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

Kim Palacios-Flores,*[,1] Jair García-Sotelo,* Alejandra Castillo,* Carina Uribe,* Luis Aguilar,* Lucía Morales,* Laura Gómez-Romero,* José Reyes,* Alejandro Garciarubio,† Margareta Boege,* and Guillermo Dávila*

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

ORCID IDs: 0000-0002-3308-2714 (K.P.-F.); 0000-0002-9462-2737 (J.G.-S.)

ABSTRACT We present a conceptually simple, sensitive, precise, and essentially nonstatistical solution for the analysis of genome variation in haploid organisms. The generation of a Perfect Match Genomic Landscape (PMGL), which computes intergenome 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 PMGL 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.

KEYWORDS genome variation; genome sequencing; reference genomes; yeast genomics; synthetic genomics

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* (Zerbino 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

doi: https://doi.org/10.1534/genetics.117.300589
Manuscript received December 5, 2017; accepted for publication January 19, 2018; published Early Online January 24, 2018.
Available freely online through the author-supported open access option.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300589/-/DC1.
[1] Corresponding author: Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Blvd. Juriquilla 3001, Querétaro, Querétaro 76230, México. E-mail: kimpalaciosflores@gmail.com

procedures as whole genome sequencing (1000 Genomes Project Consortium 2012), exome (Teer and Mullikin 2010) and transcriptome sequencing (Wang *et al.* 2009), and chromatin immunoprecipitation sequencing (Park 2009), among 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 (Kurtz *et al.* 2004; R. Q. Li *et al.* 2008; H. Li *et al.* 2008; Li and Durbin 2009; Chaisson and Tesler 2012; 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; Koboldt *et al.* 2012; Rimmer *et al.* 2014).

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 (see Supplemental Material, File S1, 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 single nucleotide variants, microindels, 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 reference genome. All modules are described in detail in File S1.

Detailed materials and methods

File S1 presents the following methodology in detail: *Saccharomyces 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, and 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 contains a detailed description of *Materials and Methods*. File S2 contains examples of variant location and characterization, including relevant subsets of the corresponding PMGL, local alignments, and Sanger sequence graphs. File S3 shows a random simulation experiment in the query genome. File S4 shows a directed simulation experiment in multiple copy regions of the reference genome. File S5 contains analyses of the S288C and BY4742 *S. cerevisiae* strains. File S6 contains analysis of the *S. cerevisiae* strain SK1. File S7 contains analysis of the *S. cerevisiae* strain Y12. File S8 shows a comparison of the as-designed synthetic chromosome III (synIII) sequence with the query genome sequence reads of *S. cerevisiae* strain HMSY011 from a previous analysis and from this study. File S9 contains alignments showing the variants uncovered for synIII. File S10 shows a 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 one nucleotide sliding window along the reference genome. The RGSL reports each reference string's starting position within a specific chromosome, its nucleotide sequence, the number of occurrences of its sequence in the entire reference genome (count reference), and the unique identifier of all reference strings sharing the same sequence. 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 one 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. The number of perfect matches associated with each reference string is normalized by its count reference.

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 1B). 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).

