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

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

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

[2] 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.

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

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

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

* **Contact information**. Tel: +34 952 137 284; Fax: +34 952 132 041; Email: claros@uma.es

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

19 **ABSTRACT**

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

36

37 **MAIN**

38 RNA and ssDNA viruses, including SARS-CoV-2, exist as highly variable and heterogeneous
39 collections of mutated and recombinant genomes known as virus quasispecies [1,2]. The term
40 'quasispecies' means that a viral genome is indeed a weighted average of a large number of different
41 but related individual genotypes, called haplotypes [2,3]. Haplotypes differ not only in point mutations,
42 since it is well known that many viruses can recombine resulting in the shuffling of beneficious and
43 deleterious variations that can be separated and tested as individual units in new combinations. High
44 variability has also been found in mitochondrial DNA (mtDNA) that has been related to human mtDNA
45 disorders that affect one of 4300 individuals [4]. The frequency in which the mtDNA variants (point

46 mutations and indels) are found determines that they are considered heteroplasmy, if different
47 variants of the mtDNA coexist in the same cell, tissue, organ or organism, or homoplasmy, in case all
48 mtDNA copies share the mutation. Heteroplasmy and homoplasmy have been related to disease and
49 ageing in humans [5].

50 Studies on virus quasiespecies usually required laborious and expensive laboratory efforts.
51 Fortunately, the advent of next-generation sequencing (NGS) technologies enabled the fast and
52 cheap detection of any minor variant within a huge population of genomes [6,7] and discern all single
53 nucleotide variations (SNVs) in any quasispecies to define a haplotype (viral variant) without the need
54 of culture or cloning prior to sequencing [8]. However, analysing NGS data can be an awkward
55 endeavour starting from the available open-source, cutting-edge software with incompatible formats to
56 study recombination, variant calling and haplotypes [9]. Additionally, quasispecies studies must deal
57 with sequencing errors, whose frequency lays close to the viral replication error rates (10^{-3}-10^{-5} errors
58 per nucleotide).

59 To fill this gap, here we describe QuasiFlow, a bioinformatic tool based on well stablished software to
60 study genetic variability from NGS data that is able to distinguishing real variations from sequencing
61 errors. Accurate and precise detection of low-frequency mutations, genetic variability and
62 recombination is achieved in viral quasispecies and other highly variable genomes like mtDNA. Its
63 companion QuasiComparer facilitates the interpreation of multiple analyses on comparable samples.
64 Both pipelines together can extract the complete genetic variability provided that more than 250
65 million nucleotides were sequenced.

66

67 **RESULTS**

68 ## Stringent filtering in QuasiFlow provides a reliable detection of low-frequency
69 genetic variations

70 The performance of QuasiFlow workflow (Fig. 1) was investigated using four datasets (mut_rec,
71 mut_REC, MUT_rec and MUT_REC, where 'MUT/mut' means mutations, 'REC/rec' means
72 recombination, lower case indicates absence, and upper case indicates presence) of synthetic reads

mimicking an Illumina sequencing project (see Materials and Methods for details). Read rejection rate was 0.72 in all four groups (Fig. 2A, RejectedReads) due to the strict requeriments in the read-preprocessing step to avoid the confusion of sequencing errors with mutation events. Consequently, the useful reads (Fig. 2A, ReferenceReads) presented a recovering index of 0.27 for recombinant free datasets (mut_rec and MUT_rec) with a slight decrease to ~0.25 where only recombinations (mut_REC) or recombinations together with mutations (MUT_REC) were present, respectively. Mapping rate of useful reads was nearly 1.00 for the samples without mutations (Fig 2A, MappedReads) with a slight decrease to 0.98 when reads contain artificial mutations, indicating that the accumulation of sequence changes causes the alignment rejection by the mapper with highly strict parameters. Concerning recombination, untangled datasets (mut_rec and MUT_rec) showed an expected 0.0 rate for true and false recombination events (Fig 2A, TrueRECs and FalseRECs) while most (0.86) recombination 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.

The whole variation profile was then investigated (Fig. 2B-D). As expected, the control group of mut_rec presented no mutations and no recombination events, being consistent with a very low frequency of genetic variability (Fig. 2B). For high variation profiles such as MUT_REC (Fig 2C), 261 new mutations with frequencies ranging from 0.31 to 1.72×10^{-4}, and 13 recombination events were detected. This illustrates that QuasiFlow can find both low- and high-frequent mutation and recombination events. A test using a real-world dataset from a RCA-amplified infectious clone of the begomovirus TYLCV-Mld (tomato yellow leaf curl virus, mild strain), where low mutation frequency is

expected as well as absence of recombination, 72 mutations with frequencies ranging from 4.2×10^{-4} to 2.2×10^{-4} were detected and no recombination events were identified (Fig. 2D). The apparent hotspot detected from the nucleotide 1110 to 1128 for the ORF so called 315-258-dir (Fig 2D, track B) is due to the quasi-dimer (1,7-mer) nature of infectious clone compared to the reference genome, being the region where digestion and fusion was produced and is not identical in all virus haplotypes. Since the RCA-amplified TYLCV-Mld (Fig 2D) presents values closer to mut_rec (Fig 2B) than to MUT_REC (Fig 2C), it can be suggested QuasiFlow was able to detect the low amount of variation introduced by bacterial cloning and RCA manipulation, and that this low level of variability would have a negligible impact on genetic variation. This also supports that mutation and recombination events and the resulting haplotypes provided by QuasiFlow should be considered reliable, derived from real biological events and not from experimental artefacts.

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

Quasiespecies reconstruction of original and recombinant geminiviruses in a tomato plant

QuasiFlow performance detecting recombinant viruses was illustrated using geminivirus. MiSeq data from a rolling circle amplification (RCA) product of a TYLCV-Mld agroinfecting *Solanum lycopersicum* plant was analysed. Viral genetic variation was evaluated with the incidence, abundance and functional diversity indexes described in methods presented in Table S1 of Supplementary File 2.

