UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS

CENTRO DE CIENCIAS GENÓMICAS

Panorama genómico de identidad absoluta: una nueva estrategia para el análisis de genomas naturales y sintéticos

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
KIM PALACIOS FLORES

DIRECTOR DE TESIS

DR. GUILLERMO DÁVILA RAMOS
LABORATORIO INTERNACIONAL DE INVESTIGACIÓN SOBRE EL GENOMA HUMANO

COMITÉ TUTOR

DR. GUILLERMO DÁVILA RAMOS
LABORATORIO INTERNACIONAL DE INVESTIGACIÓN SOBRE EL GENOMA HUMANO

DR. JULIO COLLADO VIDES
CENTRO DE CIENCIAS GENÓMICAS

DR. DANIEL PIÑERO DALMAU
INSTITUTO DE ECOLOGÍA

QUERÉTARO, MÉXICO. ABRIL DE 2018

DEDICATORIA

Este trabajo de investigación está dedicado a mi madre y a mi padre, quienes me han acompañado siempre.

AGRADECIMIENTOS

Kim Palacios Flores es una estudiante de doctorado del Doctorado en Ciencias Biomédicas de la Universidad Nacional Autónoma de México (UNAM) y recibió la beca 392315 del Consejo Nacional de Ciencia y Tecnología (CONACYT).

Agradezco a mi tutor principal, el Dr. Guillermo Dávila; a los miembros de mi Comité Tutor, los Drs. Julio Collado y Daniel Piñero; a los sinodales que revisaron mi tesis, los Drs. Antonio Peña, Esperanza Martínez, Alfredo Varela y Jean Philippe Vielle; a los colegas que me ayudaron con el desarrollo de mi proyecto de investigación, especialmente a Jair García-Sotelo, así como a Alejandra Castillo, Carina Uribe, Luis Aguilar, Lucía Morales, Laura Gómez-Romero, José Reyes, Delfino García, Alejandro Garciarubio y Margareta Boege. Agradezco también al personal administrativo, en particular a Eglee Lomelín y a Abigayl Hernández.

Este trabajo fue realizado en el Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH-UNAM). Para la realización de este trabajo se contó con el apoyo de Luis Aguilar, Alejandro de León, Carlos Flores y Jair García del Laboratorio Nacional de Visualización Científica Avanzada (LAVIS). Los procesos de secuenciación de ADN se llevaron a cabo en el Instituto Nacional de Medicina Genómica (INMEGEN). Este trabajo fue financiado parcialmente por el donativo UNAM-PAPIIT (IA207817). Las lecturas crudas de secuenciación del genoma completo de la cepa synIII (HMSY011) fueron proporcionadas por Leslie Mitchell.

ÍNDICE

RESUMEN

El Proyecto del Genoma Humano reveló la información genética completa de diferentes organismos, desde bacterias hasta humanos, y representó una revolución científica de gran trascendencia: la genómica. En su seno, una nueva revolución se está gestando, aquella que propone el diseño y la síntesis de genomas artificiales. Un aspecto fundamental de la genómica y de la biología sintética es el análisis de genomas al mayor nivel de precisión posible. Con este fin, hemos desarrollado una solución conceptualmente sencilla, sensible, precisa y esencialmente no estadística para el análisis de la variación genómica en organismos haploides. Nuestra estrategia se basa en la generación de un Panorama Genómico de Identidad Absoluta que revela firmas de variación cuando un genoma de interés difiere de un genoma de referencia. Dichas firmas codifican la ubicación precisa de diferentes tipos de variantes, introduciendo el concepto de una firma general de variación. La naturaleza de las variantes se define subsecuentemente a través de la generación de alineamientos dirigidos. De esta forma, la lógica de búsqueda de identidad absoluta entre genomas provee una plataforma unificadora para el análisis de la variación genómica. Hemos generado los perfiles de variación de genomas de levadura haploides y de siete cromosomas sintéticos del proyecto internacional Sc2.0, cuyo objetivo es diseñar y sintetizar el primer genoma eucarionte, el de la levadura *S. cerevisiae*. Nuestro enfoque ha descubierto variantes que no han sido detectadas previamente y ha permitido el refinamiento de genomas de referencia de alta calidad. Hemos propuesto una expansión de los conceptos básicos de nuestro algoritmo para abarcar el análisis de organismos diploides y de genomas más complejos.

ABSTRACT

The Human Genome Project revealed the complete genetic information of different organisms, from bacteria to human, and represented a most powerful revolution: the science of genomics. At its heart, a new revolution is awakening, that of the design and synthesis of artificial genomes. Central to genomics and synthetic biology, is the analysis of genomes at the highest level of precision. To this end, we have developed a conceptually simple, sensitive, precise, and essentially non-statistical solution for the analysis of genome variation in haploid organisms. Our strategy is based on the generation of a Perfect Match Genomic Landscape, which reveals signatures of variation wherever a query genome differs from a reference genome. Such signatures encode the precise location of different types of variants, effectively introducing the concept of a general signature of variation. The precise nature of variants is then resolved through the generation of targeted alignments. Thus, the perfect match logic provides a unified framework for the detection of genome variation. We have determined the variation profiles of natural haploid yeast genomes and of seven synthetic chromosomes from the Sc2.0 international project, which aims at constructing the first eukaryotic genome by design and synthesis, that of *S. cerevisiae*. Our approach has uncovered variants that have previously escaped detection. Moreover, our strategy is ideally suited for further refining high-quality reference genomes. On a theoretical level, we propose an expansion of our algorithm to analyze diploid organisms and more complex genomes.

INTRODUCCIÓN

La Revolución Genómica: Genomas de Referencia y la Importancia de los Perfiles de Variación Genómica

El conocer la secuencia nucleotídica de un genoma, de forma completa y precisa es de suma importancia. Prueba de ello es el legado que ha dejado la culminación del proyecto más ambicioso de las ciencias biológicas: descifrar la secuencia del genoma humano (International Human Genome Sequencing Consortium, 2004). Su impacto, tanto en el desarrollo científico como tecnológico, es hoy en día, irrefutable. Sin embargo, fervientes debates sobre la verdadera utilidad de esta información, cuya obtención representaría un altísimo costo y un enorme esfuerzo de carácter internacional, acompañaron al Proyecto del Genoma Humano (Human Genome Project, HGP por sus siglas en inglés) desde sus inicios. Muchos se preguntaban si el producto final del HGP cumpliría la promesa de revelar algunos de los misterios más interesantes de la evolución de nuestra especie, por una parte, y de iluminar caminos hacia la cura de enfermedades, por otra. Afortunadamente, la ciencia es imparable, y el HGP siguió su curso. Sin duda, los frutos de este colosal proyecto han abierto un mundo de oportunidades para la investigación de los organismos desde diversos frentes: estructural, funcional y evolutivo. Se trata de la biología molecular elevada a un nivel holístico: la genómica.

Un aspecto central del HGP fue su planteamiento y ejecución en un contexto amplio, que incluyó la elaboración de genomas de referencia de organismos modelo: bacterias (por ejemplo *H. flu* y *E. coli*), la levadura *S. cerevisiae*, el nemátodo *C. elegans*, la mosca *D. melanogaster*, la planta *A. thaliana*, y el ratón. La incorporación de organismos más sencillos en el HGP permitió: 1) desarrollar y probar nuevas metodologías de manera efectiva, lo cual sería clave para comprender su alcance en un contexto más intrincado como el del genoma humano, 2) descifrar la información genética detallada de organismos cuyo estudio ha revelado varios de los procesos más elementales de la biología y 3) analizar el genoma humano desde una perspectiva evolutiva. De hecho, el análisis inicial del genoma humano reveló aspectos tan fascinantes y diversos como la distribución no homogénea de una serie de elementos y procesos, incluyendo genes, transposones y tasa de recombinación; la disminución drástica en el estimado del número de genes codificantes para proteína y la concomitante apreciación de mecanismos como el splicing alternativo en la generación de complejidad; la generación de un catálogo extendido de arquitecturas proteicas a través de la reutilización de motivos preexistentes; la frecuencia de rearreglos estructurales grandes y recientes, como las duplicaciones segmentarias, en ciertos compartimentos del genoma; e indicios de selección positiva en elementos Alu, típicamente considerados como ADN egoísta (“selfish DNA”) (International Human Genome Sequencing Consortium, 2001).

Dado que el ensamble de estos genomas pioneros no puede partir de genomas de referencia preexistentes, se le conoce como ensamble *de novo*. Las metodologías para el ensamble *de novo* de genomas de referencia han evolucionado drásticamente, tanto en sus componentes experimentales como computacionales. Progresivamente, genomas más y más complejos han sido ensamblados *de novo* de manera más rápida y precisa y con costos que han ido disminuyendo. Existe, sin embargo, otra manera de conocer la secuencia completa de

organismos para los cuales existe ya un genoma de referencia: se les llama proyectos de re-secuenciación. El principio es claro: la comparación *in silico* de un genoma secuenciado contra un genoma de referencia genera un perfil de variación genómica entre ambos. Dicho perfil de variación genómica es la descripción de la secuencia de ADN del genoma de interés. El proyecto del genoma humano y de sus organismos modelo dotó entonces a la comunidad científica con 1) las tecnologías y los conocimientos necesarios para ensamblar genomas *de novo* y 2) con una serie de genomas de referencia de alta calidad que, a su vez, pueden guiar la elaboración de nuevos genomas de referencia de las especies correspondientes. Se trata de un círculo virtuoso que continúa revolucionando y evolucionando hoy en día, abarcando cada vez más instancias de los productos de la evolución.

Los experimentos de re-secuenciación son la base de un sinfín de proyectos genómicos. La importancia de los perfiles de variación resultantes es evidente en diversos contextos. Un caso particularmente elegante es el icónico experimento de evolución a largo plazo del Dr. Richard Lenski (Pennisi, 2013). En su laboratorio, 12 poblaciones de *E. coli*, inicialmente idénticas, han evolucionado en un medio rico constante limitado en glucosa durante más de 25 años (Tenaillon *et al.*, 2016). Aproximadamente cada 500 generaciones, muestras de las distintas poblaciones han sido congeladas, generando un tipo de "registro fósil" sin precedentes y la posibilidad de "reiniciar" caminos evolutivos tomando como punto de partida diferentes momentos en la historia del experimento. Al comparar lecturas de secuenciación de dichas muestras con el genoma de la población ancestral, se han obtenido perfiles de variación genómica que describen el proceso evolutivo. Esta descripción se da en torno a dos ejes: 1) distintos números de generaciones transcurridas y 2) distintas poblaciones evolucionando en paralelo. El experimento en su conjunto y el análisis de los perfiles de variación genómica correspondientes han recapitulado y evidenciado propiedades generales sobre el proceso evolutivo. Éstas incluyen la acumulación de mutaciones que eventualmente permiten la aparición de una innovación clave y el surgimiento y la divergencia de linajes con propiedades diferentes (Blount *et al*., 2012). Otros hallazgos han sido sumamente inesperados, en particular, la evolución observada parece poseer un componente importante de reproducibilidad y, más espectacular aún, no parece detenerse, incluso bajo condiciones constantes (Lenski *et al*., 2015).

Sin duda, el experimento de re-secuenciación más ambicioso será el que predijo Sydney Brenner en 2002, en su discurso Nobel titulado "Nature's Gift to Science" (Brenner, 2002). Al pronunciar las siguientes palabras, Brenner propone a la humanidad como el gran experimento de evolución aún por ser estudiado: "However, suppose technology existed which made it easy to characterise 30,000 genomes, perhaps even to the point of resequencing them, would we bother to do this work with mice? We could go directly to humans […]". El término "predijo" es adecuado. En efecto, diferentes tipos de acercamientos a un experimento de esta magnitud constituyen, hoy en día, la tarea cotidiana de la genómica (1000 Genomes Project Consortium, 2012). Con la cura de enfermedades como meta primordial, se analizan los diferentes perfiles de variación genómica teniendo como referencia la clasificación fenotípica de los individuos correspondientes. A lo largo de un primer eje de comparación, la variación ya existente en los "genes de la humanidad", como lo califica Brenner, está permitiendo conocer algunas de las rutas más comunes hacia ciertas enfermedades. En el caso del estudio del cáncer, que finalmente constituye un proceso de evolución dentro de un organismo, existe un segundo eje de comparación muy interesante. Se trata de contrastar tejido canceroso en distintas etapas de su

evolución con el tejido normal, ambos provenientes del mismo individuo. Este tipo de análisis forma parte de lo que se conoce como medicina personalizada: la búsqueda de estrategias terapéuticas *ad hoc* para cada individuo. El reto consiste entonces en obtener un conocimiento profundo de la biología de la persona. Nuevamente, el paso crítico es determinar su perfil de variación genómica.

Un reto continuo para la ciencia genómica es alcanzar la mayor precisión posible al descifrar la secuencia nucleotídica de genomas individuales, desde microorganismos hasta humanos. De hecho, un sólo nucleótido puede, en ocasiones, llegar a tener repercusiones importantes a nivel estructural, funcional o evolutivo. La presente tesis incluye el desarrollo de una estrategia novedosa para la generación de perfiles de variación genómica de organismos haploides, utilizando como modelo experimental la levadura *S. cerevisiae*. Además de tratarse del organismo modelo eucarionte por excelencia, *S. cerevisiae* ofrece todas las ventajas experimentales de un organismo unicelular. Más aún, el genoma de referencia de *S. cerevisiae* utilizado en este trabajo, el de la cepa S288c, ha sido perfeccionado durante los últimos veinte años (Goffeau *et al.*, 1996; Engel *et al.*, 2014), proporcionando una plataforma de muy alta calidad para la generación de perfiles de variación genómica. Finalmente, es utilizando la levadura *S. cerevisiae* que la ciencia ha dado un gran paso hacia una nueva forma de comprender la biología de organismos eucariontes: la síntesis de genomas completos (ver la sección titulada **Una Nueva Revolución Biológica: Diseño y Creación de Genomas Sintéticos**). Además de representar un avance científico *per* se, la síntesis en curso de un genoma diseñado de *S. cerevisiae*, con propiedades sumamente interesantes, constituye un experimento piloto para la continuación del HGP: la síntesis del genoma humano o HGP-write. La estrategia de análisis desarrollada en este trabajo de investigación ha hecho contribuciones en tres frentes de la genómica: 1) refinamiento de genomas de referencia, 2) generación de perfiles de variación de genomas re-secuenciados, y 3) refinamiento y generación de perfiles de variación de cromosomas sintéticos.

Panorama Genómico de Identidad Absoluta: Una Estrategia Original para la Identificación Precisa de Variación Genómica

La variación genómica reportada por la mayoría de los algoritmos actuales debe ser considerada desde un punto de vista estadístico, es decir, asociada a una cierta probabilidad de ser verdadera o falsa. Esta incertidumbre se debe al mapeo global de las lecturas de secuenciación provenientes del genoma de interés hacia el genoma de referencia para inferir su localización. Bajo este paradigma, la presencia de errores experimentales y de variación verdadera entre el genoma de interés y el genoma de referencia obliga a utilizar correspondencias ("matches") aproximadas para la búsqueda de diferencias entre ambas secuencias (Reinert *et al.*, 2015). En efecto, la mayoría de las lecturas de secuenciación no son idénticas al genoma de referencia en sus posiciones respectivas; más aún, aquellas que contienen la variación que se busca identificar, por definición, nunca lo son. Como consecuencia, la misma lectura de secuenciación podría ser alineada con una variedad de sitios del genoma de referencia, y cada alineamiento representaría entonces una descripción diferente de la variación. Un análisis comparativo entre las distintas posibilidades, que se basa en criterios estadísticos correspondientes a cada alineamiento y al conjunto de alineamientos, determina cuál es el

alineamiento óptimo, es decir, con mayor probabilidad de ser cierto. Al igual que el proceso de descubrimiento de variantes, el perfil de variación resultante es, por naturaleza, probabilístico.

Al permitir una valoración cualitativa de la variación descubierta, la estrategia de análisis desarrollada en este trabajo de investigación representa un cambio de perspectiva importante para el campo. Este cambio ha sido posible implementando una búsqueda exclusiva de correspondencias perfectas entre secuencias predeterminadas y ordenadas provenientes del genoma de referencia y secuencias provenientes del genoma de interés. Esto puede resultar intrigante, incluso paradójico. Surge la pregunta: ¿Cómo obtener información precisa sobre variación genómica utilizando únicamente segmentos idénticos entre los genomas comparados? La clave consiste en interpretar la ausencia de identidad absoluta como indicador directo de la posición precisa de la variación. La construcción de un Panorama Genómico de Identidad Absoluta (Perfect Match Genomic Landscape, PMGL por sus siglas en inglés) entre ambos genomas permite entonces identificar todos aquellos sitios afectados por algún tipo de variación. Una vez conocidos los lugares que albergan la variación, descubrir la naturaleza específica de la variación subyacente se convierte en un problema local. En otras palabras, el PMGL representa una estrategia de análisis lineal de la variación genómica, en donde las posiciones de las variantes son conocidas con precisión de un nucleótido antes de realizar alineamientos dirigidos entre el genoma de referencia y el genoma de interés.

Finalmente, dada la naturaleza cualitativa de las variantes obtenidas, es posible introducir cada variante descubierta en el genoma de referencia, generando así un genoma de referencia *ad hoc*. La construcción de un nuevo PMGL, utilizando el genoma de referencia *ad hoc* y el genoma de interés original, produce un Panorama Genómico de Identidad Absoluta uniforme cuando la variación ha sido correctamente definida. Esto implica que la estrategia PMGL de búsqueda de variación constituye su propio método de validación.

La estrategia de análisis de variación genómica por PMGL se encuentra descrita ampliamente en el artículo publicado. La figura 1 del artículo presenta un esquema general de la estrategia PMGL. La ruta computacional correspondiente ha sido completamente automatizada, y consta de 6 módulos: 1) Generación de un panorama genómico de referencia (Reference Genome Self Landscape, RGSL por sus siglas en inglés), 2) Generación de un PMGL, 3) Escaneo del PMGL, 4) Generación de un alineamiento inicial para cada firma de variación, 5) Interpretación y extensión de alineamientos, y 6) Generación de un genoma de referencia *ad hoc*. El código fuente de la ruta computacional PMGL completa ha sido depositado en el repositorio público GitHub (https://github.com/LIIGH-UNAM/PerfectMatchGenomicLandscapePipeline.git). Cada módulo se describe brevemente a continuación.

1) Generación de un RGSL

   Un RGSL es una descripción detallada de la estructura del genoma de referencia utilizado, ya sea para generar perfiles de variación de otros genomas o para su propio refinamiento. Para construir un RGSL se obtienen todas las secuencias ordenadas de tamaño k (25 nucleótidos en este estudio) que componen al genoma de referencia. Todas las secuencias adyacentes se encuentran desfasadas por un sólo nucleótido, por lo que representan una descomposición exhaustiva y ordenada del genoma de referencia. Cada

posición del genoma de referencia se encuentra entonces asociada a una secuencia de tamaño k, que puede ser única o repetida, y a un identificador único que indica la posición y el cromosoma en el que se encuentra. Para capturar la estructura del genoma de referencia en el RGSL, se busca cada secuencia en el propio genoma de referencia, y se reporta su número de ocurrencias exactas. También se reportan los identificadores únicos de todas las posiciones asociadas a la misma secuencia, generando familias de identificadores únicos si es que se trata de una secuencia repetida. La estructura de datos del RGSL es una tabla con tantos renglones como secuencias de tamaño k están presentes en el genoma de referencia, y cuatro columnas: identificador único (ID), cuenta en la referencia (count reference, CR por sus siglas en inglés), secuencia (SEQ) y familia de identificadores únicos (IDF).

2) Generación de un PMGL

Un PMGL describe la identidad absoluta que existe entre cada secuencia presente en la estructura RGSL y secuencias derivadas de las lecturas de secuenciación de un genoma de interés. Para construir un PMGL, es necesario descomponer cada lectura de secuenciación en el conjunto de secuencias de tamaño k que la constituyen. Las secuencias adyacentes derivadas de una misma lectura se encuentran desfasadas por un sólo nucleótido. Subsecuentemente, se obtiene el número de ocurrencias exactas de cada secuencia en el conjunto total de secuencias derivadas de las lecturas de secuenciación. El PMGL utiliza la estructura del RGSL para reportar el número de correspondencias perfectas entre cada secuencia del genoma de referencia y todas las secuencias del genoma de interés. Es posible obtener dicho panorama de identidad absoluta entre ambos genomas al reportar, para cada secuencia del genoma de referencia, el conteo de la misma secuencia en el genoma de interés. La estructura PMGL representa entonces una extensión del RGSL. En este contexto, la cobertura de correspondencias perfectas (perfect matches, PM por sus siglas en inglés) asociada a cada secuencia del genoma de referencia es normalizada por su nivel de repetitividad (perfect matches normalized to count reference, PMnCR por sus siglas en ingles). Los cambios en cobertura de identidad absoluta, ocasionados de manera abrupta por la presencia de variación, pueden cuantificarse obteniendo la relación (división) de PMnCRs entre nucleótidos adyacentes (signature value, SV por sus siglas en inglés). Los resultados significativamente desviados del valor basal uno, constituyen un tipo de firma de variación cuyas propiedades e implicaciones han sido desarrolladas en un nivel teórico en el artículo publicado (**ver Resultados Publicados**). La utilización formal de esta métrica como indicador general de la presencia de variación representa la avenida central de investigaciones futuras para este trabajo (**ver Perspectivas**).

3) Escaneo del PMGL

En el caso de genomas haploides en regiones únicas, la presencia de un sólo nucleótido diferente entre el genoma de referencia y el genoma de interés deja una huella en el PMGL debido a la ausencia de identidad absoluta en k nucleótidos consecutivos. Este tipo de firma de variación puede definirse algorítmicamente como una reducción en la cobertura de PMnCRs que genera una estela de valores iguales o cercanos a cero hasta la

posición n-1, seguida de su recuperación inmediata en la posición n. La estela de ceros se presenta también en el caso de deleciones, inserciones y variantes concatenadas (distanciadas por menos de k nucleótidos). La búsqueda de todas las ocurrencias de la estela de ceros a lo largo de un PMGL revela, de manera precisa, todas aquellas posiciones del genoma afectadas por estos tipos de variación. En el caso de encontrarse embebida en una zona repetida del genoma, la presencia de una estela de ceros inmediatamente indica que el genoma de interés no posee ninguna copia con la secuencia especificada por el genoma de referencia en el lugar correspondiente. Esta propiedad del PMGL es sumamente valiosa, ya que permite el refinamiento del genoma de referencia en secuencias repetidas, que son típicamente las más difíciles de ensamblar.

En un contexto más amplio y, por el momento, teórico, las firmas de variación reveladas por la métrica de SV abarcan los casos anteriormente mencionados y aquellos donde han ocurrido amplificaciones. Más aún, podrían revelar la presencia de todos los tipos de variantes en contextos más complejos, como las variantes provenientes del genoma de interés y presentes en zonas repetidas y el análisis de organismos diploides (ver **Perspectivas**). En el presente trabajo, el escaneo del PMGL en búsqueda de firmas de variación se restringe a aquellas que generan una estela de ceros.

4) Generación de un alineamiento inicial para cada firma de variación

Cada estela de ceros localizada mediante el escaneo del PMGL define una zona de variación bordeada por zonas de identidad absoluta, una ubicada río arriba y otra río abajo con respecto a la variante. La secuencia correspondiente a la zona de identidad absoluta río abajo es utilizada para la inicialización de un alineamiento entre el genoma de interés y el genoma de referencia en las posiciones inmediatamente adyacentes a la variación. La secuencia del genoma de interés que va a ser alineada se construye al identificar todas aquellas lecturas de secuenciación que contienen una correspondencia perfecta con la secuencia de identidad absoluta y un número x de nucleótidos río arriba (25 en este trabajo). En el contexto del alineamiento, la secuencia de identidad absoluta provee un punto de anclaje, y la secuencia río arriba revela la variación entre ambos genomas. Cuando la secuencia de identidad absoluta pertenece a una zona repetida se pueden generar una serie, o familia, de secuencias diferentes provenientes del genoma de interés. Para cada firma de variación, se generan independientemente uno o más alineamientos, dependiendo del tamaño de la familia correspondiente.

5) Interpretación y extensión de alineamientos

Todos los alineamientos generados para cada firma de variación entran en un proceso recursivo de interpretación y extensión. En cada iteración, cada firma de variación puede ser clasificada como solucionada, no solucionada, o en tránsito. Las firmas de variación solucionadas o no solucionadas se excluyen del proceso iterativo, que termina en cuanto todas las firmas de variación se encuentran colocadas en alguna de estas dos categorías. Todos los alineamientos correspondientes a una firma de variación en tránsito pueden ser extendidos en la siguiente iteración. Cada extensión requiere atraer un nuevo conjunto de lecturas de secuenciación. Éstas deben contener una correspondencia perfecta con la

secuencia de k nucleótidos localizada más alejada, río arriba, pero en sobreposición con la secuencia utilizada en la iteración anterior para construir la familia correspondiente. La secuencia del genoma de interés resultante es alineada con la zona, equivalentemente extendida, del genoma de referencia. Es útil imaginar el funcionamiento de este algoritmo para cada firma de variación como el crecimiento de un árbol. La secuencia de identidad absoluta represente el tronco del alineamiento, y las familias generadas en cada iteración representan ramas. Cada rama podría, a su vez, generar más ramas en la iteración siguiente. Sin embargo, todas las ramas de la iteración actual (simbólicamente localizadas a la misma altura del árbol) son evaluadas en su conjunto para determinar a qué categoría pertenece la firma de variación. Por esta razón se reporta un único alineamiento para cada firma de variación resuelta. Una firma de variación se clasifica como solucionada cuando los alineamientos generados (en lo que representa la última iteración para dicha firma) cumplen con las siguientes características: 1) comparten un motivo de variación común con respecto a la secuencia correspondiente del genoma de referencia, 2) solamente un alineamiento contiene la secuencia del genoma de interés que difiere del genoma de referencia únicamente en el motivo de variación común, y 3) dicho alineamiento es idéntico al genoma de referencia tanto río arriba como río abajo de la zona de variación. Estos requisitos aseguran que el alineamiento reportado contiene la secuencia del genoma de interés que corresponde a la zona del genoma de referencia donde se identificó la variación. Para cada firma de variación solucionada, el alineamiento final correspondiente revela directamente la naturaleza de la variación.

6) Generación de un genoma de referencia *ad hoc*

El último paso de la estrategia de análisis de variación genómica por PMGL consiste en generar un genoma de referencia *ad hoc* para validar las variantes descubiertas. Todas las variantes solucionadas se introducen, nucleótido por nucleótido, en las posiciones correspondientes de la secuencia del genoma de referencia, empezando por las posiciones más río abajo de cada cromosoma. La secuencia resultante representa un genoma de referencia refinado o el ensamble de otro genoma por re-secuenciación. Finalmente, es necesario repetir los módulos 1, 2, y 3 usando el genoma de referencia *ad hoc* y las lecturas de secuenciación originales del genoma de interés. La desaparición de las firmas de variación en los sitios correspondientes del genoma de referencia *ad hoc* confirma la naturaleza y la posición precisa de las variantes descubiertas.

Valoración del Panorama Genómico de Identidad Absoluta

La estrategia de análisis de variación genómica por PMGL ha sido evaluada en cuanto a sus dos aportaciones principales: refinamiento de genomas de referencia y generación de perfiles de variación de otros genomas. A su vez, ambos contextos han sido estudiados a través de experimentos de simulación y del análisis de cepas de levadura naturales. Los experimentos de simulación han evaluado la capacidad del método para detectar variantes provenientes tanto del genoma de interés como del genoma de referencia. Los análisis de cepas naturales han evaluado su capacidad para refinar genomas de referencia existentes, incluyendo el genoma altamente perfeccionado de la cepa S288C de *S. cerevisiae* y genomas de otras cepas ensamblados

recientemente con las últimas metodologías de la genómica (Yue, *et al*, 2017). Asimismo, el análisis de cepas naturales ha permitido evaluar la capacidad de la estrategia PMGL para generar perfiles de variación de genomas progresivamente más distantes, desde cepas directamente derivadas en el laboratorio hasta cepas ubicadas en otros clados de la filogenia de *S. cerevisiae*. Finalmente, la estrategia de análisis por PMGL ha sido evaluada en relación a su capacidad para refinar la secuencia diseñada de siete cromosomas sintéticos de levadura. La descripción detallada de esta serie de pruebas se encuentra en **Resultados Publicados** (todos los análisis de simulación, de cepas naturales y del primer cromosoma eucarionte funcional sintetizado por diseño) y en **Resultados adicionales** (el análisis de seis cromosomas eucariontes sintéticos adicionales).

Una Nueva Revolución Biológica: Diseño y Creación de Genomas Sintéticos

El HGP ha transformado el estudio de las ciencias biomédicas: ahora es posible realizar un análisis profundo de todo el material genético de cualquier individuo. No obstante, para muchos fenómenos biológicos de gran importancia todavía no hemos logrado dibujar el mapa que va de genotipo a fenotipo, conectando la estructura con la función. Una nueva revolución científica, la biología sintética, en una escala sin precedentes, propone diseñar y sintetizar genomas completos para cerrar esta brecha y entender la biología de manera fundamental. Dada la existencia de genomas de referencia de alta calidad para varios organismos, incluyendo el del humano, en principio es posible sintetizar genomas que se aparten cada vez más de este templado conocido, explorando así los efectos de cambios introducidos a voluntad. Se trata de revelar toda la potencialidad de los sustratos moleculares a través del diseño, eliminando la limitante de lo que es posible conocer a través de la secuenciación de un x número de individuos.

El objetivo del HGP-write es diseñar y sintetizar genomas tan complejos como el del humano, e introducirlos en líneas celulares en un plazo de diez años, con el fin de lograr un mayor entendimiento del "código de la vida" revelado por el HGP (Boeke *et*. al., 2016). Aunque, de ser exitoso, el HGP-write promete ser igualmente o incluso más revolucionario que el HGP, existe una diferencia fundamental entre ambos que le da al HGP-write un carácter especial. El objetivo del HGP convergió en el conocimiento de la secuencia completa del genoma humano. En cambio, la síntesis del genoma humano representa un reto científico sin fronteras, que potencialmente puede abarcar tantos proyectos como preguntas sobre la biología. Prueba de ello es la diversidad de proyectos piloto que encabezan los esfuerzos iniciales del HGP-write (GP-write Leadership Group and The GP-Write Consortium, 2016). Entre éstos figuran:

1) La generación de una línea celular "ultrasegura", resistente a ataques virales, a priones y a transformación cancerígena, y que minimice las posibilidades de rechazo inmunológico y de senescencia. Esta línea celular serviría como plataforma para diversas aplicaciones biomédicas, incluyendo la producción de agentes biológicos y distintas estrategias de terapia genómica.

2) La generación de una línea celular prototrófica capaz de sintetizar amino ácidos y vitaminas esenciales para la vida de la especie humana, y cuyas vías metabólicas están

ausentes en el genoma humano. Esta línea celular podría ayudar con los problemas de malnutrición crónica y escasez de alimentos que atacan a ciertas poblaciones.

3) Recientemente se han aislado líneas celulares humanas a partir de ovocitos haploides, generando células madre embriónicas con un karyotipo haploide normal. Estas líneas celulares poseen un perfil genómico pluripotente, incluyendo la capacidad de auto-renovación. Sorprendentemente, también permiten establecer estados somáticos diferenciados provenientes de las tres capas embriónicas germinales. Este descubrimiento ha inspirado la propuesta de sintetizar un genoma humano haploide. La complejidad del genoma humano se reduciría considerablemente, en particular gracias a la ausencia de complementación genética. El diseño riguroso de tamices ("screens") de pérdida de función en este contexto haploide abriría nuevas puertas para el estudio de la genómica funcional humana.

Aunque la mayoría de los proyectos piloto propuestos para el HGP-write se concentran en una fracción del genoma humano, típicamente en el rango de un 1% (GP-write Leadership Group and The GP-Write Consortium, 2016), existe una necesidad crítica de desarrollar metodología tanto experimental como computacional para el diseño, síntesis y valoración *in vivo* de genomas parcial o completamente artificiales. Por esta razón, la iniciativa de síntesis de genomas delineada por el HGP-write ha sido expandida para abarcar organismos modelo. Se espera que, al igual que el HGP comenzó con la secuenciación de organismos menos complejos, el HGP-write pueda beneficiarse de avances metodológicos en la síntesis de genomas más sencillos. Al proyecto en su contexto más amplio se le denomina GP-write. Algunos candidatos para la síntesis de sus genomas en un futuro cercando son la mosca *Drosophila melanogaster*, el nemátodo *Caenorhabditis elegans* y por supuesto, el ratón, que representa el modelo experimental mamífero mejor entendido.

Hoy en día, se encuentran en marcha dos proyectos de síntesis de genomas completos. El primero consiste en la síntesis de un genoma recodificado de *Escherichia coli* (Ostrov *et. al*., 2016) y el segundo consiste en la síntesis de un genoma extensamente modificado de la levadura *S. cerevisiae*. En el contexto del primer proyecto, 7 de los 64 codones que constituyen el código genético han sido seleccionados y reemplazados por codones sinónimos a lo largo de todos los genes codificantes del genoma. Se han modificado un total de 62,214 codones. Al ser validados individualmente, la mayoría de los cambios han conservado la funcionalidad de los genes alterados. Desde un punto de vista evolutivo, es interesante que han surgido, de manera natural, desviaciones del código genético universal en distintos linajes procariontes y eucariontes. En un contexto de biología sintética, es importante comprender hasta qué punto podemos alterar un genoma sin comprometer la supervivencia de un organismo. Alterar el código genético representa posiblemente uno de los cambios más fundamentales para probar su plasticidad. Este tipo de experimentos también abren una serie de oportunidades para el diseño de organismos con propiedades interesantes. Por ejemplo, una vez recodificadas todas las instancias de un codón, y eliminado su tRNA correspondiente, es posible aislar genéticamente al organismo de una serie de situaciones, incluyendo ataques virales. Esto es porque el material genético del virus no podrá ser traducido apropiadamente. Más aún, los codones eliminados y por lo tanto "liberados" de sus funciones normales pueden ser "reasignados" a aminoácidos no canónicos, expandiendo así la potencialidad del código genético. Aunque falta por ensamblar todos los componentes sintéticos

de este proyecto de recodificación de genomas, los avances hasta el momento sugieren que reescribir genomas completos, incluso de manera radical, es algo factible.

El proyecto más ambicioso de síntesis de genomas en curso es el Synthetic Yeast Project, también llamado Sc2.0 (Richardson *et al*., 2017). El objetivo del Sc2.0 es diseñar, sintetizar y ensamblar *in vivo* los 16 cromosomas nucleares de la levadura *S. cerevisiae* en su forma haploide, lo cual representa 12 Mb de ADN artificial. El Consorcio Sc2.0 ha sido creado con el fin de distribuir la construcción de cromosomas individuales en diferentes laboratorios alrededor del mundo, incluyendo Estados Unidos, el Reino Unido, Australia, Francia, Alemania, Singapur y China. Una vez completada la síntesis de cromosomas individuales, éstos serán consolidados en una misma cepa, generando así el primer genoma eucarionte por diseño.

El diseño de la secuencia del Sc2.0 propone un doble reto para la biología sintética. Por una parte, busca alejarse de algunas de las propiedades más características de los genomas conocidos, al igual que implementar sistemas inducibles de evolución acelerada. Por otra parte, requiere mantener un nivel adecuado de adaptación del organismo. La base de la secuencia del Sc2.0 parte de derivados de la cepa S288C, cuyo genoma de referencia ha sido refinado y curado durante dos décadas. El diseño integra una serie de cambios densamente distribuidos que comprenden un total de 1.1 megabases de ADN deletado, insertado o alterado para generar una cepa altamente modificada y cuyo genoma sería reducido en un 8% de su tamaño original. Entre los cambios más prominentes figuran:

- La utilización de un código genético ligeramente modificado, en donde todos los codones de paro TAG han sido reemplazados por TAA.
- La introducción de la secuencia palindrómica de recombinación loxPsym a lo largo del genoma, lo cual permite la generación de rearreglos genómicos como inversiones, deleciones y duplicaciones al administrar Cre recombinasa. A este sistema de evolución inducida se le conoce como SCRaMbLE.
- La eliminación de elementos que contribuyen a la inestabilidad del genoma, incluyendo retrotransposones (alrededor de 50 copias pertenecientes a 5 familias de retrotransposones conocidos como elementos Ty) y tRNAs. Además de representar sitios preferidos de inserción de elementos Ty, los loci de genes de tRNA pueden ocasionar eventos de colisión entre las RNA polimerasas II y III. De hecho, se ha diseñado un nuevo cromosoma, o neocromosoma para albergar todos los genes de tRNA.
- Se han eliminado la mayoría de los intrones incorporados en moléculas de pre-tRNA y pre-mRNA.

Varios mecanismos endógenos de *S. cervisiae* están siendo elegantemente utilizados para generar una cepa viviente completamente sintética. El mecanismo intrínseco de recombinación homóloga y el de mantenimiento de telómeros, permiten la construcción progresiva y consolidación de cada cromosoma sintético, respectivamente. El mecanismo de endoreduplicación, aunado a una estrategia artificial de desestabilización de cromosomas nativos, permite obtener productos haploides polisintéticos a partir de diploides heterocigotos (con un cromosoma sintético y otro nativo).

Actualmente, más de un tercio de los cromosomas sintéticos del genoma Sc2.0 se encuentran representados en cepas individuales. A pesar del gran número de cambios incorporados por diseño en los cromosomas sintéticos completados hasta el momento, synIII (Annaluru *etal*., 2014), synII (Shen *et al*., 2017), synV (Xie *et al*., 2017), synVI (Mitchell *et al*., 2017), synX (Wu *et al*., 2017) y synXII (Zhang *et al*., 2017), las cepas correspondientes presentan niveles de adaptación esencialmente normales en una serie de condiciones de crecimiento. De hecho, diferentes tipos de análisis realizados, incluyendo fenomas, transcriptomas, proteomas y monitoreo de los procesos de segregación de cromosomas y de replicación revelan, en su mayoría, el mantenimiento de perfiles genómicos muy similares a los de las cepas originales. No obstante, el proceso para obtener versiones finales de cada cromosoma con estas características de estabilidad y funcionalidad es extremadamente laborioso. En ocasiones, un solo cambio, dentro de los miles incorporados en la secuencia sintética, puede afectar sustancialmente la capacidad proliferativa del organismo, en cuyo caso debe ser precisamente identificado y corregido. El ensamble de cromosomas sintéticos, marcados con cambios sinónimos en sitios estratégicos, ha permitido el desarrollo de metodologías de alto rendimiento para mapear, masivamente, todas aquellas secuencias sintéticas que confieren fenotipos sub-óptimos. A este proceso de reparación de genomas sintéticos se le conoce como "debugging", haciendo alusión al proceso de verificación de programas computacionales.

Otro aspecto fundamental del Sc2.0 y, en general, de cualquier proyecto de biología sintética, es la fidelidad de la secuencia viviente con respecto a la secuencia diseñada. Es crucial una correspondencia perfecta entre ambas para la evaluación sistemática de los principios de diseño implementados en la síntesis de genomas. Lo último es especialmente relevante si se pretende, como es el caso, distanciarse cada vez más de las formas conocidas de la naturaleza. Dos razones principales motivan dichas incursiones hacia lo desconocido:

1) Poner a prueba nuestro conocimiento acumulado sobre los productos de la evolución y sus funciones respectivas.

2) Conferir una potencialidad extendida a los sistemas moleculares y celulares pertenecientes a una diversidad de organismos.

La importancia de ensamblar una secuencia viviente que corresponda exactamente con la secuencia diseñada se ejemplifica en la elaboración del cromosoma synV. La obtención de la versión final del cromosoma synV requirió de 22 pasos de edición, representados en una serie de cepas vivientes intermediarias. El diseño viviente obtenido de esta forma demuestra la factibilidad de generar cromosomas artificiales "perfectos". La construcción de genomas sintéticos refinados al máximo nivel requiere de estrategias de análisis genómico que permitan encontrar todas las discrepancias entre secuencias vivientes y diseñadas de manera rápida y sumamente precisa. La estrategia de búsqueda de variación genómica desarrollada en este trabajo de investigación se acopla perfectamente a este fin. De hecho, se ha establecido formalmente una colaboración con el Consorcio Sc2.0 para determinar el grado de correspondencia entre los diseños de cromosomas sintéticos y sus contrapartes vivientes. En este contexto se han analizado los cromosomas synIII, synVI, synII, synV, synX, synXII y syn XI, lo cual representa un 44% (en número de cromosomas) del genoma del Sc2.0. Los resultados del análisis del cromosoma synIII, el primer cromosoma eucarionte funcional sintetizado, se encuentran publicados (**ver**

Resultados publicados), y los resultados correspondientes al resto de los cromosomas se encuentran descritos en esta tesis (**ver Resultados Adicionales**). En el caso del cromosoma synXI, cuya síntesis final no ha sido completada aún por el proyecto Sc2.0, nuestro análisis ha servido para guiar el proceso de reparación de la secuencia viviente. Esperamos extrapolar esta colaboración a la totalidad del genoma Sc2.0 y escribir un artículo en conjunto con el Consorcio Sc2.0 delineando la metodología para este tipo de análisis de genomas sintéticos, cuya herramienta central correspondería a la estrategia PMGL desarrollada en esta tesis. Dependiendo del cromosoma analizado (**ver Resultados Adicionales**), el PMGL iguala o, frecuentemente, supera el rendimiento de los algoritmos utilizados hoy en día, que representan el estado del arte en el área de análisis de variación genómica.

