QuasiFlow: a bioinformatic tool for genetic variability analysis from next generation sequencing data

QuasiFlow for genetic variability from NGS data

Pedro Seoane[1,2,3,†], Luis Díaz-Martínez[4,5,†] , Enrique Viguera[5], M. Gonzalo Claros[1,2,3,4,*] and Ana Grande-Pérez[4,5]

[1] Department of Molecular Biology and Biochemistry, Facultad de Ciencias, Universidad de Málaga, 29071 Malaga, Spain.

[2] Institute of Biomedical Research in Málaga (IBIMA), 29010 Málaga, Spain

[3] Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Instituto de Salud Carlos III, Málaga, Spain

[4] Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora", Universidad de Málaga-Consejo Superior de Investigaciones Científicas (IHSM-UMA-CSIC), Campus de Teatinos, 29071 Malaga, Spain.

[5] Area de Genética, Facultad de Ciencias, Universidad de Málaga, 29071 Malaga, Spain

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

* **Corresponding author**: Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM-UMA-CSIC), Ampliación del Campus de Teatinos, Universidad de Málaga, 29071 Malaga, Spain. Tel: +34 952 137 284; Fax: +34 952 132 041; Email: claros@uma.es or gonzalo.claros@ ihsm.uma-csic.es

ABSTRACT

Populations of RNA and ssDNA viruses within their hosts contain a heterogeneous collection of variant genomes known as quasispecies. Large variability in mitochondrial DNA has also been found within the same organism, drawing an interesting parallel between the two situations. The advent of next-generation sequencing technologies facilitated studying genetic variation, but many open-source bioinformatic tools have to be combined in a non-trivial approach. Here it is presented QuasiFlow, a workflow based on well-stablished software that extracts reliable mutations and recombinations, even at low frequencies (~10^{-4}), provided that at least 250 million nucleotides are analysed. Accurate prediction of mutations and recombinations has been demonstrated with synthetic reads and with in vitro rolling-circle amplification of a plant geminivirus. An in-depth analysis of viral quasispecies was performed and QuasiFlow revealed the coexistence in the plant of three virus genomes and distinct recombinations between some of them. Human mitochondrial variants were also investigated and high level of heteroplasmy (75%) was confirmed, and the relation between low-frequency heteroplasmy (0.1- 0.2%) and some human diseases, regardless of sex, was established. Hence, we propose that QuasiFlow may find use with known and emerging viruses to reveal evolutionary jumps and co-infections, with mitochondrial DNA to detect relevant heteroplasmy would otherwise be elusive, or even in other population studies such as those considering single cell sequencing.

INTRODUCTION

RNA and ssDNA viruses, including SARS-CoV-2, exist as highly variable and heterogeneous collections of mutated and recombinant genomes known as virus quasispecies (Lauring and Andino 2010; Domingo et al. 2021). The term 'quasispecies' means that a viral genome is indeed a weighted average of a large number of different but

51 related individual genotypes, called haplotypes (Domingo and Perales 2019; Domingo et al.
52 2021). Haplotypes differ not only in point mutations, since it is well known that many viruses
53 can recombine resulting in the shuffling of beneficious and deleterious variations that can be
54 separated and tested as individual units in new combinations. High variability has also been
55 found in mitochondrial DNA (mtDNA) that has been related to human mtDNA disorders that
56 affect one of 4300 individuals (Gorman et al. 2015). The frequency in which the mtDNA
57 variants (point mutations and indels) are found determines that they are considered
58 heteroplasmy, if different variants of the mtDNA coexist in the same cell, tissue, organ or
59 organism, or homoplasmy, in case all mtDNA copies share the mutation. Heteroplasmy and
60 homoplasmy have been related to disease and ageing in humans (Stewart and Chinnery
61 2015).

62 Studies on virus quasiespecies usually required laborious and expensive laboratory efforts.
63 Fortunately, the advent of next-generation sequencing (NGS) technologies enabled the fast
64 and cheap detection of any minor variant within a huge population of genomes (Bull et al.
65 2016; Posada-Cespedes et al. 2017) and discern all single nucleotide variations (SNVs) in any
66 quasispecies to define a haplotype (viral variant) without the need of culture or cloning prior
67 to sequencing (Zukurov et al. 2016). However, analysing NGS data can be an awkward
68 endeavour starting from the available open-source, cutting-edge software with incompatible
69 formats to study recombination, variant calling and haplotypes (Cacciabue et al. 2020).
70 Additionally, quasispecies studies must deal with sequencing errors, whose frequency lays
71 close to the viral replication error rates (10^{-3}-10^{-5} errors per nucleotide).

72 To fill this gap, here we describe QuasiFlow, a bioinformatic tool based on well stablished
73 software to study genetic variability from NGS data that is able to distinguishing real
74 variations from sequencing errors. Accurate and precise detection of low-frequency
75 mutations, genetic variability and recombination is achieved in viral quasispecies and other

76 highly variable genomes like mtDNA. Its companion QuasiComparer facilitates the
77 interpreation of multiple analyses on comparable samples. Both pipelines together can
78 extract the complete genetic variability provided that more than 250 million nucleotides were
79 sequenced.

80

81 **RESULTS**

82 ## Stringent filtering in QuasiFlow provides a reliable detection of low-frequency
83 genetic variations

84 The performance of QuasiFlow workflow (Fig. 1) was investigated using four datasets
85 (mut_rec, mut_REC, MUT_rec and MUT_REC, where 'MUT/mut' means mutations,
86 'REC/rec' means recombination, lower case indicates absence, and upper case indicates
87 presence) of synthetic reads mimicking an Illumina sequencing project (see Materials and
88 Methods for details). Read rejection rate was 0.72 in all four groups (Fig. 2A,
89 RejectedReads) due to the strict requeriments in the read-preprocessing step to avoid the
90 confusion of sequencing errors with mutation events. Consequently, the useful reads (Fig.
91 2A, ReferenceReads) presented a recovering index of 0.27 for recombinant free datasets
92 (mut_rec and MUT_rec) with a slight decrease to ~0.25 where only recombinations
93 (mut_REC) or recombinations together with mutations (MUT_REC) were present,
94 respectively. Mapping rate of useful reads was nearly 1.00 for the samples without mutations
95 (Fig 2A, MappedReads) with a slight decrease to 0.98 when reads contain artificial
96 mutations, indicating that the accumulation of sequence changes causes the alignment
97 rejection by the mapper with highly strict parameters. Concerning recombination, untangled
98 datasets (mut_rec and MUT_rec) showed an expected 0.0 rate for true and false
99 recombination events (Fig 2A, TrueRECs and FalseRECs) while most (0.86) recombination
100 events were correctly predicted in both mut_REC and MUT_REC. The striking value of false

recombinants in the MUT_rec sample is due to one single unexpected recombination event. The few cases of false recombinations detected in MUT_REC (Fig. 2A, FalseRECs) were manually inspected revealing that they were true recombinants produced 4-5 nt apart from the expected nucleotide, indicating that stringent validations of QuasiFlow detects true recombinations, even if they are not at the expected position and should not affect when real-world data are used. Concerning mutations, it results that unmutated datasets (mut_rec and mut_REC) contained no mutation per read (Fig. 2A, TrueMUTs), while mutated datasets (MUT_rec and MUT_REC) were 0.863 and 0.831, respectively. Taken together, Fig. 2A demonstrated that QuasiFlow can predict very accurately the mutation or recombination frequency in a population using only the subset of useful reads.