130 These values are characteristic of highly diverse viral populations [10], since (i) the mutation frequency
131 was 1.008 × 10^{-2} mut/nt, two orders of magnitude higher than that of other geminiviruses (~10^{-4}
132 mut/nt) [11–14]; and (ii) the Shannon index reached 0.6. The consensus sequence differed by 32 SNVs
133 from the reference sequence (TYLCV-Mld) and its mutation frequency per nucleotide position
134 clustered into two well differentiated regions (Figure 3A). The first, from positions 1 to 1000 and
135 includes 21 changes with a mutation frequency of 0.85 ± 0.03. A second region, from position 1201 to
136 2791, included changes with mutation frequencies of 0.95 ± 0.03 and 0.69 ± 0.02. Between the two
137 regions, at positions 1000-1200 a hostpot was found with a mutation frequency of 0.45. A total of 31
138 haplotypes (Hp) were reconstructed (Figure 3B), including one (Hp1) identical to the reference
139 sequence. Hp12, Hp27 and Hp31 accounted for 80% of the total. Hp12 showed an identity of 98%
140 with TYLCV-Mld and represented 40% of the total. Interestingly, blast analysis indicated that Hp12
141 was identical to Almería strain of TYLCV (TYLCV-Almería, AJ489258.1) from nt 1-1054 whereas from
142 nt 1055-2791 was identical to Malaga strain (TYLCMaV, AF271234.1). The recombinant nature of
143 Hp12 was confirmed by RDP analysis (Fig. 3C) and suggested a triple infection within the agro-
144 infected plant. *In silico* results were confirmed by restriction analysis (Fig. 3D) of the RCA product.
145 *KpnI* releases the monomer of TYLCMaV and *ClaI* cuts TYLCV-Almería into two fragments and
146 linearises TYLCV-Mld. This assay revealed the existence of the three viruses in the tomato plant. In
147 addition, *bona fide* recombinant genomes between TYLCV-Almería and TYLCMaV (Fig. 3E) were
148 amplified with specific primers. Finally, since haplotype reconstruction showed a high frequency of
149 genomes carrying changes that belonged to the TYLCMaV sequence, the relative abundance of each
150 virus was investigated. The analysis of the coverage of each parental virus (Figure 3F), using an
151 artificial reference sequence with the genome of the three viruses, showed that most of the variants
152 belonged to TYLCMaV, followed by TYLCV-Almería, while TYLCV-Mld was almost undetectable,
153 suggesting a fitness advantage of TYLCMaV.

154

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

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

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

The impact of sex and tissue factors were inspected using QuasiComparer. No significant differences ($P < 0.05$; Table S4 in Supplementary File 2) were detected among the 47 samples for any of the factors, but differences become evident by PCA for tissue factor (Fig. 4C), that explained 48% of the total variance. In fact, hierarchical clustering grouped samples in four clusters (Fig. 4D) two of them by tissue, while no association between sex and mtDNA variants was found. The blue- and cyan-

183 boxed clusters in Fig. 4D contained mainly lymphoblastoid samples and the red-boxed cluster was
184 enriched with blood samples. Differences between these clusters were found significant (*t*-Student, P
185 < 0.05).

186

187 Genetic variability requires sequencing of 250 million nucleotides in TYLCV

188 To present a robust and accurate assessment of mutation and recombination frequencies, the
189 QuasiFlow/QuasiComparer analysis must rely on the whole genetic variability, and this is clearly
190 dependent on the number of reads for the low frequent SNVs. Hence, the number of mutations
191 detected using MiSeq or HiSeq platforms on the same sample was compared depending on the total
192 number of nucleotides sequenced. A sigmoid behavior was observed using both platforms (Figure S1
193 in Supplementary File 2), reaching a plateau (that is, mutations at all frequencies were detected)
194 when nearly 250 million nucleotides were sequenced. It is reassuring that both sequencing platforms
195 require the same number of nucleotides sequenced to saturate the analysis, corresponding to ~1.5
196 millions of true, useful 2 × 300 MiSeq reads (derived from ~3 million raw reads) per sample, or ~0.75
197 million of true 2 × 100 HiSeq reads (derived from ~1.25 millions of raw reads) per sample. However,
198 MiSeq datasets reach the plateau in ~1 815 mutations (Figure S1, yellow lines), while HiSeq saturates
199 at ~580 to ~800 (Figure S1**,** blue lines). Only the sequencing platform can be invoked as source of
200 differences since the same DNA was sequenced in both cases and the same bioinformatic pipeline
201 run. It seems that MiSeq reactions would be more prone to sequencing variations than HiSeq
202 reactions (in fact, we were verbally told by Illumina staff that the sequencing kit for 2 × 300 nt in MiSeq
203 produces more sequencing errors than other kits). Therefore, being conservative, HiSeq could be
204 considered true variations, while MiSeq is suspected to overestimate the genetic variation. The arrival
205 of NovaSeq and the new sequencing platforms might benefit the performance of QuasiFlow. In any
206 case, it seems reasonable to avoid sequencing mixtures from different platforms to measure genetic
207 variability.

208

209 **DISCUSSION**

210 The study of highly variable genetic systems was initiated in virology [15] and now extends to other
211 biological systems such as cancer [16], bacteria [17], mitochondria [18], or prions [19]. These analyses rely on
212 accurate determination of mutations for the development of effective therapies. The aim of QuasiFlow
213 (Fig. 1) is to provide sensitive, reliable, and robust analysis of complex genetic variability (including
214 both low frequency mutations and recombinations) with a highly-stringent set of well-known
215 bioinformatics tools glued with in-house code. Results are easy to interpret when compared using
216 QuasiComparer (Supplementary File 1) as illustrated with PCA and ANOVA plots grouping samples
217 by their biological origin and not by the experimental process used to produce them (Figs. 2E-F, 3F
218 and 4B-D). The capabilities of QuasiFlow/QuasiComparer couple have been demonstrated using
219 synthetic sets of data (Fig. 2) and applied to a real-world data from plant viral (Fig. 3) and human
220 mitochondrial (Fig. 4) populations.

221 The advent of the NGS (next-generation sequencing) technology has dramatically boosted the
222 intrahost viral diversity studies, even if there is still no consensus about the most appropriate
223 technology, the most appropriate software or the minimal read quality for SNV/recombination
224 validations [20]. QuasiFlow applied to NGS data enabled, using a coverage >10 000× when a threshold
225 of 20× was widely used in literature, the detection of low frequency SNVs and recombinations with a
226 small rate of false positive detection. Emerging viruses such as coronavirus and geminiviruses are
227 threatening humans and crops, respectively, due to their high frequency of mutation and
228 recombination [21]. Although some studies show that recombination is not the main phenomenon
229 involved in the evolution of these viruses [22], it is important to take this process into account due to its
230 potential to generate large evolutionary jumps in very small time frames. Thus, recombinants can lead
231 to host range expansion and increases in patogenicity. The analysis of the TYLCV-Mld infective clone
232 in tomato plants in Fig. 3A-B demonstrated a striking triple infection (Fig. 3C), revealing that studies in
233 greater depth of this type of co-infections could shed light on the reasons for fitness differences
234 between such similar virus species.

235 The reliable analyses of the genetic heterogeneity of human mtDNA (Fig. 4) confirms the wide scope
236 of QuasiFlow. In first instance, QuasiFlow confirmed the extreme genetic variation within mtDNA [23]

237 with specific variant patterns in different tissues or nuclear backgrounds (i.e. sex), becoming a highly
238 accurate and sensitive tool for the analysis of mitochondrial diseases. Interestingly, low-frequency
239 mtDNA heteroplasmies were detected and then related to diseases that might have been elusive to
240 detection by other means. Moreover, one can also envisage the utility of QuasiFlow/QuasiComparer
241 in single-cell studies [24]. In the case of tumours where genetic diversity is high from cell to cell due to a
242 high mutation rate and such tumors can be considered as a collection of subclones, the study of this
243 diversity can be facilitated using QuasiFlow: genetic diversity between patients or within a tumor can
244 lead to pinpoint actionable driver mutations to design targeted therapies [25,26].

