Article - Discoveries

Operational Gene Clusters and intrinsic uncertainty in pangenome analyses

Saioa Manzano-Morales[1,4,*], Yang Liu[1,2,*], Jaime Huerta-Cepas[1,†], Jaime Iranzo[1,3,†]

[1] Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Madrid, Spain.

[2] Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, South China Agricultural University, Guangzhou, China.

[3] Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza, Spain.

[4] Current address: Barcelona Supercomputing Centre (BSC-CNS) - Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain.

* These authors contributed equally to this work.

† Corresponding authors: Jaime Iranzo (jaime.iranzo@upm.es); Jaime Huerta-Cepas (j.huerta@csic.es)

Abstract

A key step for comparative genomics is to group open reading frames into functionally and evolutionarily meaningful gene clusters. Such clustering is complicated by intraspecific gene duplications and horizontal gene transfers, that are frequent in prokaryotes. In consequence, gene clustering methods must deal with a trade-off between identifying vertically transmitted representatives of multi-copy gene families (recognizable by synteny conservation) and retrieving complete sets of species-level orthologs. We studied the conceptual and practical implications of adopting homology, orthology, or synteny conservation as formal criteria for gene clustering by performing comparative analyses of 125 prokaryotic pangenomes. We found that clustering criteria affect pangenome functional characterization, core genome inference, and reconstruction of ancestral gene content to different extents. Pangenome and core genome sizes are affected by the same multiplicative factor in all species, which allows for robust across-species comparisons regardless of the clustering criterion. However, across-species comparisons of pangenome diversity, gene fluxes, and functional profiles are substantially affected by inconsistencies among clustering criteria. Such inconsistencies are driven not only by mobile genetic elements, but also by genes involved in defense, secondary metabolism, and other functions overrepresented in the accessory genome. In the case of pangenome diversity, the variability that can be attributed to methodological inconsistencies can exceed the effect sizes of ecological and phylogenetic variables. Our results emphasize the need to consider conceptual adequacy, not just computational performance, when designing workflows for pangenome analysis. We also provide a benchmarking dataset to assess the robustness and across-method reproducibility of future comparative studies.

Introduction

Recent advances in sequencing have revolutionized the study of microbial ecology and evolution by providing access to thousands of high-quality genomic sequences. A major consequence of the surge in microbial genomic data has been the introduction of the concept of pangenome to denote the set of all genes found in the genomes of a taxonomic group, typically defined at the species level (Tettelin et al. 2005), (Vernikos et al. 2015). Pangenomes have quickly gained importance in eco-evolutionary research because they provide valuable information about the set of functional capabilities accessible to the strains of the same species and their propensity to gain and lose genes. Therefore, the study of pangenomes can shed light on the interactions among genome plasticity, niche diversity, and adaptability to environmental changes (Shapiro 2017; Maistrenko et al. 2020; Liao et al. 2021; Whelan et al. 2021). Moreover, identifying the set of genes present in all the genomes of a species (the core genes) has proved helpful to delineate core functions and build high-resolution species trees (Lerat et al. 2003; Wu et al. 2013; Kapli et al. 2020)

Pangenome studies rely on a precise delineation of clusters of equivalent genes that represent meaningful units for functional and evolutionary analyses. The simplest class of equivalence is homology, defined for a set of genes as the property of sharing common ancestry. Beyond this qualitative definition, different types of homology can be established based on the evolutionary history of each gene. Because most gene families have experienced duplications throughout evolution, it is usual to find homologous genes with multiple representatives per genome, termed paralogs. Paralogs often display functional divergence and accelerated evolutionary rates (Petitjean

et al. 2017). As a result, homology alone does not guarantee functional and evolutionary homogeneity of gene clusters. Instead, for most purposes, it is desirable to subdivide homologous groups into a higher-resolution and more homogeneous class of equivalence, defined by orthology (Fitch 2000)(Gabaldón and Koonin 2013). Orthologs are genes that share a single common ancestor at the time of speciation (Fitch 2000; Koonin 2005). As a result, the orthology criterion can discriminate paralogs that duplicated before the last speciation event (Fig. 1), Because of the connection between orthology and speciation, orthology is the most natural grouping criterion for across-strain comparative genomics and phylogenomic analyses. In practice, however, it is challenging to apply a strict orthology criterion to pangenome studies due to the high computational burden of building thousands of phylogenetic trees including all candidate homologs and the difficulty to obtain accurate reference species trees. To circumvent these difficulties, heuristic algorithms and reference databases of orthologous genes have been developed over the last decade (Kristensen et al. 2011; Huerta-Cepas et al. 2017; Galperin et al. 2019). Together with homology and orthology, a third criterion to further refine equivalence among genes is synteny conservation. According to the synteny criterion, two genes (typically orthologs) are grouped together if they share the same gene neighborhood in different genomes. Notably, synteny conservation can help discriminate between vertically and horizontally transmitted copies of a gene, or between multiple orthologs resulting from within-species duplication (Fig. 1).

It is widely accepted that the quantitative results obtained in single-species pangenome analyses depend on the method (and parameters) used to build the pangenome. Such dependency can be traced back to the strategy adopted to identify gene clusters. As intuitively expected, the size of the pangenome increases and the size of the core genome decreases as the sequence similarity thresholds used to define homology become more and more stringent (Bayliss et al. 2019; Sitto and Battistuzzi 2020; Tonkin-Hill et al. 2020). However, comparative studies implicitly assume that qualitative trends and order relationships are robust with respect to the choice of a particular criterion. The existence of such qualitative robustness merits a critical evaluation, given the deep conceptual differences among homology, orthology, and synteny conservation, the heterogeneity of evolutionary rates across genes and species, and the high rates of horizontal gene transfer (HGT) observed in some genomes (Hao and Golding 2006; Puigbò et al. 2014), (Treangen and Rocha 2011).

The possibility that homology, orthology, and synteny-based gene clustering produce incongruent results in comparative pangenome studies is not just a technical caveat. Instead, intraspecific HGT and gene duplications raise the fundamental question of what equivalence class (homology, orthology, synteny, or something else) best captures the essentially dynamic nature of pangenome evolution (Cummins et al. 2022). This conceptual conundrum is especially clear in what concerns the splitting of multicopy orthologs based on their gene neighborhoods. Let us consider, for example, a gene that has experienced recent duplications or intra-specific HGT among some strains of the same species. For what purposes should all copies be included in the same group, following the standard orthology criterion? For what other purposes should the duplicated or horizontally transferred copy be assigned to a new cluster, based on synteny considerations? The orthology criterion appears better to assess pangenome diversity, for which it is not desirable to count a recently duplicated gene twice (Lees et al. 2019), or to study gene duplication and intraspecific HGT, which first requires classifying multicopy genes as members of the same group (Csurös 2010). In contrast, the synteny

criterion would be more appropriate to select marker genes for phylogenomic analyses or identify vertically transmitted members of mobile gene families.

In this study, we show that the choice of a particular clustering criterion can have notable consequences on downstream comparative analyses. To stress the fact that the optimal criterion is somewhat arbitrary and dependent on the main research goal, we adopt the method-agnostic term *Operational Gene Cluster* (OGC) as an umbrella that includes homologs, classical orthologs (possibly inferred through different methods), and vertically transmitted subsets of orthologs with conserved synteny. Depending on the particular choice, OGC may imply different degrees of functional equivalence and shared ancestry.

State-of-the-art methods for OGC construction implement different strategies to deal with sensitivity-vs-specificity trade-offs at manageable computational cost. The simplest approaches (e.g., MMseqs2 (Steinegger and Söding 2017) which is the OGC inference tool used by the pangenome analysis suite PanACoTA (Perrin and Rocha 2021)), build groups of homologous genes by applying a pre-defined similarity threshold to the amino acid sequences encoded by a set of open reading frames (ORFs). More sophisticated tools include additional steps to discriminate true orthology from other ways of homology. Such discrimination can be attained through two major strategies: (a) by building gene-level phylogenetic trees or sequence similarity networks, on which some heuristic rules are applied to resolve subsets of genes with shared ancestry (Altenhoff and Dessimoz 2012); and (b) by subclustering homologous groups based on their gene neighborhoods, under the assumption that synteny is locally conserved at the evolutionary timescales that are relevant to within-species diversification. By design, the former approach, which is adopted by reference databases like COG (Galperin et al. 2019), eggNOG (Huerta-Cepas et al. 2017), and the pangenome analysis tool panX (Ding et al. 2018), is better suited to assess true orthology in gene families affected by intraspecific duplications and HGT. However, its high computational burden makes it inefficient to deal with large genomic datasets. On the other side, synteny-based approaches (such as the one implemented by the popular pangenome analysis tools Roary (Page et al. 2015) and PanOCT (Fouts et al. 2012)) are faster and less resource intensive, although their results may deviate from the classical concept of orthology (by missing true orthologs) when applied to highly dynamical genomes or regions of the genome with poor synteny conservation. Phylogeny-aware and synteny-based methods are not mutually exclusive, although they are rarely performed together due to computational constraints.

Despite their distinct conceptual underpinnings, different types of OGC are sometimes just viewed as exchangeable heuristic approaches that approximate the concept of orthology in a computationally tractable manner. We tested to what extent such an assumption is true and found that some properties of the pangenome, mostly concerning its size and the identity of the core genome, are indeed robust. However, pangenome properties that are related to its fluidity can be greatly affected, leading to relatively poor correlation in the results of comparative genomic analyses conducted with different methods. Interestingly, this does not only affect mobile genetic elements, but also genes involved in non-selfish functions.

Results

Method-dependent variation and intrinsic uncertainty in pangenome size and diversity

We investigated 3 major methods for *de novo* species-wise OGC construction that qualitatively differ in the strategies for gene clustering and paralog discrimination (Table 1). All *de novo* methods start from a collection of open reading frames (ORF) and cluster them according to a predefined identity threshold that can be set to include more or less distant homologs. The resulting clusters are then processed to split paralogs into separate OGC. The three methods included in this study represent three alternative approaches implemented by some of the most popular tools for pangenome analysis: orthology-based clustering (implemented by panX), synteny-based clustering (implemented by Roary), and homology-based clustering (implemented by MMseqs2, which is the OGC construction module used by PanACoTA). For the latter two, we also explored the effect of introducing two different identity thresholds for the initial clustering.