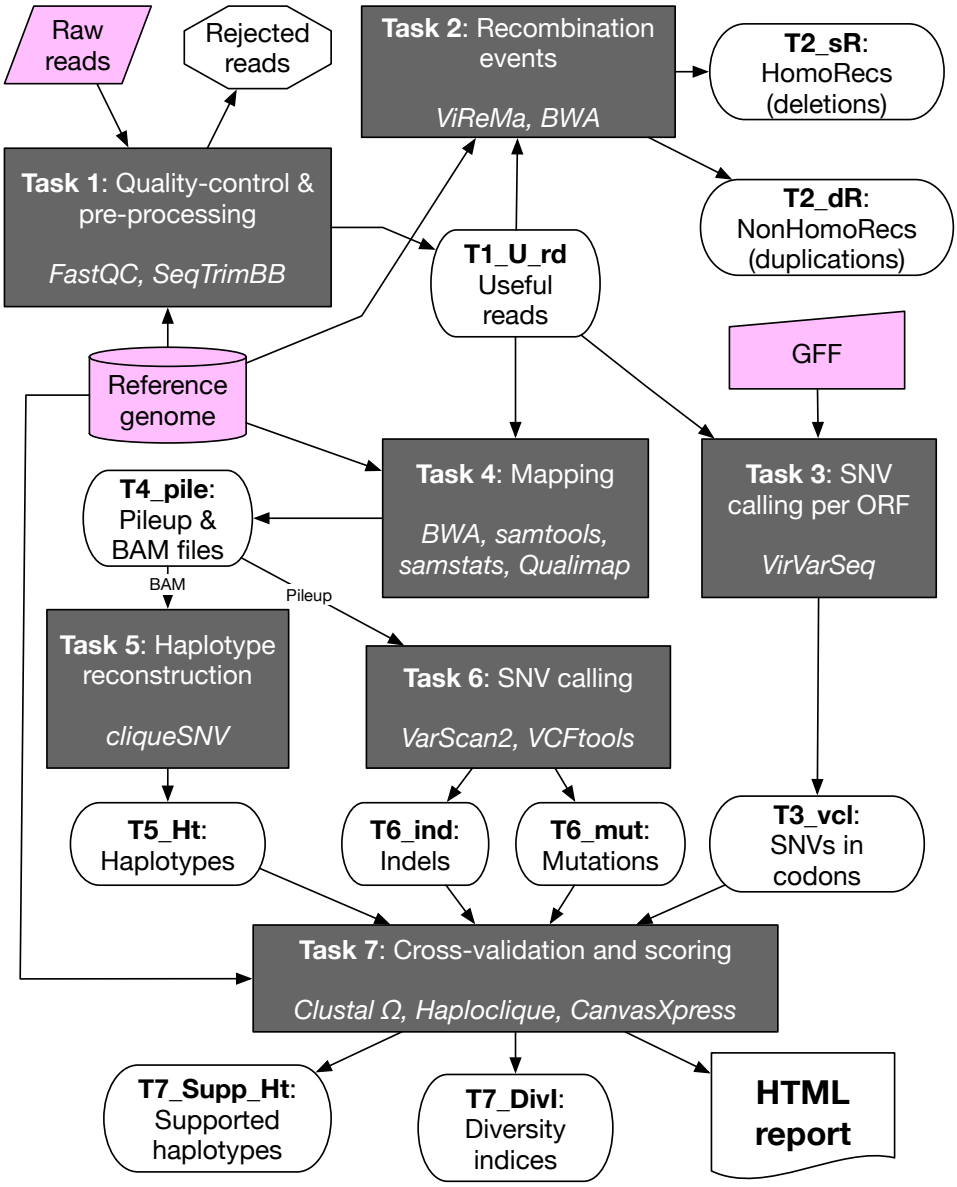

Figure 1: Conceptual representation of the QuasiFlow workflow for quantifying genetic variability using high-throughput sequencing reads. The seven tasks, described in detail in Materials and Methods, are grey rectangles, where the relevant bioinformatic tools are indicated in italics. Magenta boxes refer to QuasiFlow input files. White rounded boxes correspond to input/output data recoverable when QuasiFlow finishes; their title containing the task number will facilitate its citation throughout the text. Solid arrows indicate information flow.

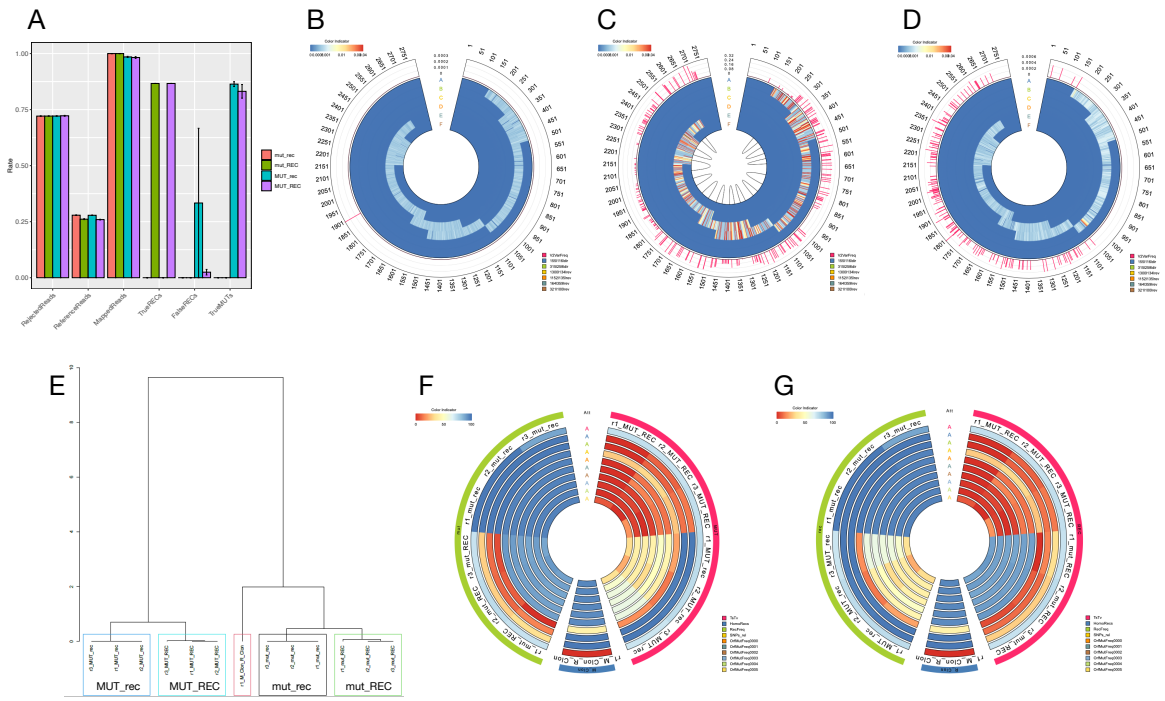

Figure 2: Quasiflow and QuasiComparer performance. (**A**) Benchmarking QuasiFlow with synthetic reads. Bars represent means of three replicates with their standard error. Colours are indicated in the legend and correspond to the four synthetic datasets. "RejectedReads" are reads removed due to quality, complexity or length issues. "ReferenceReads" are reads devoid of artefacts (useful reads) and therefore suitable for genetic variance analyses. "MappedReads" are useful reads mapped onto reference sequence of TYLCV-Mld. "TrueRECs" represent recombination events that were predicted at the expected position. "False RECs" gather those recombination events that were not detected at the expected position. "True MUTs" correspond to mutation events predicted at the expected position. "Haplotypes" are those retained after task 7 cross-validation (T1_Supp_Ht in Fig. 1). Variation profile of replicates from synthetic mut_rec (**B**) and MUT_REC (**C**), and RCA-amplified TYLCV-Mld clone (**D**), where the outer track (V2VarFreq) corresponds to the frequency of variants called by Varscan2, the other six inner tracks are VirVarSeq mutation frequencies determined in the six TYLCV-Mld frames), and the internal links represent origin and target positions of the recombinations identified by ViReMa. (**E**) Hierarchical clustering after PCA of three replicates of each synthetic dataset and the TYLCV-Mld, where the different replicates are clustered together and surrounded by color boxes, and the single TYLCV-Mld sample is boxed in red. Analysis of variance (ANOVA) for the mutation (**F**) and recombination (**G**) rates in the same samples of panel E. The outer ring groups samples by factor: in green are synthetic samples that lack of inserted mutations (mut) or recombinations (rec), in pink are gathered all samples with

136 inserted mutations (MUT) or recombinations (REC), and in blue is the TYLCV-Mld infectious clone sample. Inner rings
137 represent the most significant variables, where the colour spectrum represents the relative value of the variable in the
138 range of 0 (blue) to 100% (red), as indicated in the colour indicator. The QuasiFlow HTML report for mut_rec (*1-*
139 *mut_rec.html*) and MUT_REC (*2-MUT_REC.html*) synthetic reads, the QuasiComparer complete comparative report for
140 all synthetic reads (*3-Comparative_Report_SyntheticReads.html*), as well as for the control of TYCLV-Mld replicated in
141 vitro (*4-Control_TYLCV.html*) can be inspected in Supplementary File 1.

142

143 The whole variation profile was then investigated (Fig. 2B-D). As expected, the control group
144 of mut_rec presented no mutations and no recombination events, being consistent with a
145 very low frequency of genetic variability (Fig. 2B). For high variation profiles such as
146 MUT_REC (Fig 2C), 261 new mutations with frequencies ranging from 0.31 to 1.72×10^{-4},
147 and 13 recombination events were detected. This illustrates that QuasiFlow can find both
148 low- and high-frequent mutation and recombination events. A test using a real-world dataset
149 from a RCA-amplified infectious clone of the begomovirus TYLCV-Mld (tomato yellow leaf
150 curl virus, mild strain), where low mutation frequency is expected as well as absence of
151 recombination, 72 mutations with frequencies ranging from 4.2×10^{-4} to 2.2×10^{-4} were
152 detected and no recombination events were identified (Fig. 2D). The apparent hotspot
153 detected from the nucleotide 1110 to 1128 for the ORF so called 315-258-dir (Fig 2D, track
154 B) is due to the quasi-dimer (1,7-mer) nature of infectious clone compared to the reference
155 genome, being the region where digestion and fusion was produced and is not identical in all
156 virus haplotypes. Since the RCA-amplified TYLCV-Mld (Fig 2D) presents values closer to
157 mut_rec (Fig 2B) than to MUT_REC (Fig 2C), it can be suggested QuasiFlow was able to
158 detect the low amount of variation introduced by bacterial cloning and RCA manipulation,
159 and that this low level of variability would have a negligible impact on genetic variation. This
160 also supports that mutation and recombination events and the resulting haplotypes provided

161 by QuasiFlow should be considered reliable, derived from real biological events and not from
162 experimental artefacts.