En el contexto de esta colaboración, hemos diseñado una estrategia de análisis de variación genómica por PMGL que permite una interrogación dirigida hacia cualquier subconjunto del genoma y, en particular, hacia cromosomas individuales. Esta manera de analizar directamente secciones específicas del genoma no requiere de protocolos experimentales de captura o enriquecimiento de la muestra de ADN; se utilizan lecturas de secuenciación de genoma completo. En el caso de cromosomas sintéticos del proyecto Sc2.0, el genoma de referencia es construido a partir del genoma de referencia de la cepa S288C de *S. cerevisiae*, intercambiando la secuencia del cromosoma correspondiente por su secuencia diseñada. Se construye un RGSL utilizando la totalidad de este nuevo genoma de referencia. Se extrae un subconjunto de la estructura RGSL, seleccionando únicamente aquellos renglones que pertenecen a la secuencia del cromosoma diseñado (RGSL dirigido). Subsecuentemente, se genera un PMGL dirigido, utilizando la estructura RGSL dirigida y la totalidad de las lecturas de secuenciación de la cepa viviente que contiene al cromosoma sintético. De esta manera, se toma en cuenta la arquitectura completa del genoma de referencia, pero la operación de identidad absoluta con el genoma de interés es realizada únicamente para el cromosoma deseado. Utilizando el PMGL dirigido, la búsqueda de firmas de variación se restringe automáticamente. El resto del análisis prosigue como ha sido descrito en esta tesis. Este tipo de análisis dirigido es especialmente valioso en el contexto de la construcción progresiva de un genoma completamente sintético, ya que, en sus múltiples fases intermedias, coexisten estructuras de ADN tanto artificiales como naturales.

RESULTADOS PUBLICADOS

CARTA DE ACEPTACIÓN

January 19, 2018
RE: GENETICS/2017/300589R1

Ms. Kim Palacios Flores
Universidad Nacional Autónoma de México
Laboratorio Internacional de Investigación sobre el Genoma Humano
Blvd. Juriquilla 3001
Querétaro 76230
Mexico

Dear Dr. Palacios Flores:

Congratulations! Your manuscript entitled "**A Perfect Match Genomic Landscape Provides a Unified Framework for the Precise Detection of Variation in Natural and Synthetic Haploid Genomes**" is acceptable for publication in GENETICS.

Thanks for sending this nice story to GENETICS. I hope you will continue to submit your best work for publication in our journal.

The reviewer had a few suggestions for improving the manuscript that you may want to consider. You can view their comments at the bottom of this email.

Please visit http://www.genetics.org/content/after-acceptance to submit the final files of your manuscript for publication.

When preparing the final version of the ms., please make an effort to shorten the text. It is our experience that shortening manuscripts usually improves them. The editors urge authors to heed the advice of Strunk and White: "omit needless words".

With GENETICS' Early Online, your currently-accepted manuscript (unedited, as submitted, reviewed and accepted) will be published at GENETICS' website (and then deposited into PubMed) shortly after receipt of source files. Please include an updated PDF with your source files (we encourage you to format the PDF for ease of reading).

If you have questions about or problems while uploading your accepted manuscript files, please email the editorial office at sourcefiles@thegsajournals.org.

Sincerely,

Mark Johnston
Associate Editor
Genetics

Gary Churchill
Senior Editor
Genetics

note: Proofs will be emailed (from Dartmouth Journal Services) within 3 weeks. Please add Genetics_Specialist.djs@sheridan.com (or the domain @sheridan.com) to your email program's "safe senders" list.

ARTÍCULO PUBLICADO

Title

A Perfect Match Genomic Landscape Provides a Unified Framework for the Precise Detection of Variation in Natural and Synthetic Haploid Genomes

Authors

Kim Palacios-Flores*, Jair García-Sotelo*, Alejandra Castillo*, Carina Uribe*, Luis Aguilar*, Lucía Morales*, Laura Gómez-Romero*, José Reyes*, Alejandro Garciarubio †, Margareta Boege*, and Guillermo Dávila*

Affiliations

***Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Querétaro, Querétaro, México, 76230 † Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México, 62210**

Running title

Perfect Match Genomic Landscape

Key words

genome variation
genome sequencing
reference genomes
yeast genomics
synthetic genomics

Corresponding author

Kim Palacios Flores

Institutional Affiliation

Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Querétaro, Querétaro, México, 76230

Telephone: **+52 (442) 1926231, ext. 205**
Cell phone: **442 4266948**
Email: **kimpalaciosflores@gmail.com**

ABSTRACT

We present a conceptually simple, sensitive, precise, and essentially non-statistical solution for the analysis of genome variation in haploid organisms. The generation of a Perfect Match Genomic Landscape, which computes inter-genome identity with single nucleotide resolution, reveals signatures of variation wherever a query genome differs from a reference genome. Such signatures encode the precise location of different types of variants, including single nucleotide variants, deletions, insertions, and amplifications, effectively introducing the concept of a general signature of variation. The precise nature of variants is then resolved through the generation of targeted alignments between specific sets of sequence reads and known regions of the reference genome. Thus, the perfect match logic decouples the identification of the location of variants from the characterization of their nature, providing a unified framework for the detection of genome variation. We assessed the performance of the Perfect Match Genomic Landscape strategy via simulation experiments. We determined the variation profiles of natural genomes and of a synthetic chromosome, both in the context of haploid yeast strains. Our approach uncovered variants that have previously escaped detection. Moreover, our strategy is ideally suited for further refining high-quality reference genomes. The source codes for the automated PMGL pipeline have been deposited in a public repository.

INTRODUCTION

At the heart of genomics lies the precise determination of an organism's DNA sequence. Genome projects typically generate large amounts of sequence reads, which constitute a fragmented and unordered representation of genetic information. Sequence reads are subsequently assembled either *de novo* (Zervino and Birney, 2008) or through comparison with the ordered genetic information of a reference genome (Metzker, 2010). Reference genomes exist for different species from bacteria (Blattner *et al.*, 1997) to human (International Human Genome Sequencing Consortium, 2004). The central position of reference genomes as platforms for uncovering the nucleotide sequence of related genomes, underscores the importance of their continuous refinement according to conceptual and methodological advances (Goodwin et al., 2016). Genomic studies typically contrast the variation profiles of genomes of interest in a broad set of contexts, ranging from experimental evolution (Tenaillon *et al.*, 2016) to personalized medicine (Abrahams and Eck, 2016). The generation of both high-quality reference genomes and precise variation profiles between genomes is therefore of utmost importance.

Most current algorithms for detecting genome variation are based on mapping sequence reads from the query genome to the reference genome to infer their corresponding locations. The problem of aligning sequence reads to a reference genome is central to genomics (Pfeifer, 2017), as it is the basis for such procedures as whole genome sequencing (1000 Genomes Project Consortium, 2012), exome (Teer and Mullikin, 2010) and transcriptome sequencing (Wang *et al.*, 2009), and ChIP-Seq (Park, 2009), amongst others. A plethora of programs exist to solve this problem computationally (Mardis, 2013; Goodwin *et al.*, 2016). Most methods index the reference genome into highly optimized data structures (Li *et al.*, 2008a; Li *et al.*, 2008b; Kurtz, 2003; Chaisson and Tesler, 2012; Kurtz *et al.*, 2004; Li and Durbin, 2009; Langmead and Salzberg, 2012; Holt and McMillan, 2014), generating a variety of specialized algorithms

(Schbath *et al*., 2012). Due to experimental error (Yang et al., 2013), or true variance between the query genome and the reference genome, most sequence reads do not match exactly with the reference genome. Thus, all aligners ultimately try to solve the "approximate string matching" problem (Reinert *et al.*, 2015) using some arbitrary measure of "acceptable in-exactness". The optimal placement of sequence reads is therefore reported in conjunction with some measure of reliability. Consequently, the discovered variants and the resulting query genome sequence are likewise statistical in nature (McKenna *et al.*, 2010; Li, 2011; Rimmer *et al*., 2014; Koboldt *et al*., 2012).

We have conceptualized the analysis of genome variation from a different perspective, decomposing it into two independent processes. First, finding where the query genome and the reference genome are not identical, and second, revealing the nature of the underlying variants. The precise location of genome sites affected by variation is directly determined from a genome-wide identity landscape, or Perfect Match Genomic Landscape (PMGL). Variant characterization can thus be conducted locally, and solutions can be validated in a qualitative manner. We have previously reported the potential to precisely locate single nucleotide variants by individualizing regions of the reference genome (Reyes *et al.*, 2011), a step that is incorporated into the PMGL strategy. Most interestingly, a recent study has developed an algorithm that is based on a similar principle to that of the PMGL strategy (Audano *et al.*, 2017). Their algorithm also reduces the variant search space by first identifying regions that differ between the query genome and the reference genome. In addition to determining the variation profiles of both natural and synthetic query genomes, the non-statistical nature of the PMGL strategy is particularly suited for refining reference genomes. In fact, the PMGL strategy can penetrate both the unique and repeated compartments of a reference genome.

MATERIALS AND METHODS

General protocol for the PMGL pipeline. The reference genome sequence in fasta format is used to generate a binary database of the reference genome using Bowtie (Langmead et al 2009), and to generate the ordered set of reference strings (25-mers in this study) that constitute the entire reference genome. The number of exact occurrences of each reference string's sequence in the reference genome database is computed. A Reference Genome Self Landscape (RGSL) is generated by reporting each reference string's unique identifier, number of exact occurrences in the reference genome, sequence, and the unique identifiers of all reference strings sharing the same sequence. The raw query genome sequence reads in fastq format are used to generate a binary database of read string counts (25-mers in this study) computed by Jellyfish (Marçais and Kingsford 2011). The use of quality-trimmed sequence reads is not necessary (Figure S1). A PMGL is generated by reporting the perfect match coverage between the reference genome and the query genome at each reference string along the RGSL. The perfect match coverage is then normalized by the level of repetitiveness of each reference string in the reference genome. Finally, the normalized perfect match coverage at reference string n is divided by the normalized perfect match coverage at reference string n-1. The latter corresponds to each reference string's signature value. The PMGL is scanned to localize signatures of variation. A signature of variation is defined as a decrease in the normalized perfect match coverage that generates a trail of 0 or near-zero values terminating at position n-1, followed by its immediate recovery at

position n. The PMGL scan parameters and their relation to sequencing coverage has been experimentally addressed (Figure S1). Zero-trail signatures of variation are associated with a high signature value at position n. For SNVs, micro-indels, and indels, the reference string at position n, or downstream recovery string, corresponds to a perfect match zone that is immediately adjacent to the variation. Its sequence is used to identify the subset of query genome sequence reads that perfectly contain it. The query genome sequence(s) defined by such sequence reads is aligned with the corresponding region of the reference genome using the MUSCLE Multiple Sequence Alignment tool (Edgar 2004). The nature of the variant(s) is revealed by an iterative process of alignment interpretation and extension resulting in a single final alignment. Finally, discovered variants are introgressed into the original reference genome sequence to generate a customized reference genome. The disappearance of signatures of variation using the customized reference genome and the original query genome sequence reads as input for the PMGL pipeline validates the precise location and nature of the discovered variants.

The PMGL pipeline has been fully automated and comprises six computational modules: 1) Generation of the RGSL, 2) Generation of the PMGL, 3) Scanning of the PMGL, 4) Generation of the first alignment at each signature of variation, 5) Interpretation and extension of alignments, and 6) Generation of a customized RG. All modules are described in detail in File S1.

Detailed Materials and Methods section. File S1presents in detail the following methodology: *S. cerevisiae* strains and culture; DNA isolation and Illumina sequencing; Generation of PCR products and Sanger Sequencing; Reference genomes and query genomes; Automated PMGL pipeline; Genome-wide distribution of signature values; Random simulation in the query genome; Directed simulation in repeated regions of the reference genome.

Data availability. The source codes for the automated PMGL pipeline have been deposited in the public repository GitHub:
https://github.com/LIIGH-UNAM/PerfectMatchGenomicLandscapePipeline.git
Output files generated by intermediate steps of the automated PMGL pipeline, along with the corresponding customized genomes have been deposited in the public repository GitHub:
https://github.com/LIIGH-UNAM/GeneratingVariationProfilesRefiningReferenceGenomesUsingPMGLPipeline.git
The Illumina reads generated in this study have been deposited in the public repository GitHub:
https://github.com/LIIGH-UNAM/SequenceReads.git
File S1, detailed description of Materials and Methods. File S2, examples of variant location and characterization, including relevant subsets of the corresponding PMGL, local alignments, and Sanger sequence graphs. File S3, random simulation experiment in the query genome. File S4, directed simulation experiment in multiple copy regions of the reference genome. File S5, analyses of the S288C and BY4742 *S. cerevisiae* strains. File S6, analysis of the *S. cerevisiae* strain SK1. File S7, analysis of the *S. cerevisiae* strain Y12. File S8, comparison of the as-designed synIII sequence with the query genome sequence reads of *S. cerevisiae* strain HMSY011 from a previous analysis and from this study. File S9, alignments showing the variants uncovered for synIII. File S10, simulation experiment within the loxPsym family of synIII.

RESULTS

Rationale of the PMGL strategy. To define genome variation, our strategy exclusively utilizes perfect matches between the reference genome and the query genome. This may seem counterintuitive, as one would not expect to obtain any information about variation between genomes by focusing precisely on their invariant complement. This apparent paradox, however, can be resolved by constructing a PMGL, which reports the number of perfect matches at each successive position along the reference genome, and interpreting sudden changes in the perfect match coverage as direct indicators of the precise location of variants. Single nucleotide variants, deletions, and insertions cause a depression in the perfect match coverage. Copy number increases such as those generated by amplification cause a rise in the perfect match coverage. Importantly, changes in both directions occur sharply, between two immediately adjacent nucleotides along the reference genome, resulting in signatures of variation with single nucleotide resolution. Scanning the PMGL for signatures of variation determines the precise location of variants along the reference genome. The generation of highly targeted alignments at variation sites resolves their specific nature.

PMGL pipeline. A reference strings dataset is used to construct the RGSL structure. In turn, the RGSL and a read strings dataset are used to construct the PMGL structure (Figure 1A). The reference strings dataset contains the ordered set of overlapping DNA strings, each of size k (25 nucleotides in this work), generated using a 1 nucleotide sliding window along the reference genome. The RGSL reports each reference string's starting position within a specific chromosome (unique identifier, ID column), its nucleotide sequence (SEQ column), the number of occurrences of its sequence in the entire reference genome (count reference, CR column), and the unique identifier of all reference strings sharing the same sequence (repeat family unique identifiers, IDF column).The RGSL describes the architecture of the reference genome by continuously assessing its degree of repetitiveness. The RGSL structure derived from the *S. cerevisiae* S288C reference genome contains a total of 12,156,697 reference strings, and 93% of their sequences occur only once in the entire genome. The read strings dataset contains the set of DNA strings, each of size k, generated using a 1 nucleotide sliding window along all query genome sequence reads. The total number of occurrences of each read string is computed. The PMGL incorporates the RGSL as a structural backbone to report the number of perfect matches between the Reference Strings and the Read Strings (PM column). The number of perfect matches associated with each reference string is normalized by its count reference (PMnCR column).

Scanning the PMGL reveals the precise location of different types of variation between the query genome and the reference genome, along the reference genome (Figure 1 B). The detection of a single nucleotide variant illustrates the simple nature of the PMGL strategy (Figure 2). A single nucleotide change reduces the perfect match coverage at k successive reference strings, those overlapping with the variant. Within unique regions of a haploid genome, this sharp depression reaches 0 or near-zero values, generating a zero-trail signature of variation. The nucleotide sequence of a reference string immediately adjacent to the signature of variation (recovery string) is used to identify the subset of sequence reads that perfectly contain it. The specific nucleotide change is revealed by generating a local alignment with the corresponding region of the reference genome (Figure 1C).

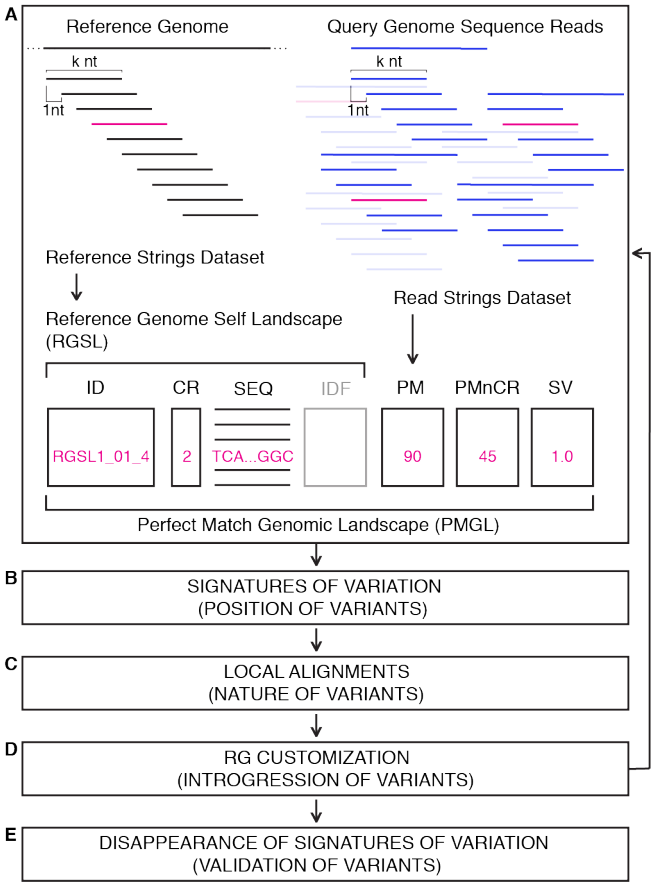


Figure 1 Pipeline of the PMGL strategy. A) Generation of the PMGL. Both the reference genome and the query genome sequence reads are decomposed into comprehensive arrays of k nucleotide-long strings, generating the reference strings dataset and the read strings dataset, respectively. The PMGL reports the number of perfect matches between the reference genome and the query genome along the RGSL structure. ID, reference string's unique identifier; SEQ, sequence; CR, count reference; IDF, repeat family unique identifiers; PM, perfect match coverage; PMnCR, perfect match coverage normalized to count reference. The signature value (SV) is the ratio of normalized perfect matches between successive reference strings. The reference genome and reference strings are shown in black; query genome sequence reads and strings are shown in blue; a specific string is shown in magenta across the different datasets. Procedures B) through E) are explained in the text and in Materials and Methods, and detailed in File S1.

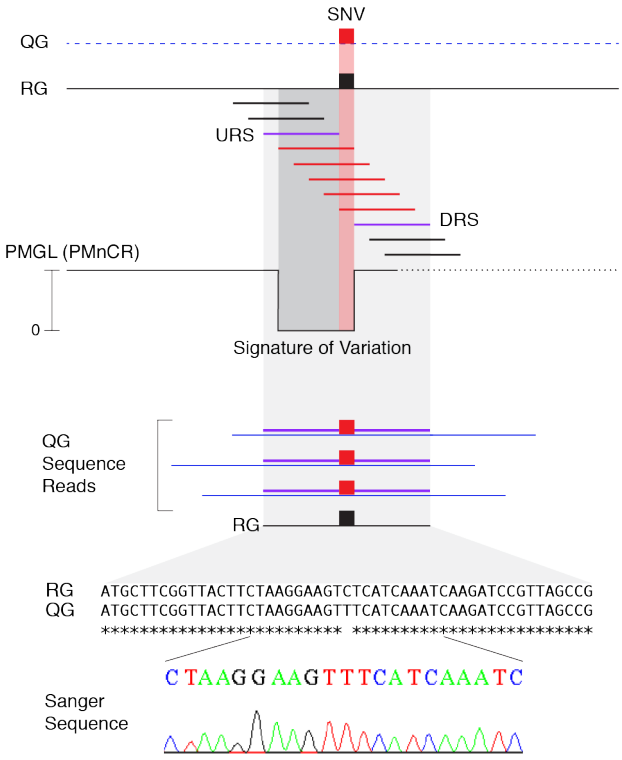


Figure 2 Detection and characterization of a single nucleotide variant. Corresponding regions of the query genome and the reference genome are represented by a dashed blue line and a solid black line, respectively, and harbor a single nucleotide variant represented by red and black boxes. Reference strings are color-coded: black strings do not include the variant, red strings incorporate the variant, and the purple strings correspond to the upstream recovery string (URS) and the downstream recovery string (DRS). The normalized perfect match coverage (PMnCR), plotted as a solid black line, generates a zero-trail signature of variation. Dark grey and red shading indicate the zero-trail zone. Red shading represents the projection of the variant into the reference genome. Light grey shading spans the reference genome segment used for the alignment. Sequence reads containing both the perfect match zone defined by the recovery strings and the variant are shown. The alignment reveals the nature of the variant at the expected site. A section of a Sanger sequence obtained from a PCR product of the corresponding region is shown.

The detection of other types of variation (Figure 3) represents an extension to the single nucleotide variant case. Indeed, the previously described signature persists, and only its shape is modified. A deletion in the query genome results in an increase in the length of the signature corresponding to the size of the deletion. The signature is generated in the presence of any type of insertion because reference strings skip the inserted sequence and do not produce perfect matches with read strings derived from the borders of the insertion. In cases where variants are separated by less than k nucleotides (concatenated variants), their corresponding signatures of variation are merged into a single signature of variation with increased length. Thus, haploid query genomes can be interrogated for the presence of single nucleotide variants, deletions, and insertions using a zero-trail scan along the PMGL (examples are provided in File S2).

The zero-trail zone generated by single nucleotide variants, deletions, and insertions is followed by a sharp rise in the perfect match coverage between two reference genome positions one nucleotide apart. Interestingly, the latter pattern is also present at the starting point of amplifications (Figure 3). The detection of amplification sites, however, has not been experimentally addressed here. These sudden changes can be quantified in a genome-wide manner by computing the ratio of normalized perfect matches between successive reference strings, herein referred to as signature value (SV column of the PMGL) (Figure 1A). Single nucleotide variants, deletions, and insertions occurring in unique regions of a haploid query genome or in both the single and repeated compartments of a reference genome (see below) generate a very high signature value at the downstream recovery string.

More generally, signature values significantly greater than 1 are indicative of variation. Indeed, sharp oscillations in perfect match coverage are not expected to occur between two immediately adjacent nucleotides in the absence of variation. By performing a genome-wide computation of the signature value, and plotting the probability that a site in the genome has a signature value of x, we have confirmed that signature values generate a very narrow distribution centered at 1 (Figure S2), providing a robust baseline for quantitation. In principle, the signature value could be used as a general metric for detecting different types of variation in a variety of contexts (see Discussion).

The PMGL strategy allows a final, qualitative validation of discovered variants. This requires the customization of the reference genome through the *in silico* introgression of each uncovered variant (Figure 1D), the construction of a new RGSL using the customized reference genome, the construction of a PMGL using the original query genome sequence reads, and the scanning of this newly generated PMGL using the original search parameters. The loss of signatures of variation at the expected sites confirms the precise nature and position of the variants originally detected (Figure 1 E).

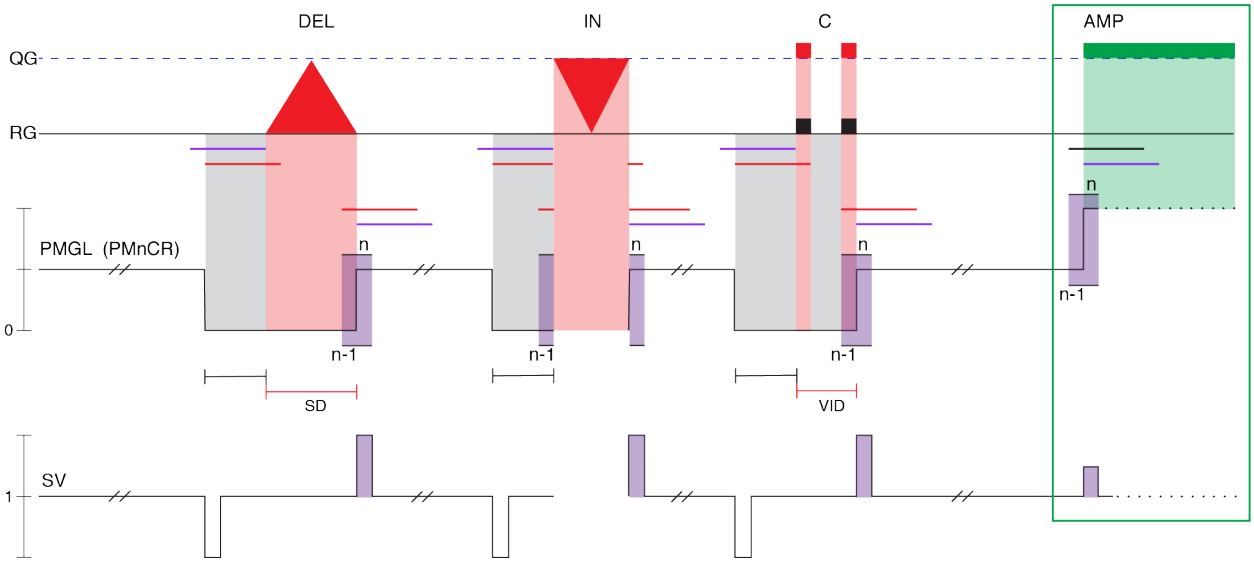


Figure 3 Detection of different types of variation. Corresponding features are color-coded as in Figure 2. DEL, deletion; IN, insertion; C, concatenated variants; AMP, amplification (shown in green). The AMP case is boxed because amplifications have not been experimentally addressed in this study. Reference strings contributing to the sudden drop and / or rise in the normalized perfect match coverage are shown for each case. The normalized perfect match coverage (PMnCR), plotted as a solid black line, reveals the corresponding signatures of variation. The relative length of each signature of variation is indicated by a black bar, which represents a constant region of about k-1 nucleotides, followed by a red bar which corresponds to a variable region contributed by the specific variant: SD, size of deletion; VID, variation inclusive distance between concatenated variants. The purple shading indicates the sharp rise in normalized perfect match coverage shared by all types of variants, and its relative magnitude (signature value, SV) is schematized at the bottom.

Performance of the PMGL Strategy Assessed through Simulation Experiments. The nature of the PMGL Strategy for detecting variation is qualitatively different from most previously described methodologies. Accordingly, we assessed its performance via *in silico* simulation experiments. Figure 4 consolidates the data from two types of simulation experiments.

To detect changes harbored in the query genome, we randomly introduced 100 variants into the reference genome of the *S. cerevisiae* yeast strain S288C. This altered genome was used to generate an artificial set of query genome sequence reads, and the unaltered S288C reference genome was used to construct the RGSL. The variants introduced included single nucleotide variants, deletions, and insertions; some were concatenated (File S3). Using the zero-trail scan, 103 signatures of variation were found. A big deletion generated more than one signature of variation (see File S1, Scanning of the PMGL). Direct inspection of the PMGL revealed the size and position of big deletions, and the search for artificial sequence reads containing both the downstream and upstream recovery string sequences revealed the breakpoint. In the case of big insertions, the search for artificial sequence reads containing either the downstream or the upstream recovery string sequence revealed the corresponding breakpoint. As expected, all zero-trail signatures of variation were associated with a high signature value at the downstream recovery string. The 3 altered sites that did not produce a zero-trail signature of variation were contained in repeated regions of the genome (see Discussion).

The simulation experiment described above was then performed introducing varying proportions of single nucleotide errors at random positions in the artificial sequence reads (Table S1). When up to 1% simulated sequencing errors were introduced, only one of the previously detected signatures of variation was lost. For all detected variation sites, the nature of the underlying variants was correctly resolved. At 2% and 3% of introduced errors, 90% and 52% of the original signatures of variation were found, respectively. The decrease in the percentage of signatures of variation recovered at high error rates is mainly due to the decrease in the absolute number of unaltered read strings that can produce perfect matches with the reference genome. In fact, when a two-fold increase in the total number of reads was implemented under the 3% errors regime, the percentage of detected signatures of variation increased to 75%. Importantly, at all percentages of introduced errors, no new signatures of variation were generated. Such absence of false positive signals highlights the robustness of the PMGL strategy to sequencing errors.

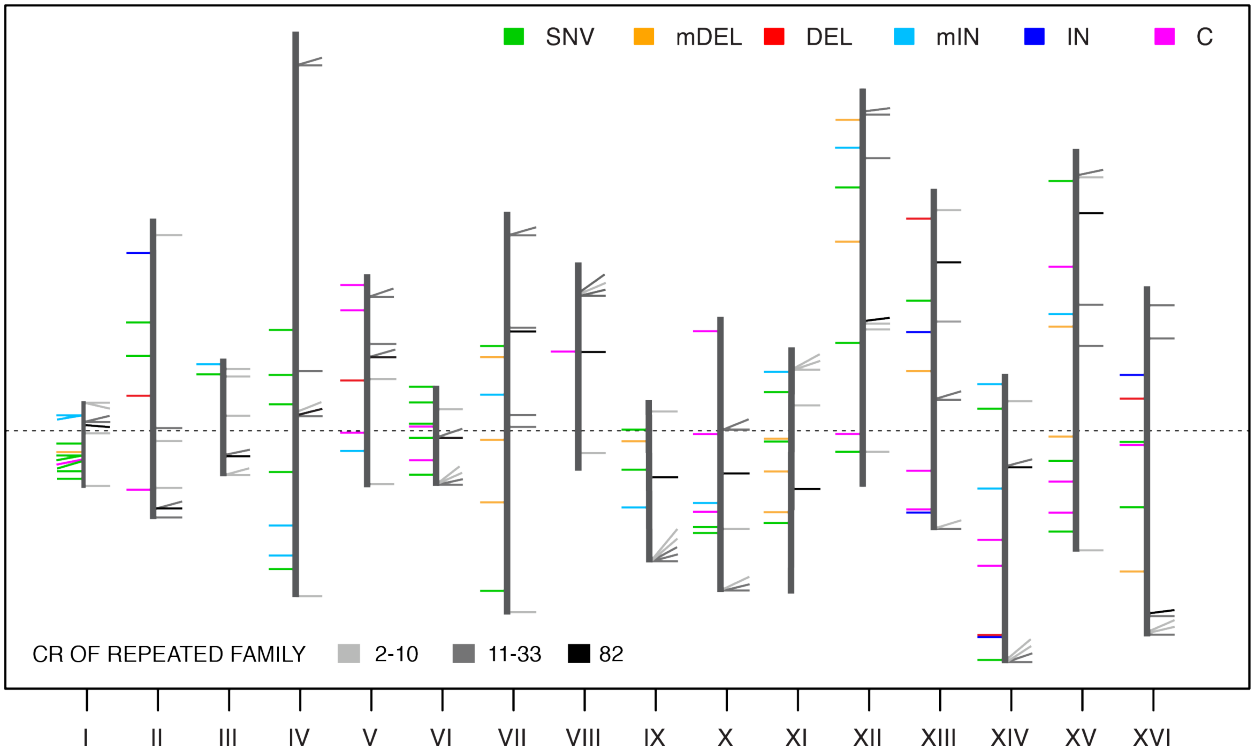


Figure 4 Assessment of the performance of the PMGL strategy via simulation experiments. The 16 nuclear chromosomes of the S288C RG are shown as vertical bars, with position 1 of each chromosome located at the bottom. The dashed black line crosses each chromosome's centromere. The left section of each chromosome presents the results for the random introduction of variants into the query genome. Color-coded lines indicate the position and type of variants that were unambiguously detected and characterized. The color code is shown at the top. SNV, single nucleotide variant; mDEL, micro-deletion; DEL, deletion; mIN, micro-insertion; IN, insertion; C, concatenated variants. The right section of each chromosome presents the results for the targeted introduction of variants into repeated regions of the reference genome. Color-coded lines indicate the copy number (CR) of the targeted family for variants that were unambiguously detected and characterized. An 82-copy region of the genome was targeted 16 times, one copy was chosen per chromosome. The color code is shown at the bottom.

We next performed a simulation experiment specifically targeting multiple copy regions of the S288C reference genome. The altered reference genome was used to construct the RGSL, and the unaltered reference genome was used to generate an artificial set of query genome sequence reads. This simulates the presence of discrepancies embedded in the reference genome. Regions of the reference genome present in multiple identical copies were selected. Each variant was introduced into one of the copies of a repeat family. Introduced variants included single nucleotide variants, micro-deletions, and micro-insertions; in some cases, these variants were concatenated. Using the zero-trail scan, the presence of 109out of 112 variants was detected exclusively at the exact copy of origin, and their nature was correctly resolved (File S4). Again, no false positive signals were generated.

For repeated regions of the genome, the generation of signatures of variation is clearly different if the variants are harbored in the query genome or in the reference genome (Figure 5). When harbored in the query genome, the location of the variant remains ambiguous but restricted to a specific position within each copy of the repeated family (Figure 5A). Furthermore, a zero-trail signature of variation is not generated, and the variant may only be detected using the signature value metric (see Discussion). In contrast, the incorporation of a discrepancy into a specific copy of the reference genome typically renders the affected copy locally unique and unable to attract any read strings. This generates a zero-trail signature of variation accompanied by a high signature value only at the affected copy (Figure 5B). In general, the presence of a zero-trail signature of variation within a repeated region directly indicates that no copies from the query genome contain the sequence specified by the reference genome at the corresponding site.

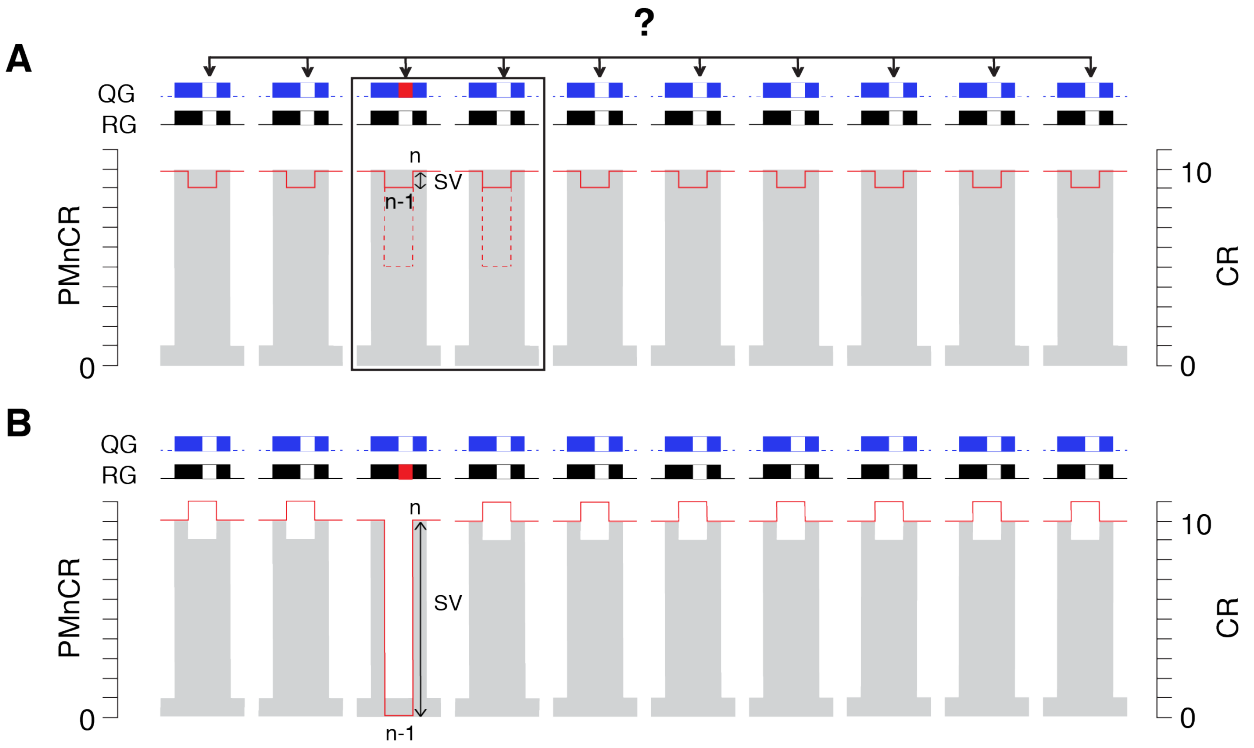


Figure 5 Impact on the PMGL for variants present in repeated regions of the query genome or the reference genome. A single nucleotide variant is harbored in a 10-copy region of either the query genome (A) or the reference genome (B). Ten copies of the repeated region, located at different positions along the genome, are shown. The query genome is represented in blue and the reference genome in black, highlighting the repeated regions with thick bars. The single nucleotide variant is shown as a red box in one of the copies, and the corresponding sites in the other copies are shown as white boxes. The normalized perfect match coverage (PMnCR) at each copy is plotted as a red line, and its relative scale is presented on the left. The corresponding copy number (CR) is shown as a gray shadow and its scale is presented on the right. SV indicates the relative magnitude of the signature value. When the variant is harbored in the query genome, the number of perfect matches decreases slightly and to the same extent in all of the copies. The signature value increases accordingly. The location of the variant remains ambiguous among the 10 copies (?). When the variant is harbored in one of the copies of the reference genome, a zero-trail signature of variation is typically generated, accompanied by a high signature value specifically at the copy of origin. The remaining copies present a slight increase in the PMnCR because their CR locally decreases by 1. The black rectangle in A) represents an alternative situation where a single nucleotide variant is located in a two-copy region of the query genome. In this case, the PMnCR decreases to about 50% at each copy (broken red line) and the signature value increases to about 2 (see Discussion). The actual position of the variant remains ambiguous but is now restricted to either of the two copies.

Generation of variation profiles of natural *S. cerevisiae* genomes. All natural genomes were analyzed using the automated PMGL pipeline. The reference genome of strain S288C, first constructed in 1996 (Goffeau *et al*., 1996), has been under continuous scrutiny since then and is considered to be of extremely good quality (Engel *et al*., 2014). Furthermore, the S288C strain and its derivatives, including BY4742, are amongst the most widely used yeast strains (Brachmann *et al*, 1998). Sequence reads from the S288C and BY4742 strains were processed to generate the corresponding read string datasets. In both cases, the RGSL structure was derived from the complete S288C reference genome. Including the analyses of both genomes (Figure 6 and File S5), a total of 153 different sites along the nuclear S288C reference genome present a zero-trail signature of variation. Only 9% of these signatures of variation were not solved. Most of the solved signatures of variation are present in both strains, and about half of these are harbored in the repeated compartment of the genome, suggesting that they represent discrepancies introduced into the reference genome sequence of strain S288C (see Figure 5B). A total of 168 individual variants were found, 163 of which were validated by customization. The remaining 5 variants, which correspond to big deletions, were detected but remained unsolved using the automated PMGL pipeline. Nevertheless, we were able to determine their position and length through direct inspection of the PMGL. Importantly, the genetic auxotrophies characteristic of theBY4742 genotype, comprising inactivating deletions at the *lys2*, *leu2, ura3*, and *his3* genes, were identified. In general, big deletions and big insertions are not directly solved by the automated PMGL pipeline (see File S1, Scanning of the PMGL, Interpretation and extension of alignments).

To further test the accuracy of the PMGL pipeline, 95regions showing signatures of variation in strain S288C were subjected to validation by PCR and Sanger sequencing. A good quality Sanger sequence was obtained for 90 such regions (57 unique and 33 repeated regions). In all cases, the Sanger sequence revealed the same variation(s) reported by the PMGL pipeline (File S5, examples are provided in File S2).

Recently, several *S. cerevisiae* strains have been assembled *de novo* utilizing state of the art methodologies including both short Illumina reads and long PacBio reads (Yue, *et al*, 2017). Using the original Illumina reads, we applied the automated PMGL pipeline to analyze strains SK1 and Y12 against their own assembled genome. We uncovered several variants that were not previously detected, 88 variants in strain SK1 (File S6) and 57 in strain Y12 (File S7). In both cases, most of the variants are contained in the repeated compartment of the genome, suggesting that such variants actually represent discrepancies introduced into the corresponding genome assemblies.

To test the scope of the PMGL strategy for generating variation profiles between more distantly related strains, we computed the number of signatures of variation reported by the automated PMGL pipeline using the S288C as reference genome and Illumina reads from strains SK1, Y12, and DBVPG6765 as query genomes (Yue, *et al*, 2017). A total of 54175, 47614, and 34028 zero-trail signatures of variation were generated for strains SK1, Y12, and DBVPG6765, respectively (Table 1). Furthermore, all of the signatures of variation present in strain SK1 were automatically analyzed to reveal the precise nature of the underlying variants; the great majority (96%) were solved and validated by customization (Table 1). As indicated in the data availability statement, the complete data sets for these analyses are available at GitHub.

Table 1. Variation profiles generated by the automated PMGL pipeline for *S. cerevisiae* strains SK1, Y12, and DBVPG6765 relative to the S288C reference genome.

	SK1		Y12	DBVPG6765
Chr	**nSV found**	**solved**	**nSV found**	**nSV found**
I	948	90 %	894	992
II	3539	97 %	2482	2995
III	1142	95 %	1444	522
IV	7109	96 %	6733	3293
V	2840	96 %	2520	1554
VI	1506	95 %	900	1103
VII	5153	96 %	4702	3097
VIII	2619	95 %	2368	1319
IX	1979	95 %	1811	1622
X	3084	97 %	2465	2375
XI	2822	96 %	2452	2577
XII	4296	97 %	3591	3353
XIII	4152	97 %	3441	2399
XIV	3315	97 %	3475	1724
XV	5009	97 %	4076	3475
XVI	4662	96 %	4260	1628
Total	54175	96 %	47614	34028

The number of signatures of variation (nSV) is reported for each strain and chromosome (Chr). For strain SK1, the percentage of solved signatures of variation is indicated.

