Perfect Match Genomic Landscape strategy: Refinement and customization of reference genomes

Kim Palacios-Flores[a,1], Jair García-Sotelo[a], Alejandra Castillo[a], Carina Uribe[a], Lucía Morales[a], Margareta Boege[b], Guillermo Dávila[a], Margarita Flores[a], and Rafael Palacios[a,2]

[a]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, 76230 Querétaro, México; and [b]Escuela Nacional de Estudios Superiores Juriquilla, Universidad Nacional Autónoma de México, 76230 Querétaro, México

Contributed by Rafael Palacios, February 17, 2021 (sent for review December 7, 2020; reviewed by Jef D. Boeke, Patricia L. Foster, and Joseph Schacherer)

When addressing a genomic question, having a reliable and adequate reference genome is of utmost importance. This drives the necessity to refine and customize reference genomes (RGs). Our laboratory has recently developed a strategy, the Perfect Match Genomic Landscape (PMGL), to detect variation between genomes [K. Palacios-Flores *et al.*. *Genetics* 208, 1631–1641 (2018)]. The PMGL is precise and sensitive and, in contrast to most currently used algorithms, is nonstatistical in nature. Here we demonstrate the power of PMGL to refine and customize RGs. As a proof-of-concept, we refined different versions of the *Saccharomyces cerevisiae* RG. We applied the automatic PMGL pipeline to refine the genomes of microorganisms belonging to the three domains of life: the archaea *Methanococcus maripaludis* and *Pyrococcus furiosus*; the bacteria *Escherichia coli*, *Staphylococcus aureus*, and *Bacillus subtilis*; and the eukarya *Schizosaccharomyces pombe*, *Aspergillus oryzae*, and several strains of *Saccharomyces paradoxus*. We analyzed the reference genome of the virus SARS-CoV-2 and previously published viral genomes from patients' samples with COVID-19. We performed a mutation-accumulation experiment in *E. coli* and show that the PMGL strategy can detect specific mutations generated at any desired step of the whole procedure. We propose that PMGL can be used as a final step for the refinement and customization of any haploid genome, independently of the strategies and algorithms used in its assembly.

microbial genomes | genome variation | mutation-accumulation experiments | experimental evolution | SARS-CoV-2

The accuracy of the nucleotide sequences of genomes is of utmost importance. In some cases, a single base pair can be relevant in regard to the function of a protein, an RNA, or a regulatory site in the DNA. Initially the nucleotide sequence of genomes was established for model organisms by directly assembling sequence reads into continuous structures, thus generating reference genomes (RGs) for different organisms. These include the bacterium *Escherichia coli* (1), the yeast *Saccharomyces cerevisiae* (2), the worm *Caenorhabditis elegans* (3), the fruit fly *Drosophila melanogaster* (4), the plant *Arabidopsis thaliana* (5), the mouse *Mus musculus* (6), and the human *Homo sapiens* (7). The RGs of different species were then used to guide the assembly of other genomes by determining the differences between the RG and the genome of interest (query genome, QG). At present, both strategies are being utilized to assembly genomes: the de novo assembly of the set of sequence reads (8) and the comparison of sequence reads with the ordered structure of an RG (9).

Most of the currently used algorithms to determine variation between genomes are based on the alignment of sequence reads to a genome used as reference. Such alignment allows a certain degree of mismatch between the sequence reads of the QG and the RG. In turn, the allowed mismatches could cause ambiguity in regard to the position of some reads. To avoid such potential problems, more precise algorithms or combinations of algorithms are continuously being developed (10). As a consequence of the mismatch problem, the alignment of sequence reads to an RG, and thus the calling of variants between genomes, is statistical in nature.

The Perfect Match Genomic Landscape (PMGL) strategy has been developed in our laboratory to determine variation between genomes (11). This strategy uses only perfect matches between strings of the RG and strings from the sequence reads of the QG. In essence, the PMGL is precise, sensitive, and nonstatistical in nature. This strategy has been applied to haploid genomes. It has been evaluated by different types of simulation experiments and has been used to determine variation in cultivated and natural yeast strains (11). Furthermore, yeast synthetic chromosomes have been generated (12, 13) and the PMGL has been used to determine the accuracy of the nucleotide sequence of yeast synthetic chromosome Syn III (11). Audano et al. (14) published an algorithm that is based on a similar principle to that of the PMGL strategy (see also ref. 11).

In this study we report the power of the PMGL to refine RGs, and to customize genomes for specific purposes. We propose that the PMGL could be used as a final step to increase the accuracy of the nucleotide sequence of any haploid genome.

Results

PMGL Strategy to Refine and Customize RGs. The PMGL strategy has been amply described and discussed by Palacios-Flores et al. (11). Fig. 1 summarizes the main steps of this strategy as applied

Significance

The accuracy of the nucleotide sequence of genomes is of utmost importance. The Perfect Match Genomic Landscape (PMGL) is a precise, sensitive, and nonstatistical strategy to detect genome variation. We used this strategy to refine reference genomes from microorganisms belonging to the three domains of life. Our studies show as well that the PMGL can be useful to detect variants in pathogen agents during a pandemic, and to isolate mutations generated during any desired stage of experimental evolution studies. We propose that the PMGL strategy could be the final step in the refinement of any haploid genome, independently of the methodology and algorithms used for its assembly.

Author contributions: K.P.-F., G.D., and R.P. designed research; J.G.S., A.C., C.U., and M.F. performed research; K.P.-F. and J.G.S. contributed new reagents/analytic tools; K.P.-F., J.G.S., A.C., C.U., L.M., M.F., and R.P. analyzed data; and L.M., M.B., and R.P. wrote the paper.

Reviewers: J.D.B., New York University School of Medicine; P.L.F., Indiana University Bloomington; and J.S., University of Strasbourg.

The authors declare no competing interest.

[1]Present address: Department of Neurobiology, Friedrich Miescher Institute for Biomedical Research, 4058 Basel, Switzerland.

[2]To whom correspondence may be addressed. Email: palacios@liigh.unam.mx.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2025192118/-/DCSupplemental.

Published March 18, 2021.

to the refinement and customization of RGs. The continuous structure (or structures in the case of several chromosomes) of the RG and the fragmented structure of the QG (sequence reads) are both divided into overlapping nucleotide strings (in this study we used strings of 25 nucleotides) (*Materials and Methods*) following a sliding window of one nucleotide. The strings of the RG are listed according to their positions in the chromosome. The strings of the QG are counted according to their sequence and the set of strings are merged with the corresponding identical strings of the RG list. The PMGL thus generated consists of four columns of data: the RG string identifier, the number of identical copies of the string in the RG, the nucleotide sequence of the string, and the normalized number of matches found in the whole set of sequence reads of the QG (Fig. 1).