163 Quasispecies studies usually require comparison of several samples. This has been
164 facilitated by another workflow called QuasiComparer. When samples mut_rec, mut_REC,
165 MUT_rec, MUT_REC and TYLCV-Mld where compared (*3-*
166 *Comparative_Report_SyntheticReads.html* in Supplementary File 1), most (67.54%) of the
167 total variance in PCA seems to depend on mutations, which explains why the TYLCV-Mld
168 sample clusters with mut_rec and mut_REC datasets (Fig. 2E), in agreement with resultis in
169 Figs. 2B and 2D. The same behaviour can be inferred after ANOVA taking into account both
170 mutations (Fig. 2F) or recombinations (Fig. 2E) as factors, the same 4 variables being highly
171 significant for both factors ($P < 0.005$). Since the TYLCV-Mld sample shows again a pattern
172 close to datasets without recombination events and without mutations, it is suggested that
173 QuasiComparer easily illustrates the classifation of known samples as expected, and serves
174 to group samples based on their biological origin, being a great help to interpret results.

175 Quasiespecies reconstruction of original and recombinant geminiviruses in a 176 tomato plant

177 QuasiFlow performance detecting recombinant viruses was illustrated using geminivirus.
178 MiSeq data from a rolling circle amplification (RCA) product of a TYLCV-Mld agroinfecting
179 *Solanum lycopersicum* plant was analysed. Viral genetic variation was evaluated with the
180 incidence, abundance and functional diversity indexes described in methods presented in
181 Table S1 of Supplementary File 2. These values are characteristic of highly diverse viral
182 populations (Gregori et al. 2016), since (i) the mutation frequency was 1.008×10^{-2} mut/nt,
183 two orders of magnitude higher than that of other geminiviruses (~10^{-4} mut/nt) (Isnard et al.
184 1998; Sanz et al. 2000; Ge et al. 2007; Duffy and Holmes 2008); and (ii) the Shannon index
185 reached 0.6. The consensus sequence differed by 32 SNVs from the reference sequence

186 (TYLCV-Mld) and its mutation frequency per nucleotide position clustered into two well
187 differentiated regions (Figure 3A). The first, from positions 1 to 1000 and includes 21
188 changes with a mutation frequency of 0.85 ± 0.03. A second region, from position 1201 to
189 2791, included changes with mutation frequencies of 0.95 ± 0.03 and 0.69 ± 0.02. Between
190 the two regions, at positions 1000-1200 a hostpot was found with a mutation frequency of
191 0.45. A total of 31 haplotypes (Hp) were reconstructed (Figure 3B), including one (Hp1)
192 identical to the reference sequence. Hp12, Hp27 and Hp31 accounted for 80% of the total.
193 Hp12 showed an identity of 98% with TYLCV-Mld and represented 40% of the total.
194 Interestingly, blast analysis indicated that Hp12 was identical to Almería strain of TYLCV
195 (TYLCV-Almería, AJ489258.1) from nt 1-1054 whereas from nt 1055-2791 was identical to
196 Malaga strain (TYLCMaV, AF271234.1). The recombinant nature of Hp12 was confirmed by
197 RDP analysis (Fig. 3C) and suggested a triple infection within the agro-infected plant. *In*
198 *silico* results were confirmed by restriction analysis (Fig. 3D) of the RCA product. *KpnI*
199 releases the monomer of TYLCMaV and *ClaI* cuts TYLCV-Almería into two fragments and
200 linearises TYLCV-Mld. This assay revealed the existence of the three viruses in the tomato
201 plant. In addition, *bona fide* recombinant genomes between TYLCV-Almería and TYLCMaV
202 (Fig. 3E) were amplified with specific primers. Finally, since haplotype reconstruction
203 showed a high frequency of genomes carrying changes that belonged to the TYLCMaV
204 sequence, the relative abundance of each virus was investigated. The analysis of the
205 coverage of each parental virus (Figure 3F), using an artificial reference sequence with the
206 genome of the three viruses, showed that most of the variants belonged to TYLCMaV,
207 followed by TYLCV-Almería, while TYLCV-Mld was almost undetectable, suggesting a
208 fitness advantage of TYLCMaV.

209

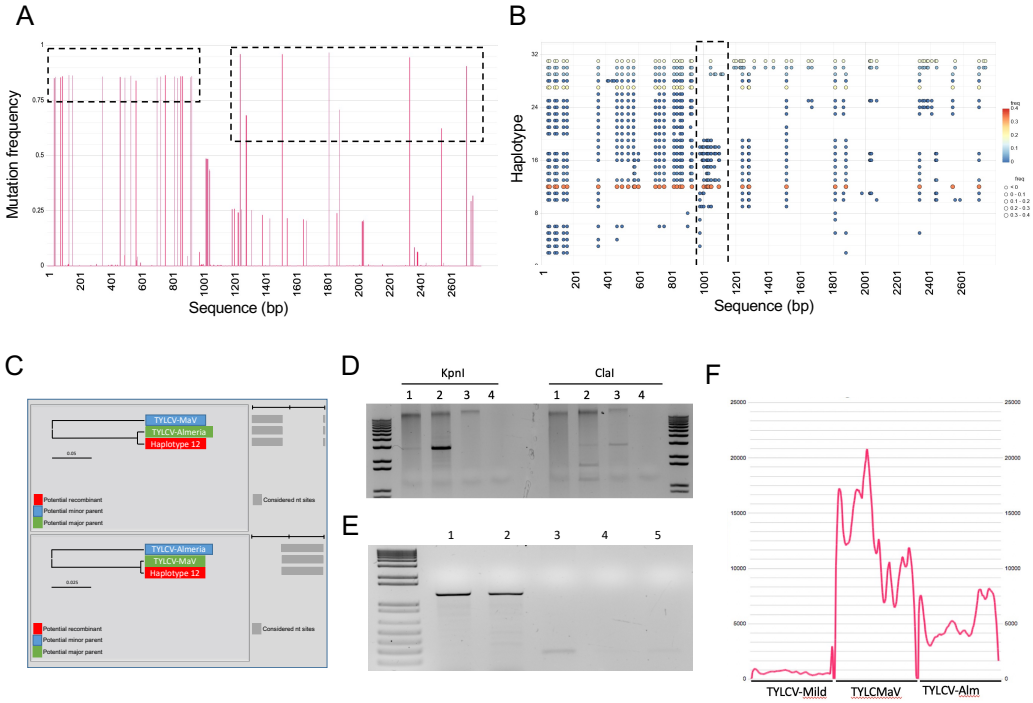

Figure 3. Variation analysis of a TYLCV-Mld population after infection of tomato plants. **(A)** Nucleotide mutation frequency profile, where black dashed lines indicate regions that group SNPs of similar frequency, one containing mutations with two different frequencies. Only mutations with a frequency over 0.5 were considered for the consensus. (**B)** Haplotype reconstruction of the TYLCV population. Circles depict mutated positions on each haplotype respect to the reference sequence TYLCV-Mld. Coloring indicates haplotype frequency. Of the 31 different reconstructed haplotypes, haplotype 1 (Hp1) is identical to the consensus sequence. The dashed squares frame regions with high variability among haplotypes. (**C)** RDP4 analysis showing Haplotype 12 as a recombinant with the first part of its sequence (gray rectangles, upper panel) of TYLCV-Almería and the second half of TYLCMaV (lower panel). **(D)** Restriction analysis of the RCA product from the tomato sample: 1/100 dilution (lanes 1) and a 1/10 dilution (lanes 2); positive control, TYLCV-Mld monoinfected tomato (lanes 3); negative control (lanes 4). (**E)** Identification of a *bona-fide* recombinant between TYLCMaV and TYLCV-Almería in the TYLCV-Mld infected tomato by PCR. A region of 1.4 kb of the recombinant was amplified in a 1/100 dilution (lane1), undiluted DNA (lane 2); positive controls: tomato monoinfected with TYLCV-Mld (lane 3), TYLCMaV (lane 4) and TYLCV-Almería-(lane 5). (**F)** Nucleotide coverage of the infected tomato sample using a TYLCV-Mld, TYLCMaV and TYLCV-Almería chimaera as a reference

225 sequence. The QuasiComparer complete comparative report for tomato infected plants with TYCLV (*5-Tomato_TYLCV-*
226 *invected.html*) can be inspected in Supplementary File 1.