In addition, we considered a fast orthology prediction method (eggNOG-mapper) based on mapping ORF to a reference database of orthologous groups. This reference-based method was not originally intended for pangenome analysis and has some important limitations. First, because the eggNOG database was built using only one representative genome per species, it does not account for variability across strains. Second, because orthologous groups in the eggNOG database were required to contain sequences from at least 3 species, mapping ORF to eggNOG automatically excludes sequences without known homologs and sets a hard limit for OGC taxonomic resolution at the genus level. Still, reference-based orthology assignments are nowadays highly efficient and scalable to large (meta)genomic datasets. Therefore, we included eggNOG-mapper in this study to assess its performance compared to *sensu stricto* pangenome reconstruction methods.

By applying these methods, we obtained 6 alternative sets of species-wise OGC for 124 bacterial and 1 archaeal species (DOI: 10.5281/zenodo.7093014). These species, defined according to the phylogenetically consistent classification scheme established by the Genome Taxonomy Database (GTDB) (Parks et al. 2018), were selected to cover every genus in the GTDB, with the condition that there were at least 15 high-quality genomes available per species. We quantified the discordance between the OGC produced by each pair of methods by means of the normalized variation of information, a measure that takes values between 0, if the two sets of OGC display a one-to-one correspondence, and 1, if they are completely independent. The hierarchical clustering of methods based on this measure reveals that the discordances, although small, are reproducible across species (Fig. 2a). The most notable differences arise between reference-based and *de novo* OGC. Among the latter, the strategy for paralog discrimination determines the resulting OGC to a greater extent than the identity threshold, at least for the relatively permissive thresholds implemented in the study.

Based on the previous finding, we restricted our analysis to a single identity threshold (80%) and investigated how paralog discrimination affects the properties of the inferred prokaryotic pangenomes. To that end, we selected 10 non-redundant quantitative features that represent different aspects of pangenome size and diversity. All *de novo* methods supported clustering these features in 4 groups (Fig. 2b) that can be interpreted in terms of (i) pangenome size, (ii) genome (and core genome) size, (iii) gene content diversity, and (iv) nucleotide diversity (evaluated in the core genes). As expected, reference-based OGC systematically produced the smallest estimates of

186 pangenome size and gene content diversity, whereas *de novo* synteny- and homology-based OGC
187 produced the largest estimates for those traits (Fig. 2c).

188 Although the differences in some estimates (especially those involving singletons) are substantial,
189 their practical impact on comparative analyses does not so much depend on their median
190 magnitude, but on the extent to which they generate unexplained variation across species. More
191 precisely, method-dependent variation appears if the choice of a particular set of OGC over another
192 does not affect all species by the same factor. Because there is no ground truth for the "correct" set
193 of OGC, method-dependent across-species variation constitutes an unavoidable source of
194 uncertainty in comparative pangenome analyses that only becomes visible when considering
195 multiple OGC construction methods. We assessed such uncertainty for 10 selected pangenome
196 features by pairing sets of OGC (corresponding to different strategies for paralog discrimination)
197 and quantifying the fraction of the total variance in the feature of interest obtained with one set
198 that remains unexplained after controlling for the values obtained with the other set. This measure
199 also serves as a proxy for the methodological inconsistency in the estimation of pangenome
200 properties. As shown in Fig. 2c-d (see also Suppl. Fig S1 and S2), core genome and pangenome sizes
201 are generally consistent across methods; that is, using one method or another affects all species by
202 the same factor. In consequence, methodological choices are not expected to affect relative
203 comparisons of core genome and pangenome sizes across species. In contrast, estimates of gene
204 content diversity display higher levels of inconsistency, with method-dependent uncertainty
205 accounting for roughly 20% of the total across-species variance (and up to 40% if using reference-
206 based OGC). The inconsistencies in genome fluidity are especially high in the case of Proteobacteria,
207 reaching around 50% of the total across-species variance.

208 A major limitation of reference-based approaches to OGC construction is that they are constrained
209 by the limited diversity of the reference database. Indeed, in most species, 5-10% of the ORF could
210 not be mapped to the eggNOG database and therefore were not assigned to any reference-based
211 OGC (Suppl Fig. S3). The fraction of missing ORF is larger in some taxa that are underrepresented or
212 absent from the reference database. The most extreme case, with >30% unmapped genes,
213 corresponds to *B. burgdorferi*, the causal agent of Lyme's disease. The poor performance of
214 reference database mapping in *B. burgdorferi* is explained by the unique structure of its genome,
215 which consists of a linear chromosome and >20 linear and circular plasmids without homologs in
216 other species (Fraser et al. 1997; Casjens et al. 2012). Despite these limitations, reference-based
217 OGC provide reasonably good estimates for the number of core genes per genome and allow
218 retrieving 85-90% of the single-copy core gene families identified by *de novo* approaches.

219 <u>Within-species paralogy and inference of core genomes</u>

220 Among the pangenome features considered in this work, the size of the core genome is the least
221 affected by the gene clustering criterion. However, a closer inspection of core gene families (defined
222 in a strict sense, given the high completeness of the genomes included in the study) reveals
223 differences in the mean copy number across methods (Fig. 3a). Almost all the core gene families
224 obtained by the synteny-based approach are single-copy, whereas those obtained by other methods
225 can sometimes harbor >1 copy per genome. Thus, the fraction of core gene families that appear as
226 single-copy ranges from 80% in reference-based to >99% in synteny-based OGC. In the case of
227 reference-based and *de novo* homology-based OGC, the distribution of the mean copy number per

228 core gene family per genome displays clear peaks at integer values, indicating the existence of
229 complete sets of duplicated paralogs that were not resolved by these methods.

230 We next focused on single-copy core (s.c.c.) genes, which are of greater practical interest as they
231 are often used to infer high-resolution phylogenies. Synteny-based methods systematically produce
232 5-10% more s.c.c. OGC than orthology-based methods (Fig. 3b). Typically, synteny-supported s.c.c.
233 genes belong to orthologous OGC that are core but not necessarily single-copy (Fig 3c). Therefore,
234 it appears that synteny criteria are effective in resolving single-copy representatives of core gene
235 families affected by within-species duplications or HGT. On the other side, s.c.c. OGC that are
236 orthologous but not supported by synteny tend to either contain incomplete ORFs (that could not
237 be processed by the synteny-based workflow) or split among 2 or more synteny-supported
238 accessory OGC (Fig. 3d). Such lack of synteny conservation could be a sign of non-vertical
239 transmission in s.c.c. genes exclusively detected through orthology criteria. However, given the
240 small fraction of these OGC in most species (<1% of all orthology-based s.c.c. OGC), the overall effect
241 of such potential non-vertical contamination in downstream analyses is possibly modest. The only
242 exception occurs in Mycoplasmatales, in which the relative contribution of non-synteny-supported
243 OGC is amplified by their small genome sizes.

244 Single-copy core genes supported by a single criterion (either synteny or orthology) are significantly
245 less conserved in terms of their average identity than those detected by both criteria (Fig. 3e);
246 Kruskal-Wallis omnibus test $p < 10^{-20}$). Among the s.c.c. genes that are method-exclusive, those
247 based on orthology are significantly less conserved than those based on synteny ($p < 10^{-8}$ for all
248 Mann-Whitney post-hoc tests with Tuckey's HSD correction). There are also significant differences
249 in the functional profiles of s.c.c. genes that are supported by only one or both criteria (Fig. 3f), with
250 overrepresentation of genes associated with mobile genetic elements, cell motility, and secondary
251 metabolism, and underrepresentation of genes involved in translation among method-exclusive
252 single-copy core genes (linear mixed effects model for isometric log-ratios; $F(2,248) > 190$, $q < 10^{-20}$
253 in all cases).

254 Systematic and species-specific biases in functional profiles

255 Estimates of genome content diversity are strongly affected by the choice of orthology- or synteny-
256 based strategies to discriminate paralogs when building *de novo* OGC. To better understand the
257 causes and implications of these differences, we classified ORF and OGC into 21 coarse-grained
258 functional categories that are representative of the main molecular and cellular processes that take
259 place in prokaryotic cells. For each functional category, we studied the agreement between
260 orthology- and synteny-based OGC by calculating the normalized variation of information, the
261 fraction of fully equivalent OGC (that is, OGC that contain exactly the same ORF regardless of the
262 strategy for paralog discrimination), and the fraction of ORF assigned to fully equivalent OGC (Fig.
263 4a and Suppl. Fig. S4). By far, the highest inconsistency occurs for mobile genetic elements, for which
264 only 25% of the OGC (encompassing 25% of the ORF) are equivalent. Besides mobile genetic
265 elements, moderate degrees of inconsistency are observed for defense systems, intracellular
266 trafficking/secretion, and replication/recombination/repair. On the other end, central cellular
267 functions, such as translation, transcription, nucleotide metabolism, and coenzyme metabolism
268 show the highest agreement, with 64-70% of the OGC (encompassing 77-84% of the ORF) being fully
269 equivalent. These trends are also manifest when looking at the absolute and relative numbers of

270 OGC per category (Fig. 4b), with synteny-based paralog discrimination producing a disproportional
271 excess of OGC associated with the mobilome. The fraction of OGC that contain ORF from more than
272 one functional category is consistently larger in the case of orthology-based OGC, although the
273 absolute differences are modest (around 0.5-1% in most categories; Suppl. Fig S4). Functionally
274 heterogeneous OGC are most often associated with signal transduction, cell cycle control/cell
275 division, mobile genetic elements, and unknown or poorly characterized functions.

276 Compositional analysis of functional profiles controlling for across-species variability confirms that
277 synteny-based paralog discrimination leads to a significant increase in the fraction of OGC
278 associated with mobile genetic elements (Fig. 4c; linear mixed effects model for isometric log-ratios;
279 ILR-balance difference = 0.49, $F(2,124) = 189$, $q < 10^{-20}$). Functional profiles derived from synteny
280 and orthology-based OGC also differ in the balance between central cellular functions (transcription,
281 translation, cell cycle, nucleotide and coenzyme metabolism) and other functional categories (ILR-
282 balance difference = -0.11, $F(2,124) = 132$, $q < 10^{-20}$); and between a set of functions including
283 secondary metabolism, carbohydrate metabolism, secretion, defense and recombination, and the
284 remaining functional categories (ILR-balance difference = 0.08, $F(2,124) = 62$, $q < 10^{-10}$). Apart from
285 these general trends, other significant differences between synteny- and orthology-based functional
286 profiles are restricted to specific categories in one or a few particular species (Fig. 4d), such as
287 defense in *Legionella pneumophila* (Z = 5.9, $q < 10^{-5}$), *Borreliella burgdorferi* (Z = 5.4, $q = 5\times10^{-5}$) and
288 *Bordetella pertussis* (Z = 4.4, $q = 0.002$), secondary metabolism in *Bacillus anthracis* (Z = 4.0, $q =$
289 0.008), and signal transduction in *Brachyspira hyodysenteriae* (Z = 3.8, $q = 0.019$).