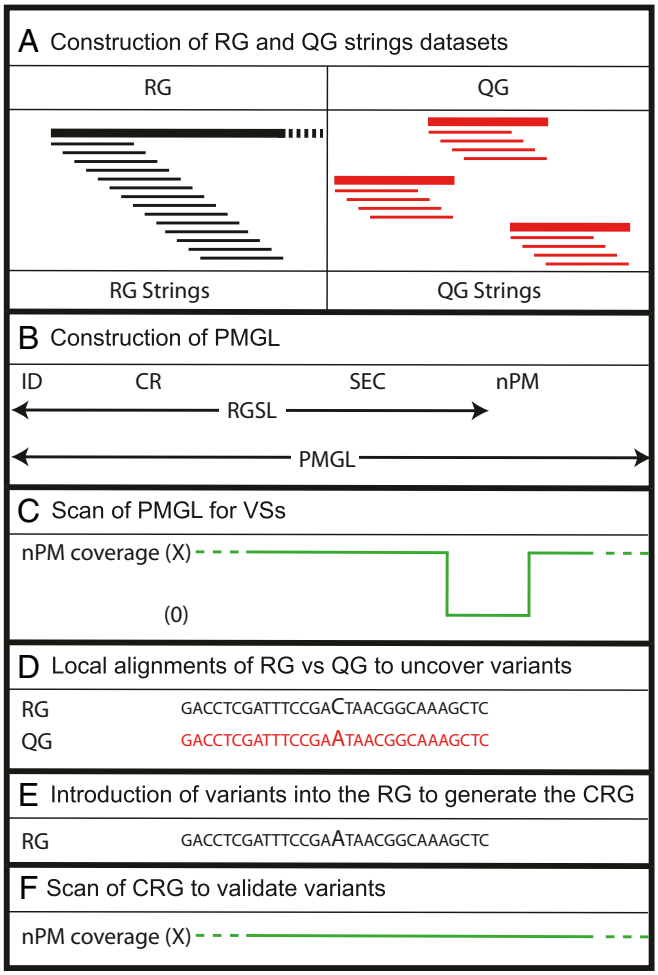

Fig. 1. PMGL pipeline to refine and customize RGs. (*A*) The continuous structure of the RG is divided in overlapping strings of 25 nucleotides (black) using a 1-nucleotide sliding window. In RGs with several chromosomes, each one is treated as a separated structure. Each sequence read of the QG is similarly divided in overlapping strings (red). (*B*) The strings of the RG are ordered according to their position in the continuous structures. For each string an identifier (ID) is assigned (first column). In the second column the number of strings presenting the exact sequence in the whole RG is indicated (count reference, CR). In the third column the nucleotide sequence of the string is indicated (SEQ). These three columns constitute the RG self-landscape (RGSL). The strings in the QG dataset are counted according to their sequence. The ordered RGSL is merged with the counts of each string in the QG; such number is normalized to the CR and indicated in the fourth column (normalized perfect matches, nPM) generating the PMGL. (*C*) The PMGL reports the coverage (green) at the start position of each string. When the RG and the QG differ, the coverage decreases to zero or near zero (in haploid genomes) for all the RG strings that are involved in the variation. Such decrease in coverage is referred to as an SV. (*D*) For each SV the RG and the QG are locally aligned to uncover the actual nature of the variation, here a single nucleotide variant. (*E*) The variants detected in the QG are introduced into RG, generating the CRG. (*F*) The CRG is scanned for SVs. The absence of the SV validates the nature of the corresponding variants.

The perfect matches of strings of the RG and the QG indicate the coverage at each nucleotide along the RG. When the RG and the QG differ, the coverage drops to zero (or very near zero) for all the strings that are involved in the variation. This drop in coverage is referred to as a variation signature (VS). Thus, the VS corresponds to a track of zero coverage. In the case of single-nucleotide variants the zero-coverage track involves 25 nucleotides. Microindels produce zero-coverage tracks of about the same size as single nucleotide variants. If another variant is present in the zero-coverage track (concatenated variants), the track increases for about 25 nucleotides upstream of such other variants (see examples of fragments of PMGL datasets harboring VSs in *SI Appendix*, Datasets S1–S4).

Large insertions of foreign DNA in the QG generate hybrid strings in the sequence reads, containing part of the original sequence present in the RG and part of the novel sequence. The hybrid strings derived from the sequence reads will not produce perfect matches with the corresponding original strings (without the insertion) present in the RG. This in turn generates a zero-coverage track that in fact constitutes a VS. The length of the VS is about the size of the strings used, and does not depend on the size of the insertion. Insertions from endogenous material in the QG, such as insertion sequences or partial or complete gene copies, will produce a similar effect, since hybrid strings will be formed in the sequence reads. In addition, the count reference of the strings involved (Fig. 1) (11) will increase according to the number of copies added. In contrast, in the case of deletions in the QG, the size of the VS increases in a number of nucleotides equal to the size of the deletion (11).

The PMGL is then scanned for VSs. Each VS indicates a site of discrepancy between the RG and the QG. The actual nature of the variants is revealed by local alignments between the region of the RG and the corresponding sequence reads of the QG. Each variant of the QG is then introduced into the original RG, generating the corresponding customized RG (CRG). The CRG is used to generate a new PMGL using the same QG. This PMGL is scanned for VSs. The absence of a VS validates the nature of the corresponding variation. In fact, if the variants introduced into the CRG corrects the local discrepancy between the RG and the QG, the corresponding VS disappears. In contrast, if a "mistaken" variant is introduced, it will not solve the conflict between the RG and the QG and the corresponding VS will remain after the scan of the CRG. Furthermore, in the original paper describing the PMGL strategy (11), the validation by disappearance of VSs in the customized genome was in turn challenged by another procedure. Regions of the RG showing VSs were analyzed by PCR and Sanger sequencing. In all cases tested (about 90) the Sanger sequence revealed the same variations reported by the PMGL pipeline.

The different steps of the PMGL algorithm have been automatized and made public. The automatic pipeline consists of six modules that must be sequentially applied (11) (*Materials and Methods*).

Historical Refinement of the *S. cerevisiae* Genome. The budding yeast *S. cerevisiae* has been used as a model for simple eukaryotic organisms in both scientific and biotechnological projects. Its genome was first sequenced in 1996 (2). In the original assembly, version 1 (V1), different *S. cerevisiae* strains were used. Since then, it has been continuously refined and it is considered a highly accurate genome (15). The most recent refinement corresponds to version 64 (V64). We selected 10 versions of the *S. cerevisiae* genome as RGs to further refine using the PMGL automatic

pipeline. As QG for all the PMGLs, we sequenced a recent stock of strain S288c obtained from the American Type Culture Collection (ATCC) (*Materials and Methods*). Each PMGL was scanned for VSs and the nature of the variants was determined by targeted alignments. All the variants found were introduced into the corresponding RG generating a CRG. The CRG was used to generate a new PMGL using the same QG. The new PMGL was scanned for VSs. The absence of VSs validated the corresponding variants.