227

228 Study of mitochondrial variants to establish a relationship between low-
229 frequency heteroplasmies and human diseases

230 Within eucariotic cells, a population of mitochondrial genomes exists, each genome
231 harboring variations called variants of mitochondrial DNA (mtDNA). We analysed MiSeq data
232 of 47 human mtDNA samples from blood, lymphoblastoid cell lines and saliva. QuasiFlow
233 detected 10 400 variants of which 14.5% (1 508 variants) were homoplasmies and 85.5%
234 (8 892 variants) corresponded to heteroplasmies. Nucleotide changes (Supplementary File
235 2, Table S2) were biased towards transitions A → G and T → C in all samples, followed by C
236 → T and G → A. No transversions were found in 12 of the 47 samples. Looking at
237 heteroplasmies, a remarkable 80% had very low frequency, that is, between 0.1% and 0.2%,
238 and the rest (20%) had a frequency no higher than 2% (Fig 4A). These results suggest that
239 mtDNA populations in these samples maintain a very low level (below 0.3%) of
240 heteroplasmies while homoplasmies were very frequent (greater than 98%) with a stricking
241 absence of intermediate cases. When the relationship between mtDNA variants and disease
242 was inspected, 451 variants were associated with different diseases (Supplementary File 2,
243 Table S3). Interestingly, more than 75% of these mtDNA variants were heteroplasmies.

244 Using the data retrieved from MITOMAP a network was constructed for each sample
245 illustrating the relationship between mtDNA variants per ORF and disease. An example is
246 shown in Figure 4B with biological replicate 1 of a female lymphoblastoid sample. Note that
247 26 homoplasmies were related to brain syndromes such as Alzheimer's, Parkinson's,
248 deafness, bipolar disorder, multiple sclerosis, Leber hereditary optic neuropathy (LHON) or
249 cerebral pseudoatrophy. Remakably, 310 heteroplasmies, of which 300 had a frequency

below 1%, were related to the same brain syndromes. These findings highlight the importance of detecting low frequency mtDNA heteroplasmies for disease association.

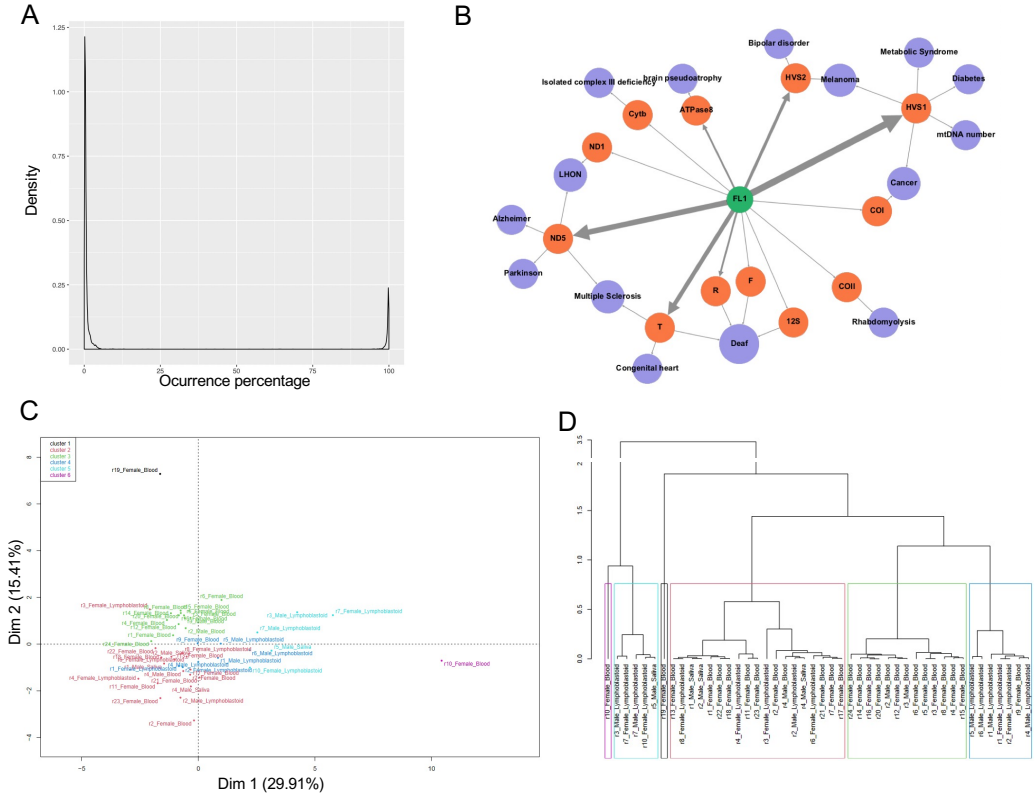

Figure 4. Variation analysis of mtDNA populations from different human tissues. **(A)** Average distribution of mtDNA variants in the 47 samples analyzed showing very low frequency of heteroplasmies (below 0.2%) and homoplasmies, but without intermediate frequency variants. **(B)** Mutation network of sample 1 of the lymphoblastoid cell line (FL1, green node). ORFs with heteroplasmies (orange nodules) are linked to diseases (purple nodules) that have been associated with these mutations. The thickness of the line is proportional to the frequency of the heteroplasmy. **(C)** Sorting the different mtDNA samples by PCA. The colors correspond to the different clusters formed from the analysis of the first four dimensions of the PCA. (**D)** Cladogram showing the five clusters from the PCA of panel D. Blood samples group in the red and black clusters, linfoblastoid samples in the blue and red clusters, and saliva samples are dispersed throughout the three main clusters. The QuasiComparer complete comparative report for the mtDNAs (*6-Comparative_Report_Mitochondria.html*) can be inspected in Supplementary File 1.

265 The impact of sex and tissue factors were inspected using QuasiComparer. No significant
266 differences (P < 0.05; Table S4 in Supplementary File 2) were detected among the 47
267 samples for any of the factors, but differences become evident by PCA for tissue factor (Fig.
268 4C), that explained 48% of the total variance. In fact, hierarchical clustering grouped
269 samples in four clusters (Fig. 4D) two of them by tissue, while no association between sex
270 and mtDNA variants was found. The blue- and cyan-boxed clusters in Fig. 4D contained
271 mainly lymphoblastoid samples and the red-boxed cluster was enriched with blood samples.
272 Differences between these clusters were found significant (*t*-Student, $P < 0.05$).

273

274 Genetic variability requires sequencing of 250 million nucleotides in TYLCV

275 To present a robust and accurate assessment of mutation and recombination frequencies,
276 the QuasiFlow/QuasiComparer analysis must rely on the whole genetic variability, and this is
277 clearly dependent on the number of reads for the low frequent SNVs. Hence, the number of
278 mutations detected using MiSeq or HiSeq platforms on the same sample was compared
279 depending on the total number of nucleotides sequenced. A sigmoid behavior was observed
280 using both platforms (Figure S1 in Supplementary File 2), reaching a plateau (that is,
281 mutations at all frequencies were detected) when nearly 250 million nucleotides were
282 sequenced. It is reassuring that both sequencing platforms require the same number of
283 nucleotides sequenced to saturate the analysis, corresponding to ~1.5 millions of true, useful
284 2 × 300 MiSeq reads (derived from ~3 million raw reads) per sample, or ~0.75 million of true
285 2 × 100 HiSeq reads (derived from ~1.25 millions of raw reads) per sample. However, MiSeq
286 datasets reach the plateau in ~1 815 mutations (Figure S1, yellow lines), while HiSeq
287 saturates at ~580 to ~800 (Figure S1, blue lines). Only the sequencing platform can be
288 invoked as source of differences since the same DNA was sequenced in both cases and the
289 same bioinformatic pipeline run. It seems that MiSeq reactions would be more prone to

290 sequencing variations than HiSeq reactions (in fact, we were verbally told by Illumina staff
291 that the sequencing kit for 2 × 300 nt in MiSeq produces more sequencing errors than other
292 kits). Therefore, being conservative, HiSeq could be considered true variations, while MiSeq
293 is suspected to overestimate the genetic variation. The arrival of NovaSeq and the new
294 sequencing platforms might benefit the performance of QuasiFlow. In any case, it seems
295 reasonable to avoid sequencing mixtures from different platforms to measure genetic
296 variability.

297

298 **DISCUSSION**

299 The study of highly variable genetic systems was initiated in virology (Domingo and Schuster
300 2016) and now extends to other biological systems such as cancer (Dagogo-Jack and Shaw
301 2018), bacteria (Chan et al. 2013), mitochondria (Aryaman et al. 2019), or prions (Mahal et
302 al. 2010). These analyses rely on accurate determination of mutations for the development
303 of effective therapies. The aim of QuasiFlow (Fig. 1) is to provide sensitive, reliable, and
304 robust analysis of complex genetic variability (including both low frequency mutations and
305 recombinations) with a highly-stringent set of well-known bioinformatics tools glued with in-
306 house code. Results are easy to interpret when compared using QuasiComparer
307 (Supplementary File 1) as illustrated with PCA and ANOVA plots grouping samples by their
308 biological origin and not by the experimental process used to produce them (Figs. 2E-F, 3F
309 and 4B-D). The capabilities of QuasiFlow/QuasiComparer couple have been demonstrated
310 using synthetic sets of data (Fig. 2) and applied to a real-world data from plant viral (Fig. 3)
311 and human mitochondrial (Fig. 4) populations.

