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:

One of the hallmarks of eukaryotic pathogens is the ability to genetically adapt to environmental change, causing for instance frequent drug resistance. Even though genome instability has been recognized as a key driver for microbial adaptation, most available computational tools focus just on one mutation type or analytical step. To overcome this limitation and better understand the role of genetic changes in enhancing microbial pathogenicity we established GIP, a novel, powerful bioinformatic pipeline for comparative genome analysis across microbial populations. GIP allows batch processing of whole genome sequencing datasets, including read alignment, normalization, quantification of chromosomes, genes and genomic bins, and detection of single nucleotide and structural variants. GIP produces a comprehensive summary report providing sample statistics together with graphical representations of genomic features and tabulated results. GIP further includes a tool-suite that enables downstream custom comparisons of

samples subsets. GIP is broadly applicable to different eukaryotic systems that exploit genome instability for adaptation, including *Leishmania*, *Plasmodium, Candida*, and cancer.

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 an number of publicly available databases have been compiled or updated (6). Parallel to data availability, many bioinformatics 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

48 detection or characterization of DNA single nucleotide variants (SNVs), copy number
49 variations (CNVs), and structural variations (SVs), but their scope is limited to the analysis of
50 one genomic feature at the time. A number of integrative WGS pipelines and workflows have
51 been established combining the execution of multiple bioinformatics algorithms serving
52 different analysis steps (12). Even though continuous progress has been made, there is no
53 standardized or unified approach for genomic investigation. For the development of improved
54 WGS data analysis pipelines, several important requirements need be considered, including
55 portability, reproducibility, scalability and compatibility with high-performance computing
56 (HPC) clusters and remote cloud computing. Here we introduce a novel genome instability
57 pipeline (GIP) that fulfills all these requirements. GIP facilitates the genome-wide detection,
58 quantification, comparison and visualization of chromosome aneuploidies, gene CNVs, SNVs
59 and SVs. GIP is implemented in Nextflow (13), a workflow language that allows to execute
60 GIP seamlessly in local workstation, on an HPC or remotely in the cloud. All required
61 environment and software dependencies of GIP are fulfilled and provided with a Singularity
62 container, thus making GIP reproducible, easy-to-install and easy-to-use. GIP allows the use of
63 giptools, a tool-suit of R-based modules for genome data exploration, enabling the comparison
64 of sample sub-sets. GIP and giptools generate a summary report with publication-quality figures
65 and spreadsheet tables. GIP and giptools constitute a single framework for WGS analysis
66 suitable both for large scale batch analysis of individual genomes and comparison of samples
67 from different experimental conditions or origins. Lastly, a key strength of GIP and giptools is
68 the general applicability to different eukaryotic species. We already successfully applied GIP
69 on the analysis of *Leishmania* genomes (14-16). In this study, we validate the use of GIP and
70 giptools using WGS data from published datasets of the three major human pathogens
71 *Leishmania infantum*, *Plasmodium vivax* and *Candida albicans* and as well as three human
72 cancer cell lines.

Results:

The GIP workflow. GIP is a tool for scientific investigation compatible with Linux systems, requiring minimal configuration and distributed as a self-contained package. GIP consists of 3 files: the Nextflow pipeline code, the configuration and the Singularity container files. The minimum required input is a paired-end WGS data set and a reference genome assembly in the standard fastq and FASTA format respectively. GIP analyses include (i) extracting genomic features such as assembly gaps or repetitive elements, (ii) mapping the reads, (iii) evaluating chromosome, gene and genomic bin copy numbers, (iv) identifying and visualizing copy number variation with respect to the reference genome, (v) identifying and quantifying gene clusters, (vi) detecting and annotating SNVs, (vii) measuring non-synonimous (N) and synonymous (S) mutations for all genes, (viii) detecting SVs including tandem duplications, deletions, inversions and break-ends translocations using split-read and read-pair orientation information, (ix) producing a report file providing summary statistics, tables and visualizations (**Fig. 1**). GIP allows to customize filtering and visualization options via the configuration file (see methods). The output of GIP can be used as input for giptools, a tool-suite to compare sample sub-sets and highlight chromosome copy number, gene copy number and SNV differences.

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,

97 research projects often involve downstream comparison between samples. Examples include
98 the comparison of gene or chromosome copy variation number between for example drug
99 resistant and drug susceptible samples, or the juxtaposition of SNVs detected in isolates from
100 different geographic areas. For this purpose, we developed giptools, a suit of thirteen modules
101 that allows to compare samples processed by GIP (**Table 1**). All modules in giptools are fully
102 embedded in the Singularity container and they are provided with their own documentation.

103 **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

104

105 To illustrate the type of exploratory data analyses and the biological questions that can be
106 addressed, we tested giptools on a previously analyzed dataset of seven clinical *Leishmania*
107 *infantum* isolates from Tunisia (14). *Leishmania* is the etiological agent of leishmaniasis, a life-
108 threatening human and veterinary disease affecting 12 million people worldwide (17). The
109 parasites were derived from patients affected with visceral leishmaniasis, expanded in cell
110 culture and their genomes were sequenced. The dataset includes four Glucantime drug
111 susceptible isolates and three isolates from relapsed patients, and thus comparison may inform
112 on genetic factors resulting in treatment failure. giptools allowed the detection and visualization
113 of pervasive intra-chromosomal CNVs across the thirty-six *Leishmania* chromosomes (**Fig.**

114 **2A**). Additionally, giptools enables targeted comparison of normalized genomic bin sequencing
115 coverage of sample pairs. We used giptools' `binCNV` module to compute the ratio between
116 corresponding genomic bins of the strains LIPA83 over ZK43, which correspond to a first-
117 episode and a relapse leishmaniasis isolate, respectively. giptools represents different
118 chromosomes as separate panels (**Fig. 2B**), as part of single genome-wide overview (**Fig. S1A**)
119 or as distinct plots (**Fig. S1B**). This analysis allowed the identification of 2 905 and 2 208 bins
120 that were respectively amplified or depleted in LIPA83 with respect to ZK43. The results are
121 returned by giptools as a Microsoft Excel table (.xlsx format) providing ratio scores at each
122 genomic position (**Table S1**). Likewise, giptools permits three-way comparisons of normalized
123 genomic bin sequencing coverage with ternary plots (**Fig. 2C**). We used this representation to
124 display the genomic bin relative abundance in samples ZK43, LIPA83 and ZK28. The analysis
125 shows important strain-specific differences in bin copy number that are visualized by shifts of
126 the signals out of the center. Similar to genomic bin analysis, giptools makes it possible to
127 compare the sequencing depth of annotated genes and thus the copy number in two or three
128 samples (**Fig. 2D** and **Fig. S2**). However, the determination of gene copy number might be
129 impeded by (i) short read length or fragment insert size, (ii) the complexity of the target genome,
130 or (iii) the presence of repetitive elements. The read map quality (MAPQ) score is a measure
131 that reflects how much each gene is supported by unambiguously mapped reads (high MAPQ)
132 in contrast to multimapping reads (low MAPQ). Together with the coverage, GIP also computes
133 the mean read MAPQ score for each gene and allows a different strategy to determine the copy
134 number of low MAPQ genes (see methods). The evaluation of LIPA83 and ZK43 gene
135 coverage ratio scores revealed 13 gene CNVs with a stringent MAPQ cutoff of 50 (**Table S2**).
136 The maximum normalized coverage value of 6.5 was observed for a putative amastin surface
137 glycoprotein (LINF_310009800). Other examples of gene CNVs in this set include the putative
138 surface antigen protein 2 (LINF_120013500) and the heat shock protein HSP33

139 (LINF_300021600) (**Table S2**). Genes falling below a user defined MAPQ score and sharing
140 high level of sequence similarity are assigned to the same gene cluster, and their measured
141 coverage scores are averaged across all members of the group. Low MAPQ scores can also be
142 associated with single genes, e.g. in the case of internal repetitive elements that cause multiple
143 ambiguous alignments inside the gene itself, or if mapping occurs in possibly misannotated
144 intergenic regions. The LIPA83/ZK43 comparison showed 27 CNV gene clusters, including
145 cluster clstr303 (3 genes with 'amastin-like' annotation) and cluster clstr16 (2 tb-292
146 membrane-associated protein-like proteins) (**Table S2**). These results demonstrate the power
147 of GIP and giptools to detect and compare intra-chromosomal CNVs in *Leishmania* at genomic
148 bin level. Conveniently, analogous two- or three-ways comparisons can be applied to reveal
149 copy number variations at individual gene or gene cluster levels.

150

151 **Comparative genomics of a *Plasmodium vivax* WGS dataset.** We next applied GIP and
152 giptools on other biological systems to demonstrate its broad applicability outside the
153 *Leishmania* field, including the human apicomplexan parasite *Plasmodium vivax*. *Plasmodium*
154 *vivax* is an obligate protist parasite and a human pathogen causing malarial illness. *Plasmodium*
155 *vivax* gives rise every year to 130 million clinical cases (18), and it is estimated that 2.5 billion
156 people are at risk of infection worldwide (19-21). We applied GIP and giptools to investigate
157 genomic variations across a sizeable dataset of 222 *Plasmodium vivax* genomes isolated from
158 clinical samples of 14 countries worldwide (22,23) (**Table S3**). The phylogenetic tree
159 reconstruction and PCA analyses (**Fig. 3A** and **B**) showed a high correlation between genotypes
160 and the geographic origin of the samples. However, we detected substantial genomic variability
161 between isolates collected at smaller geographical scale, with 14 555 SNVs (~42% of the total)
162 uniquely characterizing representative samples from five Ethiopian study sites (23) (**Fig. 3C**
163 and **D**). This result may reflect diverging evolutionary trajectories radiating from few founder

164 strains. At gene level we profiled the copy number variations of two gene panels. The first panel
165 accounts for 43 previously described genes encoding for potential erythrocyte binding proteins
166 suggested to operate at the interface of the parasite-host invasion process (23). The second panel
167 includes 2 drug resistance markers comprising the chloroquine resistance transporter
168 PVP01_0109300 and the multidrug resistance protein 1 PVP01_1010900, and 4 proteins
169 implicated in red blood cell invasion, such as the merozoite surface protein gene MSP7
170 PVP01_1219700, the reticulocyte binding protein gene 2c PVP01_0534300, the serine-repeat
171 antigen 3 PVP01_0417000, and the reticulocyte binding protein 2b PVP01_0800700) (24-33).
172 Read depth analysis indicated that 4 genes in the panel (PVP01_0623800, PVP01_1031400,
173 PVP01_1031200, PVP01_1031300) show a high degree of variability, with amplifications
174 observed in samples from distinct geographic areas thus suggesting a convergent evolutionary
175 process (**Fig. 3E**). Furthermore, 6 genes positioned on chromosome 14 are deleted in the Thai
176 strain PD0689_C as a result of the entire chromosome loss (nullisomy) (**Fig. 3E** and **F**). Finally,
177 the comparison of synonymous and non-synonymous SNVs in the panel genes revealed
178 important differences between sample groups. Our analysis indicates an overall higher number
179 of non-synonymous mutations in Ethiopian rather than Cambodian isolates, therefore
180 suggesting a stronger evolutionary pressure in the African strains (**Fig. 3G**). Taken together
181 these analyses well illustrate how GIP and giptools can be readily applied for bulk analysis of
182 *Plasmodium vivax* genomes to assess genome diversity, extract evolutionary information and
183 profile potential disease biomarkers.