The number of VSs found decreased with each version of the genome. V1 presented 1,208 VSs while V64 presented 119 VSs. Most of the variants responsible for the corresponding VS could be uncovered by the automatic PMGL pipeline (solved signatures of variation, SVs). In fact, after customization V1 presented only 34 VSs while V64 presented only 14 VSs (not solved SVs) (Fig. 2*A*). Fig. 2*C* presents a circos graph showing the position of the solved VSs in the XVI chromosomes of *S. cerevisiae*. The exact position of the VSs and the position and nature of the corresponding variants are presented in *SI Appendix*, Dataset S5. With the exception of three, all the variants found in V64 correspond to variants found as well in V1.

Another experiment was performed generating PMGLs using different versions of the *S. cerevisiae* RG and an artificial QG derived from V64. To construct such artificial QG, a total of 12 million random cuts of 100 nucleotides each, derived from the RG of V64, were used as sequence reads. This set of reads actually represents an artificial QG for V64 with a 100× coverage. About 98% of the VSs found and the variants uncovered were exactly the same as those obtained when the QG used was that from a stock of strain S288c recently obtained from ATCC (Fig. 2 *B* and *C* and *SI Appendix*, Dataset S6). As expected, no SVs were found in V64 when the QG was the artificial QG of V64. In addition, VSs found in V64 when the QG was that of S288c were not present in V1 when the QG was the artificial QG of V64.

Refinement of RGs of Organisms from Different Life Domains. We selected RGs deposited in the National Center for Biotechnology Information (NCBI) (*Materials and Methods*) from different microorganisms corresponding to the three life domains. These included the archeae *Methanococcus maripaludis* and *Pyrococcus furiosus*; the bacteria *E. coli*, *Staphylococcus aureus*, and *Bacillus subtilis*; and the eukarya *Schizosaccaromyces pombe* and *Aspergillus oryzae*. PMGLs were generated using the corresponding RGs. To generate the QGs of the different microorganisms, DNA was sequenced using the Illumina platform (*Materials and Methods*).

Fig. 3 shows the relative positions of VSs and the nature of the uncovered variants found in the original RGs, as well as the VSs remaining after customization of the corresponding genomes. The precise position of VSs and the precise positions and nature of the variants are indicated in *SI Appendix*, Datasets S7–S13. *M. maripaludis* showed one VS corresponding to one variant. *P. furiosus* showed 134 VSs, of which 130 corresponded to 190 variants; 4 VSs remained after customization. *E. coli* presented five VSs corresponding to eight variants. *S. aureus* showed one VS corresponding to one variant. *B. subtilis* showed 58 VSs, of which 55 corresponded to 60 variants; 3 VSs remained after customization. The 3 chromosomes of *S. pombe* presented 152 VSs, of which 138 corresponded to 169 variants; 14 VSs remained after customization. The 8 chromosomes of *A. oryzae* presented 49 VSs, of which 37 corresponded to 40 variants; 12 VSs remained after customization. The apparent discrepancy between the number of VSs and the number of variants is due to the fact that some VSs are generated by more than one variant (concatenated variants, see above).

Recently (16), five genomes of *Saccharomyces paradoxus* strains have been assembled using state-of-the-art methodologies, including

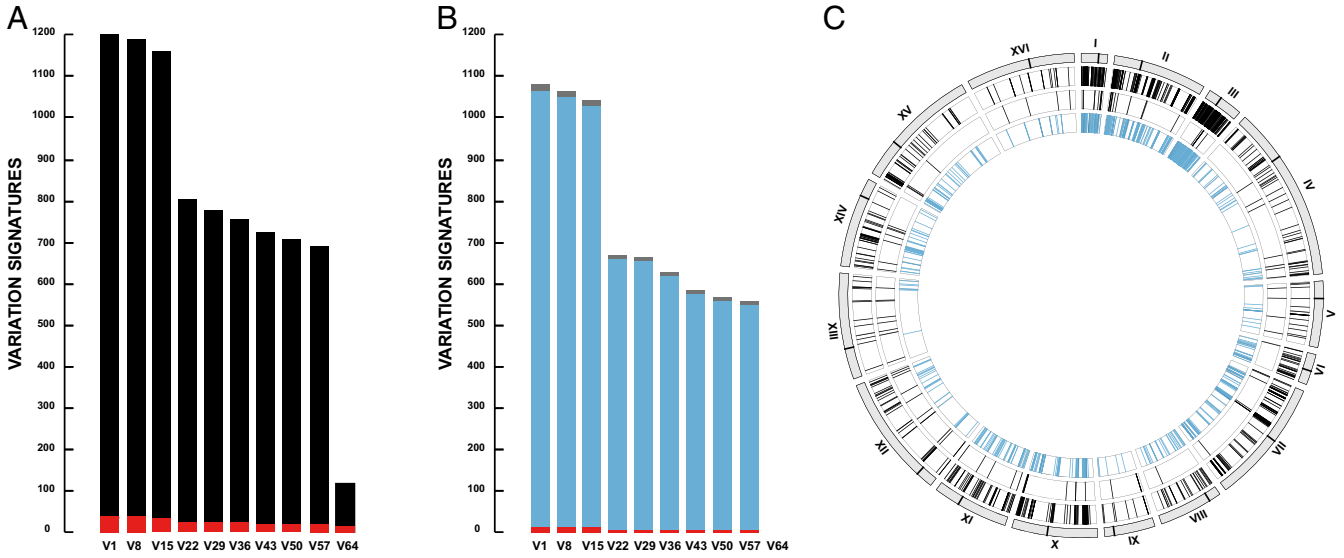

Fig. 2. Historical refinement of the *S. cerevisiae* RG. Different versions of the RG were analyzed with the PMGL pipeline. (*A*) The PMGLs were generated using the RGs of the different versions and the QG of a recent stock of strain S288c. For each version of the RG, the number of VSs found is indicated. In each bar the number of VSs that could be solved by the automatic pipeline is indicated in black and the number that were not solved is indicated in red. (*B*) The PMGLs were generated using the RGs of the different versions and an artificial QG derived from the RG of V64 (*Results*). In each bar the number of VSs that could be solved by the automatic pipeline and that are identical to those found when the QG corresponded to that of strain S288c is indicated in blue; the number of VSs that could be solved by the automatic pipeline and that are not identical to those found when the QG corresponded to that of strain S288c is indicated in gray; the number of VSs that were not solved is indicated in red. (*C*) Circos graph indicating the position of VSs in the chromosomes of *S. cerevisiae*. From outer to inner circle: chromosomes indicating the position (black line) of the centromere; position of VSs in V1 that were solved when the QG used to generate the PMGL corresponded to that of strain S288c (black); position of VSs in V64 that were solved when the QG used to generate the PMGL corresponded to that of strain S288c (black); position of VSs in V1 that were identical when the QG corresponded to either that of strain S288c or to the artificial QG generated from V64 (blue).