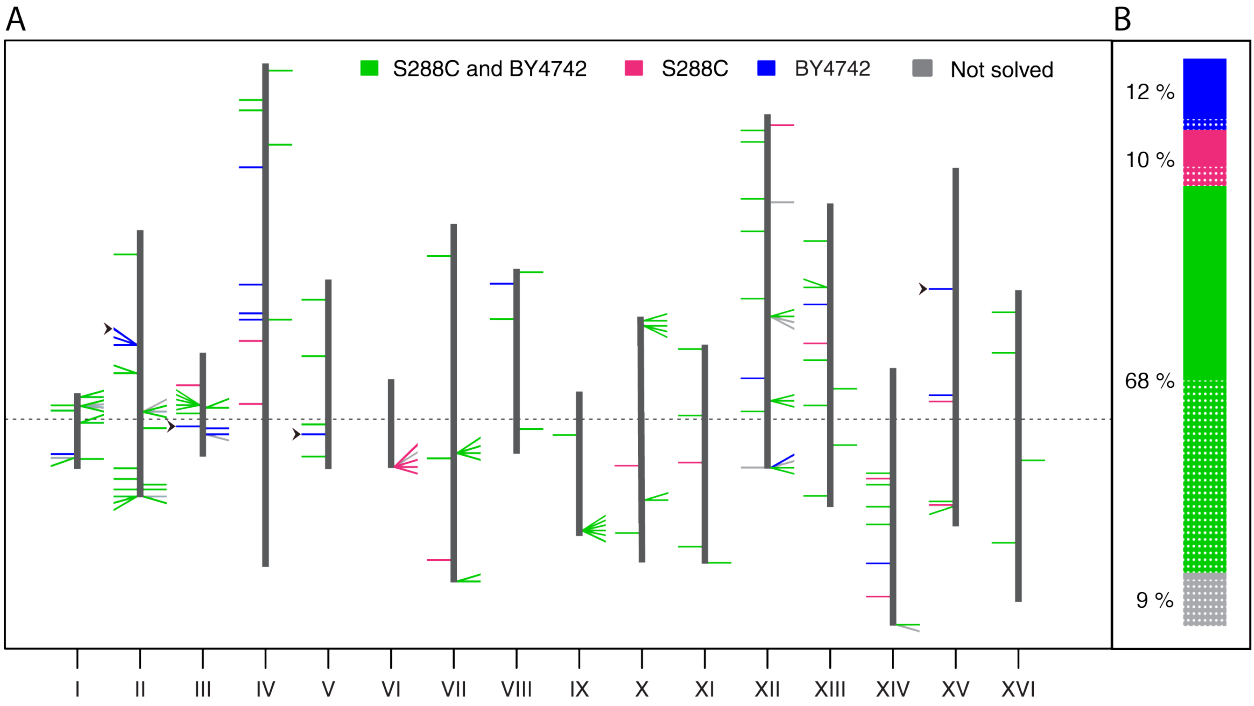


Figure 6 Analyses and comparison of genome variation profiles from yeast strains S288C and BY4742. A) The 16 nuclear chromosomes of the S288C reference genome are shown as vertical bars, with position 1 of each chromosome located at the bottom. The dashed black line crosses each chromosome's centromere. The positions of all signatures of variation are shown as lines. The left and right sections of each chromosome indicate the position of signatures of variation detected in unique or repeat regions of the genome, respectively. The color code at the top indicates whether signatures of variation are present in both genomes, are unique to either genome, or have not been solved. Arrowheads indicate the positions of the mutations underlying the auxotrophies characteristic of strain BY4742. B) Proportion of the different categories of signatures of variation. White dots indicate the fraction from each category that is harbored in repeated regions of the genome.

Analysis of a Synthetic Chromosome. The DNA sequence of synthetic chromosomes is designed *in silico*, modularly synthesized *in vitro*, and introduced *in vivo* through progressive exchange of segments of native chromosomes with the corresponding artificial segments. Living strains carrying hybrid genomes with both native and synthetic chromosomes are thus generated. Chromosome III of *S. cerevisiae* has been fully redesigned, and introduced into a living yeast to generate the HMSY011 strain (Annaluru et al, 2014). This strain has been previously analyzed to assess the degree of consistency between the as-designed sequence of synthetic chromosome III (synIII) and its physical counterpart. The result of this analysis has been previously reported and revealed 10 inconsistencies. A second version of the synIII sequence, termed the physical synIII sequence, was elaborated to match the sequence data analysis from the living strain by incorporating the corresponding changes (Annaluru et al, 2014).

We used the *S. cerevisiae* S288C reference genome as a structural scaffold for the targeted PMGL analysis of the synIII chromosome harbored in the HMSY011 strain (Figure 7A). We first exchanged chromosome three of the reference genome for the as-designed sequence of synIII and constructed the corresponding RGSL. Next, we generated a targeted RGSL that included reference strings derived from the synIII chromosome alone. We used this targeted RGSL to construct a targeted PMGL using the same whole genome sequence reads previously generated for analyzing strain HMSY011 and determining the synIII physical sequence (Annaluru et al 2014). We applied a zero-trail scan to reveal variation sites. We detected all 10 variation sites previously reported. The nature of these variants was resolved and perfectly matched the published description. Importantly, we detected the presence of 10 additional variants. These include 5 SNVs concatenated with a 2 nucleotide micro-insertion, 1 SNV found in close proximity to a previously reported missing loxPsym site, a 1-nucleotide micro-deletion, a 15-nucleotide deletion, and an 11-nucleotide insertion (Files S8 and S9).

We subsequently built a RGSL, a targeted RGSL and a targeted PMGL using the physical synIII sequence instead of the as-designed synIII sequence and performed a zero-trail scan with the same parameters as before. This revealed the disappearance of the signatures of variation corresponding to the previously reported inconsistencies, and the persistence of the signatures of variation corresponding to the newly identified ones. The latter confirmed the correct incorporation of the first ten changes, and showed the need to further modify the physical synIII sequence to obtain a more refined reference of the living synIII chromosome. Upon customization of the physical synIII sequence with our newly found variants, all signatures of variation disappeared, thus validating their precise location and nature.

The as-designed sequence of a synthetic chromosome constitutes the reference genome of the living chromosome. The highest copy number family in current synthetic yeast chromosomes corresponds to the artificially introduced loxPsym sites. These sites are 34 nucleotide long elements derived from phage P1. They allow recombination (Abremski and Hoess 1985) and reshuffling of genes (Dymond *et al*., 2011). SynIII has 198 identical loxPsym sites. To show that the PMGL strategy can unambiguously detect variants embedded in such extremely repeated regions of the reference genome we performed a series of simulation experiments (Figure 7B, File S10). We separately altered 27 copies of the loxPsym site in the as-designed synIII sequence by introducing the same variant, a deletion of the 5th nucleotide. Each altered genome was used to construct the RGSL and the unaltered genome was used to generate an artificial set of query

genome sequence reads. For each of the resulting PMGLs, only one of the 198 copies, precisely the one that was altered, produced the zero-trail signature of variation (File S10). In each case, the expected variant was found.

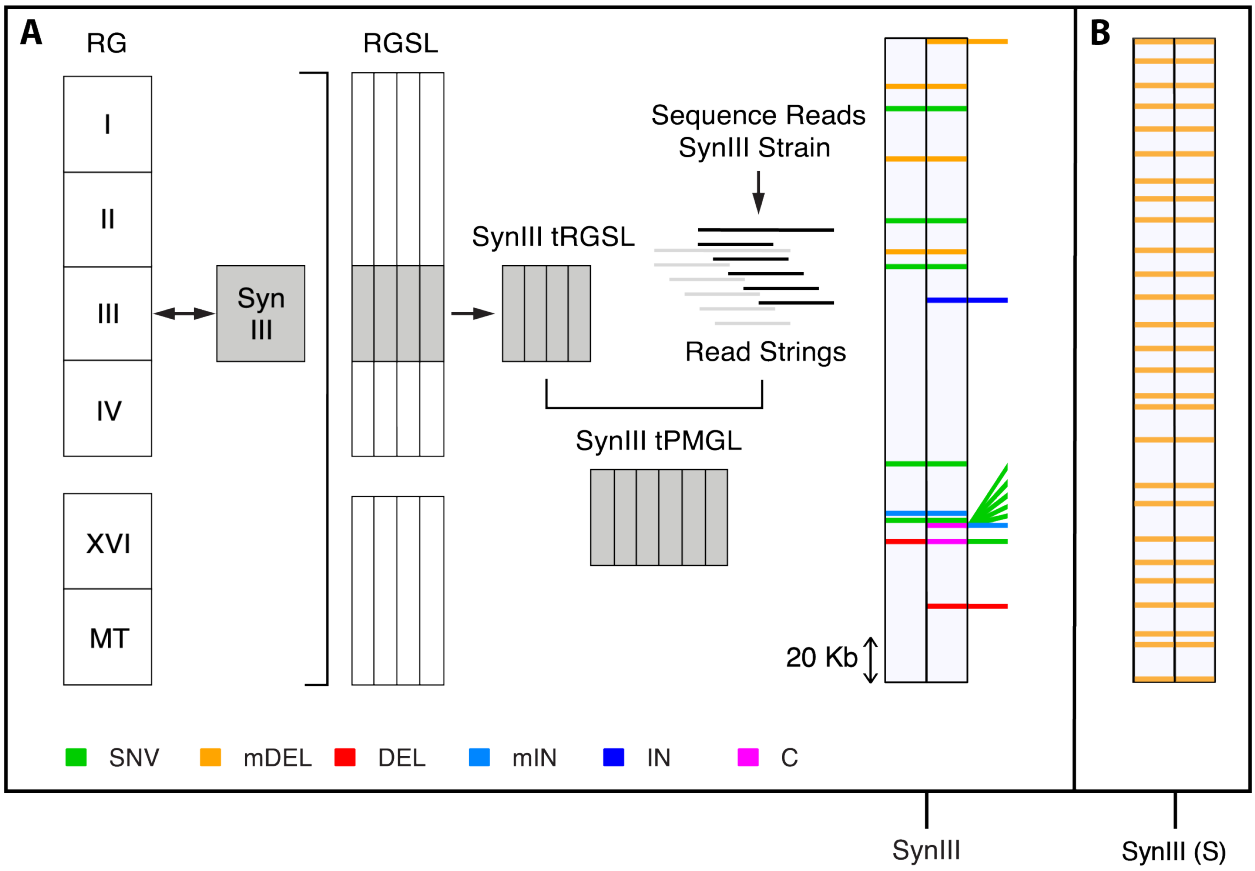


Figure 7 Analysis of a synthetic yeast chromosome. A) Targeted analysis of variation in synIII. The as-designed synIII sequence (gray shadow) is introduced into the S288C reference genome background, replacing native chromosome III. A comprehensive RGSL is generated. A targeted RGSL (tRGSL) is extracted and used to generate a targeted PMGL (tPMGL). Variants found in the living synIII chromosome relative to the as-designed synIII chromosome were revealed using the zero-trail scan (maximum normalized perfect match coverage at position n-1 = 5; the low complexity filter was not applied). The as-designed synIII chromosome is shown divided in two columns, with color-coded bars indicating the position of detected variants. Variants previously characterized are plotted on the first column. Variants characterized in this study are plotted on the second column. Variants found only with the PMGL strategy are projected to the right. The color code is shown at the bottom. SNV, single nucleotide variant; mDEL, micro-deletion; DEL, deletion; mIN, micro-insertion; IN, insertion; C, concatenated variants. B) Simulation targeting the loxPsym repeat family of synIII. The results of the 27 independent simulations are consolidated into one scheme of the synIII chromosome. The first column indicates the position of the variants introduced; the second column indicates the position of the variants found and characterized.

DISCUSSION

The PMGL strategy for the analysis of genome variation is based on the identification of signatures of variation within an identity landscape. Most interestingly, positional information on variants at the single nucleotide level is obtained prior to the generation of highly targeted alignments between the query genome and the reference genome. Importantly, the PMGL strategy allows the introgression of discovered variants into the initial reference genome sequence, creating a new reference genome over which the perfect match computation should produce a locally uniform landscape. This simple test, which provides a categorical validation for each uncovered variant, is not possible with probability based programs which always report a weighted answer.

This study directly interrogates haploid yeast genomes. Query genomes and reference genomes have been explored using the zero-trail signature of variation while simultaneously determining the signature value at the corresponding variation sites. The precise location of different types of variants has been unambiguously revealed in the unique compartment of query genomes and reference genomes and in the repeated compartment of reference genomes. In fact, both signatures appear simultaneously in these contexts, and the signature value is typically very high. The zero-trail signature of variation, however, cannot detect the presence of amplifications, of variants that occur in multiple copy regions of the query genome, or of heterozygous variants that occur in diploid genomes. In contrast, the signature value can manifest itself in all of these situations. We envision that the signature value could be used as a general signature of variation, potentially driving most types of analyses of sequenced genomes.

In theory, variants embedded in multiple copy regions of the query genome would produce signature values between 2 and 1, where 1 is the inferior limit if copies are increased to infinity. Consequently, high copy regions of the query genome may not be interrogated because the signature value would not deviate significantly from the baseline. The transition to analyzing diploid genomes should be possible, as it should not represent a major conceptual change. What would change significantly is the distribution of signature values at variation sites. In the haploid case, most variation sites generate high signature values. In contrast, for the diploid case, high signature values would be reserved for homozygous variants present in unique regions of the genome. Sites with signature values between 2 and 1 would harbor homozygous variants present in repeated regions of the genome or heterozygous variants. It will be important to assess the resolving power of the signature value metric in this context.

We have tested the performance of the PMGL strategy in a variety of ways, ranging from the determination of variation profiles of simulated query genomes and reference genomes to the genome-wide analyses of natural yeast strains and the targeted analysis of a synthetic yeast chromosome. For these experiments, we have utilized data sets generated with state-of-the-art experimental and bioinformatics methodologies. Most importantly, the PMGL strategy has contributed with the identification of novel variants in these contexts.

We have shown that high-quality reference genomes can be further refined using the PMGL strategy. Of particular relevance is the possibility to unambiguously detect discrepancies incorporated into the reference genome within identical repeats of any copy number. In most

aligners, because query genome to reference genome differences must be tolerated to some extent, all copies from a repeated family compete to attract the same subset of sequence reads. In contrast, in the PMGL strategy, if a reference genome copy harbors any discrepancy that renders it unique, that copy will not attract query genome sequences, generating a signature of variation. This important property of the PMGL strategy could be particularly useful for the challenging task of improving *de novo* genome assemblies, which provide foundational resources for future research in different organisms. In a broader context, any assembled genome, either generated *de novo* or through re-sequencing, can be conceptualized as a new reference genome and refined with the PMGL strategy, using as query genome the same set of sequence reads utilized for its assembly.

The development of highly accurate and versatile strategies for analyzing genomes is particularly relevant for the scientific revolution that is already underway: that of expanding our understanding of biology through the synthesis of entire genomes. The Genome Project-write is leading this endeavor (Boeke *et al.,* 2016), and the Sc2.0 project currently represents the largest genome synthesis initiative (Richardson *et al*., 2017). Actually, in addition to synIII (Annaluru *et al*., 2014), several synthetic yeast chromosomes have been recently completed: synII (Shen *et al*., 2017), synV (Xie *et al*., 2017), synVI (Mitchell *et al*., 2017), synX (Wu *et al*., 2017), and synXII (Zhang *et al*., 2017). Confronting designed sequences with their living representations is clearly of utmost importance. We have provided a proof of principle of the power of the PMGL strategy in this context by reanalyzing and refining the reported physical sequence of Sc2.0 synthetic chromosome III.

The perfect match logic orchestrates simple procedures in such a way that genome variation can be discovered under a different premise: that of an identity landscape containing precise information about variation. We have consolidated this concept in the PMGL pipeline to provide a unified framework for the analysis of sequenced genomes. To reach its maximum potential, this framework should be further developed, notably by exploiting the full potentiality of the signature value metric.

ACKNOWLEDGEMENTS

Kim Palacios Flores is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship392315from CONACYT. We thank Julio Collado and Daniel Piñero for helpful discussions; Delfino García for technical assistance; Leslie Mitchell for providing the raw whole genome sequence reads from the synIII strain (HMSY011); Eglee Lomelin for assistance with the artwork; the Laboratorio Nacional de Visualización Científica Avanzada (LAVIS) for providing supercomputer services; and the Instituto Nacional de Medicina Genómica (INMEGEN) for performing the Illumina sequence of strains S288c and BY4742. This work was supported in part by grant UNAM-PAPIIT(IA207817).

LITERATURE CITED

Abrahams, E., and S. L. Eck, 2016 Molecular medicine: Precision oncology is not an illusion. *Nature* 539: 357.

Abremski K, and R. Hoess, 1985 Phage P1 Cre-*lox*P Site-specific Recombination. Effects of DNA Supercoiling on Catenation and Knotting of Recombinant Products. J. Mol.Biol. 184: 211-220

Annaluru, N., H. Muller, L. A. Mitchell, S. Ramalingam, G. Stracquadanio *et al.,* 2014 Total Synthesis of a Functional Designer Eukaryotic Chromosome. Science 344: 55–58.

Audano, P., S. Ravishankar, and F. Vannberg, 2017 Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. Bioinformatics, advance access publication.

Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland *et al.*, 1997 The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1462.

Boeke, J. D., G. Church, A. Hessel, N. J. Kelley, A. Arkin *et al.*, 2016 The Genome Project-Write. Science 353: 126-127.

Brachmann, C. B., A. Davies, G. J. Cost, E. Caputo, J. Li *et al.,*1998 Designer Deletion Strains derived from *Saccharomyces cerevisiae* S288C: a Useful set of Strains and Plasmids for PCR-mediated Gene Disruption and Other Applications. Yeast 14: 115–132.

Chaisson, M. J., and G. Tesler, 2012 Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMCBioinformatics 13: 238.

Dymond, J. S., S. M. Richardson, C. E. Coombes, T. Babatz, H. Muller *et al.,*2011 Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. Nature 477: 471–476.

Edgar, R., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Engel, S.R., F. S. Dietrich, D. G. Fisk, G. Binkley, R. Balakrishnan *et al.,* 2014 The Reference Genome Sequence of *Saccharomyces cerevisiae*: Then and Now. G3: Genes, Genomes, Genetics 4: 389–398.

1000 Genomes Project Consortium, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.*, 1996 Life with 6000 Genes. Science 274: 546–567.

Goodwin, S., J. D. McPherson, and W. R. McCombie, 2016 Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17: 333-351.

Holt, J., and L. McMillan, 2014 Merging of multi-string BWTs with applications. Bioinformatics 30: 3524-3531.

International Human Genome Sequencing Consortium., 2004 Finishing the euchromatic sequence of the human genome. Nature 431, 931-945

Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan et al., 2012 VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22: 568-576.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway et al., 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12.
Kurtz, S., 2003 The Vmatch large scale sequence analysis software. Ref Type: Computer Program 412.
Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357-359.
Langmead, B., C. Trapneli, M. Pop, and S.L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.
Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987-2993.
Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.
Li, R. Q., Y. R. Li, K. Kristiansen, and J. Wang, 2008 a SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713-714.
Li ; H., J. Ruan, and R. Durbin, 2008 b Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18: 1851-1858.
Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27: 764–770.
Mardis, E. R., 2013 Next-Generation Sequencing Platforms. Annu. Rev. Anal. Chem. 6: 287-303.
McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis et al., 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297-1303.
Metzker, M. L., 2010 Sequencing technologies---the next generation. Nat. Rev. Genet. 11: 31-46.
Mitchell, L. A., A. Wang, G. Stracquadanio, Z. Kuang, X. Y. Wang et al., 2017 Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond. Science 355: eaaf4831
Park, P. J., 2009 ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. 10: 669-680.
Pfeifer, S. P., 2017 From next-generation resequencing reads to a high-quality variant data set. Heredity 118: 111-124.
Reinert, K., B. Langmead, D. Weese, and D. J. Evers, 2015 Alignment of Next-Generation Sequencing Reads. Annu. Rev. Genomics Hum. Genet. 16: 133-151.
Reyes, J., L. Gómez-Romero, X. Ibarra-Soria, K. Palacios-Flores, L. R. Arriola et al., 2011 Context-dependent individualization of nucleotides and virtual genomic hybridization allow the precise location of human SNPs. Proc. Natl. Acad. Sci. USA 108: 15294-15299.
Richardson, S. M., L. A. Mitchell, G. Stracquadanio, K. Yang, J. S. Dymond et al., 2017 Design of a synthetic yeast genome. Science 355: 1040-1044.
Rimmer, A. H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg et al., 2014 Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 46: 912-918.
Schatz, M. C., A. M. Phillippy, D. D. Sommer, A. L. Delcher, D. Puiu et al., 2013 Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. Brief.

Bioinform. 14: 213–224
Schbath, S., V. Martin, M. Zytnicki, J. Fayole, V. Loux et al., 2012 Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. J. Comput. Biol. 19: 796-813.
Shen, Y., Y. Wang, T. Chen, F. Gao, J. H. Gong et al., 2017 Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. Science 355: eaaf4791.
Teer, J. K., and J. C. Mullikin, 2010 Exome sequencing: the sweet spot before whole genomes. Hum. Mol. Genet. 19: R145-R151.
Tenaillon, O., J. E. Barrick, N. Ribeck, D. E. Deatherage, J. L. Blanchard et al., 2016 Tempo and mode of genome evolution in a 50,000-generation experiment. Nature 536: 165-170.
Wang, Z., M. Gerstein, and M. Snyder, 2009 RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10: 57-63.
Wu, Y., B. Z. Li, M. Zhao, L. A. Mitchell, Z. X. Xie et al., 2017 Bug mapping and fitness testing of chemically synthesized chromosome X. Science 355: eaaf4706.
Yang, X., S. P. Chockalingam, and S. Aluru, 2013 A survey of error-correction methods for next-generation sequencing. Brief. Bioinform. 14: 56-66.
Yue, J.X., J. Li, L. Algrain, J. Hallin, K. Persson, et al., 2017 Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat. Genet. 49: 913-924.
Xie, Z. X., B. Z. Li, L. A. Mitchell, Y. Wu, X. Qi et al., 2017 "Perfect" designer chromosome V and behavior of a ring derivative. Science 355: eaaf4704.
Zerbino, D. R. and E. Birney, 2008 Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821-829.
Zhang, W., Zhao, G., Luo, Z., Lin, Y., Wang, L., et al., 2017 Engineering the ribosomal DNA in a megabase synthetic chromosome. Science355: eaaf3981.

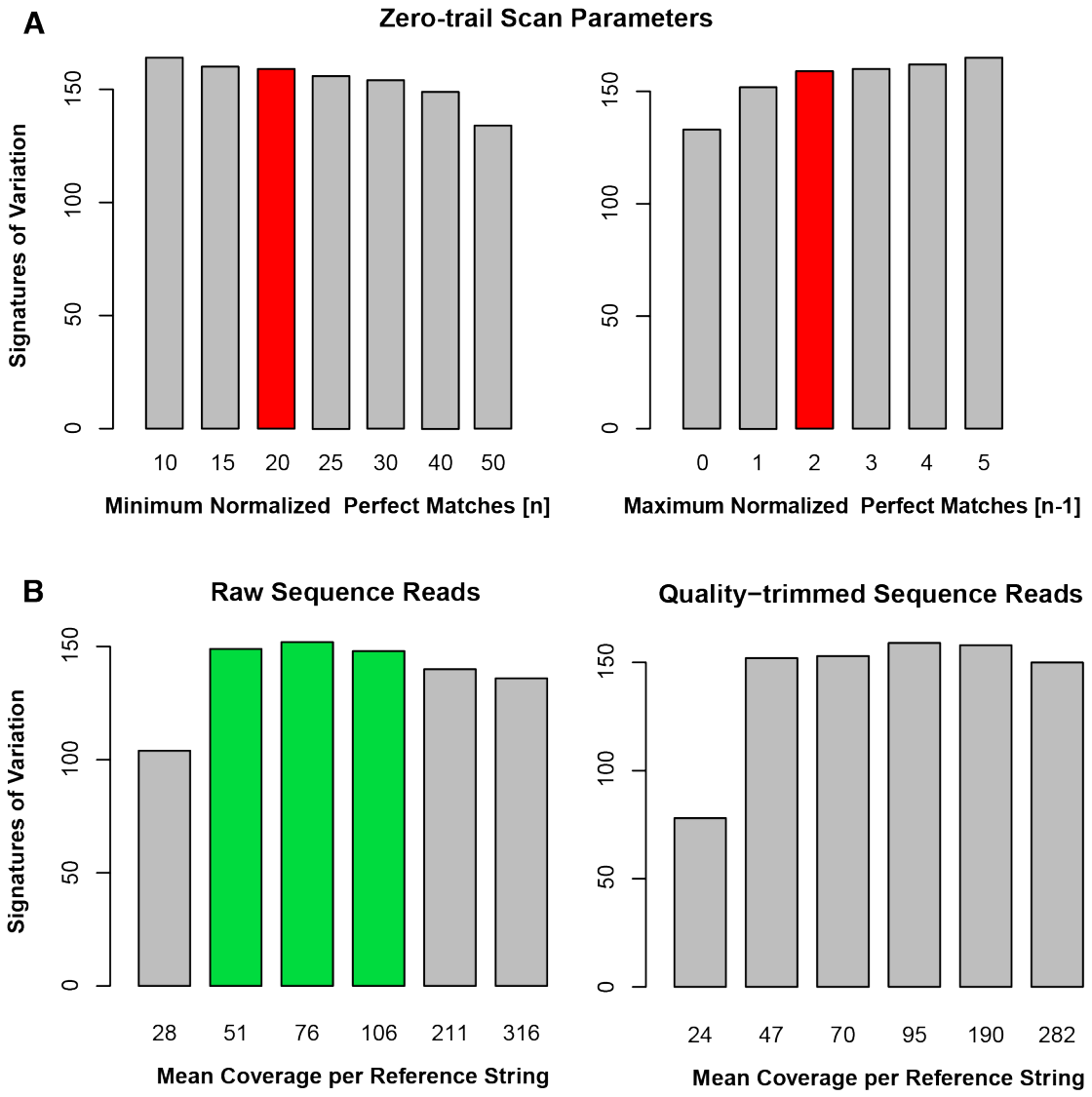


Figure S1 Calibration of the PMGL zero-trail scan parameters and sequencing coverage. A) Normalized perfect match coverage at the zero-trail signature of variation. Using the same dataset (sequenced at about 100X), the PMGL zero-trail scan was executed by independently varying the parameter values for normalized perfect match coverage at positions n and n-1. The number of recovered signatures of variation is plotted. The normalized perfect match coverage at position n represents the minimum coverage accepted in the absence of variation. The normalized perfect match coverage at position n-1 represents the maximum coverage accepted as a near-0 value. The values used throughout this study (unless otherwise specified) for both parameters are indicated with red bars. B) Impact of sequencing coverage and quality-based trimming of sequence reads on the recovery of signatures of variation. Using the parameter values highlighted in A, the PMGL zero-trail scan was implemented for varying proportions of sequence reads derived from the same dataset. Sequence reads were either trimmed or not trimmed, as indicated. Note that the trimming of sequence reads decreases the mean coverage per reference string by about 10%. The number of recovered signatures of variation is plotted. The green bars indicate the optimal raw read sequencing coverage for the PMGL scan using the parameter values highlighted in A.

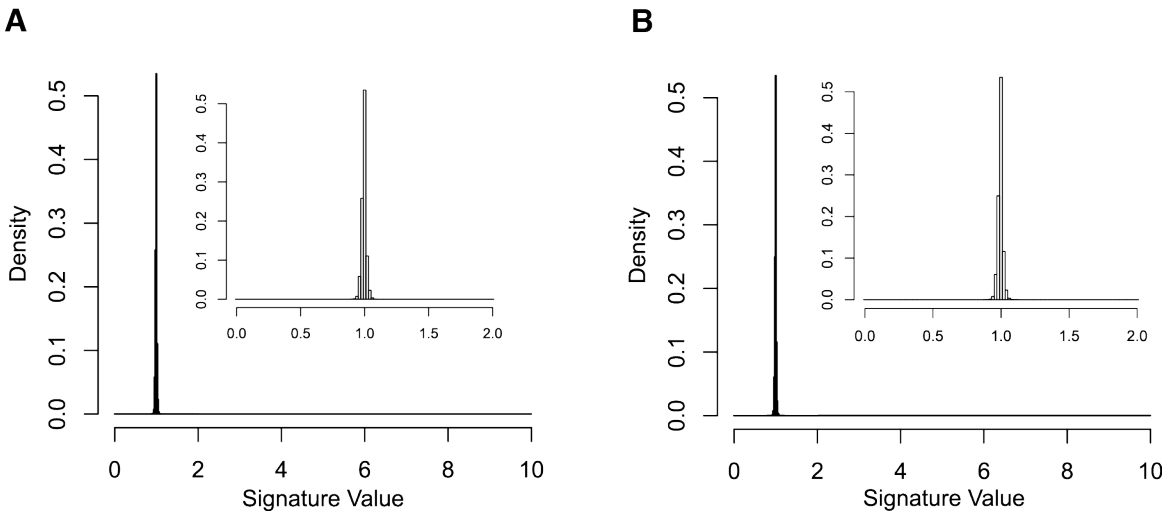


Figure S2 Genome-wide signature value distribution for strains S288C (A) and BY4742 (B). The distribution of signature values derived from the corresponding PMGL is presented. Reference strings that could generate artificial signature values were eliminated (see File S1, Genome-wide distribution of signature values). Signature values up to 10 are shown. The proportion of reference strings associated with higher signature values are not perceptible on the y-axis. An inset centered at 1 is shown for each strain.

Table S1. Random simulation in the query genome introducing random errors into artificial sequence reads.

E	R	rSV	nSV
0.0	12	100	0
0.1	12	100	0
0.3	12	99	0
1.0	12	99	0
2.0	12	90	0
3.0	12	52	0
3.0	24	75	0

E, percentage of errors introduced; R, number of artificial sequence reads used in millions; rSV, percentage of recovered signatures of variation; nSV, number of new signatures of variation generated.

File S1. Detailed description of Materials and Methods.

***S. cerevisiae* strains and culture.** Strain S288C was obtained from the American Type Culture Collection, ATCC. Strain BY4742 was obtained from the laboratory of Bernard Dujon at the Pasteur Institute, Paris, France. The strains were stored at -80°C. Both strains were cultured in rich medium (YPD) at 30°C with agitation (250 rpm).

DNA isolation and Illumina sequencing. DNA was isolated from fresh cultures using the Yeast DNA Extraction Kit from Thermo Scientific. Sequencing libraries were prepared using the TruSeq PCR-Free sample preparation Kit from Illumina. Sequencing was performed in a Next Seq 500 sequencer using the Next Seq 500/550 Mid Output Reagent, 300 cycles. Sequencing was performed at about 100X coverage.

Generation of PCR products and Sanger Sequencing. For each region to be analyzed, a PCR product was obtained using primers located at unique sequences in the genome. PCR was performed using the Verity Thermal Cycler from Applied Biosystems. The PCR protocol was: 94°C for 1 min., 34 cycles (94°C for 30 sec., 58°C for 30 sec., 68°C for 3 min.), and 72°C for 10 min. The PCR products were sequenced in Macrogen (South Korea) using appropriate primers flanking the expected region containing the variant(s).

Reference genomes and query genomes. The S288C reference genome was downloaded from NCBI and corresponds to assembly R64. The synIII as-designed reference sequence version 3.3_41 was downloaded from GenBank using the accession number KJ463385. The actual physical sequence of synIII present in the strain HMSY011, sequence version 3.3_42, was downloaded from GenBank using the accession number KC880027. Each of these synIII sequences were introduced into the S288C reference genome background by exchanging the natural chromosome III sequence with the as-designed or the physical synIII sequence, thus generating the corresponding synIII reference genomes. The SK1 and Y12 assembled genomes were downloaded from https://yjx1217.github.io/Yeast_PacBio_2016/data/. Query genomes generated in this study correspond to raw whole genome sequence reads derived from strains S288C and BY4742. Sequence reads for strain HMSY011 were provided by Leslie Mitchell and are the same as those used in the previous analysis of this strain (Annaluru et al 2014). Sequence reads for strains SK1, Y12, and DBVPG6765 were downloaded from Short Reads Archive (SRA) under accession code PRJNA340312.

Automated PMGL pipeline:

1) **Generation of the RGSL**. A fasta file is generated for each chromosome in the reference genome (chromosome fasta file). The header line of each chromosome fasta file indicates the RGSL ID and the chromosome number (e.g. >RGSL1_01). To generate the reference strings dataset, each chromosome fasta file is used to generate a fasta file containing all kmers (25-mers in this study) in order of appearance in the chromosome (kmers fasta file) and separated by 1 nucleotide from each other. The header line preceding each kmer indicates the RGSL ID, the chromosome, and the starting position of the kmer within the chromosome (e.g. >RGSL1_01_1) and corresponds to the ID column of the final RGSL structure. All chromosome fasta files are concatenated into a single fasta file (reference

genome fasta file) and the bowtie build command from bowtie-0.12.7 (Langmead et al 2009) is executed to generate a binary database comprising several index files (bowtie database). For each kmer fasta file, the bowtie command (Langmead et al 2009) is executed with the -v alignment mode parameter set to 0 and specifying –a to report all perfect-match occurrences of each kmer in the bowtie database. The bowtie output for each chromosome is used to construct a chromosome RGSL. The latter involves counting the number of occurrences of each ID (which will correspond to the CR column), constructing the IDF column (where each ID belonging to the same family is reported by order of appearance within and between chromosomes), ordering by ID, and generating the SEQ column by always reporting the forward kmer sequence (which corresponds to the sequence orientation in the reference genome). All chromosome RGSLs are concatenated to generate the final RGSL.

2) **Generation of the PMGL.** To generate the read strings dataset, the jellyfish count command from jellyfish-1.1.10 (Marçais and Kingsford 2011) is executed on the fastq file containing the sequence reads from the query genome by setting the -m parameter to k (25 in this study) and specifying --both-strands to generate a binary database (jellyfish database) of kmer counts combining the forward and reverse orientations. The kmer-cov-plot script from the AMOS repository (Schatz et al 2013) is executed on each chromosome fasta file by setting the --jellyfish and -s options and providing the jellyfish database. The latter uses the kmer counts from the jellyfish hash table to obtain the kmer coverage beginning at each base in the chromosome. The kmer-cov-plot output for each chromosome is used in conjunction with the corresponding chromosome RGSL to report the perfect match coverage associated with each position in the chromosome by directly assigning the computed kmer coverage to the corresponding ID. The resulting files are concatenated to report the ID, CR, SEQ, and PM columns of the genome-wide PMGL. The normalized perfect match coverage associated with each reference string, and corresponding to the PMnCR column of the PMGL, is obtained by dividing the values in the PM column by the corresponding values in the CR column, row by row. The SV column of the PMGL is obtained by dividing the PMnCR value at each successive position n by the PMnCR value at position n-1, where n ranges from the positions associated with the second and last reference strings of each chromosome. To avoid dividing by zero, a single count is previously added on the fly to both the PMnCR value at position n and the PMnCR value at position n-1.

3) **Scanning of the PMGL.** All signatures of variation are localized using the zero-trail scan. The complete PMnCR and CR columns of the PMGL are evaluated at positions n and n-1, where n ranges from the positions associated with the second and last reference strings of each chromosome. Zero-trail signatures of variation are defined as sites having a selected minimum value in the PMnCR column at position n, a selected maximum value in the PMnCR column at position n-1, and a value of 1 in the CR column at position n-1. The value in the CR column at position n is allowed to vary. Zero-trail signatures of variation with a value greater than 1 in the CR column at position n are either immediately flanking or embedded within a repeated region of the reference genome. Zero-trail signatures of variation containing less than 16 consecutive 0 or near-

zero values can be eliminated. Zero-trail signatures of variation associated with a low complexity downstream recovery string can be eliminated. In this pipeline, we defined low complexity strings as kmers containing equal to or more than 60% of a single nucleotide and / or less than 3 different types of nucleotides. In the case of big deletions, the zero-trail signature of variation is sometimes interrupted by reference strings that are repeated in other regions of the genome. These strings attract perfect matches derived from the other repeats, generating nested zero-trail signatures of variation within the deleted region. Direct inspection of the PMGL is necessary to infer the actual length of the deletion. In this study, the genomes of strains S288C, BY4742, SK1, Y12, and DBVPG6765 were analyzed using the following PMGL scan parameters: a minimum normalized perfect match coverage of 20 at position n, a maximum normalized perfect match coverage of 2 at position n-1, a CR value of 1 at position n-1, and both the zero-trail length filter and the low complexity filter were applied.

4) **Generation of the first alignment at each signature of variation.** For each signature of variation, the corresponding downstream recovery string is used to retrieve query genome sequence reads that contain the exact string's sequence in either the forward or reverse complementary orientation. Sequence reads containing at least 25 nucleotides upstream of the downstream recovery string are selected and cut, defining a group or groups of identical query genome sequences (read families). In turn, a section of the reference genome including the downstream recovery string plus 25 nucleotides upstream is extracted. The sequence of each read family is aligned with the corresponding section of the reference genome using the MUSCLE Multiple Sequence Alignment tool (Edgar 2004). When more than one read family is generated in this module or in module 5, the latter finds the query genome sequence derived from the variation site.

5) **Interpretation and extension of alignments.** All alignments generated for each signature of variation in module 4 enter an iterative process. At each iteration, signatures of variation are classified as solved, unsolved, or transiting. Solved and unsolved signatures of variation exit the iterative process. Solved signatures of variation generate alignments with the following characteristics: 1) all alignments at the current iteration share a common variation motif relative to the corresponding region of the reference genome, 2) a single alignment contains the query genome sequence that differs from the reference genome only at the common variation motif, and 3) the latter alignment is identical to the reference genome both upstream and downstream of the variation zone. Signatures of variation can be classified as unsolved at any iteration if no read families are generated or if an excess of read families is generated. All of the alignments corresponding to signatures of variation classified as transiting can be extended in the next iteration. For each alignment extension, a kmer located further upstream in the read family sequence (but overlapping with the kmer used in the previous iteration), is used to select a new set of sequence reads that perfectly contain it, defining a new set of read families. Each read family is used to construct an extended query genome sequence, which is aligned with the corresponding extended region of the reference genome. For each solved signature of variation, the specific nature and position of the variant(s) (common variation motif) is determined by its corresponding final alignment. This module does not solve big deletions, big insertions or complex rearrangements. Unsolved

variants, however, are precisely located along the reference genome and other methods can be used to reconstruct their sequence. For example, searching for sequence reads that contain both the downstream recovery string and the upstream recovery string could reveal the breakpoints of deletions. Searching for sequence reads that contain either the upstream recovery string or the downstream recovery string could reveal the breakpoints of insertions. In this study, the genomes of strains S288C (as compared to the S288C reference genome), BY4742 (as compared to the S288C reference genome), SK1 (as compared to the SK1 assembled genome), and Y12 (as compared to the Y12 assembled genome) were analyzed using the following alignment interpretation and extension parameters: a maximum of 4 iterations, a minimum of 20 query genome sequence reads containing each read family sequence, a maximum of 10 read families per growing query genome sequence, and a minimum of 20 perfectly matching nucleotides upstream of the variation zone in the final alignment. Furthermore, the genomes of strains SK1 (as compared to the S288C reference genome), Y12 (as compared to the S288C reference genome), and DBVPG6765 (as compared to the S288C reference genome) were analyzed using the same alignment interpretation and extension parameters as before, except for the maximum number of read families per growing query genome sequence, which was set to 3.

6) **Generation of a customized RG.** The final step of the PMGL strategy consists in customizing the reference genome to validate the variants uncovered for each signature of variation. All variants solved in step 5 are introduced into the corresponding positions of the reference genome sequence. Importantly, the variants are introduced into each chromosome in a backward direction, starting at the furthest downstream variant. After reference genome customization, steps 1, 2, and 3 of the PMGL pipeline must be repeated using the customized reference genome and the original query genome sequence reads. The disappearance of previously detected signatures of variation confirms the precise location and nature of the variants.

Genome-wide distribution of signature values. The genome wide distribution of signature values is derived from the SV column of the PMGL. Signature values corresponding to reference strings with sequences that can generate sharp artificial increases in the number of perfect matches obtained were eliminated from this analysis. These include reference strings from the S288C reference genome located on chromosome 12 from position 451418 to position 468905, which largely correspond to the RDN1 locus. The actual copy number of this rDNA region has been collapsed in the S288C reference genome. Reference strings located elsewhere in the genome but sharing the same 25-mer sequences with the latter set of strings were also eliminated. Likewise, mitochondrial DNA reference strings and reference strings sharing the same 25-mer sequences were also eliminated. In general, regions that are collapsed in the reference genome or regions that are present in multi-copy replicons should be excluded, along with reference strings located elsewhere in the genome but sharing the same sequence with the latter.

Random simulation in the query genome. A total of 100 variants were introduced into the sequence of the S288C reference genome. The nature of each single nucleotide variant, the size

of each micro-deletion, and the nature and the size of each micro-insertion were selected at random. The size of each deletion and of each insertion was selected at random from pre-established ranges. All insertions were derived from the lambda phage sequence. After determining the nature and / or size of each variant, the positions for their introduction into the reference genome were also selected at random. Variants were introduced into the reference genome in a backward direction, starting at the furthest downstream position in each chromosome. The altered reference genome was used to construct an artificial set of query genome sequence reads by randomly generating 12,000,000 fragments of 100 nucleotides each. These fragments were generated using a custom script. The RGSL was derived from the unaltered reference genome.