312 The advent of the NGS (next-generation sequencing) technology has dramatically boosted
313 the intrahost viral diversity studies, even if there is still no consensus about the most
314 appropriate technology, the most appropriate software or the minimal read quality for

315 SNV/recombination validations (McCrone and Lauring 2016). QuasiFlow applied to NGS
316 data enabled, using a coverage >10 000× when a threshold of 20× was widely used in
317 literature, the detection of low frequency SNVs and recombinations with a small rate of false
318 positive detection. Emerging viruses such as coronavirus and geminiviruses are threatening
319 humans and crops, respectively, due to their high frequency of mutation and recombination
320 (Martin et al. 2011). Although some studies show that recombination is not the main
321 phenomenon involved in the evolution of these viruses (Duffy et al. 2008), it is important to
322 take this process into account due to its potential to generate large evolutionary jumps in
323 very small time frames. Thus, recombinants can lead to host range expansion and increases
324 in patogenicity. The analysis of the TYLCV-Mld infective clone in tomato plants in Fig. 3A-B
325 demonstrated a striking triple infection (Fig. 3C), revealing that studies in greater depth of
326 this type of co-infections could shed light on the reasons for fitness differences between
327 such similar virus species.

328 The reliable analyses of the genetic heterogeneity of human mtDNA (Fig. 4) confirms the
329 wide scope of QuasiFlow. In first instance, QuasiFlow confirmed the extreme genetic
330 variation within mtDNA (Stewart and Chinnery 2021) with specific variant patterns in different
331 tissues or nuclear backgrounds (i.e. sex), becoming a highly accurate and sensitive tool for
332 the analysis of mitochondrial diseases. Interestingly, low-frequency mtDNA heteroplasmies
333 were detected and then related to diseases that might have been elusive to detection by
334 other means. Moreover, one can also envisage the utility of QuasiFlow/QuasiComparer in
335 single-cell studies (Huang et al. 2021). In the case of tumours where genetic diversity is high
336 from cell to cell due to a high mutation rate and such tumors can be considered as a
337 collection of subclones, the study of this diversity can be facilitated using QuasiFlow: genetic
338 diversity between patients or within a tumor can lead to pinpoint actionable driver mutations
339 to design targeted therapies (Mroz and Rocco 2017; Baslan et al. 2020).

340 Accurately identifying mutations and variants among the thousands of virus that arise during
341 an emerging viral infection can help understand how changes in tropism and jumps between
342 host species occur. This is of particular importance in the present COVID-19 pandemic as
343 fast haplotype reconstruction achieved by QuasiFlow can be a useful tool to detect variants
344 of unknown origin such as Omicron that may be present as minority variants in mutant
345 spectra.

346 **METHODS**

347 ## Workflow availability and requirements

348 QuasiFlow (Fig. 1), a workflow developed with the workflow manager AutoFlow (Seoane et
349 al. 2016), takes as input raw, paired-end read data generated by a NGS platform (mainly
350 Illumina, although other are allowed). Its design took advantage of computational resources
351 available and the following dependencies: Art (Huang et al. 2012), SeqTrimBB, FastQC
352 (Andrews 2010), BWA (Li and Durbin 2009), samtools (Li et al. 2009), Samstat (Li et al.
353 2009), Qualimap (Okonechnikov et al. 2015), ViReMa (Routh and Johnson 2014), VirVarSeq
354 (Verbist et al. 2015) CliqueSNV (Knyazev et al. 2021), Varscan2 (Koboldt et al. 2012),
355 VCFtools (Danecek et al. 2011), Haploclique (Töpfer et al. 2014), Clustal Ω (Sievers and
356 Higgins 2014), CanvasXpress (https://github.com/neuhausi/canvasXpress/) and FactoMineR
357 (Husson et al. 2009). It was coded and executed in Picasso supercomputer (3456 cores and
358 8.4 TB of RAM with Slurm queue system on SUSE Linux Enterprise Server 11SP2) at the
359 University of Málaga. It requires (1) a reference genome in FASTA format (the begomovirus
360 TYLCV or human mitochondrial DNA in this study), (2) a GFF file containing all open reading
361 frame (ORF) coordinates, and (3) the circular or linear nature of the reference genome.
362 Several samples can be run simultaneously using a customizable daemon script provided
363 with the code that enables (1) executing QuasiFlow with all samples, (2) monitorizing
364 QuasiFlow executions, (3) gathering results from all samples, and (4) comparative sample

365 analysis using QuasiComparer. The workflow procceeds through the following seven tasks
366 (Fig. 1):

367 ***Task 1: Read quality control and pre-processing***

368 Raw reads were quality controlled using FastQC, and pre-processed using SeqTrimBB to
369 remove low quality bases, vectors and adapters. SeqTrimBB default parameters were used,
370 except for the following: 1) minimum phred base mean quality was set to 26 to avoid
371 sequencing errors to be confounded with low frequency mutations and reduce the false
372 positive ratio on ulterior analyses (Morelli et al. 2013; Welkers et al. 2014); 2) the minimum
373 read length was increased to 100 nt to remove shorter reads confounding haplotype
374 reconstruction (Zagordi et al. 2011). Viral reads were extracted from host and contaminant
375 reads using the reference genome. This more reliable but reduced number reads (Fig 1,
376 T1_U_rd) guaranteed soundness and robustness in further steps (Daly et al. 2015).

377 ***Task 2: Recombination events***

378 The useful reads (T1_U_rd) were then treated as single reads (even though they come from
379 paired-end reads) with ViReMa (Routh and Johnson 2014) using default parameters, excepting
380 the aligner, which was configured to BWA (Li and Durbin 2009) instead of the default Bowtie
381 since BWA is widely used for recombination studies. To avoid the identification of indels or
382 misalignments as recombination events due to confounding reads, the seed minimum length
383 was set to 55, the allowed mismatches was set to 1, and the microindels size was set to 55.
384 Spurious recombination events with a frequency lower than 1×10^{-3} were removed to
385 increase reliability in predicted recombination events. Additionally, recombination events
386 located between the start and end positions of a circular reference sequence were
387 discarded. Recombination events are classified by ViReMa as HomoRecs (Fig 1, T2_sR) for
388 those resulting in a sequence deletion, or NonHomoRec (Fig 1, T2_dR) when the resulting
389 sequence presents a duplication.

Task 3: SNV calling per ORF

The SNV calling process per ORF was based on VirVarSeq (Verbist et al. 2015) using useful reads in T1_U_rd and the GFF file provided at launch containing the ORF coordinates. VirVarSeq default parameters were used, but the minimum average phred quality is consistently set to 26 to assure that high quality reads are used. Variant codons were then converted to nucleotide positions (Fig1, T3_vcl) by assigning the SNV to the three nucleotides belonging to each codon.

Task 4: Mapping

Useful reads T1_U_rd (Fig. 1) were mapped on the genome reference to prepare data for subsequent haplotype reconstruction (Task 5) and variant calling (Task 6). To ensure high quality alignments, accurate mapping scoring (preventing false variant calling), and compatibility with further tools, mapping using BWA was launched with the following command:

```
bwa mem -B 20 -A 30 -O 30 -E 3 reference_sequence pair1.fastq pair2.fastq
| samtools view -bS -q 30 -F 4 > file.bam
```

A quality control analysis of the mapping is then performed on the BAM file with both Samstat (Li et al. 2009) and Qualimap (Okonechnikov et al. 2015) tools. Finally, by means of mpileup (Li et al. 2009), a first variant dataset is provided using the following command:

```
samtools mpileup -BQ 26 -q 30 -d 10000000 -f reference_sequence file.bam
```

where the –BQ flag filters by base quality (set to 26 in order to be coherent with the previous tasks). The –q flag, used in the current and in the previous samtools command, skips the alignments with a map quality of 30 or less. Finally, the –d flag changes the default maximum read depth from 8000 to 10^7 following the samtools documentation ([CSL STYLE ERROR: reference with no printed form.]). Then, the pileup for one BAM file is saved (Fig 1, T4_pile).

Task 5: Haplotype reconstruction

The haplotype reconstruction relied on the recent CliqueSNV (Knyazev et al. 2021) that seems to outperform other haplotyping methods, even with low coverage samples where SNVs are at low frequencies. CliqueSNV opened each BAM file in T4_pile (Fig. 1) to identify groups of linked SNVs and remove sequencing errors. It resulted in a graph where SNVs are nodes and edges connect linked SNVs. Merging cliques in that graph identified haplotypes that were then saved (Fig 1, T5_Ht) together with their frequency.