184

185 **Gene CNV analysis of *Candida albicans* evolutionary adapted strains.** We next applied GIP
186 and giptools to the human fungal pathogen *Candida albicans*, an opportunistic yeast exhibiting
187 major genome plasticity (34-43) and causing hundreds of thousands of severe infections each
188 year (44). Candidemia, a bloodstream infection with *Candida*, are often associated with high

189 rates of morbidity and mortality, (15–50%) notwithstanding existing antifungal treatments
190 (45,46). We applied GIP and giptools to a *Candida albicans* WGS dataset described in a recent
191 study that covers five different progenitor strains (P75063, P75016, P78042, SC5314,
192 AMS3050) and investigates CNVs driving tolerance and resistance to anti-fungal azole drugs
193 (47). We analyzed nineteen samples, including (i) four clinical isolates (P75063, P75016,
194 P78042, SC5314), (ii) seven strains selected *in vitro* against the anti-fungal drug fluconazole
195 (FLC) (AMS4104, AMS4105, AMS4106, AMS4107, AMS4397, AMS4444, AMS4702), (iii)
196 four isogenic colonies adapted to the drug miconazole (AMS3051, AMS3052, AMS3053 and
197 AMS3054) together with their progenitor (AMS3050), and (iv) three colonies derived from a
198 miconazole-adapted population and isolated on a rich medium (AMS3092, AMS3093 and
199 AMS3094) (47-49). We confirmed gene amplification for the drug efflux pumps TAC1
200 (orf19.3188) and ERG11 (orf19.922), for the stress response proteins HSP70 (orf19.4980),
201 CGR1 (orf19.2722), ERO1 (orf19.4871), TPK1 (orf19.4892), ASR1 (orf19.2344), PBS2
202 (orf19.7388) and CRZ1 (orf19.7359), and for proteins involved in membrane and cell wall
203 integrity, including CDR3 (orf19.1313), NCP1 (orf19.2672), ECM21 (orf19.4887), MNN23
204 (orf19.4874), RHB1 (orf19.5994) and KRE6 (orf19.7363) (**Table S4**). Further, the cross
205 comparison of gene read depths across the nineteen genomes uncovered a novel set of 1 505
206 gene CNV showing correlating and anti-correlating gene dosage variations (**Fig. 4A** and **B**,
207 **Table S4**) that could be assigned to nine distinct correlation clusters (CC) (**Fig. S3A**, **Table**
208 **S5**). We verified the sequencing coverage of genomic regions encompassing gene CNVs,
209 including three regions amplified in fluconazole resistant strains (**Fig. 4B**, **Fig. S3B**) (47) and
210 a region whose amplification correlates with the level of miconazole resistance (47) (**Fig. S3C**),
211 as well as the loss of heterozygosity associated to the depletion of chromosome 3 left arm in
212 sample AMS3051 (**Fig. S3D**) (47). Eventually, by representing genes and absolute correlation
213 respectively as nodes and edges of a network, we identified 9 highly interconnected network

214 clusters (NC) (**Fig. 4C**, **Table S6**). NC7, NC8 and NC9 embody genes from individual
215 chromosomes, respectively chromosomes 1, 3 and 4. The most parsimonious explanation for
216 the high levels of correlation observed in these NCs (**Fig. 4C**) is the occurrence of sub-
217 chromosomal amplifications affecting several adjacent genes. A different scenario is pictured
218 for each of the remaining NCs (NC1-6) where the genes are located on different chromosomes
219 thus suggesting genetic interactions that causes coordinated changes in gene copy number.
220 Altogether, the reproducibility of previously published results validates GIP and giptools for
221 the analysis of fungal gene CNVs, and uncovers a role of epistatic CNV interactions similar to
222 *Leishmania* (in preparation) in driving genomic adaptation in *Candida albicans* populations
223 under drug selection.

224

225 **Exploring instability of larger genomes using cancer cell lines as a benchmark.** The larger
226 genome size, and the higher number of genes and WGS reads and can represent a challenge
227 when working with higher eukaryotes. For the purpose of comparison, the human genome is
228 ~216 times larger than the one of *Candida albicans* we analyse in this study. Therefore, we
229 sought to evaluate the applicability of the GIP and giptools framework to human data and
230 utilized a panel of cancer cell lines as a test set. In our analyses we considered publicly available
231 WGS data of the cell lines T47D, NCI_H460 and K562 (50), which respectively derive from
232 human breast, lung and blood cancers. We detected 1 647 016 SNVs (**Supplementary Data 1**)
233 and allele frequency shifts indicating repeated loss of heterozygosity events (**Fig. 5**). The
234 coverage analysis indicates pervasive presence of CNVs both at chromosomal and sub-
235 chromosomal levels (**Fig. S4**) correlating with the SNV frequency patterns (**Fig. S5** and **S6**).
236 These analyses exemplify the application of GIP and giptools to larger, more complex genomes
237 involving minor parameter adjustments while relying on the same bioinformatic tool.

238

239 **Discussion:**

240 Genome instability is a key driver of evolution for microbial pathogens and cancer cells (51)
241 and a major source of human morbidity. Here we introduce GIP and giptools, a unified
242 framework for genotype profiling of biological systems exploiting genome instability for
243 adaptation. Current analysis tools suffer from several limitations. A fist common limitation is
244 that their application is limited to individual types of mutations, but are unable to provide
245 genome-wide comprehensive reports covering multiple forms of genome instability. A second
246 limitation is when analysis tools allow to identify variations with respects to a reference
247 assembly, but do not support comparisons between samples. A third important limitation is the
248 software applicability range, where genome analysis tools are tailored for the analysis of WGS
249 data from a given species (52-56), but are not generally applicable to different biological
250 systems. Such tools cannot be applied to investigate genome variations across multiple species
251 and explore potentially conserved genomic adaptation mechanisms. A possible workaround is
252 to use different tools for different species. However, each tool has its own parametrization and
253 specificity, and as a consequence this strategy has the potential to introduce technical biases
254 making unable to perform meaningful cross-species comparisons. Furthermore, especially
255 when investigating non-model organisms, a tool may not be available for the analysis of your
256 species of interest. A forth critical limitation—with potential serious consequences for clinical
257 applications, is software portability and more in general reproducibility. The same
258 bioinformatic application may be unsuited to run on different computer environments, or worse,
259 if the software requirements are not carefully specified, it can generate irreproducible and
260 potentially flawed results. Lastly, one more common limitation is posed by software scalability.
261 In the WGS domain, with the rapid increase new samples made available and the enormous
262 amount of data generated in each sequencing run, the CPU and memory resources of local
263 workstation risk to quickly become inadequate for data analysis. Therefore, it is paramount that

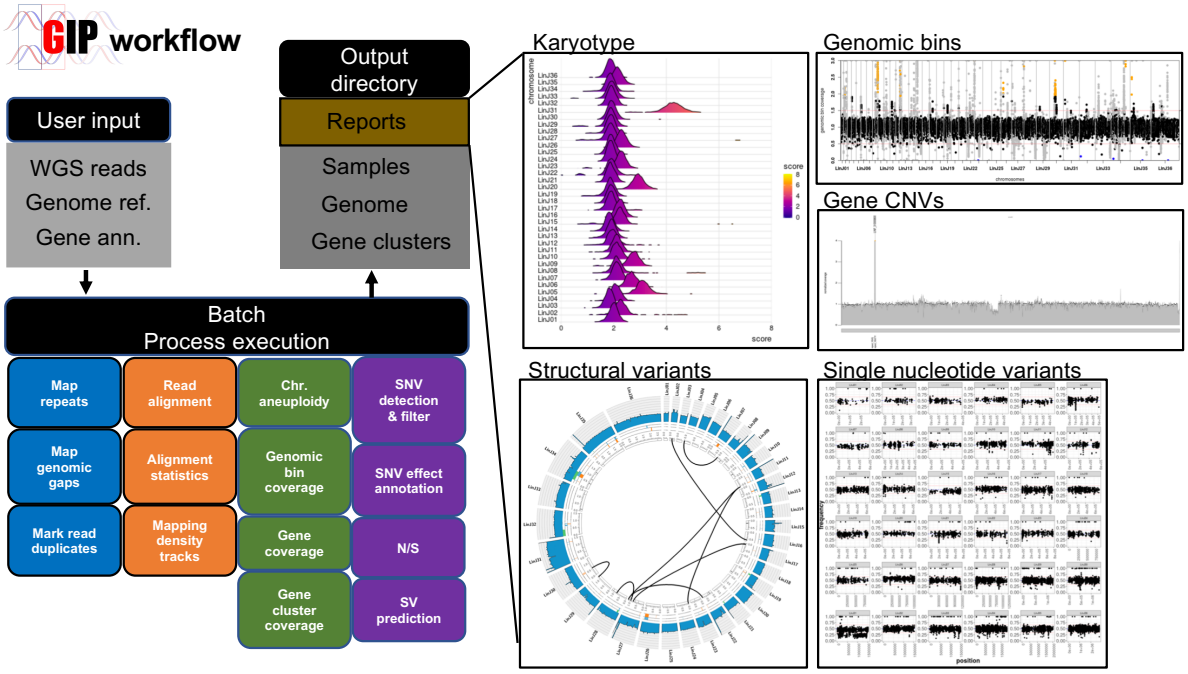
264 WGS tools are implemented to run on high-performance computing (HPC) clusters and feature
265 remote cloud computing solutions.
266 To support and ease the study of genome instability and address all these common challenges
267 we developed GIP and giptools. GIP provides karyotype, gene CNV, SNV and SV batch
268 analyses. giptools enables custom sample comparisons, to explore differences and common
269 features between genomes. Our pipeline limits as much as possible the use of hardcoded
270 parametrization which could limit its use to a specific organism. Therefore, GIP's flexible
271 design makes it adapted for the genome analysis of both model and non-model organisms,
272 including *Leishmania* or human. In this study we exemplify the ability GIP and giptools in
273 performing genomic screenings of three major pathogenic eukaryotes and human cancer cell
274 lines. Thanks to the Singularity implementation all required software are embedded and
275 provided within the container. As a consequence, a user could easily recreate the same work
276 environment just by downloading the pipeline container, and reproduce exactly the same
277 publication-quality plots and tables presented in this study (see methods). All plots and table
278 presented in this study were directly generated with GIP and giptools with only minor manual
279 layout changes to fit multiple plots in the same figure and homogenize the font sizes. The only
280 exception is the plot showing sample PD0689_C chromosome 14 nullisomy (**Fig. 3F**), which
281 was generated visualizing with IGV (57) the coverage density track produced by GIP (see
282 methods). Finally, because of its Nextflow implementation GIP can be executed on a local
283 machine, on cluster resource manager or the cloud. GIP can be executed on individual samples
284 and with no additional effort on large WGS data sets for batch computation as shown for the
285 222 *Plasmodium vivax* genomes. While originally deployed for *Leishmania* genome analysis,
286 in this study we validate the use of our pipeline on other organisms reproducing expected
287 results. For example, in *Candida albicans* we confirmed the CNVs correlating to drug
288 resistance as well as a loss of heterozygosity event (**Fig. S3B-C** and **D**). Parallel to this we also

289 show how GIP and giptools can be used for data mining and scientific discovery. New findings
290 include (i) the discovery of the convergent amplification of erythrocyte binding proteins in
291 *Plasmodium vivax* strains sampled from distinct geographic areas (**Fig. 3E**), (ii) the detection
292 of a nullisomic strain (**Fig. 3F**), and (iii) the identification of correlated copy number variations
293 between genes positioned on separate chromosomes of *Candida albicans* adapting strains.
294 These results well illustrate how GIP and giptools can be applied to perform extended genomic
295 analyses in different biological systems and drive biomedical discovery. To conclude, we
296 believe that GIP and giptools represent a step forward toward reproducible research in
297 genomics, and provide a robust framework study how microbes and tumor cells harness genome
298 instability for environmental adaptation and fitness gain.