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

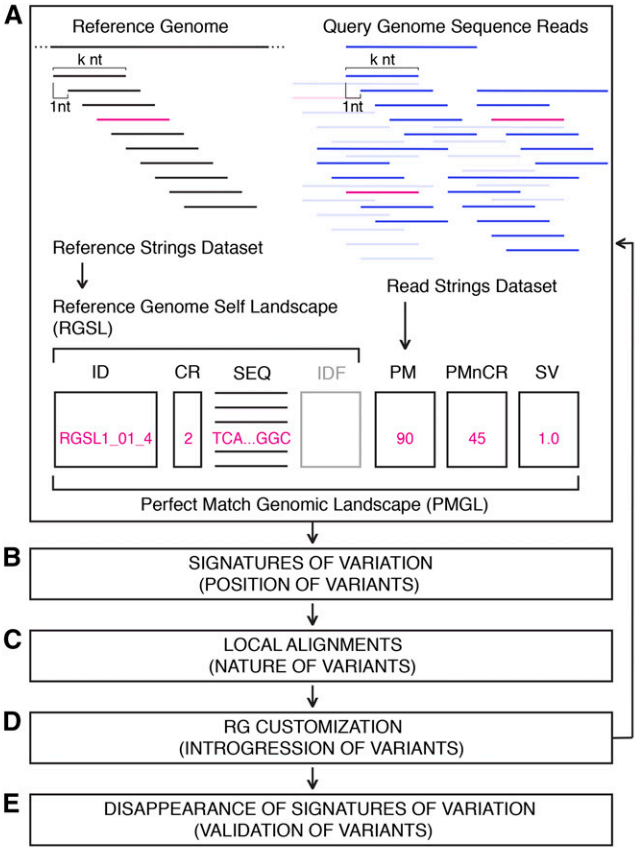

Figure 1 Pipeline of the PMGL strategy. (A) Generation of the PMGL. Both the reference genome and the query genome sequence reads are decomposed into comprehensive arrays of k nucleotide-long strings, generating the reference strings dataset and the read strings dataset, respectively. The PMGL reports the number of perfect matches between the reference genome and the query genome along the RGSL structure. 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)–(E) are explained in the text and in *Materials and Methods*, and detailed in File S1. CR, count reference; ID, reference string's unique identifier; IDF, repeat family unique identifiers; PM, perfect match coverage; PMnCR, perfect match coverage normalized to count reference; SEQ, sequence; RG, reference genome.

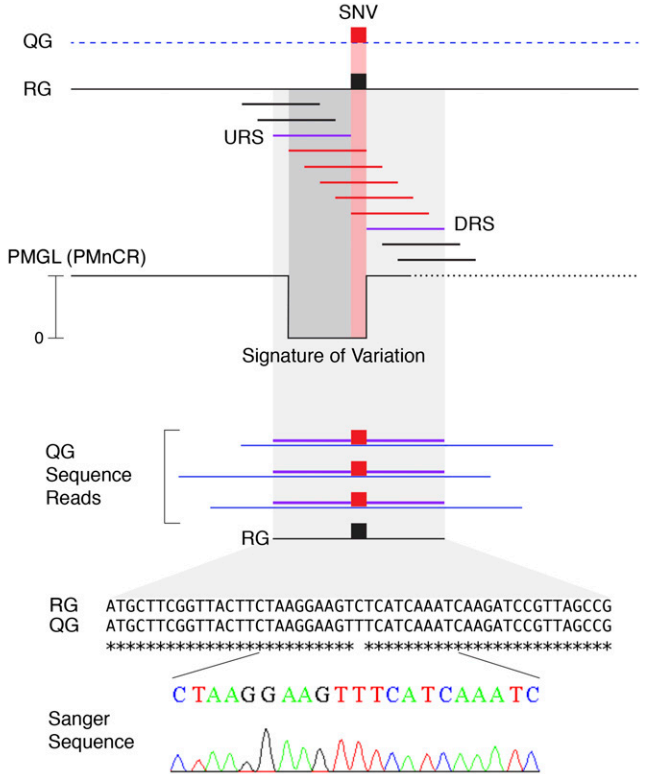
SNV
QG
RG
URS
DRS
PMGL (PMnCR)
0
Signature of Variation
QG
Sequence
Reads
RG
RG ATGCTTCGGTTACTTCTAAGGAAGTCTCATCAAATCAAGATCCGTTAGCCG
QG ATGCTTCGGTTACTTCTAAGGAAGTTTCATCAAATCAAGATCCGTTAGCCG
************************ *************************
C T A A G G A A G T T T C A T C A A A T C
Sanger
Sequence

Figure 2 Detection and characterization of a single nucleotide variant (SNV). Corresponding regions of the query genome (QG) and the reference genome (RG) 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 gray and red shading indicate the zero-trail zone. Red shading represents the projection of the variant into the reference genome. Light gray 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.

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 (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*).