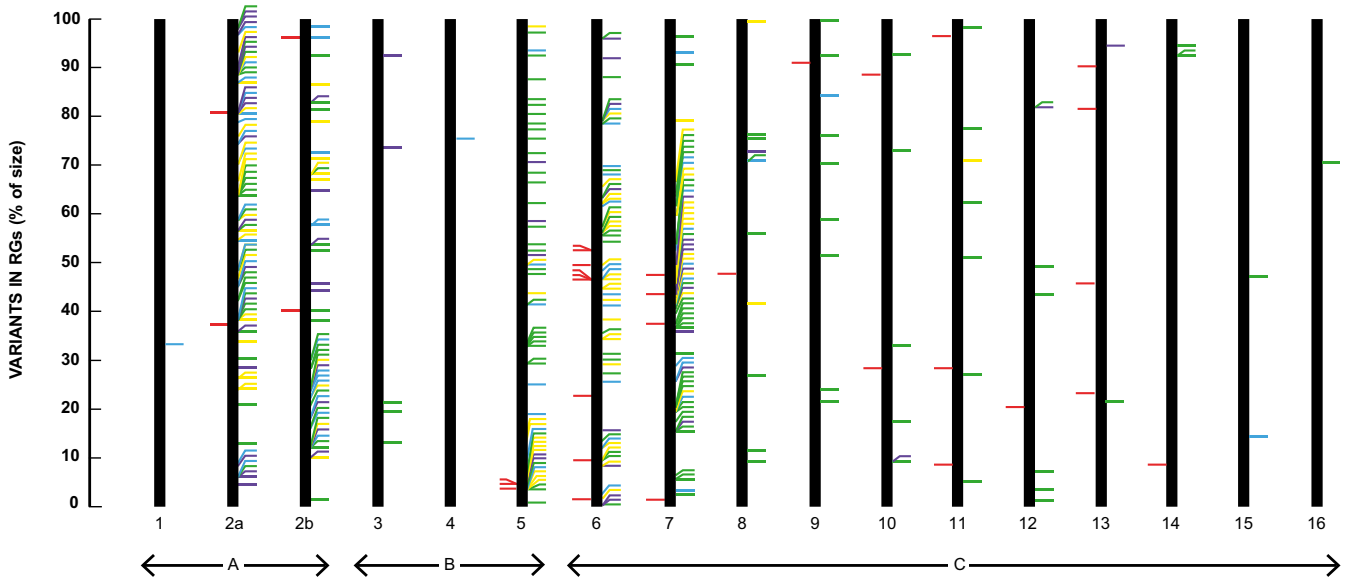

Fig. 3. Refinement of RGs of microorganisms from different life domains using the PMGL pipeline. (*A*) Archeae: 1, *M. maripaludis,* 2a and 2b; *P. furiosus* (nucleotides 1 to 954128 and nucleotides 954129 to 1908956, respectively). (*B*) Bacteria: 3, *E. coli*; 4, *S. aureus*; 5, *B. subtilis*. (*C*) Eukarya: 6 to 8 *S. pombe* (chromosomes I to III, respectively); 9 to 16, *A. oryzae* (chromosomes I to VIII, respectively). VSs that remained after customization of the respective genomes are indicated in red at the left of each bar. Variants that were uncovered are indicated at the right of each bar; color code: green, single nucleotide variants; blue, microinsertions according to the RG; yellow, microdeletions according to the RG; purple, concatenated variants. The position of the variants is indicated as percent of the size of the corresponding chromosome. In the case of *P. furiosus* the chromosome was divided in two sections; variants are indicated as percent of the size of each section.

long PacBio reads and short Illumina reads. We generated PMGLs using as RG the corresponding assembly and as QG the Illumina reads that were used to construct the same assembly (*Materials and Methods*). As shown in Table 1, the five assemblies could be further refined using the PMGL pipeline. The position of each VS and the position and nature of the variants are shown in *SI Appendix*, Datasets S14–S18.

Detection of Genome Variation in SARS-CoV-2. The SARS-CoV-2 virus, causal agent of the current pandemic of COVID-19, was first sequenced in Wuhan, China (17). This first sequence is considered the RG of the virus. As the virus spreads it acquired new mutations. A very large number of viral particles have been sequenced from human samples and several studies have focused on tracing the evolution of the virus (18). We followed the PMGL pipeline using as RG the original isolate from Wuhan and as QG the original Illumina reads used to assemble the genome. As shown in Fig. 4, we did not find any VS.

We then selected sets of previously published sequence reads from two patients, one from the United States of America (patient 1) and other from Australia (patient 2) (*Materials and Methods*). We generated PMGLs using different combinations of RGs and QGs. In type-one experiments the Wuhan genome was used as RG and the sequence reads from either patient sample were used as QG. A number of VSs were found and all their underlying variants were uncovered. Some variants were shared by the virus from each patient while others were unique to one of the patients. The variants from each patient were introduced into the Wuhan genome, thus obtaining a CRG corresponding to the nucleotide sequence of the virus from each patient. In type-two experiments, such customized genomes were used as RGs and the sequence reads from the Wuhan virus were used as QGs. The same variants were found as those found in type-one experiments, but in these cases the variants from the virus of the patients' samples were located in the RG and those from the Wuhan genome were located in the QG (Fig. 4 and *SI Appendix*, Dataset S19).

Finally, we compared the genome of the samples derived from the two patients. We generated PMGLs using as RG the customized genome of the virus from patient one and as QG the sequence reads from patient two and vice versa. As expected, the variants shared by the virus of the two patients when compared to the Wuhan virus did not appear. The variants that were unique to either patient were all present in the two versions of the experiment (Fig. 4 and *SI Appendix*, Dataset S19).