290 Variability of gene flux estimates

291 To assess whether method-dependent variation in pangenome composition propagates to
292 downstream analyses, we investigated the effect of synteny- and orthology-based paralog
293 discrimination on a quantitative study of genome dynamics. To that purpose, we used the software
294 Gloome, that infers events of gene gain and loss along a lineage taking as inputs the strain-level
295 phylogenetic tree and a binary matrix with the presence and absence profiles of each OGC. As shown
296 in Fig. 5a, using synteny-based instead of orthology-based OGC leads to a 60% increase in the
297 estimated number of gene gains and losses per lineage. Genome vs gene change ratios, measured
298 as the number of expected gains and losses per gene per core nucleotide substitution are also higher
299 (40% increase) when using synteny-based OGC. In contrast, the ratio between gene gains and losses,
300 that determines the short-term dynamics of genome size (Sela et al. 2016), displays a more complex
301 response, with synteny-based OGC producing lower or higher estimates than those obtained with
302 orthology-based OGC depending on whether a species is dominated by gains or losses. Method-
303 dependent uncertainties account for 15%, 18%, and 30% of the across-species variability in the total
304 flux, the genome vs gene change ratio, and the gain vs loss ratio, respectively. These results indicate
305 that comparative analyses of short-term genome dynamics are highly sensitive to methodological
306 choices for paralog discrimination, especially when it comes to evaluate the balance between gene
307 gain and loss.

308 A deeper analysis of gene flux by functional categories reveals that inconsistencies between gene
309 clustering criteria (quantified as one minus the squared rank correlation of species-wise estimates
310 to account for the numerous outliers) are stronger among mobile genetic elements and genes
311 involved in secondary metabolism and inorganic ion transport (Fig. 4b). Although high

312 inconsistencies are also observed in central functional categories, such as translation, the practical
313 relevance of those is lesser due to the relatively low fluxes associated with those categories. If all
314 species are jointly considered by calculating their median, flux estimates obtained from synteny
315 OGC display a systematic deviation of >1 additional event per gene in all functional categories, which
316 is evenly distributed between gains and losses (Fig. 4c).

317

318 **Discussion**

319 Orthology is generally considered the optimal criterion for clustering gene sequences for
320 comparative genomics (Gabaldón and Koonin 2013). Because orthologs tend to conserve their
321 function and evolve consistently with speciation patterns, they are, at least in theory, the most
322 fitting choice for functional and phylogenomic studies. In practice, however, the use of orthology as
323 a gold standard for pangenome analysis faces technical and conceptual challenges. On the technical
324 side, distinguishing orthologs from paralogs is computationally costly. Therefore, to deal with large
325 (meta)genomic datasets, orthology prediction tools often resort to heuristic algorithms that risk
326 missing true orthologs or including out-paralogs (paralogs that duplicated before the last common
327 ancestor of the clade of interest). On the conceptual side, intra-species gene duplication and
328 horizontal gene transfer generate multi-copy gene lineages that, despite not violating the orthology
329 criterion at the species level, may confound downstream analyses that assume vertical transmission.
330 In those cases, a stricter criterion that only clusters together vertically transmitted members of a
331 gene family may be more desirable.

332 Motivated by these caveats, we introduced the general term "operational gene cluster" (OGC) to
333 denote meaningful units for pangenome reconstruction in a manner that is agnostic of the clustering
334 criterion. The term OGC highlights the operational nature of gene clusters generated by pangenome
335 analysis tools, which, despite being inspired in many cases by the concept of orthology, do not
336 necessarily produce orthologs in a strict sense. Different clustering tools use different formal criteria
337 and heuristics to generate OGC. The purpose of this study was to assess how three major clustering
338 criteria (homology, orthology, and synteny) implemented by popular pangenome analysis tools
339 affect pangenome reconstructions and downstream phylogenomic analyses. Despite we did not
340 conduct a comprehensive comparison of heuristic implementations, we tested several identity
341 thresholds for the initial clustering step and found that, as long as they are relatively permissive,
342 their impact on the results is minor compared to that of the subsequent paralog discrimination step.
343 Therefore, it appears that the underlying formal criterion for paralog discrimination (shared
344 ancestry at speciation for orthology, conserved gene neighborhood for synteny), rather than the
345 actual implementation, is what drives qualitative differences across methods.

346 Consistent with previous works, we observed that single-species estimates of pangenome size and
347 diversity strongly depend on the method used to build OGC (Bayliss et al. 2019; Sitto and Battistuzzi
348 2020; Tonkin-Hill et al. 2020). However, when conducting comparative analyses, inconsistencies in
349 trends and relative estimates across species are of much greater concern than absolute differences
350 in species-level estimates. In that regard, trends involving pangenome and core genome sizes are
351 generally robust across methods. More precisely, choosing one or another method affects all
352 species-wise size estimates by the same constant multiplicative factor. In contrast, across-species

353 comparisons of genome plasticity and pangenome diversity are highly sensitive with respect to the
354 clustering criterion, in a way that cannot be explained by any linear or nonlinear data
355 transformation. Such inconsistencies in pangenome diversity, that reach up to 50% of the total
356 across-species variance in Proteobacteria, should be a major source of concern for studies aimed at
357 understanding the forces that shape microbial pangenomes. For comparison, it has been estimated
358 that habitat and phylogeny contribute to explain approximately 20% of the across-species variance
359 in pangenome diversity (Maistrenko et al. 2020). The contribution of a particular gene clustering
360 method to pangenome variability can only be assessed by comparison with other methods, which
361 is often unfeasible in large datasets. As a result, methodological "noise" remains unnoticed,
362 potentially contributing a relevant fraction of the unexplained variance or, in the worst-case
363 scenario, acting as a confounding factor if method-dependent biases in OGC correlate with the
364 biological variables of interest.

365 The identification of single-copy core genes is one of the main outcomes of pangenome
366 reconstruction (Harris et al. 2021). Single-copy core genes allow building high-resolution strain trees
367 that can be very robust to the effects of unbiased homologous recombination (Stott and Bobay
368 2020) . However, to obtain accurate trees, it is fundamental that the OGC that correspond to single-
369 copy core genes are not contaminated by HGT (Kapli et al. 2020). As expected, synteny is the most
370 sensitive criterion to discriminate single-copy representatives among gene families affected by
371 paralogy and intra-species duplications. Nevertheless, *de novo* homology- and orthology-based
372 methods retrieve most (90-95%) synteny-supported single-copy core OGC with low (<1%) potential
373 contamination, which makes them good alternatives for many practical applications.

374 Differences in gene clustering criteria translate to downstream phylogenomic analyses involving
375 functional profiles and gene flux estimation. In particular, synteny-based OGC display substantial
376 fragmentation of non-core gene families, including not only mobile genetic elements, but also genes
377 involved in defense, secondary metabolism, signal transduction, and secretion, among others. The
378 fact that gene family fragmentation is not homogeneous across taxa and functional categories
379 implies that the choice of an inadequate criterion can bias comparative studies of pangenome
380 function and dynamics. That said, regardless of the method used to discriminate paralogs, the
381 lowest gene fluxes are always observed in categories associated with central cell functions, whereas
382 the highest fluxes correspond to mobile genetic elements and defense systems. Therefore, the
383 inverse association between gene flux and essentiality described by previous studies (Cohen and
384 Pupko 2010; Puigbò et al. 2014; Iranzo et al. 2017) appears as a robust feature of genome plasticity.

385 As a fast alternative to *de novo* clustering methods, we evaluated the performance of OGC built by
386 annotating gene sequences with the eggNOG reference database. Although reference-based OGC
387 deviate from *de novo* OGC in many aspects, they are a reasonable option to identify single-copy core
388 genes and generate functional profiles in well-studied genera, especially if paralogs and rare genes
389 can be disregarded. In contrast, eggNOG-based OGC can miss a significant portion of species-specific
390 core genes in poorly sampled taxa, do not discriminate paralogs below the genus level, and filter
391 out gene families with narrow taxonomic distribution. Accordingly, pangenome analyses using
392 reference-based OGC should justify their conclusions accounting for these limitations. Possibly, the
393 most interesting application of reference-based OGC is found in large-scale comparative studies that
394 require tracking OGC across species. That can be achieved thanks to the hierarchical structure of the

395 eggNOG database, that allows annotating orthologs at multiple taxonomic levels without additional
396 computational cost.

397 Our findings stress that, as long as choosing an operational definition of gene cluster remains
398 arbitrary, pangenome properties affected by method-dependent variability will be subject to
399 intrinsic uncertainty. To minimize such uncertainty, methodological choices for OGC construction
400 should be guided by the nature of the research goals and their conceptual underpinnings, rather
401 than by computational considerations alone. A careful selection of clustering criteria is especially
402 required in studies that involve comparison of functional profiles or inference of gene fluxes within
403 and across pangenomes. Furthermore, future studies aimed at exploring the role of adaptive and
404 non-adaptive processes in pangenome evolution should explicitly discuss how methodological
405 choices affect the interpretation of their results. For example, choosing synteny over classical
406 orthology as the clustering criterion may inflate the proportion of mobile genetic elements in the
407 pangenome, altering the balance between selfish and potentially beneficial accessory genes. In turn,
408 measures of genome plasticity based on binary (presence or absence) phyletic profiles will
409 underestimate the contribution of intra-species paralogs to the total gene flux unless synteny is
410 used to build the OGC.

411 Reusability and meta-analysis of pangenome datasets is currently hindered by the incompatibility
412 of OGC obtained by different methods. To address that limitation and foster future research, it is
413 critical to come to a consensus on a set of methods that cover the most relevant criteria for paralog
414 discrimination. A first step in that direction would be to establish multi-method databases that
415 facilitate assessing the robustness of pangenome analyses with respect to methodological choices.
416 We expect that the datasets generated in this study (publicly available in DOI:
417 10.5281/zenodo.7093014), though relatively limited in the number of species, will contribute to
418 achieving such a goal.

