bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title: GIP: An open-source computational pipeline for mapping genomic instability from protists to cancer cells

Authors: Gerald F. Späth[1] and Giovanni Bussotti[1,2,*]

Affiliations:

[1]Institut Pasteur, INSERM U1201, Unité de Parasitologie moléculaire et Signalisation, Paris, France.

[2]Hub de Bioinformatique et Biostatistique - Département Biologie Computationnelle, Institut Pasteur, 75015 Paris, France

*corresponding author: Giovanni Bussotti, giovanni.bussotti@pasteur.fr

Abstract:

Genome instability has been recognized as a key driver for microbial and cancer adaptation and thus plays a central role in many human pathologies. Even though genome instability encompasses different types of genomic alterations, most available genome analysis software are limited to just one kind mutation or analytical step. To overcome this limitation and better understand the role of genetic changes in enhancing pathogenicity we established GIP, a novel, powerful bioinformatic pipeline for comparative genome analysis. Here we show its application to whole genome sequencing datasets of *Leishmania*, *Plasmodium, Candida*, and cancer. Applying GIP on available data sets validated our pipeline and demonstrated the power of our analysis tool to drive biological discovery. Applied to *Plasmodium vivax* genomes, our pipeline allowed us to uncover the convergent amplification of erythrocyte binding proteins and to identify a nullisomic strain. Re-analyzing genomes of drug adapted *Candida albicans* strains revealed correlated copy number variations of functionally related genes, strongly supporting a

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mechanism of epistatic adaptation through interacting gene-dosage changes. Our results illustrate how GIP can be used for the identification of aneuploidy, gene copy number variations, changes in nucleic acid sequences, and chromosomal rearrangements. Altogether, GIP can shed light on the genetic bases of cell adaptation and drive disease biomarker discovery.

One Sentence Summary:

GIP - a novel pipeline for detecting, comparing and visualizing genome instability.

Keywords:

Genome instability, aneuploidy, copy number variations, single nucleotide variation, structural variation, genetic adaptation.

Main Text:

In recent years, the field of genomics has rapidly expanded with a fast increase in the number of newly sequenced genomes (1). This surge is a direct consequence of the development of new and ever more efficient, high-throughput capable sequencing technologies (2). On the one hand, the improvement in long reads technology allowed the generation of high-quality genome assemblies (3,4). On the other hand, the decreasing costs for short-reads sequencing and the parallel increase in sequencing throughput propelled the exponential increase of available whole genome sequencing (WGS) data (5). Thanks to these advances one can reasonably expect WGS to rapidly become a key component of personalized medicine and clinical applications. In this context several consortia-based projects have been established with the goal to produce WGS for the study of different biological systems (5), and a number of publicly available databases have been compiled or updated (6). Parallel to data availability, many bioinformatics

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tools have been developed to perform specific genome analysis tasks (7,8). For instance, tools such as Freebayes (9), CNVnator (10) and DELLY (11) have been respectively used for the detection or characterization of DNA single nucleotide variants (SNVs), copy number variations (CNVs), and structural variations (SVs), but their scope is limited to the analysis of one genomic feature at the time. A number of integrative WGS pipelines and workflows have been established combining the execution of multiple bioinformatics algorithms serving different analysis steps (12). Even though continuous progress has been made, there is no standardized or unified approach for genomic investigation. For the development of improved WGS data analysis pipelines, several important requirements need be considered, including portability, reproducibility, scalability and compatibility with high-performance computing (HPC) clusters and remote cloud computing. Here we introduce a novel genome instability pipeline (GIP) that fulfills all these requirements. GIP facilitates the genome-wide detection, quantification, comparison and visualization of chromosome aneuploidies, gene CNVs, SNVs and SVs. GIP is implemented in Nextflow (13), a workflow language that allows to execute GIP seamlessly in local workstation, on an HPC or remotely in the cloud. All required environment and software dependencies of GIP are fulfilled and provided with a Singularity container, thus making GIP reproducible, easy-to-install and easy-to-use. GIP allows the use of giptools, a tool-suit of R-based modules for genome data exploration, enabling the comparison of sample sub-sets. GIP and giptools generate a summary report with publication-quality figures and spreadsheet tables. GIP and giptools constitute a single framework for WGS analysis suitable both for large scale batch analysis of individual genomes and comparison of samples from different experimental conditions or origins. Lastly, a key strength of GIP and giptools is the general applicability to different eukaryotic species. We already successfully applied GIP on the analysis of *Leishmania* genomes (14-16). In this study, we validate the use of GIP and giptools using WGS data from published datasets of the three major human pathogens

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

74 *Leishmania infantum*, *Plasmodium vivax* and *Candida albicans* and as well as three human
75 cancer cell lines. Furthermore, we demonstrate how the extensive and powerful analytical
76 approach operated by GIP and giptools can be used to find new biological signal that escaped
77 previous analyses.
78

79 ## Results:

80 **The GIP workflow.** GIP is a tool for scientific investigation compatible with Linux systems,
81 requiring minimal configuration and distributed as a self-contained package. GIP consists of
82 three files: the Nextflow pipeline code, the configuration and the Singularity container files.
83 The minimum required input is a paired-end WGS data set and a reference genome assembly
84 in the standard fastq and FASTA format, respectively. GIP analyses include (i) extracting
85 genomic features such as assembly gaps or repetitive elements, (ii) mapping the reads, (iii)
86 evaluating chromosome, gene and genomic bin copy numbers, (iv) identifying and visualizing
87 copy number variation with respect to the reference genome, (v) identifying and quantifying
88 gene clusters, (vi) detecting and annotating SNVs, (vii) measuring non-synonymous (N) and
89 synonymous (S) mutations for all genes, (viii) detecting SVs including tandem duplications,
90 deletions, inversions and break-ends translocations using split-read and read-pair orientation
91 information, and (ix) producing a report file providing summary statistics, tables and
92 visualizations (**Fig. 1**). GIP allows to customize filtering and visualization options via the
93 configuration file (see methods). The output of GIP can be used as input for giptools, a tool-
94 suite to compare sample sub-sets and highlight chromosome copy number, gene copy number
95 and SNV differences.
96

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Applying giptools on a *Leishmania infantum* case study. GIP permits the batch analysis of a set of individual samples, where each sample is considered separately and compared only with respect to the provided reference genome assembly. As a consequence, all variants and copy number alterations detected in a sample merely reflect the differences between the sequenced and the reference genomes. While this application may be sufficient in some circumstances, research projects often involve downstream comparison between samples. Examples include the comparison of gene or chromosome copy variation number between for example drug resistant and drug susceptible samples, or the juxtaposition of SNVs detected in isolates from different geographic areas. For this purpose, we developed giptools, a suit of thirteen modules that allows to compare samples processed by GIP (**Table 1**). All modules in giptools are fully embedded in the Singularity container and they are provided with their own documentation.

Table 1: giptools modules

Module name	Purpose
karyotype	Compare the chromosome sequencing coverage distributions
binCNV	Compare bin sequencing coverage in 2 samples
geCNV	Compare gene sequencing coverage in 2 samples
ternary	Compare gene sequencing coverage in 3 samples
ternaryBin	Compare bin sequencing coverage in 3 samples
SNV	Compare SNVs in multiple samples
binDensity	Density plot of bin sequencing coverage of many samples
geInteraction	Detect CNV genes in many samples and produce correlation-based networks
genomeDistance	Compare samples genomic distance
phylogeny	Extract the SNVs union and infer the phylogenetic tree
convergentCNV	Detect convergent CNV gene amplifications
overview	Overview of the sequencing coverage of chromosomes, genomic bins and genes
panel	Extract genomic information of a gene panel

To illustrate the type of exploratory data analyses and the biological questions that can be addressed, we tested giptools on a previously analyzed dataset of seven clinical *Leishmania infantum* isolates from Tunisia (14). *Leishmania* is the etiological agent of leishmaniasis, a life-threatening human and veterinary disease affecting 12 million people worldwide (17). The

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parasites were derived from patients affected with visceral leishmaniasis, expanded in cell culture and their genomes were sequenced. The dataset includes four Glucantime drug susceptible isolates and three isolates from relapsed patients, and thus comparison may inform on genetic factors resulting in treatment failure. giptools allowed the detection and visualization of pervasive intra-chromosomal CNVs across the thirty-six *Leishmania* chromosomes (**Fig. 2A**). Additionally, giptools enables targeted comparison of normalized genomic bin sequencing coverage of sample pairs. We used giptools' `binCNV` module to compute the ratio between corresponding genomic bins of the strains LIPA83 over ZK43, which correspond to a first-episode and a relapse leishmaniasis isolate, respectively. giptools represents different chromosomes as separate panels (**Fig. 2B**), as part of single genome-wide overview (**Fig. S1A**) or as distinct plots (**Fig. S1B**). This analysis allowed the identification of 2,905 and 2,208 bins that were respectively amplified or depleted in LIPA83 with respect to ZK43. The results are returned by giptools as a Microsoft Excel table (.xlsx format) providing ratio scores at each genomic position (**Table S1**). Likewise, giptools permits three-way comparisons of normalized genomic bin sequencing coverage with ternary plots (**Fig. 2C**). We used this representation to display the genomic bin relative abundance in samples ZK43, LIPA83 and ZK28. The analysis shows important strain-specific differences in bin copy number that are visualized by shifts of the signals out of the center. Similar to genomic bin analysis, giptools makes it possible to compare the sequencing depth of annotated genes and thus the copy number in two or three samples (**Fig. 2D** and **Fig. S2**). However, the determination of gene copy number might be impeded by (i) short read length or fragment insert size, (ii) the complexity of the target genome, or (iii) the presence of repetitive elements. The read map quality (MAPQ) score is a measure that reflects how much each gene is supported by unambiguously mapped reads (high MAPQ) in contrast to multimapping reads (low MAPQ). Together with the coverage, GIP also computes the mean read MAPQ score for each gene and allows a different strategy to determine the copy

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

number of low MAPQ genes (see methods). The evaluation of LIPA83 and ZK43 gene coverage ratio scores revealed 13 gene CNVs with a stringent MAPQ cutoff of 50 (**Table S2**). The maximum normalized coverage value of 6.5 was observed for a putative amastin surface glycoprotein (LINF_310009800). Other examples of gene CNVs in this set include the putative surface antigen protein 2 (LINF_120013500) and the heat shock protein HSP33 (LINF_300021600) (**Table S2**). Genes falling below a user defined MAPQ score and sharing high level of sequence similarity are assigned to the same gene cluster, and their measured coverage scores are averaged across all members of the group. Low MAPQ scores can also be associated with single genes, e.g. in the case of internal repetitive elements that cause multiple ambiguous alignments inside the gene itself, or if mapping occurs in possibly misannotated intergenic regions. The LIPA83/ZK43 comparison showed 27 CNV gene clusters, including cluster clstr303 (3 genes annotated as 'amastin-like') and cluster clstr16 (2 tb-292 membrane-associated protein-like proteins) (**Table S2**). These results demonstrate the power of GIP and giptools to detect and compare intra-chromosomal CNVs in *Leishmania* at genomic bin level. Conveniently, analogous two- or three-ways comparisons can be applied to reveal copy number variations at individual gene or gene cluster levels.