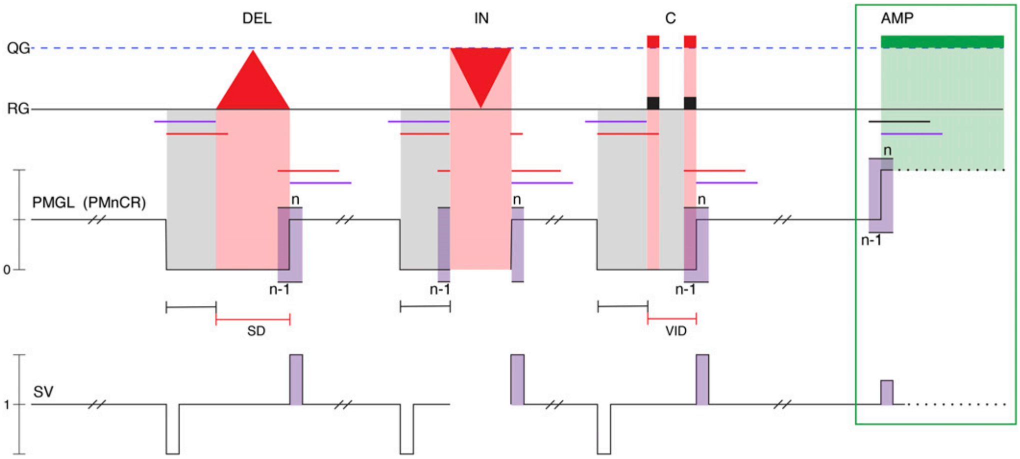

Figure 3 Detection of different types of variation. Corresponding features are color-coded as in Figure 2. The AMP case is boxed because amplifications have not been experimentally addressed in this study. Reference strings contributing to the sudden drop and/or rise in the normalized perfect match coverage 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. The purple shading indicates the sharp rise in normalized perfect match coverage shared by all types of variants, and its relative magnitude (SV) is schematized at the bottom. AMP, amplification (shown in green); C, concatenated variants; DEL, deletion; IN, insertion; SD, size of deletion; SV, signature value; VID, variation inclusive distance between concatenated variants; QG, query genome; RG, reference genome.

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 1E).

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 three 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 twofold 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.

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, microdeletions, and microinsertions; in some cases, these variants were concatenated. Using the zero-trail scan, the presence of 109 out of 112 variants was detected exclusively at the exact copy of

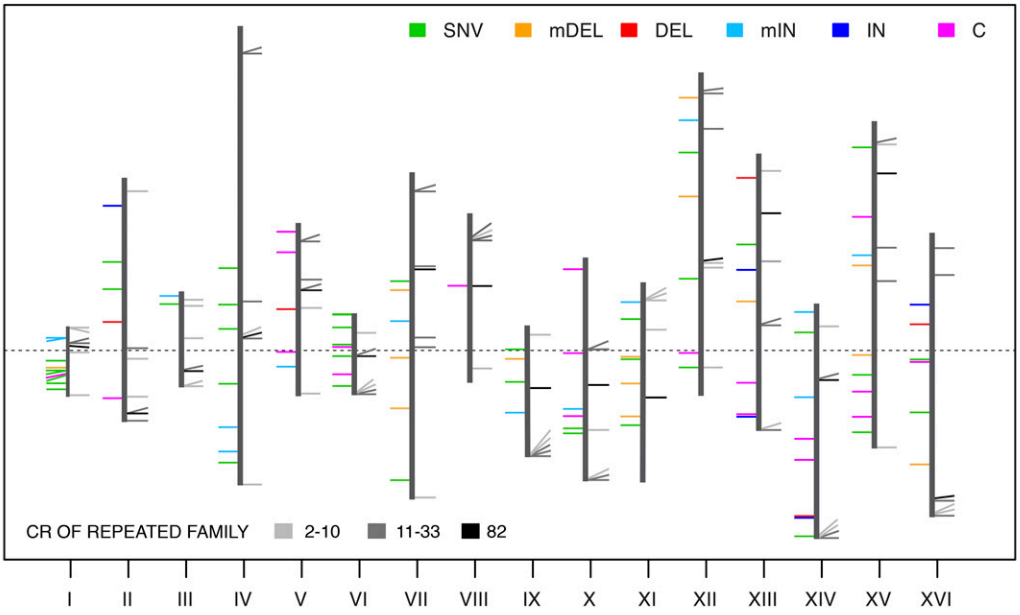

Figure 4 Assessment of the performance of the PMGL strategy via simulation experiments. The 16 nuclear chromosomes of the S288C reference genome are shown as vertical bars, with position 1 of each chromosome located at the bottom. The dashed black line crosses each chromosome's centromere. The left section 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. 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. C, concatenated variants; DEL, deletion; IN, insertion; mDEL, microdeletion; mIN, microinsertion; SNV, single nucleotide variant.

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.

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 among 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 five 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 the BY4742 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, 95 regions 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).