419 **Materials and Methods**

420 Genomic sequences

421 We parsed the metadata files of the Genome Taxonomy Database (GTDB) release 95 (Parks et al.
422 2020) to identify all the species (sensu GTDB) for which there were at least 15 high-quality available
423 genomes. High-quality genomes were defined according to the MIMAG criteria (Bowers et al. 2017)
424 plus the following more stringent filters: completeness >99% and contamination <1%, both assessed
425 through CheckM (Parks et al. 2015), mean contig length >5kb, and contig count <500. Metagenome-
426 assembled genomes (MAG) and single-amplified genomes (SAG) were not included in the analysis.
427 After applying these filters, we recovered 321 bacterial and 1 archaeal species, belonging to 125
428 different genera (sensu GTDB). To minimize possible taxonomical biases, only one representative
429 species per genus was selected for subsequent analyses. For those genera with >1 suitable species,
430 we kept the species with the highest number of high-quality genomes, as they presumably were the
431 most informative for pangenome studies.

432 To keep our analyses within a manageable computational cost, species with >100 high-quality
433 genomes were subsampled to keep at most 100 genomes per species. To that end, we separately
434 aligned the amino acid sequences of 120 nearly universal marker genes employed by the GTDB with
435 MAFFT (Katoh and Standley 2013), concatenated the alignments, and ran IQ-Tree (Nguyen et al.
436 2015) to obtain phylogenetic trees including all the strains of the same species. Then, we applied a
437 heuristic subsampling strategy that maximized the diversity of the subsampled genomes whilst
438 accurately reflecting the topology of the strain trees.

439 The final dataset comprised 6,796 bacterial genomes belonging to 124 species and 55 archaeal
440 genomes belonging to 1 species. The genomes were downloaded from the NCBI FTP site. Open
441 reading frames (ORF) were predicted with Prodigal v2.6.3 (Hyatt et al. 2010), using the "single"
442 mode that is recommended for finished genomes and quality draft genomes. All runs used the
443 bacterial, archaeal and plant plastid code 11
444 (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi), which maps UAA, UGA, and UAG to
445 stop codons, except for *Mycoplasmopsis bovis PG45* (GCF_000183385.1) and *Mycoplasma*
446 *pneumoniae FH* (GCF_001272835.1), which used code 4 and translate UGA to tryptophan.

447 De novo OGC construction

448 We generated five sets of *de novo* species-level OGC, each one representing an alternative approach
449 to identify homologous genes and discriminate within-species paralogs (Table 1). To facilitate
450 comparison among methods, we started in all cases from the same predicted ORF previously
451 obtained with Prodigal, overriding any optional ORF prediction step provided by the OGC
452 construction software.

453 Two sets of homology-based OGC were built with the sequence clustering tool MMseqs2
454 (Steinegger and Söding 2017), setting the minimum identity threshold to 50% and 80% (additional
455 options 'easy-cluster' for cascaded clustering and --cluster-mode 1 to define clusters based on
456 connected components). MMseqs2 does not perform any kind of paralog splitting; therefore, the
457 resolution of the resulting OGC only depends on the similarity threshold used for clustering.

458 Two sets of synteny-based OGC were built with the software Roary (Page et al. 2015) , setting the
459 minimum identity threshold to 80% and 95%. The Roary algorithm starts by pre-clustering highly-
460 similar protein sequences with CD-HIT (Fu et al. 2012) to obtain a smaller set of representative
461 sequences. Roary then conducts an all-against-all comparison with BLAST and filters the hits based
462 on the user-provided identity threshold. Based on the network of hits, representative sequences are
463 clustered with MCL (Enright et al. 2002) and the resulting clusters are merged with the pre-clusters.
464 As a final step, Roary uses conserved gene neighborhood information to split homologous groups
465 containing paralogs into groups of synteny-supported OGC.

466 One set of orthology-based OGC was built with the software panX (Ding et al. 2018). The panX
467 algorithm clusters sequences in orthologous groups by performing an all-against-all similarity search
468 with DIAMOND (Buchfink et al. 2021) and posterior clustering with MCL. The hits retrieved by
469 DIAMOND are filtered only in terms of statistical significance, regardless of sequence identity, which
470 allows recovering relatively divergent homologs. The sequences of these initial clusters are aligned
471 with MAFFT (Katoh and Standley 2013) and cluster-level phylogenetic trees are built with FastTree
472 (Price et al. 2010). Finally, panX obtains orthology-supported OGC by examining the resulting trees
473 and applying a set of heuristic rules to split paralogs from true species-level orthologs.

474 <u>OGC construction by reference database mapping</u>

475 Reference-based OGC were built by mapping translated ORF to the eggNOG (evolutionary genealogy
476 of genes: Non-supervised Orthologous Groups) database version 5.0 (Huerta-Cepas et al. 2019). For
477 that purpose, we ran eggNOG-mapper v2 (Cantalapiedra et al. 2021) with command line options "-
478 *m diamond --tax_scope_mode narrowest*" to search queries against eggNOG sequences with
479 DIAMOND and transfer orthologous group annotations at the highest possible taxonomic resolution
480 (which typically corresponds to the genus level).

481 <u>Pangenome features</u>

482 OGC presence-absence matrices (sometimes known as phyletic profiles) and gene-to-OCG
483 relationships were used to estimate a set of pangenome features that intend to capture both the
484 size and the diversity of the pangenome.

485 Pangenome size was calculated as the total number of OGC retrieved for each species. Because this
486 measure is positively correlated with the number of genomes sampled, an unbiased estimate was
487 obtained by randomly subsampling sets of 14 genomes and taking the average over 100 realizations
488 (we refer to this and other pangenome features obtained with the same procedure as "14-mean").
489 To ensure that the results obtained with these two measures were sufficiently representative, we
490 calculated three additional measures of pangenome size: Chao's lower bound (Chao 1987); the
491 normalized pangenome size, obtained by dividing the pangenome size (P_{tot}) by the sum of the
492 harmonic series of the number of genomes (n), such that $P_{norm} = \frac{P_{tot}}{\sum_{i=1}^{n-1} 1/i}$ (Bobay and Ochman
493 2018); and the pangenome size of the 15 most dissimilar genomes in terms of genome content.
494 After verifying that all those metrics were highly correlated across species (R>0.95), we proceeded
495 with the uncorrected and 14-mean pangenome sizes, which are simpler and more easily
496 interpretable.

497 We defined the core genome as the set of OGC contained in all the genomes of a species. This strict
498 definition is appropriate given the almost full (>99%) completeness filter imposed on high-quality
499 genomes. Pangenomes were also characterized in terms of the absolute number of single-copy core
500 OGC, accessory OGC (those that are not core), and singleton OGC (those encompassing a single
501 ORF). As measures of genome content diversity, we computed the fraction of a genome's OGC that
502 are accessory and singleton, averaged over 100 random subsamples of 14 genomes per species. In
503 addition, a measure of genomic fluidity that quantifies the average dissimilarity in gene content of
504 randomly sampled pairs of genomes (Kislyuk et al. 2011) was obtained with the 'fluidity' function of
505 the R package micropan, setting the option 'n.sim = 500' (Snipen and Liland 2015).

506 The species-wise core gene alignment similarity (CGAS) was calculated using ORFs that belong to
507 single-copy core OGC. Pairwise global alignments of all the ORFs assigned to the same OGC were
508 performed with the Needleman-Wunsch algorithm as implemented by the *needleall* program of the
509 EMBOSS suite (Madeira et al. 2019). After removing all self-alignments (alignments of ORFs with
510 themselves), the average identity of each pair of genomes was calculated as $(\sum m_i)/(\sum L_i)$, where
511 m_i and L_i represent the number of matches and total alignment length for the pair of sequences
512 from the i-th OGC and the sum extends over all single-copy core OGC. The species-wise CGAS was
513 obtained as the average over all pairs of genomes of the same species. The nucleotide sequence
514 divergence in single-copy core genes was defined as one minus the species-wise CGAS. Given the
515 high computational cost of this procedure, calculation of CGAS and core gene divergences was
516 restricted to reference-, synteny-, and orthology-based OGC.

517 <u>High-resolution species trees and inference of gene gains and losses</u>

518 High resolution trees were obtained for each species using concatenated alignments of all single-
519 copy core OGC obtained by Roary. Amino acid sequences for each single-copy core OGC were
520 aligned with mafft-linsi (L-INS-I Algorithm, default options) (Katoh and Standley 2013) and back-
521 translated to nucleotide alignments with the program pal2nal.pl using codon table 11 (except for
522 species of the order Mycoplasmatales, which use codon table 4) (Suyama et al. 2006). The
523 nucleotide alignments were concatenated and used as input for the tree construction program
524 FastTree (command line options -gtr -nt -gamma -nosupport -mlacc 2 -slownni) (Price et al. 2010).
525 The preliminary tree produced by FastTree was subsequently provided to RAxML for branch length
526 optimization (options -f e -c 25 -m GTRGAMMA) (Stamatakis 2014). The package ETE3 (Huerta-Cepas
527 et al. 2016) was used for mid-point rooting and visualization.

528 Gene gains and losses along each species tree were inferred with the phylogenomic reconstruction
529 software Gloome (Cohen et al. 2010). As the input for Gloome, we used the phyletic profiles for the
530 presence or absence of each OGC and the high-resolution species trees; options were set to optimize
531 the tree branch lengths under a genome evolution model with 4 categories of gamma-distributed
532 gain and loss rates and stationary frequencies at the root. To compare among OGC generation
533 methods, we varied the input phyletic profiles according to the desired method while keeping the
534 species trees unchanged. The Gloome optimization algorithm did not converge for *Enterobacter*
535 *himalayensis* and *Chlamydia muridarum*; therefore, those species were excluded from all the
536 analyses involving gene gains and losses.

537

538 Functional annotation and statistical analysis of functional profiles

539 Functional annotation at the gene level was done by mapping individual genes to custom-made
540 HMM profiles of the 2020 release of the COG database (Galperin et al. 2021). Functional annotation
541 at the OGC level was done by applying the majority rule to gene-level annotations. Coarse-grained
542 pangenome functional profiles were built by counting the number of OGC assigned to each of the
543 21 major prokaryotic functional categories defined in the COG database.

