Comparison of gene clustering criteria reveals intrinsic uncertainty in pangenome analyses

Saioa Manzano-Morales[1,4,*], Yang Liu[1,2,*], Sara González-Bodí[1], 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

Background

A key step for comparative genomics is to group open reading frames into functionally and evolutionarily meaningful gene clusters. Gene clustering is complicated by intraspecific 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 implications of adopting homology, orthology, or synteny conservation as formal criteria for gene clustering by performing comparative analyses of 125 prokaryotic pangenomes.

Results

Clustering criteria affect pangenome functional characterization, core genome inference, and reconstruction of ancestral gene content to different extents. Species-wise estimates of pangenome and core genome sizes change by the same factor when using different clustering criteria, which allows for robust cross-species comparisons regardless of the clustering criterion. However, cross-species comparisons of genome plasticity 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 accessory functions. In some pangenome features, the variability attributed to methodological inconsistencies can even exceed the effect sizes of ecological and phylogenetic variables.

Conclusions

Choosing an appropriate criterion for gene clustering is critical to conduct unbiased pangenome analyses. We provide practical guidelines to choose the right method depending on the research goals and the quality of genome assemblies, and a benchmarking dataset to assess the robustness and reproducibility of future comparative studies.

Keywords

pangenome, orthology, comparative genomics, homology, core gene, accessory genome, genome plasticity, MAG

Background

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, that is, the set of all genes found in the genomes of a taxonomic group [1,2]. Pangenomes have quickly gained importance in eco-evolutionary research because they provide valuable information about the functional capabilities accessible to the strains of the same species and their propensity to gain and lose genes [3] Therefore, the study of pangenomes can shed light on the interactions among genome plasticity, niche diversity, and adaptability to environmental changes [4–7]. Moreover, identifying the set of genes present in all the genomes of a species (the core genes) has proved helpful to define core functions and build high-resolution species trees [8–10].

To study the pangenome of a species, researchers first collect all available high-quality genomes and run gene prediction tools to identify open reading frames (ORF). The ORF are then clustered into

61 groups of “equivalent” genes, that serve as the basis for subsequent cross-genome comparisons. To
62 obtain meaningful results, it is critical that gene clusters represent coherent units, both from
63 functional and evolutionary perspectives. The simplest conceptual approach to gene clustering is
64 based on homology. Two sequences are homologous if they derive from a common ancestral
65 sequence. Beyond this qualitative definition, different types of homology can be established based
66 on the evolutionary history of each gene. Because most gene families have experienced duplications
67 throughout evolution, it is usual to find homologous genes with multiple representatives per
68 genome, termed paralogs. Paralogs often display functional divergence and accelerated
69 evolutionary rates [11,12]. As a result, homology alone does not guarantee functional and
70 evolutionary homogeneity of gene clusters. Instead, for most purposes, it is desirable to subdivide
71 homologs into higher-resolution and more homogeneous clusters encompassing orthologous
72 sequences [13,14]. Orthologs are genes that share a single common ancestor at the time of
73 speciation [13,15]. Accordingly, the orthology criterion discriminates paralogs that duplicated
74 before the last speciation event (Fig. 1). Because of the connection between orthology and
75 speciation, orthology is the most natural grouping criterion for cross-strain comparative genomics
76 and phylogenomic analyses. In practice, it is challenging to apply a strict orthology criterion to
77 pangenome studies due to the lack of accurate reference species trees and the high computational
78 burden of building thousands of single-gene trees, one for each group of homologs. To circumvent
79 these difficulties, heuristic algorithms and reference databases of orthologous genes have been
80 developed over the last decade [16–18]. Together with homology and orthology, a third criterion to
81 further refine equivalence among genes is synteny conservation. According to the synteny criterion,
82 two genes (typically orthologs) are grouped together if they share the same gene neighborhood in
83 different genomes. Notably, synteny conservation can help discriminate between vertically and

84 horizontally transmitted copies of a gene, or between multiple orthologs resulting from within-
85 species gene expansions (Fig. 1).

86 It is widely accepted that the quantitative results of single-species pangenome analyses depend on
87 the method (and parameters) used to build the pangenome. Such dependency can be traced back
88 to the strategy adopted to identify gene clusters. As intuitively expected, the size of the pangenome
89 (that is, the number of clusters) increases and the size of the core genome decreases as the
90 sequence similarity threshold used to define homology becomes more and more stringent [19–21].
91 However, comparative studies implicitly assume that qualitative trends and order relationships are
92 robust with respect to the choice of a particular criterion. The existence of such qualitative
93 robustness merits a critical evaluation, given the deep conceptual differences among homology,
94 orthology, and synteny conservation, the heterogeneity of evolutionary rates across genes and
95 species, and the high rates of horizontal gene transfer (HGT) observed in some genomes [22–24].

96 The possibility that homology, orthology, and synteny-based gene clustering produce incongruent
97 results in comparative pangenome studies is not just a technical caveat. Instead, intraspecific HGT
98 and gene duplications raise the fundamental question of what equivalence class (homology,
99 orthology, synteny, or something else) best captures the essentially dynamic nature of pangenome
100 evolution [25]. This conceptual conundrum is especially clear in what concerns the splitting of
101 multicopy orthologs based on their gene neighborhoods. Let us consider, for example, a gene that
102 has experienced recent duplications or intraspecific HGT. For what purposes should all copies be
103 included in the same group, following the standard orthology criterion? For what other purposes
104 should the duplicated or horizontally transferred copy be assigned to a new cluster, based on
105 synteny considerations? The orthology criterion appears better to assess pangenome diversity, for
106 which it is not desirable to count a recently duplicated gene twice [26], or to study gene duplication
107 and intraspecific HGT, which first requires classifying multicopy genes as members of the same

108 group [27]. In contrast, the synteny criterion would be more appropriate to select marker genes for
109 phylogenomic analyses or identify vertically transmitted members of mobile gene families.

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

117 State-of-the-art methods for OGC construction implement different strategies to deal with
118 sensitivity-vs-specificity trade-offs at manageable computational cost. The simplest approaches
119 (e.g., CD-HIT [28] and MMseqs2 [29]), build groups of homologous genes by applying a pre-defined
120 similarity threshold to the amino acid sequences encoded by a set of ORF. More sophisticated tools
121 include additional steps to discriminate true orthology from other ways of homology. Such
122 discrimination can be attained through two major strategies: (a) by building gene-level phylogenetic
123 trees or sequence similarity networks, on which some heuristic rules are applied to resolve subsets
124 of genes with shared ancestry [30]; and (b) by subclustering homologous groups based on their gene
125 neighborhoods, under the assumption that synteny is locally conserved at the evolutionary
126 timescales that are relevant to within-species diversification. By design, the former approach, which
127 is adopted by reference databases like COG [16] and eggNOG [18], the orthology detection tool
128 OrthoFinder [31], and the pangenome analysis suite panX [32], is better suited to assess true
129 orthology in gene families affected by intraspecific duplications and HGT. However, its high
130 computational burden makes it inefficient to deal with large genomic datasets. On the other side,
131 synteny-based approaches (such as the one implemented by the popular pangenome analysis tools

132 Roary [33] and PanOCT [34]) are faster and less resource intensive, although their results may
133 deviate from the classical concept of orthology (by missing true orthologs) when applied to highly
134 dynamical genomes or regions of the genome with poor synteny conservation. Phylogeny-aware
135 and synteny-based methods are not mutually exclusive [35], although they are rarely performed
136 together due to computational constraints. A summary of these and other popular tools for OGC
137 construction and their application to pangenome analysis is provided in Suppl. Table S1.

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

146

147 **Results**

148 <u>Method-dependent variation and intrinsic uncertainty in pangenome size and diversity</u>

149 We investigated 5 major methods for *de novo* species-wise OGC construction that qualitatively differ
150 in the strategies for gene clustering and paralog discrimination (Table 1). All *de novo* methods start
151 from a collection of open reading frames (ORF) and cluster them according to a predefined identity
152 threshold that can be set to include more or less distant homologs. The resulting clusters are then
153 processed to split paralogs into separate OGC. The methods included in this study represent three
154 alternative approaches implemented by some of the most popular tools for pangenome analysis:

155 orthology-based clustering (implemented by panX and OrthoFinder), synteny-based clustering
156 (implemented by Roary), and homology-based clustering (implemented by CD-HIT and MMseqs2,
157 which is the OGC construction module used by PanACoTA [36] and PPanGGOLiN [37]. In the case of
158 Roary, CD-HIT, and MMseqs2, we also explored the effect of introducing two different identity
159 thresholds for the initial clustering step (by design, panX and OrthoFinder do not filter gene clusters
160 based on sequence identity but on their e-value). The optional settings used with each tool are
161 detailed in the methods section and summarized in Suppl. Table S2.

162 **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	(33)
de novo	orthology-based	none (e-val < 0.001)	OrthoFinder	(32)
de novo	synteny-based	80%, 95%	roary	(34)
de novo	none	50%, 80%	MMseqs2, PanACoTA	(29,30)
de novo	none	50%, 80%	CD-HIT	(28)
reference db	n.a.	none (e-val < 0.001)	eggNOG-mapper	(59)

163

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

174 By applying these methods, we obtained 9 alternative sets of species-wise OGC for 124 bacterial
175 and 1 archaeal species (DOI:10.5281/zenodo.7387758). These species, defined according to the
176 phylogenetically consistent classification scheme established by the Genome Taxonomy Database
177 (GTDB) [38], were selected to cover every genus in the GTDB, with the condition that there were at
178 least 15 high-quality genomes available per species. We quantified the discordance between the
179 OGC produced by each pair of methods by means of the normalized variation of information (NVI),
180 a measure that takes values between 0, if the two sets of OGC display a one-to-one correspondence,
181 and 1, if they are completely independent. The hierarchical clustering of methods based on this
182 measure reveals that the discordances, although small, are reproducible across species (Fig. 2a). The
183 most notable differences arise between reference-based and *de novo* OGC. Among the latter, the
184 strategy for paralog discrimination determines the resulting OGC to a greater extent than the
185 particular tools and identity thresholds, at least for the relatively permissive thresholds
186 implemented in the study.

187 Based on the previous finding, we restricted our analysis to a single identity threshold (80%) and
188 one tool of each class. We chose MMseqs2, panX, and Roary because they were primarily designed
189 to study pangenomes or are at the core of popular workflows for pangenome analysis. Then, we
190 investigated how paralog discrimination affects the properties of the inferred prokaryotic
191 pangenomes. To that end, we selected 10 non-redundant quantitative features that represent
192 different aspects of pangenome size and diversity. All *de novo* methods supported clustering these
193 features in 4 groups (Fig. 2b) that can be interpreted in terms of (i) pangenome size, (ii) genome
194 (and core genome) size, (iii) gene content diversity, and (iv) nucleotide diversity (evaluated in the
195 core genes). As expected, reference-based OGC systematically produced the smallest estimates of
196 pangenome size and gene content diversity (around 30% lower than orthology-based OGC),

197 whereas *de novo* synteny- and homology-based OGC produced the largest estimates for those traits
198 (20-35% higher than orthology-based OGC, see Fig. 2c).

