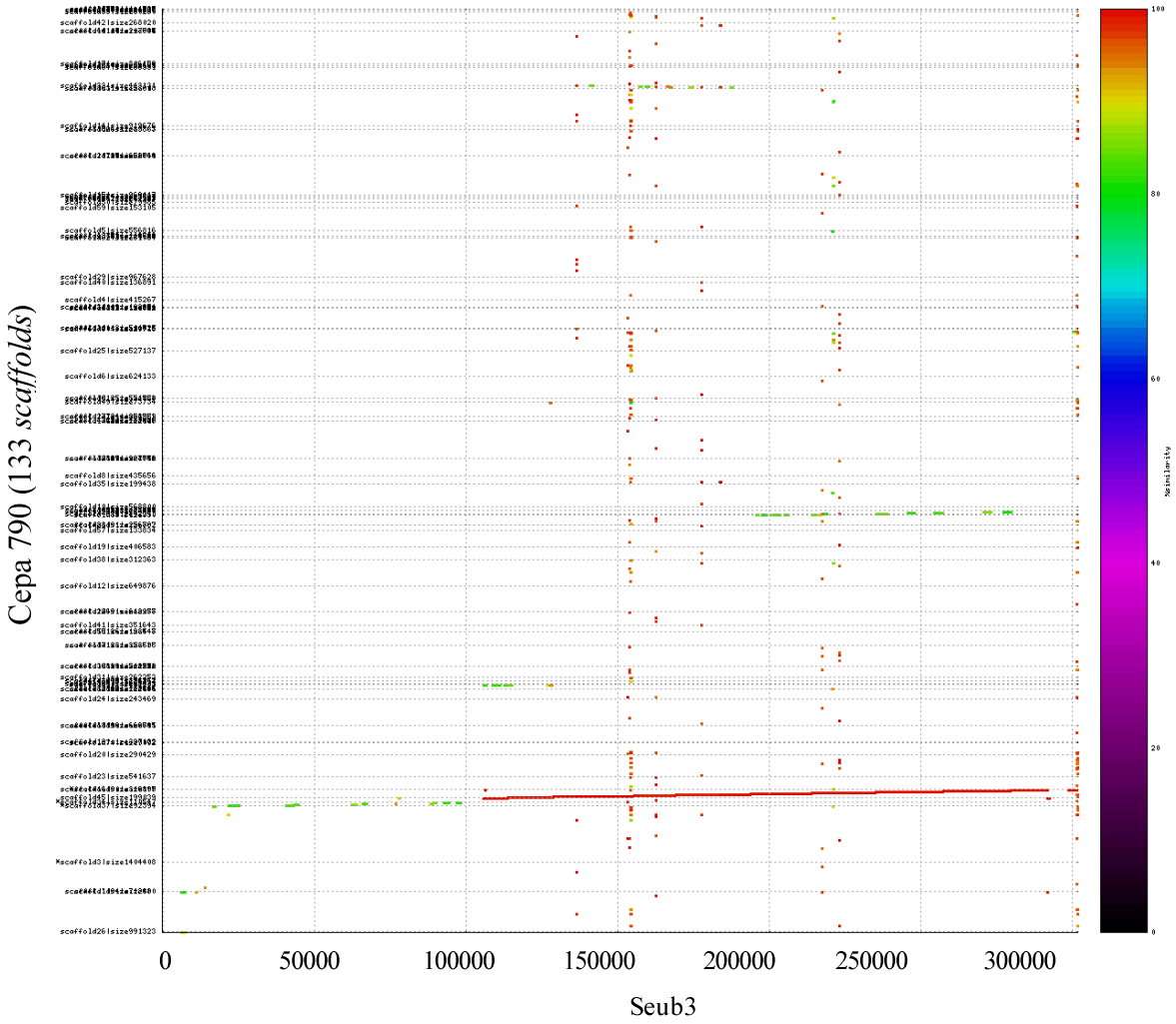
15. 133 *scaffolds* de cepa 790 vs *S. cerevisiae* S288C - Cromosoma 16 (referencia)

Figura 11: Alineamiento de los *scaffolds* de 790 (eje Y) vs *S. cerevisiae* S288C (eje X)

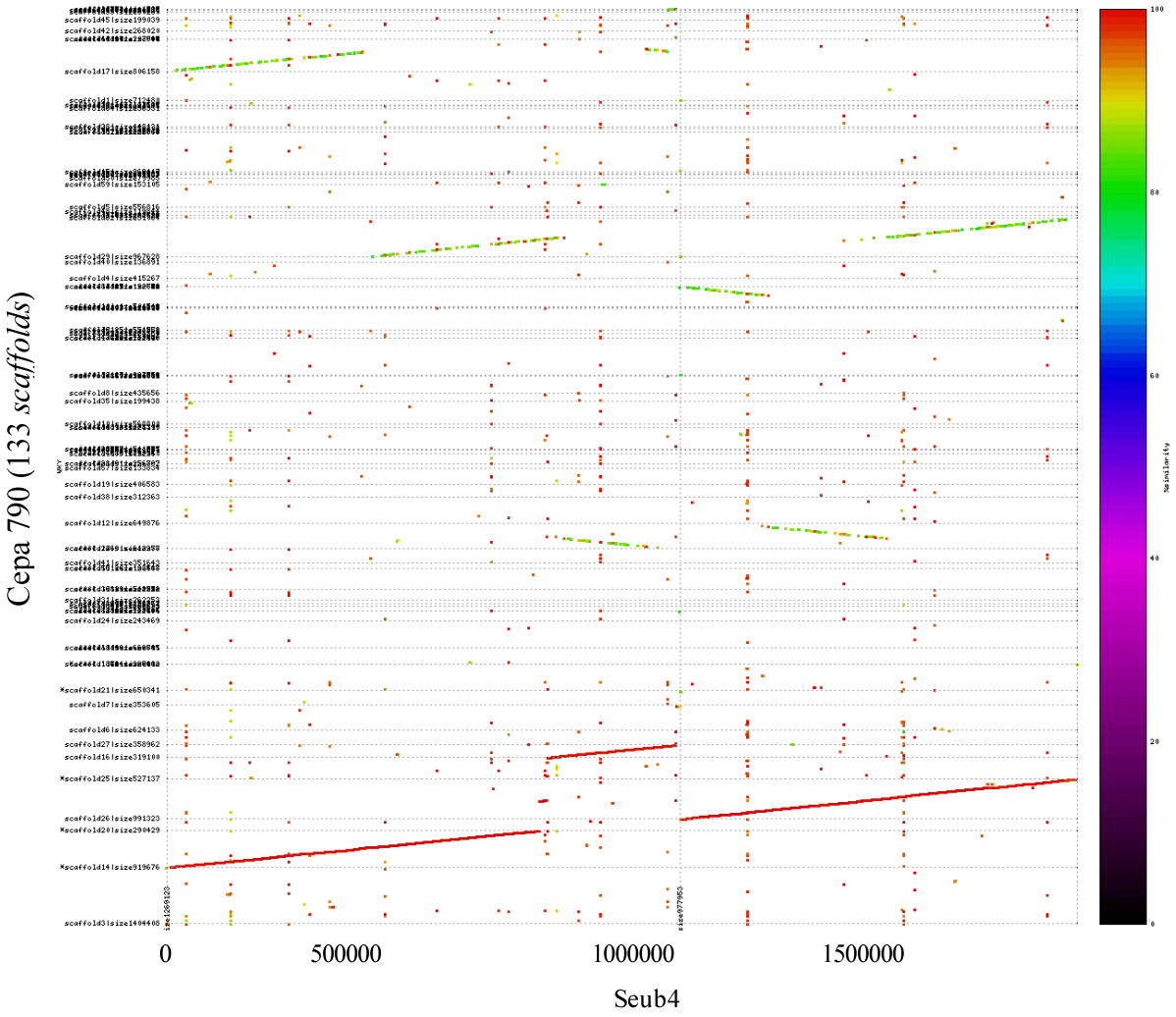
A partir de **7.3.2. 790 vs *S. eubayanus:***