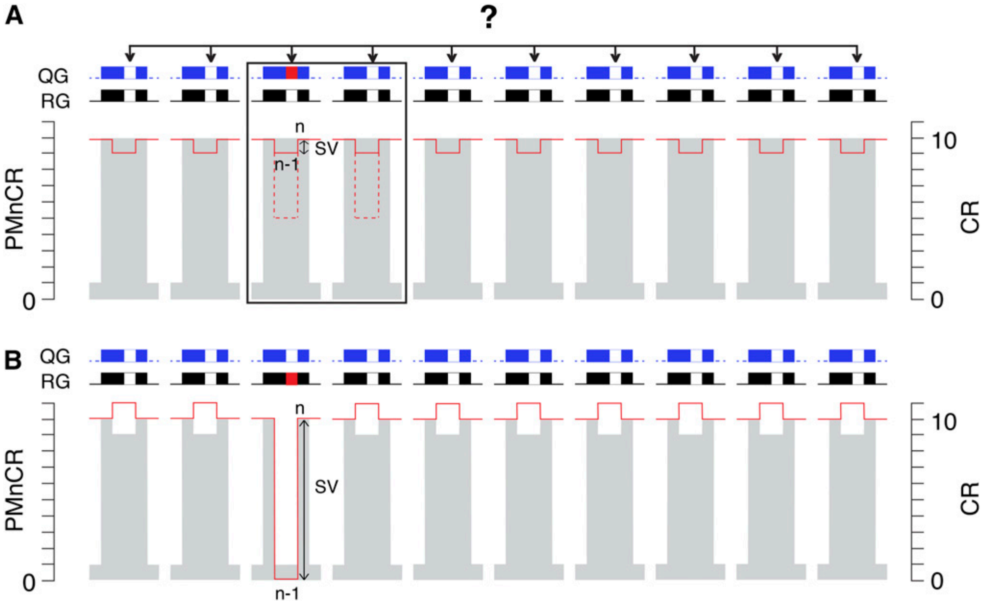

Figure 5 Impact on the PMGL for variants present in repeated regions of the query genome or the reference genome. A single nucleotide variant is harbored in a 10-copy region of either (A) the query genome (QG) or (B) the reference genome (RG). Ten copies of the repeated region, located at different positions 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 ~50% at each copy (broken red line) and the signature value increases to ~2 (see *Discussion*). The actual position of the variant remains ambiguous but is now restricted to either of the two copies.

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* section, the complete datasets for these analyses are available at GitHub.

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 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 five single nucleotide variants concatenated with a two-nucleotide microinsertion, one single nucleotide variant found in close proximity to a previously reported missing loxPsym site, a one-nucleotide microdeletion,

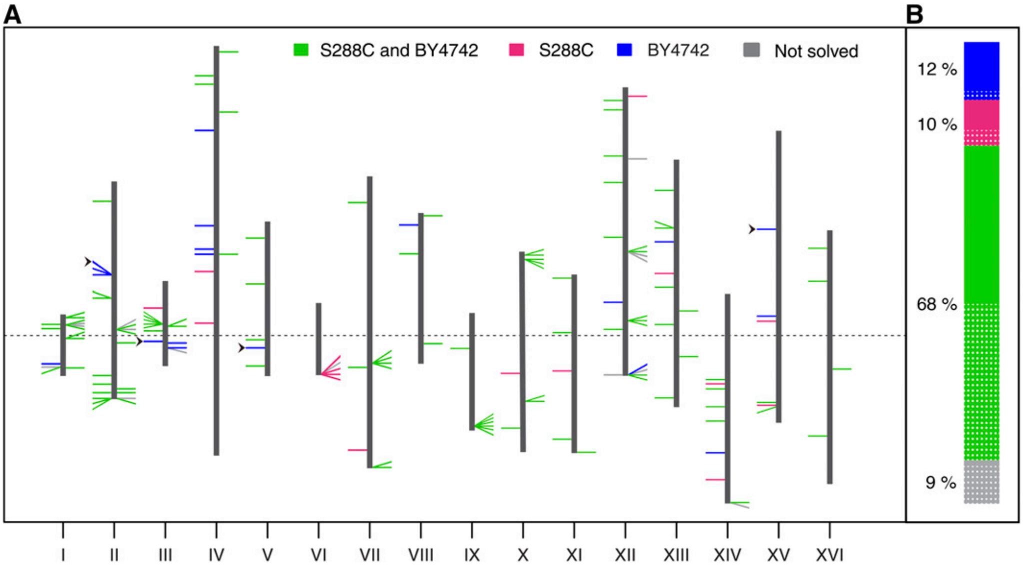

Figure 6 Analyses and comparison of genome variation profiles from yeast strains S288C and BY4742. (A) The 16 nuclear chromosomes of the S288C reference genome are shown as vertical bars, with position 1 of each chromosome located at the bottom. The dashed black line crosses each chromosome's centromere. 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.

a 15-nucleotide deletion, and an 11-nucleotide insertion (File S8 and File S9).

We subsequently built an 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 10 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 and 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 fifth 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.