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

219 Despite the known fact that sample size affects the estimation of pangenome properties, two lines
220 of evidence prove that cross-method inconsistencies are independent of the number of genomes

221 used to build the pangenome, at least within the range of 15-100 genomes explored in this study.
222 First, the same trends are detected in gross and size-corrected features (for example, the total
223 pangenome size and its average over subsets of 14 genomes). Second, cross-method scatter plots,
224 like those of Fig. 2d, show no correlation between the number of genomes per species and the
225 relative deviation of each species with respect to the regression line (Pearson's correlation, $|R| \leq$
226 0.16 and $q > 0.05$ for all features and OGC criteria included in Fig. 2c). Such lack of correlation
227 indicates that method-dependent variation is equally likely to affect species with many or just a few
228 available genomes.

229 A major limitation of reference-based approaches to OGC construction is that they are constrained
230 by the limited diversity of the reference database. Indeed, in most species, 5-10% of the ORF could
231 not be mapped to the eggNOG database and therefore were not assigned to any reference-based
232 OGC (Suppl Fig. S3). The fraction of missing ORF is larger in some taxa that are underrepresented or
233 absent from the reference database. The most extreme case, with >30% unmapped genes,
234 corresponds to *B. burgdorferi*, the causal agent of Lyme's disease. The poor performance of
235 reference database mapping in *B. burgdorferi* is explained by the unique structure of its genome,
236 which consists of a linear chromosome and >20 linear and circular plasmids without homologs in
237 other species [39,40]. Despite these limitations, reference-based OGC provide reasonably good
238 estimates for the number of core genes per genome and allow retrieving 85-90% of the single-copy
239 core gene families identified by *de novo* approaches.

240 Within-species paralogy and inference of core genomes

241 Among the pangenome features considered in this work, the size of the core genome is the least
242 affected by the gene clustering criterion. However, a closer inspection of core gene families (defined
243 in a strict sense, given the high completeness of the genomes included in the study) reveals

244 differences in the mean copy number across methods (Fig. 3a). The fraction of core gene families
245 that appear as single-copy ranges from 80% in reference-based to >99% in synteny-based OGC. In
246 the case of reference-based and *de novo* homology-based OGC, the distribution of the mean copy
247 number per core gene family per genome displays clear peaks at integer values, indicating the
248 existence of complete sets of duplicated paralogs that were not resolved by these methods.

249 We next focused on single-copy core (s.c.c.) genes, which are of greater practical interest as they
250 are often used to infer high-resolution phylogenies. Synteny-based methods systematically produce
251 5-10% more s.c.c. OGC than orthology-based methods (Fig. 3b). Typically, synteny-supported s.c.c.
252 genes belong to orthologous OGC that are core but not necessarily single-copy (Fig 3c). Therefore,
253 it appears that synteny criteria are effective in resolving single-copy representatives of core gene
254 families affected by within-species duplications or HGT. On the other side, s.c.c. OGC that are
255 orthologous but not supported by synteny tend to either contain incomplete ORFs (that could not
256 be processed by the synteny-based workflow) or split among 2 or more synteny-supported
257 accessory OGC (Fig. 3d). Such lack of synteny conservation could be a sign of non-vertical
258 transmission in s.c.c. genes exclusively detected through orthology criteria. However, given the
259 small fraction of these OGC in most species (<1% of all orthology-based s.c.c. OGC), the overall effect
260 of such potential non-vertical contamination in downstream analyses is possibly modest. The only
261 exception occurs in Mycoplasmatales, in which the relative contribution of non-synteny-supported
262 OGC is amplified by their small genome sizes.

263 Single-copy core genes supported by a single criterion (either synteny or orthology) are significantly
264 less conserved in terms of their average identity than those detected by both criteria (Fig. 3e);
265 Kruskal-Wallis omnibus test $p < 10^{-20}$). Among the s.c.c. genes that are method-exclusive, those
266 based on orthology are significantly less conserved than those based on synteny ($p < 10^{-8}$ for all
267 Mann-Whitney post-hoc tests with Tuckey's HSD correction). There are also significant differences

268 in the functional profiles of s.c.c. genes that are supported by only one or both criteria (Fig. 3f), with
269 overrepresentation of genes associated with mobile genetic elements, cell motility, and secondary
270 metabolism, and underrepresentation of genes involved in translation among method-exclusive
271 single-copy core genes (linear mixed effects model for isometric log-ratios; $F(2,248) > 190$, $q < 10^{-20}$
272 in all cases).

273 Systematic and species-specific biases in functional profiles

274 Estimates of genome content diversity are strongly affected by the choice of orthology- or synteny-
275 based strategies to discriminate paralogs when building *de novo* OGC. To better understand the
276 causes and implications of these differences, we classified ORF and OGC into 21 coarse-grained
277 functional categories that are representative of the main molecular and cellular processes that take
278 place in prokaryotic cells. For each functional category, we studied the agreement between
279 orthology- and synteny-based OGC by calculating the NVI, the fraction of fully equivalent OGC (that
280 is, OGC that contain exactly the same ORF regardless of the strategy for paralog discrimination), and
281 the fraction of ORF assigned to fully equivalent OGC (Fig. 4a and Suppl. Fig. S4). By far, the highest
282 inconsistency occurs for mobile genetic elements, for which only 25% of the OGC (encompassing
283 25% of the ORF) are equivalent. Besides mobile genetic elements, moderate degrees of
284 inconsistency are observed for defense systems, intracellular trafficking/secretion, and
285 replication/recombination/repair. On the other end, central cellular functions, such as translation,
286 transcription, nucleotide metabolism, and coenzyme metabolism show the highest agreement, with
287 64-70% of the OGC (encompassing 77-84% of the ORF) being fully equivalent. These trends are also
288 manifested when looking at the absolute and relative numbers of OGC per category (Fig. 4b), with
289 synteny-based paralog discrimination producing a disproportional excess of OGC associated with
290 the mobilome. The fraction of OGC that contain ORF from more than one functional category is
291 consistently larger in the case of orthology-based OGC, although the absolute differences are

292 modest (around 0.5-1% in most categories; Suppl. Fig S4). Functionally heterogeneous OGC are most
293 often associated with signal transduction, cell cycle control/cell division, mobile genetic elements,
294 and unknown or poorly characterized functions.

295 Compositional analysis of functional profiles controlling for between-species variability confirms
296 that synteny-based paralog discrimination leads to a significant increase in the fraction of OGC
297 associated with mobile genetic elements (Fig. 4c; linear mixed effects model for isometric log-ratios;
298 ILR-balance difference = 0.49, $F(2,124) = 189$, $q < 10^{-20}$). Functional profiles derived from synteny
299 and orthology-based OGC also differ in the balance between central cellular functions (transcription,
300 translation, cell cycle, nucleotide and coenzyme metabolism) and other functional categories (ILR-
301 balance difference = -0.11, $F(2,124) = 132$, $q < 10^{-20}$); and between a set of functions including
302 secondary metabolism, carbohydrate metabolism, secretion, defense and recombination, and the
303 remaining functional categories (ILR-balance difference = 0.08, $F(2,124) = 62$, $q < 10^{-10}$). Apart from
304 these general trends, other significant differences between synteny- and orthology-based functional
305 profiles are restricted to specific categories in one or a few particular species (Fig. 4d), such as
306 defense in *Legionella pneumophila* (Z = 5.9, $q < 10^{-5}$), *Borreliella burgdorferi* (Z = 5.4, $q = 5\times10^{-5}$) and
307 *Bordetella pertussis* (Z = 4.4, $q = 0.002$), secondary metabolism in *Bacillus anthracis* (Z = 4.0, $q =$
308 0.008), and signal transduction in *Brachyspira hyodysenteriae* (Z = 3.8, $q = 0.019$).

309 <u>Variability of gene flux estimates</u>

310 To assess whether method-dependent variation in pangenome composition propagates to
311 downstream analyses, we investigated the effect of synteny- and orthology-based paralog
312 discrimination on a quantitative study of genome dynamics. To that purpose, we used the software
313 Gloome, that infers events of gene gain and loss along a lineage taking as inputs the strain-level
314 phylogenetic tree and a binary matrix with the presence and absence profiles of each OGC. As shown

315 in Fig. 5a, using synteny-based instead of orthology-based OGC leads to a 60% increase in the
316 estimated number of gene gains and losses per lineage. Genome vs gene change ratios, measured
317 as the number of expected gains and losses per gene per core nucleotide substitution are also higher
318 (40% increase) when using synteny-based OGC. In contrast, the ratio between gene gains and losses,
319 that determines the short-term dynamics of genome size [41], displays a more complex response,
320 with synteny-based OGC producing lower or higher estimates than those obtained with orthology-
321 based OGC depending on whether a species is dominated by gains or losses. Method-dependent
322 uncertainties account for 15%, 18%, and 30% of the between-species variability in the total flux, the
323 genome vs gene change ratio, and the gain vs loss ratio, respectively. These results indicate that
324 comparative analyses of short-term genome dynamics are highly sensitive to methodological
325 choices for paralog discrimination, especially when it comes to evaluate the balance between gene
326 gain and loss.

327 A deeper analysis of gene flux by functional categories reveals that inconsistencies between gene
328 clustering criteria (quantified as one minus the squared rank correlation of species-wise estimates
329 to account for the numerous outliers) are stronger among mobile genetic elements and genes
330 involved in secondary metabolism and inorganic ion transport (Fig. 4b). Although high
331 inconsistencies are also observed in central functional categories, such as translation, the practical
332 relevance of those is lesser due to the relatively low fluxes associated with those categories. If all
333 species are jointly considered by calculating their median, flux estimates obtained from synteny
334 OGC display a systematic deviation of >1 additional event per gene in all functional categories, which
335 is evenly distributed between gains and losses (Fig. 4c).

336 <u>Estimation of pangenome properties from medium- and low-quality genome assemblies</u>

337 Pangenomes built from high-quality, nearly complete genomes represent an appropriate
338 benchmark to compare gene clustering criteria without the confounding effects of genome
339 incompleteness and contamination. However, complete genomes are relatively rare and unevenly
340 distributed among taxonomic groups. Given those limitations, incomplete genomes, often
341 assembled from metagenomes, constitute the only available choice to study the pangenomes of
342 many bacteria and archaea. To evaluate the performance of different gene clustering criteria in such
343 suboptimal scenarios, we built 96 alternative pangenomes by replacing some of the complete
344 genomes by an increasing proportion of medium- and low-quality genomes assembled from
345 metagenomes (MAG). To keep the computational cost manageable, we focused on 4 species
346 belonging to the phyla Proteobacteria, Actinobacteria, Firmicutes, and Bacteroidota, that cover a
347 broad range of pangenome sizes and gene content diversity. For each pangenome feature, we
348 quantified the uncertainty introduced by MAG as the relative amount of variation in the alternative
349 pangenomes that could not be predicted from the original high-quality pangenomes. Values greater
350 than 1 (that correspond to $R^2 < 0$) imply that pangenomes that include MAG follow qualitatively
351 different trends than those built from complete genomes. In those cases, the inclusion of MAG leads
352 not only to loss of precision, but to systematic biases.