Sequential Substitution of RGs during Mutation-Accumulation Experiments. We performed a mutation-accumulation experiment treating *E. coli* with the mutagenic agent ethyl methane sulphonate (EMS). From a culture of *E. coli* we isolated a colony derived from a single cell. This cell is the common ancestor of all the cells derived during the experiment. The colony was resuspended, an aliquot was stored at −80 °C to generate its DNA sequence (see *Materials and Methods*), and another aliquot (10^7 cells) was treated with EMS. After the treatment, colony 1, was isolated and handled again as the original colony. An aliquot was stored and another was treated with EMS. The procedure continued until colony 10 was isolated (Fig. 5*A*). The single cells that started the selected colonies were

Table 1. Variation profiles generated by the PMGL pipeline for different strains of *S. paradoxus*

Strain	VSs in original RG	VSs in CRG	VSs solved	Variants found	*SI Appendix* Dataset no.
UWOPS91-917.1	150	4	146	202	S14
CBS432	167	6	161	185	S15
N44	8	3	5	11	S16
YPS138	14	3	11	14	S17
UFRJ50816	63	4	59	74	S18

The PMGLs were generated using as RG the original assembly of the genome and as QG the sequence reads used to generate the original assembly.

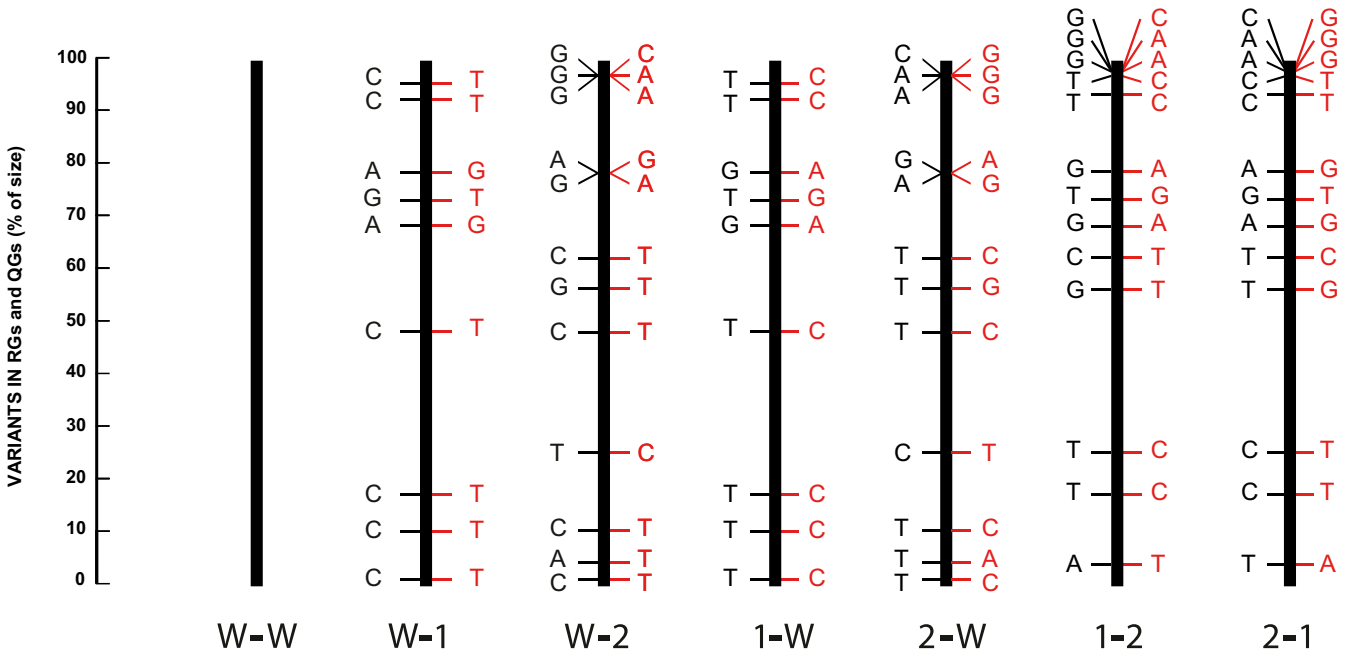

Fig. 4. SARS-CoV-2 genome variation determined with the PMGL pipeline. PMGLs were generated using RGs and QGs from different sources: W, original virus isolated from Wuhan; 1, virus from a patient from the United States; 2, virus from a patient from Australia. Each bar represents the variants found when particular combinations of RG and QG were used to construct the PMGL. At the bottom of each bar the RG (left) and the QG (right) are indicated. The scale indicates the position of the variants as percent of the virus genome. The variants are indicated in black for the RG and in red for the QG. A, adenine; C, cytosine; G, guanine, T, thymine.

ancestors of all the consecutive single cells that generated the rest of the colonies.

The genome of each of the 10 stored aliquots was sequenced generating the QGs of the corresponding colonies. The PMGL automatic pipeline was then applied. For colony 1 the PMGL was generated using as RG the CRG of *E. coli* (see above) and as QG the sequence reads of colony 1. For the other colonies the PMGL was generated using as RG the CRG of the previous colony and as QG the sequence reads of the colony. Thus, the PMGL for colony 2 was generated using as RG the CRG of colony 1 and as QG the sequence reads of colony 2. This procedure continued up to colony 10 for which the PMGL was generated using the CRG of colony 9 and the QG of colony 10 (Fig. 5*A*).

The PMGLs were scanned for VSs. The variants found in the analyses of all the PMGLs indicate the mutants generated at each step of the treatment with EMS. Including the 10 genomes analyzed, a total of 104 variants were found. After customization, no remaining VSs were present in any of the genomes. Fig. 5*B* shows the relative position and nature of the variants at each step of the experiment, and *SI Appendix*, Dataset S20 indicates the precise position of the signatures of variation and of the uncovered variants. The mutation landscape found in the experiment consists mainly (94% of the mutations found) of changes of G:C pairs to A:T pairs in the DNA. This is consistent with the mechanism of action of EMS (*Discussion*).

A PMGL was generated using the original RG of *E. coli* and the QG of colony 10. The PMGL was scanned for VSs. The uncovered variants corresponded to the addition of the variants found in the original *E. coli* RG plus those generated during each step of the experiment (*SI Appendix*, Dataset S21).

Discussion

The PMGL strategy is based on perfect matches between strings derived from the continuous structure of an RG and from the fragmented structure of the set of sequence reads of a QG. Such matches report the coverage at each nucleotide along the RG. Whenever there is a discrepancy between both genomes a VS appears, indicating the precise site of the corresponding variation. The nature of the variants is subsequently determined by targeted alignments of the corresponding region of the RG and the reads that contain the variants. The introduction of the variants into the RG produces a CRG that corresponds to the nucleotide sequence of the QG. The PMGL strategy has been presented and discussed by Palacios-Flores et al. (11).