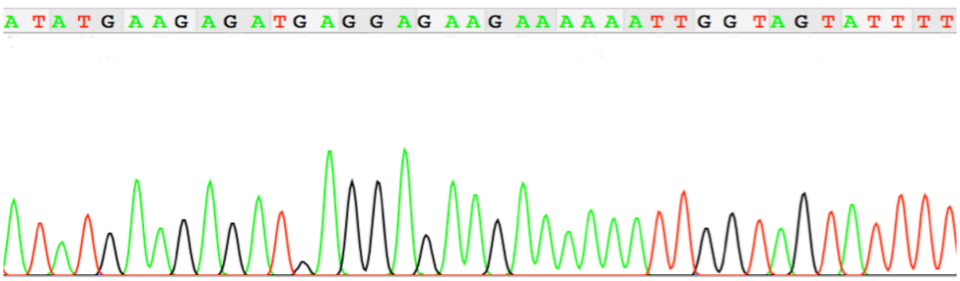
Directed simulation in repeated regions of the reference genome. A total of 112 variants were introduced into the S288C reference genome sequence. The positions for the introduction of variants were pre-determined by searching for repeats in the reference genome composed of different copy numbers, ranging from 2 to 82. Each variant was assigned to a specific position by a random procedure. Variants were introduced into the reference genome in a backward direction, starting at the furthest downstream position in each chromosome. A RGSL was generated using the altered reference genome. The unaltered reference genome was used to construct an artificial set of query genome sequence reads by randomly generating 12,000,000 fragments of 100 nucleotides each. These fragments were generated using a custom script.

File S2. Examples of variant location and characterization. The ID corresponding to the Downstream Recovery String is indicated, followed by the local PMGL, the alignment, and the Sanger sequence. ID, Reference String's unique identifier; CR, count reference; SEQ, nucleotide sequence; PM, perfect matches; PMnCR, perfect matches normalized to count reference; SV, signature value.

ID: RGSL37_02_88490

ID	CR	SEQ	PM	PMnCR	SV
RGSL37_02_88463	1	TAAGACTATATGAAGAGATGAGGAG	94	94.00	1.00
RGSL37_02_88464	1	AAGACTATATGAAGAGATGAGGAGA	95	95.00	1.01
RGSL37_02_88465	1	AGACTATATGAAGAGATGAGGAGAA	91	91.00	0.96
RGSL37_02_88466	1	GACTATATGAAGAGATGAGGAGAAG	90	90.00	0.99
RGSL37_02_88467	1	ACTATATGAAGAGATGAGGAGAAGA	88	88.00	0.98
RGSL37_02_88468	1	CTATATGAAGAGATGAGGAGAAGAG	0	0.00	0.01
RGSL37_02_88469	1	TATATGAAGAGATGAGGAGAAGAGA	0	0.00	1.00
RGSL37_02_88470	1	ATATGAAGAGATGAGGAGAAGAGAA	0	0.00	1.00
RGSL37_02_88471	1	TATGAAGAGATGAGGAGAAGAGAAG	0	0.00	1.00
RGSL37_02_88472	1	ATGAAGAGATGAGGAGAAGAGAAGA	0	0.00	1.00
RGSL37_02_88473	1	TGAAGAGATGAGGAGAAGAGAAGAA	0	0.00	1.00
RGSL37_02_88474	1	GAAGAGATGAGGAGAAGAGAAGAAA	0	0.00	1.00
RGSL37_02_88475	1	AAGAGATGAGGAGAAGAGAAGAAAA	0	0.00	1.00
RGSL37_02_88476	1	AGAGATGAGGAGAAGAGAAGAAAAA	0	0.00	1.00
RGSL37_02_88477	1	GAGATGAGGAGAAGAGAAGAAAAAA	0	0.00	1.00
RGSL37_02_88478	1	AGATGAGGAGAAGAGAAGAAAAAAT	0	0.00	1.00
RGSL37_02_88479	1	GATGAGGAGAAGAGAAGAAAAAATT	0	0.00	1.00
RGSL37_02_88480	1	ATGAGGAGAAGAGAAGAAAAAATTG	0	0.00	1.00
RGSL37_02_88481	1	TGAGGAGAAGAGAAGAAAAAATTGG	0	0.00	1.00
RGSL37_02_88482	1	GAGGAGAAGAGAAGAAAAAATTGGT	0	0.00	1.00
RGSL37_02_88483	1	AGGAGAAGAGAAGAAAAAATTGGTA	0	0.00	1.00
RGSL37_02_88484	1	GGAGAAGAGAAGAAAAAATTGGTAG	0	0.00	1.00
RGSL37_02_88485	1	GAGAAGAGAAGAAAAAATTGGTAGT	0	0.00	1.00
RGSL37_02_88486	1	AGAAGAGAAGAAAAAATTGGTAGTA	0	0.00	1.00
RGSL37_02_88487	1	GAAGAGAAGAAAAAATTGGTAGTAT	0	0.00	1.00
RGSL37_02_88488	1	AAGAGAAGAAAAAATTGGTAGTATT	0	0.00	1.00
RGSL37_02_88489	1	AGAGAAGAAAAAATTGGTAGTATTT	0	0.00	1.00
RGSL37_02_88490	1	GAGAAGAAAAAATTGGTAGTATTTT	91	91.00	92.00
RGSL37_02_88491	1	AGAAGAAAAAATTGGTAGTATTTTC	92	92.00	1.01
RGSL37_02_88492	1	GAAGAAAAAATTGGTAGTATTTTCA	93	93.00	1.01
RGSL37_02_88493	1	AAGAAAAAATTGGTAGTATTTTCAT	93	93.00	1.00
RGSL37_02_88494	1	AGAAAAAATTGGTAGTATTTTCATT	91	91.00	0.98

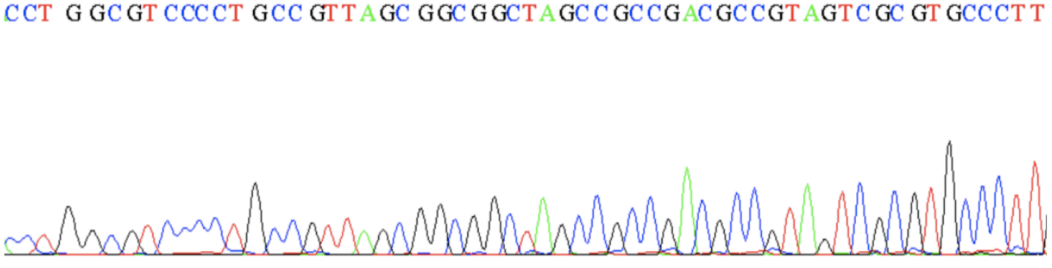
```
>Reference
AGACTATATGAAGAGATGAGGAGAAGAGAAGAAAAAATTGGTAGTATTTT
>Query
AGACTATATGAAGAGATGAG-----GAGAAGAAAAAATTGGTAGTATTTT
>Comparison
********************      *************************
```



ID: RGSL37_12_1012329

ID	CR	SEQ	PM	PMnCR	SV
RGSL37_12_1012297	1	GGACCTGGCGTCCCCTGCCGTTAGC	70	70.00	0.99
RGSL37_12_1012298	1	GACCTGGCGTCCCCTGCCGTTAGCG	69	69.00	0.99
RGSL37_12_1012299	1	ACCTGGCGTCCCCTGCCGTTAGCGG	69	69.00	1.00
RGSL37_12_1012300	1	CCTGGCGTCCCCTGCCGTTAGCGGC	67	67.00	0.97
RGSL37_12_1012301	1	CTGGCGTCCCCTGCCGTTAGCGGCG	67	67.00	1.00
RGSL37_12_1012302	1	TGGCGTCCCCTGCCGTTAGCGGCGC	1	1.00	0.03
RGSL37_12_1012303	1	GGCGTCCCCTGCCGTTAGCGGCGCT	0	0.00	0.50
RGSL37_12_1012304	1	GCGTCCCCTGCCGTTAGCGGCGCTA	0	0.00	1.00
RGSL37_12_1012305	1	CGTCCCCTGCCGTTAGCGGCGCTAG	0	0.00	1.00
RGSL37_12_1012306	1	GTCCCCTGCCGTTAGCGGCGCTAGC	0	0.00	1.00
RGSL37_12_1012307	1	TCCCCTGCCGTTAGCGGCGCTAGCC	0	0.00	1.00
RGSL37_12_1012308	1	CCCCTGCCGTTAGCGGCGCTAGCCG	0	0.00	1.00
RGSL37_12_1012309	1	CCCTGCCGTTAGCGGCGCTAGCCGA	0	0.00	1.00
RGSL37_12_1012310	1	CCTGCCGTTAGCGGCGCTAGCCGAC	0	0.00	1.00
RGSL37_12_1012311	1	CTGCCGTTAGCGGCGCTAGCCGACG	0	0.00	1.00
RGSL37_12_1012312	1	TGCCGTTAGCGGCGCTAGCCGACGC	0	0.00	1.00
RGSL37_12_1012313	1	GCCGTTAGCGGCGCTAGCCGACGCC	0	0.00	1.00
RGSL37_12_1012314	1	CCGTTAGCGGCGCTAGCCGACGCCG	0	0.00	1.00
RGSL37_12_1012315	1	CGTTAGCGGCGCTAGCCGACGCCGT	0	0.00	1.00
RGSL37_12_1012316	1	GTTAGCGGCGCTAGCCGACGCCGTA	0	0.00	1.00
RGSL37_12_1012317	1	TTAGCGGCGCTAGCCGACGCCGTAG	0	0.00	1.00
RGSL37_12_1012318	1	TAGCGGCGCTAGCCGACGCCGTAGT	0	0.00	1.00
RGSL37_12_1012319	1	AGCGGCGCTAGCCGACGCCGTAGTC	0	0.00	1.00
RGSL37_12_1012320	1	GCGGCGCTAGCCGACGCCGTAGTCG	0	0.00	1.00
RGSL37_12_1012321	1	CGGCGCTAGCCGACGCCGTAGTCGC	0	0.00	1.00
RGSL37_12_1012322	1	GGCGCTAGCCGACGCCGTAGTCGCG	0	0.00	1.00
RGSL37_12_1012323	1	GCGCTAGCCGACGCCGTAGTCGCGT	0	0.00	1.00
RGSL37_12_1012324	1	CGCTAGCCGACGCCGTAGTCGCGTG	0	0.00	1.00
RGSL37_12_1012325	1	GCTAGCCGACGCCGTAGTCGCGTGC	0	0.00	1.00
RGSL37_12_1012326	1	CTAGCCGACGCCGTAGTCGCGTGCC	0	0.00	1.00
RGSL37_12_1012327	1	TAGCCGACGCCGTAGTCGCGTGCCC	0	0.00	1.00
RGSL37_12_1012328	1	AGCCGACGCCGTAGTCGCGTGCCCT	0	0.00	1.00
RGSL37_12_1012329	1	GCCGACGCCGTAGTCGCGTGCCCTT	89	89.00	90.00
RGSL37_12_1012330	1	CCGACGCCGTAGTCGCGTGCCCTTG	90	90	1.01
RGSL37_12_1012331	1	CGACGCCGTAGTCGCGTGCCCTTGC	87	87	0.97
RGSL37_12_1012332	1	GACGCCGTAGTCGCGTGCCCTTGCC	85	85	0.98
RGSL37_12_1012333	1	ACGCCGTAGTCGCGTGCCCTTGCCC	85	85	1.00

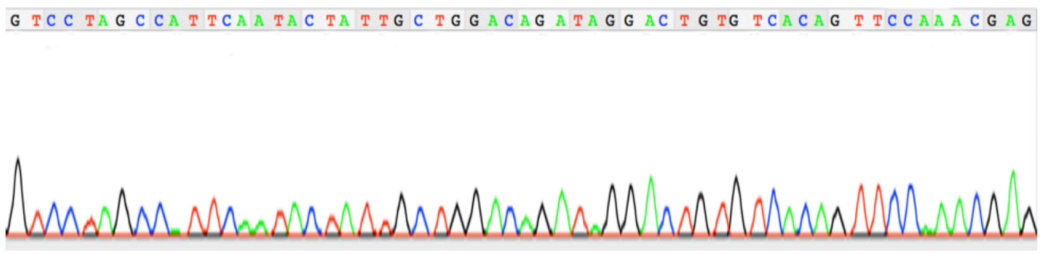
```
>Reference
GGCAGGACCTGGCGTCCCCTGCCGTTAGCGGC-GCTA---GCCGACGCCGTAGTCGCGTGCCCTT
>Query
GGCAGGACCTGGCGTCCCCTGCCGTTAGCGGCGGCTAGCCGCCGACGCCGTAGTCGCGTGCCCTT
>Comparison
********************************* ****    ************************
```



ID: RGSL37_15_79306

ID	CR	SEQ	PM	PMnCR	SV
RGSL37_15_79268	1	TAAGGTCCTAGCCATTCAATACTAT	87	87.00	0.99
RGSL37_15_79269	1	AAGGTCCTAGCCATTCAATACTATT	87	87.00	1.00
RGSL37_15_79270	1	AGGTCCTAGCCATTCAATACTATTG	86	86.00	0.99
RGSL37_15_79271	1	GGTCCTAGCCATTCAATACTATTGC	86	86.00	1.00
RGSL37_15_79272	1	GTCCTAGCCATTCAATACTATTGCT	85	85.00	0.99
RGSL37_15_79273	1	TCCTAGCCATTCAATACTATTGCTC	0	0.00	0.01
RGSL37_15_79274	1	CCTAGCCATTCAATACTATTGCTCG	0	0.00	1.00
RGSL37_15_79275	1	CTAGCCATTCAATACTATTGCTCGA	0	0.00	1.00
RGSL37_15_79276	1	TAGCCATTCAATACTATTGCTCGAC	0	0.00	1.00
RGSL37_15_79277	1	AGCCATTCAATACTATTGCTCGACA	0	0.00	1.00
RGSL37_15_79278	1	GCCATTCAATACTATTGCTCGACAG	0	0.00	1.00
RGSL37_15_79279	1	CCATTCAATACTATTGCTCGACAGA	0	0.00	1.00
RGSL37_15_79280	1	CATTCAATACTATTGCTCGACAGAA	0	0.00	1.00
RGSL37_15_79281	1	ATTCAATACTATTGCTCGACAGAAT	0	0.00	1.00
RGSL37_15_79282	1	TTCAATACTATTGCTCGACAGAATG	0	0.00	1.00
RGSL37_15_79283	1	TCAATACTATTGCTCGACAGAATGG	0	0.00	1.00
RGSL37_15_79284	1	CAATACTATTGCTCGACAGAATGGA	0	0.00	1.00
RGSL37_15_79285	1	AATACTATTGCTCGACAGAATGGAC	0	0.00	1.00
RGSL37_15_79286	1	ATACTATTGCTCGACAGAATGGACT	0	0.00	1.00
RGSL37_15_79287	1	TACTATTGCTCGACAGAATGGACTG	0	0.00	1.00
RGSL37_15_79288	1	ACTATTGCTCGACAGAATGGACTGT	0	0.00	1.00
RGSL37_15_79289	1	CTATTGCTCGACAGAATGGACTGTG	0	0.00	1.00
RGSL37_15_79290	1	TATTGCTCGACAGAATGGACTGTGT	0	0.00	1.00
RGSL37_15_79291	1	ATTGCTCGACAGAATGGACTGTGTC	0	0.00	1.00
RGSL37_15_79292	1	TTGCTCGACAGAATGGACTGTGTCA	0	0.00	1.00
RGSL37_15_79293	1	TGCTCGACAGAATGGACTGTGTCAC	0	0.00	1.00
RGSL37_15_79294	1	GCTCGACAGAATGGACTGTGTCACA	0	0.00	1.00
RGSL37_15_79295	1	CTCGACAGAATGGACTGTGTCACAG	0	0.00	1.00
RGSL37_15_79296	1	TCGACAGAATGGACTGTGTCACAGT	0	0.00	1.00
RGSL37_15_79297	1	CGACAGAATGGACTGTGTCACAGTT	0	0.00	1.00
RGSL37_15_79298	1	GACAGAATGGACTGTGTCACAGTTC	0	0.00	1.00
RGSL37_15_79299	1	ACAGAATGGACTGTGTCACAGTTCC	0	0.00	1.00
RGSL37_15_79300	1	CAGAATGGACTGTGTCACAGTTCCA	0	0.00	1.00
RGSL37_15_79301	1	AGAATGGACTGTGTCACAGTTCCAA	0	0.00	1.00
RGSL37_15_79302	1	GAATGGACTGTGTCACAGTTCCAAA	0	0.00	1.00
RGSL37_15_79303	1	AATGGACTGTGTCACAGTTCCAAAC	0	0.00	1.00
RGSL37_15_79304	1	ATGGACTGTGTCACAGTTCCAAACG	0	0.00	1.00
RGSL37_15_79305	1	TGGACTGTGTCACAGTTCCAAACGA	0	0.00	1.00
RGSL37_15_79306	1	GGACTGTGTCACAGTTCCAAACGAG	82	82.00	83.00
RGSL37_15_79307	1	GACTGTGTCACAGTTCCAAACGAGC	82	82.00	1.00
RGSL37_15_79308	1	ACTGTGTCACAGTTCCAAACGAGCC	83	83.00	1.01
RGSL37_15_79309	1	CTGTGTCACAGTTCCAAACGAGCCG	82	82.00	0.99
RGSL37_15_79310	1	TGTGTCACAGTTCCAAACGAGCCGT	82	82.00	1.00

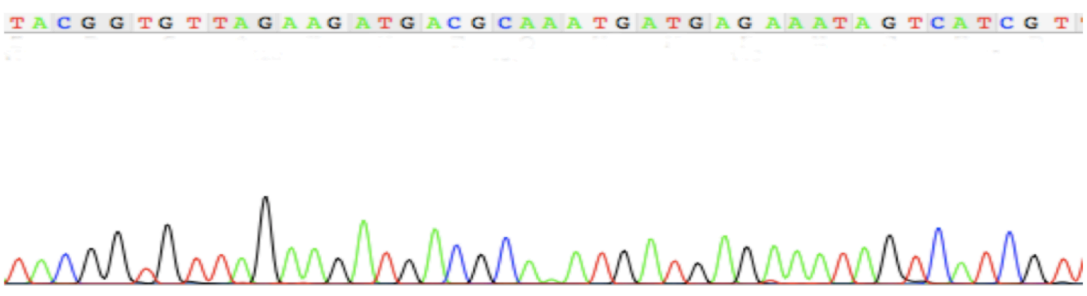
```
>Reference
CATAAGGTCCTAGCCATTCAATACTATTGCTCGACAGAATGGACTGTGTCACAGTTCCAAACGAG
>Query
CATAAGGTCCTAGCCATTCAATACTATTGCTGGACAGATAGGACTGTGTCACAGTTCCAAACGAG
>Comparison
******************************** ******  ************************
```



ID: RGSL37_07_405087

ID	CR	SEQ	PM	PMnCR	SV
RGSL37_07_405058	1	ATAATCATATACGGTGTTAGAAGAT	73	73.00	1.01
RGSL37_07_405059	1	TAATCATATACGGTGTTAGAAGATG	74	74.00	1.01
RGSL37_07_405060	1	AATCATATACGGTGTTAGAAGATGA	76	76.00	1.03
RGSL37_07_405061	85	ATCATATACGGTGTTAGAAGATGAC	9120	107.29	1.41
RGSL37_07_405062	80	TCATATACGGTGTTAGAAGATGACG	8552	106.90	1.00
RGSL37_07_405063	1	CATATACGGTGTTAGAAGATGACGG	3	3.00	0.04
RGSL37_07_405064	1	ATATACGGTGTTAGAAGATGACGGC	0	0.00	0.25
RGSL37_07_405065	1	TATACGGTGTTAGAAGATGACGGCA	0	0.00	1.00
RGSL37_07_405066	1	ATACGGTGTTAGAAGATGACGGCAA	0	0.00	1.00
RGSL37_07_405067	1	TACGGTGTTAGAAGATGACGGCAAA	0	0.00	1.00
RGSL37_07_405068	1	ACGGTGTTAGAAGATGACGGCAAAT	0	0.00	1.00
RGSL37_07_405069	1	CGGTGTTAGAAGATGACGGCAAATG	0	0.00	1.00
RGSL37_07_405070	1	GGTGTTAGAAGATGACGGCAAATGA	0	0.00	1.00
RGSL37_07_405071	1	GTGTTAGAAGATGACGGCAAATGAT	0	0.00	1.00
RGSL37_07_405072	1	TGTTAGAAGATGACGGCAAATGATG	0	0.00	1.00
RGSL37_07_405073	1	GTTAGAAGATGACGGCAAATGATGA	0	0.00	1.00
RGSL37_07_405074	1	TTAGAAGATGACGGCAAATGATGAG	0	0.00	1.00
RGSL37_07_405075	1	TAGAAGATGACGGCAAATGATGAGA	0	0.00	1.00
RGSL37_07_405076	1	AGAAGATGACGGCAAATGATGAGAA	0	0.00	1.00
RGSL37_07_405077	1	GAAGATGACGGCAAATGATGAGAAA	0	0.00	1.00
RGSL37_07_405078	1	AAGATGACGGCAAATGATGAGAAAT	0	0.00	1.00
RGSL37_07_405079	1	AGATGACGGCAAATGATGAGAAATA	0	0.00	1.00
RGSL37_07_405080	1	GATGACGGCAAATGATGAGAAATAG	0	0.00	1.00
RGSL37_07_405081	1	ATGACGGCAAATGATGAGAAATAGT	0	0.00	1.00
RGSL37_07_405082	1	TGACGGCAAATGATGAGAAATAGTC	0	0.00	1.00
RGSL37_07_405083	1	GACGGCAAATGATGAGAAATAGTCA	0	0.00	1.00
RGSL37_07_405084	1	ACGGCAAATGATGAGAAATAGTCAT	0	0.00	1.00
RGSL37_07_405085	1	CGGCAAATGATGAGAAATAGTCATC	0	0.00	1.00
RGSL37_07_405086	1	GGCAAATGATGAGAAATAGTCATCG	0	0.00	1.00
RGSL37_07_405087	3	GCAAATGATGAGAAATAGTCATCGT	277	92.33	93.33
RGSL37_07_405088	3	CAAATGATGAGAAATAGTCATCGTT	278	92.67	1.00
RGSL37_07_405089	3	AAATGATGAGAAATAGTCATCGTTT	279	93.00	1.00
RGSL37_07_405090	3	AATGATGAGAAATAGTCATCGTTTT	282	94.00	1.01
RGSL37_07_405091	3	ATGATGAGAAATAGTCATCGTTTTC	279	93.00	0.99

```
>Reference
TCATATACGGTGTTAGAAGATGACGGCAAATGATGAGAAATAGTCATCGT
>Query
TCATATACGGTGTTAGAAGATGAC-GCAAATGATGAGAAATAGTCATCGT
>Comparison
*********************** ************************
```

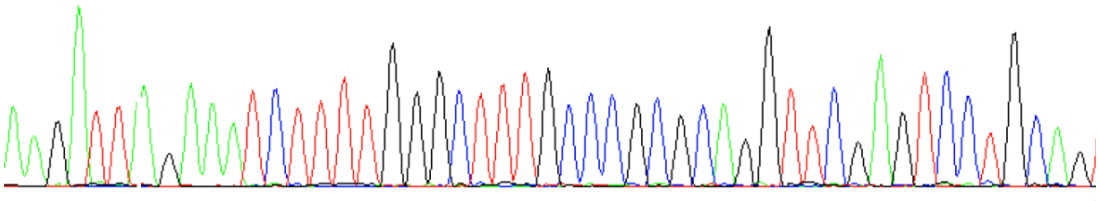


ID: RGSL37_04_1305682

ID	CR	SEQ	PM	PMnCR	SV
RGSL37_04_1305653	11	GCGAAAGATTAGAAATCTTTTGGGC	903	82.09	1.00
RGSL37_04_1305654	11	CGAAAGATTAGAAATCTTTTGGGCT	899	81.73	1.00
RGSL37_04_1305655	11	GAAAGATTAGAAATCTTTTGGGCTT	904	82.18	1.01
RGSL37_04_1305656	11	AAAGATTAGAAATCTTTTGGGCTTT	913	83.00	1.01
RGSL37_04_1305657	11	AAGATTAGAAATCTTTTGGGCTTTG	912	82.91	1.00
RGSL37_04_1305658	1	AGATTAGAAATCTTTTGGGCTTTGG	0	0.00	0.01
RGSL37_04_1305659	1	GATTAGAAATCTTTTGGGCTTTGGC	0	0.00	1.00
RGSL37_04_1305660	1	ATTAGAAATCTTTTGGGCTTTGGCC	0	0.00	1.00
RGSL37_04_1305661	1	TTAGAAATCTTTTGGGCTTTGGCCC	0	0.00	1.00
RGSL37_04_1305662	1	TAGAAATCTTTTGGGCTTTGGCCCG	0	0.00	1.00
RGSL37_04_1305663	1	AGAAATCTTTTGGGCTTTGGCCCGC	0	0.00	1.00
RGSL37_04_1305664	1	GAAATCTTTTGGGCTTTGGCCCGCG	0	0.00	1.00
RGSL37_04_1305665	1	AAATCTTTTGGGCTTTGGCCCGCGC	0	0.00	1.00
RGSL37_04_1305666	1	AATCTTTTGGGCTTTGGCCCGCGCA	0	0.00	1.00
RGSL37_04_1305667	1	ATCTTTTGGGCTTTGGCCCGCGCAG	0	0.00	1.00
RGSL37_04_1305668	1	TCTTTTGGGCTTTGGCCCGCGCAGG	0	0.00	1.00
RGSL37_04_1305669	1	CTTTTGGGCTTTGGCCCGCGCAGGT	0	0.00	1.00
RGSL37_04_1305670	1	TTTTGGGCTTTGGCCCGCGCAGGTT	0	0.00	1.00
RGSL37_04_1305671	1	TTTGGGCTTTGGCCCGCGCAGGTTC	0	0.00	1.00
RGSL37_04_1305672	1	TTGGGCTTTGGCCCGCGCAGGTTCG	0	0.00	1.00
RGSL37_04_1305673	1	TGGGCTTTGGCCCGCGCAGGTTCGA	0	0.00	1.00
RGSL37_04_1305674	1	GGGCTTTGGCCCGCGCAGGTTCGAG	0	0.00	1.00
RGSL37_04_1305675	1	GGCTTTGGCCCGCGCAGGTTCGAGT	0	0.00	1.00
RGSL37_04_1305676	1	GCTTTGGCCCGCGCAGGTTCGAGTC	0	0.00	1.00
RGSL37_04_1305677	1	CTTTGGCCCGCGCAGGTTCGAGTCC	0	0.00	1.00
RGSL37_04_1305678	1	TTTGGCCCGCGCAGGTTCGAGTCCT	0	0.00	1.00
RGSL37_04_1305679	1	TTGGCCCGCGCAGGTTCGAGTCCTG	0	0.00	1.00
RGSL37_04_1305680	1	TGGCCCGCGCAGGTTCGAGTCCTGC	0	0.00	1.00
RGSL37_04_1305681	1	GGCCCGCGCAGGTTCGAGTCCTGCA	0	0.00	1.00
RGSL37_04_1305682	11	GCCCGCGCAGGTTCGAGTCCTGCAG	863	78.45	79.45
RGSL37_04_1305683	11	CCCGCGCAGGTTCGAGTCCTGCAGT	866	78.73	1.00
RGSL37_04_1305684	11	CCGCGCAGGTTCGAGTCCTGCAGTT	862	78.36	1.00
RGSL37_04_1305685	11	CGCGCAGGTTCGAGTCCTGCAGTTG	858	78.00	1.00
RGSL37_04_1305686	11	GCGCAGGTTCGAGTCCTGCAGTTGT	864	78.55	1.01

```
>Reference
AAGATTAGAAATCTTTTGGGCTTTGGCCCGCGCAGGTTCGAGTCCTGCAG
>Query
AAGATTAGAAATCTTTTGGGCTTT-GCCCGCGCAGGTTCGAGTCCTGCAG
>Comparison
*********************** ************************
```

AAGATT AGAAATCTTTT GGGCTTT GCCC GCGCAGGTTC GAGTCCT GCAG

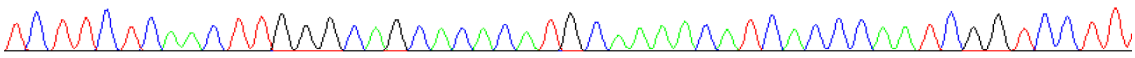


ID: RGSL37_06_3050

ID	CR	SEQ	PM	PMnCR	SV
RGSL37_06_3020	19	GTATCTCTTCTCAACTTGGGCAGCA	4290	225.79	0.99
RGSL37_06_3021	19	TATCTCTTCTCAACTTGGGCAGCAC	4291	225.84	1.00
RGSL37_06_3022	19	ATCTCTTCTCAACTTGGGCAGCACA	4286	225.58	1.00
RGSL37_06_3023	8	TCTCTTCTCAACTTGGGCAGCACAC	2388	298.50	1.32
RGSL37_06_3024	8	CTCTTCTCAACTTGGGCAGCACACA	2373	296.62	0.99
RGSL37_06_3025	1	TCTTCTCAACTTGGGCAGCACACAC	1	1.00	0.01
RGSL37_06_3026	1	CTTCTCAACTTGGGCAGCACACACG	1	1.00	1.00
RGSL37_06_3027	1	TTCTCAACTTGGGCAGCACACACGC	1	1.00	1.00
RGSL37_06_3028	1	TCTCAACTTGGGCAGCACACACGCA	1	1.00	1.00
RGSL37_06_3029	1	CTCAACTTGGGCAGCACACACGCAA	1	1.00	1.00
RGSL37_06_3030	1	TCAACTTGGGCAGCACACACGCAAA	1	1.00	1.00
RGSL37_06_3031	1	CAACTTGGGCAGCACACACGCAAAA	1	1.00	1.00
RGSL37_06_3032	1	AACTTGGGCAGCACACACGCAAAAC	1	1.00	1.00
RGSL37_06_3033	1	ACTTGGGCAGCACACACGCAAAACA	1	1.00	1.00
RGSL37_06_3034	1	CTTGGGCAGCACACACGCAAAACAT	1	1.00	1.00
RGSL37_06_3035	1	TTGGGCAGCACACACGCAAAACATC	1	1.00	1.00
RGSL37_06_3036	1	TGGGCAGCACACACGCAAAACATCA	1	1.00	1.00
RGSL37_06_3037	1	GGGCAGCACACACGCAAAACATCAC	1	1.00	1.00
RGSL37_06_3038	1	GGCAGCACACACGCAAAACATCACC	1	1.00	1.00
RGSL37_06_3039	1	GCAGCACACACGCAAAACATCACCC	1	1.00	1.00
RGSL37_06_3040	1	CAGCACACACGCAAAACATCACCCA	1	1.00	1.00
RGSL37_06_3041	1	AGCACACACGCAAAACATCACCCAA	1	1.00	1.00
RGSL37_06_3042	1	GCACACACGCAAAACATCACCCAAT	1	1.00	1.00
RGSL37_06_3043	1	CACACACGCAAAACATCACCCAATC	1	1.00	1.00
RGSL37_06_3044	1	ACACACGCAAAACATCACCCAATCG	1	1.00	1.00
RGSL37_06_3045	1	CACACGCAAAACATCACCCAATCGG	1	1.00	1.00
RGSL37_06_3046	1	ACACGCAAAACATCACCCAATCGGT	0	0.00	0.50
RGSL37_06_3047	1	CACGCAAAACATCACCCAATCGGTC	0	0.00	1.00
RGSL37_06_3048	1	ACGCAAAACATCACCCAATCGGTCC	0	0.00	1.00
RGSL37_06_3049	1	CGCAAAACATCACCCAATCGGTCCT	0	0.00	1.00
RGSL37_06_3050	8	GCAAAACATCACCCAATCGGTCCTT	2303	287.88	288.88
RGSL37_06_3051	8	CAAAACATCACCCAATCGGTCCTTT	2314	289.25	1.00
RGSL37_06_3052	8	AAAACATCACCCAATCGGTCCTTTT	2328	291.00	1.01
RGSL37_06_3053	8	AAACATCACCCAATCGGTCCTTTTT	2324	290.50	1.00
RGSL37_06_3054	8	AACATCACCCAATCGGTCCTTTTTG	2310	288.75	0.99

```
>Reference
TCTTCTCAACTTGGGCAGCACACACGCAAAACATCACCCAATCGGTCCTT
>Query
TCTTCTCAACTTGGGCAGCACACATGCAAAACATCACCCAATCGGTCCTT
>Comparison
*********************** ************************
```

T C T T C T C A A C T T G G G C A G C A C A C A T G C A A A A C A T C A C C C A A T C G G T C C T T



File S3. Random simulation in the query genome. Different types of variants were randomly introduced into the S288C reference genome to generate an artificial query genome. SNV, single nucleotide variant; MICRO-DEL, micro-deletion; MICRO-IN, micro-insertion; CON, concatenated variants; DEL, deletion; IN, insertion. All insertions correspond to fragments of the Lambda Phage Genome (LPG) at the indicated positions. RG, Reference Genome; QG, Query Genome; ID, unique identifier of the downstream recovery string; CR, count reference of the downstream recovery string; ZTL, zero-trail length; SV, signature value of the downstream recovery string. Grey rows correspond to introduced variants that did not generate a zero-trail signature of variation.

ID	CR	ZTL	SV	VARIANT TYPE	VARIANT RG/QG	POSITION	VARIANT DETECTED
RGSLS1_01_22602	1	21	68.00	SNV	T/G	22601	YES
RGSLS1_01_42728	1	25	76.00	SNV	A/C	42727	YES
RGSLS1_01_69902	1	25	73.00	SNV	A/G	69901	YES
RGSLS1_01_77088	1	33	78.00	CON	G/T	77087	YES
					A/G	77081	YES
					C/T	77079	YES
RGSLS1_01_83563	1	25	72.00	SNV	A/T	83562	YES
RGSLS1_01_86839	1	25	81.00	SNV	T/G	86838	YES
RGSLS1_01_94780	1	28	77.00	MICRO-DEL	CCAAT/00000	94775-94779	YES
RGSLS1_01_117917	1	25	73.00	SNV	A/G	117916	YES
RGSLS1_01_193534	1	23	63.00	MICRO-IN	00000/GGATG	193533	YES
RGSLS1_01_194592	1	24	73.00	MICRO-IN	000/TTG	194591	YES
RGSLS1_02_79602	1	43	71.00	CON	AAATG/00000	79598-79602	YES
					C/A	79583	YES
RGSLS1_02_334289	1	5712*	78.00	DEL	5687 N	328602-334289	YES
RGSLS1_02_441998	1	25	93.00	SNV	G/C	441997	YES
RGSLS1_02_532638	1	25	75.00	SNV	T/A	532637	YES
RGSLS1_02_720472	1	24	68.00	IN	1957 N (LPG (1-1957)	720471	YES
RGSLS1_03_275761	1	25	75.00	SNV	G/T	275760	YES
RGSLS1_03_302985	1	23	70.00	MICRO-IN	0/G	302984	YES
RGSLS1_04_75906	1	25	88.00	SNV	T/C	75905	YES
RGSLS1_04_113110	1	23	65.00	MICRO-IN	0000/AATC	113109	YES
RGSLS1_04_194190	1	24	87.00	MICRO-IN	0/A	194189	YES
RGSLS1_04_339080	1	25	65.00	SNV	G/C	339079	YES
RGSLS1_04_522427	1	25	75.00	SNV	C/G	522426	YES
RGSLS1_04_601944	1	25	69.00	SNV	C/T	601943	YES
RGSLS1_04_723652	1	25	78.00	SNV	T/C	723651	YES
RGSLS1_05_98498	1	24	100.00	MICRO-IN	0/G	98497	YES
RGSLS1_05_147838	1	28	82.00	CON	T/O	147837	YES
					T/A	147834	YES
RGSLS1_05_288901	1	212	76.00	DEL	187 N	288714-288901	YES
RGSLS1_05_479008	1	42	83.00	CON	G/T	479007	YES
					T/G	478990	YES
RGSLS1_05_547490	1	39	83.00	CON	00/AC	547489	YES
					A/C	547475	YES
RGSLS1_06_30519	1	25	77.00	SNV	A/C	30518	YES
RGSLS1_06_69695	1	46	74.00	CON	A/T	69694	YES
					T/G	69673	YES
RGSLS1_06_130133	1	25	74.00	SNV	A/T	130132	YES
RGSLS1_06_162048	1	8	78.00	CON	0/C	162047	YES
					T/A	162039	YES
RGSLS1_06_168804	1	25	73.00	SNV	A/C	168803	YES
RGSLS1_06_226057	1	25	67.00	SNV	G/A	226056	YES
RGSLS1_06_268788	1	25	86.00	SNV	A/T	268787	YES
RGSLS1_07_63686	1	25	77.00	SNV	A/C	63685	YES
RGSLS1_07_304341	1	25	82.00	MICRO-DEL	TC/00	304340-304341	YES
RGSLS1_07_473665	1	26	68.00	MICRO-DEL	GG/00	473663-473664	YES

RGSLS1_07_595745	1	24	76.00	MICRO-IN	00000/TTGAA	595744	YES
RGSLS1_07_697372	1	25	89.00	MICRO-DEL	AC/00	697370-697371	YES
RGSLS1_07_727256	1	25	74.00	SNV	G/T	727255	YES
RGSLS1_08_320725	1	26	106.00	CON	T/A	320724	YES
					G/T	320723	YES
RGSLS1_09_148566	1	23	85.00	MICRO-IN	0/C	148565	YES
RGSLS1_09_251315	1	25	86.00	SNV	G/A	251314	YES
RGSLS1_09_328160	1	25	77.00	MICRO-DEL	T/0	328159	YES
RGSLS1_09_358843	1	25	75.00	SNV	A/C	358842	YES
RGSLS1_10_160406	1	25	76.00	SNV	A/G	160405	YES
RGSLS1_10_176952	1	25	82.00	SNV	A/C	176951	YES
RGSLS1_10_217860	1	27	82.00	CON	0/C	217859	YES
					T/A	217857	YES
RGSLS1_10_241712	1	24	75.00	MICRO-IN	0/C	241711	YES
	2		1.93	SNV	T/G	367197	NO
RGSLS1_10_428096	1	42	68.00	CON	000/AAT	428095	YES
					G/T	428078	YES
RGSLS1_10_706986	1	28	82.00	CON	G/A	706985	YES
					A/C	706982	YES
RGSLS1_11_191275	1	25	61.00	SNV	C/T	191274	YES
RGSLS1_11_220608	1	24	74.00	MICRO-DEL	TA/00	220606-220607	YES
RGSLS1_11_330654	1	26	72.00	MICRO-DEL	GTT/000	330652-330654	YES
RGSLS1_11_415100	1	25	76.00	SNV	G/C	415099	YES
RGSLS1_11_419194	1	24	79.00	MICRO-DEL	C/0	419193	YES
RGSLS1_11_545707	1	25	67.00	SNV	A/G	545706	YES
RGSLS1_11_600747	1	24	81.00	MICRO-IN	0/T	600746	YES
	19		1.06	MICRO-IN	0000/TTAG	8795	NO
RGSLS1_12_94754	1	11	70.00	SNV	A/C	94753	YES
RGSLS1_12_143254	1	36	82.00	CON	A/T	143253	YES
					A/C	143252	YES
					T/G	143242	YES
RGSLS1_12_389577	1	25	72.00	SNV	T/G	389576	YES
RGSLS1_12_664063	1	25	72.00	MICRO-DEL	G/0	664062	YES
RGSLS1_12_810888	1	25	77.00	SNV	T/A	810887	YES
RGSLS1_12_918559	1	23	76.00	MICRO-IN	0000/TAGC	918558	YES
RGSLS1_12_993809	1	25	94.00	MICRO-DEL	G/0	993808	YES
RGSLS1_13_47543	1	24	103.00	IN	5895 N (LPG, 2201-8095)	47542	YES
RGSLS1_13_55935	1	41	65.00	CON	C/0	55934	YES
					C/G	55918	YES
RGSLS1_13_160786	1	40	78.00	CON	G/T	160785	YES
					G/C	160784	YES
					A/T	160783	YES
					T/A	160770	YES
RGSLS1_13_430776	1	27	76.00	MICRO-DEL	ATT/000	430775-430777	YES
RGSLS1_13_536125	1	24	77.00	IN	136 N (LPG, 2001-2136)	536124	YES
RGSLS1_13_620952	1	25	78.00	SNV	T/G	620951	YES
RGSLS1_13_844534	1	2376	68.00	DEL	2351 N	842183-844533	YES
RGSLS1_14_9097	1	4	78.00	SNV	G/A	9096	YES
RGSLS1_14_74131	1	24	92.00	IN	48 N (LPG, 14896-14943)	74130	YES
RGSLS1_14_76825	1	779	69.00	DEL	754 N	76071-76824	YES
RGSLS1_14_264223	1	44	65.00	CON	GC/00	264223-264224	YES
					A/T	264203	YES
RGSLS1_14_334544	1	46	77.00	CON	A/C	334543	YES
					T/G	334542	YES
					C/T	334522	YES
RGSLS1_14_473311	1	24	75.00	MICRO-IN	00/CT	473310	YES
RGSLS1_14_689709	1	25	76.00	SNV	A/C	689708	YES

RGSLS1_14_756031	1	24	62.00	MICRO-IN	000/GGC	756030	YES
	2		1.70	SNV	T/G	783956	NO
RGSLS1_15_54291	1	25	85.00	SNV	T/C	54290	YES
RGSLS1_15_105891	1	36	70.00	CON	0/G	105890	YES
					A/T	105879	YES
RGSLS1_15_189936	1	38	76.00	CON	ACTC/0000	189932-189935	YES
					A/T	189922	YES
RGSLS1_15_245928	1	25	86.00	SNV	G/T	245927	YES
RGSLS1_15_311893	1	28	65.00	MICRO-DEL	GACTT/00000	311888-311892	YES
RGSLS1_15_609609	1	25	56.00	MICRO-DEL	C/0	609608	YES
RGSLS1_15_643394	1	24	80.00	MICRO-IN	00/AG	643393	YES
RGSLS1_15_771734	1	30	76.00	CON	G/A	771733	YES
					C/T	771732	YES
					A/G	771728	YES
RGSLS1_15_1003054	1	25	78.00	SNV	T/G	1003053	YES
RGSLS1_16_176098	1	24	70.00	MICRO-DEL	T/0	176097	YES
RGSLS1_16_349903	1	25	89.00	SNV	T/G	349902	YES
RGSLS1_16_518294	1	34	76.00	CON	T/C	518293	YES
					C/T	518284	YES
RGSLS1_16_526756	1	25	76.00	SNV	G/C	526755	YES
RGSLS1_16_644103	1	86	63.00	DEL	62 N	644041-644102	YES
RGSLS1_16_708023	1	23	73.00	IN	683 N (LPG, 16001-16683)	708022	YES

* The zero-trail was interrupted in 6 places (see File S1)

File S4. Directed simulation in multiple copy regions of the reference genome. Different types of variants were introduced into repeated regions of the S288C reference genome. SNV, single nucleotide variant; MICRO-DEL, micro-deletion; MICRO-IN, micro-insertion. RG, reference genome; QG, query genome; ID, unique identifier of the downstream recovery string; CR, count reference of the downstream recovery string; ZTL, zero-trail length. Grey rows correspond to introduced variant that did not generate a zero-trail signature of variation.