353 Estimates of pangenome size are relatively unaffected by the inclusion of medium- and low-quality
354 MAG, especially if pangenome size is calculated by averaging over multiple subsamples (Fig. 6a, top;
355 Suppl. Fig. S5a). By using reference, homology, or synteny OGC, the unexplained variance in
356 pangenome size can be kept below 10% even if half of the genomes are incomplete. In contrast,
357 estimates of gene content diversity are much more sensitive to genome incompleteness (Fig. 6a,
358 middle; Suppl. Fig. S5a and S6), with systematic deviations in fluidity appearing with only a 5% of
359 MAG. As expected, incomplete genomes severely impair the direct detection of core genes (Suppl.
360 Fig. S7), although consistent estimates of the core genome size (with up to 50% medium- and low-

361 quality MAG) can be obtained with mOTUpan [42], a specialized tool that uses Bayesian inference
362 techniques to predict core genes while accounting for genome incompleteness (Fig 6a, bottom;
363 Suppl. Fig. S5a).

364 Inspired by the good performance of mOTUpan, we explored the potential of Bayesian approaches
365 to produce more robust measures of gene content diversity. Thus, we recalculated the fraction of
366 accessory genes per genome by combining mOTUpan's predictions for the number of core genes
367 and the mean genome size (see Methods). The new metrics is more consistent than fluidity,
368 although still far from reaching the high robustness of pangenome size estimates (Suppl. Fig. S6).
369 Taken together, the higher performance of mOTUpan for the estimation of core genes and this
370 "mOTUpan-based" metrics for gene content diversity illustrate how a probabilistic framework can
371 help overcoming some of the limitations imposed by incomplete and poorly assembled genomes.

372 To better understand variations in performance across methods, we looked at genes from complete
373 genomes and studied whether the inclusion of MAG leads to their being reassigned to different
374 clusters (Fig. 6b; Suppl. Fig. S5b). Because reference-based methods operate gene by gene, the
375 addition of new genomes does not have any effect on how a given gene is clustered. Such
376 "structural" robustness explains why reference-based OGC provide the most consistent estimates
377 in the presence of MAG. In stark contrast, clustering based on synteny and orthology is substantially
378 affected by the addition of incomplete or poorly assembled genomes (the within-method NVI after
379 adding 5% of MAG becomes as large as the cross-method NVI in high-quality pangenomes, shown
380 in Fig. 4a). The fact that this phenomenon is not observed for homology-based OGC indicates that
381 the paralog discrimination algorithms used by synteny- and orthology-based methods are prone to
382 misclassify genes from high-quality genomes if the dataset also includes lower-quality genomes.

383

384 **Discussion**

385 Orthology is generally considered the optimal criterion for clustering gene sequences for
386 comparative genomics [14]. Because orthologs tend to conserve their function and evolve
387 consistently with speciation patterns, they are, at least in theory, the most fitting choice for
388 functional and phylogenomic studies. In practice, however, the use of orthology as a gold standard
389 for pangenome analysis faces technical and conceptual challenges. On the technical side,
390 distinguishing orthologs from paralogs is computationally costly. Therefore, to deal with large
391 (meta)genomic datasets, orthology prediction tools often resort to heuristic algorithms that risk
392 missing true orthologs or including out-paralogs (paralogs that duplicated before the last common
393 ancestor of the clade of interest). On the conceptual side, intra-species gene duplication and
394 horizontal gene transfer generate multi-copy gene lineages that, despite not violating the orthology
395 criterion at the species level, may confound downstream analyses that assume vertical transmission.
396 In those cases, a stricter criterion that only clusters together vertically transmitted members of a
397 gene family may be more desirable.

398 Motivated by these caveats, we assessed how three major clustering criteria (homology, orthology,
399 and synteny) implemented by five popular pangenome analysis tools affect pangenome
400 reconstructions and downstream phylogenomic analyses. Although we only tested a limited number
401 of tools and parameter settings, our results suggest that the underlying formal criterion for paralog
402 discrimination (shared ancestry at speciation for orthology, conserved gene neighborhood for
403 synteny), rather than the actual implementation, is what drives qualitative differences across
404 methods.

405 Previous works had pointed out that single-species estimates of pangenome size and diversity
406 strongly depend on the method used to cluster genes [19–21]. We confirmed such observations and

407 expanded on the implications for comparative pangenome analyses. When conducting comparative
408 studies, inconsistencies in relative differences and cross-species trends are of much greater concern
409 than absolute differences in single-species estimates. In that regard, trends involving pangenome
410 and core genome sizes are generally robust across methods. More precisely, choosing one or
411 another method affects all species-wise size estimates by the same constant multiplicative factor.
412 In contrast, cross-species comparisons of genome plasticity and pangenome diversity are highly
413 sensitive with respect to the clustering criterion, in a way that cannot be explained by any linear or
414 nonlinear data transformation. Such inconsistencies in pangenome diversity, that reach up to 50%
415 of the total between-species variance in Proteobacteria, should be a major source of concern for
416 studies aimed at understanding the forces that shape microbial pangenomes. For comparison, it has
417 been estimated that habitat and phylogeny contribute to explain approximately 20% of the
418 between-species variance in pangenome diversity [6]. That said, regardless of the method used to
419 discriminate paralogs, the lowest rates of gain and loss are always observed in genes associated with
420 central cell functions, whereas the highest rates correspond to mobile genetic elements and defense
421 systems. Therefore, the inverse association between gene flux and essentiality described by
422 previous studies [43–46] appears as a robust feature of genome plasticity.

423 The contribution of a particular gene clustering method to pangenome variability can only be
424 assessed by comparison with other methods, which is often unfeasible in large datasets. As a result,
425 methodological "noise" remains unnoticed, potentially contributing a relevant fraction of the
426 unexplained variance or, in the worst-case scenario, acting as a confounding factor if methodological
427 biases correlate with the biological variables of interest.

428 Technical choices during genome assembly and ORF prediction can also cause biases that propagate
429 downstream affecting pangenome analyses [35]. For example, variations in assembly completeness

430 and contamination will have a direct repercussion on the fraction of core genes and genes with no
431 homologs in the pangenome. By studying the interplay between assembly quality and gene
432 clustering criteria, we found that reference-based clustering produces the most robust estimates of
433 pangenome properties in pangenomes that include incomplete and poorly assembled genomes. On
434 the other side, orthology-based methods often lead to inconsistent results even with relatively low
435 proportions of medium-quality assemblies. The most likely cause for such poor performance is that
436 incomplete genomes interfere with the construction of the strain-level reference trees that serve as
437 guides for the identification of orthologs. For example, the default for panX is to build guide trees
438 from strict core genes, so that the method fails if no core genes are found (which typically occurs if
439 >5% of all genomes are incomplete). More permissive methods that take into account the "soft"
440 core, such as OrthoFinder or panX with customized core thresholds, can handle higher proportions
441 of incomplete genomes. However, their accuracy will still be compromised if the guide tree is
442 incorrect.

443 Practical guidelines

444 Our findings stress that, as long as choosing an operational definition of gene cluster remains
445 arbitrary, pangenome properties affected by method-dependent variability will be subject to
446 intrinsic uncertainty. To minimize such uncertainty, the selection of tools for gene clustering should
447 be primarily guided by the nature of the research goals, rather than by computational
448 considerations such as runtime and memory usage. In the next paragraphs, we provide some
449 practical guidelines to help selecting optimal clustering criteria for different purposes. A summary
450 of those guidelines can be found in Table 2.

451

452 **Table 2:** Practical guidelines to select optimal gene clustering criteria for pangenome analysis.

OGC	Strengths	Limitations	Recommended use
Ref-db	- Annotation and classification at once. - The most robust method in pangenomes with incomplete genomes.	- Loss of all gene families that are not represented in the ref-db. - Discrimination of paralogs is limited by the taxonomic resolution of the ref-db.	- Functional profiling and detection of single-copy core genes (if missing taxon-specific genes is not a problem). - Analysis of pangenomes with a large fraction of incomplete genomes. - Cross-species tracking of pangenome contents.
Homology	- Faster than any other method.	- Setting a hard similarity threshold can lead to merging paralogs and/or missing distant orthologs.	- Quick detection of single-copy core genes. - Fast de novo clustering of sequences without homologs in ref-db. - Analysis of very large pangenomes.
Orthology	- Evolutionary consistent clustering of mobile gene families. - Discrimination of in- and out-paralogs.	- Computationally costly (panX becomes prohibitive with >200 genomes). - Poor performance if the pangenome contains incomplete genomes.	- Unbiased functional profiling. - Study of genome plasticity and gene flux in high-quality pangenomes. - Assessment of within-species gene expansions.
Synteny	- Accurate discrimination of vertically transmitted gene families. - The most sensitive method for single-copy core genes.	- The fragmentation of mobile gene families in multiple OGC can bias functional profiles and quantitative estimates of genome plasticity.	- Identification of vertically transmitted orthologs. - Generation of high-resolution phylogenetic trees. - Analysis of pangenomes with up to 20% of incomplete genomes.

453

454 The identification of single-copy core genes is one of the main outcomes of pangenome
455 reconstruction [47]. Single-copy core genes allow building high-resolution strain trees that can be
456 very robust to the effects of unbiased homologous recombination [48]. However, to obtain accurate
457 trees, it is fundamental that the OGC that correspond to single-copy core genes are not
458 contaminated by HGT [10]. As expected, synteny is the most sensitive criterion to discriminate
459 single-copy representatives among gene families affected by paralogy and intra-species
460 duplications. Nevertheless, *de novo* homology- and orthology-based methods retrieve most (90-

461 95%) synteny-supported single-copy core OGC with low (<1%) potential contamination, which
462 makes them good alternatives for many practical applications. Regardless of the criterion used to
463 obtain the OGC, if the pangenome includes incomplete genomes, inference of core genes can be
464 notably improved by applying probabilistic approaches, such as those implemented by mOTUpan
465 [42].

466 Synteny-based gene clusters display substantial fragmentation of non-core gene families, that
467 affects not only mobile genetic elements but also genes involved in defense, secondary metabolism,
468 signal transduction, and secretion, among others. Because such fragmentation is not homogeneous
469 across taxa and functional categories, synteny-based clustering can introduce biases in comparative
470 studies of pangenome function and dynamics. Functional profiles based on orthology (rather than
471 synteny) are better choices, as they are more robust with respect to the amplification of mobile
472 genetic elements and other accessory gene families that are characteristic of some species
473 [22,49,50]. Likewise, inference of gene fluxes within and across pangenomes with tools that account
474 for gene copy number (e.g., the phylogenomic software COUNT [27]) should be performed with
475 orthology-based gene clusters. In turn, analyses of genome plasticity based on binary (presence or
476 absence) phyletic profiles may be more accurate if using synteny-based gene clusters, because they
477 better capture the contribution of intra-species paralogs to the total gene flux. Future studies aimed
478 at exploring the role of adaptive and non-adaptive processes in pangenome evolution should
479 explicitly discuss how methodological choices affect the interpretation of their results. For example,
480 choosing synteny over classical orthology as the clustering criterion may inflate the proportion of
481 mobile genetic elements in the pangenome, altering the balance between selfish and potentially
482 beneficial accessory genes. Estimates of gene content diversity are extremely sensitive to the
483 presence of incomplete genomes, even if they only constitute a small fraction of the whole
484 pangenome. Therefore, to study pangenome plasticity, it is preferable to work with a reduced set

485 of high-quality genomes than with a larger set of lower-quality genomes. If incomplete genomes
486 cannot be avoided, we recommend using reference-based OGC and measuring gene content
487 diversity as the predicted (rather than observed) percentage of accessory genes per genome.