The term RG usually refers to a genome that represents a particular group of evolutionary close organisms. In this study we have used the term RG in its broader context, referring to a genome that has been assembled in a continuous structure (or structures in the case of genomes with several chromosomes). If both the RG and the QG of the same organism are available, a PMGL can be generated between them. Such PMGL can be used to refine and customize the RG. After the introduction into the original genome of the variants found by the PMGL pipeline, the result is a customized genome. Such a customized genome can be considered as a refined RG. In some cases, the customized genome is the result of a particular manipulation, such as the isolation of a single cell or a particular event during experimental evolution; in such cases the resulting customized genome does not necessarily represent a new refined genome for a particular organism.

The RG of the yeast *S. cerevisiae* was first sequenced in 1996 (2). Since then it has been continuously refined, generating a total of 64 versions. As a proof-of-concept of the power of the PMGL to refine genomes, we were able to refine the 10 different versions used here, including V1 and V64. We found 1,208 and 119 VSs in the RGs of V1 and of V64, respectively. After customization, the V1 genome presented only 34 VSs. Thus, using the same set of sequence reads (*Materials and Methods*) we could generate a more accurate genome from V1 than that corresponding to the original V64. This is particularly interesting since

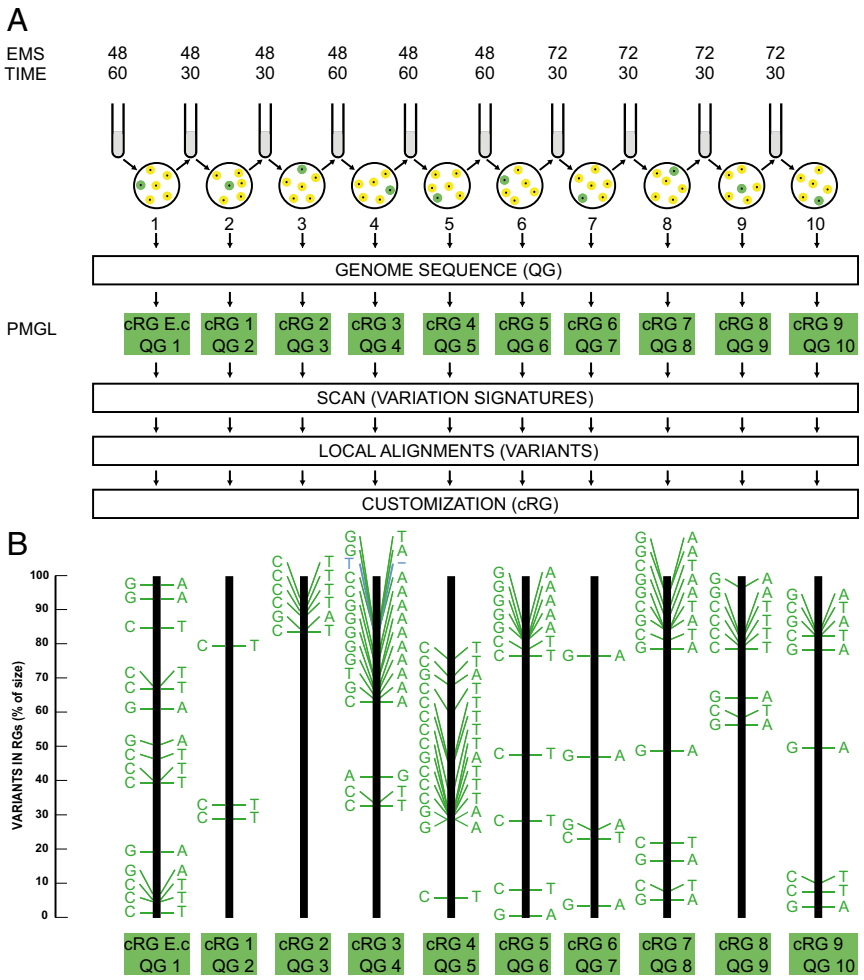

Fig. 5. Sequential substitution of RGs in a mutation-accumulation experiment. (*A*) Overall scheme of the experiment. A culture of *E. coli* was plated at high dilution to generate isolated colonies. One colony was resuspended in Luria medium and an aliquot containing 10^7 cells was incubated at 37 °C in the presence of EMS at the indicated concentration (in micrograms per milliliter) for the indicated time in minutes (TIME). The culture was diluted and an aliquot was plated in solid Luria medium and incubated at 37 °C for 3 d. An isolated colony, colony 1, was resuspended and treated with EMS at the concentration and during the time indicated. The culture was diluted, plated and incubated as indicated above for colony 1, generating colony 2. Colony 2 was treated with EMS at the concentration and time indicated and colony 3 was isolated as indicated for colony 2. The procedure was continued until colony 10 was isolated. The complete genome sequence of each of the 10 colonies was obtained. PMGLs were generated using the RGs and QGs indicated for each step (green boxes). The PMGLs were scanned for VSs. At each VS local alignments between the RG and QG uncovered the corresponding variants. The variants found in the QG were introduced into the RG to generate the CRG. (*B*) New variants found in each step of the mutation-accumulation experiment. Each step of the experiment (1 to 10) is presented as a black bar indicating at the bottom the RG and QG (green boxes) used to generate the PMGL. The position of the variants is indicated as percent of the size of the genome of *E. coli*. The variants found for each VS in the targeted alignments between the RG and the QG are shown. The nucleotide found in the RG is indicated at the left of the bar; the nucleotide found in the QG is indicated at the right of the bar. Color code of the variants: green, single-nucleotide variants; blue, microinsertion according to the RG. A, adenine; C, cytosine; G, guanine; T, thymine.

V64 was generated about 20 y later than V1 and after 63 refinement steps. In turn, V64 was further refined with PMGL, generating a new genome that decreased the sites of discrepancy with the QG from 119 to 14. It must be emphasized that the PMGL automatic pipeline reveals the exact position of the sites of discrepancy (VSs). The knowledge of the actual positions of the discrepancies allows the use of different targeted techniques to solve some or all of the remaining discrepancies after customization of the genome.

It must be taken into account that the variations found between the RG and the QG could be due to either mistakes in the assembly of the RG or to mutations that occurred in the strain used as QG. To assess the contribution of each of the two sources of variation, we performed the experiment presented in Fig. 2 *B* and *C* and *SI Appendix*, Dataset S6. An artificial QG was generated from V64 of the *S. cerevisiae* RG (*Results*). A PMGL was then generated using the RG of different versions of the genome and the artificial QG of V64. The fact that most (about 98%) of the solved VSs and the uncovered variants were identical to those found in the PMGL generated from the RG of V1 and the QG of the S288c strain, indicates that the variation found was mostly due to mistakes in the RG. This experiment ascertains that different versions of the genome (V1, V8, V15, V22, V29, V36, V43, V50, V57) were actually refined. We do not have a direct proof of the origin of the variants found in V64. They could be variants introduced in the assembly of V1 or differences between the V1 assembly and the isolate of strain S288c maintained at ATCC. If the latter is correct, the CRG of V64 would represent the refined RG of the isolate of strain S288c maintained at ATCC (*Materials and Methods*).