299

300 **Figures**

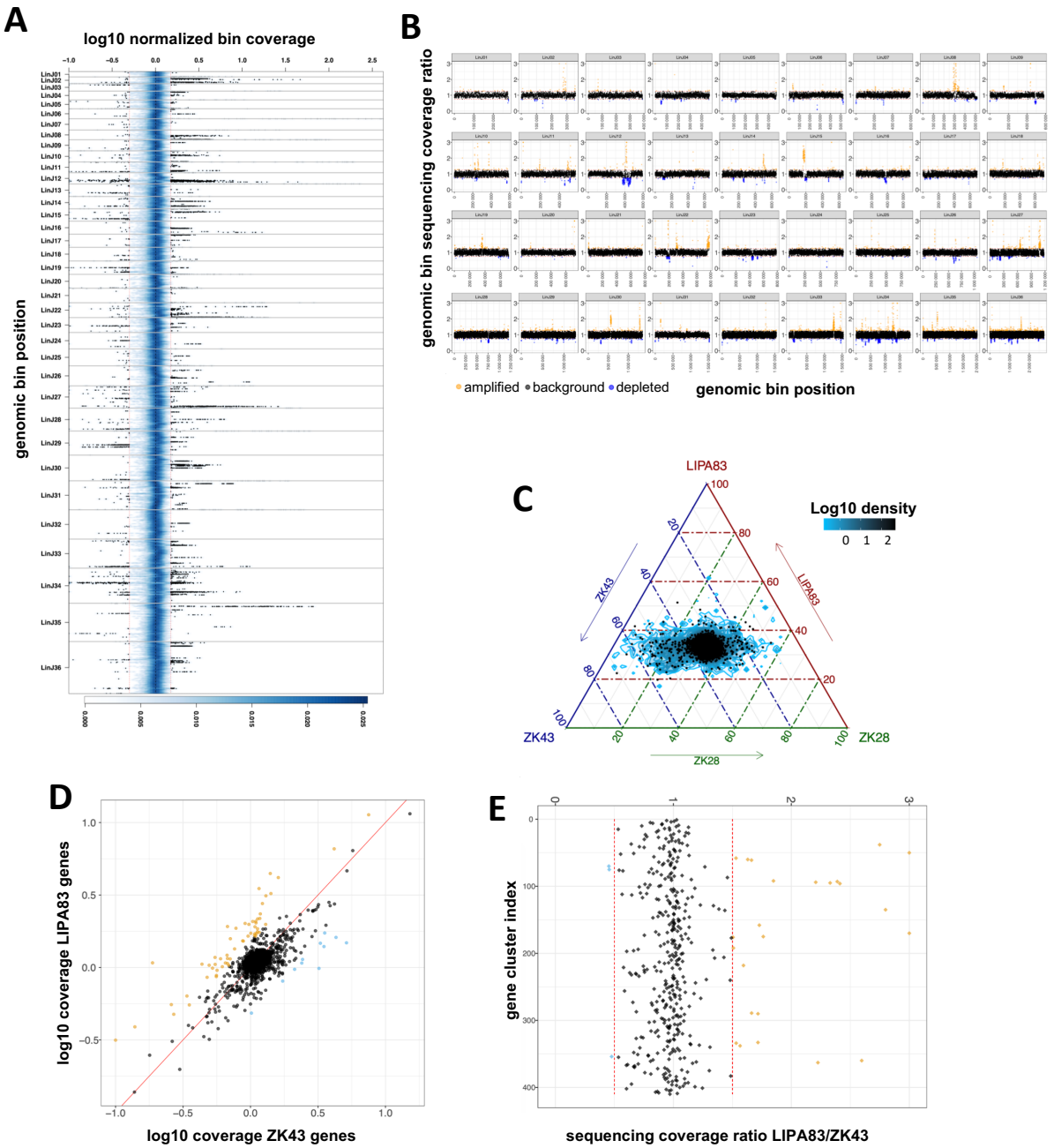
301 **Fig. 1**



302

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

315 **Fig. 2:**

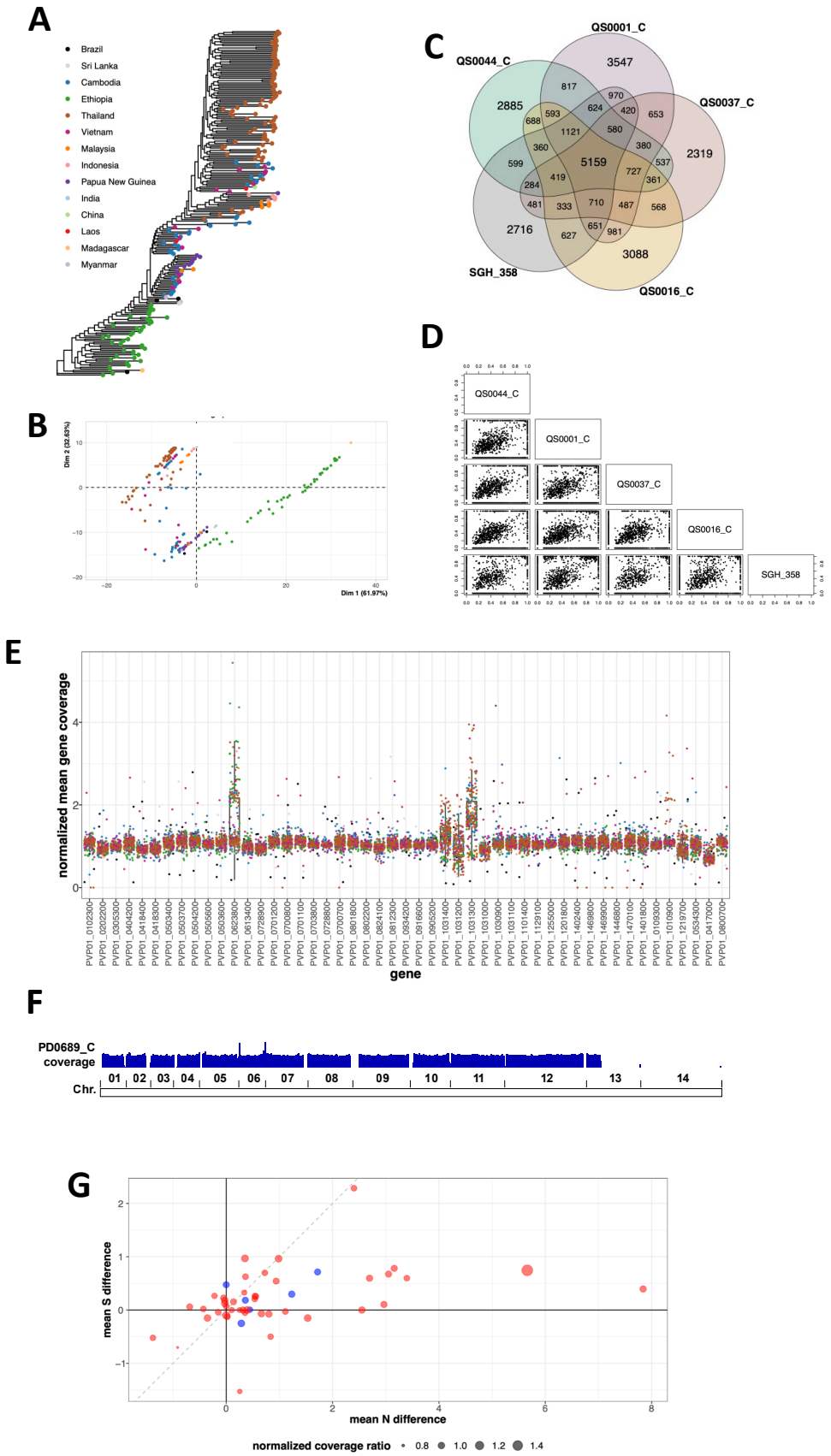


316

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

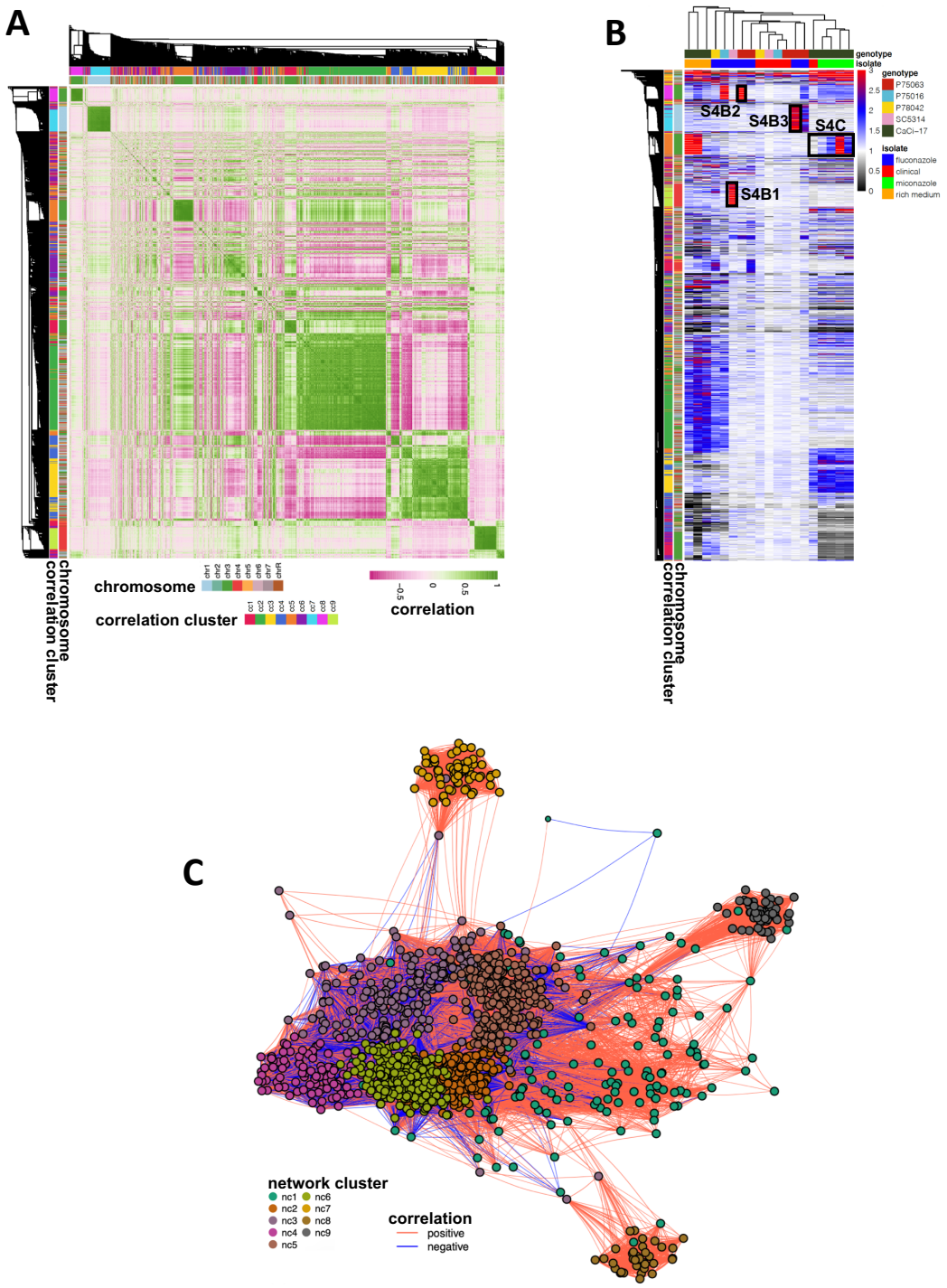
322 coverage values. A selection of 50 000 bins with coverage > 1.5 or < 0.5 is shown as black dots.
323 (**B**) Scatterplot of the genomic bin normalized sequencing coverage ratio of samples LIPA83
324 over ZK43. The x any y axes show the ratio score and the genomic position respectively. Ratio
325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

Fig. 3:


A
Brazil
Sri Lanka
Cambodia
Ethiopia
Thailand
Vietnam
Malaysia
Indonesia
Papua New Guinea
India
China
Laos
Madagascar
Myanmar
B
Dim 2 (22.63%)
Dim 1 (61.97%)
C
QS0001_C
QS0044_C
QS0037_C
SGH_358
QS0016_C