488 As a fast alternative to *de novo* clustering methods, we evaluated the performance of OGC built by
489 annotating gene sequences with the eggNOG reference database. Although reference-based OGC
490 deviate from *de novo* OGC in many aspects, they are a reasonable option to identify single-copy core
491 genes and generate functional profiles in well-studied genera, especially if paralogs and rare genes
492 can be disregarded. Reference-based OGC are also the most robust option to deal with pangenomes
493 that contain a large fraction of incomplete or poorly assembled genomes. On the other side,
494 eggNOG-based OGC can miss a significant portion of species-specific core genes in poorly sampled
495 taxa, do not discriminate paralogs below the genus level, and filter out gene families with narrow
496 taxonomic distribution. Accordingly, pangenome analyses using reference-based OGC should justify
497 their conclusions accounting for these limitations. Possibly, the most interesting application of
498 reference-based OGC is found in large-scale comparative studies that require tracking OGC across
499 species. That can be achieved thanks to the hierarchical structure of the eggNOG database, that
500 allows annotating orthologs at multiple taxonomic levels without additional computational cost.

501

502 **Conclusions**

503 There is no "one size fits all" gene clustering method for pangenome analyses. Such lack of a
504 universal method does not only reflect the practical limitations of current algorithms, but a deeper
505 conceptual trade-off. For applications that involve tracking vertically transmitted genes, the best
506 OGC are those that effectively discriminate among in-paralogs, which is generally achieved by

507 applying synteny criteria. However, those same OGC are not optimal to study within-species
508 expansions and contractions of accessory gene families. (For that purpose, OGC should keep in-
509 paralogs together, as dictated by classical orthology.) In practice, the best choice also depends on
510 the quality of the genomic assemblies, with homology and reference-based OGC producing the most
511 robust results in medium- and low-quality pangenomes.

512 Reusability and meta-analysis of pangenome datasets is currently hindered by the incompatibility
513 of OGC obtained by different methods. To address that limitation and foster future research, it is
514 critical to come to a consensus on a set of methods that cover the most relevant criteria for paralog
515 discrimination. To maximize consistency and reusability, an optimal solution should leverage the
516 hierarchical nature of gene clustering criteria, with synteny OGC nested within orthology OGC, and
517 the latter nested within homology OGC. Such multi-level scheme would facilitate selecting the best
518 set of OGC for each purpose and assessing the robustness of pangenome properties in the face of
519 methodological uncertainty. We expect that the multi-method dataset generated in this study
520 (publicly available in https://dx.doi.org/10.5281/zenodo.7387758), though not strictly nested, will
521 constitute a first step in that direction.

522

523 **Methods**

524 <u>High-quality genomic sequences</u>

525 We parsed the metadata files of the Genome Taxonomy Database (GTDB) release 95 [51] to identify
526 all the species (sensu GTDB) for which there were at least 15 high-quality available genomes. High-
527 quality genomes were defined according to the MIMAG criteria [52] plus the following more
528 stringent filters: completeness >99% and contamination <1%, both assessed through CheckM [53],

529 mean contig length >5kb, and contig count <500. Metagenome-assembled genomes (MAG) and
530 single-amplified genomes (SAG) were not included in the analysis. After applying these filters, we
531 recovered 321 bacterial and 1 archaeal species, belonging to 125 different genera (sensu GTDB). To
532 minimize possible taxonomical biases, only one representative species per genus was selected for
533 subsequent analyses. For those genera with >1 suitable species, we kept the species with the highest
534 number of high-quality genomes, as they presumably were the most informative for pangenome
535 studies.

536 To keep our analyses within a manageable computational cost, species with >100 high-quality
537 genomes were subsampled to keep at most 100 genomes per species. To that end, we separately
538 aligned the amino acid sequences of 120 nearly universal marker genes employed by the GTDB with
539 MAFFT [54], concatenated the alignments, and ran IQ-Tree [55] to obtain phylogenetic trees
540 including all the strains of the same species. Then, we applied a heuristic subsampling strategy that
541 maximized the diversity of the subsampled genomes whilst accurately reflecting the topology of the
542 strain trees.

543 The final dataset comprised 6,796 bacterial genomes belonging to 124 species and 55 archaeal
544 genomes belonging to 1 species. The genomes were downloaded from the NCBI FTP site. Open
545 reading frames (ORF) were predicted with Prodigal v2.6.3 [56], using the "single" mode that is
546 recommended for finished genomes and quality draft genomes. All runs used the bacterial, archaeal
547 and plant plastid code 11 (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi), which
548 maps UAA, UGA, and UAG to stop codons, except for *Mycoplasmopsis bovis PG45*
549 (GCF_000183385.1) and *Mycoplasma pneumoniae FH* (GCF_001272835.1), which used code 4 and
550 translate UGA to tryptophan.

551 Metagenome-assembled genomes

552 Medium- and low-quality MAG for *Escherichia coli*, *Cutibacterium acnes*, *Bacteroides uniformis*, and
553 *Staphylococcus epidermidis* were downloaded from the Global Microbial Genomic Bins database
554 (GMBC 1.0) [57], and their ORF were predicted with Prodigal using the same options as with high-
555 quality genomes.

556 For each species, we built 24 mixed pangenomes organized in two series, one for low-quality MAG
557 and the other for medium-quality MAG. Each series consists of 12 pangenomes, with 5%, 20%, 50%,
558 and 100% MAG (the rest being high-quality genomes), and three random replicates for each
559 composition. To build each pangenome, we randomly ampled the required number of genomes
560 from the collection of MAG and high-quality genomes. All mixed pangenomes contained a total of
561 100 genomes, except for *Bacteroides uniformis*, for which only 65 high-quality genomes were
562 available, and *Staphylococcus epidermidis,* for which only 88 and 84 medium and low quality MAGs
563 were available, respectively.

564 <u>De novo OGC construction</u>

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

570 Four sets of homology-based OGC were built with the sequence clustering tools MMseqs2 [29] and
571 CD-HIT [28], setting the minimum identity threshold to 50% and 80%. MMseqs2 was run with the
572 options 'easy-cluster' for cascaded clustering and 'cluster-mode 1' to define clusters based on
573 connected components. The options for CD-HIT were set so that sequence homology was calculated
574 locally, the alignments covered >80% of the longest sequence, and sequences were assigned to the

575 best-matching cluster (-G 0 -aL 0.8 -g 1 -M 8000). The word size for CD-HIT (option -n) was set to 5
576 for minimum identity 80% and 3 for minimum identity 50%, as recommended by the developers.
577 MMseqs2 and CD-HIT do not perform any kind of paralog splitting; therefore, the resolution of the
578 resulting OGC only depends on the similarity threshold used for clustering.

579 Two sets of synteny-based OGC were built with the software Roary [33] , setting the minimum
580 identity threshold to 80% and 95%. The Roary algorithm starts by pre-clustering highly-similar
581 protein sequences with CD-HIT [28] to obtain a smaller set of representative sequences. Roary then
582 conducts an all-against-all comparison with BLAST and filters the hits based on the user-provided
583 identity threshold. Based on the network of hits, representative sequences are clustered with MCL
584 [58] and the resulting clusters are merged with the pre-clusters. As a final step, Roary uses
585 conserved gene neighborhood information to split homologous groups containing paralogs into
586 groups of synteny-supported OGC.

587 Two sets of orthology-based OGC were built with the software panX [32] and OrthoFinder [31]. Both
588 algorithms initially cluster sequences in orthologous groups by performing an all-against-all
589 similarity search with DIAMOND [59] and posterior clustering with MCL. The hits retrieved by
590 DIAMOND are filtered only in terms of statistical significance, regardless of sequence identity, which
591 allows recovering relatively divergent homologs. In the panX algorithm, the sequences of these
592 initial clusters are aligned with MAFFT [54] and cluster-level phylogenetic trees are built with
593 FastTree [60]. Finally, panX obtains orthology-supported OGC by examining the resulting trees and
594 applying a set of heuristic rules to split paralogs from true species-level orthologs. OrthoFinder
595 directly builds phylogenetic trees from the DIAMOND scores, infers the root based on gene
596 duplication patterns, and generates OGC that are compatible with the rooted trees.

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

598 Reference-based OGC were built by mapping translated ORF to the eggNOG (evolutionary genealogy
599 of genes: Non-supervised Orthologous Groups) database version 5.0 [61]. For that purpose, we ran
600 eggNOG-mapper v2 [62] with command line options "*-m diamond --tax_scope_mode narrowest*" to
601 search queries against eggNOG sequences with DIAMOND and transfer orthologous group
602 annotations at the highest possible taxonomic resolution (which typically corresponds to the genus
603 level).

604 Pangenome features

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

608 Pangenome size was calculated as the total number of OGC retrieved for each species. Because this
609 measure is positively correlated with the number of genomes sampled, an unbiased estimate was
610 obtained by randomly subsampling sets of 14 genomes and taking the average over 100 realizations
611 (we refer to this and other pangenome features obtained with the same procedure as "14-mean").
612 To ensure that the results obtained with these two measures were sufficiently representative, we
613 calculated three additional measures of pangenome size: Chao's lower bound [63]; the normalized
614 pangenome size, obtained by dividing the pangenome size (P_{tot}) by the sum of the harmonic series
615 of the number of genomes (n), such that $P_{norm} = \frac{P_{tot}}{\sum_{i=1}^{n-1} 1/i}$ [64]; and the pangenome size of the 15
616 most dissimilar genomes in terms of genome content. After verifying that all those metrics were
617 highly correlated across species (R>0.95), we proceeded with the uncorrected and 14-mean
618 pangenome sizes, which are simpler and more easily interpretable.

619 We defined the core genome as the set of OGC contained in all the genomes of a species. This strict
620 definition is appropriate given the almost full (>99%) completeness filter imposed on high-quality

621 genomes. Pangenomes were also characterized in terms of the absolute number of single-copy core
622 OGC, accessory OGC (those that are not core), and singleton OGC (those encompassing a single
623 ORF). As measures of genome content diversity, we computed the mean percentage of OGC that
624 are accessory and singleton with respect to all the OGC present in each genome, averaged over 100
625 random subsamples of 14 genomes per species. In addition, a measure of genomic fluidity that
626 quantifies the average dissimilarity in gene content of randomly sampled pairs of genomes [65] was
627 obtained with the 'fluidity' function of the R package micropan [66]. To ensure that variations in
628 fluidity were not simply due to the randomness of the subsampling procedure, we conducted a
629 series of preliminary tests to determine the minimum number of sampled pairs required for
630 convergence in fluidity estimates. We found that using 500 random pairs (instead of only 10 pairs
631 that is the default in micropan) reduced the relative variability across repeated estimates below
632 0.5%. Therefore, we set the option 'n.sim = 500' for all fluidity calculations.