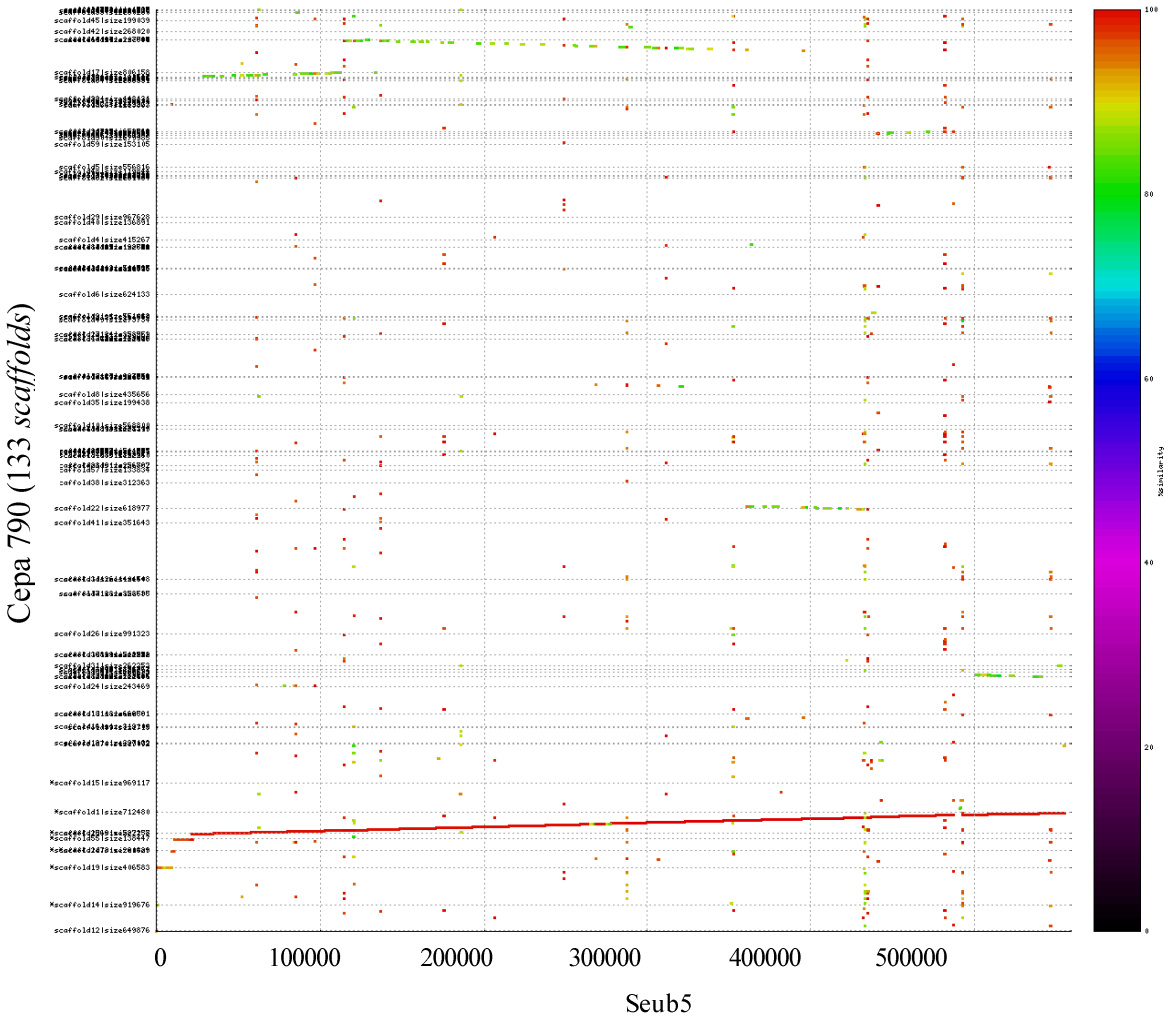
16. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 2 (referencia)



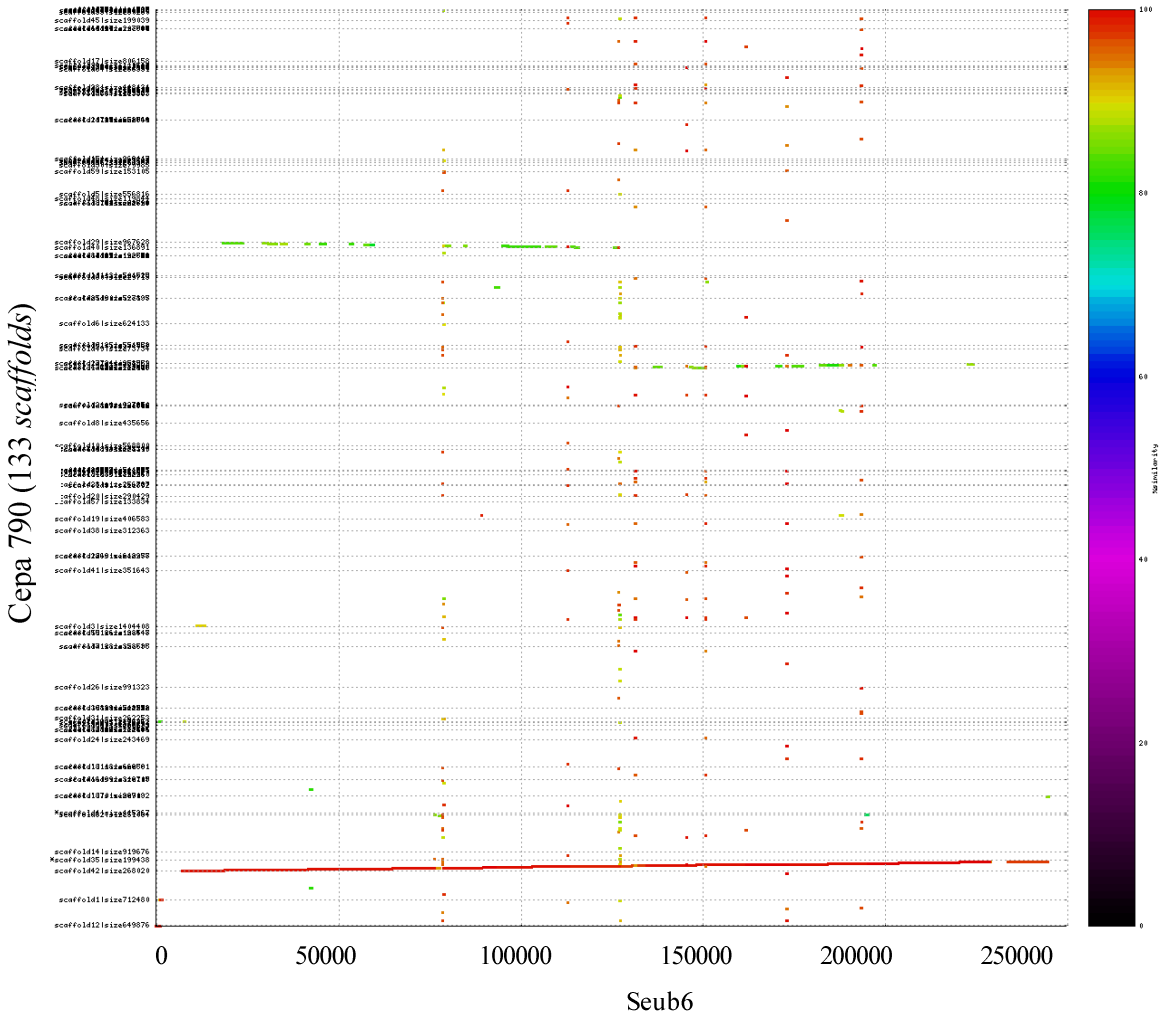
17. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 3 (referencia)