3547
2885
2319
2716
3088
5159
D
QS0044_C
QS0001_C
QS0037_C
QS0016_C
SGH_358
E
normalized mean gene coverage
gene
F
PD0689_C
coverage
Chr.
01
02
03
04
05
06
07
08
09
10
11
12
13
14
G
mean S difference
mean N difference
normalized coverage ratio
0.8
1.0
1.2
1.4

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

357 **Fig. 4:**

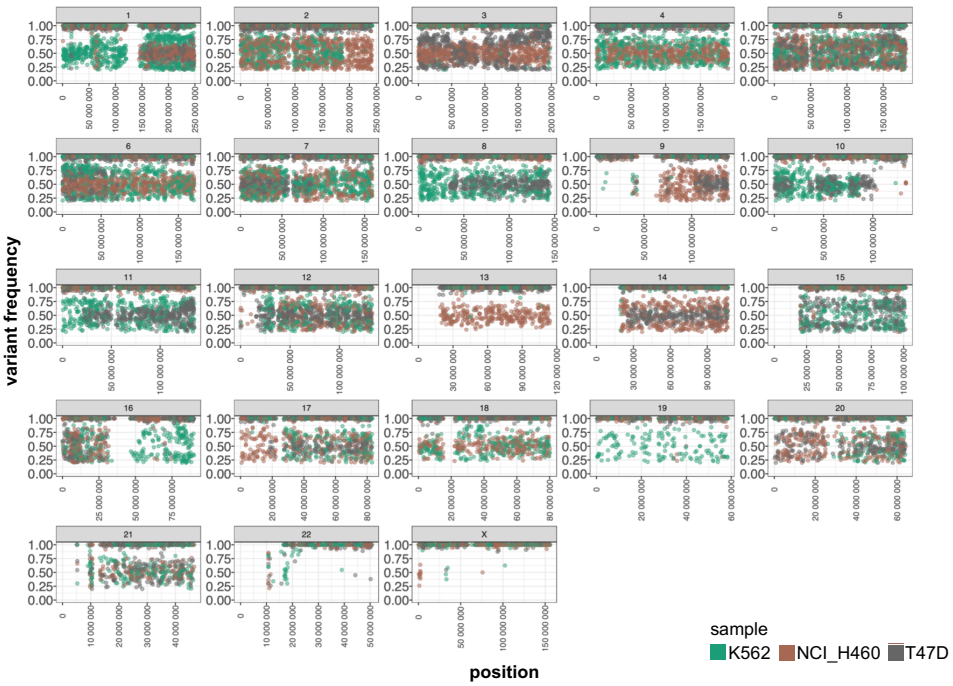


358

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

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

375 **Fig. 5:**



376

377 SNV heterogeneity of cancer cell lines. The x-axis indicates the genomic position. The y-axis

378 indicates the variant allele frequency. The dots indicate SNVs and they are coloured according

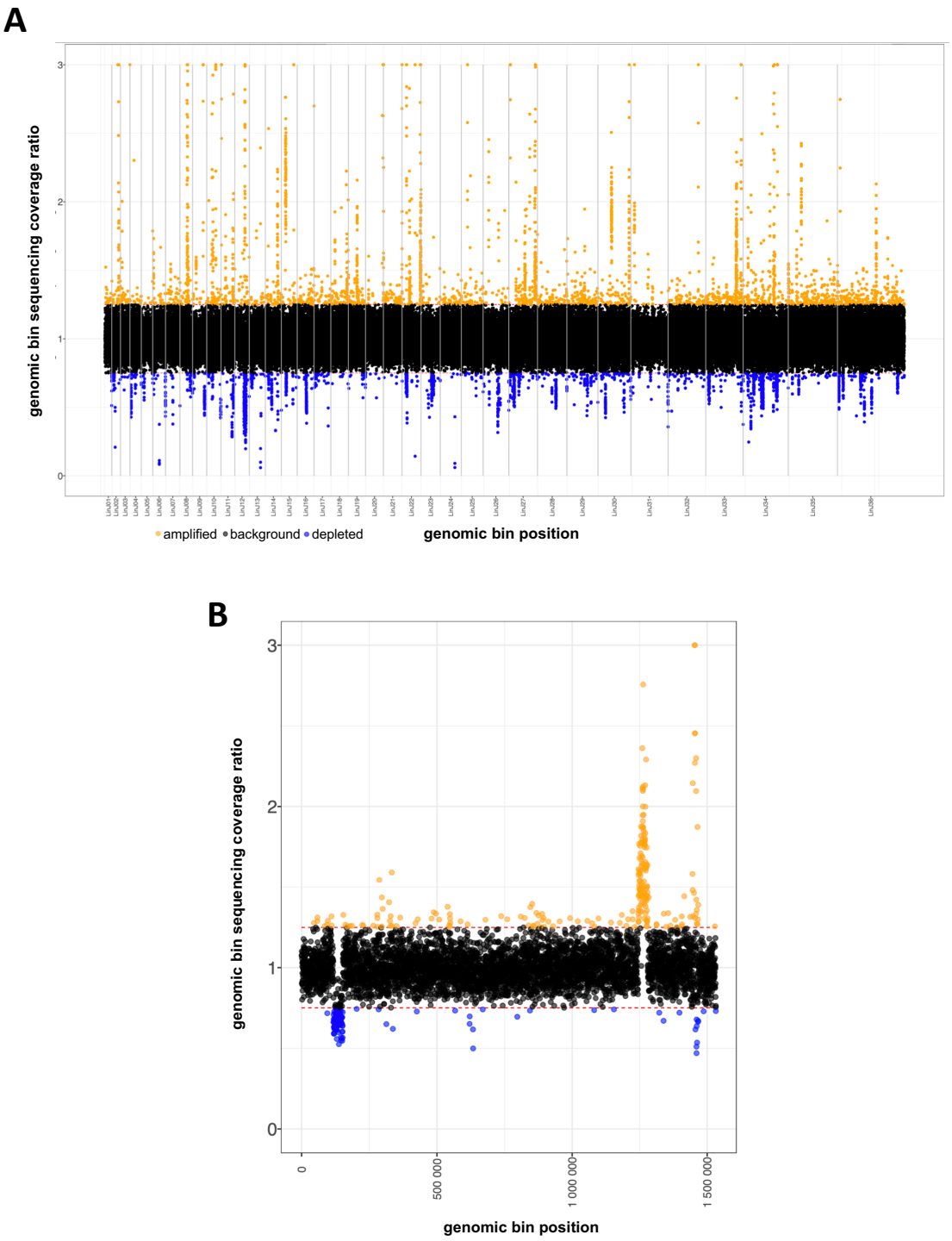
379 to the sample. K562, green; NCI_H460, brown; T47D, grey. Different panels show different

380 chromosomes.