ID	CR	ZTL	SV	VARIANT TYPE	VARIANT QG/RG	POSITION	VARIANT DETECTED
RGSLA28_01_13870	3	24	69.33	MICRO-DEL	A/0	13870	YES
RGSLA28_01_139807	2	24	172	MICRO-IN	000/GGA	139805	YES
RGSLA28_01_160464	82	27	90.26	SNVs	CCA/TTG	160459	YES
RGSLA28_01_163881	33	25	83.7	SNV	C/A	163878	YES
RGSLA28_01_164831	29	23	83.07	MICRO-DEL	ATA/000	164830	YES
RGSLA28_01_213003	2	25	75	SNV	A/T	213003	YES
RGSLA28_01_218320	3	25	89.33	MICRO-IN	0/A	218320	YES
RGSLA28_02_3466	19	25	76.37	SNV	C/G	3465	YES
RGSLA28_02_29741	82	27	90.38	SNVs	AAA/TGG	29738	YES
RGSLA28_02_31326	13	23	78.62	MICRO-DEL	T/0	31326	YES
RGSLA28_02_88581	2	24	67	MICRO-DEL	AAC/000	88582	YES
RGSLA28_02_221742	3	25	76	SNV	C/T	221745	YES
RGSLA28_02_260663	29	27	75.55	MICRO-IN	000/CGT	260663	YES
RGSLA28_02_810730	4	25	101.33	MICRO-IN	0/T	810730	YES
RGSLA28_03_4794	2	25	74.5	SNV	C/A	4793	YES
RGSLA28_03_12399	3	22	75.67	MICRO-DEL	T/0	12401	YES
RGSLA28_03_82792	82	27	91.68	SNVs	TGA/GAC	82790	YES
RGSLA28_03_86475	13	25	83.5	SNV	C/A	86475	YES
RGSLA28_03_199545	3	23	110.5	MICRO-DEL	TCA/000	199546	YES
RGSLA28_03_294421	3	26	82.33	MICRO-IN	000/AGG	294423	YES
RGSLA28_03_314990	2	25	83.5	MICRO-IN	0/A	314990	YES
RGSLA28_04_2480	2	25	78.5	MICRO-IN	0/A	2479	YES
RGSLA28_04_516969	13	25	83	SNV	C/A	516967	YES
RGSLA28_04_519534	82	27	91.65	SNVs	TTC/GAT	519530	YES
RGSLA28_04_527578	3	22	78.67	MICRO-DEL	AGC/000	527578	YES
RGSLA28_04_650108	29	24	76.83	MICRO-DEL	C/0	650110	YES
RGSLA28_04_1527883	11	25	80.2	SNV	G/T	1527885	YES
RGSLA28_04_1531040	19	27	77.42	MICRO-IN	000/AAG	1531040	YES
RGSLA28_05_19962	2	21	75	MICRO-DEL	A/0	19962	YES
RGSLA28_05_363588	2	24	71.5	MICRO-DEL	AGA/000	363589	YES
RGSLA28_05_443619	82	26	89.1	SNVs	ACT/TTA	443620	YES
RGSLA28_05_447991	29	25	79.82	SNV	T/G	447994	YES
RGSLA28_05_496130	33	27	84.13	MICRO-IN	000/TTG	496131	YES
RGSLA28_05_572687	11	25	102	SNV	T/A	572687	YES
RGSLA28_05_575855	19	25	81.28	MICRO-IN	0/T	575855	YES
RGSLA28_06_3731	19	25	77.21	SNV	G/C	3730	YES
RGSLA28_06_3754	19	23	85.18	MICRO-DEL	G/0	3755	YES
RGSLA28_06_7131	4	25	144.5	SNV	A/G	7131	YES
RGSLA28_06_8080	2	24	69	MICRO-IN	000/TTA	8080	YES
RGSLA28_06_138019	82	27	83.43	SNVs	CAT/TGA	138014	YES
RGSLA28_06_139577	13	25	90.45	MICRO-IN	0/T	139574	YES
RGSLA28_06_220671	2	21	78.5	MICRO-DEL	ACC/000	220669	YES
RGSLA28_07_7759	2	25	76	MICRO-IN	0/C	7758	YES
RGSLA28_07_537998	33	23	79.47	MICRO-DEL	AAC/000	537997	YES
RGSLA28_07_572014	13	25	88.55	SNV	A/T	572015	YES
RGSLA28_07_811548	82	27	84.63	SNVs	TCA/GGT	811547	YES
RGSLA28_07_822167	29	27	82.69	MICRO-IN	000/AAC	822166	YES
RGSLA28_07_1085292	19	25	85.82	SNV	T/G	1085290	YES

RGSLA28_07_1086756	11	24	90.33	MICRO-DEL	C/0	1086755	YES
RGSLA28_08_85986	3	25	74	SNV	A/C	85985	YES
	2			MICRO-IN	000/CCG	104992	NO
RGSLA28_08_389403	82	27	87.79	SNVs	CTA/TCG	389397	YES
RGSLA28_08_547252	33	25	78.34	SNV	A/G	547248	YES
RGSLA28_08_548394	29	24	77.64	MICRO-DEL	G/0	548391	YES
RGSLA28_08_553058	2	25	71.5	MICRO-IN	0/G	553055	YES
RGSLA28_08_561361	19	24	91.75	MICRO-DEL	GCC/000	561358	YES
RGSLA28_09_1099	19	22	82.06	MICRO-IN	000/TTA	1096	YES
RGSLA28_09_4042	11	24	73.91	MICRO-DEL	C/0	4039	YES
RGSLA28_09_6026	19	25	88.59	SNV	T/C	6023	YES
RGSLA28_09_7073	3	25	69.67	MICRO-IN	A	7070	YES
RGSLA28_09_10076	2	25	89	SNV	T/G	10072	YES
RGSLA28_09_246323	82	27	87.73	SNVs	TAG/AGC	246317	YES
RGSLA28_09_432696	2	22	83	MICRO-DEL	ATG/000	432695	YES
RGSLA28_10_1083	19	25	78.11	SNV	A/T	1082	YES
RGSLA28_10_4018	11	24	82	MICRO-DEL	CCC/000	4018	YES
RGSLA28_10_10033	2	24	175	MICRO-DEL	C/0	10036	YES
RGSLA28_10_198893	3	27	76	MICRO-IN	000/GGC	198894	YES
RGSLA28_10_355066	82	27	86.42	SNVs	TAA/CGT	355064	YES
RGSLA28_10_479326	29	25	79.71	SNV	G/C	479326	YES
RGSLA28_10_480309	33	25	84.23	MICRO-IN	0/T	480309	YES
RGSLA28_11_314295	82	27	85.78	SNVs	GTC/TCT	314292	YES
	2			MICRO-IN	0/C	432619	NO
RGSLA28_11_555091	2	24	70.5	MICRO-DEL	G/0	555090	YES
	2			SNV	A/G	658781	NO
RGSLA28_11_659868	2	27	89	MICRO-IN	000/TTA	659865	YES
RGSLA28_11_662601	2	25	77	SNV	A/G	662597	YES
RGSLA28_11_664927	2	24	72.5	MICRO-DEL	GTT/000	664924	YES
RGSLA28_12_95874	2	26	90	MICRO-IN	000/CCA	95872	YES
RGSLA28_12_451996	3	25	77.67	SNV	G/A	451992	YES
RGSLA28_12_468951	4	24	82.75	MICRO-IN	0/G	468947	YES
RGSLA28_12_476207	82	27	85.29	SNVs	AAT/GCA	476200	YES
RGSLA28_12_942834	13	22	88	MICRO-DEL	CTA/000	942831	YES
RGSLA28_12_1067412	11	25	73.27	SNV	T/G	1067410	YES
RGSLA28_12_1075173	19	24	77.63	MICRO-DEL	A/0	1075172	YES
RGSLA28_13_4596	19	25	80.83	SNP	A/G	4595	YES
RGSLA28_13_7956	4	25	96.33	MICRO-IN	000/GCC	7953	YES
RGSLA28_13_376344	33	25	80.87	SNV	T/C	376340	YES
RGSLA28_13_377488	29	23	75.38	MICRO-DEL	ACT/000	377486	YES
RGSLA28_13_599840	2	24	83.5	MICRO-DEL	C/0	599840	YES
RGSLA28_13_768645	82	27	88.84	SNVs	AAT/GTC	768643	YES
RGSLA28_13_918565	3	25	80.33	MICRO-IN	0/C	918565	YES
RGSLA28_14_3643	11	22	91.44	SNV	G/T	3642	YES
RGSLA28_14_5639	19	24	83.94	MICRO-IN	0/T	5638	YES
RGSLA28_14_9058	4	25	99.67	SNV	C/A	9056	YES
RGSLA28_14_9983	2	23	139	MICRO-DEL	TTT/000	9982	YES
RGSLA28_14_562245	82	27	91.68	SNVs	AGC/tcg	562244	YES
RGSLA28_14_565335	13	23	77	MICRO-DEL	G/0	565337	YES
RGSLA28_14_751394	3	25	87	MICRO-IN	TTC	751394	YES
RGSLA28_15_2193	2	25	81.5	SNV	T/A	2192	YES
RGSLA28_15_596095	29	22	76.9	MICRO-DEL	CAT/000	596096	YES
RGSLA28_15_707356	13	25	83.17	MICRO-IN	0/T	707358	YES
RGSLA28_15_970386	82	27	86.56	SNVs	AGT/TTA	970385	YES
RGSLA28_15_1079008	3	27	83.33	MICRO-IN	000/GTC	1079007	YES
RGSLA28_15_1086999	11	25	81.2	SNV	C/A	1086997	YES
RGSLA28_15_1090178	19	24	86.53	MICRO-DEL	A/0	1090177	YES
RGSLA28_16_3547	11	25	104.12	SNV	T/G	3546	YES
RGSLA28_16_9568	3	27	74.67	MICRO-IN	000/GCA	9565	YES

RGSLA28_16_13776	2	24	70	MICRO-DEL	ACT/000	13773	YES
RGSLA28_16_58681	33	25	79.22	SNV	A/C	58680	YES
RGSLA28_16_62134	82	27	91.66	SNVs	AGA/TCC	62131	YES
RGSLA28_16_855413	29	23	79.44	MICRO-DEL	G/0	855413	YES
RGSLA28_16_947016	19	25	97.67	MICRO-IN	0/T	947016	YES

File S5. Analysis of the S288C and the BY4742 *S. cerevisiae* strains. The query genome sequece reads of strains S288C and BY4742 were compared with the reference genome of strain S288C. ID, unique identifier of the downstream recovery string; CR, count reference of the downstream recovery string; SEQ, nucleotide sequence of the downstream recovery string ; ZTL, zero-trail length; SV, signature value of the downstream recovery string; RG, reference genome; QG, query genome; AUX, auxothrophy; SNV, single nucleotide variant; m-DEL, micro-deletion; m-IN, micro-insertion; DEL, deletion; CON, concatenated variants. Strains in parenthesis indicate that the signature of variation was not detected by the corresponding automated PMGL scan, but that direct inspection of the PMGL revealed a signature of variation that was marginally filtered out by one of the PMGL scan parameters. Grey rows indicate signatures of variation that were not solved.

ID	CR	SEQ	ZTL S288C	ZTL BY4742	SV S288C	SV BY4742	VARIANT TYPE	VARIANT (RG/QG)	POSITION	STRAIN	PCR AND SANGER SEQUENCE	READABLE SANGER SEQUENCE	CONFIRMED BY SANGER SEQUENCE	OBSERVATIONS
RGSL37_01_19972	3	TTTCAACTTAGTGTGAAAGCCGAAC	24	24	89.67	97.67	m-IN	00/CT	19972	S288C, BY4742	YES	YES	YES	
RGSL37_01_25341	1	AGACGTGATAAAGCTGGTGATTTGT	25	25	78	88	SNV	C/A	25340	S288C, BY4742	YES	YES	YES	
RGSL37_01_25516	3	GTTGTAACCAAACCTTCACTGGTTG	21	21	69.83	36.33				S288C, BY4742				
RGSL37_01_41476	1	AACGTAGGTCTCGTTGGCCACAGAA	0	25	1.00	85	SNV	G/T	41475	BY4742	NO			
RGSL37_01_141031	2	CCCAAGTAAGATTCGGCAGTTTCCT	25	25	85	99	SNV	T/A	141030	S288C, BY4742	YES	YES	YES	
RGSL37_01_141136	2	GGCTTACCGTCAACATCGATCAACT	25	25	88	85.5	SNV	A/T	141135	S288C, BY4742	YES	YES	YES	
RGSL37_01_183389	1	CTTCAAGTATATTCTGTATACCTAA	24	24	85	116	m-DEL	C/0	183388	S288C, BY4742	YES	YES	YES	
RGSL37_01_199812	1	ATTGCAAAAATTTATATATTTGGAG	25	25	95	102	SNV	T/G	199811	S288C, BY4742	YES	YES	YES	
RGSL37_01_205051	14	CTCAGCCATGGAACGACACTTTTAC	25	25	6.31	23.93	SNV	C/A	205050	(S288C), BY4742	NO			
RGSL37_01_205520	2	CGGTTTGCCAACTGATGAGACCATC	45	45	21	18				S288C, (BY4742)				
RGSL37_01_206124	3	ACTACAACGGAACCATGGACCGGTA	22	22	27.33	35.33				S288C, BY4742				
RGSL37_01_206180	5	CGTCACCGGTACCAACGGCGTTCCA	27	27	32.6	38.6	CON	A/C C/A G/C	206179 206178 206177	S288C, BY4742	YES	NO		
RGSL37_01_212089	2	TCCCGAGGATTATAATTGTTCGGCT	24	24	88.5	106	m-DEL	T/0	212088	S288C, BY4742	YES	YES	YES	
RGSL37_01_214062	2	ATGGTCGATAACTTGCAAGAGAGAT	29	29	100	99	CON	C/A C/A	214061 214057	S288C, BY4742	YES	YES	YES	
RGSL37_02_7512	4	CGGTATTTATATCATCAAAAAAAGT	26	26	102.25	25.25	CON	G/A G/A	7511 7510	S288C, (BY4742)	NO			
RGSL37_02_7914	8	AAGCAGTAGCAGCGATGGCAGCGAC	23	23	91.38	96.25				S288C, BY4742				
RGSL37_02_15102	1	GTCATGATTTTGAATAATATTATCT	25	25	88	91	SNV	T/G	15101	S288C, BY4742	YES	YES	YES	
RGSL37_02_15519	1	ATCAATGTGATCAACTTCTAGATCC	33	33	100	104	CON	A/G G/T	15518 15510	S288C, BY4742	YES	YES	YES	
RGSL37_02_16393	1	AGGAAATCACCACAATATGCACAGC	37	37	100	107	CON	A/G C/T	16392 16380	S288C, BY4742	YES	YES	YES	
RGSL37_02_30013	1	TCATCAAATCAAGATCCGTTAGCCG	24	24	99	55.5	SNV	C/T	30012	S288C, BY4742	YES	YES	YES	
RGSL37_02_30211	4	CCAAACAAAGCTATGGCTTCTAACT	25	25	79.25	83.5	SNV	T/C	30210	S288C, BY4742	YES	YES	YES	
RGSL37_02_35486	5	CCCTTTCATGGATTCCTAAATCCTT	18	18	109.6	91.6	SNV	C/T	35485	S288C, BY4742	YES	YES	YES	
RGSL37_02_59756	1	TGTTATTTATAAAGAAAATATATAT	25	25	105	95	SNV	T/C	59755	S288C, BY4742	NO			
RGSL37_02_88490	1	GAGAAGAAAAAATTGGTAGTATTTT	22	22	92	95	m-DEL	GAGAA/00000	88485-88489	S288C, BY4742	YES	YES	YES	
RGSL37_02_222305	2	TACGCACGTCATCGATGTATACATA	26	26	97.5	75	CON	G/C C/G	222304 222303	S288C, BY4742	YES	YES	YES	
RGSL37_02_261410	33	TAAAGGTATTGCACACTCCTAACAT	19	25	109.3	101.79	SNV	T/A	261409	S288C, BY4742	NO			
RGSL37_02_263170	9	TCATCAGAAAGTAATTCACTACACC	47	47	40.28	78.44				S288C, BY4742				
RGSL37_02_264269	3	ACGACATCCTTGGCTTAGAAATCAA	25	25	113.67	118	SNV	C/T	264268	S288C, BY4742	YES	YES	YES	
RGSL37_02_265383	53	GAGGATTCCTATATCCTCGAGGAGA	25	25	110.45	50.85				S288C, BY4742				
RGSL37_02_383177	1	GGTATATCAAACGGTATTTGATTCC	22	22	65	75	m-IN	0/G	383177	S288C, BY4742	YES	YES	YES	
RGSL37_02_389431	1	TAGCGCTAAAGACAGCTATAGTGGA	29	29	87	63	CON	G/C C/G CT/00	389430 389429 389422-389423	S288C, BY4742	YES	YES	YES	
RGSL37_02_474240	1	AAATCGAATAATCAGTTTCAATGAA	0	4539	1.00	70	DEL	DEL 4514 N.	469725-474239	BY4742	NO			AUX LYS 2
RGSL37_02_474545	1	CGATTGATGAGCATACTTACCCCAG	0	25	0.99	87	SNV	A/G	474544	BY4742	NO			
RGSL37_02_474921	1	TCGTTACCATCAGCAGGTCTCCATA	0	25	0.98	65	SNV	T/C	474920	BY4742	NO			
RGSL37_02_746370	1	ATCTGCAATCTTAGTTTCACTGCCT	20	20	39.5	93	m-DEL	ATCTGCA/0000000	746363-746369	S288C, BY4742	YES	YES	YES	
RGSL37_03_84809	2	TCTGTTGGAATAAAAATCAACTATC	0	53	0.99	48				BY4742				
RGSL37_03_85134	3	TTCTCTTCATGGTAGCGCCTATGCT	0	20	1.22	61.67	SNV	G/A	85133	BY4742	NO			
RGSL37_03_86750	13	CTAGCTCCCATAGTCAAACATGGAG	0	25	0.97	75.77	SNV	C/A	86749	BY4742	NO			
RGSL37_03_90858 RGSL37_03_92481	1	CGACACGAAATTACAAAATGGAATA	0	1732	1.03	106	DEL	DEL 1708 N	90772-92480	BY4742	NO			AUX LEU 2
RGSL37_03_143132	1	GGCCGATGCTGTTGCTATAGCTCGG	25	25	97	112	SNV	T/C	143131	S288C, BY4742	YES	YES	YES	
RGSL37_03_149510 RGSL37_03_149665 RGSL37_03_149920 RGSL37_03_151275 RGSL37_03_151477 RGSL37_03_151517	2	GAATTGTTGGAATAAAAATCCACTA	2926	2926	44.5	41.5	DEL	DEL 2901 N	148615-151516	S288C, BY4742	NO			
RGSL37_03_151555	48	ATAGTTATATTATCAATATATTATC	22	22	15	33.62	m-IN	0/A		(S288C), BY4742	NO			
RGSL37_03_162362	1	AACTATAAGTTGTGCATACTCATAC	25	25	101	111	SNV	T/C	162361	S288C, BY4742	YES	YES	YES	
RGSL37_03_162397	1	ACGGCCTAATCTCTGTAAGGATTGT	25	25	87	103	SNV	G/C	162396	S288C, BY4742	YES	YES	YES	
RGSL37_03_162641	1	AACAATTTGGATTTTACGGGATCTG	25	25	40	89	SNV	T/G	162640	S288C, BY4742	YES	YES	YES	
RGSL37_03_163060	1	AAAGTCATTTATGATAGATTAAAAG	25	25	80	40	SNV	T/C	163059	S288C, BY4742	YES	YES	YES	
RGSL37_03_163193	1	TTTGCTCTTCCTTCAAAGTGTGTAG	21	21	71	100	m-IN	0/T	163193	S288C, BY4742	NO			
RGSL37_03_212960	1	CGAGATTCTGCAAAAAAGTTGCGAC	25	0	79	0.99	SNV	G/C	212959	S288C	YES	YES	YES	
RGSL37_04_507644	1	AGGATGTCAGAAAAAGCAGTTTAAT	25	0	50	1.00	SNV	A/T	507643	S288C	YES	YES	YES	
RGSL37_04_696582	1	CTACTGGCGATCCCTCAGAAAAGGC	25	0	100	1.00	SNV	A/C	696581	S288C	NO			
RGSL37_04_758311	4	CTGATGTAGAATTCTTGAAGGAGAA	25	25	50.5	20.12	SNV	T/C	758310	S288C, BY4742	YES	NO		

RGSL37_04_769329	1	GGCGGTGGAAACCCCTAAGGGGCGA	0	25	0.97	31	SNV	G/A	769328	BY4742	NO			
RGSL37_04_783599	1	GCGTTTCCATTGCGGGCATCTTCCT	0	25	1.00	83	SNV	G/A	783598	BY4742	NO			
RGSL37_04_864220	1	CAAAAGCACGCGGAAAAATCATTTG	0	25	1.00	62	SNV	A/G	864319	BY4742	NO			
RGSL37_04_1225550	1	ACTTCTTTTGTTGAACGCCGCGTCA	0	25	1.00	76	SNV	G/C	1225549	BY4742	NO			
RGSL37_04_1305682	11	GCCCGCGCAGGTTCGAGTCCTGCAG	24	24	79.45	27.58	m-DEL	G/0	1305681	S288C, BY4742	YES	YES	YES	
RGSL37_04_1402306	1	AAAGGTCACCAAGAAGCACGGTGCC	25	25	93	60	SNV	C/T	1402305	S288C, BY4742	YES	YES	YES	
RGSL37_04_1433711	1	TATGCCCTTGGTGCCATTCAGCTTC	25	25	89	70	SNV	A/T	1433710	S288C, BY4742	YES	YES	YES	
RGSL37_04_1524966	3	TTGAGATGATATATACTGTAGCATC	16	16	28.5	67.67	CON	T/C G/0	1524965 1524964	S288C, BY4742	YES	NO		
RGSL37_05_48385	1	TTTATTGTTTCTTTGGCGAAAAGGC	25	25	89	108	SNV	T/C	48384	S288C, BY4742	YES	YES	YES	
RGSL37_05_117046	1	CCCGGGAATCTCGGTCGTAATGATT	0	1084	1.03	72	DEL	DEL 1096 N	115949-117045	BY4742	NO			AUX URA 3
RGSL37_05_154532	1	CTTTGAGATTTGAAGGGAATACTTT	25	25	76	72	SNV	T/A	154531	S288C, BY4742	YES	YES	YES	
RGSL37_05_352395	1	AAAACAAGAAGAGCTTGCTACCAAG	25	25	87	78	SNV	A/G	352394	S288C, BY4742	YES	YES	YES	
RGSL37_05_517530	1	GGATAAGCCTTAGAAGCCTTAGAAG	25	25	83	94	SNV	T/C	517529	S288C, BY4742	YES	YES	YES	
RGSL37_06_3050	8	GCAAAACATCACCCAATCGGTCCTT	25	0	288.88	0.86	SNV	C/T	3049	S288C	YES	YES	YES	
RGSL37_06_3652	19	CATTGCACCAGTTCTGAATTTTACC	25	0	236.63	1.23	SNV	T/C	3651	S288C	NO			
RGSL37_06_4166	7	ACTTTCGTTTTCAATTTCCATGGTG	25	0	236.35	1.83	SNV	C/T	4165	S288C	NO			
RGSL37_06_4232	17	TTCGTGGATATTCTCTTGGATACTT	20	0	137.32	1.03				S288C				
RGSL37_06_4258	5	AAACATGGACCTACGCTTAGCCTGC	25	0	72.2	0.94	SNV	C/T	4257	S288C	NO			
RGSL37_07_215	3	CCATATCCAACTCCACTACCATTAC	22	21	34.84	30.5	m-IN	0/C	215	S288C, BY4742	NO			
RGSL37_07_8152	2	ACATATGGTGGTAATTTGTATAAGG	23	23	80.5	45.25	m-IN	0/A	8152	S288C, BY4742	YES	YES	YES	
RGSL37_07_76883	1	CTTTCGCCCGAATGTCTGAATTGAA	25	0	68	0.99	SNV	T/C	76882	S288C	YES	YES	YES	
RGSL37_07_386980	1	TTTCGCCTTAGAATGTTGGGGTGGT	26	26	96	60	CON	G/C C/G	386979 386978	S288C, BY4742	YES	YES	YES	
RGSL37_07_404475	2	GAAGCCTCGAAAAGTCGATTCGTGT	23	23	91.5	93.5	m-IN	0/G	404475	S288C, BY4742	YES	YES	YES	
RGSL37_07_404523	2	GTTCATTGCCTACCTATTATGGTAG	23	23	44.5	100.5	m-IN	0/G	404523	S288C, BY4742	YES	YES	YES	
RGSL37_07_404937	2	ATCAGCCCCAATGAAAATTAATCAG	26	26	43.75	95.5	CON	G/C C/G	404936 404935	S288C, BY4742	YES	YES	YES	
RGSL37_07_405087	3	GCAAATGATGAGAAATAGTCATCGT	24	24	93.33	97.67	m-DEL	G/0	405086	S288C, BY4742	YES	YES	YES	
RGSL37_07_1006280	1	TATCTTAATTACTCTGGCTGCCAGG	27	23	91	85	CON	G/C C/G G/C	1006279 1006278 1006277	S288C, BY4742	YES	YES	YES	
RGSL37_08_85511	25	TATATATTATATAAAATATGAATCA	19	19	21.78	27.59	m-DEL	TA/00	85509-85510	(288C),BY4742	NO			
RGSL37_08_417055	1	AAAAATTTTTCATTTGGAGGAGTTG	25	25	63	72	SNV	G/A	417054	S288C, BY4742	YES	YES	YES	
RGSL37_08_524448	1	AGGTAAGAGGGTTTCATATTTTAGA	0	25	1.02	82	SNV	C/G	524447	BY4742	NO			
RGSL37_08_557839	19	GCGTAAGTTTGAAAAAGCAAACTTT	23	23	99.13	47.81	m-IN	0/G	557839	S288C, BY4742	NO			
RGSL37_09_18864	3	CCATCTGCAAGCAGACCCCATTATT	23	23	37.16	82.67	m-DEL	C/0	18863	S288C, BY4742	NO			
RGSL37_09_22274	2	CTGCGGGTTCTCTTAGTAAATGAGG	23	23	78	86	m-IN	0/C	22274	S288C, BY4742	YES	YES	YES	
RGSL37_09_23084	2	GGAAATTGCACTAGAAGCCTTGGAA	23	23	81	44	m-DEL	G/0	23083	S288C, BY4742	YES	YES	YES	
RGSL37_09_23186	2	GTGACGGTTGAAGTAGCTTCAGGTA	23	23	85	68.5	m-IN	0/G	23186	S288C, BY4742	YES	YES	YES	
RGSL37_09_23215	2	AGAGGTGATTATCTTGACGTTACAA	25	25	86.5	71	SNV	T/G	23214	S288C, BY4742	YES	YES	YES	
RGSL37_09_23284	2	CTTAGGGGCAGCACTGGTTTCTGAA	55	55	80.5	88	CON	C/0 C/0 C/0	23283 23263 23252	S288C, BY4742	YES	YES	YES	
RGSL37_09_318695	1	ACTTGAACTATCAGTCGTTGTTTCT	25	25	74	84	SNV	T/A	318694	S288C, BY4742	YES	YES	YES	
RGSL37_10_97479	1	AACCGCGTGATCTTGTGCACTGTTA	25	25	90	90	SNV	G/A	97478	S288C, BY4742	YES	YES	YES	
RGSL37_10_199171	3	GTAAACATTACCAATGACAAAACCT	25	25	83.33	75	SNV	A/C	199170	S288C, BY4742	YES	YES	YES	
RGSL37_10_200220	3	CCTGGAACACAAAAGTACAGGTAAA	25	25	91	78	SNV	A/G	200219	S288C, BY4742	YES	YES	YES	
RGSL37_10_297733	1	CTCCTTCTGTTGCTACTAAAGGAAC	25	0	36	1.02	SNV	T/G	297732	S288C	YES	YES	YES	
RGSL37_10_733266	4	ACAAGGTTTCACCAGAGGATCCATG	25	25	88.5	77	SNV	G/A	733265	S288C, BY4742	YES	YES	YES	
RGSL37_10_733999	2	CCAGTTACTATATGAAGAAGGTATA	25	25	108	49.75	SNV	C/T	733998	S288C, BY4742	YES	YES	YES	
RGSL37_10_734784	2	TACAGTAAAAAGACCTCATATAGGT	25	25	94	84.5	SNV	G/A	734783	S288C, BY4742	YES	NO		
RGSL37_10_740239	2	CGACCTGTCGCGATGTTGCAAGCGT	25	25	89.5	88	SNV	T/C	740238	S288C, BY4742	YES	YES	YES	
RGSL37_10_743852	6	GAAACGCTAAGGACTTGTAAATTTC	26	26	90.17	86	CON	G/C C/G	743851 743850	S288C, BY4742	NO			
RGSL37_11_458	3	GTATATACCACTCTCAACTTACCCT	24	24	101	87.67	m-IN	0/A	458	S288C, BY4742	YES	YES	YES	
RGSL37_11_68469	1	CGCCGGAGGAGGCGGCGGAGGCGAC	38	38	81	64	CON	0/G 00/CC	68469 68455	S288C, BY4742	YES	YES	YES	
RGSL37_11_323020	1	ATTTGTATATAACAAGCAACTGCAG	25	0	64	0.99	SNV	A/T	323019	S288C	YES	YES	YES	
RGSL37_11_451575	1	GAACACGAACAGAAACTGCTGGCCA	19	19	67	33.5	m-DEL	GAACAC/000000	451569-451574	S288C, BY4742	NO			
RGSL37_11_666558	1	GTAAGTTGTGAGACAGGTTCATCAG	25	25	60	39	m-DEL	T/0	666557	S288C, BY4742	YES	YES	YES	
RGSL37_12_554	18	TCGCTATCCAGCTCTTTGTGAACCG	24	24	261.5	123.89	m-DEL	T/0	553	S288C, BY4742	NO			
RGSL37_12_591	2	TGGTACCCTCGTGTTATCTGCAGCG	22	22	82	41.5	m-IN	0000000/TGTACAG	591	S288C, BY4742	YES	NO		
RGSL37_12_5712	1	CACACCCACACACTCTCTTACATCT	0	21	2.36	40				BY4742				
RGSL37_12_5748	3	CGCTGTCACACCTTACCCGGCTTTC	0	31	1.22	26.67				BY4742				
RGSL37_12_5808	3	TTCTTTAGCCCTACAACACTTTTAC	0	25	1.07	41.33	CON	C/G C/T	5807 5803	BY4742	NO			
RGSL37_12_187139	1	TTGATTGTCTGTTGGGATGTGAACT	23	23	93	40	m-DEL	T/0	187138	S288C, BY4742	YES	YES	YES	
RGSL37_12_216702	25	ATGCAAGTGACAGGGATGTCATTAA	25	21	130.92	58.5	SNV	G/A	216701	S288C, BY4742	YES	YES	YES	
RGSL37_12_216825	5	CACGTTTCCTGTTAAAGATAAACAT	16	16	110.2	49.90	SNV	C/T	216824	S288C, (BY4742)	YES	YES	YES	
RGSL37_12_217019	10	GTTCGTATTGGTTTGATTGATTTTA	25	25	105.4	97.3	SNV	A/T	217018	S288C, BY4742	YES	YES	YES	

RGSL37_12_291387	1	TCGTAAATGTTTACTTTTGGGCGCT	0	27	0.99	86	CON	A/G A/G	291386 291384	BY4742	NO			
RGSL37_12_472056	4	CCCGATAGGTGAGACAATTCTTGAA	25	25	114.25	102	SNV	T/G	472055	S288C, BY4742	NO			
RGSL37_12_474418	4	CCTTCGGAGGACTAAGCGAGATCTC	22	22	58.62	55.25	m-IN	0/C	474418	S288C, BY4742	NO			
RGSL37_12_475972	2	TGTGAGAAATGGGTGAATGTTGAGA	22	22	24.75	22.5				S288C, BY4742				
RGSL37_12_481898	4	CTTTGAGACATTGTGAGACCCTCCG	23	23	58	51.75				S288C, BY4742				
RGSL37_12_528750	1	ACGAAGAAGCAATTTTTCACAGCAT	25	25	100	82	SNV	T/A	528749	S288C, BY4742	YES	YES	YES	
RGSL37_12_725937	1	TAAAGTTCGATCGTGACTTCCTGTT	25	25	101	122	SNV	G/A	725936	S288C, BY4742	YES	YES	YES	
RGSL37_12_818470	1	TTTTTATACCTTTAACTATCATCAT	19	19	122	91				S288C, BY4742				
RGSL37_12_828904	1	GAAGACGAAAATGACCAGCAGATGC	25	25	84	64	SNV	C/T	828903	S288C, BY4742	YES	YES	YES	
RGSL37_12_1012329	1	GCCGACGCCGTAGTCGCGTGCCCTT	27	27	90	82	CON	000/GCC 0/G	1012329 1012325	S288C, BY4742	YES	YES	YES	
RGSL37_12_1039436	1	TCAACCGATTTGGTATCATTTTGGG	25	25	96	112	SNV	T/A	1039435	S288C, BY4742	YES	YES	YES	
RGSL37_12_1061529	2	ACAAAGCTTGGAATGCTGAGTACTT	22	0	42	0.97	SNV	A/G	1061528	S288C	NO			
RGSL37_13_34474	1	CCATCTCACCTATTGATGGGATCTC	25	25	44	118	SNV	T/G	34473	S288C, BY4742	YES	YES	YES	
RGSL37_13_198514	32	GGTCTAGTGCTATTGACTATCAATG	25	25	57.77	106.69	SNV	A/T	198513	S288C, BY4742	NO			
RGSL37_13_325906	1	TGATAAGGGCACTTTTGGTGAATAC	25	25	82	75	SNV	T/A	325905	S288C, BY4742	YES	YES	YES	
RGSL37_13_361551	4	AGTAGATTTATTACGTTTTGATTCC	25	25	83.5	38.62	SNV	G/A	361550	S288C, BY4742	YES	YES	YES	
RGSL37_13_448335	1	ACGAACCGATATGTGCATGCGTTAT	26	26	71	52	CON	A/G G/A	448334 448333	S288C, BY4742	YES	YES	YES	
RGSL37_13_502025	1	AACCTCTTCTGCTTTCGCGGGTCCT	25	0	104	1.00	SNV	T/C	502024	S288C	YES	YES	YES	
RGSL37_13_630506	1	TACTCAGCTACCCTACATATCGGTT	0	25	0.98	45	SNV	A/G	630505	BY4742	NO			
RGSL37_13_672130	1	GATGAAGCCGACAGCTTTAATGGCC	25	25	77	59	SNV	T/C	672129	S288C, BY4742	NO			
RGSL37_13_680941	1	AAGTTTGTAAGGTCTTTCCTCCGCA	29	29	89	80	CON	C/T T/C	680940 680936	S288C, BY4742	YES	YES	YES	
RGSL37_13_809199	1	AGAACCCTTAGTATATTCTGTGTCC	25	25	109.6	82	SNV	A/G	809198	S288C, BY4742	YES	YES	YES	
RGSL37_14_7479	7	CGTCATACCACTATGCTCTATTCCA	23	23	80.25	60.85				S288C, BY4742				
RGSL37_14_14334	2	CTAAAGGGCACAATCGTTCACGTAT	23	23	110	53	m-IN	0/C	14334	S288C, BY4742	YES	YES	YES	
RGSL37_14_89504	1	GTTGCCACTGGTCCATTTCAGTGTT	25	0	84	1.03	SNV	G/C	89503	S288C	YES	YES	YES	
RGSL37_14_189586	1	CATTCAGAAAAAGGAACAGTTAGAA	0	25	1.00	83	SNV	C/T	189585	BY4742	NO			
RGSL37_14_312332	1	ACGACGTAGTTGGGGGTTTGAAAAT	21	21	67	78	m-DEL	ACG/000	312329-312331	S288C, BY4742	YES	YES	YES	
RGSL37_14_377888	1	TGATATGATACCCATGCTTGCTCTT	25	25	74	85	SNV	G/T	377887	S288C, BY4742	YES	YES	YES	
RGSL37_14_439398	1	TTTTTTCCTTTACCCGGCAACCATA	18	18	86.5	89	m-IN	0/T	439398	S288C, BY4742	YES	YES	YES	
RGSL37_14_449107	1	GAACAATGCTAAAAGAGTTGCCCCA	25	0	80	1.02	SNV	A/T	449106	S288C	NO			
RGSL37_14_471768	1	TGAACAAATTCTCGAGGTTTTTTGA	25	25	58	77	SNV	A/G	471767	S288C, BY4742	YES	YES	YES	
RGSL37_15_60241	1	CTTCATTTTGACTTTTCCTTCTGTG	20	19	52	50	SNV	C/T	60240	S288C, BY4742	NO			
RGSL37_15_66023	1	GAGAGATTGAATAAGCACTTTGGTA	25	0	73	1.01	SNV	A/T	66022	S288C	YES	YES	YES	
RGSL37_15_79306	1	GGACTGTGTCACAGTTCCAAACGAG	33	33	83	84	CON	T/A A/T C/G	79305 79304 79297	S288C, BY4742	YES	YES	YES	
RGSL37_15_389238	1	AAGGCGAATGTTGTAAGGACTCTTG	25	0	70	1.03	SNV	T/G	389237	S288C	YES	YES	YES	
RGSL37_15_412142	1	GGTGAAGAAGACATAGAGCTAGAAG	0	25	1.00	78	SNV	G/A	412141	BY4742	NO			
RGSL37_15_722440	1	AAGCTTTGCAGAGGCTAGCAGAATT	0	207	0.98	62	DEL	DEL 182 N	722257-722439	BY4742	NO			AUX HIS 3
RGSL37_16_191945	1	ACCGTTTGATTTTAAGGAGCATAAA	25	25	90	74	SNV	C/T	191944	S288C, BY4742	YES	YES	YES	
RGSL37_16_441576	2	TTTTGCACTGGACTTAAGTTGTTGT	24	24	93	75.5	m-IN	0/G	441576	S288C, BY4742	YES	YES	YES	
RGSL37_16_759401	1	GGAATTTACCTGGATGCCGACAATG	25	25	34	85	SNV	T/G	759400	S288C, BY4742	YES	YES	YES	
RGSL37_16_890347	1	TAAGAATAATTTCGATGCCTTGGCG	25	25	70	86	SNV	T/A	890346	S288C, BY4742	YES	YES	YES	
RGSL37_17_20347	1	AACAAAGTAAGTGAAGGAGATATCT	22	23	942	271.5	m-DEL	A/0	20346	S288C, (BY4742)	No			
RGSL37_17_20935	1	CAGGAGGTGGTGACCCAATCTTATA	25	25	590	481.5	SNV	T/G	20934	S288C, BY4742	No			
RGSL37_17_39518	1	TGGTTTATTTGAAGGAGATGGTTGA	25	25	1031	580	SNV	G/T	39517	S288C, BY4742	No			
RGSL37_17_79563	1	TGTACTATTAGGTGCATGTTGACCA	0	25	0.99	1260	SNV	A/C	79562	BY4742	No			

File S6. Analysis of the SK1 *S. cerevisiae* strain. The query genome sequence reads of strain SK1 were compared with the assembled genome of strain SK1. ID, unique identifier of the downstream recovery string; CR, count reference of the downstream recovery string; SEQ, nucleotide sequence of the downstream recovery string; ZTL, zero-trail length; SV, signature value of the downstream recovery string; RG, reference genome; QG, query genome. Rows in gray indicate signatures of variation that were not solved.