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

250

251 ONLINE CONTENT

252 Any methods, additional references, Nature Research reporting summaries, source data, extended
253 data, supplementary information, acknowledgements, peer review information; details of author
254 contributions and competing interests; and statements of data and code availability are available at
255 *QuasiFlow_ONLINE.pdf* document.

256

257 REFERENCES

258 1. Lauring, A. S. & Andino, R. Quasispecies theory and the behavior of RNA viruses. *PLoS*
259 *Pathogens* vol. 6 1–8 (2010).

260 2. Domingo, E., García-Crespo, C. & Perales, C. Historical Perspective on the Discovery of the
261 Quasispecies Concept. *Annu. Rev. Virol.* **8**, 51–72 (2021).

262 3. Domingo, E. & Perales, C. Viral quasispecies. *PLOS Genetics* **15**, e1008271 (2019).

263 4. Gorman, G. S. *et al.* Prevalence of nuclear and mitochondrial DNA mutations related to adult
264 mitochondrial disease. *Annals of Neurology* **77**, 753–759 (2015).

265 5. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy:
266 implications for human health and disease. *Nat Rev Genet* **16**, 530–542 (2015).

267 6. Bull, R. A. *et al.* A method for near full-length amplification and sequencing for six hepatitis C
268 virus genotypes. *BMC Genomics* **17**, (2016).

269 7. Posada-Cespedes, S., Seifert, D. & Beerenwinkel, N. Recent advances in inferring viral
270 diversity from high-throughput sequencing data. *Virus Research* vol. 239 17–32 (2017).

271 8. Zukurov, J. P. *et al.* Estimation of genetic diversity in viral populations from next generation
272 sequencing data with extremely deep coverage. *Algorithms for Molecular Biology* **11**, (2016).

273 9. Cacciabue, M., Currá, A., Carrillo, E., König, G. & Gismondi, M. I. A beginner's guide for
274 FMDV quasispecies analysis: sub-consensus variant detection and haplotype reconstruction using
275 next-generation sequencing. *Briefings in Bioinformatics* bbz086 (2019) doi:10.1093/bib/bbz086.

276 10. Gregori, J. *et al.* Viral quasispecies complexity measures. *Virology* vol. 493 227–237 (2016).

277 11. Isnard, M., Granier, M., Frutos, R., Reynaud, B. & Peterschmitt, M. Quasispecies nature of
278 three maize streak virus isolates obtained through different modes of selection from a population used
279 to assess response to infection of maize cultivars. *Journal of General Virology* **79**, 3091–3099 (1998).

280 12. Sanz, A. I. *et al.* Multiple infection, recombination and genome relationships among
281 begomovirus isolates found in cotton and other plants in Pakistan. *Journal of General Virology* **81**,
282 1839–1849 (2000).

283 13. Ge, L., Zhang, J., Zhou, X. & Li, H. Genetic structure and population variability of tomato
284 yellow leaf curl China virus. *Journal of virology* **81**, 5902–5907 (2007).

285 14. Duffy, S. & Holmes, E. C. Phylogenetic evidence for rapid rates of molecular evolution in the
286 single-stranded DNA begomovirus tomato yellow leaf curl virus. *J Virol* **82**, 957–965 (2008).

287 15. Domingo, E. & Schuster, P. *Quasispecies: from theory to experimental systems*. (Springer,
288 2016).

289 16. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat*
290 *Rev Clin Oncol* **15**, 81–94 (2018).

291 17. Chan, K.-H. *et al.* Comparison of Pyrosequencing, Sanger Sequencing, and Melting Curve
292 Analysis for Detection of Low-Frequency Macrolide-Resistant Mycoplasma pneumoniae Quasispecies
293 in Respiratory Specimens. *J Clin Microbiol* **51**, 2592–2598 (2013).

294 18. Aryaman, J., Johnston, I. G. & Jones, N. S. Mitochondrial Heterogeneity. *Front. Genet.* **9**, 718
295 (2019).

296 19. Mahal, S. P., Browning, S., Li, J., Suponitsky-Kroyter, I. & Weissmann, C. Transfer of a prion
297 strain to different hosts leads to emergence of strain variants. *Proc. Natl. Acad. Sci. U.S.A.* **107**,
298 22653–22658 (2010).

299 20. McCrone, J. T. & Lauring, A. S. Measurements of Intrahost Viral Diversity Are Extremely
300 Sensitive to Systematic Errors in Variant Calling. *J Virol* **90**, 6884–6895 (2016).

301 21. Martin, D. P. *et al.* Complex Recombination Patterns Arising during Geminivirus Coinfections
302 Preserve and Demarcate Biologically Important Intra-Genome Interaction Networks. *PLoS Pathog* **7**,
303 e1002203 (2011).

304 22. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns
305 and determinants. *Nat Rev Genet* **9**, 267–276 (2008).

306 23. Stewart, J. B. & Chinnery, P. F. Extreme heterogeneity of human mitochondrial DNA from
307 organelles to populations. *Nat Rev Genet* **22**, 106–118 (2021).

308 24. Huang, R. H., Wang, L. X., He, J. & Gao, W. Application and prospects of single cell
309 sequencing in tumors. *Biomark Res* **9**, 88 (2021).

310 25. Mroz, E. A. & Rocco, J. W. The challenges of tumor genetic diversity. *Cancer* **123**, 917–927
311 (2017).