381

382 **Supplementary figures**

383 **Fig. S1**:



384

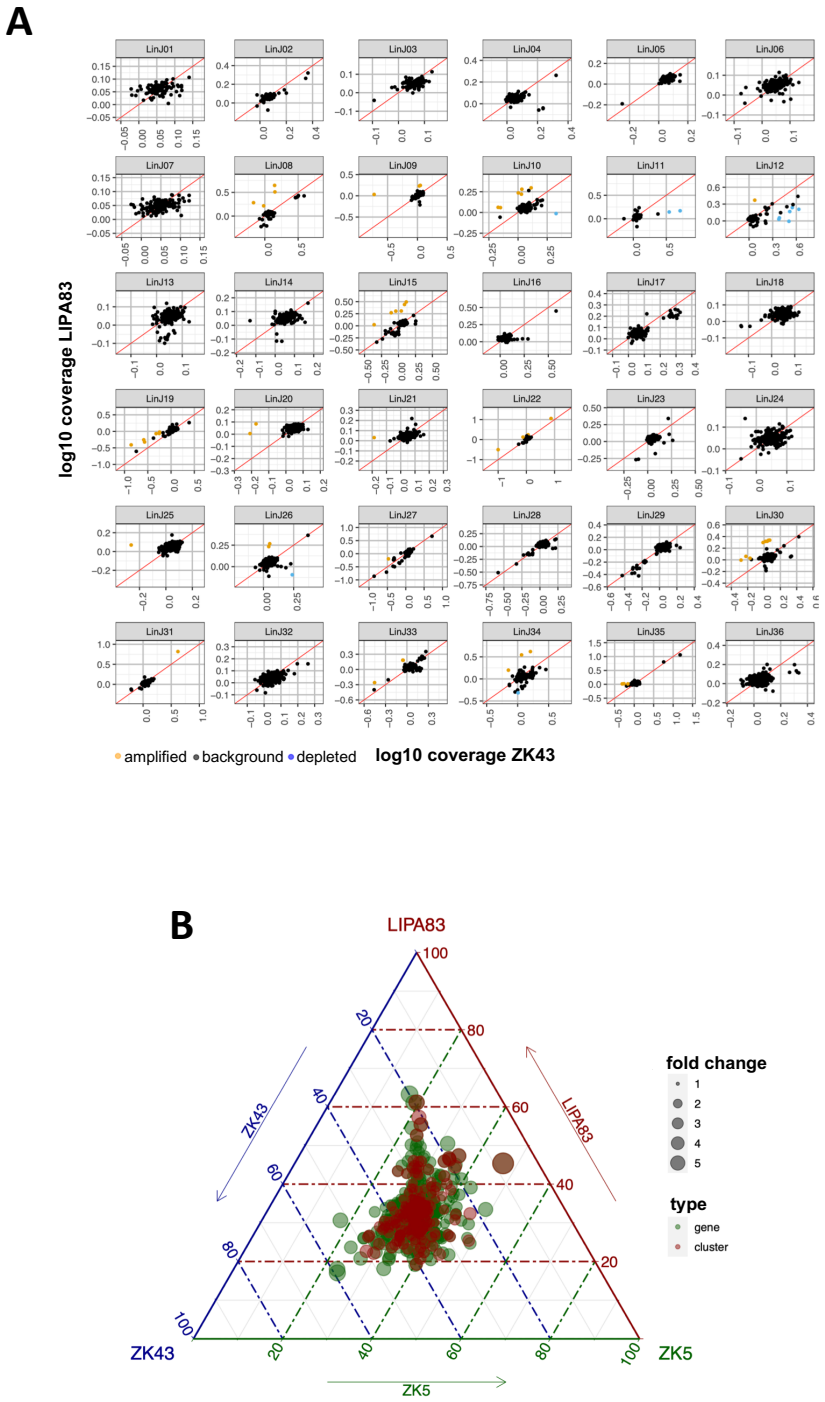
385 Bin coverage ratio visualizations. The figure shows alternative representations produced for the

386 genomic bin normalized sequencing coverage ratio of samples ZK43 and LIPA83. Same layout

387 as **Fig. 2B**. (**A**) Whole genome overview. (**B**) All the individual chromosomes separately. The

388 plot shows the example of chromosome 33.

389 **Fig. S2**:

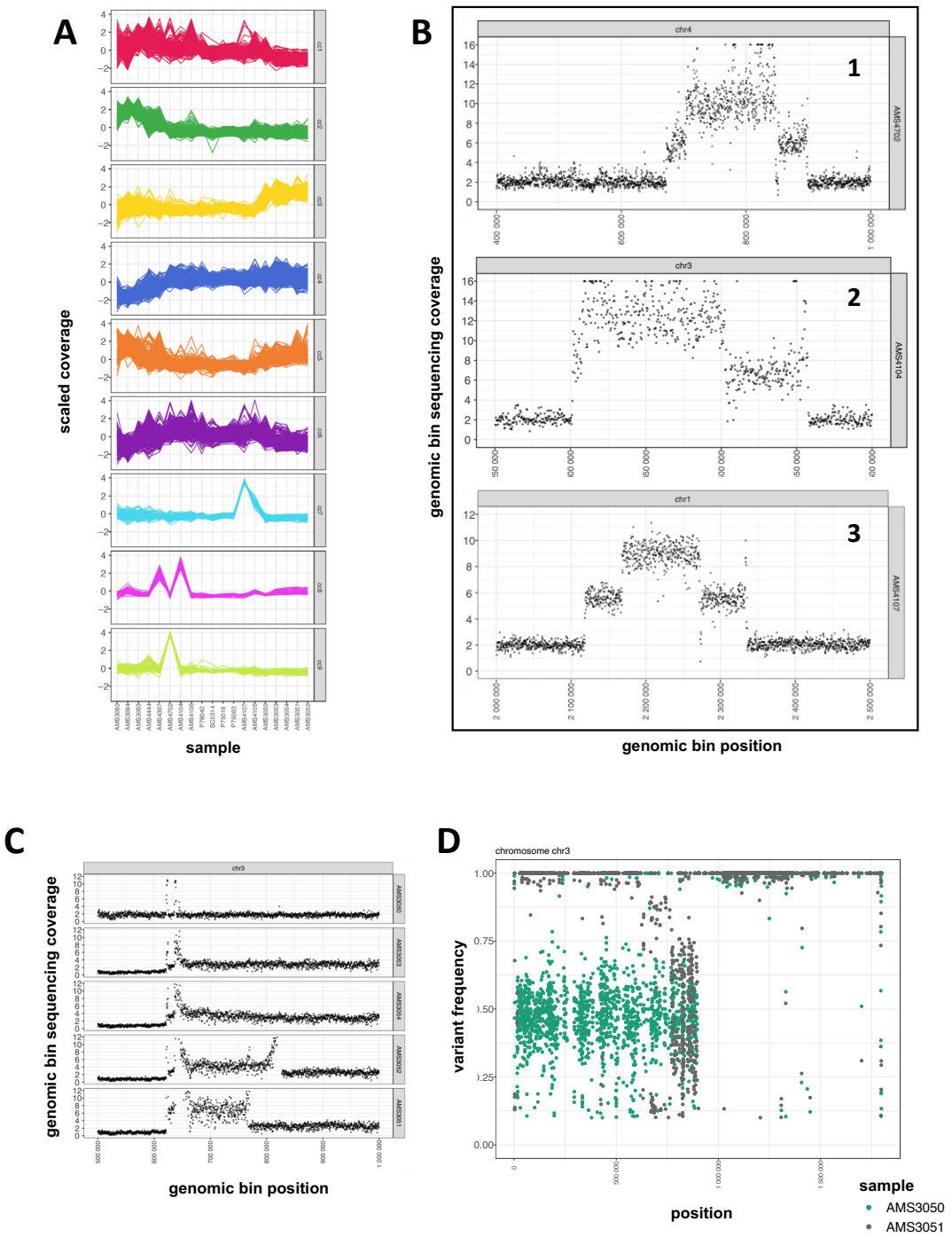


390

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

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

400 **Fig. S3:**

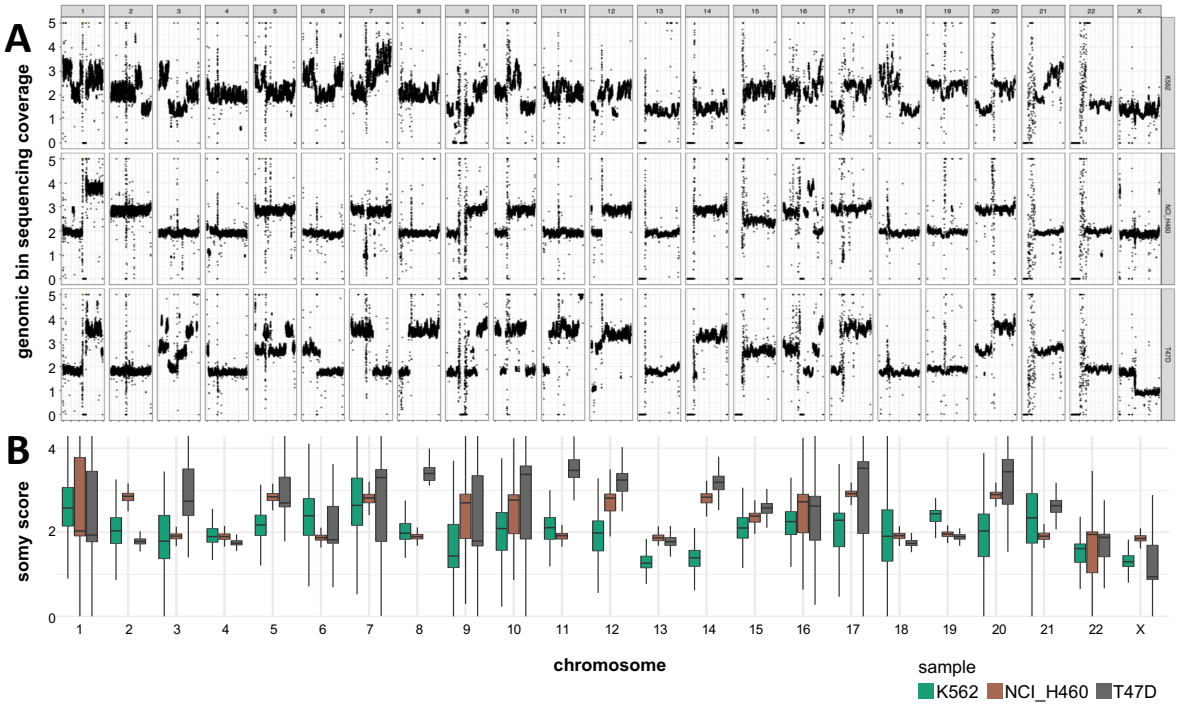


401

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

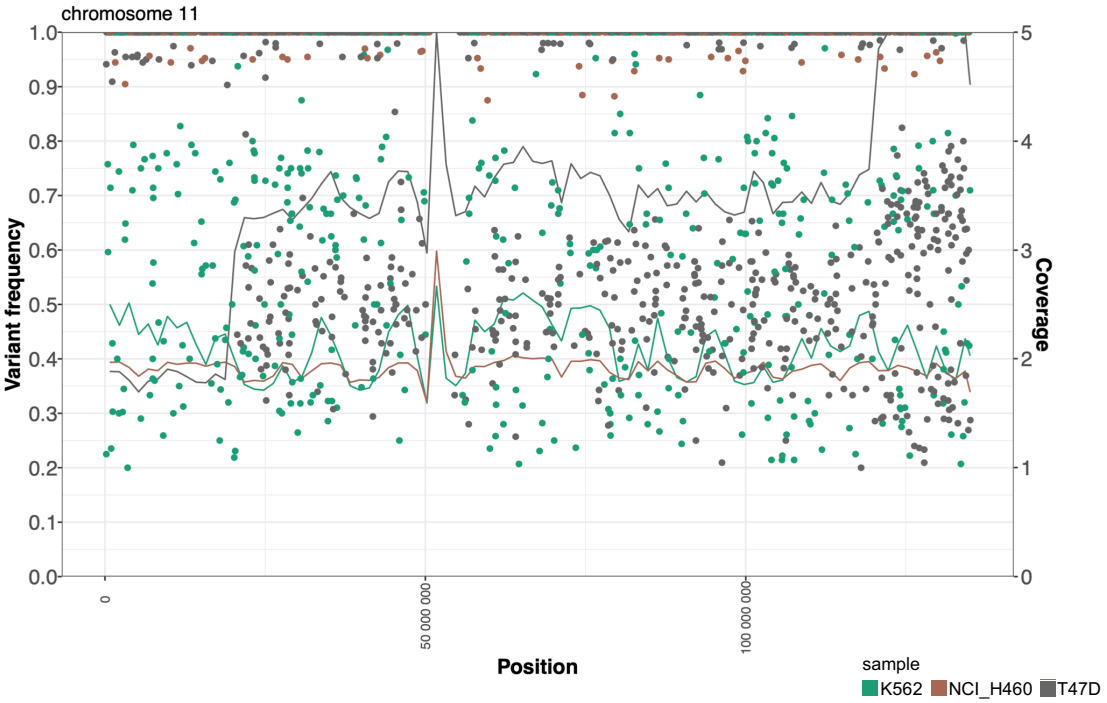
406 of three genomic regions (chr4:400 000-1 000 000; chr3:1 250 000-1 500 000; chr1:2 000 000-
407 2 500 000) in three different samples AMS4702, AMS4104 and AMS4107. The x-axis shows
408 the genomic position. Dots represent genomic bins. (**C**) Scatterplot showing the bin normalized
409 sequencing coverage of a genomic region (chr3:500 000-1 000 000) in five different samples
410 represented as separate panels. From top to bottom the samples are: AMS3050, AMS3053,
411 AMS3054, AMS3052 and AMS3051. The x and y axes show respectively the genomic position
412 and the normalized sequencing coverage. Dots represent bins. (**D**) Comparative analysis of
413 chromosome 3 SNVs. AMS3050 SNVs are shown in green, while AMS3051 SNVs are shown
414 in grey. The x and the y axes show respectively the genomic position and the variant allele
415 frequency.

Fig. S4:



Genomic coverage analysis of three cancer cell lines. (**A**) The x-axis indicates the genomic position. The y-axis indicates the normalized genomic bin sequencing coverage values. Dots demonstrate the genomic bins. To ease visualization, coverage values greater than 5 are reported as 5. Different panels show different chromosomes. The three panel rows indicate top to bottom the following samples: K562, NCI_H460 and T47D. (**B**) Chromosome coverage analysis. The x-axis reports the chromosomes. The y-axis reports the estimated somy score. The boxes show the somy score distributions. Green, brown, and grey boxes indicate respectively samples K562, NCI_H460 and T47D.

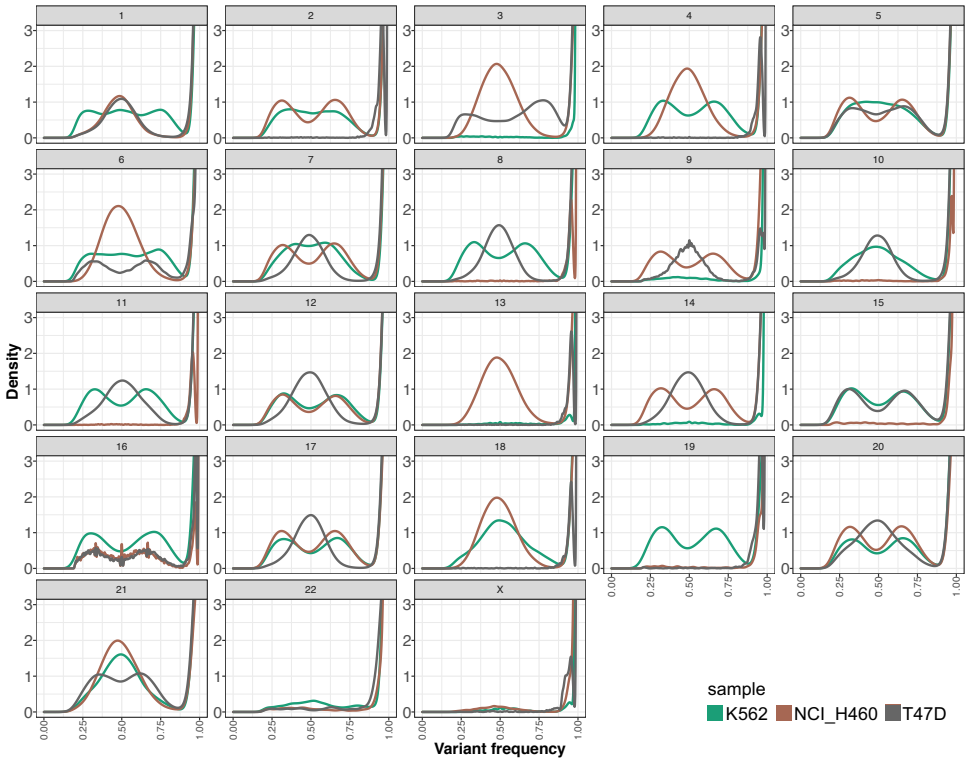
426 **Fig. S5:**



427

428 Chromosome 11 combined SNV and bin coverage plot. To ease visualization, giptools allows
429 the simultaneous displaying of variant allele frequencies (y-axis, left) and sequencing coverage
430 (y-axis, right). Dots represent SNVs. The lines represent the normalized bin sequencing
431 coverage. The x-axis indicates the genomic position. Coverage values greater than 5 are shown
432 as 5. Green, brown, and grey colours indicate respectively samples K562, NCI_H460 and
433 T47D.

434 **Fig. S6:**



435

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

439

440

441 **Methods:**

442 **GIP and giptools.** All results presented in this study were generated using GIP and giptools
443 version 1.0.9. GIP code is available upon request. The GIP configuration files (**Supplementary**
444 **Data 2**) and the giptools command options used to generate all results (**Supplementary Data**
445 **3**) are provided.

446 **Read alignment.** WGS reads were downloaded from the Sequence Read Archive (SRA) (58)
447 and the European Nucleotide Archive (ENA) (59) repositories and the Encyclopedia of DNA
448 Elements (ENCODE) dashboard (60) (**Table S3**). For *Leishmania infantum* the
449 GCA_900500625 genome reference and gene annotations available from the ENSEMBL
450 protists (61) server (release-48) were used. For *Candida albicans* the assembly 21 of the
451 SC5314 strain genome reference and gene annotations available from the *Candida* Genome
452 Database (CGD) (62) were used. For *Plasmodium vivax* the P01 reference genome and gene
453 annotations available from PlasmoDB (63) (release-50) were used. For the cancer cell lines the
454 human genome GRCh38 primary assembly and gene annotations available from ENSEMBL
455 (release-102) were used. The repetitive elements of reference genomes were soft-masked by
456 GIP using Red (64). WGS reads were mapped by GIP using BWA-mem (version 0.7.17)
457 (65,66) run with option -M to label shorter split hits as secondary. Then the alignment files
458 were sorted, indexed and reformatted by GIP using Samtools (version 1.8) (67). Finally, read
459 duplicates were removed by GIP using Picard MarkDuplicates
460 (http://broadinstitute.github.io/picard) (version 2.18.9) with the option
461 "VALIDATION_STRINGENCY=LENIENT." In the four considered datasets, WGS reads
462 were aligned against full assemblies, including unsorted contigs if present. However just the
463 canonical assembled chromosomes were considered for all downstream analyses (`chrs` option,
464 **Supplementary Data 2**). A minimum read alignment MAPQ score was adopted to select genes
465 for cluster analysis, and to call for SNVs and SVs (`MAPQ` option, **Supplementary Data 2**).
466 Altogether a total of 6 306 951 266 reads were aligned to the respective reference genomes.
467 The `giptools overview` module was run to gather the alignment statistics as estimated by
468 Picard CollectAlignmentSummaryMetrics (**Table S7**).

469 **Genomic bins and genes quantification.** GIP was used to evaluate the mean sequencing
470 coverage and the mean read MAPQ of genomic bins and genes. For genomic bins, GIP

471 partitioned the input genomes into adjacent windows of user defined lengths (`binSize` option,
472 **Supplementary Data 2**). The coverage GC-content score bias was corrected (`CGcorrect`
473 option, **Supplementary Data 2**) fitting a LOESS regression with a 5-fold cross validation to
474 optimize the model span parameter. A larger window length was utilized to bin the reference
475 genomes for Circos plot representations (`binSizeCircos` option, **Supplementary Data 2**). In
476 **Fig. 1** ("Genomic bins" and "Gene CNVs" plots), **Fig. 2**, **Fig. S1**, **Fig. S2** and **Fig. 4** bins and
477 genes coverage scores were normalized by median chromosome coverage to highlight
478 amplifications or depletions with respect to the chromosome copy number. In **Fig. 1**
479 ("Structural variants" plot), **Fig. S3B-C**, **Fig. 3E-G**, and **Fig. S4A** bins and genes coverage
480 scores were normalized by median genome coverage to account for sequencing library size
481 differences. GIP evaluated statistically significant copy number variant bins and genes (**Fig. 1**
482 "Genomic bins" and "Gene CNVs" plots) using a p-value threshold of 0.001
483 (`covPerBinSigOPT` and `covPerGeSigOPT` options, **Supplementary Data 2**). Estimated p-
484 values for bins and genes CNVs were corrected for multiple testing using the Benjamini –
485 Yekutieli (`--padjust BY `) and the Benjamini – Hochberg (`--padjust BH`) methods. The somy
486 scores shown in **Fig. 1** (karyotype plot) and **Fig. S4B** were computed multiplying the median
487 genome coverage normalized bin coverage by 2. GIP enabled the CNV analysis of genes
488 sharing high sequence identity by clustering the nucleotide sequences of the genes with low
489 mean MAPQ score into groups with cd-hit-est (version 4.8.1) (68) with options `-s 0.9 -c 0.9 -
490 r 0 -d 0 -g 1`. Then for each gene cluster GIP computed the mean gene coverage normalized by
491 median chromosome coverage (**Fig. 2E**, **Fig S2B**).

492 **Sequencing coverage density estimates.** GIP was used to convert the read alignment files
493 (.bam format) in binary data files reflecting sequencing coverage (.bigWig format). The
494 coverage file were produced using bamCoverage from the deepTools2 suite (69) (version 3.5.1)
495 with options "--normalizeUsing RPKM --ignoreDuplicates --binSize 10 --smoothLength 30"

496 (`bigWigOPT` option, **Supplementary Data 2**). The coverage track of sample PD0689_C was
497 visualized with IGV using the `Bar Chart`, `Autoscale` and windowing function `Mean`
498 options.