ID	CR	SEQ	ZTL	SV	VARIANT (RG/QG)	POSITION
RGSL38_01_297	9	CCCGGAGATGAACATTCACAAAGAT	21	205.04	0/C	297
RGSL38_02_169	5	TCAGCATCGACAGGAATGCCGTCCA	24	357.5	T/0	168
RGSL38_02_484	8	AAGTTGCTCCATACTTTATAATACA	23	144.75	A/0	483
RGSL38_02_602	17	TAATGGCCAGGTACCAAGCATAATT	22	155.86	T/0	601
RGSL38_02_1094	20	CATCTTTATTGGCGTCCTCCTTGGC	22	188.68	0/C	1094
RGSL38_02_819104	2	TTTTTTTCAAGTTGAAGAACTGCTG	18	185.5	T/0	819104
RGSL38_02_821548	2	CTGTTGGTAAATGCGAATCTATCAC	23	116.5	0/C	821548
RGSL38_02_821982	2	TTTTTTCGGGTTCTTCATCATCAAT	19	226	T/0	821981
RGSL38_02_822566	2	GGTAGATAGTAGGTCCATCCGTTGA	23	94.75	G/0	822565
RGSL38_02_822594	2	TTATACGGACAGTATGCACATGCGT	27	178.5	0/A T/G	822594 822591
RGSL38_02_826053	1	CAACGCACACACCGACACCCACTAC	17	15.67		
RGSL38_02_826080	1	CGCAACAGCAACACCCGCACCGACT	23	43		
RGSL38_02_826118	1	CACACACGCCAACATCAACATCAAC	22	201		
RGSL38_03_205155	2	GGGAAACTGTATAAAACTTCCAAAA	21	105	0/G	205155
RGSL38_03_205219	2	TCAAGAAGGACAACATGGATGATAT	34	79.5	T/A 0/A T/A	205218 205212 205209
RGSL38_03_205608	2	GCATTACTCCACTTCAAGTAAGAGT	23	54	0/G	205608
RGSL38_03_205669	2	CTATACTAACAATTTGTAGTTCATA	23	37.5	0/C	205669
RGSL38_03_205742	2	AATAGCATAGTCGGGTTTTTCTTTT	51	90.5	G/A GTGA/0000 C/T C/A C/0 T/A	205741 205736-205739 205735 205733 205726 205715
RGSL38_03_205777	2	TTCGCGCAACAGTATAATTTTATAA	21	117		
RGSL38_04_6305	21	ATAAAAAGTTATAAATTACATTTCC	20	87.47	000/ATA	6305
RGSL38_04_6434	27	CGAATAAGAAAACAGACCCATTCAC	23	347.07	0/C	6434
RGSL38_05_609	8	AAAACATATACTATACACAATACAT	21	146.25	A/0	608
RGSL38_05_2893	18	TTTTTTGATGATGATCTGAACTCGC	18	139.57	T/0	2892
RGSL38_05_3928	23	TTTTGATCCATTTCCACATTGCAAC	21	136.14	T/0	3927
RGSL38_07_566670	1	ATCGCACATAATCAGACTCTTTATT	20	37	0000000000/ATCGTACATA	566670
RGSL38_08_535978	2	AAAAAAATAGAAAGTCTTGGTCTGT	18	101.5	A/C GC/00 A/G	535977 535975-535976 535974
RGSL38_08_537924	2	AAAAAGGAGTACCAGGCACACAATA	20	268.5	A/0	537923
RGSL38_08_538312	3	TTTTGCGATATGGCAAACCATTTTT	21	275.33	T/0	538312
RGSL38_08_538461	3	AAGACCATATCCATCATACCATTCA	23	231.67	A/0	538460
RGSL38_08_538856	3	GACGACCAGTGAGCCATATTGTGGT	24	205	G/0	538855
RGSL38_08_538931	2	AAAATGAAGTTGAGAAAGGATACTA	21	230	A/0	538930
RGSL38_08_539047	2	AAAAAAATAATACGATTGTTTCGTT	18	120	A/0	539046
RGSL38_08_539697	2	TTTGGTTAATGCAGGTACCATTCCC	34	200.5	T/0 A/0	539696 539686
RGSL38_08_539748	2	AAAATCGGAAGCCAAACAGAGATTT	37	123.25	A/0 T/0	539747 539734
RGSL38_08_541694	2	AAATTTTTTGCGGAACACTTAATCG	35	82.5	A/0 G/C C/G G/C 00/AT A/C	541693 541691 541689 541687 541683 541682

RGSL38_08_541879	5	TTTGTTTATTACAAATTAAAAGAAT	21	30.4	0/T	541879
RGSL38_08_541961	6	AGACTACAAGACCGCACTAGACTTC	21	59.89	00/AG	541961
RGSL38_08_542215	5	CTGTTGGGTGGTATCGGCTTCATTC	22	138.5	00/CT	542215
RGSL38_09_4393	5	CTGATTCTGTGGCAGAAGATGAACC	56	264.1		
RGSL38_09_230984	2	AAGATTTATTAAAATGTTAGATTTA	31	357		
RGSL38_09_446415	10	TTTTTTCTGCATACCAAGCAAGTTT	18	246.2	T/0 A/0	446414 446395
RGSL38_09_446495	13	AACCCATTGAGATAAAGTACTTTTC	19	357.92	0/A	446495
RGSL38_09_447045	24	GGCTATGGTAAGACGGAGTTATTTC	22	184.6	0/G	447045
RGSL38_09_447183	24	TTGCTTGAATGTGGCCCCTGTAAGA	23	141.83	T/0	447182
RGSL38_09_447231	24	ATGGCGTTACTGATTTATACGTGGG	25	221.15	C/0	447230
RGSL38_09_447521	6	GGGACTGGCCAAGAAGTCGATGGAC	21	266.83	G/0	447520
RGSL38_09_447611	19	TTTAATCTAATCAAGGAGAAATCCG	22	154.25	A/0	447610
RGSL38_09_447755	20	AGTTGCAAGCACAACCAACGAAGTG	24	407.35	A/0	447754
RGSL38_09_449279	19	TTTTCCTCTATGGTATTGACATATA	19	164.61	T/0	449278
RGSL38_09_449354	20	AAAGATGTTCGAGCTCTGTGTCTGT	20	269.05	0/A	449354
RGSL38_12_49498	1	AAAGAAAAGGGTGTCTAAGATGTAC	16	39	T/A	49497
RGSL38_12_407176	2	TATTTTAGCTGGTTTACTATTTAAT	16	410.5		
RGSL38_12_444013	1	TCTTTTTTCATATTTGAGTAGATCT	23	223	0/T	444013
RGSL38_14_790439	2	AAAAAAACTTCTTTTTTTTGATGAT	18	80	A/0	790438
RGSL38_14_790554	2	ACATAAGAAAGTAACTTTTACACTA	23	85.5	0/A	790554
RGSL38_14_790643	5	GAAGCGTATGACTTCGCAGCACTTT	23	271.2	0/G	790643
RGSL38_14_790697	6	TTTTTCATTTTTTTGGCGCGTCGCG	19	316.17	0/T	790697
RGSL38_14_790824	5	TTTTTGACACGTAAAATATCAAGAT	19	387.6	0/T	790824
RGSL38_15_1048309	3	AAAAAAATTCTCAGAAAATCTTTTA	17	99.33	A/0	1048308
RGSL38_15_1049212	11	GAATTACGATCTAGCTTATATGTCA	24	51.88	G/0	1049211
RGSL38_15_1049444	3	TAGCGAGATGATAATTGATGCCATA	35	62.67	T/0 A/0	1049443 1049429
RGSL38_15_1049506	3	ACAATAGTAGGTTTGTAAGGGTATA	24	38.67	A/0	1049505
RGSL38_15_1049617	3	TTCACGTTTTATTTTCTGTTATAAA	23	53.33	T/0	1049616
RGSL38_15_1049703	3	AATAATTTGATATAACATATTGAAC	22	72.33	0/A	1049703
RGSL38_15_1050167	3	TGGATTCCTAATTCCTCGAGGAGAA	29	77		
RGSL38_15_1050428	3	CCCTCGATTTTAACAAAATGGCAGG	25	194.33	T/0	1050427
RGSL38_15_1050552	3	AAAAAAGTGCGTACTCGATGACAGC	19	105.17	A/0	1050551
RGSL38_15_1050593	3	GTACCGTCATAAAACGTGTCCAAGC	23	219	0/G	1050593
RGSL38_15_1050739	3	TTTCGCTTATAAGCTAAGTAGAAGC	36	207.67	T/0 0/G	1050738 1050725
RGSL38_15_1050878	3	CAACTTTTTAGGGTCTACATATGCT	25	262.33	A/C 00/CA	1050877 1050877
RGSL38_15_1050943	3	AAACTGAGTGGACTGACTCATATAT	38	223.33	A/0 T/0	1050942 1050928
RGSL38_15_1051466	3	TTTGTGCAGTGCAAGTCCTATTGAG	21	207.67	0/T	1051466
RGSL38_15_1052869	3	GGATATTCCTTGCCAGAGCACCTAA	22	79.5	0/G	1052869
RGSL38_15_1052975	1	AGGCATTCGGTGGAAATAATTCTGT	24	113	0/C	1052975
RGSL38_16_940044	4	GGATGGATCGTGGTTCGGAGTGGCA	20	391		

File S7. Analysis of the Y12 *S. cerevisiae* strain. The query genome sequence reads of strain Y12 were compared with the assembled genome of strain Y12. ID, unique identifier of the downstream recovery string; CR, count reference of the downstream recovery string; SEQ, nucleotide sequence of the downstream recovery string; ZTL, zero-trail length; SV, signature value of the downstream recovery string; RG, reference genome; QG, query genome. Rows in gray indicate signatures of variation that were not solved.

ID	CR	SEQ	ZTL	SV	VARIANT (RG/QG)	POSITION
RGSL39_04_944773	1	TTTGTAGCTTGTCTTAACATATTTA	21	89	0/T	944773
RGSL39_04_944954	1	GGGGTTTATTCAAAAAAAAAGTCC	21	98	G/0	944953
RGSL39_04_1497133	2	CCAATCGGTGCGTTTGCGTTCATAA	23	164.5	C/0	1497132
RGSL39_04_1497195	2	AAAAAATGTGGCTACCATCAAAAAT	19	170	A/0	1497194
RGSL39_04_1497283	2	TACATTTGAAGTACTTGTATTTAAA	23	162		
RGSL39_04_1497364	2	TCATTTGGTGGAAACAACTTCCCAT	22	154.5	0000/TCTT	1497364
RGSL39_06_44	5	CCTGATTTCACCTGTCTCTCAATTT	43	49.8		
RGSL39_06_88	3	CTACTCGTTACCCTGTCTCATTCAA	26	74		
RGSL39_06_1257	5	TTTCTAGTTACAGTTACACAAAAAA	22	242.8	T/0	1256
RGSL39_06_1277	5	AAAAAACTGTGCCAACCCAAAAATT	19	133.6	A/0 T/0	1276 1256
RGSL39_06_1561	6	TTTTTCTAGAATAGTGTAAAAGTTT	20	71	T/0	1560
RGSL39_07_568237	2	CTTATCGCACATAATCTGACTCTTT	17	193	G/C	568236
RGSL39_07_1098647	8	TTTTTTATAGTGGGACATTGCAAAT	18	462.62	0/T	1098647
RGSL39_07_1098725	9	GGGTAATACATTTTGAGGGAAGGTT	21	366.44	0/G	1098725
RGSL39_07_1099275	9	TATTGATGGATTTCTTGTCAAAAAG	26	374.44	000/TTA T/G	1099275 1099273
RGSL39_07_1099622	10	AGACGAACCAGATTTCCAGGGCGGA	23	245.35	0/A	1099622
RGSL39_07_1099981	10	GCGAAAAGAAAGTCGACACAGAGCG	23	463	0/G	1099981
RGSL39_07_1100096	9	GGTGATATTTTGCATATTATCTTGC	22	207.22	0/G	1100096
RGSL39_07_1100147	10	GGTCCCCGTGGTGGCGCTGGTGACG	26	432.8	C/G G/C	1100146 1100145
RGSL39_07_1100554	10	GAGAGTTGAACAACTGCTTTTATAT	25	229.25	A/0	1100553
RGSL39_07_1101008	8	TTTGTCACTGTGCTAAGGCAAATGC	21	564.88	0/T	1101008
RGSL39_07_1101157	8	TTCTGAATACCGCGAAGGGCTGTCT	43	291.44	0/T 0/A 000/GAG	1101157 1101147 1101136
RGSL39_07_1101445	8	GCTTCCGAGCGCTGGATTCAGTGGT	23	251.06	00/GA	1101445
RGSL39_07_1102266	8	AAGATCTCAGCAGAGGTCTATCCAG	23	285.12	A/0	1102265
RGSL39_07_1102711	8	GGGGCCTCTGTTATCTATTATCTAG	20	201.81	0/G	1102711
RGSL39_07_1104078	8	AAGAGAAACAATTAGTGTATTGGAT	23	173.87	00/AG	1104078
RGSL39_07_1104158	8	AAAGAGTTGTATTATAAAGTATGGA	22	227.56	A/0	1104157
RGSL39_07_1105066	6	AAGAGGATTTCATTTTCATTTTTTT	20	238.91	000/AAG	1105066
RGSL39_07_1105228	2	TGTGTATATCTATGTCACCTTATTG	33	67		
RGSL39_07_1105386	9	TGGGACATTGCAAATCAAGGAAGTA	24	429.78	0/G	1105386
RGSL39_07_1105603	9	GAAGAAGTTGTAGGCTAAGCGCAGG	23	341.11	0/G	1105603
RGSL39_07_1106420	9	TTGATGGTCTACACGTTGTTTCAAG	23	470.67	T/0	1106419
RGSL39_07_1107053	10	TGTGCTTGTACCGCAAGGGATTTAG	23	150	0/T	1107053
RGSL39_07_1107359	10	TTTTGTTCAGAGAACCGAAGCGAGG	24	509	0/G	1107359
RGSL39_07_1107623	9	AACGTATACGTCGATGATACAACAA	23	166.96	A/0	1107622
RGSL39_08_208402	4	TTTTTTTTAAAATTTCCAAAATCTT	17	46.5	T/0	208401
RGSL39_09_233276	2	CATATTATAAATAGACAAAAGAGTC	16	60		
RGSL39_11_16007	2	TTCAATTATTTATATTATTTAAATG	37	36.5		
RGSL39_11_16292	1	TTCACTTAATTTGTTGTTATAATAA	16	22		
RGSL39_12_649	8	AAATATCACCCAATCGGTCCTTTTT	21	210.25	0/A	649
RGSL39_12_974	10	TTATAGTATCCTCTGTTGAGGTAAA	22	124.73	T/0	973
RGSL39_12_1337	9	TTTTGTTATTTAGAGCCGACTCAAA	20	154.52	0/T	1337
RGSL39_12_1788	9	AACCAGGATTCTGTTCGTTGCTCAG	23	122.04	A/0	1787
RGSL39_12_3214	8	TCCCAGTTCAAAAAGTACTGCAGCA	24	214.25	T/0	3213
RGSL39_12_3448	8	TTTCAAATATCCTACAGGGTCCCCA	22	244.12	T/0	3447
RGSL39_12_3681	8	ATAGAGGAAAAGATGTTTAATTTCG	24	222.42	A/0	3680
RGSL39_12_3783	8	TCTCTAGCAAAACCATTTGACTCCC	24	244.83	T/0	3782
RGSL39_12_4053	5	GTAGTAGCATTAGTGCTAGAGTTGA	88	579.2	DEL 64 N	3986-4052
RGSL39_12_4365	8	ACCTGTCACTGCTATTGCTCTCCTG	36	686.12		
RGSL39_12_4403	8	TAACGCAACGATCGACATGGAAGCT	24	337.25	T/0	4402
RGSL39_12_4470	7	TTCCACTGTGTCAGCAGACAGGTCT	23	337.21	T/0	4469
RGSL39_12_4503	8	AGCCACAGCATCCAACATGCTGGCC	24	272.56	A/0	4502
RGSL39_12_4817	8	AAACTCCTTTGTGCGAGACACCTTT	22	290.06	A/0	4816

RGSL39_12_4978	8	AAAAGAAGCTTCAGTGCTTCTTCGG	21	692.12	A/0	4977
RGSL39_14_101551	19	TACCTGATACAAGAACTCGACAAGA	23	97.11	0/T	101551
RGSL39_14_101656	19	AAACAGATTTTTTGGTACTAAAGCA	22	78.05	A/0	101655
RGSL39_14_101881	62	TTATATTATCAATATATTATCATAT	23	20.09	T/0	101880
RGSL39_14_102175	19	GTAGCGCCTGTGCTTCGGTTACTTC	23	172.63	0/G	102175
RGSL39_14_102329	7	CCCCATCATGCCTCTCCTCAACCTG	20	151.86	0/C	102329
RGSL39_14_102503	7	GGCCACAATCACAGTTTCCGCAGTA	22	54.38	0/G	102503
RGSL39_14_103044	8	AACAAGGTCGCATGCCAATTAATTA	24	89.56	0/T	103044

File S8. Comparison of the as-designed synIII sequence with the query genome sequence reads of strain HMSY011 from a previous and the current studies. The targeted RGSL used (RGSL6) corresponds to the as-designed synIII sequence. The targeted PMGL was constructed using the complete set of sequence reads of strain HMSY011.

ID	POSITION OF VARIANT	As-designed synIII	QG	PREVIOUSLY DETECTED	DETECTED IN THIS STUDY
RGSL6_03_32234	32219 to 32233	NA	15 N Deletion (CCAAATGCTACGCAG)	NO	YES
RGSL6_03_59578			Absence of loxPsym (upstream signal)		
RGSL6_03_59621	59583 to 59616	loxPsym	Absence of loxPsym (downstream signal) (ATAACTTCGTATAATGTACATTATACGAAGTTAT)	YES	YES
RGSL6_03_59621	59620	T	C	NO	YES
RGSL6_03_67441	67425	NA	2 N Insertion (TT)	NO	YES
RGSL6_03_67441	67427	G	T	NO	YES
RGSL6_03_67441	67431	G	T	NO	YES
RGSL6_03_67441	67434	A	C	NO	YES
RGSL6_03_67441	67437	G	T	NO	YES
RGSL6_03_67441	67440	A	C	NO	YES
RGSL6_03_68619	68618	G	T	YES	YES
RGSL6_03_71594	71594	NA	1 N Insertion (C)	YES	YES
RGSL6_03_92549	92548	C	A	YES	YES
RGSL6_03_161849	161849	NA	11 N Insertion (TAAAAATTGGT)	NO	YES
RGSL6_03_176112	176111	A	G	YES	YES
RGSL6_03_182369	182368	T	1 N Deletion (T)	YES	YES
RGSL6_03_195514	195513	G	C	YES	YES
RGSL6_03_221702	221701	T	1 N Deletion (T)	YES	YES
RGSL6_03_243007	243006	T	C	YES	YES
RGSL6_03_252435	252434	A	1 N Deletion (A)	YES	YES
RGSL6_03_272536	272535	G	1 N Deletion (G)	NO	YES

File S9. Alignments showing the variants uncovered in synIII. Reference corresponds to the as-designed synIII sequence; Query corresponds to the sequence reads of strain HMSY011.

```
RGSL6_03_32234

Reference       TGGCATTGAACCTAACGCCACTACTCCAAATGCTACGCAGCCAAATGCTACGCAGCCAAATACTA
Query           TGGCATTGAACCTAACGCCACTACT---------------CCAAATGCTACGCAGCCAAATACTA
                ************************               ************************


RGSL6_03_59621

Reference       GTCACTATTAGAGTCAGTTCGACATAACTTCGTATAATGTACATTATACGAAGTTATTGCTTAGAAGAACTGCTGGTTGTCAGGAT
Query           GTCACTATTAGAGTCAGTTCGAC---------------------------------TGCCTAGAAGAACTGCTGGTTGTCAGGAT
                **********************                                 *** ************************


RGSL6_03_67441

Reference       TGTGCATCCAGCTTACCTTCACCAA--TGGGGGCGATAGCGACCCAAAATTGGGCAATAAGAATTAA
Query           TGTGCATCCAGCTTACCTTCACCAATTTTGGGTCGCTATCGCCCCAAAATTGGGCAATAAGAATTAA
                *************************  * *** ** ** ** ************************


RGSL6_03_68619

Reference       CCATAATAGCGTAACCAATAAGTAGGCCAGCTGGGCCCGCATGAACCAACG
Query           CCATAATAGCGTAACCAATAAGTAGTCCAGCTGGGCCCGCATGAACCAACG
                ************************* ************************


RGSL6_03_71594

Reference       CAGGGAATTTAGCAGCTCTCGAAAA-CGAAGCAGCTTCTTGTTGAGCAGCA
Query           CAGGGAATTTAGCAGCTCTCGAAAACCGAAGCAGCTTCTTGTTGAGCAGCA
                ************************ ************************


RGSL6_03_92549

Reference       TTATTATGGAAGTAATGGAAATGCCCTGATAAATATGTTTAGTTGACCTAT
Query           TTATTATGGAAGTAATGGAAATGCCATGATAAATATGTTTAGTTGACCTAT
                ************************ ************************


RGSL6_03_161849

Reference       CGTTAGTGTCATTGAGAATTGTGCA-----------TAAAAATTGGTTAAAAATTGGACTA
Query           CGTTAGTGTCATTGAGAATTGTGCATAAAAATTGGTTAAAAATTGGTTAAAAATTGGACTA
                *************************          ************************


RGSL6_03_176112

Reference       ATAGGTCGTCTTGTTCAAGAAGGTAAGCGAGGACATTATCTATCAGTACAA
Query           ATAGGTCGTCTTGTTCAAGAAGGTAGGCGAGGACATTATCTATCAGTACAA
                ************************ ************************

RGSL6_03_182369

Reference       CACAACCATTTTAGATGTCTGCAGCTTTTTTTTTTTTGATTTTTTTACTAA
Query           CACAACCATTTTAGATGTCTGCAGC-TTTTTTTTTTTTGATTTTTTTACTAA
                ************************ ************************
```

RGSL6_03_195514

```
Reference       TCTCTTATTTGATCTTGGATTTCAGGATGTGCATACACTAACGATGCCAAT
Query           TCTCTTATTTGATCTTGGATTTCAGCATGTGCATACACTAACGATGCCAAT
                ************************* *************************
```

RGSL6_03_221702

```
Reference       CCTTGCTAACCAAGATATTCAAATCTGTTTCTTTTGTTGTTATATCATACA
Query           CCTTGCTAACCAAGATATTCAAATC-GTTTCTTTTGTTGTTATATCATACA
                ************************ *************************
```

RGSL6_03_243007

```
Reference       AGACTCCTTCCCCAGCACCTGCTGCTAAGATTTCTTCCCGTGTTAACGACA
Query           AGACTCCTTCCCCAGCACCTGCTGCCAAGATTTCTTCCCGTGTTAACGACA
                ************************ *************************
```

RGSL6_03_252435

```
Reference       CTCCTAGCTCCCCAAGTTATATAGCAAAGGACAGTAGAAACCTGAGTAATG
Query           CTCCTAGCTCCCCAAGTTATATAGC-AAGGACAGTAGAAACCTGAGTAATG
                ************************* *************************
```

RGSL6_03_272536

```
Reference       AGACAGGATTCAAAACCGATTAATAGGCGGCCGCCACCGCGGTGGAGCTCG
Query           AGACAGGATTCAAAACCGATTAATA-GCGGCCGCCACCGCGGTGGAGCTCG
                ************************* *************************
```

File S10. Simulation experiment in the loxPsym repeat family of synIII. The experiment is described in the text. CHR, chromosome; CR, count reference.

CHR	POSITION	CR	ALTERED RGSL	VARIATION	VARIANT FOUND	VARIANT FOUND IN OTHER COPIES
SYN III	381	198	RGSLA1	C/0	YES	NO
SYN III	9294	198	RGSLA2	C/0	YES	NO
SYN III	19489	198	RGSLA3	C/0	YES	NO
SYN III	28203	198	RGSLA4	C/0	YES	NO
SYN III	37793	198	RGSLA5	C/0	YES	NO
SYN III	48192	198	RGSLA6	C/0	YES	NO
SYN III	59587	198	RGSLA7	C/0	YES	NO
SYN III	67142	198	RGSLA8	C/0	YES	NO
SYN III	75890	198	RGSLA9	C/0	YES	NO
SYN III	88691	198	RGSLA10	C/0	YES	NO
SYN III	98713	198	RGSLA11	C/0	YES	NO
SYN III	108333	198	RGSLA12	C/0	YES	NO
SYN III	119861	198	RGSLA13	C/0	YES	NO
SYN III	129917	198	RGSLA14	C/0	YES	NO
SYN III	139008	198	RGSLA15	C/0	YES	NO
SYN III	149830	198	RGSLA16	C/0	YES	NO
SYN III	154365	198	RGSLA17	C/0	YES	NO
SYN III	168302	198	RGSLA18	C/0	YES	NO
SYN III	187480	198	RGSLA19	C/0	YES	NO
SYN III	195044	198	RGSLA20	C/0	YES	NO
SYN III	209869	198	RGSLA21	C/0	YES	NO
SYN III	219658	198	RGSLA22	C/0	YES	NO
SYN III	227410	198	RGSLA23	C/0	YES	NO
SYN III	237619	198	RGSLA24	C/0	YES	NO
SYN III	249680	198	RGSLA25	C/0	YES	NO
SYN III	254242	198	RGSLA26	C/0	YES	NO
SYN III	269811	198	RGSLA27	C/0	YES	NO

PMGL of loxPsym simulation experiment 15. Variant introduced at position 139008.

ID	CR	SEQ	PM	PMnCR	SV
RGSLA15_03_138979	1	TAGAAGGAGATAAGAAACAAATGGA	84	84.00	1.00
RGSLA15_03_138980	1	AGAAGGAGATAAGAAACAAATGGAA	84	84.00	1.00
RGSLA15_03_138981	1	GAAGGAGATAAGAAACAAATGGAAT	86	86.00	1.02
RGSLA15_03_138982	1	AAGGAGATAAGAAACAAATGGAATA	86	86.00	1.00
RGSLA15_03_138983	1	AGGAGATAAGAAACAAATGGAATAA	85	85.00	0.99
RGSLA15_03_138984	1	GGAGATAAGAAACAAATGGAATAAT	0	0.00	0.01
RGSLA15_03_138985	1	GAGATAAGAAACAAATGGAATAATT	0	0.00	1.00
RGSLA15_03_138986	1	AGATAAGAAACAAATGGAATAATTC	0	0.00	1.00
RGSLA15_03_138987	1	GATAAGAAACAAATGGAATAATTCG	0	0.00	1.00
RGSLA15_03_138988	1	ATAAGAAACAAATGGAATAATTCGT	0	0.00	1.00
RGSLA15_03_138989	1	TAAGAAACAAATGGAATAATTCGTA	0	0.00	1.00
RGSLA15_03_138990	1	AAGAAACAAATGGAATAATTCGTAT	0	0.00	1.00
RGSLA15_03_138991	1	AGAAACAAATGGAATAATTCGTATA	0	0.00	1.00
RGSLA15_03_138992	1	GAAACAAATGGAATAATTCGTATAA	0	0.00	1.00
RGSLA15_03_138993	1	AAACAAATGGAATAATTCGTATAAT	0	0.00	1.00
RGSLA15_03_138994	1	AACAAATGGAATAATTCGTATAATG	0	0.00	1.00
RGSLA15_03_138995	1	ACAAATGGAATAATTCGTATAATGT	0	0.00	1.00
RGSLA15_03_138996	1	CAAATGGAATAATTCGTATAATGTA	0	0.00	1.00
RGSLA15_03_138997	1	AAATGGAATAATTCGTATAATGTAC	0	0.00	1.00
RGSLA15_03_138998	1	AATGGAATAATTCGTATAATGTACA	0	0.00	1.00
RGSLA15_03_138999	1	ATGGAATAATTCGTATAATGTACAT	0	0.00	1.00
RGSLA15_03_139000	1	TGGAATAATTCGTATAATGTACATT	0	0.00	1.00
RGSLA15_03_139001	1	GGAATAATTCGTATAATGTACATTA	0	0.00	1.00
RGSLA15_03_139002	1	GAATAATTCGTATAATGTACATTAT	0	0.00	1.00
RGSLA15_03_139003	1	AATAATTCGTATAATGTACATTATA	0	0.00	1.00
RGSLA15_03_139004	1	ATAATTCGTATAATGTACATTATAC	0	0.00	1.00
RGSLA15_03_139005	1	TAATTCGTATAATGTACATTATACG	0	0.00	1.00
RGSLA15_03_139006	1	AATTCGTATAATGTACATTATACGA	0	0.00	1.00
RGSLA15_03_139007	1	ATTCGTATAATGTACATTATACGAA	0	0.00	1.00
RGSLA15_03_139008	197	TTCGTATAATGTACATTATACGAAG	15380	78.07	79.07
RGSLA15_03_139009	197	TCGTATAATGTACATTATACGAAGT	15371	78.03	1.00
RGSLA15_03_139010	197	CGTATAATGTACATTATACGAAGTT	15337	77.85	1.00
RGSLA15_03_139011	197	GTATAATGTACATTATACGAAGTTA	15328	77.81	1.00
RGSLA15_03_139012	197	TATAATGTACATTATACGAAGTTAT	15324	77.79	1.00

PMGL of loxPsym simulation experiment 27. No variant introduced at position 139008.

ID	CR	SEQ	PM	PMnCR	SV
RGSLA27_03_138979	1	TAGAAGGAGATAAGAAACAAATGGA	84	84.00	1.00
RGSLA27_03_138980	1	AGAAGGAGATAAGAAACAAATGGAA	84	84.00	1.00
RGSLA27_03_138981	1	GAAGGAGATAAGAAACAAATGGAAT	86	86.00	1.02
RGSLA27_03_138982	1	AAGGAGATAAGAAACAAATGGAATA	86	86.00	1.00
RGSLA27_03_138983	1	AGGAGATAAGAAACAAATGGAATAA	85	85.00	0.99
RGSLA27_03_138984	1	GGAGATAAGAAACAAATGGAATAAC	86	86.00	1.01
RGSLA27_03_138985	1	GAGATAAGAAACAAATGGAATAACT	85	85.00	0.99
RGSLA27_03_138986	1	AGATAAGAAACAAATGGAATAACTT	87	87.00	1.02
RGSLA27_03_138987	1	GATAAGAAACAAATGGAATAACTTC	90	90.00	1.03
RGSLA27_03_138988	1	ATAAGAAACAAATGGAATAACTTCG	91	91.00	1.01
RGSLA27_03_138989	1	TAAGAAACAAATGGAATAACTTCGT	91	91.00	1.00
RGSLA27_03_138990	1	AAGAAACAAATGGAATAACTTCGTA	91	91.00	1.00
RGSLA27_03_138991	1	AGAAACAAATGGAATAACTTCGTAT	91	91.00	1.00
RGSLA27_03_138992	1	GAAACAAATGGAATAACTTCGTATA	90	90.00	0.99
RGSLA27_03_138993	1	AAACAAATGGAATAACTTCGTATAA	90	90.00	1.00
RGSLA27_03_138994	1	AACAAATGGAATAACTTCGTATAAT	90	90.00	1.00
RGSLA27_03_138995	1	ACAAATGGAATAACTTCGTATAATG	87	87.00	0.97
RGSLA27_03_138996	1	CAAATGGAATAACTTCGTATAATGT	87	87.00	1.00
RGSLA27_03_138997	1	AAATGGAATAACTTCGTATAATGTA	87	87.00	1.00
RGSLA27_03_138998	1	AATGGAATAACTTCGTATAATGTAC	86	86.00	0.99
RGSLA27_03_138999	1	ATGGAATAACTTCGTATAATGTACA	85	85.00	0.99
RGSLA27_03_139000	2	TGGAATAACTTCGTATAATGTACAT	154	77.00	0.91
RGSLA27_03_139001	2	GGAATAACTTCGTATAATGTACATT	156	78.00	1.01
RGSLA27_03_139002	12	GAATAACTTCGTATAATGTACATTA	914	76.17	0.98
RGSLA27_03_139003	58	AATAACTTCGTATAATGTACATTAT	4524	78.00	1.02
RGSLA27_03_139004	197	ATAACTTCGTATAATGTACATTATA	15324	77.79	1.00
RGSLA27_03_139005	197	TAACTTCGTATAATGTACATTATAC	15328	77.81	1.00
RGSLA27_03_139006	197	AACTTCGTATAATGTACATTATACG	15337	77.85	1.00
RGSLA27_03_139007	197	ACTTCGTATAATGTACATTATACGA	15371	78.03	1.00
RGSLA27_03_139008	197	CTTCGTATAATGTACATTATACGAA	15380	78.07	1.00
RGSLA27_03_139009	197	TTCGTATAATGTACATTATACGAAG	15380	78.07	1.00
RGSLA27_03_139010	197	TCGTATAATGTACATTATACGAAGT	15371	78.03	1.00
RGSLA27_03_139011	197	CGTATAATGTACATTATACGAAGTT	15337	77.85	1.00
RGSLA27_03_139012	197	GTATAATGTACATTATACGAAGTTA	15328	77.81	1.00

PMGL of loxPsym simulation experiment 27. Variant introduced at position 269811.

ID	CR	SEQ	PM	PMnCR	SV
RGSLA27_03_269782	1	AGGATGGAGGAAGAAATCGATACCA	80	80.00	0.99
RGSLA27_03_269783	1	GGATGGAGGAAGAAATCGATACCAA	79	79.00	0.99
RGSLA27_03_269784	1	GATGGAGGAAGAAATCGATACCAAT	79	79.00	1.00
RGSLA27_03_269785	1	ATGGAGGAAGAAATCGATACCAATA	78	78.00	0.99
RGSLA27_03_269786	1	TGGAGGAAGAAATCGATACCAATAA	79	79.00	1.01
RGSLA27_03_269787	1	GGAGGAAGAAATCGATACCAATAAT	0	0.00	0.01
RGSLA27_03_269788	1	GAGGAAGAAATCGATACCAATAATT	0	0.00	1.00
RGSLA27_03_269789	1	AGGAAGAAATCGATACCAATAATTC	0	0.00	1.00
RGSLA27_03_269790	1	GGAAGAAATCGATACCAATAATTCG	0	0.00	1.00
RGSLA27_03_269791	1	GAAGAAATCGATACCAATAATTCGT	0	0.00	1.00
RGSLA27_03_269792	1	AAGAAATCGATACCAATAATTCGTA	0	0.00	1.00
RGSLA27_03_269793	1	AGAAATCGATACCAATAATTCGTAT	0	0.00	1.00
RGSLA27_03_269794	1	GAAATCGATACCAATAATTCGTATA	0	0.00	1.00
RGSLA27_03_269795	1	AAATCGATACCAATAATTCGTATAA	0	0.00	1.00
RGSLA27_03_269796	1	AATCGATACCAATAATTCGTATAAT	0	0.00	1.00
RGSLA27_03_269797	1	ATCGATACCAATAATTCGTATAATG	0	0.00	1.00
RGSLA27_03_269798	1	TCGATACCAATAATTCGTATAATGT	0	0.00	1.00
RGSLA27_03_269799	1	CGATACCAATAATTCGTATAATGTA	0	0.00	1.00
RGSLA27_03_269800	1	GATACCAATAATTCGTATAATGTAC	0	0.00	1.00
RGSLA27_03_269801	1	ATACCAATAATTCGTATAATGTACA	0	0.00	1.00
RGSLA27_03_269802	1	TACCAATAATTCGTATAATGTACAT	0	0.00	1.00
RGSLA27_03_269803	1	ACCAATAATTCGTATAATGTACATT	0	0.00	1.00
RGSLA27_03_269804	1	CCAATAATTCGTATAATGTACATTA	0	0.00	1.00
RGSLA27_03_269805	1	CAATAATTCGTATAATGTACATTAT	0	0.00	1.00
RGSLA27_03_269806	1	AATAATTCGTATAATGTACATTATA	0	0.00	1.00
RGSLA27_03_269807	1	ATAATTCGTATAATGTACATTATAC	0	0.00	1.00
RGSLA27_03_269808	1	TAATTCGTATAATGTACATTATACG	0	0.00	1.00
RGSLA27_03_269809	1	AATTCGTATAATGTACATTATACGA	0	0.00	1.00
RGSLA27_03_269810	1	ATTCGTATAATGTACATTATACGAA	0	0.00	1.00
RGSLA27_03_269811	197	TTCGTATAATGTACATTATACGAAG	15380	78.07	79.07
RGSLA27_03_269812	197	TCGTATAATGTACATTATACGAAGT	15371	78.03	1.00
RGSLA27_03_269813	197	CGTATAATGTACATTATACGAAGTT	15337	77.85	1.00
RGSLA27_03_269814	197	GTATAATGTACATTATACGAAGTTA	15328	77.81	1.00
RGSLA27_03_269815	197	TATAATGTACATTATACGAAGTTAT	15324	77.79	1.00

PMGL of loxPsym simulation experiment 15. No variant introduced at position 269811.

ID	CR	SEQ	PM	PMnCR	SV
RGSLA15_03_269782	1	GGATGGAGGAAGAAATCGATACCAA	79	79.00	0.99
RGSLA15_03_269783	1	GATGGAGGAAGAAATCGATACCAAT	79	79.00	1.00
RGSLA15_03_269784	1	ATGGAGGAAGAAATCGATACCAATA	78	78.00	0.99
RGSLA15_03_269785	1	TGGAGGAAGAAATCGATACCAATAA	79	79.00	1.01
RGSLA15_03_269786	1	GGAGGAAGAAATCGATACCAATAAC	81	81.00	1.02
RGSLA15_03_269787	1	GAGGAAGAAATCGATACCAATAACT	84	84.00	1.04
RGSLA15_03_269788	1	AGGAAGAAATCGATACCAATAACTT	81	81.00	0.96
RGSLA15_03_269789	1	GGAAGAAATCGATACCAATAACTTC	79	79.00	0.98
RGSLA15_03_269790	1	GAAGAAATCGATACCAATAACTTCG	77	77.00	0.97
RGSLA15_03_269791	1	AAGAAATCGATACCAATAACTTCGT	76	76.00	0.99
RGSLA15_03_269792	1	AGAAATCGATACCAATAACTTCGTA	77	77.00	1.01
RGSLA15_03_269793	1	GAAATCGATACCAATAACTTCGTAT	78	78.00	1.01
RGSLA15_03_269794	1	AAATCGATACCAATAACTTCGTATA	77	77.00	0.99
RGSLA15_03_269795	1	AATCGATACCAATAACTTCGTATAA	76	76.00	0.99
RGSLA15_03_269796	1	ATCGATACCAATAACTTCGTATAAT	76	76.00	1.00
RGSLA15_03_269797	1	TCGATACCAATAACTTCGTATAATG	75	75.00	0.99
RGSLA15_03_269798	1	CGATACCAATAACTTCGTATAATGT	74	74.00	0.99
RGSLA15_03_269799	1	GATACCAATAACTTCGTATAATGTA	74	74.00	1.00
RGSLA15_03_269800	1	ATACCAATAACTTCGTATAATGTAC	75	75.00	1.01
RGSLA15_03_269801	1	TACCAATAACTTCGTATAATGTACA	74	74.00	0.99
RGSLA15_03_269802	2	ACCAATAACTTCGTATAATGTACAT	172	86.00	1.16
RGSLA15_03_269803	2	CCAATAACTTCGTATAATGTACATT	171	85.50	0.99
RGSLA15_03_269804	11	CAATAACTTCGTATAATGTACATTA	807	73.36	0.86
RGSLA15_03_269805	58	AATAACTTCGTATAATGTACATTAT	4524	78.00	1.06
RGSLA15_03_269806	197	ATAACTTCGTATAATGTACATTATA	15324	77.79	1.00
RGSLA15_03_269807	197	TAACTTCGTATAATGTACATTATAC	15328	77.81	1.00
RGSLA15_03_269808	197	AACTTCGTATAATGTACATTATACG	15337	77.85	1.00
RGSLA15_03_269809	197	ACTTCGTATAATGTACATTATACGA	15371	78.03	1.00
RGSLA15_03_269810	197	CTTCGTATAATGTACATTATACGAA	15380	78.07	1.00
RGSLA15_03_269811	197	TTCGTATAATGTACATTATACGAAG	15380	78.07	1.00
RGSLA15_03_269812	197	TCGTATAATGTACATTATACGAAGT	15371	78.03	1.00
RGSLA15_03_269813	197	CGTATAATGTACATTATACGAAGTT	15337	77.85	1.00
RGSLA15_03_269814	197	GTATAATGTACATTATACGAAGTTA	15328	77.81	1.00
RGSLA15_03_269815	197	TATAATGTACATTATACGAAGTTAT	15324	77.79	1.00

CÓDIGO FUENTE DE LA RUTA COMPUTACIONAL PMGL

Nota: el código fuente se encuentra disponible en el repositorio público GitHub:
https://github.com/LIIGH-UNAM/PerfectMatchGenomicLandscapePipeline.git

README

PerfectMatchGenomicLandscapePipeline
Precise detection of genome variation using the Perfect Match Genomic Landscape Pipeline

---
AUTHORS

Kim Palacios-Flores
Jair Garcia-Sotelo

Contact information:

Kim Palacios Flores, kimpalaciosflores@gmail.com
Jair Santiago Garcia Sotelo, jsgarcia@liigh.unam.mx

PMGL PIPELINE

This pipeline is used to reveal signatures of variation embedded in a Perfect Match Genomic Landscape (PMGL), discover the underlying variants, generate a customized Reference Genome, and validate the precise location and nature of the introgressed variants. The PMGL pipeline is divided into six modules:

1) Generation of a Reference Genome Self Landscape (RGSL).
2) Generation of a Perfect Match Genomic Landscape (PMGL).
3) Scanning of the PMGL using the zero-trail scan. Pinpoints the position of variants.
4) Generation of the first alignment at each signature of variation using the MUSCLE Multiple Sequence Alignment tool.
5) Orchestrates the interpretation and extension of alignments. Discovers the nature of variants.
6) Generation of a customized Reference Genome. The customized Reference Genome sequence is validated by performing steps 1), 2), and 3) using the customized Reference Genome and the original Query Genome sequence reads.