Comparative genomics of a *Plasmodium vivax* WGS dataset. We next applied GIP and giptools on other biological systems to demonstrate its broad applicability outside the *Leishmania* field, including the human apicomplexan parasite *Plasmodium vivax*. *Plasmodium vivax* is a protist parasite and a human pathogen causing malaria. *Plasmodium vivax* gives rise every year to 130 million clinical cases (18), and it is estimated that 2.5 billion people are at risk of infection worldwide (19-21). We applied GIP and giptools to investigate genomic variations across a sizeable dataset of 222 *Plasmodium vivax* genomes isolated from clinical samples of 14 countries worldwide (22,23) (**Table S3**). The phylogenetic tree reconstruction

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and PCA analyses (**Fig. 3A** and **B**) showed a high correlation between genotypes and the geographic origin of the samples. However, we detected substantial genomic variability between isolates collected at smaller geographical scale, with 14,555 SNVs (~42% of the total) uniquely characterizing representative samples from five Ethiopian study sites (23) (**Fig. 3C** and **D**). This result may reflect diverging evolutionary trajectories radiating from few founder strains. At gene level we profiled the copy number variations of two gene panels. The first panel accounts for 43 previously described genes encoding for potential erythrocyte binding proteins suggested to operate at the interface of the parasite-host invasion process (23). The second panel includes two drug resistance markers comprising the chloroquine resistance transporter PVP01_0109300 and the multidrug resistance protein 1 PVP01_1010900, and four proteins implicated in red blood cell invasion, such as the merozoite surface protein gene MSP7 PVP01_1219700, the reticulocyte binding protein gene 2c PVP01_0534300, the serine-repeat antigen 3 PVP01_0417000, and the reticulocyte binding protein 2b PVP01_0800700) (24-33). Read depth analysis indicated that four genes in the panel (PVP01_0623800, PVP01_1031400, PVP01_1031200, PVP01_1031300) show a high degree of variability, with amplifications observed in samples from distinct geographic (**Fig. 3E**). This convergence is sign of strong natural selection, which further sustains the functional importance of these genes in the infection process. Furthermore, 6 genes positioned on chromosome 14 are absent in the Thai strain PD0689_C as a result of the loss of this chromosome (nullisomy) (**Fig. 3E** and **F**). Finally, the comparison of synonymous and non-synonymous SNVs in the panel of genes revealed important differences between sample groups. Our analysis indicates an overall higher number of non-synonymous mutations in Ethiopian compared to Cambodian isolates, therefore suggesting a stronger evolutionary pressure acting on the African strains (**Fig. 3G**). Taken together these analyses well illustrate how GIP and giptools can be readily applied for bulk

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

188 analysis of *Plasmodium vivax* genomes to assess genome diversity, extract evolutionary
189 information and identify potential disease biomarkers.

190

191 **Gene CNV analysis of *Candida albicans* evolutionary adapted strains.** We next applied GIP
192 and giptools to the human fungal pathogen *Candida albicans*, an opportunistic yeast exhibiting
193 major genome plasticity (34-43) and causing hundreds of thousands of severe infections each
194 year (44). Candidemia, a bloodstream infection with *Candida*, are often associated with high
195 rates of morbidity and mortality (15–50%) notwithstanding existing antifungal treatments
196 (45,46). We applied GIP and giptools to a *Candida albicans* WGS dataset described in a recent
197 study that covers five different progenitor strains (P75063, P75016, P78042, SC5314,
198 AMS3050) and investigates CNVs driving tolerance and resistance to anti-fungal azole drugs
199 (47). We analyzed nineteen samples, including (i) four clinical isolates (P75063, P75016,
200 P78042, SC5314), (ii) seven strains selected *in vitro* against the anti-fungal drug fluconazole
201 (FLC) (AMS4104, AMS4105, AMS4106, AMS4107, AMS4397, AMS4444, AMS4702), (iii)
202 four isogenic colonies adapted to the drug miconazole (AMS3051, AMS3052, AMS3053 and
203 AMS3054) together with their progenitor (AMS3050), and (iv) three colonies derived from a
204 miconazole-adapted population and isolated on a rich medium (AMS3092, AMS3093 and
205 AMS3094) (47-49). GIP and giptools were able to reproduce previous observations of the
206 amplification of the genes for the drug efflux pumps TAC1 (orf19.3188) and ERG11
207 (orf19.922), for the stress response proteins HSP70 (orf19.4980), CGR1 (orf19.2722), ERO1
208 (orf19.4871), TPK1 (orf19.4892), ASR1 (orf19.2344), PBS2 (orf19.7388) and CRZ1
209 (orf19.7359), and for proteins involved in membrane and cell wall integrity, including CDR3
210 (orf19.1313), NCP1 (orf19.2672), ECM21 (orf19.4887), MNN23 (orf19.4874), RHB1
211 (orf19.5994) and KRE6 (orf19.7363) (**Table S4**). Furthermore, the powerful comparative
212 approach of our pipeline permitted the discovery of 1,505 genes showing correlating or anti-

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

213 correlating copy number variations (**Fig. 4A** and **B**, **Table S4**), which could be assigned to nine
214 distinct correlation clusters (CC) (**Fig. S3A**, **Table S5**) that escaped previous analyses. We
215 verified the sequencing coverage of genomic regions encompassing gene CNVs, including
216 three regions amplified in fluconazole resistant strains (**Fig. 4B**, **Fig. S3B**) (47) and a region
217 whose amplification correlates with the level of miconazole resistance (47) (**Fig. S3C**), as well
218 as the loss of heterozygosity associated to the depletion of chromosome 3 left arm in sample
219 AMS3051 (**Fig. S3D**) (47). Eventually, by representing genes and absolute correlation
220 respectively as nodes and edges of a network, we identified 9 highly interconnected network
221 clusters (NC) (**Fig. 4C**, **Table S6**). NC7, NC8 and NC9 embody genes from individual
222 chromosomes, respectively chromosomes 1, 3 and 4. The most parsimonious explanation for
223 the high levels of correlation observed in these NCs (**Fig. 4C**) is the occurrence of sub-
224 chromosomal amplifications affecting several adjacent genes. A different scenario is pictured
225 for each of the remaining NCs (NC1-6) where the genes are located on different chromosomes
226 thus suggesting genetic interactions that causes coordinated changes in gene copy number. The
227 gene ontology (GO) and metabolic pathway analyses revealed a significant functional
228 enrichment of genes expressed on the cell surface and involved in the interaction with the host
229 (NC2), gibberellin biosynthesis (NC3), transmembrane nucleobase transporters (NC4) and
230 gluconeogenesis (NC5) (**Table S7**). Altogether, GIP and giptools are validated by reproducing
231 previously published results, and beyond that can drive new biological findings as documented
232 by the discovery of a network of epistatic CNV interactions supporting genomic adaptation in
233 *Candida albicans* populations under drug selection.
234
235 **Exploring instability of larger genomes using cancer cell lines as a benchmark.** The larger
236 genome size, and the higher number of genes and WGS reads can represent a challenge when
237 working with higher eukaryotes. For the purpose of comparison, the human genome is ~216
238 times larger than the one of *Candida albicans* we analyse in this study. Therefore, we sought

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to evaluate the applicability of the GIP and giptools framework to human data and utilized a panel of genomes from cancer cell lines as a test set. In our analyses we considered publicly available WGS data of the cell lines T47D, NCI_H460 and K562 (50), which respectively derive from human breast, lung and blood cancers. The karyotype analysis revealed aneuploidy for all chromosomes except chromosome 4 (**Fig. 5A**). The observed heterogeneity in read depth across chromosomes, illustrated by large interquartile range in the boxplot, suggests sub-chromosomal or episomal copy number variations, or the co-existence of karyotypically different sub-populations. Indeed, the coverage analysis confirmed the pervasive presence of CNVs both at chromosomal and sub-chromosomal levels (**Fig. S4**) with remarkable instability observed for specific chromosomes, e.g. chromosomes 6, 9, 10 and 16 (**Fig. 5B**). Overall, we detected 1,647,016 SNVs (**Supplementary Data 1**) and allele frequency shifts with respect to the reference genome, suggesting haplotype selection and the preferential expression of distinct alleles in different cell lines (**Fig. 5C** and **Fig. S5**). Furthermore, we identified repeated loss of heterozygosity events and uneven distribution of SNVs that form "patches" of high frequency correlating with chromosomal and sub-chromosomal CNVs (**Fig. 5D**, **Fig. 5E** and **Fig. S6**). These results identify GIP and giptools as a powerful new platform to reveal loci, genes or alleles that are under natural selection in cancer cells, thus allowing important new insight into the genetic basis of tumor development, cancer cell evolution and drug resistance.

Discussion:

Genome instability is a key driver of evolution for microbial pathogens and cancer cells (51) and a major source of human morbidity. Here we introduce GIP and giptools, an integrated framework for the genotype profiling of biological systems exploiting genome instability for adaptation. We document the power and versatility of GIP and giptools by performing genomic screenings of three major pathogenic eukaryotes and human cancer cell lines. While originally

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

deployed for *Leishmania* genome analysis, in this study we validate the use of our pipeline on other organisms reproducing expected results. For example, in *Candida albicans* we confirmed the CNVs correlating to drug resistance as well as a loss of heterozygosity event (**Fig. S3B-C** and **D**). Parallel to this we also show how GIP and giptools can be used for data mining and scientific discovery. New findings include (i) the discovery of the convergent amplification of erythrocyte binding proteins in *Plasmodium vivax* strains sampled from distinct geographic areas (**Fig. 3E**), (ii) the detection of a nullisomic strain (**Fig. 3F**), (iii) the identification of correlated copy number variations between genes positioned on separate chromosomes of *Candida albicans* adapting strains, and (iv) the functional association of such genes, strongly supporting a mechanism of epistatic interactions exerted through gene-dosage changes, and corroborating previous reports on adapting *Leishmania* populations (52).

Importantly, GIP and giptools overcome key limitations of current analysis tools, such as the breadth of analysis that is often limited to individual types of mutations, and the lack of genome-wide, comprehensive reports. To ease genome instability investigations our pipeline offers a single solution to karyotype, gene CNV, SNV and SV batch analyses, providing summary reports and high-quality, genome-wide visualizations. Furthermore, many current tools identify variations with respects to a reference assembly only, which leaves the between samples comparisons to external tools that need installing, may be incompatible in terms of file format, and may rely on different analytical assumptions. To address this limitation giptools enables custom sample comparisons, to explore differences and common features between genomes, and provides a vast choice of analytical tools with compatible features. Likewise, current tools are often restricted to the analysis of data from one or few species only (53-57), but are not generally applicable to different biological systems, which interferes with the investigation of genome variations across multiple species and the exploration potentially conserved genomic adaptation mechanisms. By contrast, our pipeline limits as much as possible

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

289 the use of hardcoded parametrization, which could limit its use to a specific organism.
290 Therefore, GIP's flexible design makes it adapted for the genome analysis of both model and
291 non-model organisms, including *Leishmania* or human.

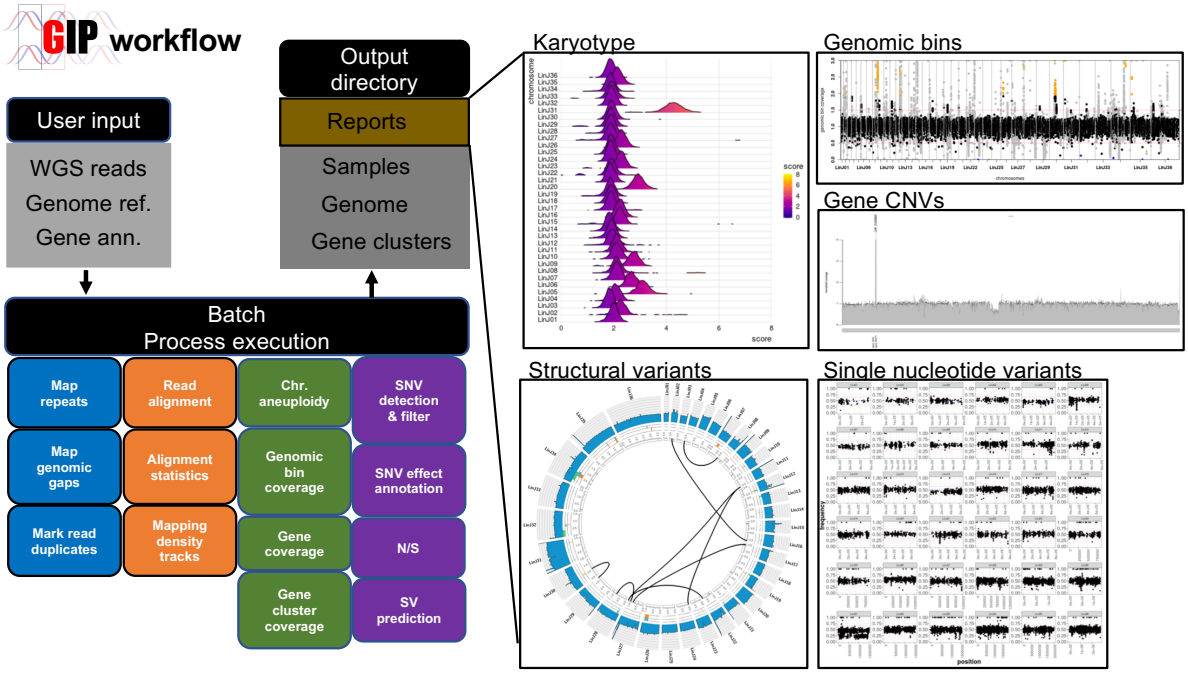
292 Many current tools are further limited in software portability and reproducibility across
293 different computer environments, which can produce faulty results calling their clinical
294 application into question. Conversely, thanks to the Singularity implementation all required
295 software are embedded and provided within the software container. As a consequence, users
296 can easily recreate the same work environment just by downloading the pipeline container, and
297 reproduce exactly the same publication-quality plots and tables presented in this study. Lastly,
298 one more common limitation is posed by software scalability. In the WGS domain, with the
299 rapid increase new samples made available and the enormous amount of data generated in each
300 sequencing run, the CPU and memory resources of local workstation risk to quickly become
301 inadequate for data analysis. Therefore, it is paramount that WGS tools are implemented to run
302 on high-performance computing (HPC) clusters and feature remote cloud computing solutions.
303 Because of its Nextflow implementation GIP can be executed on a local machine, on cluster
304 resource manager or the cloud. GIP can be applied on individual samples and without additional
305 effort on large WGS data sets for batch computation as shown for the 222 *Plasmodium vivax*
306 genomes.