633 The species-wise core gene alignment similarity (CGAS) was calculated using ORFs that belong to
634 single-copy core OGC. Pairwise global alignments of all the ORFs assigned to the same OGC were
635 performed with the Needleman-Wunsch algorithm as implemented by the *needleall* program of the
636 EMBOSS suite [67]. After removing all self-alignments (alignments of ORFs with themselves), the
637 average identity of each pair of genomes was calculated as $\sum m_i / \sum L_i$ where m_i and L_i represent
638 the number of matches and total alignment length for the pair of sequences from the i-th OGC and
639 the sum extends over all single-copy core OGC. The species-wise CGAS was obtained as the average
640 over all pairs of genomes of the same species. The nucleotide sequence divergence in single-copy
641 core genes was defined as one minus the species-wise CGAS. Given the high computational cost of
642 this procedure, calculation of CGAS and core gene divergences was restricted to reference-, synteny-
643 , and orthology-based OGC.

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

645 High resolution trees were obtained for each species using concatenated alignments of all single-
646 copy core OGC obtained by Roary. Amino acid sequences for each single-copy core OGC were
647 aligned with mafft-linsi (L-INS-I Algorithm, default options) [54] and back-translated to nucleotide
648 alignments with the program pal2nal.pl using codon table 11 (except for species of the order
649 Mycoplasmatales, which use codon table 4) [68]. The nucleotide alignments were concatenated and
650 used as input for the tree construction program FastTree (command line options -gtr -nt -gamma -
651 nosupport -mlacc 2 -slownni) [60]. The preliminary tree produced by FastTree was subsequently
652 provided to RAxML for branch length optimization (options -f e -c 25 -m GTRGAMMA) [69]. The
653 package ETE3 [70] was used for mid-point rooting and visualization.

654 Gene gains and losses along each species tree were inferred with the phylogenomic reconstruction
655 software Gloome [71]. As the input for Gloome, we used the phyletic profiles for the presence or
656 absence of each OGC and the high-resolution species trees; options were set to optimize the tree
657 branch lengths under a genome evolution model with 4 categories of gamma-distributed gain and
658 loss rates and stationary frequencies at the root. To compare among OGC generation methods, we
659 varied the input phyletic profiles according to the desired method while keeping the species trees
660 unchanged. The Gloome optimization algorithm did not converge for *Enterobacter himalayensis* and
661 *Chlamydia muridarum*; therefore, those species were excluded from all the analyses involving gene
662 gains and losses.

663 Functional annotation and statistical analysis of functional profiles

664 Functional annotation at the gene level was done by mapping individual genes to custom-made
665 HMM profiles of the 2020 release of the COG database [72]. Functional annotation at the OGC level
666 was done by applying the majority rule to gene-level annotations. Coarse-grained pangenome

667 functional profiles were built by counting the number of OGC assigned to each of the 21 major
668 prokaryotic functional categories defined in the COG database.

669 The statistical analysis of functional profiles was conducted by applying the phylofactorization
670 framework [73]. First, to account for the constant-sum constraint that complicates the statistical
671 analysis of compositional data, we applied the isometric log-ratio (ILR) transform to the species-wise
672 functional profiles. Informative ILR balances were defined following a guide tree that was obtained
673 by calculating the mean differences between synteny and orthology OGC for each functional
674 category and performing hierarchical clustering of the functional categories based on such
675 differences. Then, linear mixed effects models were set out for each of the 20 ILR balances, with
676 OGC clustering criteria as fixed effects and species as random effects. Model fitting was performed
677 with the R package lmerTest [74] and contrasts among balances were conducted with the R package
678 phylofactor (https://github.com/reptalex/phylofactor) , which ranks the balances based on the
679 fraction of the total variance that is explained by the model. Statistical significance was calculated
680 by using Satterthwaite's approach to estimate the degrees of freedom of the F-statistic and applying
681 Bonferroni correction to account for multiple comparisons.

682 Comparison of high-quality and MAG-derived pangenomes

683 To quantify the inconsistency between pangenomes built from high-quality genomes and
684 pangenomes built from mixtures of MAG and high-quality genomes, we calculated the fraction of
685 unexplained variance in the latter that cannot be explained by the former. If the value of a given
686 pangenome feature in a given species is represented as a pair (x, y), where x and y represent,
687 respectively, the value in high-quality and mixed pangenomes, then the fraction of unexplained
688 variance is equal to $\sum(y-x)^2 / \sum(y-\bar{y})^2$, where the sum extends to all species and replicates and
689 $\bar{y}$ is the mean of y. Formally, this value is related to the coefficient of determination of a 1:1 fit as

690 $1 - R^2$. Therefore, inconsistencies greater than 1 are equivalent to $R^2 < 0$, implying that the (x, y)
691 pairs strongly deviate from a 1:1 trend (more precisely, the y-values are better predicted by their
692 mean than by the x-values).

693 A robust estimate of the core genome size in mixed pangenomes was obtained with the command
694 mOTUpan [42], using as inputs the OGC presence-absence matrices (option '-c') and the
695 completeness and contamination values reported in the GTDB and GMBC metadata files (option '-
696 k'). Moreover, mOTUpan estimates of the mean genome size and the core genome size were used
697 to derive an alternative measure of gene content diversity that accounts for genome
698 incompleteness. Given a predicted core genome with size C and mean predicted genome size G, the
699 gene content diversity measure is obtained as $1 - C/G$. Assuming that the true genome sizes are
700 similar among strains of the same species, this value is approximately equal to the proportion of
701 genes within a genome that are accessory. Ready-to-use code to calculate the predicted percentage
702 of accessory genes per genome from the output of mOTUpan is provided in Additional File 1.

703

704 **Declarations**

705 Ethics approval and consent to participate

706 Not applicable

707 Consent for publication

708 Not applicable

709 Availability of data and materials

710 The datasets generated and analyzed in this study are available in Zenodo, at
711 https://dx.doi.org/10.5281/zenodo.7387758

712 Competing interests

713 The authors declare that they have no competing interests.

714 Funding

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

723 Authors' contributions

724 JI and JHC designed the study; SMM and SGB generated the datasets for each clustering criterion;
725 SMM, YL, SGB, and JI generated and compared the pangenomes; SMM, YL, JHC, and JI wrote the
726 manuscript. All authors read and approved the final manuscript.

727 Acknowledgements

728 We thank J. Botelho and L. Dudbridge for their feedback and critical read of the manuscript.

729

730 **Figure legends**

731 **Figure 1: Homology, species-level orthology, and synteny conservation.** The phylogeny of a gene
732 family that has experienced intra-species HGT (*) and inter-species HGT (**) is superimposed on a

733 species tree (colored ribbons). In the species tree, speciation is followed by diversification into two
734 strains (blue), and a sister species is represented in green. Gene neighborhoods are displayed at the
735 leaves of the gene tree, with the gene of interest in the middle (labeled A-D, larger size, and red
736 border). All sequences (A-D) are homologs. However, only A, B, and C are species-level orthologs,
737 because they descend from the same ancestral sequence at the time of speciation. Of those, A and
738 C were vertically transmitted and share conserved gene neighborhoods (synteny). Sequences B and
739 C are in-paralogs, whereas C and D are out-paralogs.

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

754 **Figure 3: Comparison of core gene sets obtained from different OGC building methods.** (a)
755 Distribution of the mean copy number per core OGC per genome. (b) Fraction of all single-copy core
756 (scc) gene families that are exclusively recovered through orthology- or synteny-based paralog

757 splitting. (c) Distribution of synteny-exclusive, single-copy core OGC among orthology-based OGC.
758 The large bar at "1" implies that most synteny-exclusive single-copy core OGC are subsets of larger
759 orthology-based OGC that are core but not single-copy. (d) Distribution of orthology-exclusive,
760 single-core OGC among synteny-based OGC. Most of the single-copy core OGC exclusively supported
761 by orthology combine 2 accessory synteny-based OGC. In (c) and (d), gray bars denote method-
762 exclusive single-copy core OGC that include sequences that could not be processed by the other
763 method. (e) Nucleotide sequence divergence in single-copy core OGC exclusively supported by
764 orthology, synteny, or both criteria. To account for between-species differences in evolutionary
765 rates and phylogenetic tree spans, values were normalized by the species-level mean divergence
766 (calculated from all single-copy core OGC that were supported by both synteny and orthology).
767 Vertical lines indicate the distribution means for each group of single-copy core OGC. (f) Functional
768 differences among single-copy core OGC supported by different criteria. Each set of box plots
769 represents the balance (measured as the isometric log-ratio) between the relative frequencies of a
770 given functional category (x-axis) and the remaining categories not considered in previous sets (e.g.,
771 the second set of box plots corresponds to the balance between functional category N and all the
772 rest except X). The figure shows the 4 ILR balances with the greatest variation across methods.
773 Abbreviations of functional categories, X: mobilome; N: cell motility; Q: 2º metabolites biosynthesis,
774 transport and catabolism; J: translation, ribosomal structure and biogenesis. In (b) and (f), each data
775 point corresponds to one species; boxes span the 25-75 percentiles; the central line indicates the
776 median; whiskers extend to the most extreme data points that are not outliers; isolated points
777 denote outliers; notches (only in f) show the 95% confidence interval of the median.

778 **Figure 4: Systematic and specific biases in functional profiles associated with paralog splitting**
779 **criteria.** (a) Inconsistency of ORF assignations into OGC (normalized variation of information, top),
780 and fraction of OGC that exactly contain the same ORFs (bottom) under synteny and orthology

781 splitting criteria, stratified by functional category. (b) Absolute number (top) and relative fraction
782 (bottom) of synteny- and orthology-based OGC associated with each functional category. (c)
783 Balances (quantified as isometric log-ratios) for the functional categories that show the greatest
784 systematic variation between paralog splitting criteria. Each set of boxplots represents the balance
785 between the relative abundances of a group of functional categories (shown below) and all the
786 remaining categories not considered in previous sets. Each data point corresponds to the
787 pangenome of one species; boxes span the 25-75 percentiles; the central line indicates the median;
788 whiskers extend to the most extreme data points that are not outliers; isolated points denote
789 outliers. (d) Standardized residuals (Z-scores) of the linear mixed effects model used to infer the
790 systematic differences shown in (c). Each row corresponds to the pangenome of one species, sorted
791 according to the GDTB species tree [75] (phyla colored as in Fig 1). Colored cells indicate a significant
792 excess of synteny- (green) or orthology-based (purple) OGC from a given category in a specific
793 pangenome that is not explained by the general trends in (c). Abbreviations of functional categories,
794 C: energy production and conversion; D: cell cycle control, cell division, chromosome partitioning;
795 E: amino acid transport and metabolism; F: nucleotide transport and metabolism; G: carbohydrate
796 transport and metabolism; H: coenzyme transport and metabolism; I: lipid transport and
797 metabolism; J: translation, ribosomal structure and biogenesis; K: transcription; L: replication,
798 recombination and repair; M: cell wall/membrane/envelope biogenesis; N: cell motility; O:
799 posttranslational modification, protein turnover, chaperones; P: inorganic ion transport and
800 metabolism; Q: secondary metabolites biosynthesis, transport and catabolism; R: general function
801 prediction only; S: function unknown; T: signal transduction mechanisms; U: intracellular trafficking,
802 secretion, and vesicular transport; V: defense mechanisms; X: mobilome: prophages, transposons.