544 The statistical analysis of functional profiles was conducted by applying the phylofactorization
545 framework (Washburne et al. 2017). First, to account for the constant-sum constraint that
546 complicates the statistical analysis of compositional data, we applied the isometric log-ratio (ILR)
547 transform to the species-wise functional profiles. Informative ILR balances were defined following a
548 guide tree that was obtained by calculating the mean differences between synteny and orthology
549 OGC for each functional category and performing hierarchical clustering of the functional categories
550 based on such differences. Then, linear mixed effects models were set out for each of the 20 ILR
551 balances, with OGC clustering criteria as fixed effects and species as random effects. Model fitting
552 was performed with the R package lmerTest (Kuznetsova et al. 2017) and contrasts among balances
553 were conducted with the R package phylofactor (https://github.com/reptalex/phylofactor) , which
554 ranks the balances based on the fraction of the total variance that is explained by the model.
555 Statistical significance was calculated by using Satterthwaite's approach to estimate the degrees of
556 freedom of the F-statistic and applying Bonferroni correction to account for multiple comparisons.

Acknowledgements

The work of Y.L. was supported by the China Scholarship Council (CSC Grant No. 202008440425). J.I. is supported by the Ramón y Cajal Programme of the Spanish Ministry of Science (Grant No. RYC-2017-22524), the Agencia Estatal de Investigación of Spain (Grant No. PID2019-106618GA-I00), the Severo Ochoa Programme for Centres of Excellence in R&D of the Agencia Estatal de Investigación of Spain (Grant No. SEV-2016-0672 (2017–2021) to the CBGP), the Comunidad de Madrid (through the call Research Grants for Young Investigators from Universidad Politécnica de Madrid, Grant No. M190020074JIIS); and the National Programme for Fostering Excellence in Scientific and Technical Research (grant No. PGC2018-098073-A-I00 MCIU/AEI/FEDER, UE; to J.H-C.)

Data Availability

The data underlying this article are available in Zenodo, at https://dx.doi.org/10.5281/zenodo.7093014

570 **Figure legends**

571 **Figure 1: Homology, species-level orthology, and synteny conservation.** The phylogeny of a gene
572 family that has experienced intra-species duplication (†), intra-species HGT (*), and inter-species
573 HGT (**) is superimposed on a species tree (colored ribbons). In the species tree, speciation is
574 followed by diversification into two strains (blue), and a sister species is represented in green. Gene
575 neighborhoods are displayed at the leaves of the gene tree.

576 **Figure 2: Method-dependent variation and uncertainty in pangenome features.** (a) Consensus
577 similarity tree of OGC building methods based on the species-wise normalized variation of
578 information for the assignation of ORF to OGC. Labels indicate the number of species (out of 125)
579 that support each branch. (b) Consensus similarity tree of different pangenome features based on
580 pairwise, unsigned correlations. Labels indicate the number of OGC building methods (out of 4) that
581 support each branch (values <3 are not shown). In all cases when the support is not complete,
582 reference database mapping is the method that disagrees. (c) Quantitative comparison of
583 pangenome estimates among methods. Left: relative differences in pangenome estimates; right:
584 relative contribution of methodological choices to across-species variance. Note the different color
585 scale for Proteobacteria. Highlighted cells correspond to the features shown in (d). (d) Species-wise
586 comparison of selected pangenome features (pangenome size, number of single-copy core OGC,
587 fluidity, and nucleotide sequence diversity in core genes) inferred from orthology- and synteny-
588 based OGC. Black lines show the orthogonal least squares fit; gray lines indicate the 1:1 trend. Each
589 point corresponds to the pangenome of one species, colored according to its phylum.

590 **Figure 3: Comparison of core gene sets obtained from different OGC building methods.** (a)
591 Distribution of the mean copy number per core OGC per genome. (b) Fraction of all single-copy core
592 (scc) gene families that are exclusively recovered through orthology- or synteny-based paralog
593 splitting. (c) Distribution of synteny-exclusive, single-copy core OGC among orthology-based OGC.
594 The large bar at "1" implies that most synteny-exclusive single-copy core OGC are subsets of larger
595 orthology-based OGC that are core but not single-copy. (d) Distribution of orthology-exclusive,
596 single-core OGC among synteny-based OGC. Most of the single-copy core OGC exclusively supported
597 by orthology combine 2 accessory synteny-based OGC. In (c) and (d), gray bars denote method-
598 exclusive single-copy core OGC that include sequences that could not be processed by the other
599 method. (e) Nucleotide sequence divergence in single-copy core OGC exclusively supported by
600 orthology, synteny, or both criteria. To account for across-species differences in evolutionary rates
601 and phylogenetic tree spans, values were normalized by the species-level mean divergence
602 (calculated from all single-copy core OGC that were supported by both synteny and orthology).
603 Vertical lines indicate the distribution means for each group of single-copy core OGC. (f) Functional
604 differences among single-copy core OGC supported by different criteria. Each set of box plots
605 represents the balance (measured as the isometric log-ratio) between the relative frequencies of a
606 given functional category (x-axis) and the remaining categories not considered in previous sets (e.g.,
607 the second set of box plots corresponds to the balance between functional category N and all the
608 rest except X). The figure shows the 4 ILR balances with the greatest variation across methods.
609 Abbreviations of functional categories, X: mobilome; N: cell motility; Q: 2º metabolites biosynthesis,
610 transport and catabolism; J: translation, ribosomal structure and biogenesis. In (b) and (f), each data
611 point corresponds to one species; boxes span the 25-75 percentiles; the central line indicates the

612 median; whiskers extend to the most extreme data points that are not outliers; isolated points
613 denote outliers; notches (only in f) show the 95% confidence interval of the median.

614 **Figure 4: Systematic and specific biases in functional profiles associated with paralog splitting**
615 **criteria.** (a) Inconsistency of ORF assignations into OGC (normalized variation of information, top),
616 and fraction of OGC that exactly contain the same ORFs (bottom) under synteny and orthology
617 splitting criteria, stratified by functional category. (b) Absolute number (top) and relative fraction
618 (bottom) of synteny- and orthology-based OGC associated with each functional category. (c)
619 Balances (quantified as isometric log-ratios) for the functional categories that show the greatest
620 systematic variation between paralog splitting criteria. Each set of boxplots represents the balance
621 between the relative abundances of a group of functional categories (shown below) and all the
622 remaining categories not considered in previous sets. Each data point corresponds to the
623 pangenome of one species; boxes span the 25-75 percentiles; the central line indicates the median;
624 whiskers extend to the most extreme data points that are not outliers; isolated points denote
625 outliers. (d) Standardized residuals (Z-scores) of the linear mixed effects model used to infer the
626 systematic differences shown in (c). Each row corresponds to the pangenome of one species, sorted
627 according to the GDTB species tree (Parks et al. 2022) (phyla colored as in Fig 1). Colored cells
628 indicate a significant excess of synteny- (green) or orthology-based (purple) OGC from a given
629 category in a specific pangenome that is not explained by the general trends in (c). Abbreviations of
630 functional categories, C: energy production and conversion; D: cell cycle control, cell division,
631 chromosome partitioning; E: amino acid transport and metabolism; F: nucleotide transport and
632 metabolism; G: carbohydrate transport and metabolism; H: coenzyme transport and metabolism; I:
633 lipid transport and metabolism; J: translation, ribosomal structure and biogenesis; K: transcription;
634 L: replication, recombination and repair; M: cell wall/membrane/envelope biogenesis; N: cell
635 motility; O: posttranslational modification, protein turnover, chaperones; P: inorganic ion transport
636 and metabolism; Q: secondary metabolites biosynthesis, transport and catabolism; R: general
637 function prediction only; S: function unknown; T: signal transduction mechanisms; U: intracellular
638 trafficking, secretion, and vesicular transport; V: defense mechanisms; X: mobilome: prophages,
639 transposons.

640 **Figure 5: Effect of paralog splitting criteria on the inference of gene flux**. (a) Species-wise
641 comparison of the total gene flux (gains and losses along the species tree, left), genomic content vs
642 gene sequence evolution ratio (middle), and total gain vs loss ratio (right) inferred from synteny-
643 and orthology-based OGC. The method-dependent uncertainty is equal to $1 - R^2$. (b) Flux per gene
644 per functional category inferred from orthology- (top) and synteny-based (middle) OGC. Each data
645 point corresponds to one species; boxes span the 25-75 percentiles; the central line indicates the
646 median; whiskers extend to the most extreme data points that are not outliers; isolated points
647 denote outliers. The bar plot at the bottom shows the inconsistency between methods, quantified
648 as one minus the squared rank correlation. Abbreviations of functional categories as in Fig. 2. (c)
649 Median flux per gene per category, calculated over all the species, for orthology-based (x-axis) and
650 synteny-based (y-axis) OGC. Similar trends are observed for gains, losses, and the combination of
651 both.

652 **Table 1:** OGC generation strategies and tools used in this study.