307 These results well illustrate how GIP and giptools can be applied to perform extended
308 genomic analyses in different biological systems and drive biomedical discovery. To conclude,
309 we believe that GIP and giptools represent a step forward toward reproducible research in
310 genomics, and provide a robust computational framework to study how microbes and tumor
311 cells harness genome instability for environmental adaptation and fitness gain.
312

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

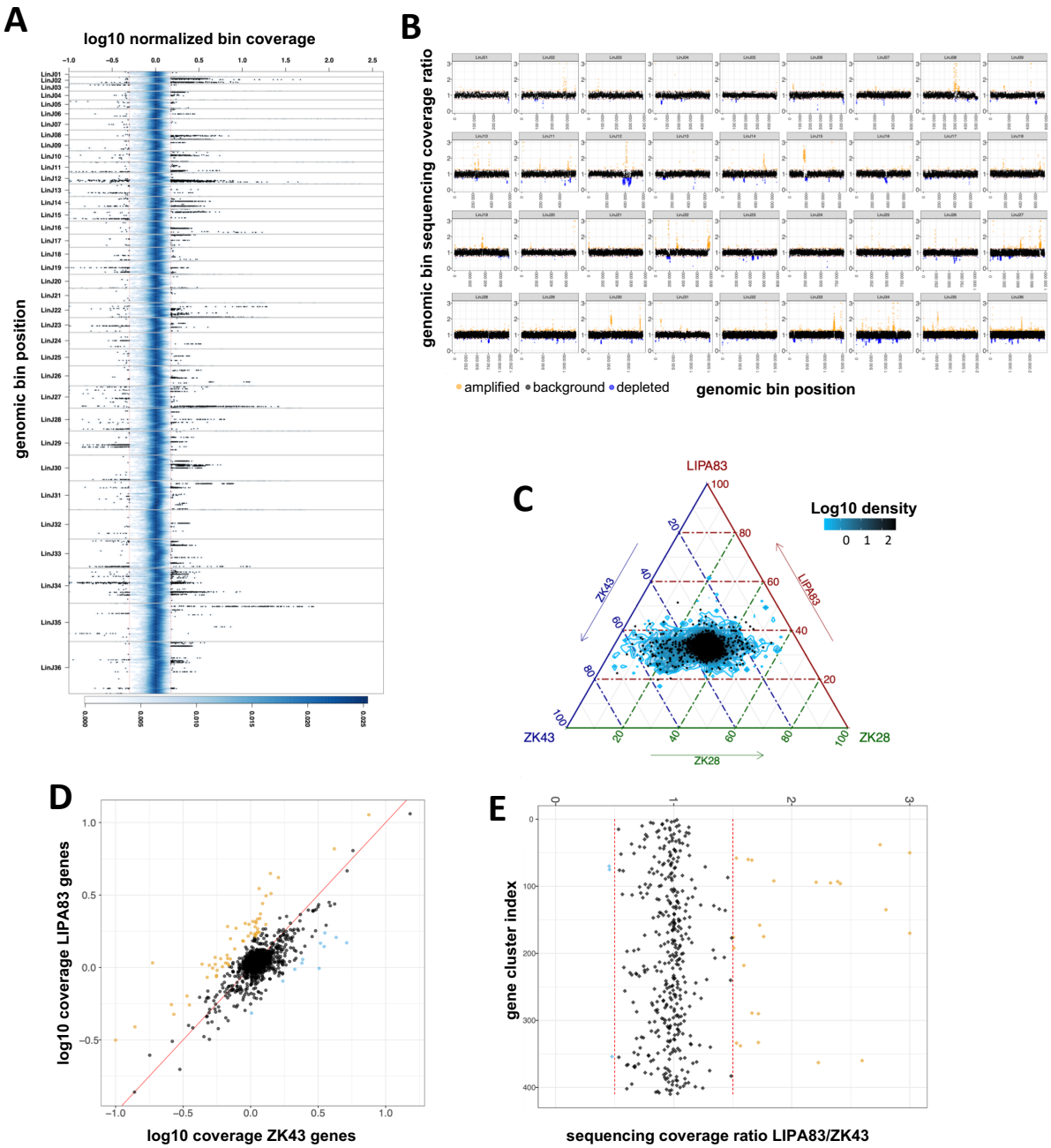
Fig. 1



GIP workflow. The schema on the left recapitulates the GIP inputs, processes and outputs (see methods). Blue, orange, green and purple boxes indicate genome reference, read mapping, quantification and variants computation steps, respectively. The panels on the right demonstrate example plots included in the GIP report computed for individual samples. The "Karyotype" plot shows the coverage distributions for each chromosome (y-axis). The "Genomic bins" plot shows the genomic position (x-axis) and the normalized genomic bin sequencing coverage (y-axis). The "Gene CNVs" plot shows the normalized gene sequencing coverage. The "Structural variants" panel shows a Circos plot representing break-end translocations (black links in the inmost part of the plot), and other possible structural variations in the outer tracks, including insertions, duplications, deletions and inversions. The outmost track shows the normalized sequencing coverage. The "Single nucleotide variants" plot shows on the x and y axes respectively the genomic position and variant allele frequency of detected SNVs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

328 **Fig. 2:**



329

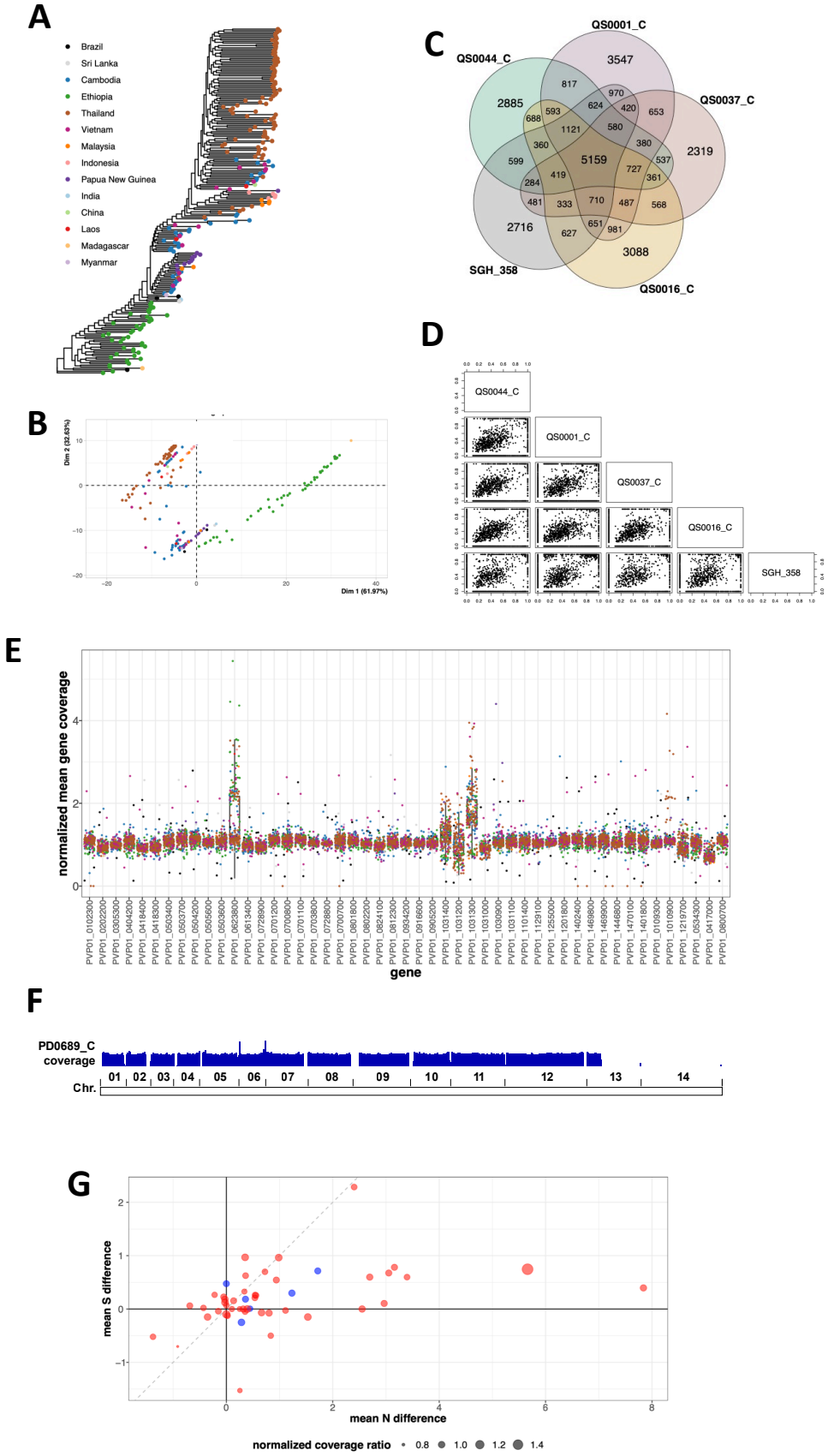
330 Comparing *Leishmania infantum* genomes with giptools. (**A**) Density plot representing the
331 genomic coverage of the seven *Leishmania infantum* isolates. The x-axis shows the log 10
332 normalized coverage of genomic bins. The y-axis reflects the genomic position. The thirty-six
333 different chromosomes are materialized as separate panels. The blue shading indicates the (2D)
334 kernel density estimates of genomic bins. The two red vertical lines mark the 1.5 and 0.5

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

coverage values. A selection of 50 000 bins with coverage > 1.5 or < 0.5 is shown as black dots. (**B**) Scatterplot of the genomic bin normalized sequencing coverage ratio of samples LIPA83 over ZK43. The x any y axes show the ratio score and the genomic position respectively. Ratio scores > 1.25 are labelled in orange and indicate genomic bin amplification. Ratio scores < 0.75 are labelled in blue and indicate genomic bin depletion. (**C**) Ternary comparison showing the relative abundance of the genomic in samples LIPA83, ZK43 and ZK28. The axes report the fraction of the bins normalized sequencing coverage in the three strains. The blue contour indicates the log10 bin density. A subset of 5,000 bins is shown as black dots. Each given point in the plot is adding up to 100. The density area at the center of the plot indicates bins with equal copy number and thus a ~33 distribution across the three axes. (**D**) Scatterplot showing the log10 normalized sequencing coverage of annotated genes in ZK43 (x-axis) and LIPA83 (y-axis). The red line indicates the bisector. Dots represent individual genes. (**E**) Sequencing coverage ratio of gene clusters in samples LIPA83 and ZK43. Dots represent gene clusters. For plots **D** and **E,** the ratio scores > 1.5 or < 0.5 are labelled in orange and blue respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 3:

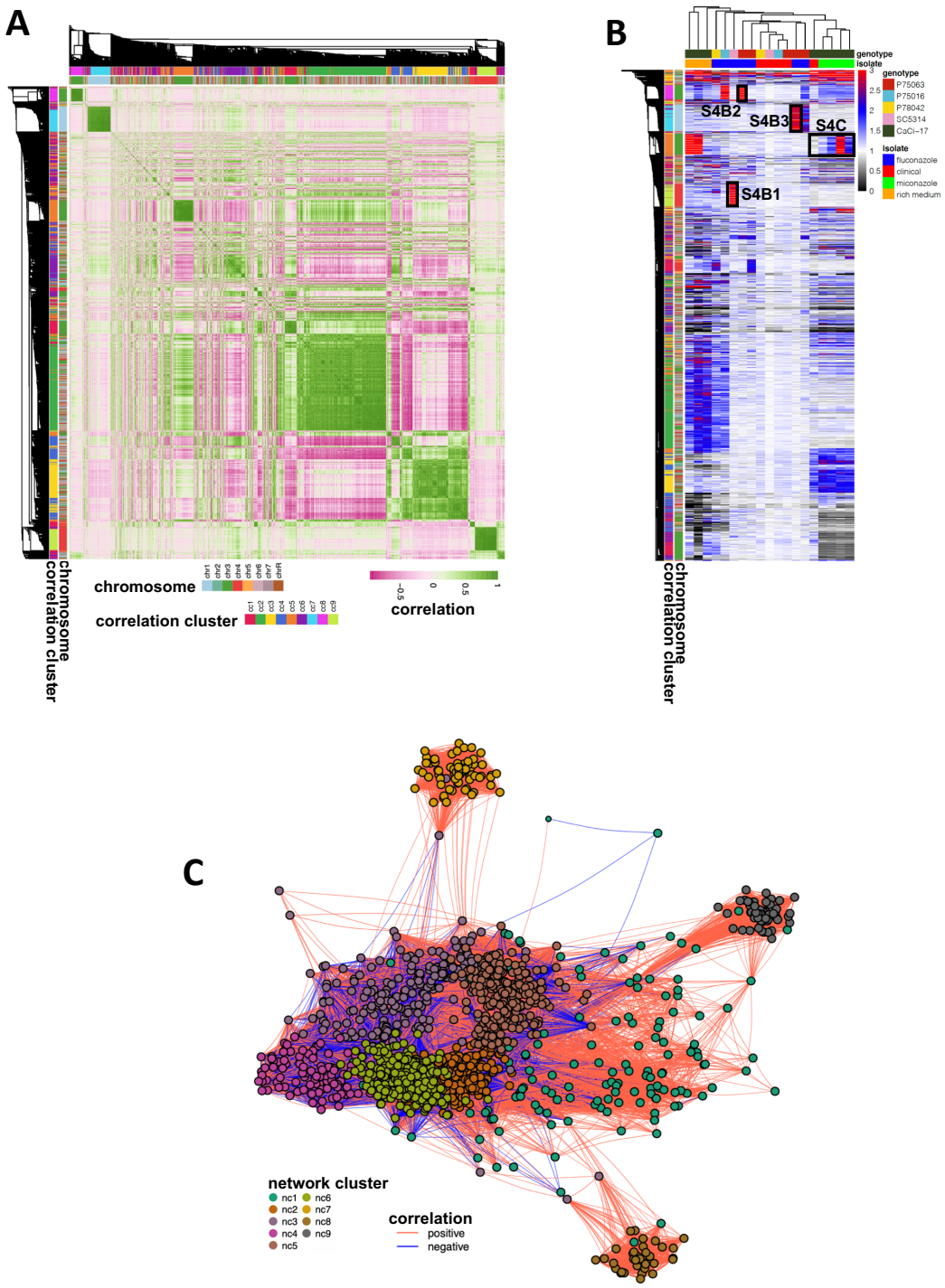


bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

351 *Plasmodium vivax* genomic diversity. (**A**) Predicted maximum likelihood phylogenetic tree
352 reconstruction. (**B**) PCA analysis of the phylogenetic distances estimated from the tree in (**A**).
353 Each dot indicates a sample. The colour code reflects the geographic origin of the samples and
354 matches with the colours of the legend in (**A**). (**C**) Venn diagram comparing the SNVs of five
355 representative Ethiopian strains. (**D**) Pairwise scatterplot comparing the variant allele frequency
356 of all detected SNVs in the five Ethiopian strains. (**E**) Gene panel analysis. The x-axis reports
357 a set of genes of interest. The y-axis indicates the normalized mean gene coverage. The boxplots
358 demonstrate the coverage values distributions for each gene across all samples. Each dot
359 represents the coverage of the indicated gene in a given sample. Dot colours reflect the sample
360 geographic origin as in (**A**). (**F**) Reads per kilo base per million mapped reads (RPKM)
361 normalized sequencing coverage density track of sample PD0689_C. The boundaries of the 14
362 chromosomes are shown on the bottom. (**G**) Comparison of non-synonymous (N) and
363 synonymous (S) mutations between Ethiopia and Cambodia sample groups. Dots represent
364 genes. The x-axis represents the difference between the mean non-synonymous mutation count
365 in the two sample groups. The y-axis represents the difference between the mean synonymous
366 mutation count in the two sample groups. The dot size demonstrates the ratio of the mean
367 normalized sequencing coverage between the two sample groups for each gene. Red and blue
368 dot colors indicate genes belonging to the 43 genes panel (23) and the custom 6 genes panel
369 respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

370 **Fig. 4:**



371

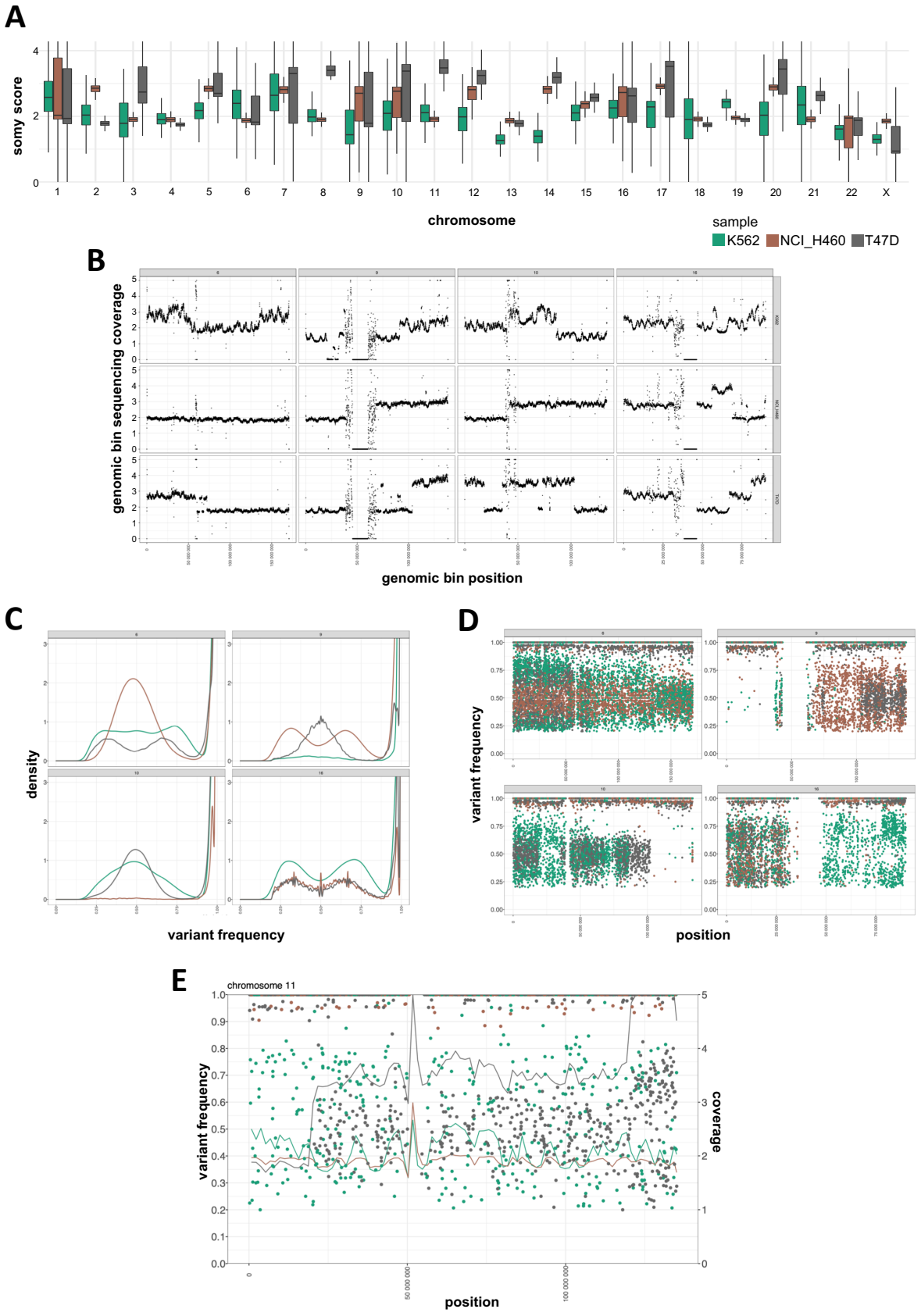
372 Gene CNVs interactions. (**A**) All-vs-all normalized sequencing coverage correlation heatmap.
373 The heatmap is symmetrical along its diagonal and reports both on the rows and the columns
374 the detected gene CNVs. The colour scale indicates with green and pink high levels of positive
375 and negative Pearson correlation, respectively. The side ribbons demonstrate in different

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

colours the chromosome and the correlation cluster of each gene. (**B**) Gene CNV heatmap. The columns and the rows report respectively the samples and the detected gene CNVs. The colour scale indicates the normalized sequencing coverage of the genes. To ease visualization, coverage values greater than 3 are reported as 3 (red). Black boxes highlight the genomic regions shown in **Fig. S3B** (panels 1, 2 and 3) and **Fig. S3C**. The ribbons on the left indicate the chromosome and the correlation cluster of each gene. Top ribbons indicate the genotype and the strains resulting from the different evolutionary experiments. (**C**) Gene interaction network. Nodes indicate gene CNVs. Edges reflect the absolute Pearson correlation value. The closer the nodes are, the higher is the correlation. Only significant interactions (Benjamini-Hochberg adjusted p-value < 0.01) are shown. The colour of the edges indicates in red and blue respectively positive and negative correlations. The colour of the nodes denotes the predicted network cluster for each gene.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

388 **Fig. 5**:



389

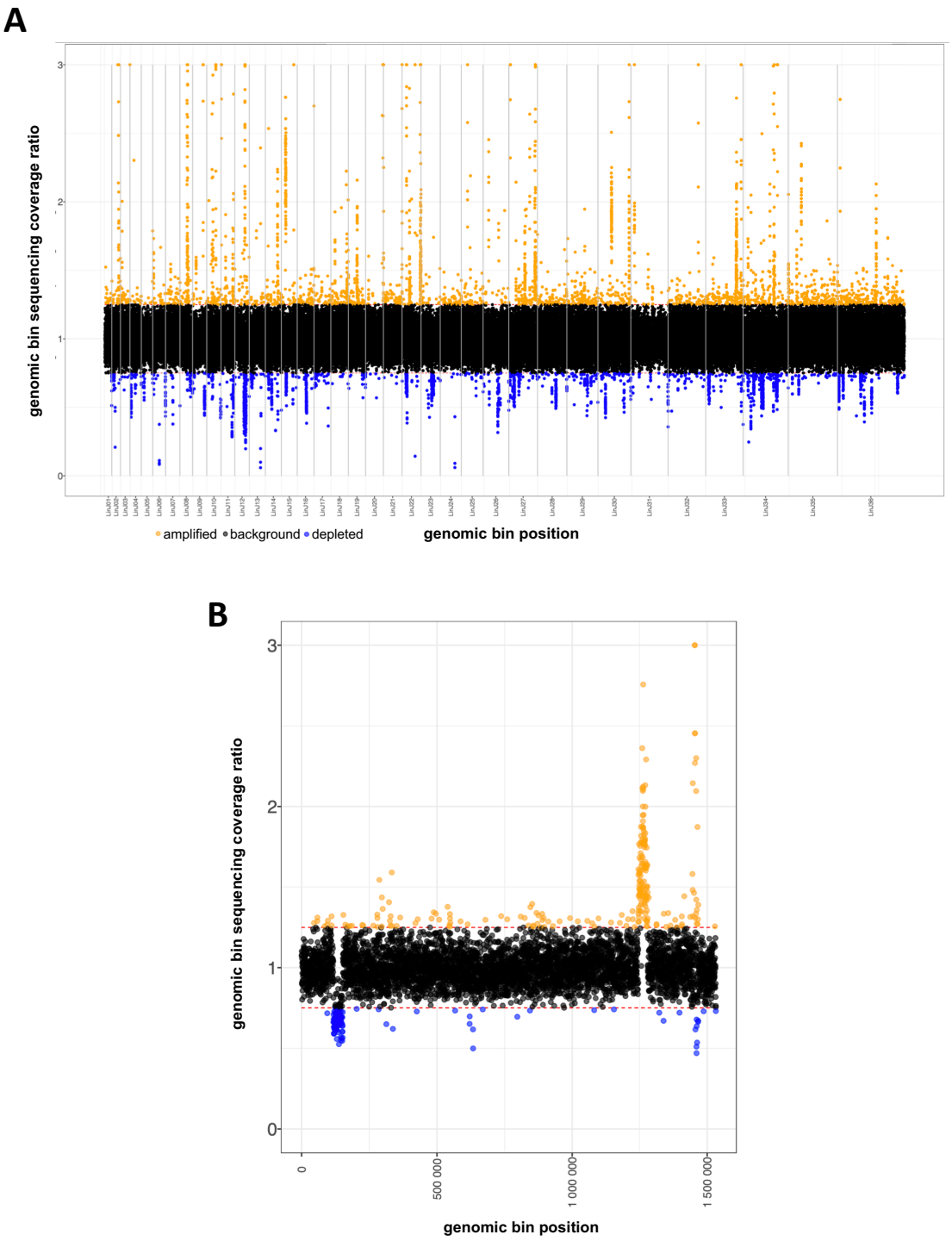
bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cancer cell lines genome instability. Green, brown, and grey colours indicate respectively samples K562, NCI_H460 and T47D. (**A**) Chromosome coverage analysis. The x-axis reports the chromosomes. The y-axis reports the estimated somy score. The boxes show the somy score distributions. (**B**) Sub-chromosomal copy number variation. Dots indicate genomic bins. Different panels indicate different chromosomes. The panel columns indicate from left to right four selected chromosomes: 6, 9, 10 and 16. The panel rows show top to bottom the samples K562, NCI_H460 and T47D. The x-axis indicates the genomic position. The y-axis indicates the normalized genomic bin sequencing coverage values. Coverage values greater than 5 are reported as 5. (**C**) SNV frequency density plots. The four different panels represent different selected chromosomes: 6, 9, 10 and 16. The x-axis reports the variant allele frequency. The y-axis the estimated kernel density between 0 and 3. (**D**) SNV frequency scatter plots. The four different panels represent different selected chromosomes: 6, 9, 10 and 16. The x-axis indicates the genomic position. The y-axis indicates the variant allele frequency. (**E**) Chromosome 11 combined SNV and bin coverage plot. To ease visualization, giptools allows the simultaneous displaying of variant allele frequencies (y-axis, left) and sequencing coverage (y-axis, right). Dots represent SNVs. The lines represent the normalized bin sequencing coverage. The x-axis indicates the genomic position. Coverage values greater than 5 are shown as 5.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary figures

Fig. S1:



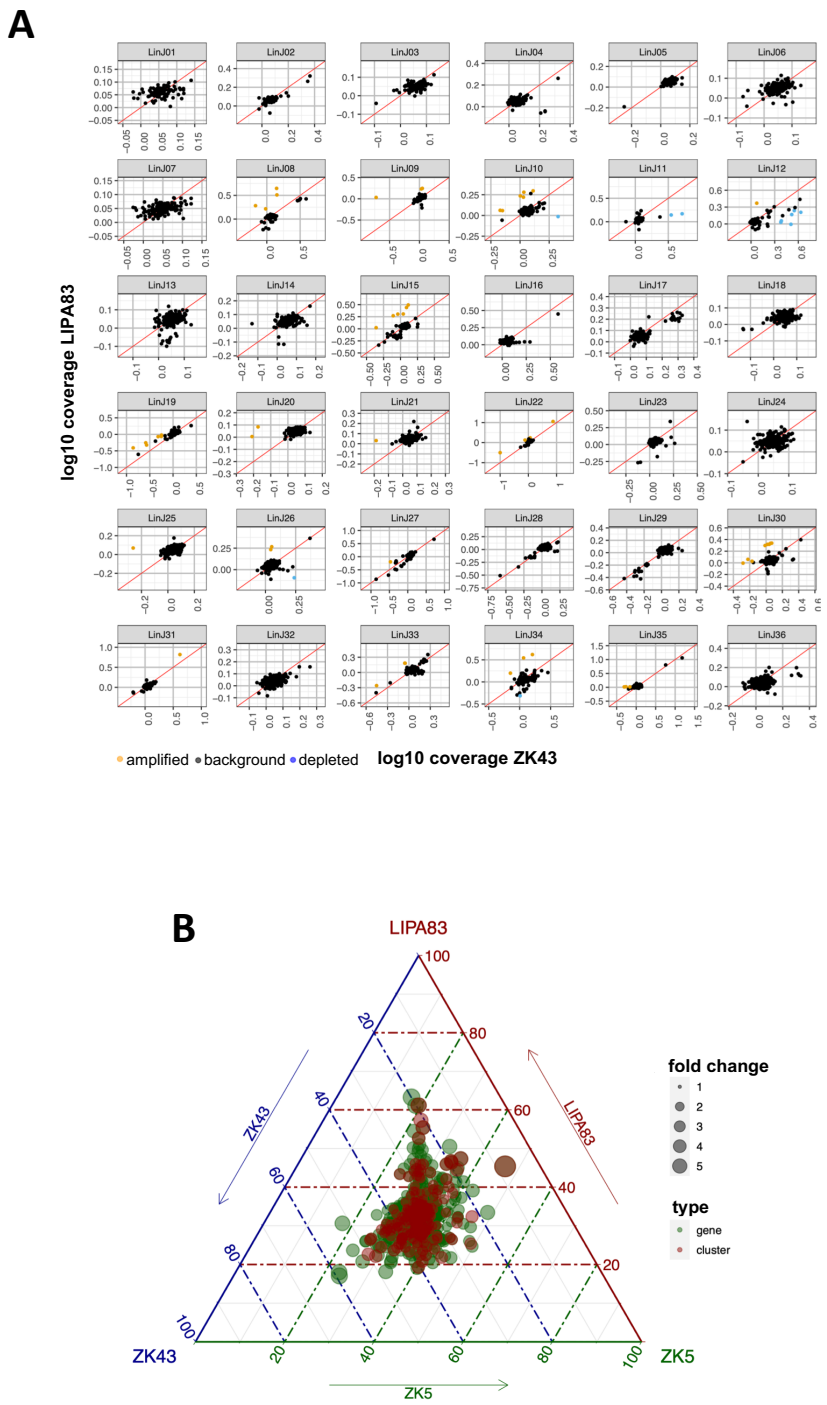
Bin coverage ratio visualizations. The figure shows alternative representations produced for the genomic bin normalized sequencing coverage ratio of samples ZK43 and LIPA83. Same layout

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as **Fig. 2B**. (**A**) Whole genome overview. (**B**) All the individual chromosomes separately. The plot shows the example of chromosome 33.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

415 **Fig. S2**:



416

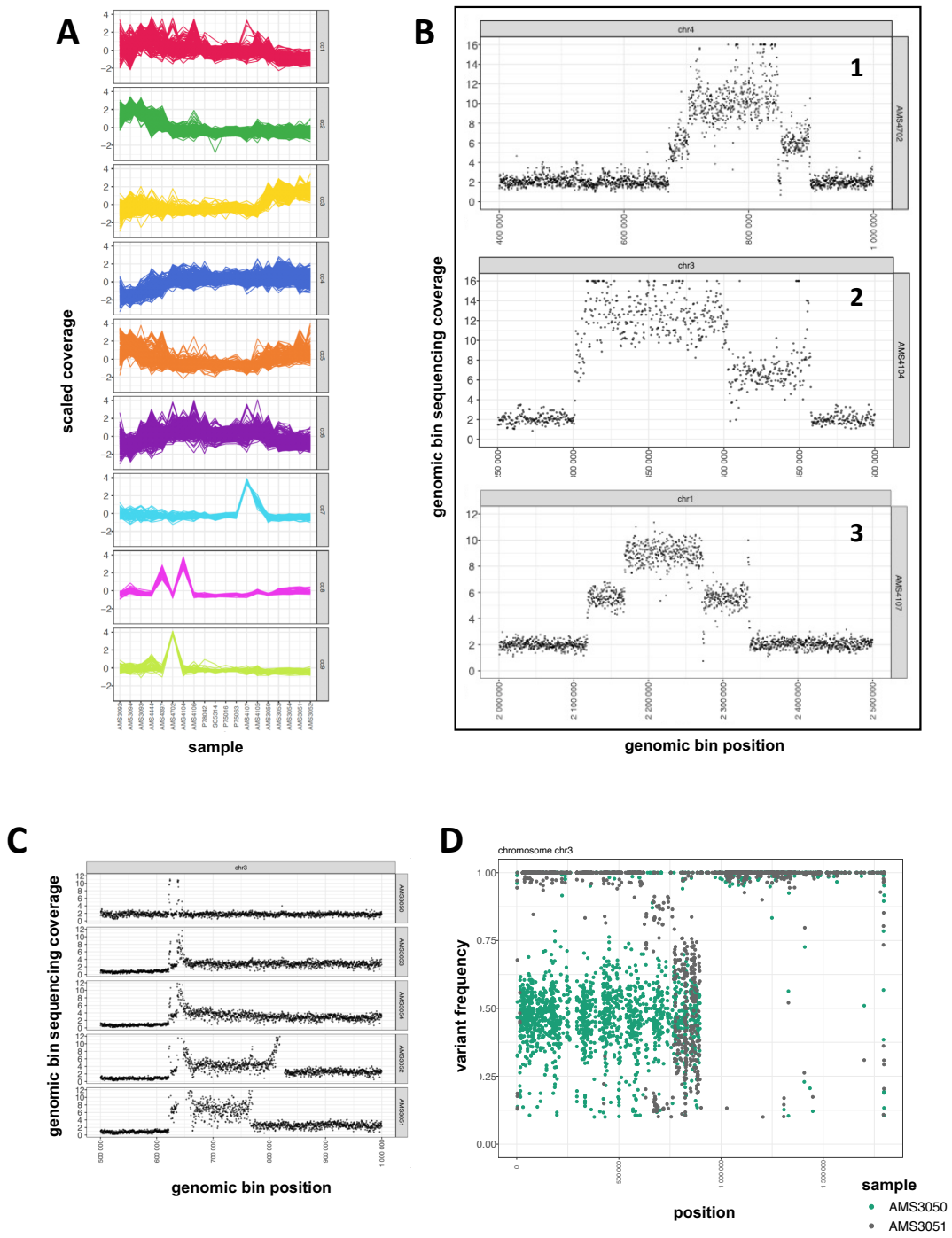
417 Two-ways and three-ways gene coverage comparisons. (**A**) Log10 normalized sequencing
418 coverage comparison of genes in ZK43 (x-axis) and LIPA83 (y-axis). Same layout a **Fig. 2D**
419 but representing the different chromosomes in individual panels. (**B**) Ternary plot showing the
420 relative abundance of genes (green dots) and gene clusters (red dots) in samples LIPA83, ZK43