Cluster configuration

System of Cluster Administration: Bright Cluster Manager 7.1
S.O. Centos 7 x86_64
Resources Scheduler: SGE (Sun Grid Engine) 2011.11p1
Environment Modules: 3.2.10
Internal Management Network: 1GbE
Internal Infiniband Network: Infiniband 40Gbps

Requirements

R version 3.3.2
R libraries:
- stringr
- seqinr
- stringi
- readtext
- optparse

Perl version 5.16.3
Perl libraries:
- Getopt::Long;
- LWP::Simple;
- File::Copy;
- Math::Round;
- Math::BigFloat;

Bowtie 0.12.7
Jellyfish 1.1.10

Internet access is required to execute the Client of MUSCLE Multiple Sequence Alignment tool. IMPORTANT NOTE: The Client of MUSCLE Multiple Sequence Alignment tool was modified by Jair Santiago Garcia Sotelo and Kim Palacios Flores on 02/11/2017 to re-execute MUSCLE using a new JOB ID when it is not responding. The modified Client of MUSCLE Multiple Sequence Alignment tool is integrated into the PMGL pipeline scripts.

Generate RGSL

DESCRIPTION

Generates a Reference Genome Self Landscape (RGSL) from a Reference Genome sequence in fasta format.

USAGE

The only script that needs to be executed by the user for this module is the following:

perl generateRGSL_v1.0.pl -binDir /url/ -fastaDir /url/RG#_#.fasta -bowtieDir /url/ -outputDir /url/ -kmerLength # -rgslId RGSL# -memory #

OUTPUT

RGSL#.tab file. The RGSL reports each Reference String's unique identifier (ID), the number of times its sequence is present in the entire Reference Genome (CR), its DNA sequence (SEQ), and the unique identifiers of all Reference Strings in the entire Reference Genome that share the same sequence (IDF).

Generates four types of output directories:

bowtie: Contains the comparison files between the bowtie database and the chromosome kmers fasta files.
fasta: Contains Reference Genome fasta file.
kmers: Contains the chromosome kmers fasta files.
LN: Contains the chromosome landscapes.
RGSL: Contains the Reference Genome Self Landscape.

Scripts comprising the Generate RGSL module:

generateRGSL_v1.0.pl
makeSGE_v1.0.pl
makeRGKmers_v1.0.pl
makeRGSLFile_v1.0.pl
orderFamily_v1.0.pl

Generate PMGL

DESCRIPTION

Generates a Perfect Match Genomic Landscape (PMGL) from a Reference Genome Self Landscape (RGSL) and a Query Genome (sequence reads in fastq format).

USAGE

The only script that needs to be executed by the user for this module is the following:

perl generatePMGL_v1.0.pl -binDir /url/ -fastqFile /url/QG#.fastq -fastaRgDir /url/RG#_#.fasta -lnRgDir /url/ -outputDir /url/ -kmerLength # -pmglId PMGL# -memory #

OUTPUT

PMGL#_RG#_all_PMnCR_SV.tab file. The PMGL reports the RGSL (except the IDF column) plus the number of perfect match occurrences of each Reference String in
the Read Strings Dataset (PM), each Reference String's PM normalized by its CR (PMnCR), each Reference String's signature value (SV).

Generates four types of output directories:

BD: Contains the jellyfish database.
kmerCovPlot: Contains the comparison files between the Reference Genome chromosome fasta files and the jellyfish database.
landscape: Contains the chromosome landscapes.
PMGL: Contains the Perfect Match Genomic Landscape.

Scripts comprising the Generate PMGL module:

generatePMGL_v1.0.pl
makeSGE_v1.0.pl
jointRefKmer-cov_v1.0.pl
normalizedByCountReference_v1.0.pl
signatureValue_v1.0.pl

Generate PMGL Zero Trail Scan

DESCRIPTION

Locates signatures of variation along a Perfect Match Genomic Landscape (PMGL) using the zero-trail scan.

USAGE

The only script that needs to be executed by the user for this module is the following:

perl PMGLZeroTrailScan.pl -inputFile PMGL#_RG#_all_PMnCR_SV.tab -minPMn # -maxPMn_1 # -CRn_1 # -lowComplexity -zeroTrail

OUTPUT

PMGL#_RG#_all_PMnCR_SV_ZeroTrailScan.tab File. Contains the PMGL rows corresponding to the Downstream Recovery String of each signature of variation plus its zero-trail length (last column).

Scripts comprising the PMGL Zero Trail Scan module:

PMGLZeroTrailScan_V0.3.pl

Generate First Alignment

DESCRIPTION

For each signature of variation located using the zero-trail scan, generates an alignment anchored by the Downstream Recovery String.

USAGE

The only script that needs to be executed by the user for this module is the following:

perl generateAligment_v1.0.pl -binDir /url/ -outputDir /url/ -zeroTrailScanFile PMGL#_RG#_all_PMnCR_SV_ZeroTrailScan.tab -rgslFile /URL/RGSL#.tab -fastqFile /url/QG#.fastq -rawRgDir /url/ -rgId RG# -minCountFamily # -memory # -kmerLength # -maxFamily #

OUTPUT

Generates four types of output directories:

alignment: Contains the formatted alignment files.
bash: Contains the script files for JOB execution.
family: Contains the different Query Genome sequences (Read Families) that contain a perfect match with the Downstream Recovery String, and that have a minimum number of occurrences in the Query Genome.
muscle: Contains the alignment files generated by the MUSCLE Multiple Sequence Alignment tool.
reads: Contains the Query Genome sequence reads containing a perfect match with the Downstream Recovery String.

Scripts comprising the Generate Alignment module:

generateAlignment_v1.0.pl
makeSGE_v1.0.pl
findReads_v1.0.sh
catAlignment_v1.0.pl
muscle_lwp.pl

Interpretation and Extension of Alignments

DESCRIPTION

Orchestrates iterative alignment extensions at signatures of variation to uncover the underlying variants. Signatures of variation are classified as either solved or unsolved.

USAGE

The only script that needs to be executed by the user for this module is the following:

Rscript --vanilla /url/organizer_v1.0.R --binDir /url/ --rawRgDir /url/ --readsDir /url/ --familyDir /url/ --muscleDir /url/ --alignmentDir /url/ --organizerDir /url/ --rgslFile /url/RGSL#.tab --fastqFile /url/QG#.fastq --organizerName organizer --rgId RG# --kmerLength # --minCountFamily # --maxFamily # --minAnchor # --maxPmWalk #

OUTPUT

1) A file reporting the final status (solved or unsolved plus additional information) of each signature of variation.

2) An R object containing the final status report for all signatures of variation, the final alignment report for solved signatures of variation, and a partial report for unsolved signatures of variation.

Scripts comprising the Interpretation and Extension of Alignments module:

organizer_v1.0.R

Note: It is recommended to execute this script in a computer node.

Customization

DESCRIPTION

Generates a new, customized Reference Genome using the Common Variation Motifs found for solved signatures of variation by the organizer_v1.0.R program.

USAGE

The only script that needs to be executed by the user for this module is the following:

Rscript --vanilla /url/customization_v1.0.R --rawRgDir /url/ --custRgDir /url/ --organizerFile /url/organizer.rds --kmerLength # --rgId RG# --custRgId RG#

OUTPUT

1) A file reporting all sequence changes performed by the customization process decomposed into single nucleotide changes, deletions, or insertions.
Sequence changes are reported in the order that they were effectuated, from the more downstream positions to the more upstream ones within each chromosome. NOTE: In the case where an alignment has invaded another alignment located further upstream, and each alignment instructs a different change to be made on the same nucleotide(s) from the Reference Genome to match the Query Genome, such nucleotide(s) is not customized. Additionally, a file reporting all uncustomized Reference Genome positions due to conflicting alignment changes is generated.

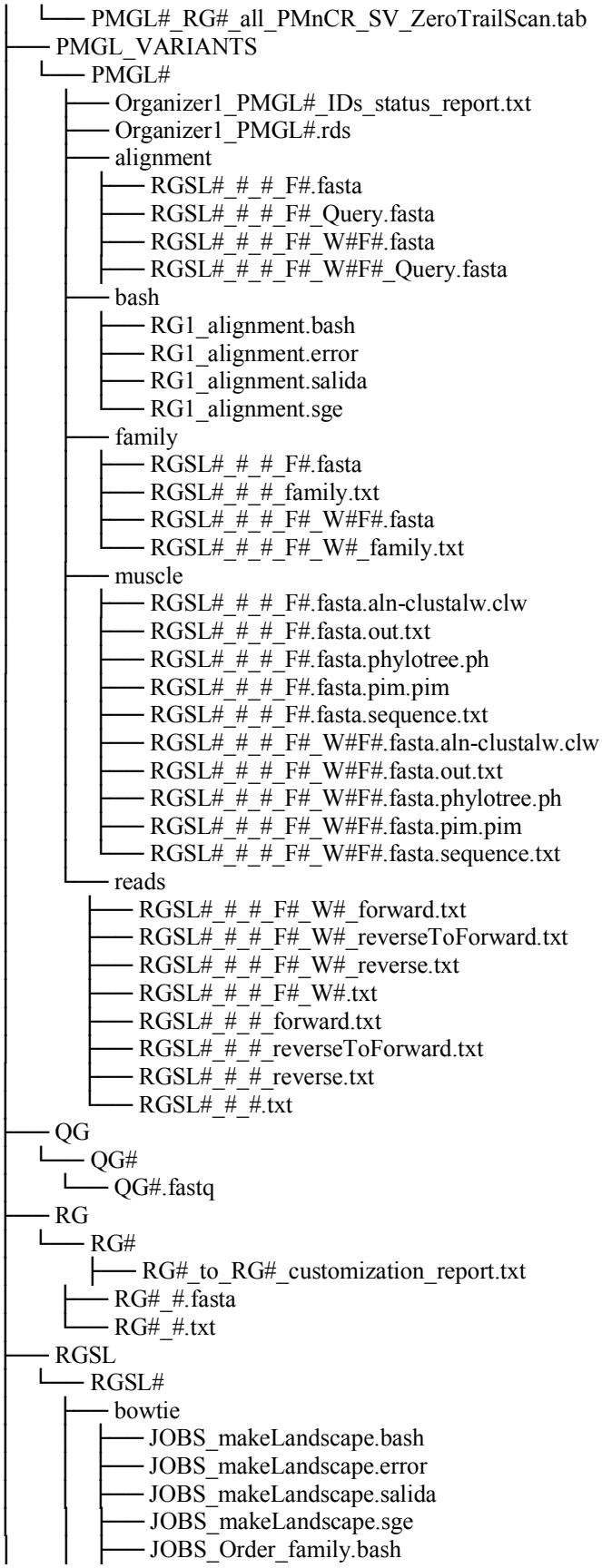
2) The new, customized Reference Genome sequence per chromosome in fasta and txt format.

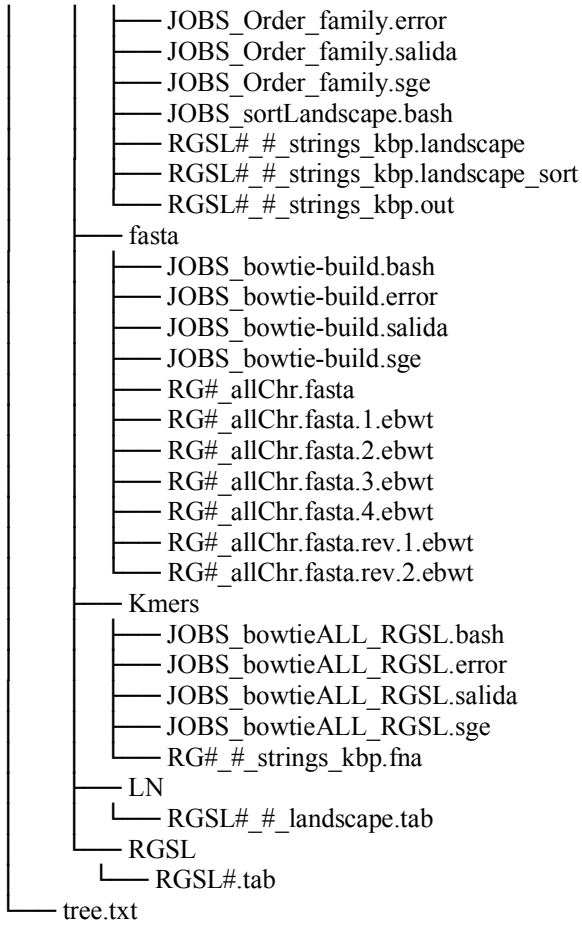
Scripts comprising the Customization module:

customization_v1.0.R

Directory structure

```
.
├── bin
│   ├── catAlignment_v1.0.pl
│   ├── customization_v1.0.R
│   ├── findReads_v1.0.sh
│   ├── generateAlignment_v1.0.pl
│   ├── generateBashAligment_v1.0.pl
│   ├── generatePMGL_v1.0.pl
│   ├── generatePMWalk_v1.0.pl
│   ├── generateRGSL_v1.0.pl
│   ├── jointRefKmer-cov_v1.0.pl
│   ├── library.pm
│   ├── makeRGKmers_v1.0.pl
│   ├── makeRGSLFile_v1.0.pl
│   ├── makeSGE_v1.0.pl
│   ├── muscle_lwp.pl
│   ├── normalizedByCountReference_v1.0.pl
│   ├── orderFamily_v1.0.pl
│   ├── organizer_v1.0.R
│   ├── PMGLZeroTrailScan_V0.3.pl
│   └── signatureValue_v1.0.pl
├── PMGL
│   └── PMGL#
│       ├── BD
│       │   ├── JOBS_makeDB_Jellifish.bash
│       │   ├── JOBS_makeDB_Jellifish.error
│       │   ├── JOBS_makeDB_Jellifish.salida
│       │   ├── JOBS_makeDB_Jellifish.sge
│       │   └── PMGL#_BD_0
│       ├── kmerCovPlot
│       │   ├── JOBS_Final_LN.bash
│       │   ├── JOBS_Final_LN.error
│       │   ├── JOBS_Final_LN.salida
│       │   ├── JOBS_Final_LN.sge
│       │   ├── JOBS_kmer-cov-plot.bash
│       │   ├── JOBS_kmer-cov-plot.salida
│       │   └── PMGL#_RG#_#.kmer-cov-plot
│       ├── landscape
│       │   └── PMGL#_RG#_#_landscape.tab
│       └── PMGL
│           ├── JOBS_normalized_string.bash
│           ├── JOBS_normalized_string.error
│           ├── JOBS_normalized_string.salida
│           ├── JOBS_normalized_string.sge
│           ├── JOBS_signature_value.bash
│           ├── JOBS_signature_value.error
│           ├── JOBS_signature_value.salida
│           ├── JOBS_signature_value.sge
│           ├── PMGL#_RG#_all_PMnCR_SV.tab
│           ├── PMGL#_RG#_all_PMnCR.tab
│           └── PMGL#_RG#_all.tab
└── PMGL_SCAN
```


└── PMGL#_RG#_all_PMnCR_SV_ZeroTrailScan.tab
PMGL_VARIANTS
└── PMGL#
Organizer1_PMGL#_IDs_status_report.txt
Organizer1_PMGL#.rds
alignment
RGSL#_#_#_F#.fasta
RGSL#_#_#_F#_Query.fasta
RGSL#_#_#_F#_W#F#.fasta
RGSL#_#_#_F#_W#F#_Query.fasta
bash
RG1_alignment.bash
RG1_alignment.error
RG1_alignment.salida
RG1_alignment.sge
family
RGSL#_#_#_F#.fasta
RGSL#_#_#_family.txt
RGSL#_#_#_F#_W#F#.fasta
RGSL#_#_#_F#_W#_family.txt
muscle
RGSL#_#_#_F#.fasta.aln-clustalw.clw
RGSL#_#_#_F#.fasta.out.txt
RGSL#_#_#_F#.fasta.phylotree.ph
RGSL#_#_#_F#.fasta.pim.pim
RGSL#_#_#_F#.fasta.sequence.txt
RGSL#_#_#_F#_W#F#.fasta.aln-clustalw.clw
RGSL#_#_#_F#_W#F#.fasta.out.txt
RGSL#_#_#_F#_W#F#.fasta.phylotree.ph
RGSL#_#_#_F#_W#F#.fasta.pim.pim
RGSL#_#_#_F#_W#F#.fasta.sequence.txt
reads
RGSL#_#_#_F#_W#_forward.txt
RGSL#_#_#_F#_W#_reverseToForward.txt
RGSL#_#_#_F#_W#_reverse.txt
RGSL#_#_#_F#_W#.txt
RGSL#_#_#_forward.txt
RGSL#_#_#_reverseToForward.txt
RGSL#_#_#_reverse.txt
RGSL#_#_#.txt
QG
└── QG#
└── QG#.fastq
RG
└── RG#
RG#_to_RG#_customization_report.txt
RG#_#.fasta
RG#_#.txt
RGSL
└── RGSL#
bowtie
JOBS_makeLandscape.bash
JOBS_makeLandscape.error
JOBS_makeLandscape.salida
JOBS_makeLandscape.sge
JOBS_Order_family.bash


JOBS_Order_family.error
JOBS_Order_family.salida
JOBS_Order_family.sge
JOBS_sortLandscape.bash
RGSL#_#_strings_kbp.landscape
RGSL#_#_strings_kbp.landscape_sort
RGSL#_#_strings_kbp.out
fasta
JOBS_bowtie-build.bash
JOBS_bowtie-build.error
JOBS_bowtie-build.salida
JOBS_bowtie-build.sge
RG#_allChr.fasta
RG#_allChr.fasta.1.ebwt
RG#_allChr.fasta.2.ebwt
RG#_allChr.fasta.3.ebwt
RG#_allChr.fasta.4.ebwt
RG#_allChr.fasta.rev.1.ebwt
RG#_allChr.fasta.rev.2.ebwt
Kmers
JOBS_bowtieALL_RGSL.bash
JOBS_bowtieALL_RGSL.error
JOBS_bowtieALL_RGSL.salida
JOBS_bowtieALL_RGSL.sge
RG#_#_strings_kbp.fna
LN
RGSL#_#_landscape.tab
RGSL
RGSL#.tab
tree.txt

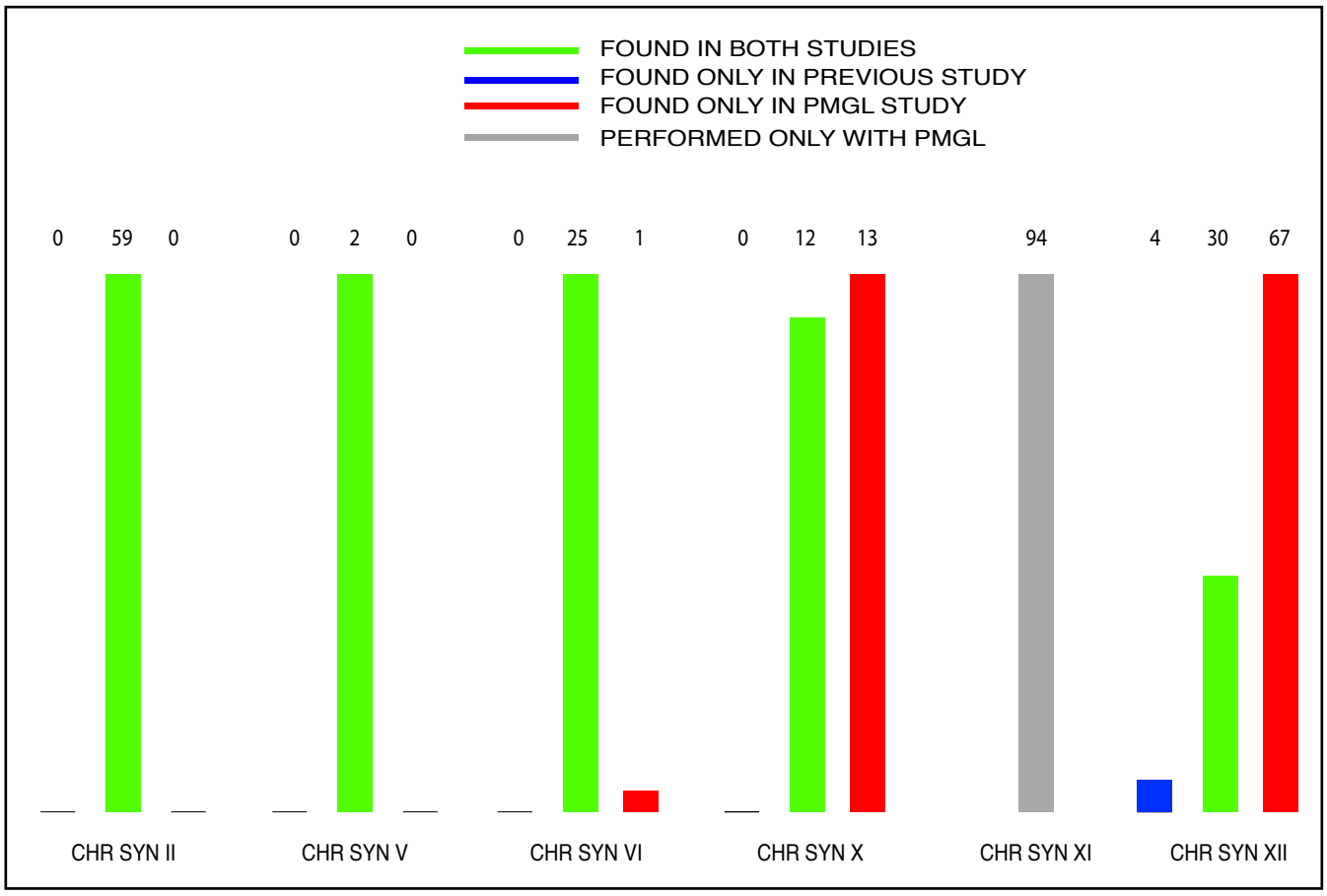
RESULTADOS NO PUBLICADOS

En el trabajo publicado "**A Perfect Match Genomic Landscape Provides a Unified Framework for the Precise Detection of Variation in Natural and Synthetic Haploid Genomes"** por **Kim Palacios-Flores *et al***., se ha demostrado que la estrategia PMGL es particularmente adecuada para el análisis de cromosomas sintéticos. En el contexto de la publicación se analizó el perfil de variación del cromosoma syn III incorporado en la cepa HMSY011 de *S. cerevisiae* en relación a su propia secuencia diseñada. Como se menciona en el artículo, el PMGL reveló todas las variantes anteriormente reportadas por el grupo que sintetizó dicho cromosoma (Annaluru et al., 2014). El que la estrategia PMGL haya, adicionalmente, descubierto variantes previamente no identificadas causó gran interés dentro de la comunidad de científicos trabajando en la construcción de distintos cromosomas sintéticos como parte del proyecto internacional Sc2.0 (Richardson et al., 2017). Este proyecto tiene como objetivo la síntesis completa del genoma de *S. cerevisiae.* El Dr. Boeke, director del proyecto Sc2.0, solicitó a la autora de la presente tesis su colaboración en el análisis de distintos cromosomas sintéticos.

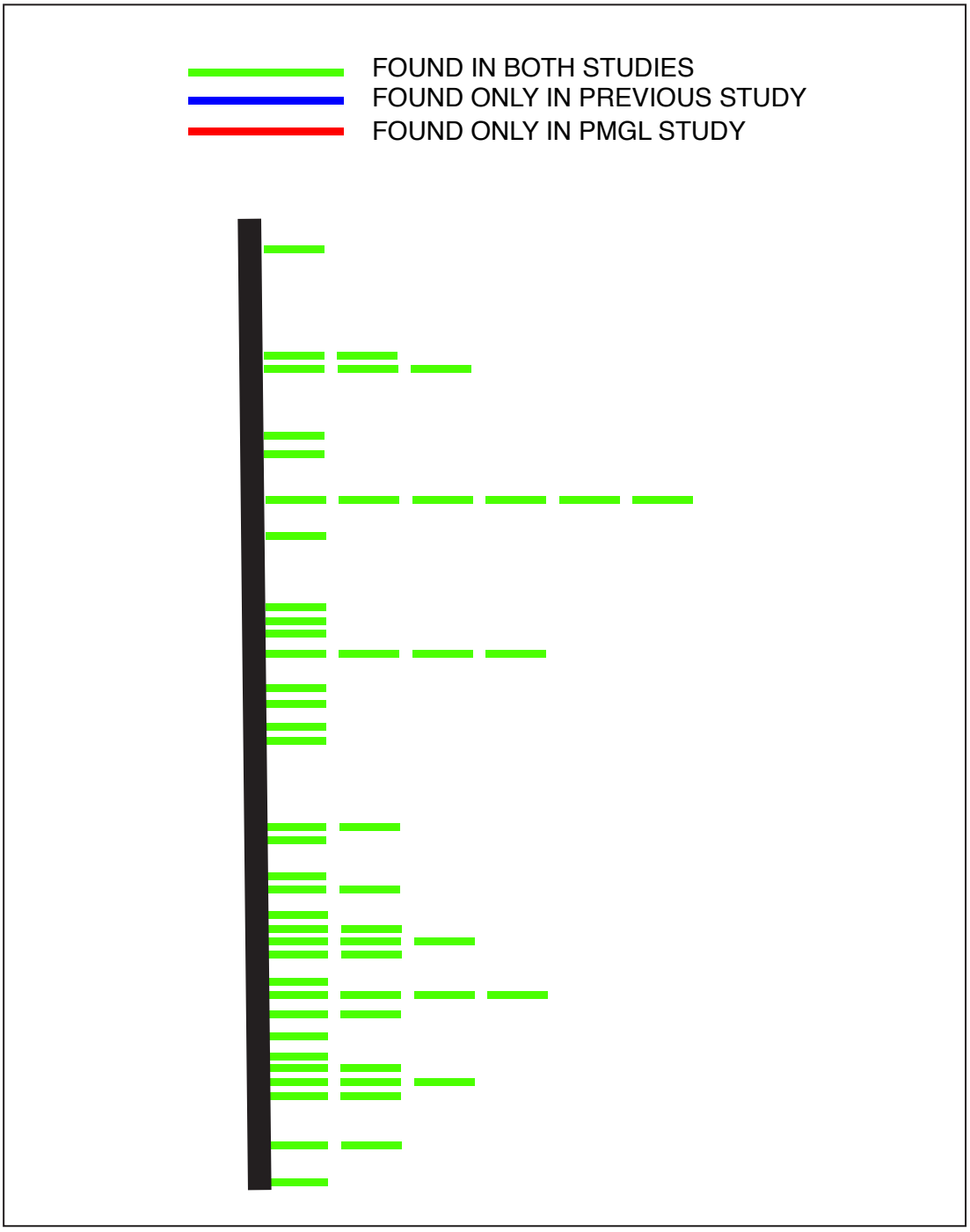
Hasta la fecha de la entrega de la tesis, hemos analizado, además del cromosoma syn III, los cromosomas syn II, syn V, syn VI, syn X, syn XI y syn XII. La Figura 1 de la sección de Resultados No Publicados muestra un panorama general de los datos obtenidos para todos los cromosomas. Es pertinente aclarar que, en el área de biología sintética, el término "secuencia viviente" o "secuencia física" se utiliza para distinguir entre la secuencia del cromosoma sintetizado que reside en la célula y la secuencia del cromosoma tal como fue diseñada *in silico*. El número de variantes entre ambas secuencias refleja el nivel de exactitud de los procesos experimentales que conllevan a la creación de cromosomas sintéticos (Xie *et al*., 2017). Para algunos de los cromosomas sintéticos analizados, la secuencia del cromosoma viviente muestra una identidad casi absoluta con la secuencia del cromosoma diseñado. En efecto, el análisis del cromosoma syn V (Xie *et al*., 2017), reveló tan sólo dos variantes, ambas encontradas en el estudio original y en el análisis por PMGL. El cromosoma syn II (Shen et al., 2017) presentó 59 variantes, todas encontradas en el estudio original y en nuestro análisis. La totalidad de las variantes encontradas en estudios anteriores para los cromosomas syn VI (Mitchell et al., 2017) y syn X (Wu et al., 2017) fueron reveladas por la estrategia PMGL. Adicionalmente, el PMGL reveló nuevas variantes. En el caso del cromosoma syn XII (Zhang *et al*., 2017), el análisis anterior encontró 34 variantes. El PMGL encontró 30 de éstas últimas. Las 4 variantes restantes no fueron encontradas debido a los parámetros utilizados para el escaneo del genoma (ver explicación en la nota de la tabla 6 de la sección de Resultados No Publicados). Sin embargo, para este cromosoma, el PMGL descubrió 67 variantes previamente no reportadas. El cromosoma syn XI es un caso interesante, ya que el grupo encargado de su síntesis solicitó a la autora de la tesis analizarlo por la estrategia PMGL con objeto de guiar el proceso de reparación ("debugging") del cromosoma viviente. En este caso, el PMGL reveló la presencia de 94 variantes. En resumen, es posible afirmar que el PMGL representa una estrategia ideal para determinar el perfil de variación de cromosomas artificiales que coexisten con cromosomas naturales, formando parte de genomas híbridos de microorganismos. La estrategia de análisis de variación genómica por PMGL permite, tanto asistir en la construcción de un cromosoma viviente que posea una secuencia de ADN idéntica a la secuencia diseñada, como refinar la secuencia diseñada para que represente fielmente la secuencia del cromosoma viviente.

Las Figuras 2-7 de la sección de Resultados No Publicados muestran la posición de las variantes encontradas en estudios previos y utilizando la estrategia PMGL en los cromosomas

analizados. La naturaleza específica de las variantes se describe en las tablas 1-6 de la sección de Resultados No Publicados. Tanto las figuras como las tablas se presentan en el idioma inglés, ya que se espera que formen parte de publicaciones futuras.



Unpublished Figure 1. Six synthetic chromosomes were analyzed using the PMGL strategy. Five of them have been previously analyzed. Chromosome XI has only been analyzed using the PMGL strategy. The columns for each chromosome schematize the number of variants found according to the color code presented at the top. For each chromosome, the column presenting the highest number of variants was adjusted to the same size. The actual number of variants for each column is indicated.



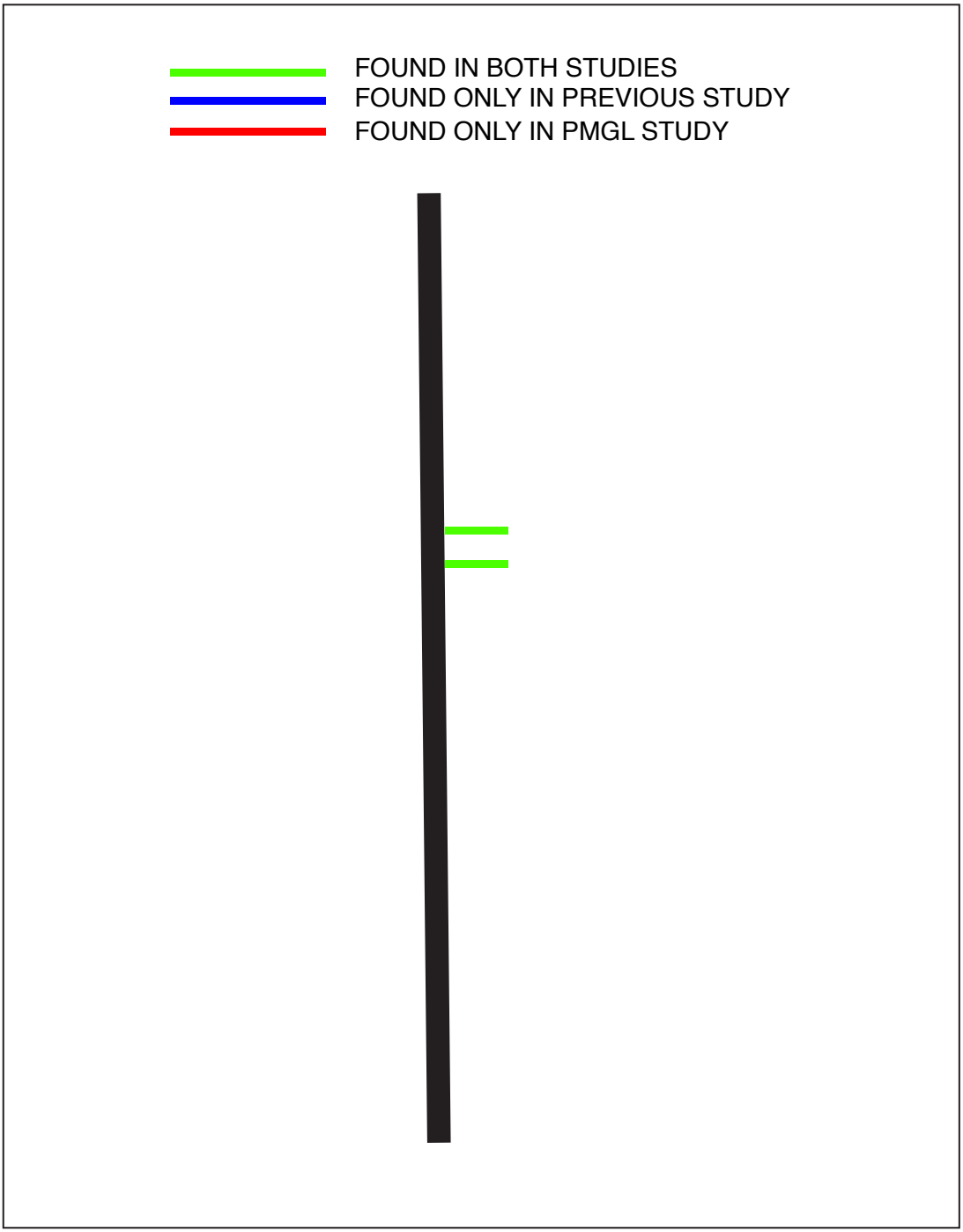
Unpublished Figure 2. Comparison of the physical syn II with the as-designed syn II. The data indicate the position of variants found in a previous study and in the present study (PMGL). Each line represents the position of one variant along the chromosome. The as-designed chromosome is represented as a black bar; nucleotide 1 (bottom); nucleotide 770,036 (top). Closely spaced variants are represented by rows of lines. The color code indicates the study in which the variant was detected. The nature of the variants is indicated in Unpublished Table 1.

Unpublished Table 1. Comparison of the as-designed synII sequence with the physical synII sequence from a previous analysis and in this study. The parameters used for the zero-trail scan were: min PMn at position n = 15; max PMn at position n-1 = 5

ID	POSITION OF VARIANT	As-designed synII	Physical synII	PREVIOUSLY DETECTED	DETECTED IN THIS STUDY
RGSL13_02_16710	16709	G	1 nucleotide deletion (G)	YES	YES
RGSL13_02_38933	38932	A	G	YES	YES
RGSL13_02_39074	39073	T	C	YES	YES
RGSL13_02_79643	79639	T	G	YES	YES
	79642	G	A	YES	YES
RGSL13_02_86949	86939	C	T	YES	YES
	86945	G	A	YES	YES
	86948	C	T	YES	YES
RGSL13_02_95248	95247	A	G	YES	YES
RGSL13_02_95581	95580	C	T	YES	YES
RGSL13_02_102041	102040	C	T	YES	YES
RGSL13_02_126625	126624	C	T	YES	YES
RGSL13_02_135178	135176	G	T	YES	YES
	135177	A	C	YES	YES
RGSL13_02_151431	151427	T	C	YES	YES
	151430	G	A	YES	YES
*RGSL13_02_151947	151950-151983	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
*RGSL13_02_151984					
RGSL13_02_153843	153850-153883	LoxP site	Absence of loxPsym (34 nucleotide deletion) (ATAACTTCGTATAATGTACATTATACGAAGTTAT)	YES	YES
RGSL13_02_153884					
RGSL13_02_170545	170544	T	C	YES	YES
RGSL13_02_192638	192637	G	A	YES	YES
RGSL13_02_196198	196197	T	A	YES	YES
RGSL13_02_201211	201210	A	G	YES	YES
RGSL13_02_201455	201454	T	C	YES	YES
RGSL13_02_201609	201608	T	A	YES	YES
RGSL13_02_211369	211365	C	T	YES	YES
	211368	G	C	YES	YES
RGSL13_02_214998	214998	NA	1 nucleotide insertion (G)	YES	YES
RGSL13_02_242570	242569	A	T	YES	YES
RGSL13_02_242749	242746 - 242748	TCA	3 nucleotide deletion (TCA)	YES	YES
RGSL13_02_252861	252860	G	A	YES	YES
RGSL13_02_286955	286954	C	T	YES	YES
RGSL13_02_294245	294238	C	A	YES	YES
	294244	G	A	YES	YES
RGSL13_02_354879	354878	G	A	YES	YES
*RGSL13_02_362016	362021-362054	LoxP site	Absence of loxPsym (34 nucleotide deletion) (ATAACTTCGTATAATGTACATTATACGAAGTTAT)	YES	YES
*RGSL13_02_362055					
RGSL13_02_391271	391270	C	A	YES	YES
RGSL13_02_408366	408365	C	T	YES	YES

RGSL13_02_434557	434547	C	T	YES	YES
	434550	T	C	YES	YES
	434553	G	C	YES	YES
	434556	C	A	YES	YES
RGSL13_02_449842	449847	A	G	YES	YES
RGSL13_02_457693	457692	G	A	YES	YES
*RGSL13_02_458553	458561-458594	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
*RGSL13_02_458595					
RGSL13_02_519808	519807	C	T	YES	YES
RGSL13_02_552000	551987	G	T	YES	YES
	551990	C	A	YES	YES
	551993	T	A	YES	YES
	551996	G	A	YES	YES
	551999	C	A	YES	YES
*RGSL13_02_552501	552505-552538	LoxP site	Absence of loxPsym (34 nucleotide deletion) (ATAACTTCGTATAATGTACATTATACGAAGTTAT)	YES	YES
*RGSL13_02_552539					
RGSL13_02_587180	587179	T	G	YES	YES
RGSL13_02_604062	604061	C	G	YES	YES
RGSL13_02_645857	645856	C	T	YES	YES
RGSL13_02_649536	649535	C	T	YES	YES
RGSL13_02_649656	649655	A	G	YES	YES
RGSL13_02_657735	657734	G	A	YES	YES
RGSL13_02_658248	658247	T	G	YES	YES
RGSL13_02_747961	747960	T	1 nucleotide deletion (T)	YES	YES

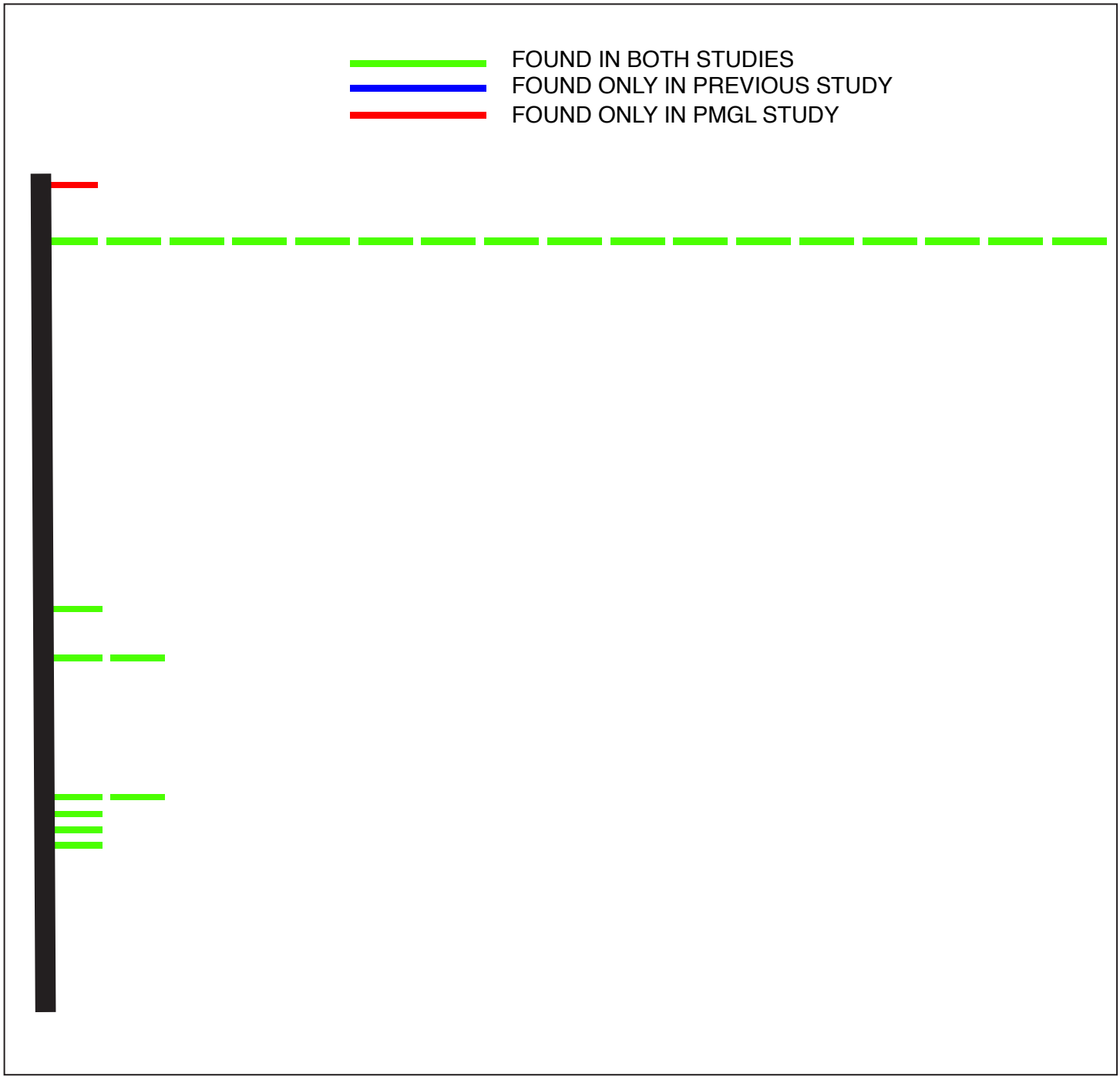
*The absence of a loxPsym site usually produces two signatures of variation when the PMGL pipeline is used



Unpublished Figure 3. Comparison of the physical syn V with the as-designed syn V. The data indicate the position of variants found in a previous study and in the present study (PMGL). Each line represents the position of one variant along the chromosome. The as-designed chromosome is represented as a black bar; nucleotide 1 (bottom); nucleotide 536,025 (top). Closely spaced variants are represented by rows of lines. The color code indicates the study in which the variant was detected. The nature of the variants is indicated in Unpublished Table 2.