Clustering	Paralog splitting	Identity threshold	Software	Ref.
de novo	orthology-based	none (e-val < 0.001)	panX	Ding et al. (2018)
de novo	synteny-based	80%, 95%	roary	Page et al. (2015); Ding et al. (2018)
de novo	none	50%, 80%	MMseqs2, PanACoTA	Steinegger and Söding (2017); Perrin and Rocha (2021)
reference db	n.a.	none (e-val < 0.001)	eggNOG-mapper	Cantalapiedra et al. (2021)

653

654 **References**

655 Altenhoff AM, Dessimoz C. 2012. Inferring orthology and paralogy. *Methods Mol. Biol.* 855:259–
656 279. Available from: http://dx.doi.org/10.1007/978-1-61779-582-4_9

657 Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. 2019. PIRATE: A fast and scalable
658 pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* 8. Available
659 from: http://dx.doi.org/10.1093/gigascience/giz119

660 Bobay L-M, Ochman H. 2018. Factors driving effective population size and pan-genome evolution
661 in bacteria. *BMC Evol. Biol.* 18:153. Available from: http://dx.doi.org/10.1186/s12862-018-
662 1272-4

663 Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J,
664 Rivers AR, Eloe-Fadrosh EA, et al. 2017. Minimum information about a single amplified
665 genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.
666 *Nat. Biotechnol.* 35:725–731. Available from: http://dx.doi.org/10.1038/nbt.3893

667 Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using
668 DIAMOND. *Nat. Methods* 18:366–368. Available from: http://dx.doi.org/10.1038/s41592-
669 021-01101-x

670 Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2:
671 Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic
672 Scale. *Mol. Biol. Evol.* 38:5825–5829. Available from:
673 http://dx.doi.org/10.1093/molbev/msab293

674 Casjens SR, Mongodin EF, Qiu W-G, Luft BJ, Schutzer SE, Gilcrease EB, Huang WM, Vujadinovic M,
675 Aron JK, Vargas LC, et al. 2012. Genome stability of Lyme disease spirochetes: comparative
676 genomics of Borrelia burgdorferi plasmids. *PLoS One* 7:e33280. Available from:
677 http://dx.doi.org/10.1371/journal.pone.0033280

678 Chao A. 1987. Estimating the population size for capture-recapture data with unequal catchability.
679 *Biometrics* 43. Available from: https://pubmed.ncbi.nlm.nih.gov/3427163/

680 Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. 2010. GLOOME: gain loss mapping engine.
681 *Bioinformatics* 26:2914–2915. Available from:
682 http://dx.doi.org/10.1093/bioinformatics/btq549

683 Cohen O, Pupko T. 2010. Inference and characterization of horizontally transferred gene families
684 using stochastic mapping. *Mol. Biol. Evol.* 27:703–713. Available from:
685 http://dx.doi.org/10.1093/molbev/msp240

686 Csurös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and
687 likelihood. *Bioinformatics* 26:1910–1912. Available from:
688 http://dx.doi.org/10.1093/bioinformatics/btq315

689 Cummins EA, Hall RJ, McInerney JO, McNally A. 2022. Prokaryote pangenomes are dynamic
690 entities. *Curr. Opin. Microbiol.* 66:73–78. Available from:
691 http://dx.doi.org/10.1016/j.mib.2022.01.005

692 Ding W, Baumdicker F, Neher RA. 2018. panX: pan-genome analysis and exploration. *Nucleic Acids*
693 *Res.* 46:e5. Available from: http://dx.doi.org/10.1093/nar/gkx977

694 Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of
695 protein families. *Nucleic Acids Res.* 30. Available from:
696 https://pubmed.ncbi.nlm.nih.gov/11917018/

697 Fitch WM. 2000. Homology a personal view on some of the problems. *Trends Genet.* 16:227–231.
698 Available from: http://dx.doi.org/10.1016/s0168-9525(00)02005-9

699 Fouts DE, Brinkac L, Beck E, Inman J, Sutton G. 2012. PanOCT: automated clustering of orthologs
700 using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely
701 related species. *Nucleic Acids Res.* 40:e172. Available from:
702 http://dx.doi.org/10.1093/nar/gks757

703 Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson
704 R, Hickey EK, et al. 1997. Genomic sequence of a Lyme disease spirochaete, Borrelia
705 burgdorferi. *Nature* 390:580–586. Available from: http://dx.doi.org/10.1038/37551

706 Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation
707 sequencing data. *Bioinformatics* 28:3150–3152. Available from:
708 https://academic.oup.com/bioinformatics/article-pdf/28/23/3150/18529929/bts565.pdf

709 Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nat.*
710 *Rev. Genet.* 14:360–366. Available from: http://dx.doi.org/10.1038/nrg3456

711 Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. 2019. Microbial genome analysis:
712 the COG approach. *Brief. Bioinform.* 20:1063–1070. Available from:
713 http://dx.doi.org/10.1093/bib/bbx117

714 Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. 2021. COG database
715 update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic*
716 *Acids Res.* 49:D274–D281. Available from: http://dx.doi.org/10.1093/nar/gkaa1018

717 Hao W, Golding GB. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation
718 or death. *Genome Res.* 16:636–643. Available from: http://dx.doi.org/10.1101/gr.4746406

719 Harris CD, Torrance EL, Raymann K, Bobay L-M. 2021. CoreCruncher: Fast and Robust Construction
720 of Core Genomes in Large Prokaryotic Data Sets. *Mol. Biol. Evol.* 38:727–734. Available from:
721 http://dx.doi.org/10.1093/molbev/msaa224

722 Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast
723 Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper.
724 *Mol. Biol. Evol.* 34:2115–2122. Available from: http://dx.doi.org/10.1093/molbev/msx148

725 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of
726 Phylogenomic Data. *Mol. Biol. Evol.* 33:1635–1638. Available from:
727 http://dx.doi.org/10.1093/molbev/msw046

728 Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic

729 I, Rattei T, Jensen LJ, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically
730 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*
731 47:D309–D314. Available from: http://dx.doi.org/10.1093/nar/gky1085

732 Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene
733 recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. Available
734 from: http://dx.doi.org/10.1186/1471-2105-11-119

735 Iranzo J, Cuesta JA, Manrubia S, Katsnelson MI, Koonin EV. 2017. Disentangling the effects of
736 selection and loss bias on gene dynamics. *Proc. Natl. Acad. Sci. U. S. A.* 114:E5616–E5624.
737 Available from: http://dx.doi.org/10.1073/pnas.1704925114

738 Kapli P, Yang Z, Telford MJ. 2020. Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.*
739 21:428–444. Available from: http://dx.doi.org/10.1038/s41576-020-0233-0

740 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
741 improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780. Available from:
742 http://dx.doi.org/10.1093/molbev/mst010

743 Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. 2011. Genomic fluidity: an integrative view of
744 gene diversity within microbial populations. *BMC Genomics* 12:1–10. Available from:
745 https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-32

746 Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39:309–338.
747 Available from: http://dx.doi.org/10.1146/annurev.genet.39.073003.114725

748 Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. 2011. Computational methods for Gene
749 Orthology inference. *Brief. Bioinform.* 12:379–391. Available from:
750 http://dx.doi.org/10.1093/bib/bbr030

751 Kuznetsova A, Brockhoff PB, Christensen RHB. 2017. LmerTest package: Tests in linear mixed
752 effects models. *J. Stat. Softw.* 82. Available from: http://www.jstatsoft.org/v82/i13/

753 Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher
754 NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.*
755 29:304–316. Available from: http://dx.doi.org/10.1101/gr.241455.118

756 Lerat E, Daubin V, Moran NA. 2003. From gene trees to organismal phylogeny in prokaryotes: the
757 case of the gamma-Proteobacteria. *PLoS Biol.* 1:E19. Available from:
758 http://dx.doi.org/10.1371/journal.pbio.0000019

759 Liao J, Guo X, Weller DL, Pollak S, Buckley DH, Wiedmann M, Cordero OX. 2021. Nationwide
760 genomic atlas of soil-dwelling Listeria reveals effects of selection and population ecology on
761 pangenome evolution. *Nat Microbiol* 6:1021–1030. Available from:
762 http://dx.doi.org/10.1038/s41564-021-00935-7

763 Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC,
764 Finn RD, et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic*
765 *Acids Res.* 47:W636–W641. Available from:
766 https://academic.oup.com/nar/article/47/W1/W636/5446251

767 Maistrenko OM, Mende DR, Luetge M, Hildebrand F, Schmidt TSB, Li SS, Rodrigues JFM, von
768 Mering C, Pedro Coelho L, Huerta-Cepas J, et al. 2020. Disentangling the impact of
769 environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J.*
770 14:1247–1259. Available from: http://dx.doi.org/10.1038/s41396-020-0600-z

771 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic
772 algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32:268–274.
773 Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu300

774 Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA,
775 Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*
776 31:3691–3693. Available from: http://dx.doi.org/10.1093/bioinformatics/btv421

777 Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete
778 domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* 38:1079–1086.
779 Available from: https://www.nature.com/articles/s41587-020-0501-8

780 Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. 2022. GTDB: an ongoing
781 census of bacterial and archaeal diversity through a phylogenetically consistent, rank
782 normalized and complete genome-based taxonomy. *Nucleic Acids Res.* 50:D785–D794.
783 Available from: http://dx.doi.org/10.1093/nar/gkab776

784 Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A
785 standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of
786 life. *Nat. Biotechnol.* 36:996–1004. Available from: http://dx.doi.org/10.1038/nbt.4229

787 Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the
788 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
789 *Genome Res.* 25:1043–1055. Available from: http://dx.doi.org/10.1101/gr.186072.114

790 Perrin A, Rocha EPC. 2021. PanACoTA: a modular tool for massive microbial comparative
791 genomics. *NAR Genom Bioinform* 3:lqaa106. Available from:
792 http://dx.doi.org/10.1093/nargab/lqaa106

793 Petitjean C, Makarova KS, Wolf YI, Koonin EV. 2017. Extreme Deviations from Expected
794 Evolutionary Rates in Archaeal Protein Families. *Genome Biol. Evol.* 9:2791–2811. Available
795 from: http://dx.doi.org/10.1093/gbe/evx189

796 Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large
797 alignments. *PLoS One* 5. Available from: https://pubmed.ncbi.nlm.nih.gov/20224823/

798 Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil:
799 quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* 12:66. Available
800 from: http://dx.doi.org/10.1186/s12915-014-0066-4

801 Sela I, Wolf YI, Koonin EV. 2016. Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U.*
802 *S. A.* 113:11399–11407. Available from: http://dx.doi.org/10.1073/pnas.1614083113

803 Shapiro BJ. 2017. The population genetics of pangenomes. *Nat Microbiol* 2:1574. Available from:
804 http://dx.doi.org/10.1038/s41564-017-0066-6

805 Sitto F, Battistuzzi FU. 2020. Estimating Pangenomes with Roary. *Mol. Biol. Evol.* 37:933–939.
806 Available from: http://dx.doi.org/10.1093/molbev/msz284

807 Snipen L, Liland KH. 2015. micropan: an R-package for microbial pan-genomics. *BMC*
808 *Bioinformatics* 16:79. Available from: http://dx.doi.org/10.1186/s12859-015-0517-0

809 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
810 phylogenies. *Bioinformatics* 30:1312–1313. Available from:
811 http://dx.doi.org/10.1093/bioinformatics/btu033

812 Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the
813 analysis of massive data sets. *Nat. Biotechnol.* 35:1026–1028. Available from:
814 http://dx.doi.org/10.1038/nbt.3988

815 Stott CM, Bobay L-M. 2020. Impact of homologous recombination on core genome phylogenies.
816 *BMC Genomics* 21:829. Available from: http://dx.doi.org/10.1186/s12864-020-07262-x