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

421 and ZK28. The axes report the fraction of the genes and genes clusters normalized sequencing
422 coverage in the three strains. Each given point in the plot adding up to 100. Genes with equal
423 copy number are shown in the center of the ternary plot, while copy number variations are
424 visualized by shifts of the dots out of the center. The dots size reflects the fold change variation
425 between the maximum and the minimum observed normalized coverage value.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

426 **Fig. S3:**



427

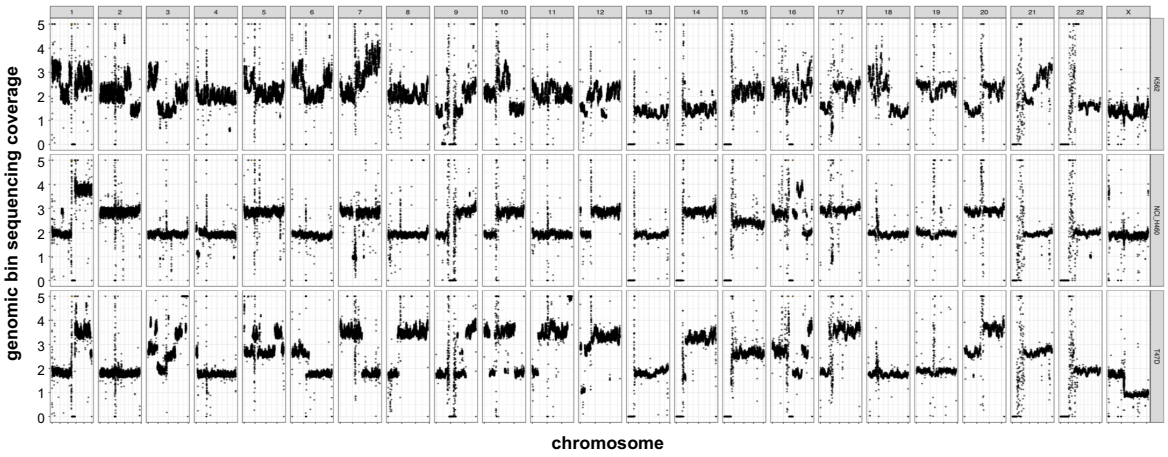
428 Copy number variations of evolutionary adapted *Candida albicans* strains. (**A**) Gene CNV
429 correlation clusters. The x and y axes indicate respectively the samples and the scaled
430 normalized sequencing coverage of the genes in each cluster. Different panels indicate different
431 correlation clusters. (**B**) Scatterplots showing the bin normalized sequencing coverage (y-axis)

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of three genomic regions (chr4:400,000-1,000,000; chr3:1,250,000-1,500,000; chr1:2,000,000-2,500,000) in three different samples AMS4702, AMS4104 and AMS4107. The x-axis shows the genomic position. Dots represent genomic bins. (**C**) Scatterplot showing the bin normalized sequencing coverage of a genomic region (chr3:500,000-1,000,000) in five different samples represented as separate panels. From top to bottom the samples are: AMS3050, AMS3053, AMS3054, AMS3052 and AMS3051. The x and y axes show respectively the genomic position and the normalized sequencing coverage. Dots represent bins. (**D**) Comparative analysis of chromosome 3 SNVs. AMS3050 SNVs are shown in green, while AMS3051 SNVs are shown in grey. The x and the y axes show respectively the genomic position and the variant allele frequency.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

442 **Fig. S4:**

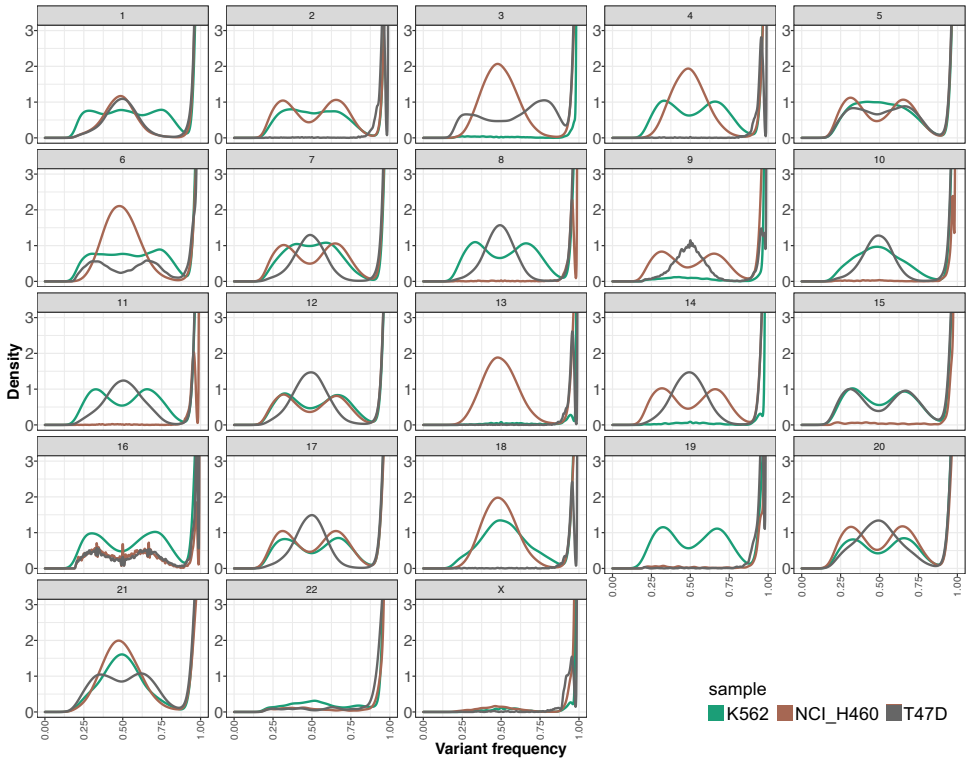


443

444 Genomic coverage analysis of three cancer cell lines. The x-axis indicates the genomic position.
445 The y-axis indicates the normalized genomic bin sequencing coverage values. Dots demonstrate
446 the genomic bins. To ease visualization, coverage values greater than 5 are reported as 5.
447 Different panels show different chromosomes. The three panel rows indicate top to bottom the
448 following samples: K562, NCI_H460 and T47D.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

449 **Fig. S5:**



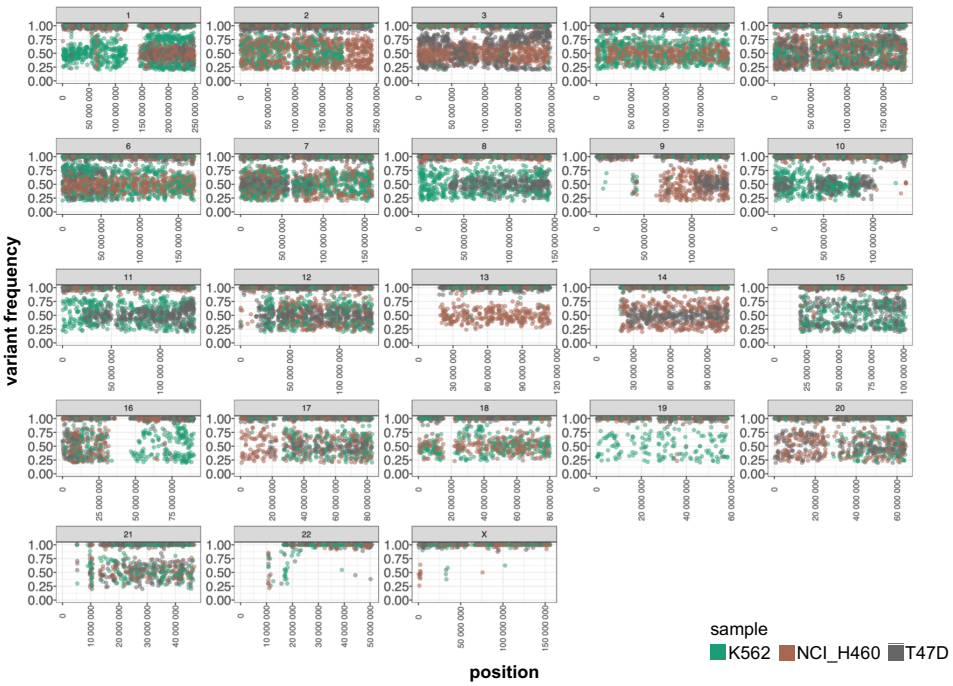
450

451 SNV frequency density plot. The x-axis shows the variant allele frequency. The y-axis the
452 estimated kernel density between 0 and 3. Green, brown, and grey slopes indicate respectively
453 samples K562, NCI_H460 and T47D.

454

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

455 **Fig. S6:**



456

457 SNV heterogeneity of cancer cell lines. The x-axis indicates the genomic position. The y-axis
458 indicates the variant allele frequency. The dots indicate SNVs and they are coloured according
459 to the sample. K562, green; NCI_H460, brown; T47D, grey. Different panels show different
460 chromosomes.

461

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods:

GIP and giptools. All results presented in this study were generated using GIP and giptools version 1.0.9. GIP code is maintained and freely distributed at the github page: https://github.com/giovannibussotti/GIP. giptools container is accessible from the Singularity cloud at https://cloud.sylabs.io/library/giovannibussotti/default/giptools. The GIP configuration files (**Supplementary Data 2**) and the giptools command options used to generate all results (**Supplementary Data 3**) are provided. The full documentation of GIP and giptools including a description of all options is available from https://gip.readthedocs.io/en/latest/.

Read alignment. WGS reads were downloaded from the Sequence Read Archive (SRA) (58) and the European Nucleotide Archive (ENA) (59) repositories and the Encyclopedia of DNA Elements (ENCODE) dashboard (60) (**Table S3**). For *Leishmania infantum* the GCA_900500625 genome reference and gene annotations available from the ENSEMBL protists (61) server (release-48) were used. For *Candida albicans* the assembly 21 of the SC5314 strain genome reference and gene annotations available from the *Candida* Genome Database (CGD) (62) were used. For *Plasmodium vivax* the P01 reference genome and gene annotations available from PlasmoDB (63) (release-50) were used. For the cancer cell lines the human genome GRCh38 primary assembly and gene annotations available from ENSEMBL (release-102) were used. The repetitive elements of reference genomes were soft-masked by GIP using Red (64). WGS reads were mapped by GIP using BWA-mem (version 0.7.17) (65,66) run with option -M to label shorter split hits as secondary. Then the alignment files were sorted, indexed and reformatted by GIP using Samtools (version 1.8) (67). Finally, read duplicates were removed by GIP using Picard MarkDuplicates (http://broadinstitute.github.io/picard) (version 2.18.9) with the option

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

"VALIDATION_STRINGENCY=LENIENT." In the four considered datasets, WGS reads were aligned against full assemblies, including unsorted contigs if present. However just the canonical assembled chromosomes were considered for all downstream analyses (`chrs` option, **Supplementary Data 2**). A minimum read alignment MAPQ score was adopted to select genes for cluster analysis, and to call for SNVs and SVs (`MAPQ` option, **Supplementary Data 2**). Altogether a total of 6,306,951,266 reads were aligned to the respective reference genomes. The `giptools overview` module was run to gather the alignment statistics as estimated by Picard CollectAlignmentSummaryMetrics (**Table S8**).