In this work we applied the automatic PMGL pipeline to refine the published RGs of different microorganisms. The microorganisms

were selected to have examples of each of the three domains of life: the archaea *M. maripaludis* and *P. furiosus*; the bacteria *E. coli*, *S. aureus*, and *B. subtilis*; and the eukarya *S. pombe* and *A. oryzae*. As shown in Fig. 3 and in *SI Appendix*, Datasets S7–S13, we could further refine all the RGs analyzed. These results suggest that, independently of the method used to generate the nucleotide sequence of any haploid RG, it could be further refined by applying the PMGL pipeline. This is further supported by the experiment presented in Table 1 and *SI Appendix*, Datasets S14–S18. *S. paradoxus* genomes assembled using state-of-the-art methodologies could be refined by the PMGL pipeline using as RGs each one of the corresponding assemblies and as QGs the Illumina reads used to generate such assemblies.

As discussed above, the variants found when comparing an RG with a QG could be derived from mistakes introduced when the genome used as RG was assembled and from mutations that occurred in the strain used as QG. In any case, the CRG represents an update of the nucleotide sequence of the QG being analyzed.

The experiments presented in Fig. 4 and *SI Appendix*, Dataset S19 using nucleotide sequences of SARS-CoV-2 (*Materials and Methods*) illustrate the accuracy, the reproducibility, and the flexibility of the PMGL strategy. The scan of the PMGL generated between the RG of the virus and the sequence reads used for its assembly did not produce any VS. The absence of VSs indicates that each nucleotide along the entire RG is identical to the corresponding nucleotide present in the sequence reads. This ascertains the accuracy of the RG of the virus. We then selected two patients, for which the sequence reads had been previously deposited in the NCBI (*Materials and Methods*). PMGLs were constructed using different combinations of RGs and QGs. The PMGL approach can transform a set of sequence reads (QG) into a CRG. This CRG can in turn be used as the RG to investigate the variation of other sets of reads. Experiments can be performed between different genomes, using them as either RGs or QGs. If the experiment is performed in the two directions, a double check of the accuracy of the variants found is possible. It must be highlighted that all the variants found in the PMGLs generated with SARS-CoV-2 sequences were perfectly congruent among them.

To generate a PMGL, the raw sequence reads are used, without statistically selecting or trimming for quality (11). Furthermore, since only perfect matches are allowed, the presence of sequences from human or other microorganisms in the reads should not be a problem when using the PMGL. Actually, we have checked all the strings of SARS-CoV-2 RG and did not found any perfect match with the human, the *E. coli*, or the *S. aureus* RGs. The complete automatic PMGL pipeline, from the crude sequence reads to the generation of a CRG of the virus present in a sample from a patient, takes only about 1 h. The time that the PMGL pipeline uses is not directly proportional to the size of the genome; actually, in the case of bacterial genomes, of about 5 Mb, the procedure takes about 2 h. With the appropriate computation infrastructure several genomes can be simultaneously analyzed. Thus, the PMGL strategy could be a useful tool to trace the evolution of a pathogenic agent during the course of an epidemic or pandemic condition.

Mutation accumulation experiments (19) are a particular case of experimental evolution experiments. A population derived from a single cell is cultivated under certain stress conditions. According to the stress applied, mutations appear in different cells of the culture. At certain time points a single cell is selected and used to continue the experiment and to analyze the mutants that have been generated. In the experiment presented in Fig. 5 and *SI Appendix*, Dataset S20, we used the mutagenic agent EMS to treat *E. coli* cells. The genomes of colonies derived from single cells were sequenced and customized using the PMGL pipeline. The number of variants increased at each step of the experiment. The variants found in steps 1 to 10 were congruent with the mechanism of action of EMS (20). In fact, EMS alkylates guanine in the DNA generating the abnormal base O-6-ethylguanine. During a first replication, such an abnormal base is paired with thymine instead of cytosine. In a second replication thymine pairs with adenine. Thus, after two replications, EMS induces base substitutions of guanine–cytosine (G/C) pairs to adenine–thymine (A/T) pairs in the DNA.

The generation of PMGLs using the customized genome of the previous stage actually "erased" the mutations that appeared in the previous stages and allowed the revelation of the mutations that appeared only in each particular step of the experiment. Since these types of studies are in some cases very complex, the isolation of the desired steps can provide a simpler approach to analyze the acquired mutations at different stages of the experiments. Thus, the PMGL can be a precise and sensitive strategy in different types of experimental evolution experiments.

We propose that the PMGL strategy is a powerful tool that can be used as a final step in the refinement of any haploid genome, independently of the methodology used to assembly such genome, either de novo assembly or by comparison with an RG using different types of algorithms. Even in the cases where highly accurate RGs exist, the genome of the actual organism used could have acquired new mutations and a good practice is to customize the genome for specific experiments. The PMGL strategy is also a powerful tool to accomplish this task. Finally, the PMGL strategy can also be used to follow the evolution of a pathogenic agent or to analyze experimental evolution experiments.

Materials and Methods

Strains and Culture Conditions. *S. cerevisiae* strain S288C, *S. pombe* strain 972h, and *E. coli* strain K12, substrain MG1655 were obtained from the ATCC and stored at −80 °C. *S. cerevisiae* and *S. pombe* strains were cultured in liquid YPD medium at 30 °C with agitation (250 rpm). *E. coli* was cultured in Luria medium either in liquid or in agar plates at 37 °C.

RGs of Microorganisms. *S. cerevisiae* V1, V8, V15, V22, V29, V36, V43, V50, V57, and V64 were obtained from the *Saccharomyces* Genome Database (http://sgd-archive.yeastgenome.org/sequence/S288C_reference/genome_releases/?fbclid=IwAR1aUGohre32X-yhh9nxyWsgzkBP49aXAzkG_1iUdjbvwt4786ScCaGv1d4).