Unpublished Table 2. Comparison of the as-designed synV sequence with the physical synV sequence from a previous analysis and in this study. The parameters used for the zero-trail scan were: min PMn at position n = 15; max PMn at position n-1 = 2

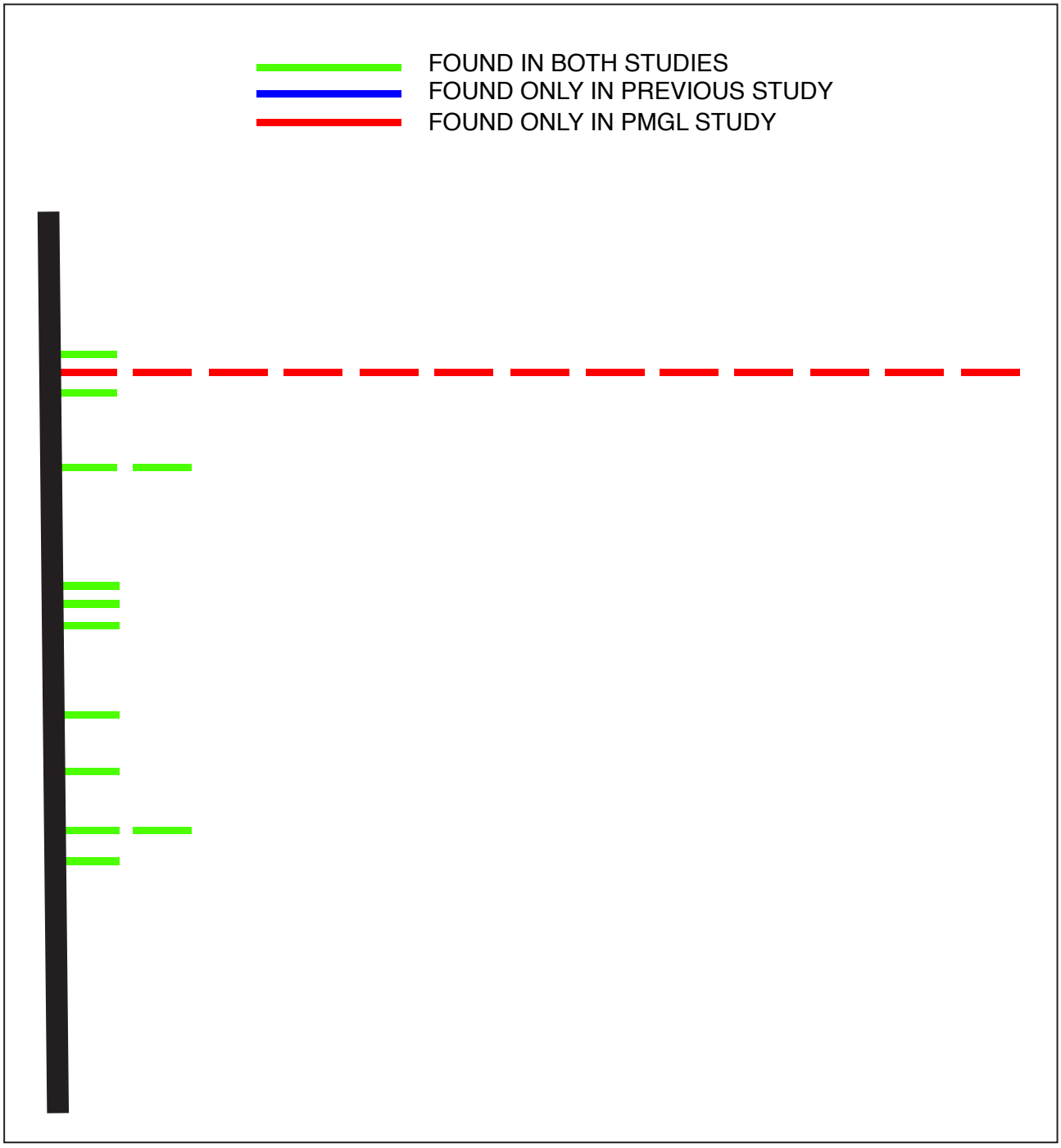
ID	POSITION OF VARIANT	As-designed synV	Physical synV	PREVIOUSLY DETECTED	DETECTED IN THIS STUDY
RGSL27_05_325793	325792	G	T	YES	YES
RGSL27_05_345194	345193	G	A	YES	YES



Unpublished Figure 4. Comparison of the physical syn VI with the as-designed syn VI. The data indicate the position of variants found in a previous study and in the present study (PMGL). Each line represents the position of one variant along the chromosome. The as-designed chromosome is represented as a black bar; nucleotide 1 (bottom); nucleotide 242,746 (top). Closely spaced variants are represented by rows of lines. The color code indicates the study in which the variant was detected. The nature of the variants is indicated in Unpublished Table 3.

Unpublished Table 3. Comparison of the as-designed synVI sequence with the physical synVI sequence from a previous analysis and in this study. The parameters used for the zero-trail scan were: min PMn at position n = 20; max PMn at position n-1 = 1

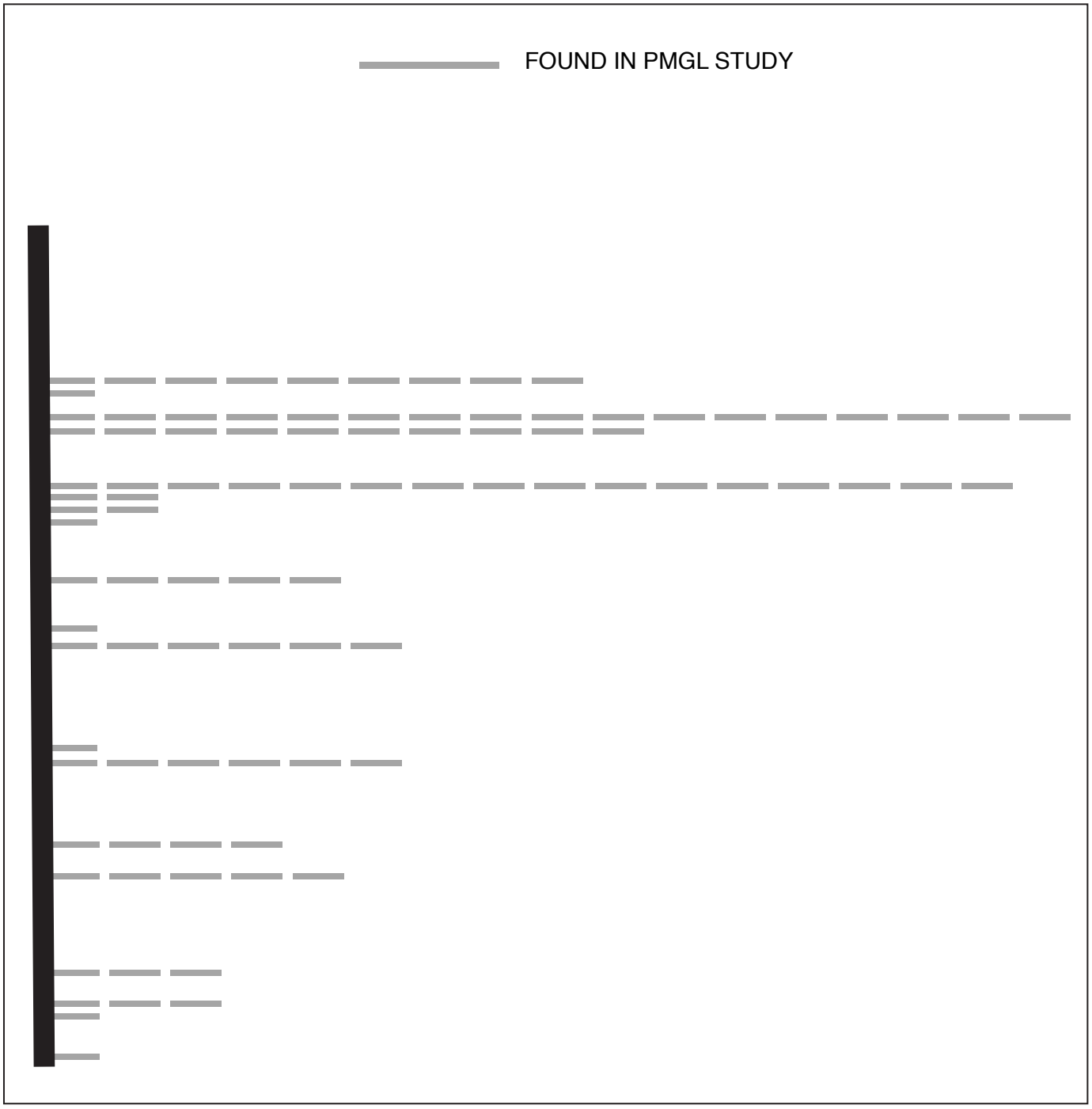
ID	POSITION OF VARIANT	As-designed synVI	Physical synVI	PREVIOUSLY DETECTED	DETECTED IN THIS STUDY
RGSL8_06_47428	47427	A	G	YES	YES
RGSL8_06_50461	50460	C	T	YES	YES
RGSL8_06_54045	54044	A	G	YES	YES
RGSL8_06_55666	55665	A	G	YES	YES
RGSL8_06_56219	56218	A	G	YES	YES
RGSL8_06_103869	103865	C	A	YES	YES
	103868	G	T	YES	YES
RGSL8_06_119663	119662	A	G	YES	YES
RGSL8_06_224819	224818	T	C	YES	YES
RGSL8_06_224954	224926	A	G	YES	YES
	224931	A	G	YES	YES
	224932	G	T	YES	YES
	224933	C	G	YES	YES
	224934	T	A	YES	YES
	224935	A	G	YES	YES
	224938	A	G	YES	YES
	224943	G	T	YES	YES
	224944	G	A	YES	YES
	224945	C	G	YES	YES
	224946	T	A	YES	YES
	224947	G	A	YES	YES
	224948	C	G	YES	YES
	224949	T	A	YES	YES
	224950	T	A	YES	YES
	224953	A	G	YES	YES
RGSL8_06_240435	240434	A	C	NO	YES



Unpublished Figure 5. Comparison of the physical syn X with the as-designed syn X. The data indicate the position of variants found in a previous study and in the present study (PMGL). Each line represents the position of one variant along the chromosome. The as-designed chromosome is represented as a black bar; nucleotide 1 (bottom); nucleotide 707,460 (top). Closely spaced variants are represented by rows of lines. The color code indicates the study in which the variant was detected. The nature of the variants is indicated in Unpublished Table 4.

Unpublished Table 4. Comparison of the as-designed synX sequence with the physical synX sequence from a previous analysis and in this study. The parameters used for the zero-trail scan were: min PMn at position n = 15; max PMn at position n-1 = 5

ID	POSITION OF VARIANT	As-designed synX	Physical synX	PREVIOUSLY DETECTED	DETECTED IN THIS STUDY
RGSL21_10_204035	204034	T	C	YES	YES
RGSL21_10_225234	225233	C	T	YES	YES
RGSL21_10_227312	227311	T	C	YES	YES
RGSL21_10_272049	272048	A	G	YES	YES
RGSL21_10_311659	311658	C	A	YES	YES
RGSL21_10_387192	387191	T	C	YES	YES
RGSL21_10_392862	392865-392898	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL21_10_392899					
RGSL21_10_408450	408449	T	C	YES	YES
RGSL21_10_507478	507487-507520	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL21_10_507521					
RGSL21_10_511672	511671	C	T	YES	YES
RGSL21_10_564353	564352	C	G	YES	YES
RGSL21_10_572039	572011	00	AG	NO	YES
	572011	G	A	NO	YES
	572012	C	A	NO	YES
	572013	T	C	NO	YES
	572016	A	C	NO	YES
	572021	A	G	NO	YES
	572022	A	G	NO	YES
	572024-572025	CT	1 nucleotide deletion (CT)	NO	YES
	572026	A	G	NO	YES
	572029	G	A	NO	YES
	572032	A	C	NO	YES
	572035	G	A	NO	YES
	572038	G	A	NO	YES
RGSL21_10_580755	580754	G	T	YES	YES

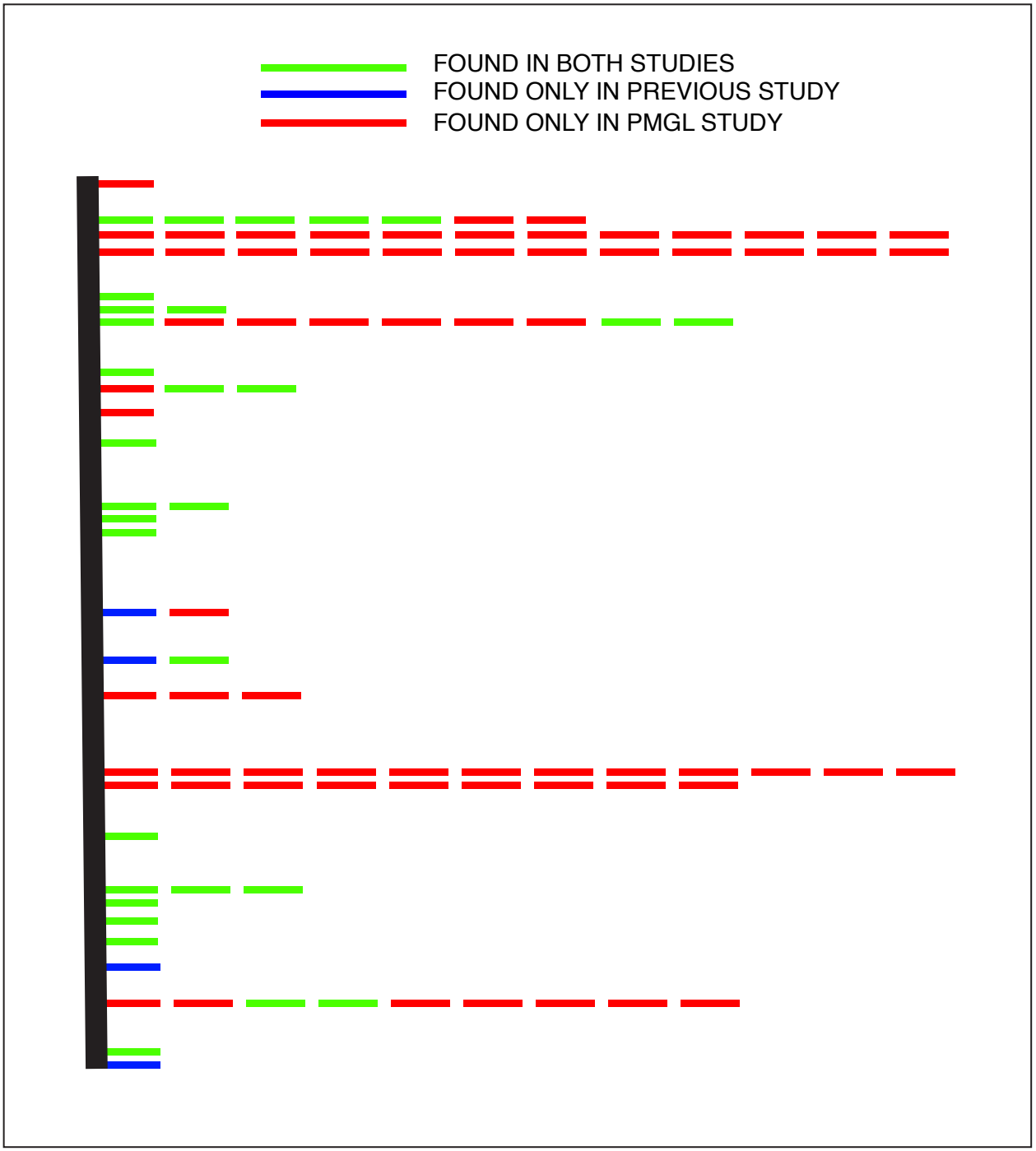


Unpublished Figure 6. Comparison of the physical syn XI with the as-designed syn XI. This analysis has only been performed in this study using the PMGL pipeline. The data indicate the position of variants found. Each gray line represents the position of one variant along the chromosome. The as-designed chromosome is represented as a black bar; nucleotide 1 (bottom); nucleotide 659,584 (top). Closely spaced variants are represented by rows of lines. The nature of the variants is indicated in Unpublished Table 5.

Unpublished Table 5. Comparison of the as-designed synXI sequence with the physical synXI sequence. The analysis has been performed only with the PMGL Pipeline (this study). The parameters used for the zero-trail scan were: min PMn at position n = 15; max PMn at position n-1 = 2

ID	POSITION OF VARIANT	As-designed synXI	Physical synXI
RGSL30_11_4790	4789	T	C
RGSL30_11_37066	37065	T	C
RGSL30_11_43203	43197	C	T
	43200	G	C
	43202	G	A
RGSL30_11_70968	70965	G	A
	70966	C	G
	70967	T	A
RGSL30_11_154695	154682	G	T
	154685	C	T
	154688	C	T
	154691	G	C
	154694	C	A
RGSL30_11_176630	176620	C	T
	176621	C	T
	176623	T	A
	176629	C	T
RGSL30_11_238870	238855	G	T
	238858	C	T
	238861	C	T
	238864	G	T
	238867	C	T
	238869	G	A
RGSL30_11_245336	245335	C	A
RGSL30_11_328408	328395	C	T
	328397	G	A
	328398	G	A
	328404	A	C
	328406	G	A
	328407	G	A
RGSL30_11_343171	343170	G	T
RGSL30_11_380034	380023	G	T
	380024	C	T
	380027	A	T
	380029	G	A
	380033	C	T
RGSL30_11_425992	425991	C	T
RGSL30_11_436063	436067 to 436100	ATAACTTCGTATAATGTACATTATACGAAGTTAT (loxPsym)	ACCTTATTTAGTATTGGACCATTGAGGTATTAGG
RGSL30_11_436101			
RGSL30_11_436359	between 436358 and 436359	NA	Presence of loxPsym (34 nucleotide insertion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)
RGSL30_11_442588	442587	C	T
RGSL30_11_444206	444205	C	T
RGSL30_11_448095	448094	G	C

RGSL30_11_466269	446241	C	A
	446244	T	C
	446247	A	T
	446250	T	A
	446251	A	T
	446252	G	C
	446253	C	G
	446254	T	C
	446256	G	A
	446259	G	A
	446260	T	A
	446261	C	G
	446262	A	C
	446265	T	A
	446268	C	T
RGSL30_11_493864	493836	T	G
	493839	T	C
	493842	A	G
	493845	G	T
	493848	A	T
	493851	G	A
	493854	G	A
	493857	A	T
	493860	A	G
	493863	C	T
RGSL30_11_493939	493911	G	A
	493914	A	G
	493915	C	G
	493916	T	A
	493917	G	A
	493918	C	G
	493919	T	A
	493920	G	A
	493923	A	G
	493926	G	A
	493927	C	G
	493928	T	A
	493929	G	A
	493932	T	G
	493933	G	C
	493934	A	T
	493938	A	G
RGSL30_11_526663 RGSL30_11_526702	526668 to 526701	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)
RGSL30_11_536987	536959	C	T
	536962	A	C
	536968	C	T
	536970	A	G
	536971	C	T
	536973	A	G
	536976	T	G
	536977	G	A
	536986	T	C



Unpublished Figure 7. Comparison of the physical syn XII with the as-designed syn XII. The data indicate the position of variants found in a previous study and in the present study (PMGL). Each line represents the position of one variant along the chromosome. The as-designed chromosome is represented as a black bar; nucleotide 1 (bottom); nucleotide 999,407 (top). Closely spaced variants are represented by rows of lines. The color code indicates the study in which the variant was detected. The nature of the variants is indicated in Unpublished Table 6. See explanation for blue bars in the footnote of Unpublished Table 6.

Unpublished Table 6. Comparison of the as-designed synXII sequence with the physical synXII sequence from a previous analysis and in this study. The parameters used for the zero-trail scan were: min PMn at position n = 12; max PMn at position n-1 = 2

ID	Variation Site	As-designed synXII	Physical synXII	PREVIOUSLY DETECTED	DETECTED IN THIS STUDY
RGSL17_12_27814	27813	A	T	YES	YES
RGSL17_12_82405	82404	C	T	NO	YES
RGSL17_12_82691	82695	A	G	NO	YES
	82699 - 82732	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL17_12_83203	83170 - 83203	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL17_12_83260	83247	G	A	NO	YES
	83250	C	T	NO	YES
	83253	G	A	NO	YES
	83256	G	A	NO	YES
	83259	C	A	NO	YES
RGSL17_12_142536	142535	C	T	YES	YES
RGSL17_12_159309	159308	C	T	YES	YES
RGSL17_12_186198	186197	T	C	YES	YES
RGSL17_12_200689	200692 - 200725	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL17_12_200726					YES
RGSL17_12_204940	204939	T	C	YES	YES
RGSL17_12_256543	256542	NA	1 nucleotide insertion (A)	YES	YES
RGSL17_12_327746	327721	A	G	NO	YES
	327724	C	T	NO	YES
	327727	G	T	NO	YES
	327730	A	G	NO	YES
	327733	G	A	NO	YES
	327736	A	C	NO	YES
	327739	G	C	NO	YES
	327742	G	A	NO	YES
	327745	A	G	NO	YES
RGSL17_12_327941	327913	G	A	NO	YES
	327916	A	G	NO	YES
	327919	A	T	NO	YES
	327922	G	A	NO	YES
	327925	A	C	NO	YES
	327928	A	T	NO	YES
	327931	C	T	NO	YES
	327934	C	T	NO	YES
	327936	A	G	NO	YES
	327937	T	C	NO	YES
	327939	A	G	NO	YES
	327940	G	A	NO	YES
RGSL17_12_425591	425586	C	G	NO	YES
	425587	T	G	NO	YES
	425590	NA	3 nucleotide insertion (CGT)	NO	YES
RGSL17_12_463605	463604	T	A	YES	YES
RGSL17_12_519403	519402	NA	Presence of loxPsym (34 nucleotide insertion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	NO	YES
RGSL17_12_601172	601138 - 601171	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL17_12_624342	624337	GGACC	AGATT	YES	YES
RGSL17_12_634261	634260	A	C	YES	YES
RGSL17_12_634675	634674	T	C	YES	YES
RGSL17_12_700345	700344	C	A	YES	YES
RGSL17_12_737011	737010	A	1 nucleotide deletion (A)	NO	YES
RGSL17_12_769259	769256 - 769258	TCA	3 nucleotide deletion (TCA)	NO	YES
RGSL17_12_770903	770911 - 770944	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL17_12_770944					YES

RGSL17_12_780626	780625	T	C	YES	YES
RGSL17_12_840365	840332 - 840365	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL17_12_844322	844312	G	T	NO	YES
	844315	C	A	NO	YES
	844318	G	A	NO	YES
	844319	C	G	NO	YES
	844320	T	A	NO	YES
	844321	G	C	NO	YES
RGSL17_12_844501	844507 - 844540	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL17_12_844541					YES
RGSL17_12_850440	850446 - 850479	LoxP site	Absence of loxPsym (34 nucleotide deletion) (TATAACTTCGTATAATGTACATTATACGAAGTTA)	YES	YES
RGSL17_12_850480					YES
RGSL17_12_861020	861019	A	1 nucleotide deletion (A)	YES	YES
RGSL17_12_869950	869922	G	T	NO	YES
	869925	A	G	NO	YES
	869928	A	G	NO	YES
	869931	A	G	NO	YES
	869934	G	A	NO	YES
	869937	G	A	NO	YES
	869940	A	T	NO	YES
	869943	A	G	NO	YES
	869944	C	G	NO	YES
	869945	T	A	NO	YES
	869946	G	A	NO	YES
	869949	G	A	NO	YES
RGSL17_12_920944	920916	T	C	NO	YES
	920919	C	T	NO	YES
	920923	A	T	NO	YES
	920924	G	C	NO	YES
	920925	C	T	NO	YES
	920928	C	T	NO	YES
	920931	A	G	NO	YES
	920934	T	C	NO	YES
	920937	T	C	NO	YES
	920938	T	C	NO	YES
	920940	G	T	NO	YES
	920943	C	T	NO	YES
RGSL17_12_921031	921022	T	A	YES	YES
	921023	C	G		
	921024	A	T		
	921027	G	A		
	921030	T	C		
RGSL17_12_921280	921250	GACATCTTGTGGGTTGAGGACAGTTCAGGTAC	GATATTCTTTGGGTGGAGGATTCCAGTGGGAC	NO	YES
RGSL17_12_921319	921318	A	G	NO	YES
RGSL17_12_986569	986568	A	1 nucleotide deletion (A)	NO	YES

Important note: variants reported by Boeke at positions 2180, 113931, 462443, and 517167 were excluded from our analysis because they have very low counts of supporting sequence reads (7,4,5, and 6, respectively). Normally, we use a minimum of 15 supporting sequence reads containing the perfect match zone adjacent to the variant site. For this analysis, we lowered the threshold to a minimum of 12 supporting sequence reads. If we lower the threshold further down to 4 (to include all variants reported by Boeke), we increase the number of potential variant sites from 44 to 82. In our opinion, this extended list would include false positives.

CONCLUSIONES

- En esta tesis se desarrolló una estrategia original para el análisis de la variación genómica: "Panorama Genómico de Identidad Absoluta" o PMGL.

- Las propiedades principales de la estrategia de análisis de variación genómica por PMGL son las siguientes:

    - Se revela directamente la presencia de variación a través de la construcción de un panorama de identidad absoluta entre el genoma de referencia y el genoma de interés.

    - Se detecta la posición precisa de las variantes, independientemente del tipo de variación.

    - Se revela la naturaleza de las variantes a través de la generación de alineamientos altamente dirigidos entre el genoma de referencia y el genoma de interés.

    - Se establece un marco teórico para el desarrollo de una firma general de variación.

- Se generó una ruta computacional totalmente automatizada para el análisis de variación genómica por PMGL. Esta ruta computacional permite identificar las variantes contenidas en un proyecto de secuenciación en relación a un genoma previamente ensamblado, o genoma de referencia. Además de caracterizar con precisión dichas variantes, la misma ruta computacional permite valorarlas internamente.

- La ruta computacional automatizada está constituida por los siguientes procesos:

    - Generación de un RGSL

    - Generación de un PMGL

    - Escaneo del PMGL

    - Generación de un alineamiento inicial para cada firma de variación

    - Interpretación y extensión de alineamientos

    - Generación de un genoma de referencia *ad hoc*

    - Validación de las variantes obtenidas (reutilizando los tres primeros procesos)

- Con base en la estrategia de análisis de variación genómica por PMGL, en esta tesis se han logrado los siguientes objetivos específicos:

  -Refinamiento de los genomas de referencia de las siguientes cepas de la levadura *S. cerevisiae*:

  -S288C
  -SK1
  -Y12

  -Generación de perfiles de variación de los genomas de las siguientes cepas de *S. cerevisiae* por comparación con el genoma de referencia de la cepa S288C:

  -BY4742
  -SK1

  - Identificación de los sitios de variación del genoma de las siguientes cepas de *S. cerevisiae* por comparación con el genoma de referencia de la cepa S288C:

  -Y12
  -DBVPG6765

  - Caracterización de la variación de los siguientes cromosomas sintéticos vivientes en relación a sus respectivas secuencias diseñadas:

  -*S.cerevisiae* syn II
  -*S.cerevisiae* syn III
  -*S.cerevisiae* syn V
  -*S.cerevisiae* syn VI
  -*S.cerevisiae* syn X
  -*S.cerevisiae* syn XI
  -*S.cerevisiae* syn XII

- La presente tesis ha generado los siguientes productos de investigación:

  - Artículo publicado en la revista de prestigio internacional GENETICS con Factor de Impacto 5.963:

  A Perfect Match Genomic Landscape Provides a Unified Framework for the Precise Detection of Variation in Natural and Synthetic Haploid Genomes
  Kim Palacios-Flores, Jair García-Sotelo, Alejandra Castillo, Carina Uribe, Luis Aguilar, Lucía Morales, Laura Gómez-Romero, José Reyes, Alejandro Garciarubio, Margareta Boege, and Guillermo Dávila
  GENETICS 2018 (en prensa)

- Software público:

Depositado en el repositorio público: GitHub
https://github.com/LIIGH-UNAM/PerfectMatchGenomicLandscapePipeline.git
Autores: Kim Palacios-Flores y Jair García-Sotelo
Descripción: Precise detection of genome variation using the Perfect Match Genomic Landscape Pipeline

- Presentaciones orales por invitación en congresos internacionales:

> - A Perfect Match Genomic Landscape (PMGL) provides a unified framework for the detection of variation in synthetic genomes
> Kim Palacios-Flores
> LIIGH-UNAM
> Seventh International Meeting on Synthetic Biology
> National University of Singapore
> June 13-16, 2017
>
> - A new algorithm for verifying synthetic chromosome sequences and detecting genome alterations
> Kim Palacios-Flores
> LIIGH-UNAM
> Sixth Annual Sc2.0 Meeting
> Singapore
> June 17, 2017

- Además de los datos publicados, se cuenta con datos relevantes que se espera incorporar en publicaciones futuras. Éstos comprenden el análisis de los siguientes cromosomas sintéticos vivientes en relación a sus respectivas secuencias diseñadas: *S. cerevisiae* syn II, syn V, syn VI, syn X, syn XI y syn XII.

- Invitación a formar parte del Consorcio Internacional Sc2.0. Por su participación en el análisis de cromosomas sintéticos, la autora de la tesis, Kim Palacios-Flores y la Institución en la que se realizaron dichos análisis, el Laboratorio Internacional de Investigación sobre el Genoma Humano de la UNAM (LIIGH), han sido invitados a formar parte del Consorcio Internacional Sc2.0. Éste último tiene como objetivo el diseño y la síntesis completa del primer genoma eucarionte, el de la levadura *S. cerevisiae*.

PERSPECTIVAS

Aplicaciones futuras de la estrategia PMGL

En su forma actual, la ruta computacional desarrollada en esta tesis puede ser extrapolada a una variedad de aplicaciones, entre las cuales figuran las siguientes:

- Refinamiento de genomas de células individuales. En la actualidad, una serie de estudios genómicos utilizan la secuenciación de células individuales. La metodología para secuenciar y analizar estos genomas aún no se encuentra totalmente optimizada y, por lo general, genera un número elevado de errores. El genoma individual resultante puede ser considerado como el genoma de referencia, y las lecturas de secuenciación como el genoma de interés. En este contexto, es posible utilizar la estrategia PMGL para refinar el genoma y limitar, pensamos que sustancialmente, el nivel de errores del ensamble original.

- Análisis en tiempo real de la variación acumulada en experimentos de evolución a largo plazo. El paradigma de este tipo de experimentos ha sido establecido por el Dr. Richard Lenski, quien ha determinado la variación genética de poblaciones de *E. coli* crecidas con limitación de glucosa durante decenas de miles de generaciones. Es importante mencionar que hemos iniciado pláticas con el Dr. Lenski para analizar algunas de sus cepas con la ruta computacional PMGL.

- Análisis de la variación en secuencias de ADN específicas de distintos genomas presentes en un experimento de metagenómica. Las propiedades de la estrategia de análisis de variación genómica por PMGL permitirían realizar el siguiente tipo de experimento: 1) generar una metaestructura RGSL (metaRGSL) utilizando los genomas concatenados de todos los organismos por ser analizados, 2) generar un metaRGSL dirigido al extraer todos aquellos renglones asociados a una secuencia única, 3) utilizar el metaRGSL dirigido resultante y las lecturas de secuenciación totales para generar un metaPMGL dirigido, 4) escanear el metaPMGL dirigido en búsqueda de firmas de variación y 5) caracterizar las variantes subyacentes. Cada una de las variantes resueltas podría ser rastreada a un genoma específico dentro de la población estudiada. Un análisis evolutivo en tiempo real podría ser particularmente interesante, especialmente tratándose de comunidades que coexisten en distintos nichos del organismo humano.

Hacia una firma universal de variación

El marco teórico desarrollado en el artículo "**A Perfect Match Genomic Landscape Provides a Unified Framework for the PreciseDetection of Variation in Natural and Synthetic Haploid Genomes"** propone una firma universal de variación para el análisis de la variación genómica. Esta firma de variación se manifiesta como cambios repentinos, entre dos nucleótidos consecutivos, en la cobertura de correspondencias perfectas reportada por la estructura PMGL. La generalidad de esta firma recae sobre su potencialidad para revelar la presencia de cambios de un nucleótido, deleciones, inserciones y amplificaciones, en una

diversidad de contextos, incluyendo zonas multicopia del genoma de interés y genomas diploides. El objetivo futuro más importante para la línea de investigación presentada en esta tesis es consolidar el concepto de la firma universal de variación a través de su implementación en el análisis de genomas más complejos. Paralelamente, la ruta computacional desarrollada en este trabajo tendrá que evolucionar. En particular, será necesario incluir módulos computacionales que permitan la determinación, de manera automatizada, de los puntos de inflexión de variaciones estructurales, como lo son las deleciones e inserciones grandes. También, será necesario proveer una infraestructura que permita paralelizar los procesos de caracterización y validación de las variantes una vez encontradas sus posiciones precisas. En su forma actual, la ruta computacional desarrollada permite paralelizar el análisis al nivel de cromosomas completos. Sin embargo, una partición más ambiciosa del problema permitirá analizar genomas más grandes y con un porcentaje de secuencias repetidas más elevado en un tiempo óptimo. Una vez alcanzadas estas metas será posible incorporarse a una serie de proyectos que requieren del conocimiento de la secuencia de ADN de diferentes genomas con un nivel de precisión y extensión sin precedentes. La relevancia de la estrategia PMGL ha sido demostrada a través de nuestra participación en el análisis de cromosomas sintéticos del proyecto internacional Sc2.0, que constituye un proyecto piloto para el HGP-write. A su vez, esperamos, en un futuro relativamente cercano, colaborar con el análisis de las fases intermedias y finales de la síntesis del genoma humano.

BIBLIOGRAFÍA

Abrahams, E., and S. L. Eck, 2016 Molecular medicine: Precision oncology is not an illusion. *Nature* 539: 357.

Abremski K, and R. Hoess, 1985 Phage P1 Cre-*lox*P Site-specific Recombination. Effects of DNA Supercoiling on Catenation and Knotting of Recombinant Products. J. Mol.Biol. 184: 211-220

Annaluru, N., H. Muller, L. A. Mitchell, S. Ramalingam, G. Stracquadanio *et al.,* 2014 Total Synthesis of a Functional Designer Eukaryotic Chromosome. Science 344: 55–58.

Audano, P., S. Ravishankar, and F. Vannberg, 2017 Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. Bioinformatics, advance access publication.

Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland *et al.*, 1997 The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1462.

Blount, Z., Barrick, J., Davidson, C. and Lenski, R., 2012 Genomic analysis of a key innovation in an experimental Escherichia coli population. Nature 489: 513–518.

Boeke, J. D., G. Church, A. Hessel, N. J. Kelley, A. Arkin *et al.*, 2016 The Genome Project-Write. Science 353: 126-127.

Brachmann, C. B., A. Davies, G. J. Cost, E. Caputo, J. Li *et al.,* 1998 Designer Deletion Strains derived from *Saccharomyces cerevisiae* S288C: a Useful set of Strains and Plasmids for PCR-mediated Gene Disruption and Other Applications. Yeast 14: 115–132.

Brenner, S., 2002 Nature's gift to science. Nobel Lecture

Chaisson, M. J., and G. Tesler, 2012 Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMCBioinformatics 13: 238.

Dymond, J. S., S. M. Richardson, C. E. Coombes, T. Babatz, H. Muller *et al.,* 2011 Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. Nature 477: 471–476.

Edgar, R., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Engel, S.R., F. S. Dietrich, D. G. Fisk, G. Binkley, R. Balakrishnan *et al.,* 2014 The Reference Genome Sequence of *Saccharomyces cerevisiae*: Then and Now. G3: Genes, Genomes, Genetics 4: 389–398.

1000 Genomes Project Consortium, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.*, 1996 Life with 6000 Genes. Science 274: 546–567.

Goodwin, S., J. D. McPherson, and W. R. McCombie, 2016 Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17: 333-351.

GP-write Leadership Group and The GP-Write Consortium, 2016 Genome Project-write: A Grand Challenge Using Synthesis, Gene Editing and Other Technologies to Understand, Engineer and Test Living Systems.
(White Paper available at http://engineeringbiologycenter.org/).

Holt, J., and L. McMillan, 2014 Merging of multi-string BWTs with applications. Bioinformatics 30: 3524-3531.

International Human Genome Sequencing Consortium., 2001 Initial sequencing and analysis of the human genome. Nature 409, 860-921
International Human Genome Sequencing Consortium., 2004 Finishing the euchromatic sequence of the human genome. Nature 431, 931-945
Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan et al., 2012 VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22: 568-576.
Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway et al., 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12.
Kurtz, S., 2003 The Vmatch large scale sequence analysis software. Ref Type: Computer Program 412.
Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357-359.
Langmead, B., C. Trapneli, M. Pop, and S.L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.
Lenski, R., Wiser, M., Ribeck, N., Blount, Z., Nahum, J. et al., 2015 Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with Escherichia coli. Proc. R. Soc. B **282**: 20152292.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987-2993.
Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.
Li, R. Q., Y. R. Li, K. Kristiansen, and J. Wang, 2008 a SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713-714.
Li ; H., J. Ruan, and R. Durbin, 2008 b Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18: 1851-1858.
Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27: 764–770.
Mardis, E. R., 2013 Next-Generation Sequencing Platforms. Annu. Rev. Anal. Chem. 6: 287-303.
McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis et al., 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297-1303.
Metzker, M. L., 2010 Sequencing technologies---the next generation. Nat. Rev. Genet. 11: 31-46.
Mitchell, L. A., A. Wang, G. Stracquadanio, Z. Kuang, X. Y. Wang et al., 2017 Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond. Science 355: eaaf4831
Ostrov, N., Landon, M., Guell, M., Kuznetsov, G., Teramoto, J., et al., 2016 Design, synthesis, and testing toward a 57-codon genome. Science **353**: 819–822.
Park, P. J., 2009 ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. 10: 669-680.
Pennisi, E., 2013. The Man Who Bottled Evolution. Science. 342, 790–793

Pfeifer, S. P., 2017 From next-generation resequencing reads to a high-quality variant data set. Heredity 118: 111-124.
Reinert, K., B. Langmead, D. Weese, and D. J. Evers, 2015 Alignment of Next-Generation Sequencing Reads. Annu. Rev. Genomics Hum. Genet. 16: 133-151.
Reyes, J., L. Gómez-Romero, X. Ibarra-Soria, K. Palacios-Flores, L. R. Arriola et al., 2011 Context-dependent individualization of nucleotides and virtual genomic hybridization allow the precise location of human SNPs. Proc. Natl. Acad. Sci. USA 108: 15294-15299.
Richardson, S. M., L. A. Mitchell, G. Stracquadanio, K. Yang, J. S. Dymond et al., 2017 Design of a synthetic yeast genome. Science 355: 1040-1044.
Rimmer, A. H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg et al., 2014 Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 46: 912-918.
Schatz, M. C., A. M. Phillippy, D. D. Sommer, A. L. Delcher, D. Puiu et al., 2013 Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. Brief. Bioinform. 14: 213–224
Schbath, S., V. Martin, M. Zytnicki, J. Fayole, V. Loux et al., 2012 Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. J. Comput. Biol. 19: 796-813.
Shen, Y., Y. Wang, T. Chen, F. Gao, J. H. Gong et al., 2017 Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. Science 355: eaaf4791.
Teer, J. K., and J. C. Mullikin, 2010 Exome sequencing: the sweet spot before whole genomes. Hum. Mol. Genet. 19: R145-R151.
Tenaillon, O., J. E. Barrick, N. Ribeck, D. E. Deatherage, J. L. Blanchard et al., 2016 Tempo and mode of genome evolution in a 50,000-generation experiment. Nature 536: 165-170.
Wang, Z., M. Gerstein, and M. Snyder, 2009 RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10: 57-63.
Wu, Y., B. Z. Li, M. Zhao, L. A. Mitchell, Z. X. Xie et al., 2017 Bug mapping and fitness testing of chemically synthesized chromosome X. Science 355: eaaf4706.
Yang, X., S. P. Chockalingam, and S. Aluru, 2013 A survey of error-correction methods for next-generation sequencing. Brief. Bioinform. 14: 56-66.
Yue, J.X., J. Li, L. Algrain, J. Hallin, K. Persson, et al., 2017 Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat. Genet. 49: 913-924.
Xie, Z. X., B. Z. Li, L. A. Mitchell, Y. Wu, X. Qi et al., 2017 “Perfect” designer chromosome V and behavior of a ring derivative. Science 355: eaaf4704.
Zerbino, D. R. and E. Birney, 2008 Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821-829.
Zhang, W., Zhao, G., Luo, Z., Lin, Y., Wang, L., et al., 2017 Engineering the ribosomal DNA in a megabase synthetic chromosome. Science 355: eaaf3981.