26. Baslan, T. *et al.* Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing. *eLife* **9**, e51480 (2020).

FIGURES

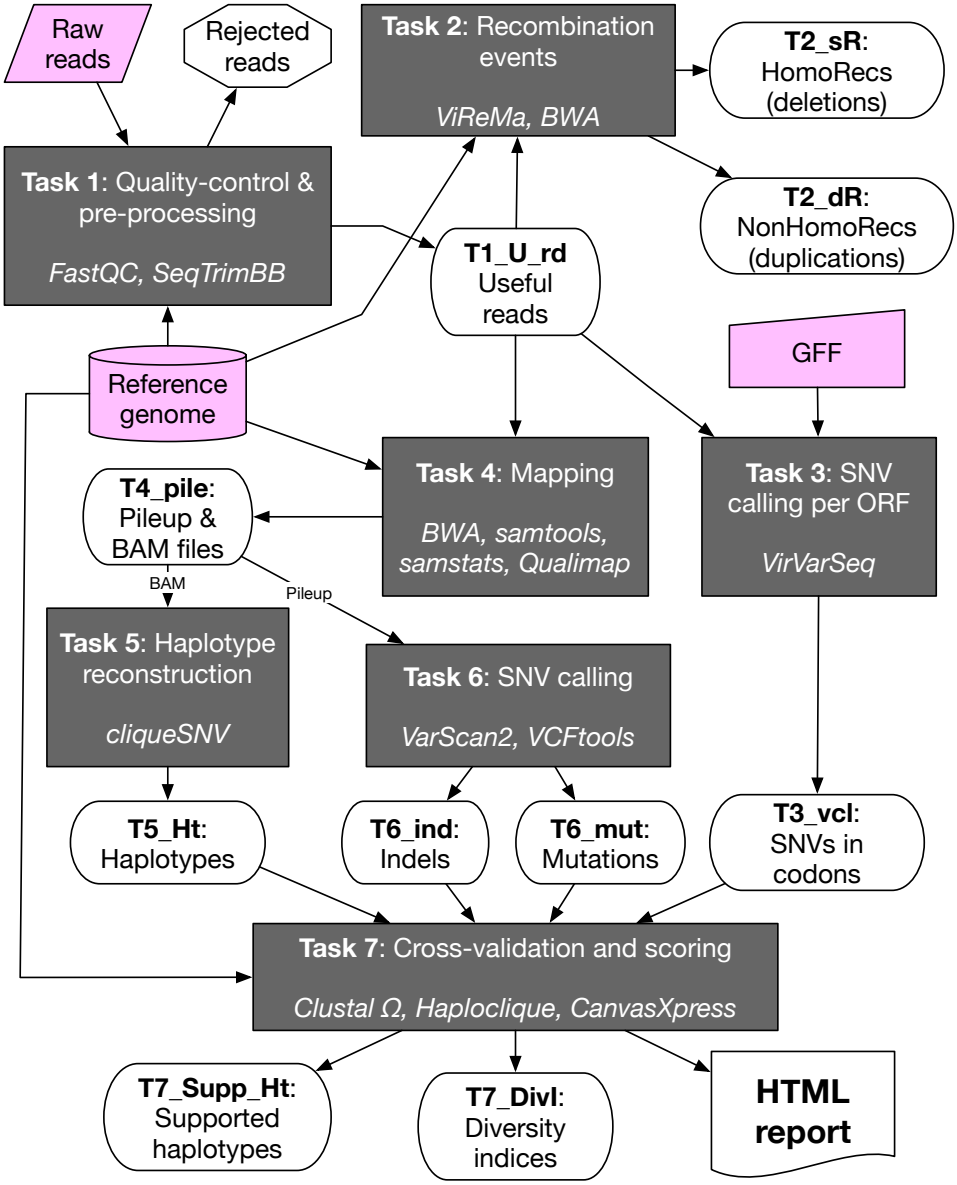

Figure 1: Conceptual representation of the QuasiFlow workflow for quantifying genetic variability using high-throughput sequencing reads. The seven tasks, described in detail in Materials and Methods, are grey rectangles, where the relevant bioinformatic tools are indicated in italics. Magenta boxes 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.

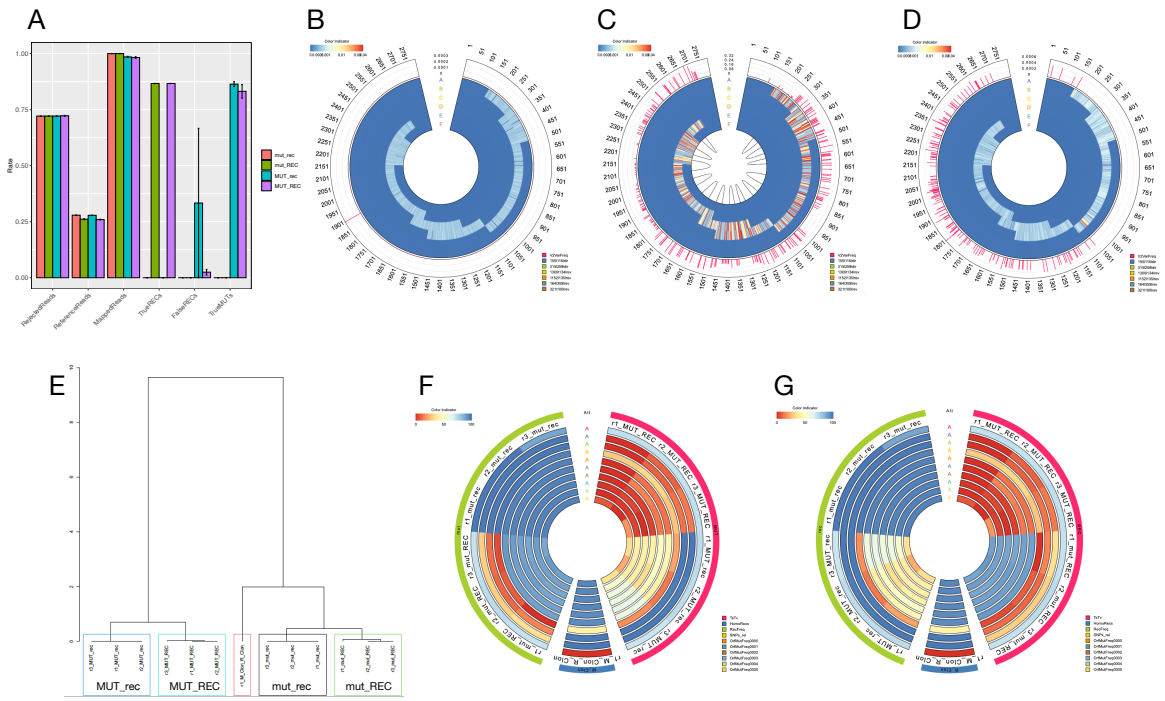

Figure 2: Quasiflow and QuasiComparer performance. (**A**) Benchmarking QuasiFlow with synthetic reads. Bars represent means of three replicates with their standard error. Colours are indicated in the legend and correspond to the four synthetic datasets. "RejectedReads" are reads removed due to quality, 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 inserted mutations (MUT) or recombinations (REC), and in blue is the TYLCV-Mld infectious clone sample. Inner rings represent the most significant variables, where the colour spectrum represents the relative value of the variable in the range of 0 (blue) to 100% (red), as indicated in the colour indicator. The QuasiFlow HTML report for mut_rec (*1-mut_rec.html*) and MUT_REC (*2-MUT_REC.html*) synthetic reads, the QuasiComparer complete comparative report for all synthetic reads (*3-Comparative_Report_SyntheticReads.html*), as well as for the control of TYCLV-Mld replicated in vitro (*4-Control_TYLCV.html*) can be inspected in Supplementary File 1.