Task 6: SNV calling

SNV calling was based on VarScan2 (Verbist et al. 2015). Mutations were called using the pileup file from Task 4 (T4_pile) with the following command

```
varscan mpileup2snp file.mpileup --min-coverage 10000 --min-reads2 10 --
min-var-freq 7e-6 --min-avg-qual 30
```

Indels were obtained with the same command but substituting `mpileup2snp` for `mpileup2indel`. The `min-avg-qual` removed low quality alignments, ensuring high quality calling SNVs, and the `min-coverage` and `min-reads` parameters selected only SNVs supported by 10 reads with a minimum read depth of 10 000. The `min-var-freq` parameter allows detection of real SNVs with frequencies as low as 7×10^{-6}. The resulting files of mutations (Fig 1, T6_mut) and indels (Fig 1, T6_ind) were processed with the command vcf-stats from the VCFtools suite (Danecek et al. 2011) to obtain parameters such as variant counts, transition/transversion frequencies, as well as computing the frequencies of each type of base substitution that will be used below in QuasiComparer.

Task 7: Cross-validation and scoring

Taking in mind that haplotype reconstruction was poor when sequence divergence is low, failing to recover rare haplotypes (Schirmer et al. 2014), QuasiFlow was designed with a final task to remove unreliable haplotypes and increase the consistency with previously

440 detected SNVs. To do so, haplotypes in T5_Ht were aligned to the genome reference using
441 Clustal Ω, and the variable positions were cross-validated with previously obtained SNVs
442 (T6_mut and T6_ind). Variability not supported by a SNV was edited, and haplotypes whose
443 variations were all edited were then rejected as due to sequencing errors or untrusted
444 variations. The resulting SNV-supported haplotypes (Fig. 1, T7_Supp_Ht) were realigned
445 again and haplotype frequency was used to score diversity indices described by Gregori et
446 al (Gregori et al. 2016), such as normalized Shannon, Simpson, Gini-Simpson, Hill numbers,
447 sample nucleotide diversity, etc (Fig. 1, T7_DivI; see Supplementary Table S5 for a detailed
448 description). In addition, haplotypes and SNVs were represented as a network where each
449 haplotype was connected with all its SNVs to facilitate the interpretation of the genetic
450 complexity and heterogeneity of the sample (see reports in Supplementary File 1). To
451 facilitate result interpretation and visual user inspection an HTML report based on the
452 javascript scientific library CanvasXpress (https://github.com/neuhausi/canvasXpress/)
453 including the values for the parameters of Supplementary Table S5 as well as interactive
454 plots and graphs is provided. It should be noted that most diversity indices (but Shannon)
455 are provided for illustrative purposes since they are depending on the selected haplotypes.

456 ***QuasiComparer***

457 The variability analysis usually implies comparison of several samples or conditions. Since
458 QuasiFlow can only manage one sequencing library (sample) at a time, inter-sample
459 comparisons required the development of a dedicated comparative system called
460 QuasiComparer. This analysis can be accomplished automatically by the same daemon
461 script mentioned above provided that a description file with the following data is supplied: 1)
462 the sample name, 2) a replicate identifier that allows to group the replicates that belongs to
463 the same sample, and 3) the experimental group to which the sample is assigned,
464 depending on the user experimental design (e.g., an experimental factor such as days post
465 infection with three sample replicates taken at 7 days and other three replicates at 28 days).

466 QuasiComparer uses the parameters indicated in Supplementary Table S5 to perform
467 correlation, ANOVA and principal component analysis (PCA) (Seoane et al. 2018) (Fig. 2B-
468 G) to provide a new, interactive, compartive report in HTML (HTML reports included in
469 Supplementary File 1) incluiding linear correlation plots, impact of each parameter on the
470 experimental factors defined by the user by SNVs and by recombination events. The
471 ANOVA is presented as a table and some circular plots (see Fig. 2F-G, one by each
472 experimental factor described by the user) to illustrate the weight of each significant variable
473 on each sample. The PCA also includes tables with information about its first two
474 dimensions displaying the weight of each variable in PCA dimensions as well as plots (Fig.
475 4C), a PDF containing correlation analysis between the variables, and hierarchical clustering
476 (Fig. 2E and 4D). All this information is intended to help users to see which samples have a
477 different behavior in comparison to the rest, and correlate this behavior with the experimental
478 factors used in the study.

479 ***Synthetic and experimental datasets for testing***

480 To test QuasiFlow and QuasiComparer performance, twelve independent synthetic read
481 datasets were generated using Art (Huang et al. 2012) with the virus reference sequence of
482 tomato yellow leaf curl virus Mild (TYLCV-Mld, AC# AF071228). The daemon was loaded
483 with two parametres: 1) the presence of mutations and 2) the presence of recombinations,
484 and called an in-house script that generates a FASTA file with mutated, recombinant and
485 original virus sequences. Recombinations were generated making 30 recombinant events in
486 a single sequence of TYLCV-Mld separated by 60 nt. The process was repeated until the
487 recombinant sequences represented 20% of the total. To generate mutations, random
488 changes were included in 10% of the nucleotide pool of TYLCV-Mld sequences. The FASTA
489 file generated was then provided to Art, setting the read coverage to 300 000×, the
490 maximum read length in 250 nt and a fragment size of 300 nt, with a standard deviation of
491 10 nt and using the MiSeq version 3.0 built-in error profile.

492 Four (mut_rec, mut_REC, MUT_rec and MUT_REC) synthetic sample datasets were
493 generated by triplicate. "mut_rec" corresponded to reads where no mutation and no
494 recombination were introduced and was considered as control group. "mut_REC" cotains
495 reads where only recombinations were introduced and were considered as control for
496 recombination. "MUT_rec" contains reads where only mutations were introduced, which
497 makes it suitable as a control for mutation. Finally, the dataset "MUT_REC" contains reads
498 where both mutations and recombinations were introduced and was considered the testing
499 group.

500 As real-world dataset, an infectious clone of TYLCV-Mld was sequenced as a control before
501 its passage through a plant to test the ability to detect mutations or recombination events
502 introduced during high-throughput sequencing. This dataset was sequenced with a paired-
503 end layout 2 × 300 nt resulting in 1 106 300 reads. After pre-processing and filtering to keep
504 only TYLCV-Mld useful reads, this number decreased to 218 884.

505 ***Samples, virus source and viral inoculation***

506 Thirty day-old *Arabidopsis thaliana* Col-0 plants were inoculated with *Agrobacterium*
507 *tumefaciens* LBA4404 containing the infectious clone (p1.7SP72/9) of TYLCV-Mld isolate
508 [ES:72:97] (AC# AF071228). The construction harbours 1.7 genome-length of TYLCV-Mild
509 (genome region from positions 161 to 1026 nt is only once) flanked by two intergenic (IR)
510 regions that allow the excision of the viral monomer. Each *A. thaliana* plant was inoculated
511 by the stem-puncture method using 20 µl of *A. tumefaciens* liquid culture adjusted to an
512 OD_{600} of 1.00. After agroinoculation, plants were maintained for 28 days in a growth
513 chamber (22 °C during the day and 20 °C at night, 70% relative humidity, with an 8-16 h
514 day-night photoperiod of 8-16 h).

515 Dr. Enrique Moriones (IHSM "La Mayora") kindly provided the tomato (*Solanum*
516 *lycopersicum)* plants agroinoculated with TYLCV-Mld with disease symptoms and that were
517 positive for the infection of begomoviruses of the TYLC-disease complex.

518 ***DNA extraction, virus amplification, cloning, and sequencing***

519 DNA from *A. thaliana* Col-0 and *S. lycopersicum* plants were extracted following the
520 Edwards' procedure (Edwards et al. 1991) with some modifications. From each sample,
521 150 mg of fresh tissue were ground with liquid nitrogen, and 400 µl of extraction buffer
522 [200 mM Tris-HCl, 200 mM NaCl, 25 mM EDTA, SDS 0.5 % (w/v), pH 7.5, 0.1 % β-
523 mercaptoethanol] added. After vortexing at 37 °C for 2 min, extracts were centrifuged at
524 12 000 × *g* at room temperature for 5 min. The supernatants were transferred to a new tube
525 and mixed with 1 volume of phenol:chloroform (1:1). Following centrifugation at 12 000 × *g*
526 at room temperature for 5 min, the aqueous phase was transferred to a new tube and mixed
527 with 1 volume of isopropanol. After 2 min at room temperature, nucleic acids were recovered
528 by centrifugation at 12 000 × *g* at room temperature for 5 min. The pellet was washed with
529 70% ethanol, dried and resuspended in 200 µl of RTE buffer (50 mM Tris-HCl pH 8.0,
530 10 mM EDTA pH 8.0, 0.1 % RNase A) to a concentration of 130 ng/µl by incubation at 65 °C
531 for 10 min.

532 DNA samples from tomato and *A. thaliana* agroinfected plants were treated with DpnI
533 endonuclease to remove *A. tumefaciens* viral DNA remnants but not viral DNA replicated in
534 the plants. Endonuclease-treated DNA was purified with StrataClean Resin kit (Agilent
535 Technologies), following the instructions given by the manufacturer.