803 **Figure 5: Effect of paralog splitting criteria on the inference of gene flux**. (a) Species-wise
804 comparison of the total gene flux (gains and losses along the species tree, left), genomic content vs

805 gene sequence evolution ratio (middle), and total gain vs loss ratio (right) inferred from synteny-
806 and orthology-based OGC. The method-dependent uncertainty is equal to $1 – R^2$. (b) Flux per gene
807 per functional category inferred from orthology- (top) and synteny-based (middle) OGC. Each data
808 point corresponds to one species; boxes span the 25-75 percentiles; the central line indicates the
809 median; whiskers extend to the most extreme data points that are not outliers; isolated points
810 denote outliers. The bar plot at the bottom shows the inconsistency between methods, quantified
811 as one minus the squared rank correlation. Abbreviations of functional categories as in Fig. 2. (c)
812 Median flux per gene per category, calculated over all the species, for orthology-based (x-axis) and
813 synteny-based (y-axis) OGC. Similar trends are observed for gains, losses, and the combination of
814 both.

815 **Figure 6: Estimation of pangenome properties from incomplete genomes.** (a) Comparison between
816 pangenome properties inferred from high-quality genomes (x-axis) and mixtures of medium-quality
817 MAG and high-quality genomes (y-axis). From left to right, the scatter plots correspond to
818 pangenomes with 5%, 20%, 50% and 100% of MAG. Each point in the scatter plots corresponds to
819 one pangenome, with different symbols and colors used to distinguish among species and OGC
820 generation methods, respectively. Each scatter plot combines data from 4 species, 6 gene clustering
821 methods, and 3 random subsamples. The bar plots on the right summarize the observed
822 inconsistencies, with color intensities representing the fraction of MAG. Note that panX fails to
823 produce results in pangenomes that contain >5% of MAG (purple asterisks). (b) Sensitivity of gene
824 clustering methods to the addition of MAGs, calculated by comparing the cluster assignations of
825 genes from high-quality genomes before and after adding MAG. NVI: normalized variation of
826 information. Different OGC generation methods are color coded, with color intensities indicating
827 the fraction of MAG. Note that reference-database mapping methods produce an NVI equal to zero
828 (double asterisk).

829 **References**

830 1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of
831 multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-
832 genome." Proc Natl Acad Sci U S A. 2005;102:13950–5.

833 2. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin
834 Microbiol. 2015;23:148–54.

835 3. Collins RE, Higgs PG. Testing the infinitely many genes model for the evolution of the bacterial
836 core genome and pangenome. Mol Biol Evol. 2012;29:3413–25.

837 4. Liao J, Guo X, Weller DL, Pollak S, Buckley DH, Wiedmann M, et al. Nationwide genomic atlas of
838 soil-dwelling Listeria reveals effects of selection and population ecology on pangenome evolution.
839 Nat Microbiol. 2021;6:1021–30.

840 5. Whelan FJ, Hall RJ, McInerney JO. Evidence for Selection in the Abundant Accessory Gene Content
841 of a Prokaryote Pangenome. Mol Biol Evol. 2021;38:3697–708.

842 6. Maistrenko OM, Mende DR, Luetge M, Hildebrand F, Schmidt TSB, Li SS, et al. Disentangling the
843 impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME
844 J. 2020;14:1247–59.

845 7. Shapiro BJ. The population genetics of pangenomes. Nat Microbiol. 2017. p. 1574.

846 8. Wu D, Jospin G, Eisen JA. Systematic identification of gene families for use as "markers" for
847 phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major
848 subgroups. PLoS One. 2013;8:e77033.

849 9. Lerat E, Daubin V, Moran NA. From gene trees to organismal phylogeny in prokaryotes: the case
850 of the gamma-Proteobacteria. PLoS Biol. 2003;1:E19.

851 10. Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet.
852 2020;21:428–44.

853 11. Petitjean C, Makarova KS, Wolf YI, Koonin EV. Extreme Deviations from Expected Evolutionary
854 Rates in Archaeal Protein Families. Genome Biol Evol. 2017;9:2791–811.

855 12. Ahrens JB, Teufel AI, Siltberg-Liberles J. A Phylogenetic Rate Parameter Indicates Different
856 Sequence Divergence Patterns in Orthologs and Paralogs. J Mol Evol. 2020;88:720–30.

857 13. Fitch WM. Homology a personal view on some of the problems. Trends Genet. 2000;16:227–31.

858 14. Gabaldón T, Koonin EV. Functional and evolutionary implications of gene orthology. Nat Rev
859 Genet. 2013;14:360–6.

860 15. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309–38.

861 16. Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. Microbial genome analysis: the
862 COG approach. Brief Bioinform. 2019;20:1063–70.

863 17. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for Gene Orthology
864 inference. Brief Bioinform. 2011;12:379–91.

865 18. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-
866 Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol Biol Evol.
867 2017;34:2115–22.

868 19. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished
869 prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21:180.

870 20. Sitto F, Battistuzzi FU. Estimating Pangenomes with Roary. Mol Biol Evol. 2020;37:933–9.

871 21. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics
872 toolbox for clustering diverged orthologues in bacteria. Gigascience. 2019;8.

873 22. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification
874 of genome dynamics in prokaryote supergenomes. BMC Biol. 2014;12:66.

875 23. Hao W, Golding GB. The fate of laterally transferred genes: life in the fast lane to adaptation or
876 death. Genome Res. 2006;16:636–43.

877 24. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein
878 families in prokaryotes. PLoS Genet. 2011;7:e1001284.

879 25. Cummins EA, Hall RJ, McInerney JO, McNally A. Prokaryote pangenomes are dynamic entities.
880 Curr Opin Microbiol. 2022;66:73–8.

881 26. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial
882 genomic epidemiology with PopPUNK. Genome Res. 2019;29:304–16.

883 27. Csurös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood.
884 Bioinformatics. 2010;26:1910–2.

885 28. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing
886 data. Bioinformatics. 2012;28:3150–2.

887 29. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis
888 of massive data sets. Nat Biotechnol. 2017;35:1026–8.

889 30. Altenhoff AM, Dessimoz C. Inferring orthology and paralogy. Methods Mol Biol. 2012;855:259–
890 79.

891 31. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics.
892 Genome Biol. 2019;20:238.

893 32. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. Nucleic Acids Res.
894 2018;46:e5.

895 33. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale
896 prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.

897 34. Fouts DE, Brinkac L, Beck E, Inman J, Sutton G. PanOCT: automated clustering of orthologs using
898 conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related
899 species. Nucleic Acids Res. 2012;40:e172.

900 35. Zhou Z, Charlesworth J, Achtman M. Accurate reconstruction of bacterial pan- and core genomes
901 with PEPPAN. Genome Res. 2020;30:1667–79.

902 36. Perrin A, Rocha EPC. PanACoTA: a modular tool for massive microbial comparative genomics.
903 NAR Genom Bioinform. 2021;3:lqaa106.

904 37. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, et al. PPanGGOLiN: Depicting
905 microbial diversity via a partitioned pangenome graph. PLoS Comput Biol. 2020;16:e1007732.

906 38. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized
907 bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat
908 Biotechnol. 2018;36:996–1004.

909 39. Casjens SR, Mongodin EF, Qiu W-G, Luft BJ, Schutzer SE, Gilcrease EB, et al. Genome stability of
910 Lyme disease spirochetes: comparative genomics of Borrelia burgdorferi plasmids. PLoS One.
911 2012;7:e33280.

912 40. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, et al. Genomic sequence of
913 a Lyme disease spirochaete, Borrelia burgdorferi. Nature. 1997;390:580–6.

914 41. Sela I, Wolf YI, Koonin EV. Theory of prokaryotic genome evolution. Proc Natl Acad Sci U S A.
915 2016;113:11399–407.

916 42. Buck M, Mehrshad M, Bertilsson S. mOTUpan: a robust Bayesian approach to leverage
917 metagenome-assembled genomes for core-genome estimation. NAR Genom Bioinform.
918 2022;4:lqac060.

919 43. Iranzo J, Wolf YI, Koonin EV, Sela I. Gene gain and loss push prokaryotes beyond the homologous
920 recombination barrier and accelerate genome sequence divergence. Nat Commun. 2019;10:5376.

921 44. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification
922 of genome dynamics in prokaryote supergenomes. BMC Biol. 2014;12:66.

923 45. Cohen O, Pupko T. Inference and characterization of horizontally transferred gene families using
924 stochastic mapping. Mol Biol Evol. 2010;27:703–13.

925 46. Sela I, Wolf YI, Koonin EV. Selection and genome plasticity as the key factors in the evolution of
926 bacteria. Phys Rev X. 2019;9.

927 47. Harris CD, Torrance EL, Raymann K, Bobay L-M. CoreCruncher: Fast and Robust Construction of
928 Core Genomes in Large Prokaryotic Data Sets. Mol Biol Evol. 2021;38:727–34.

929 48. Stott CM, Bobay L-M. Impact of homologous recombination on core genome phylogenies. BMC
930 Genomics. 2020;21:829.

931 49. Iranzo J, Cuesta JA, Manrubia S, Katsnelson MI, Koonin EV. Disentangling the effects of selection
932 and loss bias on gene dynamics. Proc Natl Acad Sci U S A. 2017;114:E5616–24.

933 50. Iranzo J, Gómez MJ, López de Saro FJ, Manrubia S. Large-scale genomic analysis suggests a
934 neutral punctuated dynamics of transposable elements in bacterial genomes. PLoS Comput Biol.
935 2014;10:e1003680.

936 51. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-
937 to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38:1079–86.

938 52. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum
939 information about a single amplified genome (MISAG) and a metagenome-assembled genome
940 (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35:725–31.

941 53. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality
942 of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res.
943 2015;25:1043–55.

944 54. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements
945 in performance and usability. Mol Biol Evol. 2013;30:772–80.

946 55. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic
947 algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.

948 56. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene
949 recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

950 57. Coelho LP, Alves R, Del Río ÁR, Myers PN, Cantalapiedra CP, Giner-Lamia J, et al. Towards the
951 biogeography of prokaryotic genes. Nature. 2022;601:252–6.

952 58. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein
953 families. Nucleic Acids Res. 2002;30.

954 59. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using
955 DIAMOND. Nat Methods. 2021;18:366–8.

956 60. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large
957 alignments. PLoS One. 2010;5.

958 61. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG
959 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090
960 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–14.

961 62. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2:
962 Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale.
963 Mol Biol Evol. 2021;38:5825–9.

964 63. Chao A. Estimating the population size for capture-recapture data with unequal catchability.
965 Biometrics. 1987;43.

966 64. Bobay L-M, Ochman H. Factors driving effective population size and pan-genome evolution in
967 bacteria. BMC Evol Biol. 2018;18:153.

968 65. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene

969 diversity within microbial populations. BMC Genomics. 2011;12:1–10.

970 66. Snipen L, Liland KH. micropan: an R-package for microbial pan-genomics. BMC Bioinformatics.
971 2015;16:79.

972 67. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and
973 sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47:W636–41.