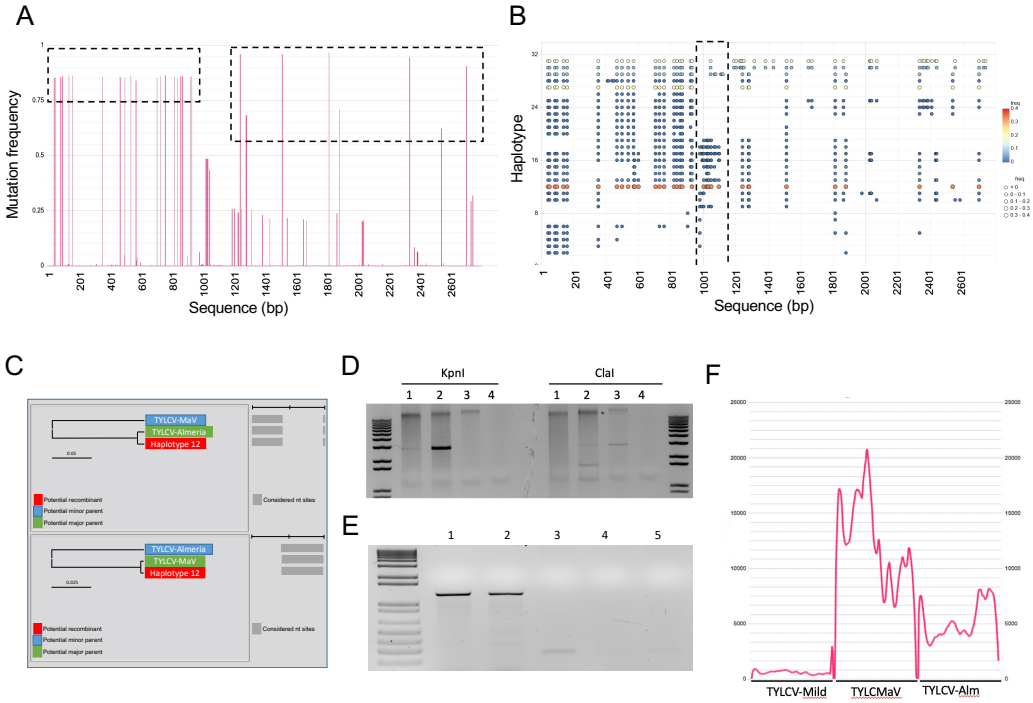

Figure 3. Variation analysis of a TYLCV-Mld population after infection of tomato plants. **(A)** Nucleotide mutation frequency profile, where black dashed lines indicate regions that group SNPs of similar frequency, one containing mutations with two different frequencies. Only mutations with a frequency 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 sequence. The QuasiComparer complete comparative report for tomato infected plants with TYCLV (*5-Tomato_TYLCV-invected.html*) can be inspected in Supplementary File 1.

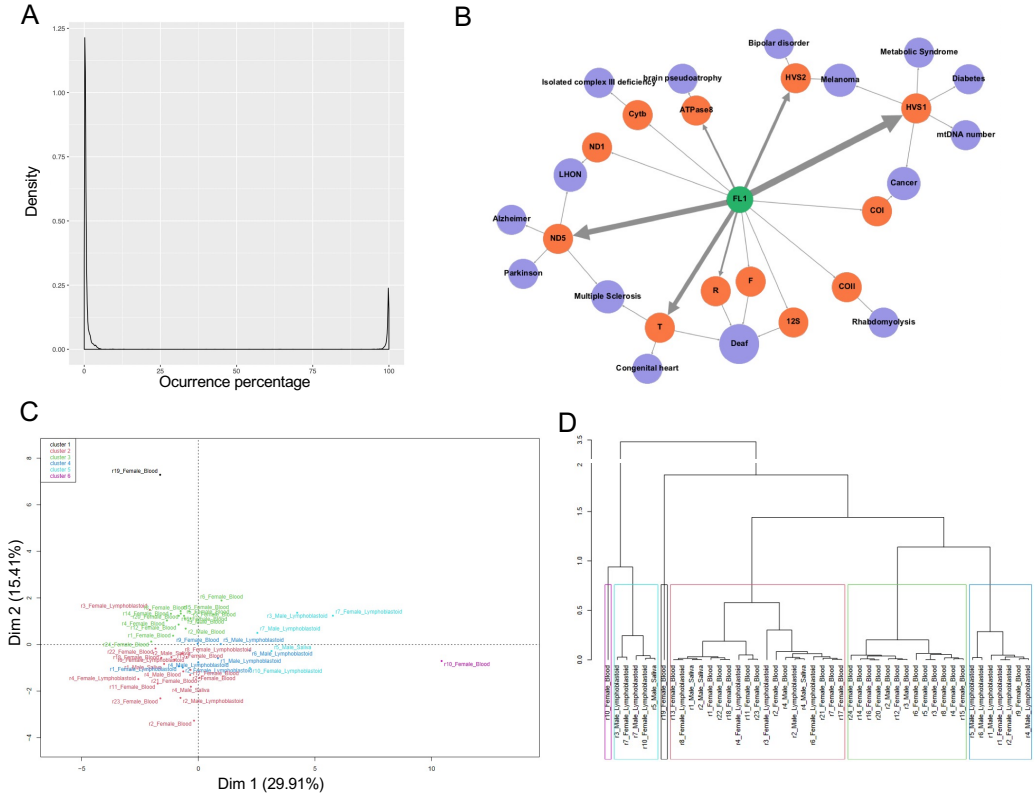

Figure 4. Variation analysis of mtDNA populations from different human tissues. **(A)** Average distribution of mtDNA variants in the 47 samples analyzed showing very low frequency of heteroplasmies (below 0.2%) and homoplasmies, but without intermediate frequency variants. **(B)** Mutation network of 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.