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

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
Chromosome	nSV found	Solved (%)	nSV found	nSV found
I	948	90	894	992
II	3,539	97	2,482	2,995
III	1,142	95	1,444	522
IV	7,109	96	6,733	3,293
V	2,840	96	2,520	1,554
VI	1,506	95	900	1,103
VII	5,153	96	4,702	3,097
VIII	2,619	95	2,368	1,319
IX	1,979	95	1,811	1,622
X	3,084	97	2,465	2,375
XI	2,822	96	2,452	2,577
XII	4,296	97	3,591	3,353
XIII	4,152	97	3,441	2,399
XIV	3,315	97	3,475	1,724
XV	5,009	97	4,076	3,475
XVI	4,662	96	4,260	1,628
Total	54,175	96	47,614	34,028

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

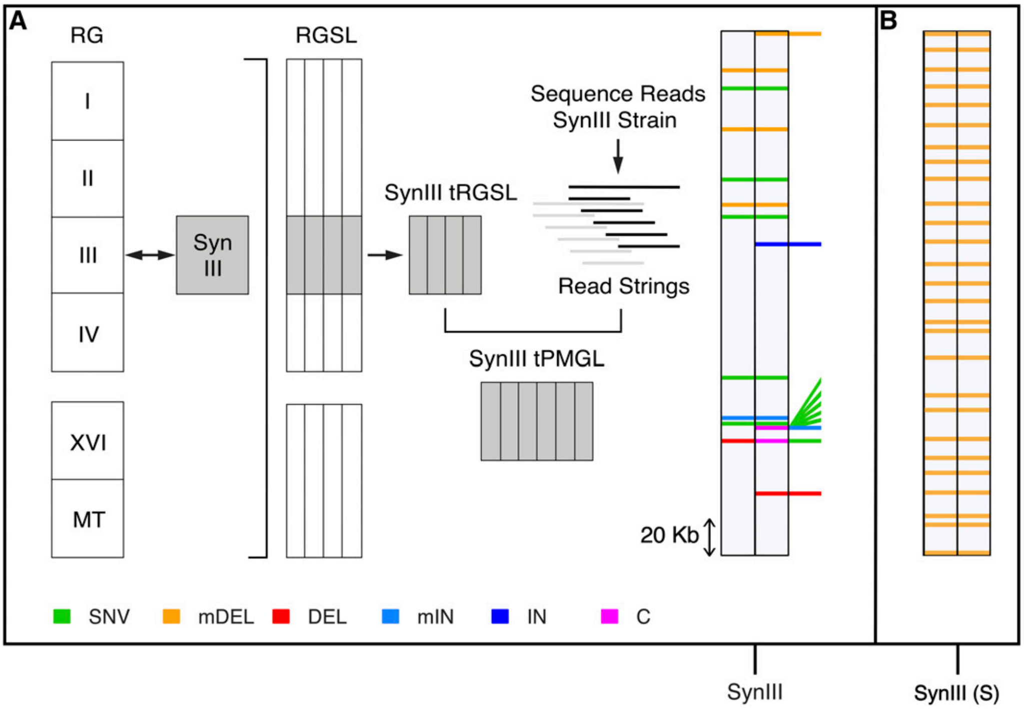

Figure 7 Analysis of a synthetic yeast chromosome. (A) Targeted analysis of variation in synIII. The as-designed synIII sequence (gray shadow) is introduced into the S288C reference genome (RG) background, replacing native chromosome III. A comprehensive RGSL is generated. A targeted RGSL (tRGSL) is 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. (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. C, concatenated variants; DEL, deletion; IN, insertion; mDEL, microdeletion; mIN, microinsertion; SNV, single nucleotide variant.

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 two and one, where one 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 two and one 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 datasets generated with state of the art experimental and bioinformatics methodologies. Most importantly, the PMGL strategy has contributed to 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 resequencing, 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 synIII.

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.

Acknowledgments

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. K.P.-F. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship grant 392315 from Consejo Nacional de Ciencia y Tecnología (CONACYT). This work was supported in part by grant IA207817 from the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) from the Universidad Nacional Autónoma de México (UNAM).

Literature Cited

1000 Genomes Project Consortium Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

Abrahams, E., and S. L. Eck, 2016 Molecular medicine: precision oncology is not an illusion. Nature 539: 357.10.1038/539357e

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. doi: 10.1093/bioinformatics/btx753

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. BMC Bioinformatics 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 (Bethesda) 4: 389–398.

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.

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 SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713–714.

Li, H., J. Ruan, and R. Durbin, 2008 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. (Palo Alto, Calif.) 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., 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.

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.

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.

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.

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., G. Zhao, Z. Luo, Y. Lin, L., Wang *et al.*, 2017 Engineering the ribosomal DNA in a megabase synthetic chromosome. Science 355: eaaf3981.

Communicating editor: M. Johnston