817 Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments
818 into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612. Available from:
819 http://dx.doi.org/10.1093/nar/gkl315

820 Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones
821 AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus
822 agalactiae: implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci. U. S. A.*
823 102:13950–13955. Available from: http://dx.doi.org/10.1073/pnas.0506758102

824 Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, Gladstone RA, Lo S, Beaudoin
825 C, Floto RA, et al. 2020. Producing polished prokaryotic pangenomes with the Panaroo
826 pipeline. *Genome Biol.* 21:180. Available from: http://dx.doi.org/10.1186/s13059-020-02090-
827 4

828 Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein
829 families in prokaryotes. *PLoS Genet.* 7:e1001284. Available from:
830 http://dx.doi.org/10.1371/journal.pgen.1001284

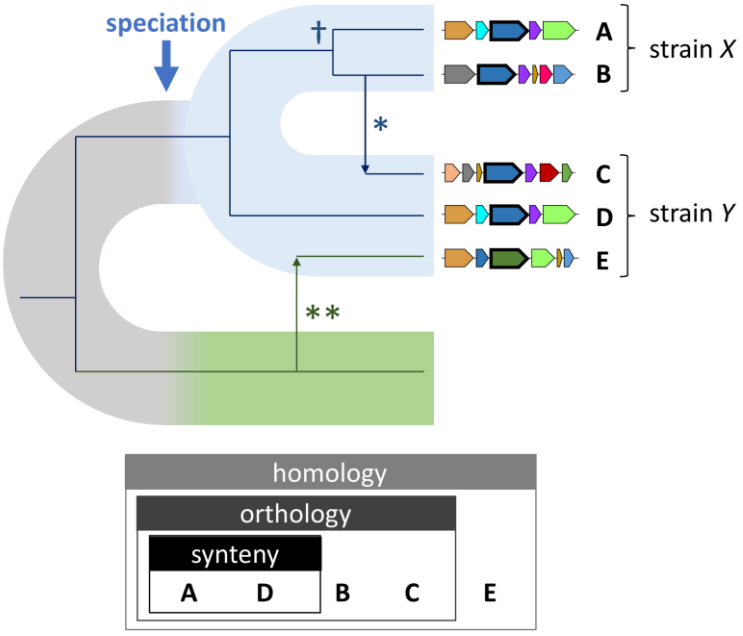
831 Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. *Curr. Opin.*
832 *Microbiol.* 23:148–154. Available from: http://dx.doi.org/10.1016/j.mib.2014.11.016

833 Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. 2017.
834 Phylogenetic factorization of compositional data yields lineage-level associations in
835 microbiome datasets. *PeerJ* 5:e2969. Available from: http://dx.doi.org/10.7717/peerj.2969

836 Whelan FJ, Hall RJ, McInerney JO. 2021. Evidence for Selection in the Abundant Accessory Gene
837 Content of a Prokaryote Pangenome. *Mol. Biol. Evol.* 38:3697–3708. Available from:
838 http://dx.doi.org/10.1093/molbev/msab139

839 Wu D, Jospin G, Eisen JA. 2013. Systematic identification of gene families for use as "markers" for
840 phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major
841 subgroups. *PLoS One* 8:e77033. Available from:
842 http://dx.doi.org/10.1371/journal.pone.0077033