499 **Single-nucleotide variant analysis.** GIP was used to call SNVs using Freebayes (version
500 1.3.2) (`freebayesOPT` option, **Supplementary Data 2**) and filter its output
501 (`filterFreebayesOPT` option, **Supplementary Data 2**). Filters included the minimum allele
502 frequency (`--minFreq`), the minimum number of reads supporting the alternative alleles (`--
503 minAO`) and minimum mean mapping quality of for the reads supporting the reference (`--
504 minMQMR`) or the alternative allele (`--minMQM`). A higher number of reads supporting the
505 variants was requested for predictions positioned inside simple repeats of the same nucleotide
506 (homopolymers) (`--minAOhomopolymer`). The homopolymers were defined as the DNA
507 region spanning ±5 bases from the SNV (`--contextSpan 5`), with over 40% of identical
508 nucleotides (--homopolymerFreq 0.4 `). Further, GIP discarded SNVs with sequencing
509 coverage above or below 4 median absolute deviations (MADs) from the median chromosome
510 coverage (`--MADrange`). snpEff (version 4.3t) (70) was used to predict the impact of SNVs
511 on coding sequence. The predicted effects that GIP considered synonymous mutations are:
512 “synonymous_variant", "stop_retained_variant" and "start_retained". The predicted effects that
513 GIP considered non-synonymous mutations are: "missense_variant", "start_lost",
514 "stop_gained", "stop_lost" and "coding_sequence_variant". The phylogenetic tree was
515 computed by the giptools module `phylogeny` using IQtree2 (version 2.1.2) (71,72) with
516 options `--seqtype DNA --alrt 1000 -B 1000`. The Venn-diagram comparison considered the
517 strains QS0044_C, QS0001_C, QS0037_C, QS0016_C and SGH_358 that were sampled from
518 different locations in Ethiopia, respectively Habala, Badowacho, Arbaminch, Hawassa and
519 Jimma. The strains were selected to have comparable average genome coverage (23). To infer
520 the tree GIP considered the set of filtered SNV and adopted the IUPAC ambiguous notation for

521 the positions with allele frequency below 70%. The tree was visualized by giptools using the
522 ggtree R-package (73).

523 **Analysis of structural variants.** GIP was used to detect structural variants including insertions,
524 tandem duplications, deletions, inversions and break-end translocations with DELLY (version
525 0.8.7) (11). To reduce incorrect predictions the DELLY output was additionally filtered
526 (`filterDellyOPT` option, **Supplementary Data 2**). GIP discarded poor predictions with
527 DELLY label "LowQual" (`--rmLowQual`) and low median MAPQ score of mapping reads (`-
528 -minMAPQ`). SVs positioned in proximity of chromosome ends were removed (--
529 chrEndFilter) to limit false predictions caused by potential misassembled regions close to the
530 telomeric ends. To ease visualization and limit the analysis only to best supported SVs GIP
531 limited the output only to the top predictions (`--topHqPercentIns`, `--topHqPercentDel`, `--
532 topHqPercentDup` and `--topHqPercentInv`) based on the SV support score as in **Formula 1**,
533 where *DV*, *DR*, *RV*, and *RR* are respectively the number of high-quality variant pairs, reference
534 pairs, variant junction reads, and reference junction reads.

535 $$\frac{\boldsymbol{DV + RV}}{\boldsymbol{DV + RV + DR + RR}} * \mathbf{100}$$

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

537 The predicted structural variants were represented with Circos (version 0.69-9) (74).

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

540

541 **Acknowledgments:**

542 **Funding:** This study was supported by a seeding grant from the Institut Pasteur International
543 Department to the LeiSHield Consortium and the EU H2020 project LeiSHield-MATI - REP-
544 778298-1.

545 **Author contributions:** Conceptualization, G.B. and G.F.S.; Methodology, G.B. and G.F.S.;
546 Software, G.B.; Formal Analysis, G.B., and G.F.S.; Investigation, G.B. and G.F.S.; Writing –
547 Original Draft, G.B. and G.F.S.; Writing – Review & Editing, G.B. and G.F.S.; Supervision,
548 G.B. and G.F.S.; Funding Acquisition, G.F.S. and G.B.

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

550

551 ## Supplementary Tables:

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

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

554 **Table S3:** Sample information.

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

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

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

558 **Table S7:** Mapping statistics.

559

560 ## Supplementary Data:

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

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

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

564

565 **References:**

566 1. Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock,
567 G.M., Isaacs, F., Rozowsky, J. *et al.* (2016) The real cost of sequencing: scaling
568 computation to keep pace with data generation. *Genome Biol*, **17**, 53.
569 2. Pareek, C.S., Smoczynski, R. and Tretyn, A. (2011) Sequencing technologies and
570 genome sequencing. *J Appl Genet*, **52**, 413-435.
571 3. Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum,
572 A., Copeland, A., Huddleston, J., Eichler, E.E. *et al.* (2013) Nonhybrid, finished
573 microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*, **10**,
574 563-569.
575 4. Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A.,
576 Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A. *et al.* (2016) Phased
577 diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*, **13**,
578 1050-1054.
579 5. Reuter, J.A., Spacek, D.V. and Snyder, M.P. (2015) High-throughput sequencing
580 technologies. *Mol Cell*, **58**, 586-597.
581 6. Aurrecoechea, C., Barreto, A., Basenko, E.Y., Brestelli, J., Brunk, B.P., Cade, S.,
582 Crouch, K., Doherty, R., Falke, D., Fischer, S. *et al.* (2017) EuPathDB: the eukaryotic
583 pathogen genomics database resource. *Nucleic Acids Res*, **45**, D581-D591.
584 7. Dolled-Filhart, M.P., Lee, M., Jr., Ou-Yang, C.W., Haraksingh, R.R. and Lin, J.C.
585 (2013) Computational and bioinformatics frameworks for next-generation whole exome
586 and genome sequencing. *ScientificWorldJournal*, **2013**, 730210.
587 8. Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler,
588 B., Speicher, M.R., Zschocke, J. and Trajanoski, Z. (2014) A survey of tools for variant
589 analysis of next-generation genome sequencing data. *Brief Bioinform*, **15**, 256-278.
590 9. Garrison, E. and Marth, G. (2012), *arXiv preprint arXiv:1207.3907 [q-bio.GN]*.
591 10. Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M. (2011) CNVnator: an approach
592 to discover, genotype, and characterize typical and atypical CNVs from family and
593 population genome sequencing. *Genome Res*, **21**, 974-984.
594 11. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V. and Korbel, J.O. (2012),
595 *Bioinformatics*, Vol. 28, pp. i333-i339.
596 12. Hwang, K.B., Lee, I.H., Li, H., Won, D.G., Hernandez-Ferrer, C., Negron, J.A. and
597 Kong, S.W. (2019) Comparative analysis of whole-genome sequencing pipelines to
598 minimize false negative findings. *Sci Rep*, **9**, 3219.
599 13. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. and Notredame,
600 C. (2017) Nextflow enables reproducible computational workflows. *Nat Biotechnol*, **35**,
601 316-319.
602 14. Bussotti, G., Benkahla, A., Jeddi, F., Souiai, O., Aoun, K., Spath, G.F. and Bouratbine,
603 A. (2020) Nuclear and mitochondrial genome sequencing of North-African Leishmania
604 infantum isolates from cured and relapsed visceral leishmaniasis patients reveals
605 variations correlating with geography and phenotype. *Microb Genom*, **6**.
606 15. Bussotti, G., Gouzelou, E., Cortes Boite, M., Kherachi, I., Harrat, Z., Eddaikra, N.,
607 Mottram, J.C., Antoniou, M., Christodoulou, V., Bali, A. *et al.* (2018) Leishmania
608 Genome Dynamics during Environmental Adaptation Reveal Strain-Specific
609 Differences in Gene Copy Number Variation, Karyotype Instability, and Telomeric
610 Amplification. *MBio*, **9**.
611 16. Prieto Barja, P., Pescher, P., bussotti, g., Dumetz, F., Imamura, H., Kedra, D.,
612 Domagalska, M., Chaumeau, V., Himmelbauer, H., Pages, M. *et al.* (2017), *Nat Ecol*
613 *Evol*, Vol. 1, pp. 1961-1969.

614 17. Alvar, J., Velez, I.D., Bern, C., Herrero, M., Desjeux, P., Cano, J., Jannin, J., den Boer,
615 M. and Team, W.H.O.L.C. (2012) Leishmaniasis worldwide and global estimates of its
616 incidence. *PLoS One*, **7**, e35671.
617 18. Organization, W.H. (2018), WHO, Geneva. .
618 19. Price, R.N., Tjitra, E., Guerra, C.A., Yeung, S., White, N.J. and Anstey, N.M. (2007)
619 Vivax malaria: neglected and not benign. *Am J Trop Med Hyg*, **77**, 79-87.
620 20. Gething, P.W., Elyazar, I.R., Moyes, C.L., Smith, D.L., Battle, K.E., Guerra, C.A., Patil,
621 A.P., Tatem, A.J., Howes, R.E., Myers, M.F. *et al.* (2012) A long neglected world
622 malaria map: Plasmodium vivax endemicity in 2010. *PLoS Negl Trop Dis*, **6**, e1814.
623 21. Battle, K.E., Gething, P.W., Elyazar, I.R., Moyes, C.L., Sinka, M.E., Howes, R.E.,
624 Guerra, C.A., Price, R.N., Baird, K.J. and Hay, S.I. (2012) The global public health
625 significance of Plasmodium vivax. *Adv Parasitol*, **80**, 1-111.
626 22. Pearson, R.D., Amato, R., Auburn, S., Miotto, O., Almagro-Garcia, J., Amaratunga, C.,
627 Suon, S., Mao, S., Noviyanti, R., Trimarsanto, H. *et al.* (2016) Genomic analysis of
628 local variation and recent evolution in Plasmodium vivax. *Nat Genet*, **48**, 959-964.
629 23. Ford, A., Kepple, D., Abagero, B.R., Connors, J., Pearson, R., Auburn, S., Getachew,
630 S., Ford, C., Gunalan, K., Miller, L.H. *et al.* (2020) Whole genome sequencing of
631 Plasmodium vivax isolates reveals frequent sequence and structural polymorphisms in
632 erythrocyte binding genes. *PLoS Negl Trop Dis*, **14**, e0008234.
633 24. Singh, V., Gupta, P. and Pande, V. (2014) Revisiting the multigene families:
634 Plasmodium var and vir genes. *J Vector Borne Dis*, **51**, 75-81.
635 25. Rayner, J.C., Tran, T.M., Corredor, V., Huber, C.S., Barnwell, J.W. and Galinski, M.R.
636 (2005) Dramatic difference in diversity between Plasmodium falciparum and
637 Plasmodium vivax reticulocyte binding-like genes. *Am J Trop Med Hyg*, **72**, 666-674.
638 26. Rahul, C.N., Shiva Krishna, K., Pawar, A.P., Bai, M., Kumar, V., Phadke, S. and
639 Rajesh, V. (2014) Genetic and structural characterization of PvSERA4: potential
640 implication as therapeutic target for Plasmodium vivax malaria. *J Biomol Struct Dyn*,
641 **32**, 580-590.
642 27. Rahul, C.N., Shiva Krishna, K., Meera, M., Phadke, S. and Rajesh, V. (2015)
643 Plasmodium vivax: N-terminal diversity in the blood stage SERA genes from Indian
644 isolates. *Blood Cells Mol Dis*, **55**, 30-35.
645 28. Luo, Z., Sullivan, S.A. and Carlton, J.M. (2015) The biology of Plasmodium vivax
646 explored through genomics. *Ann N Y Acad Sci*, **1342**, 53-61.
647 29. Lin, J.T., Patel, J.C., Kharabora, O., Sattabongkot, J., Muth, S., Ubalee, R., Schuster,
648 A.L., Rogers, W.O., Wongsrichanalai, C. and Juliano, J.J. (2013) Plasmodium vivax
649 isolates from Cambodia and Thailand show high genetic complexity and distinct
650 patterns of P. vivax multidrug resistance gene 1 (pvmdr1) polymorphisms. *Am J Trop*
651 *Med Hyg*, **88**, 1116-1123.
652 30. Gunalan, K., Niangaly, A., Thera, M.A., Doumbo, O.K. and Miller, L.H. (2018)
653 Plasmodium vivax Infections of Duffy-Negative Erythrocytes: Historically Undetected
654 or a Recent Adaptation? *Trends Parasitol*, **34**, 420-429.
655 31. Costa, G.L., Amaral, L.C., Fontes, C.J.F., Carvalho, L.H., de Brito, C.F.A. and de
656 Sousa, T.N. (2017) Assessment of copy number variation in genes related to drug
657 resistance in Plasmodium vivax and Plasmodium falciparum isolates from the Brazilian
658 Amazon and a systematic review of the literature. *Malar J*, **16**, 152.
659 32. Cornejo, O.E., Fisher, D. and Escalante, A.A. (2014) Genome-wide patterns of genetic
660 polymorphism and signatures of selection in Plasmodium vivax. *Genome Biol Evol*, **7**,
661 106-119.
662 33. Chen, E., Salinas, N.D., Huang, Y., Ntumngia, F., Plasencia, M.D., Gross, M.L.,
663 Adams, J.H. and Tolia, N.H. (2016) Broadly neutralizing epitopes in the Plasmodium