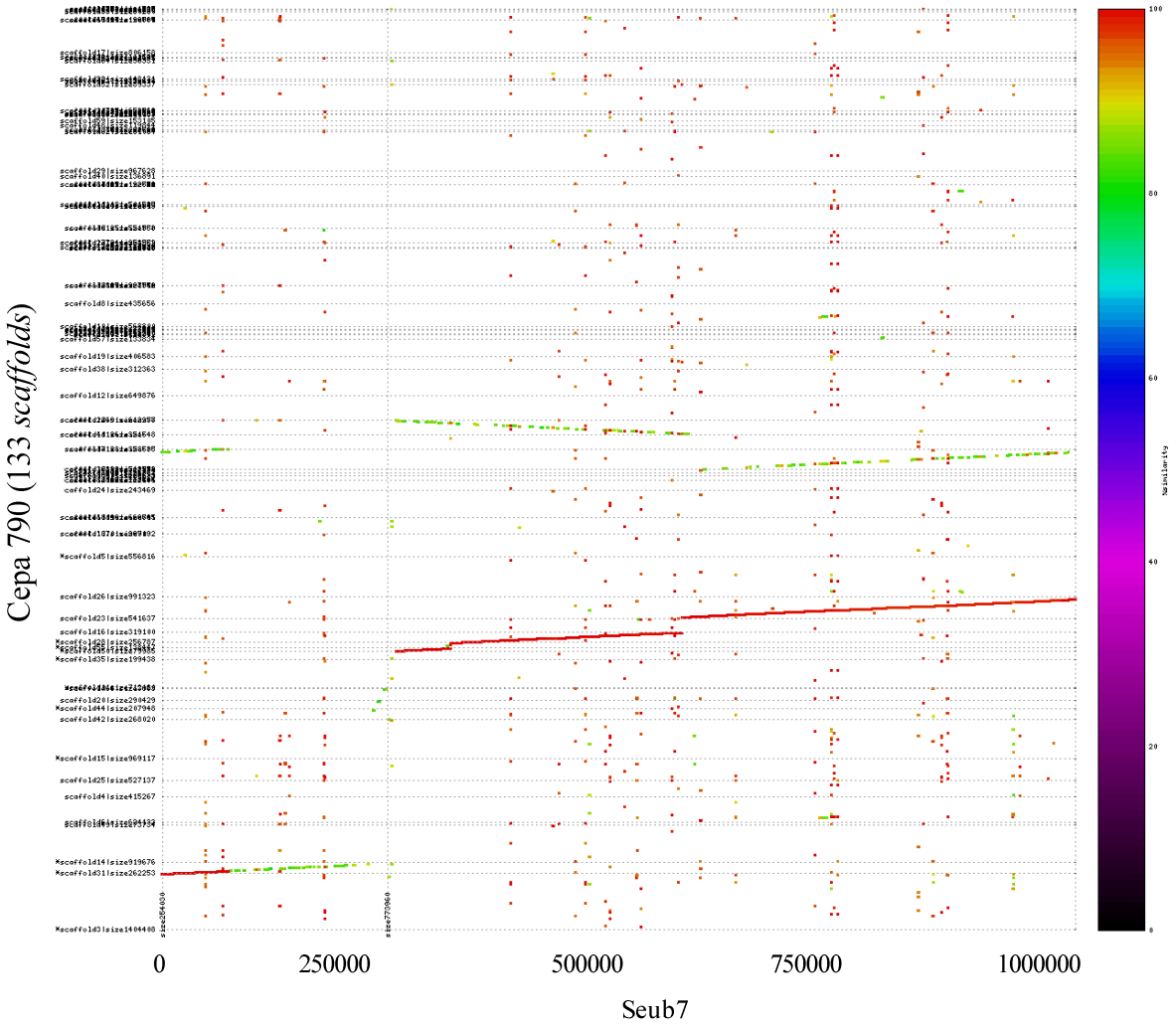
18. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 4 (referencia)