536 Verification of the existence of the bioinformatically detected geminivirus species of tomato
537 and agroinoculated *A. thaliana* plants was done by rolling-circle amplification (RCA) and
538 restriction analysis. For this, 130 ng of DNA from young non-inoculated leaves were
539 amplified for 18 h with the Illustra TempliPhi kit (GE Healthcare, UK) following

540 manufacturer's instructions. Amplified DNA was digested with single cutter (in viral genome)
541 restriction enzyme BamHI to obtain the full-length DNA (2.7 kb). Restriction enzymes KpnI
542 and ClaI were used to distinguish between TYLCV-Mld, TYLCV-Almería (AJ489258.1) and
543 TYLCMaV (AF271234.1) in DNA extracts from tomato plants.

544 The presence of recombinant genomes between TYLCV-Almería and TYLCMaV in extracts
545 form tomato plants was verified by PCR using reverse primer MAV-1271_1289-RW (5'-
546 TTTTCTCAACTTCCGCATC) that binds to TYLCMaV sequence and forward primer IL-
547 41_55-FW (5'-CCCACGAGGGTTCC) that binds to TYLCV-Almería sequence. RDP4
548 program (Martin et al. 2015) was used to analyse the recombination event of the most
549 frequent haplotype. Digestions and PCR results were analysed by agarose (Agarose SPI,
550 Duchefa) gel electrophoresis 1 % (w/v) with TBE buffer.

551 To generate DNA for Illumina sequencing viral DNA from total DNA plants extracts and the
552 infectious clone of TYLCV-Mld [ES:72:97] were amplified by RCA. To avoid creating a
553 population bottleneck during the RCA amplification step two microliters of the plant DNA
554 extract (260 ng total) were used as template. Replicates of each amplified sample were
555 pooled to reduce the accumulation of random errors due to Phi29 Polymerase. Phi29
556 polymerase has an associated 3′-5′ exonuclease "proofreading" activity and has a reported
557 error rate of 3-5 $\times$ 10^{-6}. The contribution of the Phi29 polymerase error during RCA is about
558 100-fold lower than the observed mutation frequency of begomoviruses (around 10^{-4}
559 mutations per nucleotide or higher).

560 MiSeq and HiSeq sequencing was performed at Macrogen (Seoul, South Korea). For MiSeq,
561 equimolar sequencing libraries of tomato and nine *A. thaliana* samples were done with
562 Truseq DNA PCR-free (550 bp insert) kit and sequenced on an Illumina MiSeq as 2 $\times$ 300 bp
563 reads. For HiSeq, libraries were done for three *A. thaliana* samples with Truseq DNA PCR-
564 free (350 bp insert) kit and sequenced on an Illumina HiSeq2000 as 2 $\times$ 100 bp reads. The

565 number of reads belonging to geminiviruses that exceeded the initial quality filter was
566 293 668, with an average coverage of 14 772× per nucleotide.

567 ***Mitocondrial sequences for genetic variability analyses***

568 Human mtDNA sequences BioProject PRJEB5005 contained blood, lymphoblastoid cell
569 lines and saliva samples from males and females. The set comprised 47 libraries of 2 × 300
570 bp paired-end reads obtained with the Illumina MiSeq platform with a number of reads
571 between 1.38×10^5 and 4.32×10^5. Samples from the same tissue and sex were considered
572 biological replicates. The Revised Cambridge Reference Sequence ("rCRS") of 16 569
573 nucleotides (AC# NC_012920) was used as standard reference for human mtDNA. Since
574 the mean coverage was about 1500×, variants with a frequency below 9×10^{-4} (0.09 %)
575 were filtered out by means of QuasiFlow parameters `--min-var-freq to 9x10e-4` and
576 `the --min-coverage to 1000` in Task 6. The biological significance of mitochondrial
577 variants was obtained using MITOMAP database (Lott et al. 2013)
578 (https://www.mitomap.org). To assess tissue origin differences between the lymphoblastoid
579 cell lines and blood a *t*-Student was performed using FactoMineR for each axis of the PCA.

580

581 **DATA ACCESS**

582 **Data**: Sequencing reads can be found at BioProject PRJNA823046. Human mtDNA
583 sequences were obtained from BioProject PRJEB5005.

584 **Code**: QuasiFlow, QuasiComparer and the daemon script (daemon.sh) that controls their
585 execution can be downloaded from *https://bitbucket.org/seoanezonjic/quasiflow/src/master/*.
586 The resulting reports are attached as Supplementary File 1.

587

COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGEMENTS

This work was supported by Consejería de Economía, Innovación y Ciencia, Junta de Andalucía with assistance from the European Regional Development Fund (ERDF) and the European Social Fund (ESF) [grant to Research Groups BIO-264 and BIO-267; P10-CVI-6561 and UMA18-FEDERJA178 to A.G.-P., E.V.M; UMA20-FEDERJA-029 to MGC; and a predoctoral fellowship to L.D.-M. A.G-P. acknowledges support of Plan Propio de Investigación y Transferencia of the University of Málaga. Open access charges was funded by University of Málaga. The authors also thankfully acknowledge the computer resources and the technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga.

AUTHOR CONTRIBUTIONS

PS designed QuasiFlow and QuasiComparer and constructed the synthetic data. LDM analysed the TYCLV and mitochondrial datasets. EV interpreted the mitochondrial data. MGC supervised the algorithm implementation and execution, and wrote the manuscript. AGP conceived the research, supervised the viral interpretations, and wrote the manuscript.

SUPPLEMENTARY INFORMATION

Supplementary File 1 (DOI 10.6084/m9.figshare.19518976): Compressed zip file containing the following HTML reports

611 • *1-mut_rec.html*: QuasiFlow report for mut_rec synthetic reads.
612 • *2-MUT_REC.html*: QuasiFlow report for MUT_REC synthetic reads.
613 • *3-Comparative_Report_SyntheticReads.html*: QuasiComparer comparative report for
614 all synthetic reads
615 • *4-Control_TYLCV.html*: QuasiComparer comparative report for the control of TYCLV-
616 Mld replicated in vitro.
617 • *5-Tomato_TYLCV-invected.html*: QuasiFlow report for TYCLV samples from tomato-
618 infected cells.
619 • *6-Comparative_Report_Mitochondria.html:* QuasiComparer comparative report for
620 mtDNAs.

621 **Supplementary File 2** (DOI 10.6084/m9.figshare.19519003): Complementary data related
622 to:

623 • *Table S1:* Evaluation of TYLCV-Mld genetic variation.
624 • *Table S2:* Nucleotide changes detected in the mtDNA samples used to validate
625 QuasiFlow analyses.
626 • *Table S3:* mtDNA variants detected that have a known relation with a human
627 disease.
628 • *Table S4:* ANOVA test for the 47 samples considering the different factors evaluating
629 genetic variation provided by QuasiFlow
630 • *Figure S1:* Plot of the number of SNPs detected depending on the number of
631 nucleotides analysed. Nine MiSeq datasets (yellow lines; ranging from 115 to 335
632 million reads) and three HiSeq dataset (blue lines, ranging from 500 to 501.5 million
633 reads) were used. Every dataset was randomly subset in bulks of 50 000 read folds
634 and plotted against the number of SNPs detected in it by Varscan2.
635 • *Table S5:* Variables calculated during QuasiFlow execution

636

637 **REFERENCES**

638 Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. *Babraham*
639 *Bioinformatics*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

640 Aryaman J, Johnston IG, Jones NS. 2019. Mitochondrial Heterogeneity. *Front Genet* **9**: 718.

641 Baslan T, Kendall J, Volyanskyy K, McNamara K, Cox H, D'Italia S, Ambrosio F, Riggs M, Rodgers L,
642 Leotta A, et al. 2020. Novel insights into breast cancer copy number genetic heterogeneity revealed
643 by single-cell genome sequencing. *eLife* **9**: e51480.

644 Bull RA, Eltahla AA, Rodrigo C, Koekkoek SM, Walker M, Pirozyan MR, Betz-Stablein B, Toepfer A,
645 Laird M, Oh S, et al. 2016. A method for near full-length amplification and sequencing for six hepatitis
646 C virus genotypes. *BMC Genomics* **17**: 247.

647 Cacciabue M, Currá A, Carrillo E, König G, Gismondi MI. 2020. A beginner's guide for FMDV
648 quasispecies analysis: sub-consensus variant detection and haplotype reconstruction using next-
649 generation sequencing. *Briefings in Bioinformatics* **21**: 1766–1775.

650 Chan K-H, To KKW, Chan BWK, Li CPY, Chiu SS, Yuen K-Y, Ho P-L. 2013. Comparison of
651 Pyrosequencing, Sanger Sequencing, and Melting Curve Analysis for Detection of Low-Frequency
652 Macrolide-Resistant Mycoplasma pneumoniae Quasispecies in Respiratory Specimens. *J Clin*
653 *Microbiol* **51**: 2592–2598.