Genomic bins and genes quantification. GIP was used to evaluate the mean sequencing coverage and the mean read MAPQ of genomic bins and genes. For genomic bins, GIP partitioned the input genomes into adjacent windows of user defined lengths (`binSize` option, **Supplementary Data 2**). The coverage GC-content score bias was corrected (`CGcorrect` option, **Supplementary Data 2**) fitting a LOESS regression with a 5-fold cross validation to optimize the model span parameter. A larger window length was utilized to bin the reference genomes for Circos plot representations (`binSizeCircos` option, **Supplementary Data 2**). In **Fig. 1** ("Genomic bins" and "Gene CNVs" plots), **Fig. 2**, **Fig. S1**, **Fig. S2** and **Fig. 4** bins and genes coverage scores were normalized by median chromosome coverage to highlight amplifications or depletions with respect to the chromosome copy number. In **Fig. 1** ("Structural variants" plot), **Fig. S3B-C**, **Fig. 3E-G**, and **Fig. S4A** bins and genes coverage scores were normalized by median genome coverage to account for sequencing library size differences. GIP evaluated statistically significant copy number variant bins and genes (**Fig. 1** "Genomic bins" and "Gene CNVs" plots) using a p-value threshold of 0.001 (`covPerBinSigOPT` and `covPerGeSigOPT` options, **Supplementary Data 2**). Estimated p-values for bins and genes CNVs were corrected for multiple testing using the Benjamini – Yekutieli (`--padjust BY `) and the Benjamini – Hochberg (`--padjust BH`) methods. The somy

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

scores shown in **Fig. 1** (karyotype plot) and **Fig. S4B** were computed multiplying the median genome coverage normalized bin coverage by 2. GIP enabled the CNV analysis of genes sharing high sequence identity by clustering the nucleotide sequences of the genes with low mean MAPQ score into groups with cd-hit-est (version 4.8.1) (68) with options `-s 0.9 -c 0.9 -r 0 -d 0 -g 1`. Then for each gene cluster GIP computed the mean gene coverage normalized by median chromosome coverage (**Fig. 2E**, **Fig S2B**).

Gene ontology and metabolic pathway enrichment. The FungiDB online tool (Release 52, 20 May 2021) (69) was used to evaluate the functional enrichment of network clusters genes. For the gene ontology analysis, the biological process (BP), molecular function (MF) and cellular compartment (CC) terms enrichments were tested, considering both computed and curated evidences and a p-value cutoff of 0.05. For the metabolic pathway enrichment, both KEGG (70) and MetaCyc (71) pathway sources were considered with a p-value cutoff of 0.05. Terms and pathways with Benjamini – Hochberg adjusted p-values < 0.05 were considered statistically significant.

Sequencing coverage density estimates. GIP was used to convert the read alignment files (.bam format) in binary data files reflecting sequencing coverage (.bigWig format). The coverage file were produced using bamCoverage from the deepTools2 suite (72) (version 3.5.1) with options "--normalizeUsing RPKM --ignoreDuplicates --binSize 10 --smoothLength 30" (`bigWigOPT` option, **Supplementary Data 2**). The coverage track of sample PD0689_C was visualized with IGV using the `Bar Chart`, `Autoscale` and windowing function `Mean` options.

Single-nucleotide variant analysis. GIP was used to call SNVs using Freebayes (version 1.3.2) (`freebayesOPT` option, **Supplementary Data 2**) and filter its output (`filterFreebayesOPT` option, **Supplementary Data 2**). Filters included the minimum allele

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

frequency (`--minFreq`), the minimum number of reads supporting the alternative alleles (`--minAO`) and minimum mean mapping quality of for the reads supporting the reference (`--minMQMR`) or the alternative allele (`--minMQM`). A higher number of reads supporting the variants was requested for predictions positioned inside simple repeats of the same nucleotide (homopolymers) (`--minAOhomopolymer`). The homopolymers were defined as the DNA region spanning ±5 bases from the SNV (`--contextSpan 5`), with over 40% of identical nucleotides (--homopolymerFreq 0.4 `). Further, GIP discarded SNVs with sequencing coverage above or below 4 median absolute deviations (MADs) from the median chromosome coverage (`--MADrange`). snpEff (version 4.3t) (73) was used to predict the impact of SNVs on coding sequence. The predicted effects that GIP considered synonymous mutations are: “synonymous_variant", "stop_retained_variant" and "start_retained". The predicted effects that GIP considered non-synonymous mutations are: "missense_variant", "start_lost", "stop_gained", "stop_lost" and "coding_sequence_variant". The phylogenetic tree was computed by the giptools module `phylogeny` using IQtree2 (version 2.1.2) (74,75) with options `--seqtype DNA --alrt 1000 -B 1000`. The Venn-diagram comparison considered the strains QS0044_C, QS0001_C, QS0037_C, QS0016_C and SGH_358 that were sampled from different locations in Ethiopia, respectively Habala, Badowacho, Arbaminch, Hawassa and Jimma. The strains were selected to have comparable average genome coverage (23). To infer the tree GIP considered the set of filtered SNV and adopted the IUPAC ambiguous notation for the positions with allele frequency below 70%. The tree was visualized by giptools using the ggtree R-package (76).

Analysis of structural variants. GIP was used to detect structural variants including insertions, tandem duplications, deletions, inversions and break-end translocations with DELLY (version 0.8.7) (11). To reduce incorrect predictions the DELLY output was additionally filtered (`filterDellyOPT` option, **Supplementary Data 2**). GIP discarded poor predictions with

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

560 DELLY label "LowQual" (`--rmLowQual`) and low median MAPQ score of mapping reads (`-
561 -minMAPQ`). SVs positioned in proximity of chromosome ends were removed (--
562 chrEndFilter) to limit false predictions caused by potential misassembled regions close to the
563 telomeric ends. To ease visualization and limit the analysis only to best supported SVs GIP
564 limited the output only to the top predictions (`--topHqPercentIns`, `--topHqPercentDel`, `--
565 topHqPercentDup` and `--topHqPercentInv`) based on the SV support score as in **Formula 1**,
566 where *DV*, *DR*, *RV*, and *RR* are respectively the number of high-quality variant pairs, reference
567 pairs, variant junction reads, and reference junction reads.

568 $$\frac{\boldsymbol{DV + RV}}{\boldsymbol{DV + RV + DR + RR}} * \mathbf{100}$$

569 **Formula 1:** SV support score.

570 The predicted structural variants were represented with Circos (version 0.69-9) (77).

571 **Data and materials availability.** All data is available in the main text or the supplementary
572 materials.

573

574 ## Acknowledgments:

575 **Funding:** This study was supported by a seeding grant from the Institut Pasteur International
576 Department to the LeiSHield Consortium and the EU H2020 project LeiSHield-MATI - REP-
577 778298-1.

578 **Author contributions:** Conceptualization, G.B. and G.F.S.; Methodology, G.B. and G.F.S.;
579 Software, G.B.; Formal Analysis, G.B., and G.F.S.; Investigation, G.B. and G.F.S.; Writing –
580 Original Draft, G.B. and G.F.S.; Writing – Review & Editing, G.B. and G.F.S.; Supervision,
581 G.B. and G.F.S.; Funding Acquisition, G.F.S. and G.B.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

582 **Competing interests:** Authors declare no competing interests.

583

584 ## Supplementary Tables:

585 **Table S1:** Genomic bin coverage ratio of sample LIPA83 over ZK43.

586 **Table S2:** Gene coverage ratio of sample LIPA83 over ZK43.

587 **Table S3:** Sample information.

588 **Table S4:** *Candida albicans* gene CNVs.

589 **Table S5:** Gene correlation clusters in *Candida albicans*.

590 **Table S6:** Network analysis of gene CNVs in *Candida albicans*.

591 **Table S7:** Gene ontology and metabolic pathway enrichment analyses.

592 *Footnote: Terms and pathways with Benjamini – Hochberg adjusted p-values < 0.05 are*

593 *labeled in red.*

594 **Table S8:** Mapping statistics.

595

596 ## Supplementary Data:

597 **Supplementary Data 1:** Cancer cell lines SNVs.

598 **Supplementary Data 2:** GIP configuration files.

599 **Supplementary Data 3:** giptools commands.

600

601 ## References:

602 1. Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock,
603 G.M., Isaacs, F., Rozowsky, J. *et al.* (2016) The real cost of sequencing: scaling
604 computation to keep pace with data generation. *Genome Biol*, **17**, 53.
605 2. Pareek, C.S., Smoczynski, R. and Tretyn, A. (2011) Sequencing technologies and
606 genome sequencing. *J Appl Genet*, **52**, 413-435.
607 3. Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum,
608 A., Copeland, A., Huddleston, J., Eichler, E.E. *et al.* (2013) Nonhybrid, finished

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*, **10**, 563-569.

4. Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*, **13**, 1050-1054.
5. Reuter, J.A., Spacek, D.V. and Snyder, M.P. (2015) High-throughput sequencing technologies. *Mol Cell*, **58**, 586-597.
6. Aurrecoechea, C., Barreto, A., Basenko, E.Y., Brestelli, J., Brunk, B.P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S. *et al.* (2017) EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic Acids Res*, **45**, D581-D591.
7. Dolled-Filhart, M.P., Lee, M., Jr., Ou-Yang, C.W., Haraksingh, R.R. and Lin, J.C. (2013) Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. *ScientificWorldJournal*, **2013**, 730210.
8. Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J. and Trajanoski, Z. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*, **15**, 256-278.
9. Garrison, E. and Marth, G. (2012), *arXiv preprint arXiv:1207.3907 [q-bio.GN]*.
10. Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*, **21**, 974-984.
11. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V. and Korbel, J.O. (2012), *Bioinformatics*, Vol. 28, pp. i333-i339.
12. Hwang, K.B., Lee, I.H., Li, H., Won, D.G., Hernandez-Ferrer, C., Negron, J.A. and Kong, S.W. (2019) Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. *Sci Rep*, **9**, 3219.
13. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. and Notredame, C. (2017) Nextflow enables reproducible computational workflows. *Nat Biotechnol*, **35**, 316-319.
14. Bussotti, G., Benkahla, A., Jeddi, F., Souiai, O., Aoun, K., Spath, G.F. and Bouratbine, A. (2020) Nuclear and mitochondrial genome sequencing of North-African Leishmania infantum isolates from cured and relapsed visceral leishmaniasis patients reveals variations correlating with geography and phenotype. *Microb Genom*, **6**.
15. Bussotti, G., Gouzelou, E., Cortes Boite, M., Kherachi, I., Harrat, Z., Eddaikra, N., Mottram, J.C., Antoniou, M., Christodoulou, V., Bali, A. *et al.* (2018) Leishmania Genome Dynamics during Environmental Adaptation Reveal Strain-Specific Differences in Gene Copy Number Variation, Karyotype Instability, and Telomeric Amplification. *MBio*, **9**.
16. Prieto Barja, P., Pescher, P., bussotti, g., Dumetz, F., Imamura, H., Kedra, D., Domagalska, M., Chaumeau, V., Himmelbauer, H., Pages, M. *et al.* (2017), *Nat Ecol Evol*, Vol. 1, pp. 1961-1969.
17. Alvar, J., Velez, I.D., Bern, C., Herrero, M., Desjeux, P., Cano, J., Jannin, J., den Boer, M. and Team, W.H.O.L.C. (2012) Leishmaniasis worldwide and global estimates of its incidence. *PLoS One*, **7**, e35671.
18. Organization, W.H. (2018), WHO, Geneva. .
19. Price, R.N., Tjitra, E., Guerra, C.A., Yeung, S., White, N.J. and Anstey, N.M. (2007) Vivax malaria: neglected and not benign. *Am J Trop Med Hyg*, **77**, 79-87.
20. Gething, P.W., Elyazar, I.R., Moyes, C.L., Smith, D.L., Battle, K.E., Guerra, C.A., Patil, A.P., Tatem, A.J., Howes, R.E., Myers, M.F. *et al.* (2012) A long neglected world malaria map: Plasmodium vivax endemicity in 2010. *PLoS Negl Trop Dis*, **6**, e1814.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