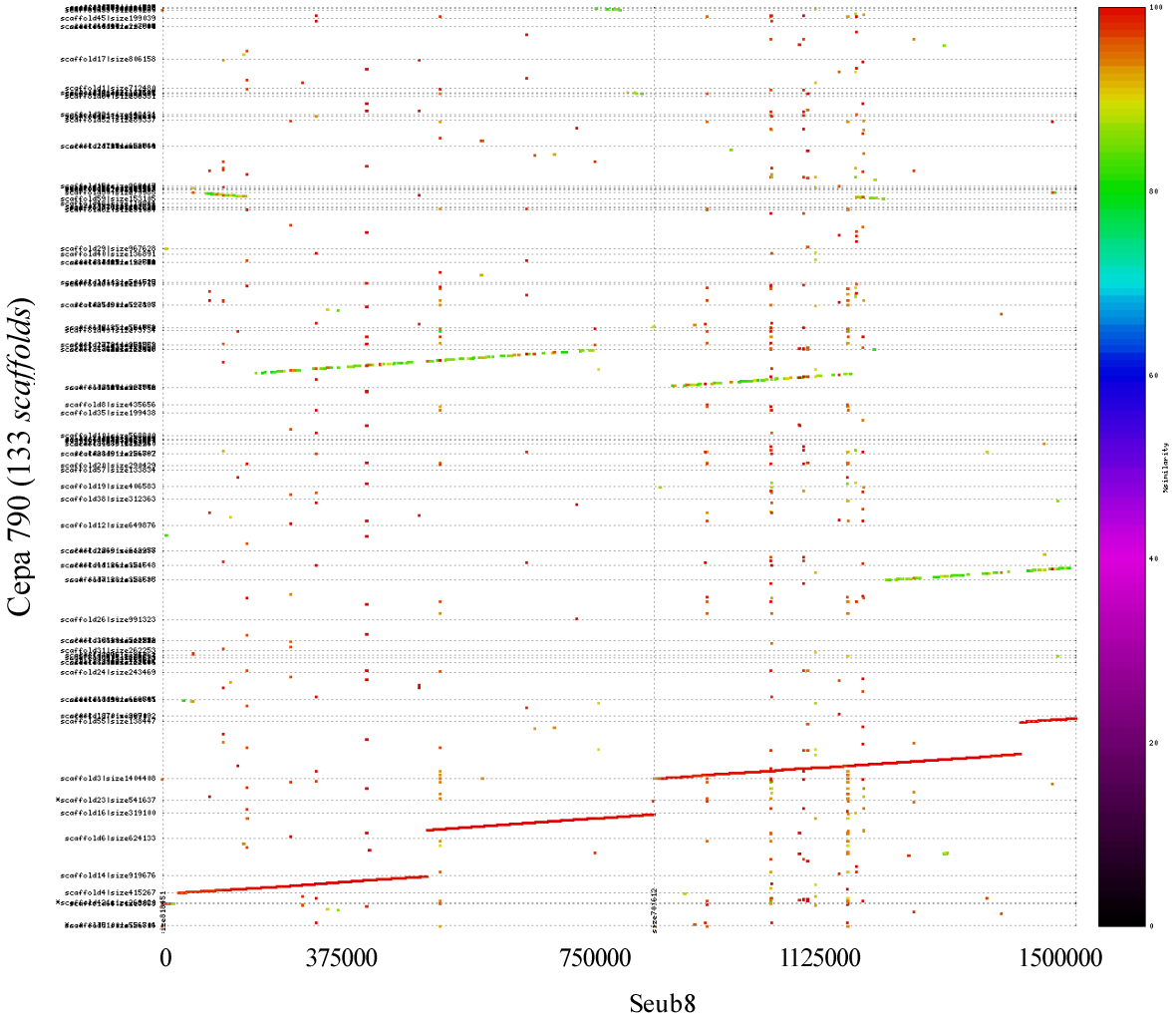
19. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 5 (referencia)



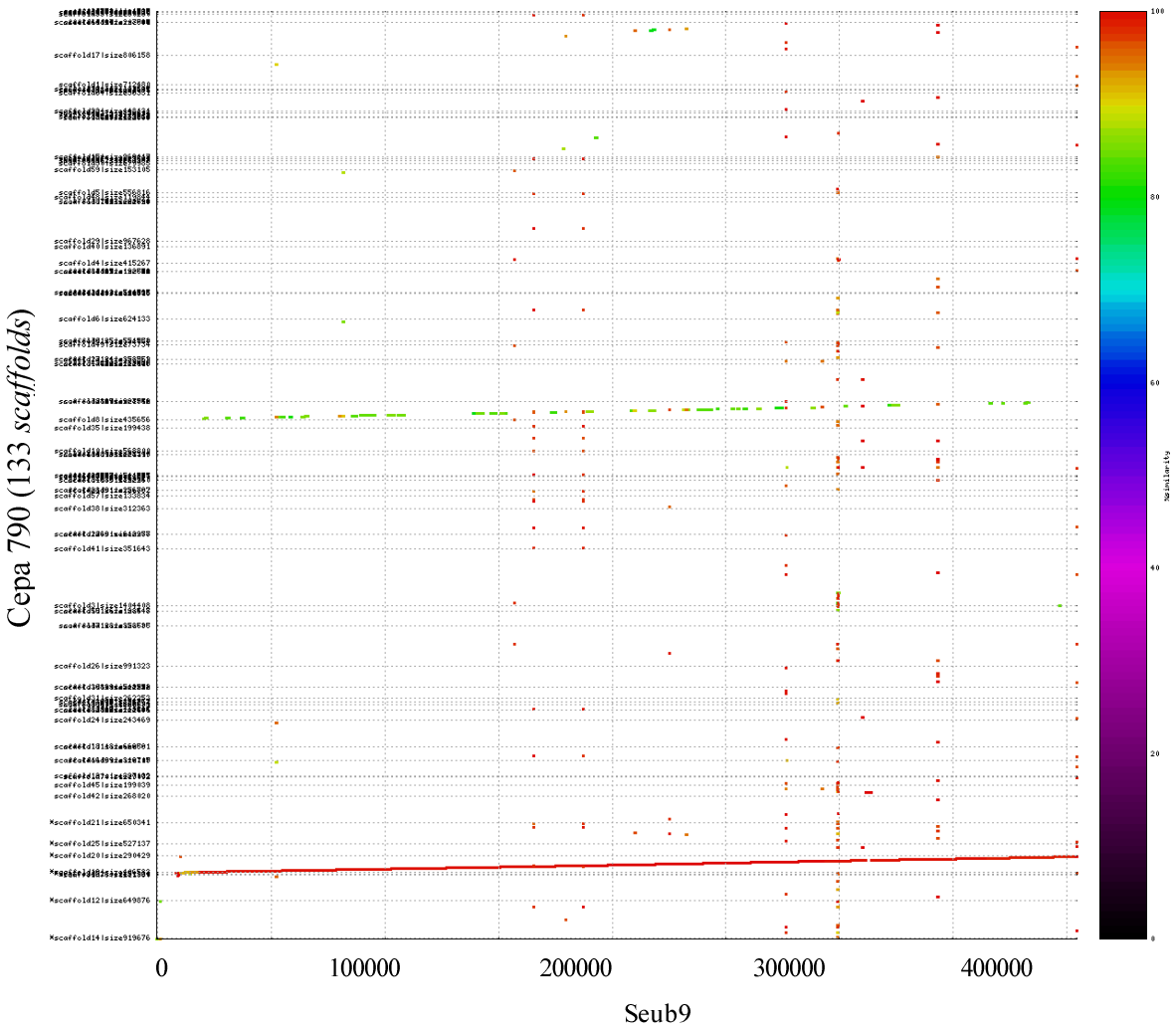
20. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 6 (referencia)