386

387 **METHODS ONLINE**

388 ## Workflow availability and requirements

389 QuasiFlow (Fig. 1), a workflow developed with the workflow manager AutoFlow [1], takes as input raw,
390 paired-end read data generated by a NGS platform (mainly Illumina, although other are allowed). Its
391 design took advantage of computational resources available and the following dependencies: Art [2],
392 SeqTrimBB, FastQC [3], BWA [4], samtools [5], Samstat [5], Qualimap [6], ViReMa [7], VirVarSeq [8] CliqueSNV
393 [9], Varscan2 [10], VCFtools [11], Haploclique [12], Clustal Ω [13], CanvasXpress [14] and FactoMineR [15]. It was
394 coded and executed in Picasso supercomputer (3456 cores and 8.4 TB of RAM with Slurm queue
395 system on SUSE Linux Enterprise Server 11SP2) at the University of Málaga. It requires (1) a
396 reference genome in FASTA format (the begomovirus TYLCV or human mitochondrial DNA in this
397 study), (2) a GFF file containing all open reading frame (ORF) coordinates, and (3) the circular or
398 linear nature of the reference genome. Several samples can be run simultaneously using a
399 customizable daemon script provided with the code that enables (1) executing QuasiFlow with all
400 samples, (2) monitorizing QuasiFlow executions, (3) gathering results from all samples, and (4)
401 comparative sample analysis using QuasiComparer. The workflow procceeds through the following
402 seven tasks (Fig. 1):

403 ### *Task 1: Read quality control and pre-processing*

404 Raw reads were quality controlled using FastQC, and pre-processed using SeqTrimBB to remove low
405 quality bases, vectors and adapters. SeqTrimBB default parameters were used, except for the
406 following: 1) minimum phred base mean quality was set to 26 to avoid sequencing errors to be
407 confounded with low frequency mutations and reduce the false positive ratio on ulterior analyses [16,17];
408 2) the minimum read length was increased to 100 nt to remove shorter reads confounding haplotype
409 reconstruction [18]. Viral reads were extracted from host and contaminant reads using the reference
410 genome. This more reliable but reduced number reads (Fig 1, T1_U_rd) guaranteed soundness and
411 robustness in further steps [19].

Task 2: Recombination events

The useful reads (T1_U_rd) were then treated as single reads (even though they come from paired-end reads) with ViReMa [7] using default parameters, excepting the aligner, which was configured to BWA [4] instead of the default Bowtie since BWA is widely used for recombination studies. To avoid the identification of indels or misalignments as recombination events due to confounding reads, the seed minimum length was set to 55, the allowed mismatches was set to 1, and the microindels size was set to 55. Spurious recombination events with a frequency lower than 1×10^{-3} were removed to increase reliability in predicted recombination events. Additionally, recombination events located between the start and end positions of a circular reference sequence were discarded. Recombination events are classified by ViReMa as HomoRecs (Fig 1, T2_sR) for those resulting in a sequence deletion, or NonHomoRec (Fig 1, T2_dR) when the resulting sequence presents a duplication.

Task 3: SNV calling per ORF

The SNV calling process per ORF was based on VirVarSeq [8] 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 [5] and Qualimap [6] tools. Finally, by means of mpileup [5], 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 [20]. 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 [9] 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 [8]. 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 [11] 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.

465 ***Task 7: Cross-validation and scoring***

466 Taking in mind that haplotype reconstruction was poor when sequence divergence is low, failing to
467 recover rare haplotypes [21], QuasiFlow was designed with a final task to remove unreliable haplotypes
468 and increase the consistency with previously detected SNVs. To do so, haplotypes in T5_Ht were
469 aligned to the genome reference using Clustal Ω, and the variable positions were cross-validated with
470 previously obtained SNVs (T6_mut and T6_ind). Variability not supported by a SNV was edited, and
471 haplotypes whose variations were all edited were then rejected as due to sequencing errors or
472 untrusted variations. The resulting SNV-supported haplotypes (Fig. 1, T7_Supp_Ht) were realigned
473 again and haplotype frequency was used to score diversity indices described by Gregori et al [22], such
474 as normalized Shannon, Simpson, Gini-Simpson, Hill numbers, sample nucleotide diversity, etc (Fig.
475 1, T7_Divl; see Table MM1 below for a detailed description). In addition, haplotypes and SNVs were
476 represented as a network where each haplotype was connected with all its SNVs to facilitate the
477 interpretation of the genetic complexity and heterogeneity of the sample (see reports in
478 Supplementary File 1). To facilitate result interpretation and visual user inspection an HTML report
479 based on the javascript scientific library CanvasXpress [14] including the values for the parameters of
480 Table MM1 as well as interactive plots and graphs is provided. It should be noted that most diversity
481 indices (but Shannon) are provided for illustrative purposes since they are depending on the selected
482 haplotypes.

Table MM1: Variables calculated during QuasiFlow execution.

Variable name	**Definition**	**Software[a]**
Variables used in PCA QuasiComparer		
HomoRecs	Number of nomologous recombinations causing deletions	ViReMa
NonHomoRecs	Number of non homologous recombinations causing duplications	ViReMa
ShannonDivI	Shannon Index: measure of diversity based on haplotype frequencies	
SampNucDivEntity	Sample nucleotide diversity: average number of substitutions among pairs of haplotypes in the multiple alignment, which results of dividing the functional attribute diversity by the number of haplotype pairs	
Haplot_tot	Amount of haplotypes reconstructed after verification of all sequence changes	CliqueSNV
SNPs_rel	Number of mutations normalized by the total amount of mapped nucleotides	VarScan2/BWA
InDels_rel	Number of indels normalized by the total amount of mapped nucleotides	VarScan2/BWA
Transits_rel	Number of transitions normalized by the total amount of mapped nucleotides	VarScan2/BWA
Transvers_rel	Number of transversions normalized by the total amount of mapped nucleotides	VarScan2/BWA
TsTv_ratio	Ratio between transitions and transversions	VarScan2
OrfMutFreqX	Mutation frequencies detected for each ORF (encoded as 000X code, where X is the number of the ORF), which are computed after translating codons to amino acids	VirVarSeq
Rec_freq	Recombination frecuency calculated as the number of recombinant reads divided by the total number of reads analysed by ViReMa	
Variables not used for PCA classification		

PopNucDivI	Population nucleotide diversity: average number of nucleotide differences between two genomes of the quasispecies	
SimpsonDivI	Simpson index: probability that two randomly selected genomes from a viral population belong to the same haplotype	
GinSimpDivI	Ginni-Simpson index: probability that two genomes taken at random belong to a different haplotype	
HillQ1DivI	Hill index Q1: diversity measure in units of equally abundant species assuming a contribution of the different haplotypes to the diversity $q = 1$	
HillQ2DivI	Hill index Q2: diversity measure in units of equally abundant species assuming a contribution of the different haplotypes to the diversity $q = 2$	
FuncAttDiv	Functional attribute diversity: it is an incidence-based functional diversity index equal to the sum of the elements in the matrix of dissimilarities of pair-wise genetic distances between haplotypes	
MutFreqEntity	Mutation frequency: fraction of mutated residues in the alignment with respect to the reference	
VFPKM	Virus fragments per kilobase and million reads	
InDels_tot	Number of detected indels	VarScan2
SNPs_tot	Number of detected mutations	VarScan2
Transits_tot	Number of detected transitions	VarScan2
Transvers_tot	Number of detected transversions	VarScan2
X2Y	Number of changes of nucleotide X into Y	VarScan2
X2Y_rel	Number of changes of nucleotide X into Y normalized by the total amount of mapped nucleotides against the reference	VarScan2
Ambigous_recombination_events_by_Virema	Number of homologous recombinations in which the strand sense cannot be determined	ViReMa
Indels_by_Virema	Total amount of indels	ViReMa
Total_reads_analysed_by_Virema	Number of reads processed in the recombination analysis	ViReMa