974 68. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into
975 the corresponding codon alignments. Nucleic Acids Res. 2006;34:W609–12.

976 69. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
977 phylogenies. Bioinformatics. 2014;30:1312–3.

978 70. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of
979 Phylogenomic Data. Mol Biol Evol. 2016;33:1635–8.

980 71. Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. GLOOME: gain loss mapping engine.
981 Bioinformatics. 2010;26:2914–5.

982 72. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database
983 update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids
984 Res. 2021;49:D274–81.

985 73. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, et al. Phylogenetic
986 factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ.
987 2017;5:e2969.

988 74. Kuznetsova A, Brockhoff PB, Christensen RHB. LmerTest package: Tests in linear mixed effects
989 models. J Stat Softw. 2017;82.

990 75. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. GTDB: an ongoing
991 census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized
992 and complete genome-based taxonomy. Nucleic Acids Res. 2022;50:D785–94.

993 76. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic
994 genomes. Genome Res. 2003;13:2178–89.

995 77. Dessimoz C, Cannarozzi G, Gil M, Margadant D, Roth A, Schneider A, et al. OMA, A
996 Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome
997 Data: Introduction and First Achievements. Comparative Genomics. Springer Berlin Heidelberg;
998 2005. p. 61–72.

999 78. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin EV, Mushegian A. A low-
1000 polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric
1001 best matches. Bioinformatics 2010;26:1481-7.

1002 79. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished
1003 prokaryotic pangenomes with the Panaroo pipeline. Genome Biology 2020;21:180.

1004 80. Peng Y, Tang S, Wang D, Zhong H, Jia H, Cai X. MetaPGN: a pipeline for construction and graphical

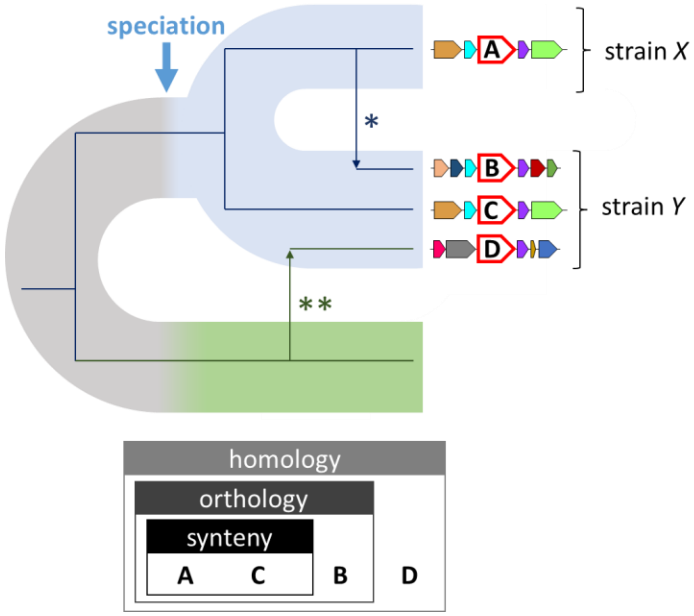
1005 visualization of annotated pangenome networks. GigaScience 2018;7:giy121.

1006 81. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: pan-genomes analysis pipeline. Bioinformatics
1007 2012;28:416-8.

1008 82. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL. Automatic clustering of orthologs and inparalogs
1009 shared by multiple proteomes. Bioinformatics 2006;22:e9-e15.