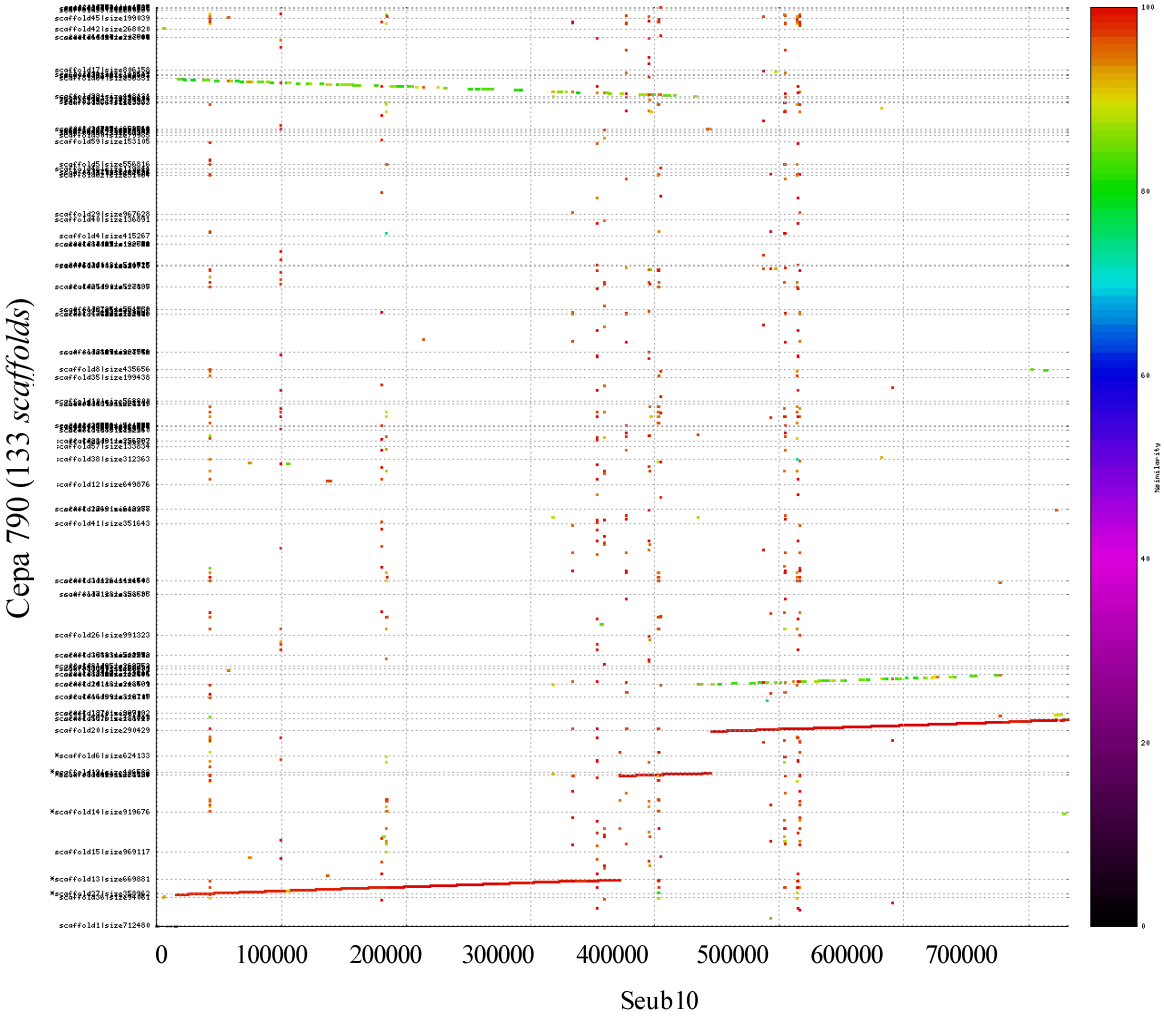
21. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 7 (referencia)



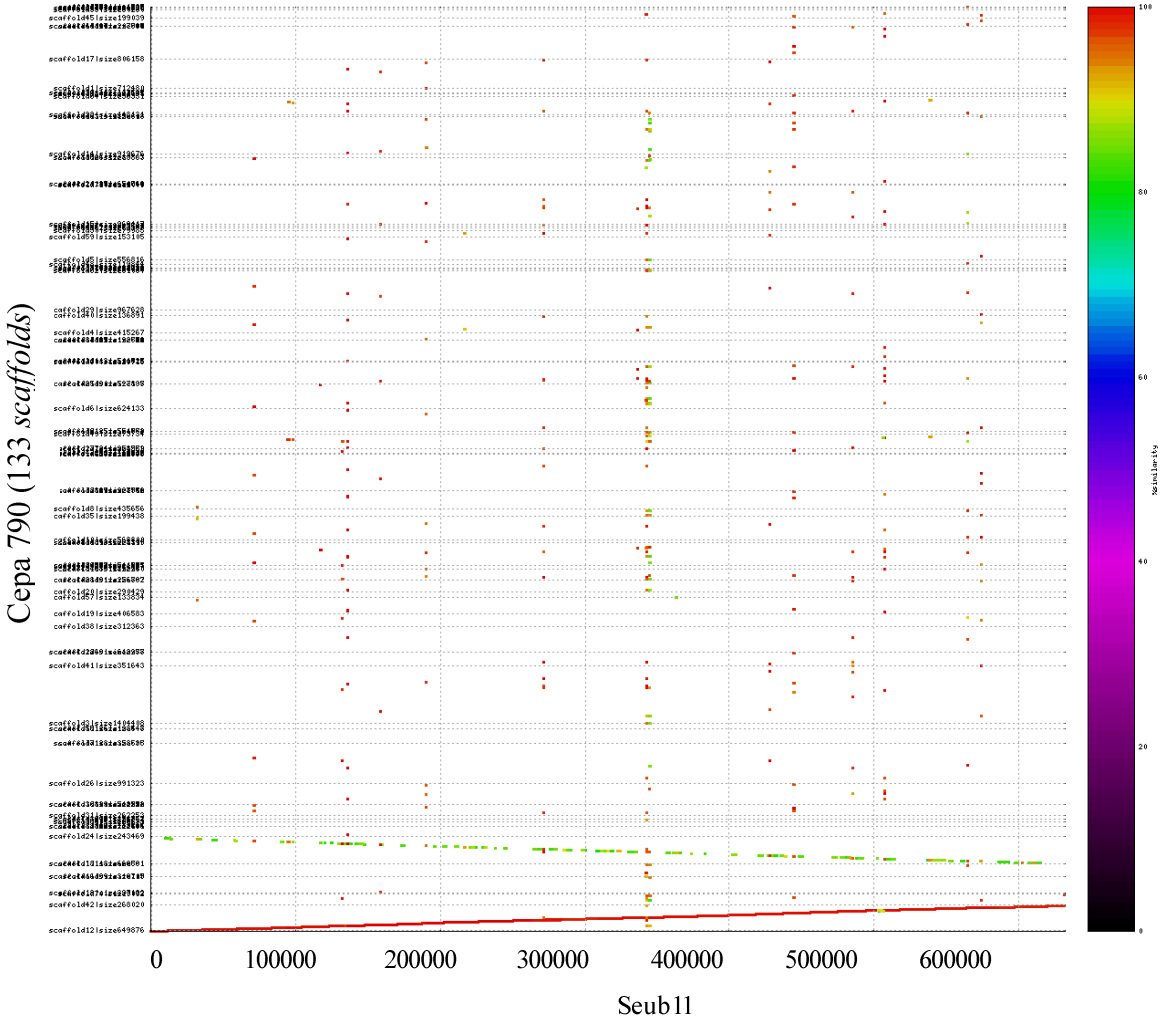
23. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 8 (referencia)