RawReads_tot	Number of raw reads in original fastq files	SeqTrimBB
RejectedReads_tot	Number of reads rejected due to low quality/short length	SeqTrimBB
ReferenceReads_tot	Number of reads that are expected to map to the reference sequence	SeqTrimBB
ReferenceReads_perc	Reads asigned to reference normalized to input read number.	SeqTrimBB
HostReads_tot	Number of useful reads not assigned to the reference	SeqTrimBB
ntMapped2Reference_tot	Total amount of mapped nucleotides against reference. Used in normalization of relative metrics	BWA
mapping_coverage	Proportion of the reference covered by the analysed reads. It is averaged when the reference is comprised by several sequences.	Qualimap
Haplot_raw	Number of detected haplotypes without any further verification	CliqueSNV
Reads_tot_(init/final)	Number of input reads	FastQC
MaxLengthReads_(init/final)	Maximum length of input reads	FastQC
MinLengthReads_(init/final)	Minimum length of input reads	FastQC
GC_(init/final)	GC content	FastQC
MeanQVtotal_(init/final)	Average qual across all the nucleotides	FastQC
QVinLowQ_(init/final)	Quality presented by the nucleotide at the lower quartile	FastQC
QVin10D_(init/final)	Qualilty presented by the nucleotide at the 10th decile. It should be less or equal to the previous one	FastQC
MeanQVperRead_(init/final)	Average nucleotide quality weigthed by the qual frecuency in reads	FastQC
NsFreq_(init/final)	Number of nucleotide indeterminations normalized by the total amount of nucleotides	FastQC
ReadWLength_(init/post)	Read length weigthed by the length dsitribution observed in the input reads	FastQC

Variables describing information using virus reference coordinates		
V2VarFreq	Positions in which VarScan2 detected variants and their frecuency.	VarScan2
i|j|(dir/rev)	The name is the coordinate position of each ORF analysed by VirVarSeq (*i* is the initial nt of the ORF and *j* the length of the ORF in aminoacids, *dir/rev* is the ORF strand, direct or reverse), representing variant frequency at codon level	VirVarSeq
Homologous Strand recombination	Origin/Target coordinates of homologus recombinations	ViReMa
Complementary Strand recombination	Origin/Target coordinates of complementary (non homologous) recombinations	ViReMa

[a] An empty cell indicates that it is an in house function of QuasiFlow

QuasiComparer

The variability analysis usually implies comparison of several samples or conditions. Since QuasiFlow can only manage one sequencing library (sample) at a time, inter-sample comparisons required the development of a dedicated comparative system called QuasiComparer. This analysis can be accomplished automatically by the same daemon script mentioned above provided that a description file with the following data is supplied: 1) the sample name, 2) a replicate identifier that allows to group the replicates that belongs to the same sample, and 3) the experimental group to which the sample is assigned, depending on the user experimental design (e.g., an experimental factor such as days post infection with three sample replicates taken at 7 days and other three replicates at 28 days). QuasiComparer uses the parameters indicated in Table MM1 to perform correlation, ANOVA and principal component analysis (PCA) [23] (Fig. 2B-G) to provide a new, interactive, compartive report in HTML (HTML reports included in Supplementary File 1) incluiding linear correlation plots, impact of each parameter on the experimental factors defined by the user by SNVs and by recombination events. The ANOVA is presented as a table and some circular plots (see Fig. 2F-G, one by each experimental factor described by the user) to illustrate the weight of each significant variable on each sample. The PCA also includes tables with information about its first two dimensions displaying the weight of each variable in PCA dimensions as well as plots (Fig. 4C), a PDF containing correlation analysis between the variables, and hierarchical clustering (Fig. 2E and 4D). All this information is intended to help users to see which samples have a different behavior in comparison to the rest, and correlate this behavior with the experimental factors used in the study.

Synthetic and experimental datasets for testing

To test QuasiFlow and QuasiComparer performance, twelve independent synthetic read datasets were generated using Art [2] with the virus reference sequence of tomato yellow leaf curl virus Mild (TYLCV-Mld, AC# AF071228). The daemon was loaded with two parametres: 1) the presence of mutations and 2) the presence of recombinations, and called an in-house script that generates a FASTA file with mutated, recombinant and original virus sequences. Recombinations were generated making 30 recombinant events in a single sequence of TYLCV-Mld separated by 60 nt. The process was repeated until the recombinant sequences represented 20% of the total. To generate mutations,

random changes were included in 10% of the nucleotide pool of TYLCV-Mld sequences. The FASTA file generated was then provided to Art, setting the read coverage to 300 000×, the maximum read length in 250 nt and a fragment size of 300 nt, with a standard deviation of 10 nt and using the MiSeq version 3.0 built-in error profile.

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

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

Samples, virus source and viral inoculation

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

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

DNA extraction, virus amplification, cloning, and sequencing

DNA from *A. thaliana* Col-0 and *S. lycopersicum* plants were extracted following the Edwards' procedure [24] with some modifications. From each sample, 150 mg of fresh tissue were ground with liquid nitrogen, and 400 µl of extraction buffer [200 mM Tris-HCl, 200 mM NaCl, 25 mM EDTA, SDS 0.5 % (w/v), pH 7.5, 0.1 % β-mercaptoethanol] added. After vortexing at 37 °C for 2 min, extracts were centrifuged at 12 000 × *g* at room temperature for 5 min. The supernatants were transferred to a new tube and mixed with 1 volume of phenol:chloroform (1:1). Following centrifugation at 12 000 × *g* at room temperature for 5 min, the aqueous phase was transferred to a new tube and mixed with 1 volume of isopropanol. After 2 min at room temperature, nucleic acids were recovered by centrifugation at 12 000 × *g* at room temperature for 5 min. The pellet was washed with 70% ethanol, dried and resuspended in 200 µl of RTE buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0, 0.1% RNase A) to a concentration of 130 ng/µl by incubation at 65 °C for 10 min.

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

Verification of the existence of the bioinformatically detected geminivirus species of tomato and agroinoculated *A. thaliana* plants was done by rolling-circle amplification (RCA) and restriction analysis. For this, 130 ng of DNA from young non-inoculated leaves were amplified for 18 h with the Illustra TempliPhi kit (GE Healthcare, UK) following manufacturer's instructions. Amplified DNA was digested with single cutter (in viral genome) restriction enzyme BamHI to obtain the full-length DNA (2.7 kb). Restriction enzymes KpnI and ClaI were used to distinguish between TYLCV-Mld, TYLCV-Almería (AJ489258.1) and TYLCMaV (AF271234.1) in DNA extracts from tomato plants.