1010

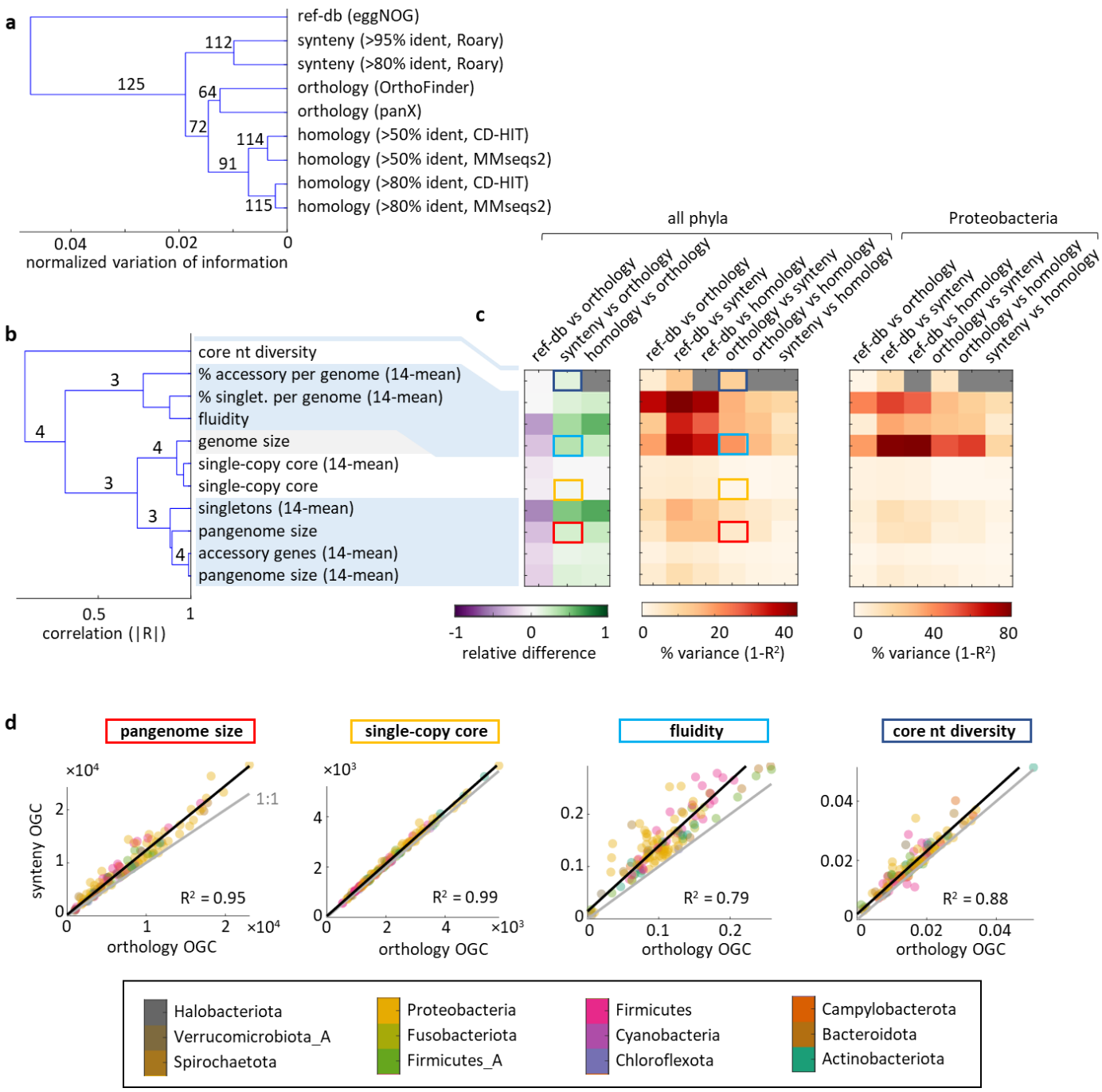
1011 **Figure 1**



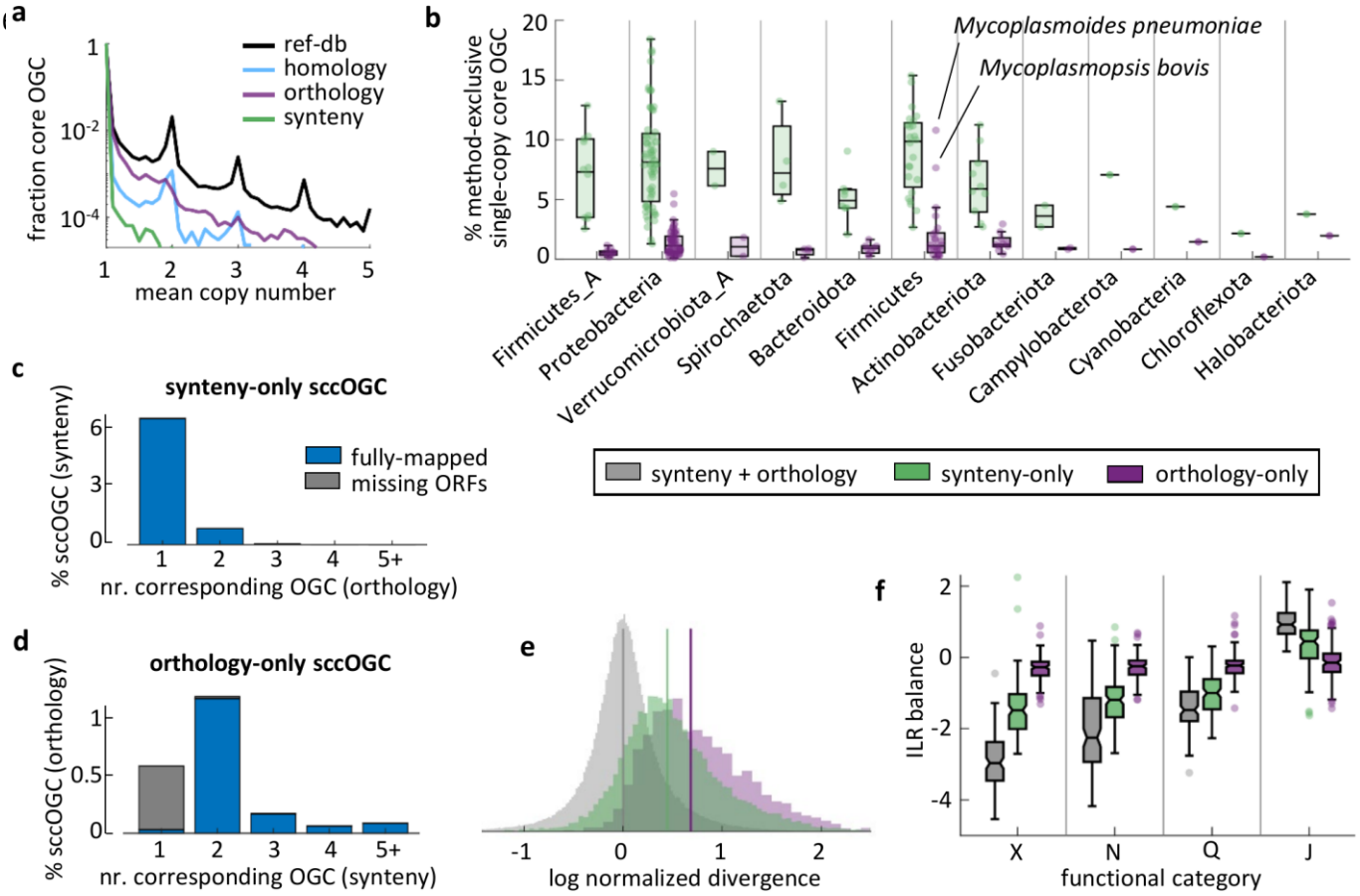
1012

1013 **Figure 2**

1014

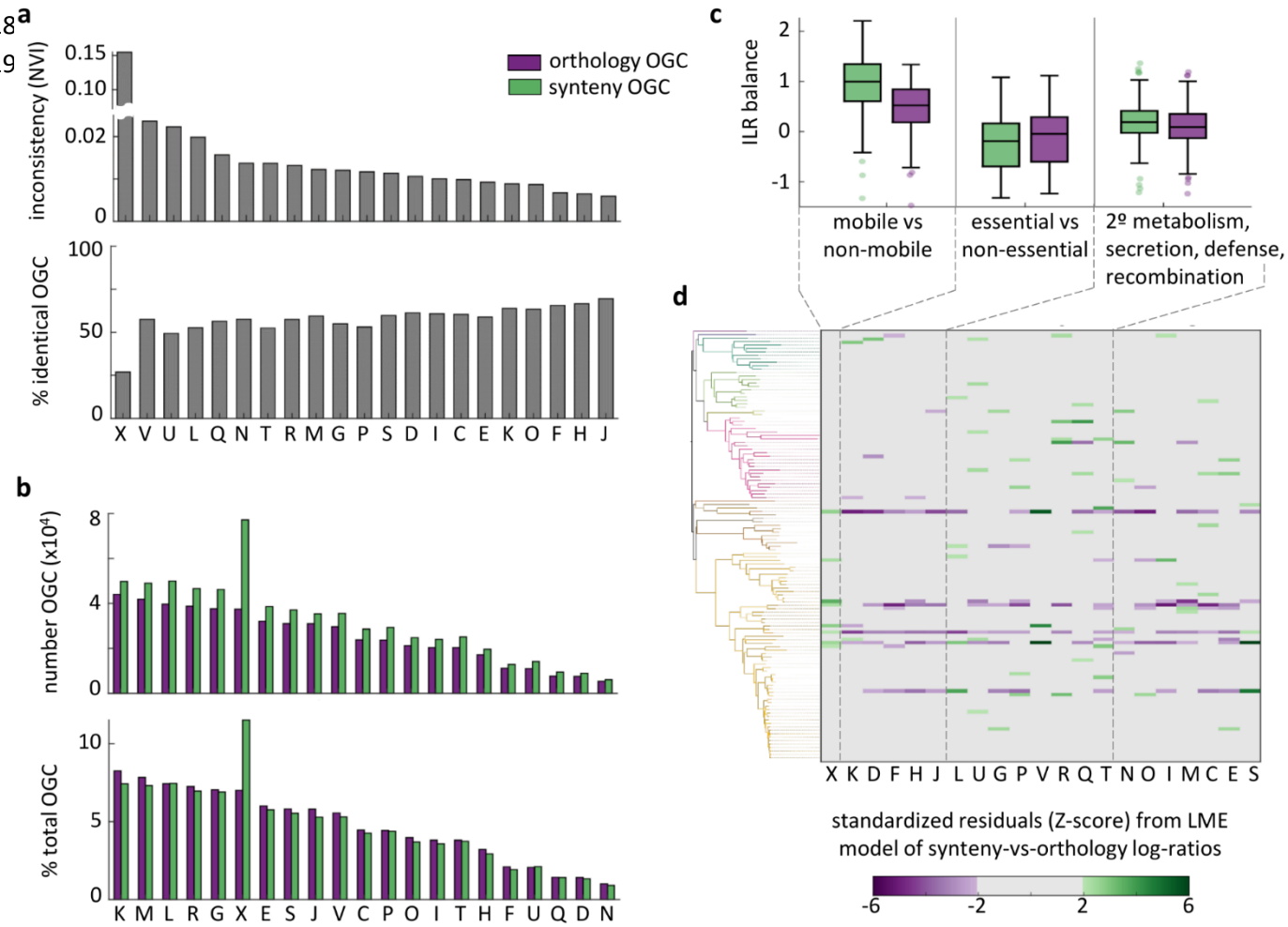


1015 **Figure 3**

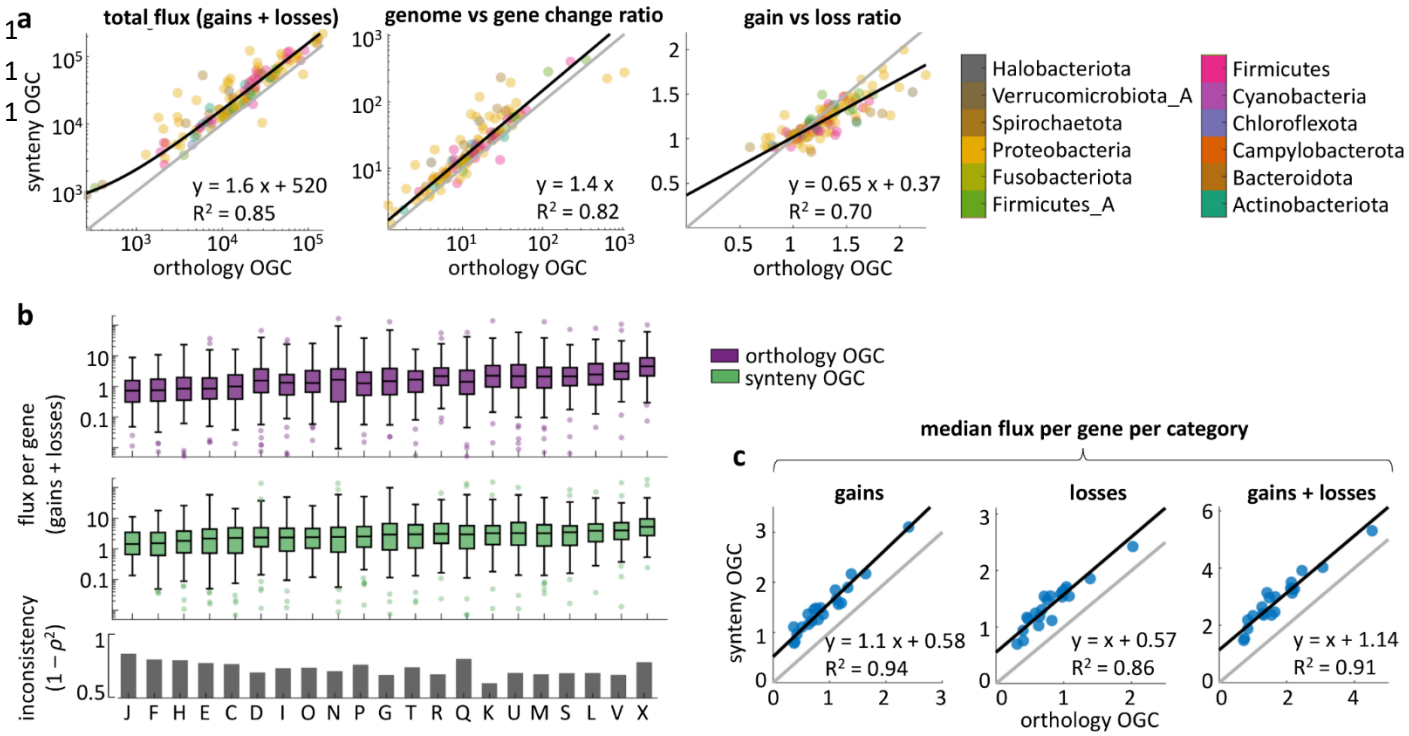
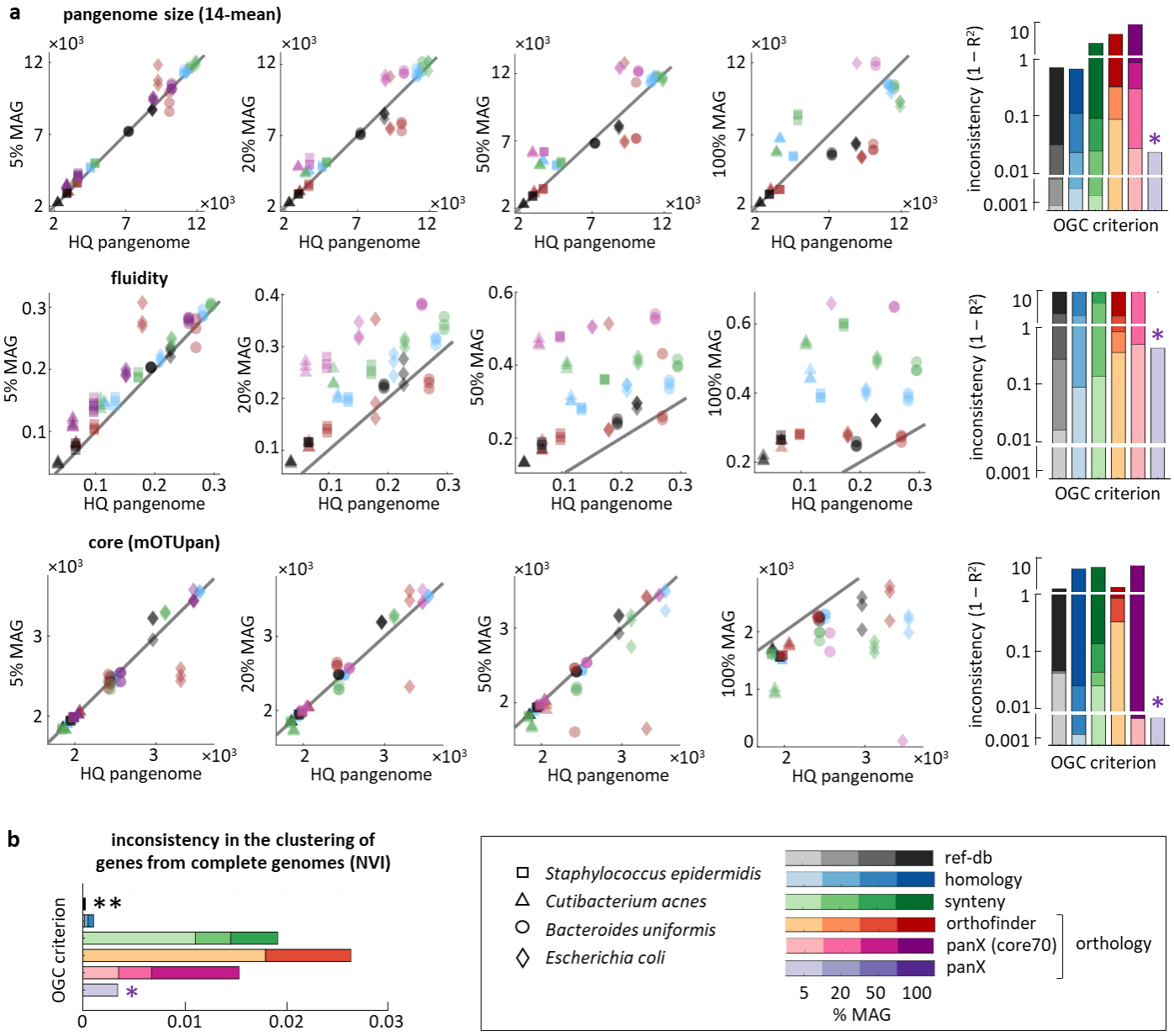


1017 **Figure 4**

1018 1019



1020 **Figure 5**



1024 **Figure 6**