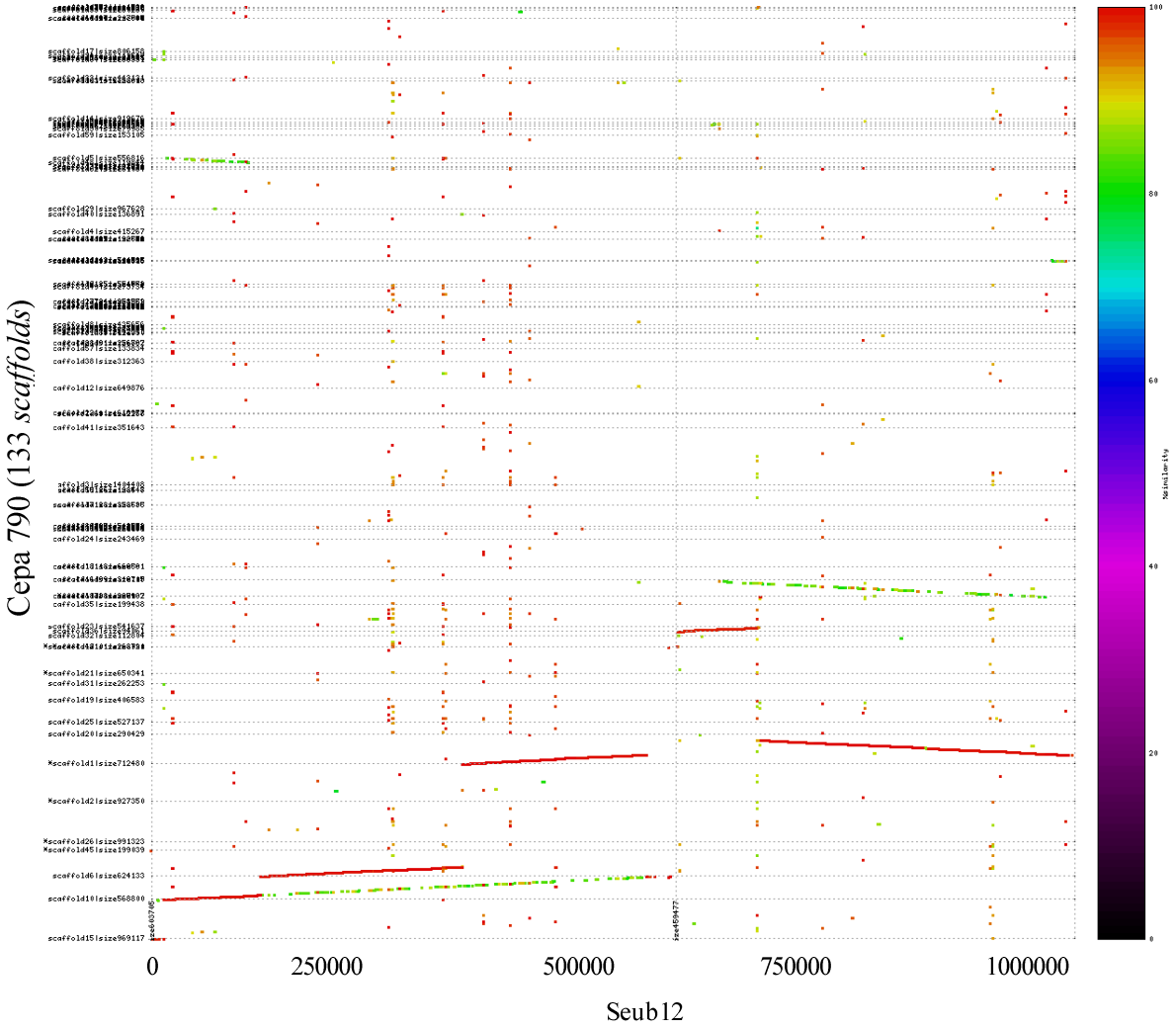
24. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 9 (referencia)



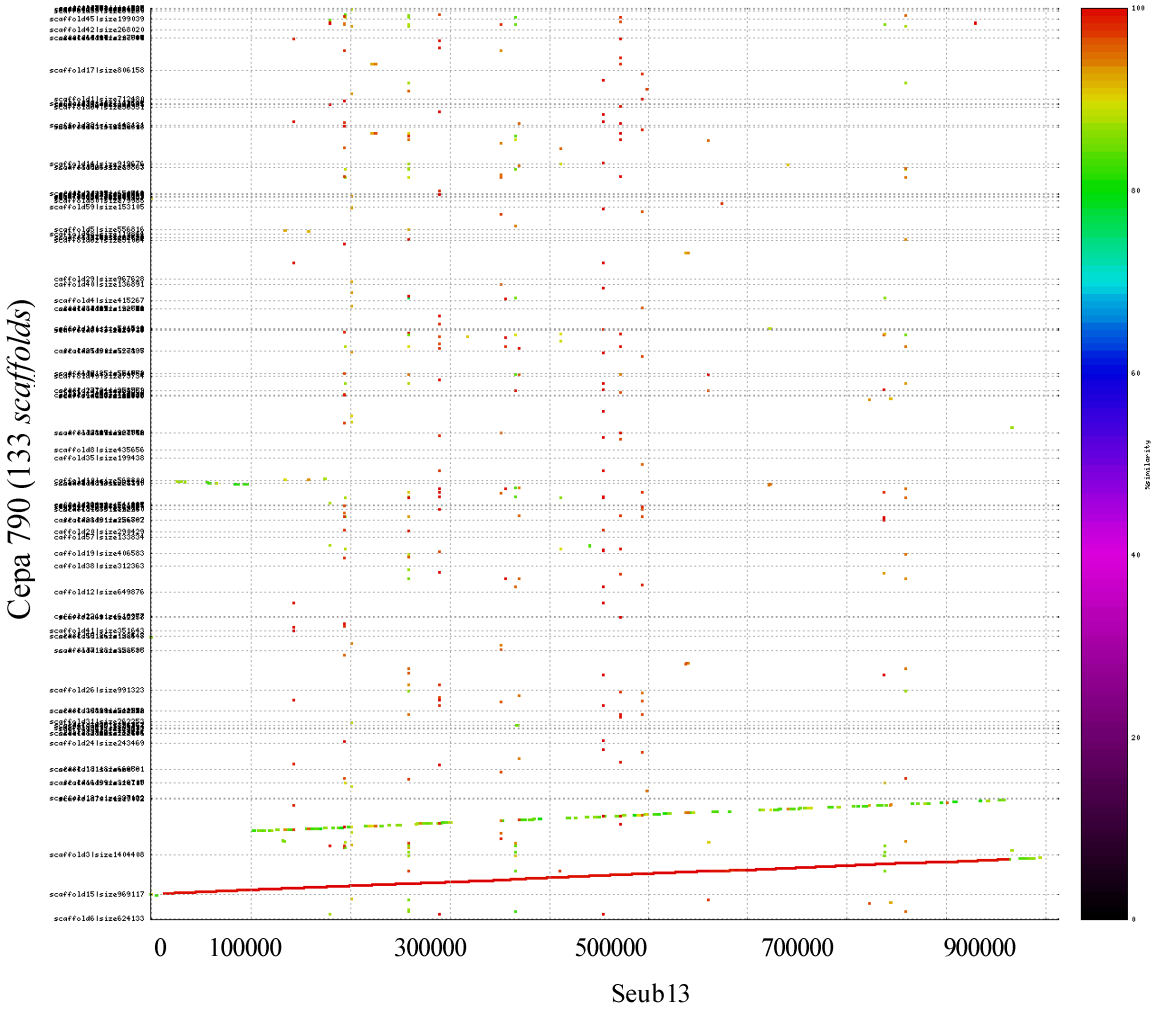
25. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 10 (referencia)