654 Dagogo-Jack I, Shaw AT. 2018. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev*
655 *Clin Oncol* **15**: 81–94.

656 Daly GM, Leggett RM, Rowe W, Stubbs S, Wilkinson M, Ramirez-Gonzalez RH, Caccamo M, Bernal
657 W, Heeney JL. 2015. Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery
658 Using Next Generation Sequencing Data. *PloS one* **10**: e0129059.

659 Danecek P, Auton A, Abecasis G, Albers C a, Banks E, DePristo M a, Handsaker RE, Lunter G,

660 Marth GT, Sherry ST, et al. 2011. The variant call format and VCF tools. *Bioinformatics (Oxford,*
661 *England)* **27**: 2156–8.

662 Domingo E, García-Crespo C, Perales C. 2021. Historical Perspective on the Discovery of the
663 Quasispecies Concept. *Annu Rev Virol* **8**: 51–72.

664 Domingo E, Perales C. 2019. Viral quasispecies. *PLOS Genetics* **15**: e1008271.

665 Domingo E, Schuster P. 2016. *Quasispecies: from theory to experimental systems*. Springer.

666 Duffy S, Holmes EC. 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-
667 stranded DNA begomovirus tomato yellow leaf curl virus. *J Virol* **82**: 957–965.

668 Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and
669 determinants. *Nat Rev Genet* **9**: 267–276.

670 Edwards K, Johnstone C, Thompson C. 1991. A simple and rapid method for the preparation of plant
671 genomic DNA for PCR analysis. *Nucleic Acids Res* **19**: 1349–1349.

672 Ge L, Zhang J, Zhou X, Li H. 2007. Genetic structure and population variability of tomato yellow leaf
673 curl China virus. *Journal of virology* **81**: 5902–5907.

674 Gorman GS, Schaefer AM, Ng Y, Gomez N, Blakely EL, Alston CL, Feeney C, Horvath R, Yu-Wai-
675 Man P, Chinnery PF, et al. 2015. Prevalence of nuclear and mitochondrial DNA mutations related to
676 adult mitochondrial disease. *Annals of Neurology* **77**: 753–759.

677 Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. 2016. Viral quasispecies
678 complexity measures. *Virology* **493**: 227–237.

679 Huang RH, Wang LX, He J, Gao W. 2021. Application and prospects of single cell sequencing in
680 tumors. *Biomark Res* **9**: 88.

681 Huang W, Li L, Myers JR, Marth GT. 2012. ART: A next-generation sequencing read simulator.
682 *Bioinformatics* **28**: 593–594.

683 Husson F, Josse J, Pagès L. 2009. FactoMineR, An R package dedicated to exploratory multivariate

684 analysis. *Journal of Statistical Software* **25**: 1–18.

685 Isnard M, Granier M, Frutos R, Reynaud B, Peterschmitt M. 1998. Quasispecies nature of three maize
686 streak virus isolates obtained through different modes of selection from a population used to assess
687 response to infection of maize cultivars. *Journal of General Virology* **79**: 3091–3099.

688 Knyazev S, Tsyvina V, Shankar A, Melnyk A, Artyomenko A, Malygina T, Porozov YB, Campbell EM,
689 Switzer WM, Skums P, et al. 2021. Accurate assembly of minority viral haplotypes from next-
690 generation sequencing through efficient noise reduction. *Nucleic Acids Research* **49**: e102–e102.

691 Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson
692 RK. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome
693 sequencing. *Genome Research* **22**: 568–576.

694 Lauring AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathogens*
695 **6**: 1–8.

696 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
697 *Bioinformatics* **25**: 1754–1760.

698 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009.
699 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

700 Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, Procaccio V, Wallace DC. 2013.
701 mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr Protoc Bioinformatics* **44**: 1.23.1-
702 26.

703 Mahal SP, Browning S, Li J, Suponitsky-Kroyter I, Weissmann C. 2010. Transfer of a prion strain to
704 different hosts leads to emergence of strain variants. *Proc Natl Acad Sci USA* **107**: 22653–22658.

705 Martin DP, Lefeuvre P, Varsani A, Hoareau M, Semegni J-Y, Dijoux B, Vincent C, Reynaud B, Lett J-
706 M. 2011. Complex Recombination Patterns Arising during Geminivirus Coinfections Preserve and
707 Demarcate Biologically Important Intra-Genome Interaction Networks ed. P.D. Nagy. *PLoS Pathog* **7**:
708 e1002203.

709 Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of
710 recombination patterns in virus genomes. *Virus Evolution* **1**: vev003.

711 McCrone JT, Lauring AS. 2016. Measurements of Intrahost Viral Diversity Are Extremely Sensitive to
712 Systematic Errors in Variant Calling ed. T.S. Dermody. *J Virol* **90**: 6884–6895.

713 Morelli MJ, Wright CF, Knowles NJ, Juleff N, Paton DJ, King DP, Haydon DT. 2013. Evolution of foot-
714 and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts.
715 *Veterinary Research* **44**: 12.

716 Mroz EA, Rocco JW. 2017. The challenges of tumor genetic diversity. *Cancer* **123**: 917–927.

717 Okonechnikov K, Conesa A, García-Alcalde F. 2015. Qualimap 2: Advanced multi-sample quality
718 control for high-throughput sequencing data. *Bioinformatics* **32**: 292–294.

719 Posada-Cespedes S, Seifert D, Beerenwinkel N. 2017. Recent advances in inferring viral diversity
720 from high-throughput sequencing data. *Virus Research* **239**: 17–32.

721 Routh A, Johnson JE. 2014. Discovery of functional genomic motifs in viruses with ViReMa-a virus
722 recombination mapper-for analysis of next-generation sequencing data. *Nucleic Acids Research* **42**:
723 e11.

724 Sanz AI, Fraile A, García-Arenal F, Zhou X, Robinson DJ, Khalid S, Butt T, Harrison BD. 2000.
725 Multiple infection, recombination and genome relationships among begomovirus isolates found in
726 cotton and other plants in Pakistan. *Journal of General Virology* **81**: 1839–1849.

727 Schirmer M, Sloan WT, Quince C. 2014. Benchmarking of viral haplotype reconstruction programmes:
728 An overview of the capacities and limitations of currently available programmes. *Briefings in*
729 *Bioinformatics* **15**: 431–442.

730 Seoane P, Espigares M, Carmona R, Polonio Á, Quintana J, Cretazzo E, Bota J, Pérez-García A,
731 Dios Alché J de, Gómez L, et al. 2018. TransFlow: a modular framework for assembling and
732 assessing accurate de novo transcriptomes in non-model organisms. *BMC Bioinformatics* **19**: 416.

733 Seoane P, Ocaña S, Carmona R, Bautista R, Madrid E, Torres AM, Claros MG. 2016. AutoFlow, a

734 Versatile Workflow Engine Illustrated by Assembling an Optimised de novo Transcriptome for a Non-
735 Model Species, such as Faba Bean (*Vicia faba*). *Current Bioinformatics* **11**: 440–450.

736 Sievers F, Higgins DG. 2014. Clustal Omega. *Current Protocols in Bioinformatics* **2014**: 3.13.1-
737 3.13.16.

738 Stewart JB, Chinnery PF. 2021. Extreme heterogeneity of human mitochondrial DNA from organelles
739 to populations. *Nat Rev Genet* **22**: 106–118.

740 Stewart JB, Chinnery PF. 2015. The dynamics of mitochondrial DNA heteroplasmy: implications for
741 human health and disease. *Nat Rev Genet* **16**: 530–542.

742 Töpfer A, Marschall T, Bull RA, Luciani F, Schönhuth A, Beerenwinkel N. 2014. Viral quasispecies
743 assembly via maximal clique enumeration. *PLoS Comput Biol* **10**: e1003515.

744 Verbist BMP, Thys K, Reumers J, Wetzels Y, Van Der Borght K, Talloen W, Aerssens J, Clement L,
745 Thas O. 2015. VirVarSeq: A low-frequency virus variant detection pipeline for Illumina sequencing
746 using adaptive base-calling accuracy filtering. *Bioinformatics* **31**: 94–101.

747 Welkers MRA, Jonges M, Jeeninga RE, Koopmans MPG, de Jong MD. 2014. Improved detection of
748 artifactual viral minority variants in high-throughput sequencing data. *Frontiers in microbiology* **5**: 804.

749 Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. 2011. ShoRAH: estimating the genetic
750 diversity of a mixed sample from next-generation sequencing data. *BMC bioinformatics* **12**: 119.

751 Zukurov JP, Nascimento-Brito S, Volpini AC, Oliveira GC, Janini LMR, Antoneli F. 2016. Estimation of
752 genetic diversity in viral populations from next generation sequencing data with extremely deep
753 coverage. *Algorithms for Molecular Biology* **11**: 2.

754
755

756

757