664 vivax vaccine candidate Duffy Binding Protein. *Proc Natl Acad Sci U S A*, **113**, 6277-
665 6282.
666 34. Zolan, M.E. (1995) Chromosome-length polymorphism in fungi. *Microbiol Rev*, **59**,
667 686-698.
668 35. Suzuki, T., Nishibayashi, S., Kuroiwa, T., Kanbe, T. and Tanaka, K. (1982) Variance
669 of ploidy in Candida albicans. *J Bacteriol*, **152**, 893-896.
670 36. Sionov, E., Chang, Y.C. and Kwon-Chung, K.J. (2013) Azole heteroresistance in
671 Cryptococcus neoformans: emergence of resistant clones with chromosomal disomy in
672 the mouse brain during fluconazole treatment. *Antimicrob Agents Chemother*, **57**, 5127-
673 5130.
674 37. Shin, J.H., Chae, M.J., Song, J.W., Jung, S.I., Cho, D., Kee, S.J., Kim, S.H., Shin, M.G.,
675 Suh, S.P. and Ryang, D.W. (2007) Changes in karyotype and azole susceptibility of
676 sequential bloodstream isolates from patients with Candida glabrata candidemia. *J Clin*
677 *Microbiol*, **45**, 2385-2391.
678 38. Selmecki, A., Forche, A. and Berman, J. (2010) Genomic plasticity of the human fungal
679 pathogen Candida albicans. *Eukaryot Cell*, **9**, 991-1008.
680 39. Magee, B.B. and Magee, P.T. (2000) Induction of mating in Candida albicans by
681 construction of MTLa and MTLalpha strains. *Science*, **289**, 310-313.
682 40. Gerstein, A.C., Fu, M.S., Mukaremera, L., Li, Z., Ormerod, K.L., Fraser, J.A., Berman,
683 J. and Nielsen, K. (2015) Polyploid titan cells produce haploid and aneuploid progeny
684 to promote stress adaptation. *mBio*, **6**, e01340-01315.
685 41. Croll, D. and McDonald, B.A. (2012) The accessory genome as a cradle for adaptive
686 evolution in pathogens. *PLoS Pathog*, **8**, e1002608.
687 42. Chibana, H., Beckerman, J.L. and Magee, P.T. (2000) Fine-resolution physical mapping
688 of genomic diversity in Candida albicans. *Genome Res*, **10**, 1865-1877.
689 43. Bravo Ruiz, G., Ross, Z.K., Holmes, E., Schelenz, S., Gow, N.A.R. and Lorenz, A.
690 (2019) Rapid and extensive karyotype diversification in haploid clinical Candida auris
691 isolates. *Curr Genet*, **65**, 1217-1228.
692 44. Brown, G.D. and Netea, M.G. (2012) Exciting developments in the immunology of
693 fungal infections. *Cell Host Microbe*, **11**, 422-424.
694 45. Pfaller, M.A., Diekema, D.J., Turnidge, J.D., Castanheira, M. and Jones, R.N. (2019)
695 Twenty Years of the SENTRY Antifungal Surveillance Program: Results for Candida
696 Species From 1997-2016. *Open Forum Infect Dis*, **6**, S79-S94.
697 46. Pfaller, M.A., Castanheira, M., Messer, S.A., Moet, G.J. and Jones, R.N. (2010)
698 Variation in Candida spp. distribution and antifungal resistance rates among
699 bloodstream infection isolates by patient age: report from the SENTRY Antimicrobial
700 Surveillance Program (2008-2009). *Diagn Microbiol Infect Dis*, **68**, 278-283.
701 47. Todd, R.T. and Selmecki, A. (2020) Expandable and reversible copy number
702 amplification drives rapid adaptation to antifungal drugs. *Elife*, **9**.
703 48. Mount, H.O., Revie, N.M., Todd, R.T., Anstett, K., Collins, C., Costanzo, M., Boone,
704 C., Robbins, N., Selmecki, A. and Cowen, L.E. (2018) Global analysis of genetic
705 circuitry and adaptive mechanisms enabling resistance to the azole antifungal drugs.
706 *PLoS Genet*, **14**, e1007319.
707 49. Hirakawa, M.P., Martinez, D.A., Sakthikumar, S., Anderson, M.Z., Berlin, A., Gujja,
708 S., Zeng, Q., Zisson, E., Wang, J.M., Greenberg, J.M. *et al.* (2015) Genetic and
709 phenotypic intra-species variation in Candida albicans. *Genome Res*, **25**, 413-425.
710 50. Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardimci, G.G.,
711 Chakraborty, A., Bann, D.V., Wang, Y. *et al.* (2018) Integrative detection and analysis
712 of structural variation in cancer genomes. *Nat Genet*, **50**, 1388-1398.
713 51. Yao, Y. and Dai, W. (2014) Genomic Instability and Cancer. *J Carcinog Mutagen*, **5**.

714 52. Abbey, D.A., Funt, J., Lurie-Weinberger, M.N., Thompson, D.A., Regev, A., Myers,
715 C.L. and Berman, J. (2014) YMAP: a pipeline for visualization of copy number
716 variation and loss of heterozygosity in eukaryotic pathogens. *Genome Med*, **6**, 100.
717 53. Bogaerts, B., Delcourt, T., Soetaert, K., Boarbi, S., Ceyssens, P.J., Winand, R., Van
718 Braekel, J., De Keersmaecker, S.C.J., Roosens, N.H.C., Marchal, K. *et al.* (2021) A
719 bioinformatics WGS workflow for clinical Mycobacterium tuberculosis complex isolate
720 analysis, validated using a reference collection extensively characterized with
721 conventional methods and in silico approaches. *J Clin Microbiol*.
722 54. Bogaerts, B., Winand, R., Fu, Q., Van Braekel, J., Ceyssens, P.J., Mattheus, W.,
723 Bertrand, S., De Keersmaecker, S.C.J., Roosens, N.H.C. and Vanneste, K. (2019)
724 Validation of a Bioinformatics Workflow for Routine Analysis of Whole-Genome
725 Sequencing Data and Related Challenges for Pathogen Typing in a European National
726 Reference Center: Neisseria meningitidis as a Proof-of-Concept. *Front Microbiol*, **10**,
727 362.
728 55. Ellison, M.A., Walker, J.L., Ropp, P.J., Durrant, J.D. and Arndt, K.M. (2020)
729 MutantHuntWGS: A Pipeline for Identifying Saccharomyces cerevisiae Mutations. *G3*
730 *(Bethesda)*, **10**, 3009-3014.
731 56. Quijada, N.M., Rodriguez-Lazaro, D., Eiros, J.M. and Hernandez, M. (2019) TORMES:
732 an automated pipeline for whole bacterial genome analysis. *Bioinformatics*, **35**, 4207-
733 4212.
734 57. Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2013), *Brief. Bioinformatics*, Vol.
735 14, pp. 178-192.
736 58. Leinonen, R., Sugawara, H., Shumway, M. and Collaboration, I.N.S.D. (2011), *Nucleic*
737 *Acids Res.*, Vol. 39, pp. D19-21.
738 59. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y.,
739 Cleland, I., Faruque, N., Goodgame, N., Gibson, R. *et al.* (2011) The European
740 Nucleotide Archive. *Nucleic Acids Res*, **39**, D28-31.
741 60. Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C.,
742 Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T. *et al.* (2016) ENCODE data at the
743 ENCODE portal. *Nucleic Acids Res*, **44**, D726-732.
744 61. Howe, K.L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J.,
745 Alvarez-Jarreta, J., Barba, M., Bolser, D.M., Cambell, L. *et al.* (2020) Ensembl
746 Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res*, **48**,
747 D689-D695.
748 62. Arnaud, M.B., Costanzo, M.C., Skrzypek, M.S., Binkley, G., Lane, C., Miyasato, S.R.
749 and Sherlock, G. (2005) The Candida Genome Database (CGD), a community resource
750 for Candida albicans gene and protein information. *Nucleic Acids Res*, **33**, D358-363.
751 63. Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M.J., Gajria, B., Grant, G.R., Ginsburg, H.,
752 Gupta, D., Kissinger, J.C., Labo, P. *et al.* (2003) PlasmoDB: the Plasmodium genome
753 resource. A database integrating experimental and computational data. *Nucleic Acids*
754 *Res*, **31**, 212-215.
755 64. Girgis, H.Z. (2015) Red: an intelligent, rapid, accurate tool for detecting repeats de-
756 novo on the genomic scale. *BMC Bioinformatics*, **16**, 227.
757 65. Li, H. (2013), *arXiv:1303.3997v1 [q-bio.GN]*.
758 66. Li, H. and Durbin, R. (2009), *Bioinformatics*. 2009/05/20 ed, Vol. 25, pp. 1754-1760.
759 67. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
760 Abecasis, G., Durbin, R. and Subgroup, G.P.D.P. (2009), *Bioinformatics*, Vol. 25, pp.
761 2078-2079.
762 68. Li, W. and Godzik, A. (2006), *Bioinformatics*, Vol. 22, pp. 1658-1659.

69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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.