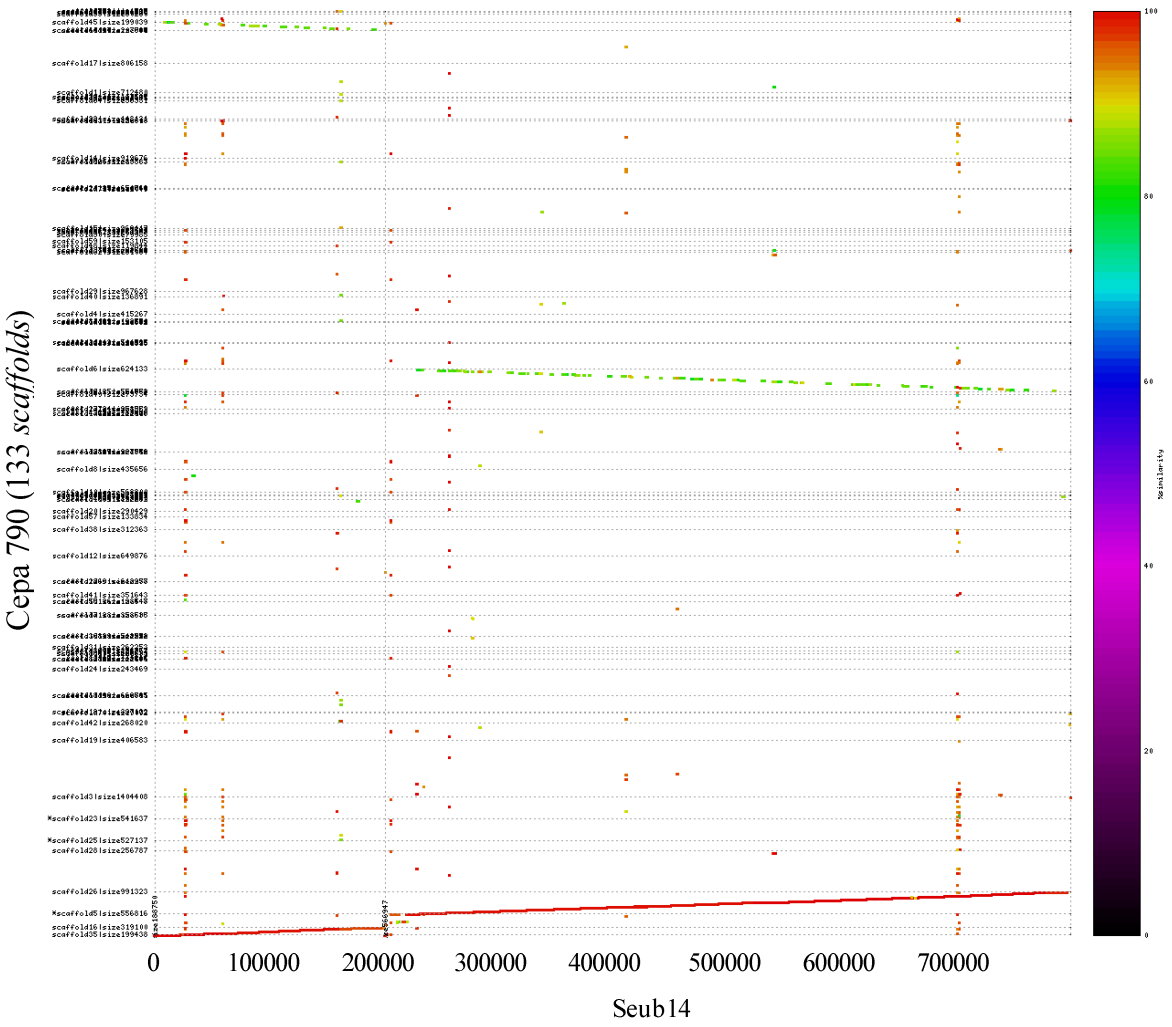
26. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 11 (referencia)