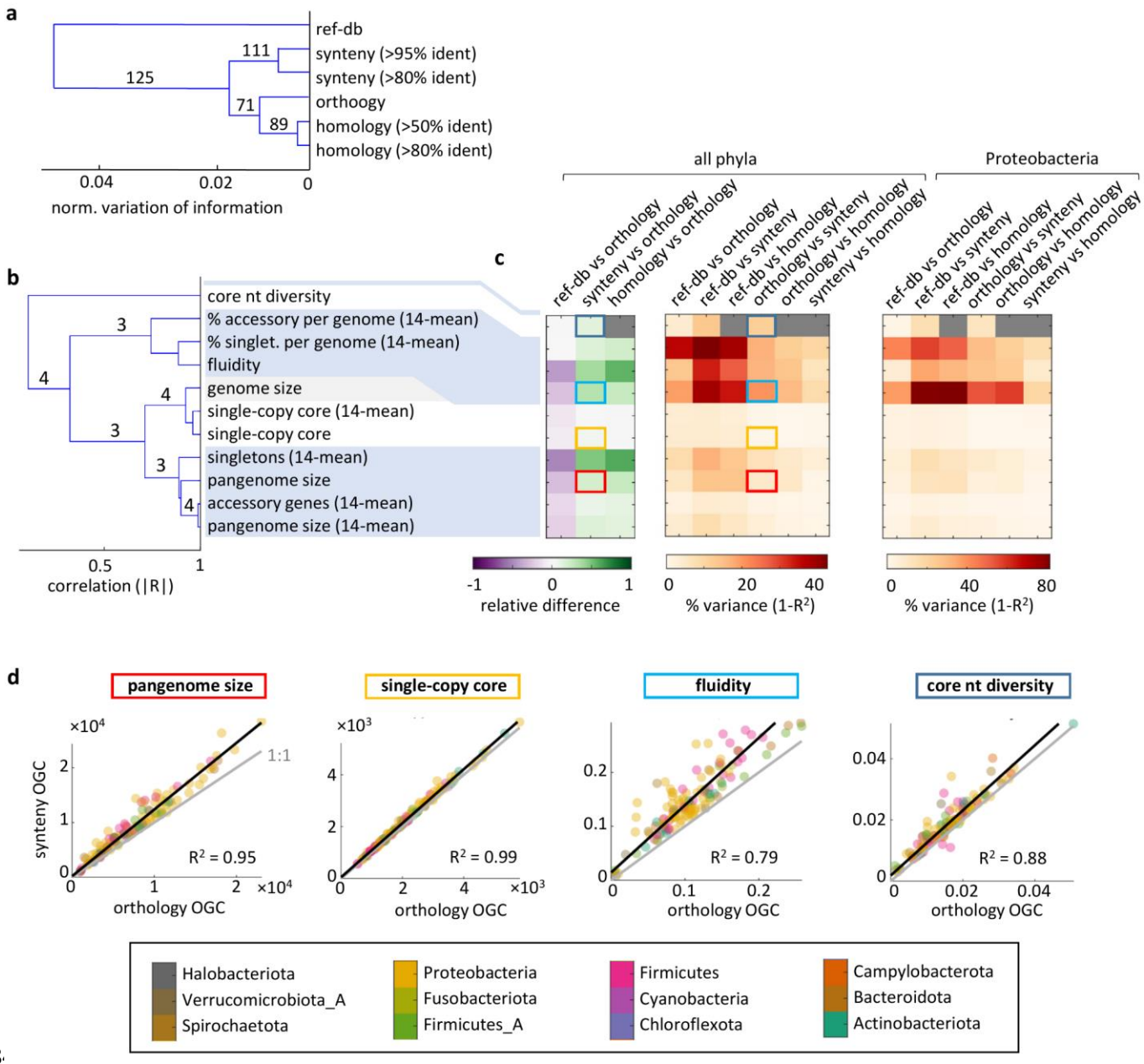
843 **Figure 1**



844

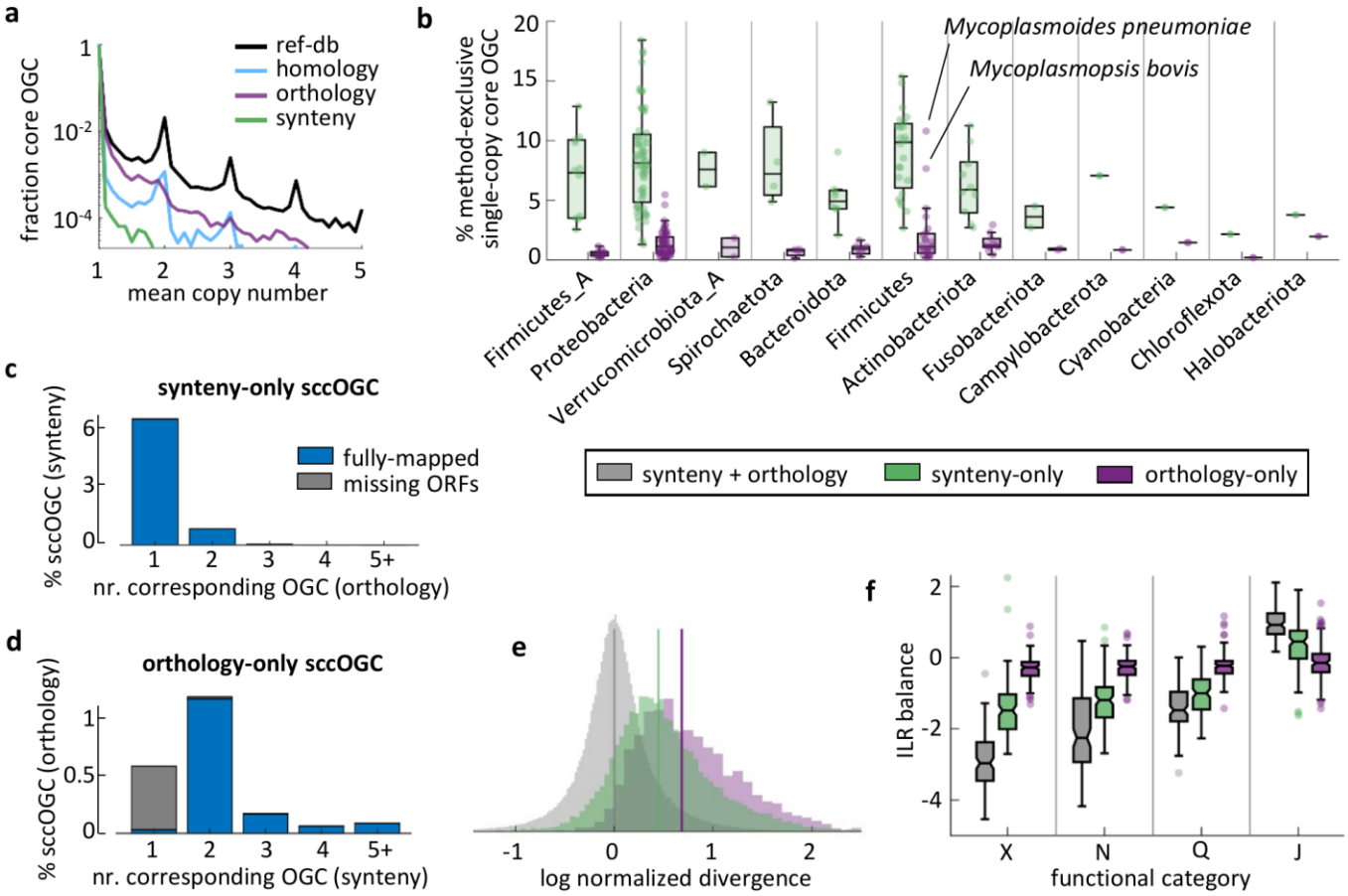
845 **Figure 2**

846

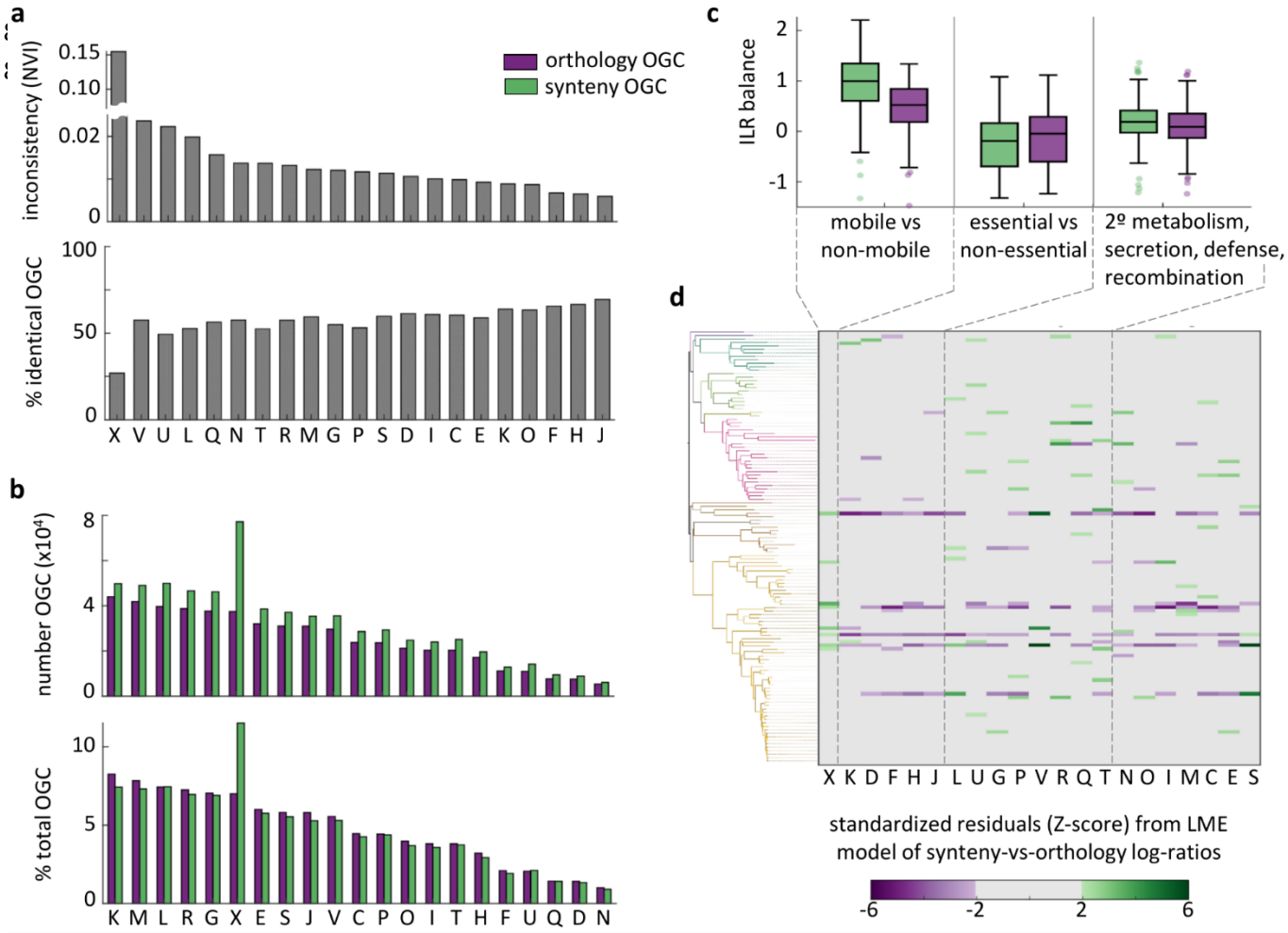


8
848

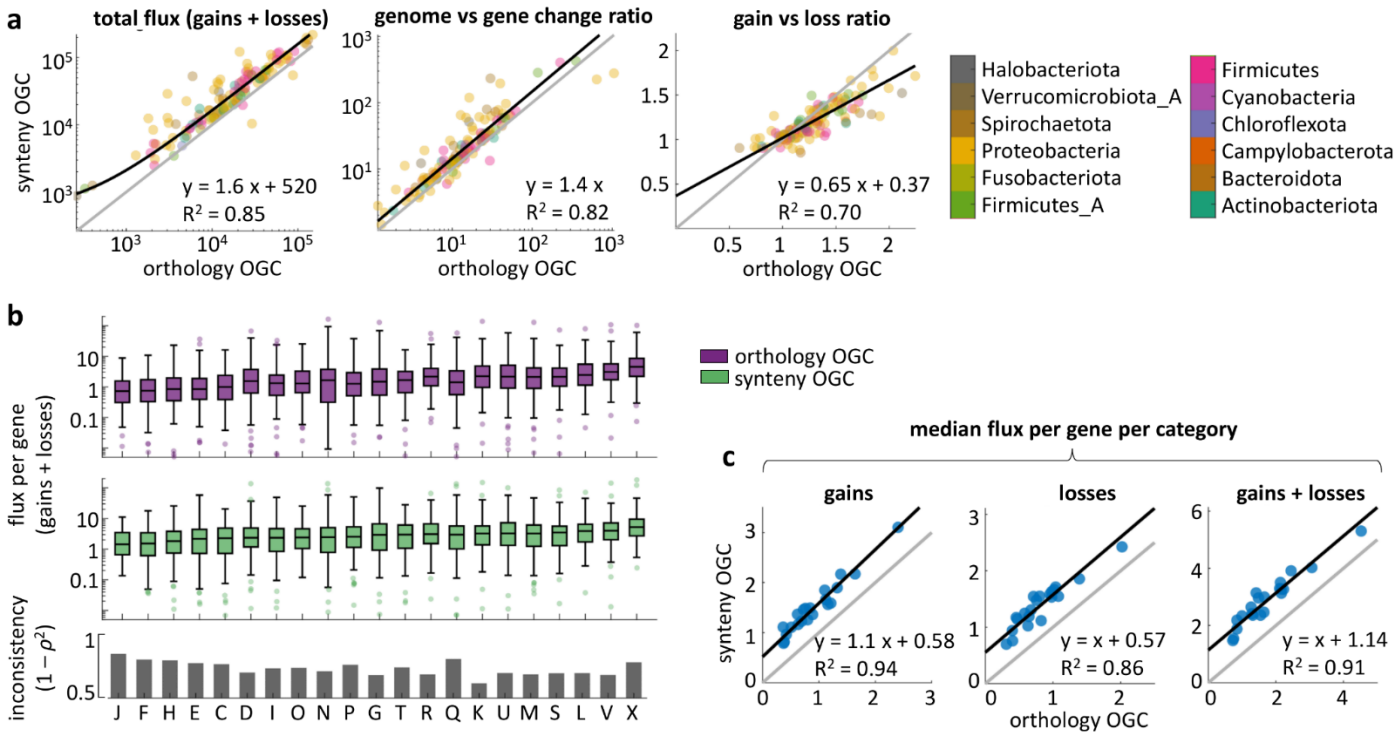
849 **Figure 3**


a
fraction core OGC
mean copy number
ref-db
homology
orthology
synteny
b
% method-exclusive single-copy core OGC
Mycoplasmoides pneumoniae
Mycoplasmopsis bovis
Firmicutes_A
Proteobacteria
Verrucomicrobiota_A
Spirochaetota
Bacteroidota
Firmicutes
Actinobacteriota
Fusobacteriota
Campylobacterota
Cyanobacteria
Chloroflexota
Halobacteriota
synteny + orthology
synteny-only
orthology-only
c
synteny-only sccOGC
% sccOGC (synteny)
fully-mapped
missing ORFs
nr. corresponding OGC (orthology)
d
orthology-only sccOGC
% sccOGC (orthology)
nr. corresponding OGC (synteny)
e
log normalized divergence
f
ILR balance
functional category

851 **Figure 4**


a
inconsistency (NVI)
0.15
0.10
0.02
0
orthology OGC
synteny OGC
% identical OGC
100
50
0
X V U L Q N T R M G P S D I C E K O F H J
b
number OGC (x10⁴)
8
4
0
% total OGC
10
5
0
K M L R G X E S J V C P O I T H F U Q D N
c
ILR balance
2
1
0
-1
mobile vs non-mobile
essential vs non-essential
2º metabolism, secretion, defense, recombination
d
X K D F H J L U G P V R Q T N O I M C E S
standardized residuals (Z-score) from LME model of synteny-vs-orthology log-ratios
-6
-2
2
6

854 **Figure 5**


a
total flux (gains + losses)
genome vs gene change ratio
gain vs loss ratio
synteny OGC
orthology OGC
y = 1.6 x + 520
R² = 0.85
y = 1.4 x
R² = 0.82
y = 0.65 x + 0.37
R² = 0.70
Halobacteriota
Verrucomicrobiota_A
Spirochaetota
Proteobacteria
Fusobacteriota
Firmicutes_A
Firmicutes
Cyanobacteria
Chloroflexota
Campylobacterota
Bacteroidota
Actinobacteriota
b
flux per gene (gains + losses)
inconsistency (1 − ρ²)
J F H E C D I O N P G T R Q K U M S L V X
orthology OGC
synteny OGC
c
median flux per gene per category
gains
losses
gains + losses
synteny OGC
orthology OGC
y = 1.1 x + 0.58
R² = 0.94
y = x + 0.57
R² = 0.86
y = x + 1.14
R² = 0.91