The genomes of other microorganisms were downloaded from the NCBI: *M. maripaludis* C7 (https://www.ncbi.nlm.nih.gov/nuccore/CP000745.1?report=fasta); *P. furiosus* DSM 3638 (https://www.ncbi.nlm.nih.gov/nuccore/AE009950.1?report=fasta); *E. coli* MG 1655 (Migula) Castellani and Chalmers (ATCC 700926) (https://www.ncbi.nlm.nih.gov/nuccore/U00096.3?report=fasta); *S. aureus* subsp. *aureus* TCH1516 (https://www.ncbi.nlm.nih.gov/nuccore/CP000730.1?report=fasta); *B. subtilis* subsp. *subtilis* 168 (https://www.ncbi.nlm.nih.gov/nuccore/AL009126.3?report=fasta&log$=seqview); *A. oryzae* RIB40 (https://www.ncbi.nlm.nih.gov/genome/526); *S. pombe* 972h (https://www.ncbi.nlm.nih.gov/genome/?term=schizosaccharomyces%20pombe).

The RGs of the different *S. paradoxus* strains used in this work were downloaded from https://yjx1217.github.io/Yeast_PacBio_2016/data/.

The Illumina sequencing reads were downloaded from the NCBI Short Reads Archive (SRA) under accession no. PRJNA340312 (https://www.ncbi.nlm.nih.gov/sra?term=PRJNA340312).

Genomes of SARS-CoV-2. The RG and the sequence reads from Wuhan, as well as the sequence reads from patient samples, were downloaded from the NCBI as follows: genome assembly from Wuhan (https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3?report=fasta&log$=seqview); sequence reads from Wuhan (https://www.ncbi.nlm.nih.gov/sra/?term=SRR10971381%20); sequence reads from patient 1 (https://www.ncbi.nlm.nih.gov/sra/?term=SRR12348890); sequence reads from patient 2 (https://www.ncbi.nlm.nih.gov/sra/?term=SRR12162149).

DNA. *S. cerevisae* S288C DNA was isolated using the Yeast DNA Extraction Kit from Thermo Scientific. *S. pombe* 972h DNA was isolated using the YeaStar Genomic DNA Kit from Zymo Research. *E. coli* strain K-12 substrain MG1655 DNA was isolated using the QIAamp DNA Mini Kit from Qiagen. DNA from

M. maripaludis C7, *P. furiosus* DSM 3638, *S. aureus* subsp. *aureus* TCH1516, *B. subtilis* subsp. *subtilis* 168, and *A. oryzae* RIB40, were acquired from the ATCC.

DNA Sequence. To generate the QGs of the strains that were grown in the laboratory, aliquots of the corresponding ATCC stock were cultivated as a lawn in solid medium and then DNA was purified.

To generate all the QGs, libraries were prepared using the TruSeq DNA PCR-Free from Illumina. Sequencing was performed in a NextSEq. 500 Mid Output v2 Kit (300 cycles) from Illumina.

Size of the RG and QG Strings. The size of the strings must be identical in both the RG and the QG. The length must be large enough to have a negligible probability to generate at random specific nucleotide sequences of the corresponding size. This probability increases as the genome size increase. accordingly, with large-size genomes the strings might be longer. On the other hand, the size of the strings must be adequate to cover the sequence reads with a sufficient amount of strings. In the case of short sequence reads, the strings should be of a relatively small size. In general, we have found that for small genomes as those from bacteria and yeast, using Illumina reads, strings from 20 to 25 nucleotides are a good choice. In this work we used strings of 25 nucleotides.

PMGL Pipeline. The PMGL pipeline has been automatized and deposited in GitHub: https://github.com/LIIGH-UNAM/PerfectMatchGenomicLandscapePipeline. It has been thoroughly described and discussed by Palacios-Flores et al. (11). This pipeline consists of six automatized computational modules: 1) Generation of the RG self-landscape; 2) generation of the PMGL; 3) scanning of the PMGL for SVs; 4) generation of the first alignment at each SV; 5) interpretation and extension of alignments; 6) generation of a CRG. Variants are validated by the disappearance of SVs in a PMGL generated between the CRG and the original QG. The published automatic pipeline contains seven columns. In this study we used only four columns, as described in Fig. 1 and examples shown in *SI Appendix*, Datasets S1–S4.

Data Availability. The CRGs generated in this study for the different microorganisms, as well as the sequence reads used as QGs to generate the corresponding PMGLs, have been deposited in GitHub: https://github.com/LIIGH-UNAM/RefinementOfReferenceGenomesByThePerfectMatchGenomeLandscape/.

ACKNOWLEDGMENTS. We thank Eglee Lomelin for assistance with the artwork. This work received support from Luis Aguilar, Alejandro de León, and Carlos S. Flores of the Laboratorio Nacional de Visualización Científica Avanzada in providing supercomputer services, and from the Instituto Nacional de Medicina Genómica in performing the Illumina sequence.

1. F. R. Blattner *et al*., The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
2. A. Goffeau *et al*., Life with 6000 genes. *Science* **274**, 546, 563–567 (1996).
3. *C. elegans* Sequencing Consortium, Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science* **282**, 2012–2018 (1998).
4. M. Adams *et al*., The *Drosophila melanogaster* genome. *Science* **287**, 2185–2195 (2000).
5. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
6. R. H. Waterston *et al*.; Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
7. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
8. D. R. Zerbino, E. Birney, Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
9. M. L. Metzker, Sequencing technologies—The next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
10. K. Reinert, B. Langmead, D. Weese, D. J. Evers, Alignment of next-generation sequencing reads. *Annu. Rev. Genomics Hum. Genet.* **16**, 133–151 (2015).
11. K. Palacios-Flores *et al*., A perfect match genomic landscape provides a unified framework for the precise detection of variation in natural and synthetic haploid genomes. *Genetics* **208**, 1631–1641 (2018).
12. N. Annaluru *et al*., Total synthesis of a functional designer eukaryotic chromosome. *Science* **344**, 55–58 (2014).
13. S. M. Richardson *et al*., Design of a synthetic yeast genome. *Science* **355**, 1040–1044 (2017).
14. P. A. Audano, S. Ravishankar, F. O. Vannberg, Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. *Bioinformatics* **34**, 1659–1665 (2018).
15. S. R. Engel *et al*., The reference genome sequence of *Saccharomyces cerevisiae*: Then and now. *G3 (Bethesda)* **4**, 389–398 (2014).
16. J. X. Yue *et al*., Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.* **49**, 913–924 (2017).
17. F. Wu *et al*., A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
18. A. Gómez-Carballa, X. Bello, J. Pardo-Seco, F. Martinón-Torres, A. Salas, Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 superspreaders. *Genome Res.* **30**, 1434–1448 (2020).
19. H. Lee, E. Popodi, H. Tang, P. L. Foster, Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2774–E2783 (2012).
20. D. Shah *et al*., Mutagenic action of ethyl methane sulphonate (EMS): A review. *J. Res. Dev. (Srinagar)* **16**, 63–68 (2016).