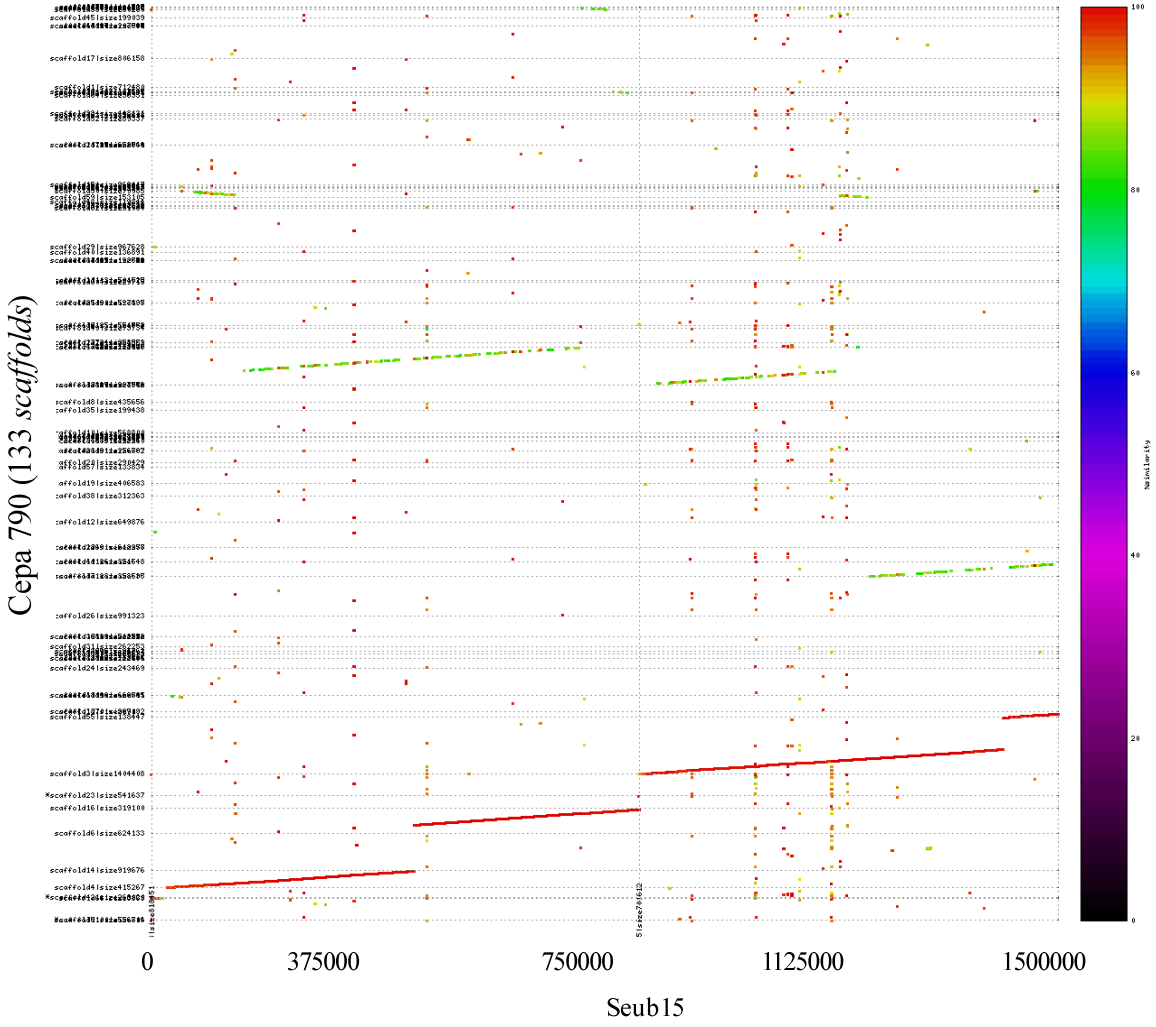
27. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 12 (referencia)



28. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 13 (referencia)



29. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 14 (referencia)



30. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 15 (referencia)

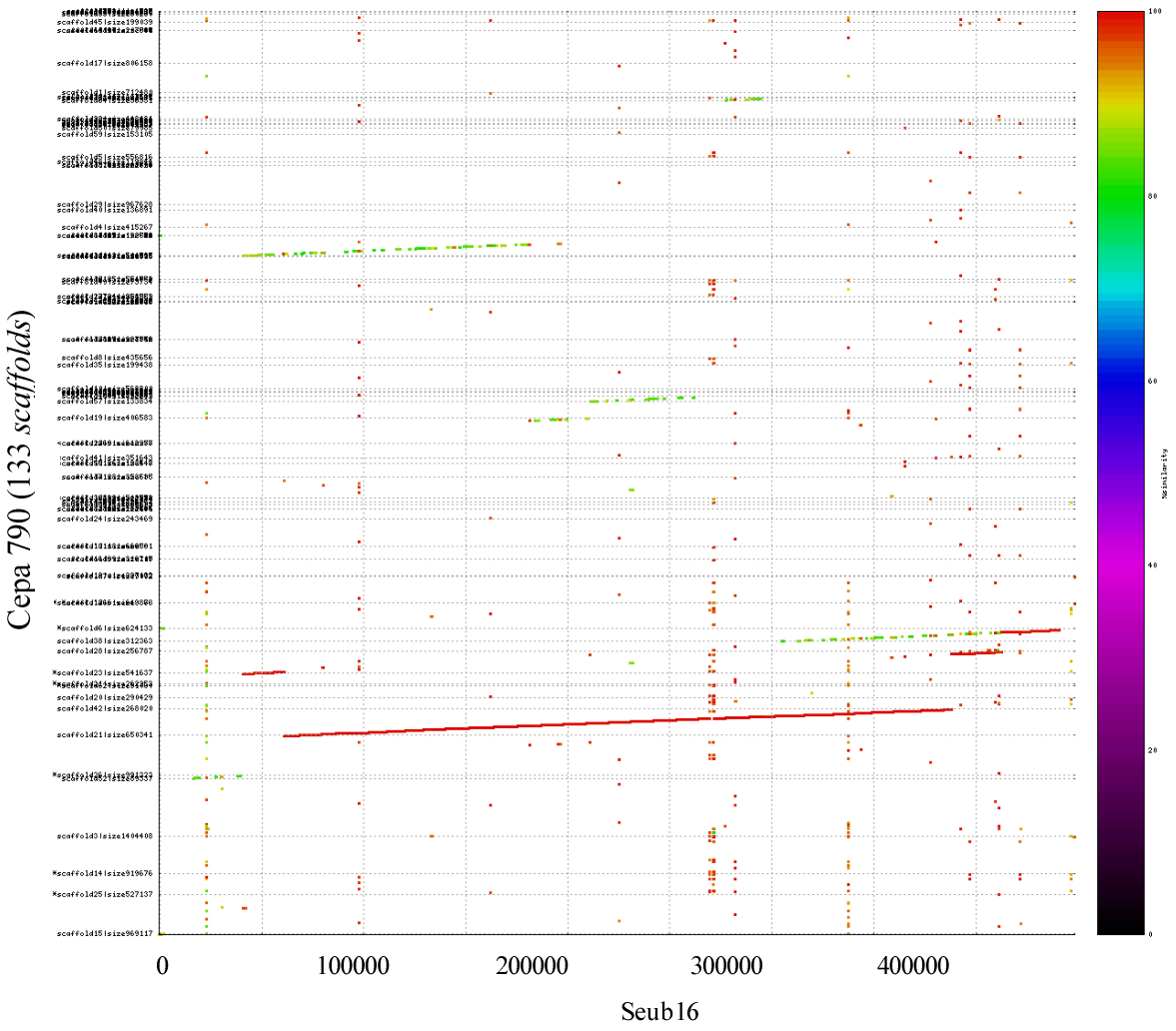


Figura 12: Alineamiento de los *scaffolds* de 790 (eje Y) vs *S. eubayanus* (eje X)

31. 133 *scaffolds* de cepa 790 vs *S. eubayanus* - Cromosoma 16 (referencia)

RESUMEN BIOGRÁFICO

Ramiro Elizondo González

Candidato para el Grado de

Maestro en Ciencias con Acentuación en Microbiología

Tesis: Ensamblaje del genoma de una levadura cervecera tipo lager (*Saccharomyces cerevisiae* var. *uvarum)*

Campo de estudio: Bioinformática

Datos personales: Nacido en Monterrey, N.L. el 19 de agosto de 1986, hijo de Julián Elizondo Molina y Susana Guadalupe González Guzmán

Educación: Egresado de la Universidad Autónoma de Nuevo León. Grado obtenido de Lic. en Biotecnología Genómica en 2009.