The presence of recombinant genomes between TYLCV-Almería and TYLCMaV in extracts form tomato plants was verified by PCR using reverse primer MAV-1271_1289-RW (5'-TTTTCTCAACTTCCGCATC) that binds to TYLCMaV sequence and forward primer IL-41_55-FW (5'-CCCACGAGGGTTCC) that binds to TYLCV-Almería sequence. RDP4 program [25] was used to analyse the recombination event of the most frequent haplotype. Digestions and PCR results were analysed by agarose (Agarose SPI, Duchefa) gel electrophoresis 1% (w/v) with TBE buffer.

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

MiSeq and HiSeq sequencing was performed at Macrogen (Seoul, South Korea). For MiSeq, equimolar sequencing libraries of tomato and nine *A. thaliana* samples were done with Truseq DNA PCR-free (550 bp insert) kit and sequenced on an Illumina MiSeq as 2 $\times$ 300 bp reads. For HiSeq, libraries were done for three *A. thaliana* samples with Truseq DNA PCR-free (350 bp insert) kit and sequenced on an Illumina HiSeq2000 as 2 $\times$ 100 bp reads. The number of reads belonging to geminiviruses that exceeded the initial quality filter was 293 668, with an average coverage of 14 772$\times$ per nucleotide.

Mitocondrial sequences for genetic variability analyses

Human mtDNA sequences BioProject PRJEB5005 contained blood, lymphoblastoid cell lines and saliva samples from males and females. The set comprised 47 libraries of 2 $\times$ 300 bp paired-end reads obtained with the Illumina MiSeq platform with a number of reads between 1.38 $\times$ 10^5 and 4.32 $\times$ 10^5. Samples from the same tissue and sex were considered biological replicates. The Revised Cambridge Reference Sequence ("rCRS") of 16 569 nucleotides (AC# NC_012920) was used as

standard reference for human mtDNA. Since the mean coverage was about 1500×, variants with a frequency below 9×10^{-4} (0.09%) were filtered out by means of QuasiFlow parameters `--min-var-freq to 9x10e-4 and the --min-coverage to 1000` in Task 6. The biological significance of mitochondrial variants was obtained using MITOMAP database [26] (https://www.mitomap.org). To assess tissue origin differences between the lymphoblastoid cell lines and blood a *t*-Student was performed using FactoMineR for each axis of the PCA.

DATA AVAILABILITY

Sequencing reads can be found at BioProject PRJNA823046. Human mtDNA sequences were obtained from BioProject PRJEB5005.

CODE AVAILABILITY

QuasiFlow, QuasiComparer and the daemon script (daemon.sh) that controls their execution can be downloaded from *https://bitbucket.org/seoanezonjic/quasiflow/src/master/*.

The resulting reports are attached as Supplementary File 1.

REFERENCES FOR METHODS

1. Seoane, P. *et al.* Autoflow, a versatile workflow engine illustrated by assembling an optimised de novo transcriptome for a non-model species, such as faba bean (Vicia faba). *Current Bioinformatics* **11**, (2016).

2. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: A next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).

3. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Online* (2010).

4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

5. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

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

7. Routh, A. & Johnson, J. E. Discovery of functional genomic motifs in viruses with ViReMa-a virus recombination mapper-for analysis of next-generation sequencing data. *Nucleic Acids Research* **42**, (2014).

8. Verbist, B. M. P. *et al.* VirVarSeq: A low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics* **31**, 94–101 (2015).

9. Knyazev, S. *et al.* Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Research* **49**, e102–e102 (2021).

10. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22**, 568–576 (2012).

11. Danecek, P. *et al.* The variant call format and VCF tools. *Bioinformatics (Oxford, England)* **27**, 2156–8 (2011).

12. Topfer, A. *et al.* Viral quasispecies assembly via maximal clique enumeration. *PLoS Comput Biol* **10**, e1003515 (2014).

13. Sievers, F. & Higgins, D. G. Clustal Omega. *Current Protocols in Bioinformatics* **2014**, 3.13.1-3.13.16 (2014).

14. Neuhaus, I. CanvasXpress.

15. Husson, F., Josse, J. & Pagès, L. FactoMineR, An R package dedicated to exploratory multivariate analysis. *Journal of Statistical Software* **25**, 1–18 (2009).

16. Morelli, M. J. *et al.* Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. *Veterinary Research* **44**, 12 (2013).

17. Welkers, M. R. A., Jonges, M., Jeeninga, R. E., Koopmans, M. P. G. & de Jong, M. D. Improved detection of artifactual viral minority variants in high-throughput sequencing data. *Frontiers in microbiology* **5**, 804 (2014).

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

19. Daly, G. M. *et al.* Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data. *PloS one* **10**, e0129059 (2015).

20. Samtools Documentation.

21. Schirmer, M., Sloan, W. T. & Quince, C. Benchmarking of viral haplotype reconstruction programmes: An overview of the capacities and limitations of currently available programmes. *Briefings in Bioinformatics* **15**, 431–442 (2014).

22. Gregori, J. *et al.* Viral quasispecies complexity measures. *Virology* vol. 493 227–237 (2016).

23. Seoane, P. *et al.* TransFlow: a modular framework for assembling and assessing accurate de novo transcriptomes in non-model organisms. *BMC Bioinformatics* **19**, 416 (2018).

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

25. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution* **1**, vev003 (2015).

26. Lott, M. T. *et al.* mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr Protoc Bioinformatics* **44**, 1.23.1-26 (2013).

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.

ETHICS DECLARATIONS

Competing interests

The authors declare no competing interests.

AUTHOR INFORMATION

Affiliations

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

1) Department of Molecular Biology and Biochemistry, Facultad de Ciencias, Universidad de Málaga, Malaga, 29071, Spain.
Pedro Seoane & M. Gonzalo Claros

2) 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.

Luis Díaz-Martínez, M. Gonzalo Claros & Ana Grande-Pérez

3) Area de Genética, Facultad de Ciencias, Universidad de Málaga, 29071 Malaga, Spain

Luis Díaz-Martínez, Enrique Viguera & Ana Grande-Pérez

4) Institute of Biomedical Research in Málaga (IBIMA), 29010 Málaga, Spain

Pedro Seoane & M. Gonzalo Claros

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

Pedro Seoane & M. Gonzalo Claros

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

- *1-mut_rec.html*: QuasiFlow report for mut_rec synthetic reads.
- *2-MUT_REC.html*: QuasiFlow report for MUT_REC synthetic reads.
- *3-Comparative_Report_SyntheticReads.html*: QuasiComparer comparative report for all synthetic reads

- *4-Control_TYLCV.html*: QuasiComparer comparative report for the control of TYCLV-Mld replicated in vitro.
- *5-Tomato_TYLCV-invected.html*: QuasiFlow report for TYCLV samples from tomato-infected cells.
- *6-Comparative_Report_Mitochondria.html:* QuasiComparer comparative report for mtDNAs.

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

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