21. Battle, K.E., Gething, P.W., Elyazar, I.R., Moyes, C.L., Sinka, M.E., Howes, R.E., Guerra, C.A., Price, R.N., Baird, K.J. and Hay, S.I. (2012) The global public health significance of Plasmodium vivax. *Adv Parasitol*, **80**, 1-111.
22. Pearson, R.D., Amato, R., Auburn, S., Miotto, O., Almagro-Garcia, J., Amaratunga, C., Suon, S., Mao, S., Noviyanti, R., Trimarsanto, H. *et al.* (2016) Genomic analysis of local variation and recent evolution in Plasmodium vivax. *Nat Genet*, **48**, 959-964.
23. Ford, A., Kepple, D., Abagero, B.R., Connors, J., Pearson, R., Auburn, S., Getachew, S., Ford, C., Gunalan, K., Miller, L.H. *et al.* (2020) Whole genome sequencing of Plasmodium vivax isolates reveals frequent sequence and structural polymorphisms in erythrocyte binding genes. *PLoS Negl Trop Dis*, **14**, e0008234.
24. Singh, V., Gupta, P. and Pande, V. (2014) Revisiting the multigene families: Plasmodium var and vir genes. *J Vector Borne Dis*, **51**, 75-81.
25. Rayner, J.C., Tran, T.M., Corredor, V., Huber, C.S., Barnwell, J.W. and Galinski, M.R. (2005) Dramatic difference in diversity between Plasmodium falciparum and Plasmodium vivax reticulocyte binding-like genes. *Am J Trop Med Hyg*, **72**, 666-674.
26. Rahul, C.N., Shiva Krishna, K., Pawar, A.P., Bai, M., Kumar, V., Phadke, S. and Rajesh, V. (2014) Genetic and structural characterization of PvSERA4: potential implication as therapeutic target for Plasmodium vivax malaria. *J Biomol Struct Dyn*, **32**, 580-590.
27. Rahul, C.N., Shiva Krishna, K., Meera, M., Phadke, S. and Rajesh, V. (2015) Plasmodium vivax: N-terminal diversity in the blood stage SERA genes from Indian isolates. *Blood Cells Mol Dis*, **55**, 30-35.
28. Luo, Z., Sullivan, S.A. and Carlton, J.M. (2015) The biology of Plasmodium vivax explored through genomics. *Ann N Y Acad Sci*, **1342**, 53-61.
29. Lin, J.T., Patel, J.C., Kharabora, O., Sattabongkot, J., Muth, S., Ubalee, R., Schuster, A.L., Rogers, W.O., Wongsrichanalai, C. and Juliano, J.J. (2013) Plasmodium vivax isolates from Cambodia and Thailand show high genetic complexity and distinct patterns of P. vivax multidrug resistance gene 1 (pvmdr1) polymorphisms. *Am J Trop Med Hyg*, **88**, 1116-1123.
30. Gunalan, K., Niangaly, A., Thera, M.A., Doumbo, O.K. and Miller, L.H. (2018) Plasmodium vivax Infections of Duffy-Negative Erythrocytes: Historically Undetected or a Recent Adaptation? *Trends Parasitol*, **34**, 420-429.
31. Costa, G.L., Amaral, L.C., Fontes, C.J.F., Carvalho, L.H., de Brito, C.F.A. and de Sousa, T.N. (2017) Assessment of copy number variation in genes related to drug resistance in Plasmodium vivax and Plasmodium falciparum isolates from the Brazilian Amazon and a systematic review of the literature. *Malar J*, **16**, 152.
32. Cornejo, O.E., Fisher, D. and Escalante, A.A. (2014) Genome-wide patterns of genetic polymorphism and signatures of selection in Plasmodium vivax. *Genome Biol Evol*, **7**, 106-119.
33. Chen, E., Salinas, N.D., Huang, Y., Ntumngia, F., Plasencia, M.D., Gross, M.L., Adams, J.H. and Tolia, N.H. (2016) Broadly neutralizing epitopes in the Plasmodium vivax vaccine candidate Duffy Binding Protein. *Proc Natl Acad Sci U S A*, **113**, 6277-6282.
34. Zolan, M.E. (1995) Chromosome-length polymorphism in fungi. *Microbiol Rev*, **59**, 686-698.
35. Suzuki, T., Nishibayashi, S., Kuroiwa, T., Kanbe, T. and Tanaka, K. (1982) Variance of ploidy in Candida albicans. *J Bacteriol*, **152**, 893-896.
36. Sionov, E., Chang, Y.C. and Kwon-Chung, K.J. (2013) Azole heteroresistance in Cryptococcus neoformans: emergence of resistant clones with chromosomal disomy in

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the mouse brain during fluconazole treatment. *Antimicrob Agents Chemother*, **57**, 5127-5130.

37. Shin, J.H., Chae, M.J., Song, J.W., Jung, S.I., Cho, D., Kee, S.J., Kim, S.H., Shin, M.G., Suh, S.P. and Ryang, D.W. (2007) Changes in karyotype and azole susceptibility of sequential bloodstream isolates from patients with Candida glabrata candidemia. *J Clin Microbiol*, **45**, 2385-2391.
38. Selmecki, A., Forche, A. and Berman, J. (2010) Genomic plasticity of the human fungal pathogen Candida albicans. *Eukaryot Cell*, **9**, 991-1008.
39. Magee, B.B. and Magee, P.T. (2000) Induction of mating in Candida albicans by construction of MTLa and MTLalpha strains. *Science*, **289**, 310-313.
40. Gerstein, A.C., Fu, M.S., Mukaremera, L., Li, Z., Ormerod, K.L., Fraser, J.A., Berman, J. and Nielsen, K. (2015) Polyploid titan cells produce haploid and aneuploid progeny to promote stress adaptation. *mBio*, **6**, e01340-01315.
41. Croll, D. and McDonald, B.A. (2012) The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog*, **8**, e1002608.
42. Chibana, H., Beckerman, J.L. and Magee, P.T. (2000) Fine-resolution physical mapping of genomic diversity in Candida albicans. *Genome Res*, **10**, 1865-1877.
43. Bravo Ruiz, G., Ross, Z.K., Holmes, E., Schelenz, S., Gow, N.A.R. and Lorenz, A. (2019) Rapid and extensive karyotype diversification in haploid clinical Candida auris isolates. *Curr Genet*, **65**, 1217-1228.
44. Brown, G.D. and Netea, M.G. (2012) Exciting developments in the immunology of fungal infections. *Cell Host Microbe*, **11**, 422-424.
45. Pfaller, M.A., Diekema, D.J., Turnidge, J.D., Castanheira, M. and Jones, R.N. (2019) Twenty Years of the SENTRY Antifungal Surveillance Program: Results for Candida Species From 1997-2016. *Open Forum Infect Dis*, **6**, S79-S94.
46. Pfaller, M.A., Castanheira, M., Messer, S.A., Moet, G.J. and Jones, R.N. (2010) Variation in Candida spp. distribution and antifungal resistance rates among bloodstream infection isolates by patient age: report from the SENTRY Antimicrobial Surveillance Program (2008-2009). *Diagn Microbiol Infect Dis*, **68**, 278-283.
47. Todd, R.T. and Selmecki, A. (2020) Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs. *Elife*, **9**.
48. Mount, H.O., Revie, N.M., Todd, R.T., Anstett, K., Collins, C., Costanzo, M., Boone, C., Robbins, N., Selmecki, A. and Cowen, L.E. (2018) Global analysis of genetic circuitry and adaptive mechanisms enabling resistance to the azole antifungal drugs. *PLoS Genet*, **14**, e1007319.
49. Hirakawa, M.P., Martinez, D.A., Sakthikumar, S., Anderson, M.Z., Berlin, A., Gujja, S., Zeng, Q., Zisson, E., Wang, J.M., Greenberg, J.M. *et al.* (2015) Genetic and phenotypic intra-species variation in Candida albicans. *Genome Res*, **25**, 413-425.
50. Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardimci, G.G., Chakraborty, A., Bann, D.V., Wang, Y. *et al.* (2018) Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet*, **50**, 1388-1398.
51. Yao, Y. and Dai, W. (2014) Genomic Instability and Cancer. *J Carcinog Mutagen*, **5**.
52. Giovanni Bussotti, L.P., Pascale Pescher, Malgorzata Anna Domagalska, K. Shanmugha Rajan, Tirza Doniger, Disha Gajanan Hiregange, Peter J Myler, Ron Unger, Shulamit J Michaeli, Gerald F. Spaeth. (2021) Genome instability drives epistatic adaptation in the human pathogen Leishmania. *bioRxiv*
53. Abbey, D.A., Funt, J., Lurie-Weinberger, M.N., Thompson, D.A., Regev, A., Myers, C.L. and Berman, J. (2014) YMAP: a pipeline for visualization of copy number variation and loss of heterozygosity in eukaryotic pathogens. *Genome Med*, **6**, 100.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

54. Bogaerts, B., Delcourt, T., Soetaert, K., Boarbi, S., Ceyssens, P.J., Winand, R., Van Braekel, J., De Keersmaecker, S.C.J., Roosens, N.H.C., Marchal, K. *et al.* (2021) A bioinformatics WGS workflow for clinical Mycobacterium tuberculosis complex isolate analysis, validated using a reference collection extensively characterized with conventional methods and in silico approaches. *J Clin Microbiol*.
55. Bogaerts, B., Winand, R., Fu, Q., Van Braekel, J., Ceyssens, P.J., Mattheus, W., Bertrand, S., De Keersmaecker, S.C.J., Roosens, N.H.C. and Vanneste, K. (2019) Validation of a Bioinformatics Workflow for Routine Analysis of Whole-Genome Sequencing Data and Related Challenges for Pathogen Typing in a European National Reference Center: Neisseria meningitidis as a Proof-of-Concept. *Front Microbiol*, **10**, 362.
56. Ellison, M.A., Walker, J.L., Ropp, P.J., Durrant, J.D. and Arndt, K.M. (2020) MutantHuntWGS: A Pipeline for Identifying Saccharomyces cerevisiae Mutations. *G3 (Bethesda)*, **10**, 3009-3014.
57. Quijada, N.M., Rodriguez-Lazaro, D., Eiros, J.M. and Hernandez, M. (2019) TORMES: an automated pipeline for whole bacterial genome analysis. *Bioinformatics*, **35**, 4207-4212.
58. Leinonen, R., Sugawara, H., Shumway, M. and Collaboration, I.N.S.D. (2011), *Nucleic Acids Res.*, Vol. 39, pp. D19-21.
59. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R. *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Res*, **39**, D28-31.
60. Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T. *et al.* (2016) ENCODE data at the ENCODE portal. *Nucleic Acids Res*, **44**, D726-732.
61. Howe, K.L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., Alvarez-Jarreta, J., Barba, M., Bolser, D.M., Cambell, L. *et al.* (2020) Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res*, **48**, D689-D695.
62. Arnaud, M.B., Costanzo, M.C., Skrzypek, M.S., Binkley, G., Lane, C., Miyasato, S.R. and Sherlock, G. (2005) The Candida Genome Database (CGD), a community resource for Candida albicans gene and protein information. *Nucleic Acids Res*, **33**, D358-363.
63. Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M.J., Gajria, B., Grant, G.R., Ginsburg, H., Gupta, D., Kissinger, J.C., Labo, P. *et al.* (2003) PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res*, **31**, 212-215.
64. Girgis, H.Z. (2015) Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics*, **16**, 227.
65. Li, H. (2013), *arXiv:1303.3997v1 [q-bio.GN]*.
66. Li, H. and Durbin, R. (2009), *Bioinformatics*. 2009/05/20 ed, Vol. 25, pp. 1754-1760.
67. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, G.P.D.P. (2009), *Bioinformatics*, Vol. 25, pp. 2078-2079.
68. Li, W. and Godzik, A. (2006), *Bioinformatics*, Vol. 22, pp. 1658-1659.
69. Basenko, E.Y., Pulman, J.A., Shanmugasundram, A., Harb, O.S., Crouch, K., Starns, D., Warrenfeltz, S., Aurrecoechea, C., Stoeckert, C.J., Jr., Kissinger, J.C. *et al.* (2018) FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. *J Fungi (Basel)*, **4**.
70. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27-30.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.15.448580; this version posted July 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

71. Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P. and Karp, P.D. (2020) The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res*, **48**, D445-D453.
72. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F. and Manke, T. (2016), *Nucleic Acids Res.*, Vol. 44, pp. W160-165.
73. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80-92.
74. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. and Lanfear, R. (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol*, **37**, 1530-1534.
75. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. and Vinh, L.S. (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol*, **35**, 518-522.
76. Yu, G., Lam, T.T., Zhu, H. and Guan, Y. (2018) Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol*, **35**, 3041-3043.
77. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009), *Genome Res.*, Vol. 19, pp. 1639-1645.