Genotyping sequence-resolved copy number variation using pangenomes reveals paralog-specific global diversity and expression divergence of duplicated genes

Walfred Ma[1], Mark JP Chaisson[1,2,3]

1. Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA.
2. Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA
3. Corresponding author

Abstract

Copy number variant (CNV) genes are important in evolution and disease, yet sequence variation in CNV genes remains a blind spot in large-scale studies. We present ctyper, a method that leverages pangenomes to produce allele-specific copy numbers with locally phased variants from next-generation sequencing (NGS) reads. Benchmarking on 3,351 CNV genes, including *HLA*, *SMN*, and *CYP2D6*, and 212 challenging medically relevant (CMR) genes that are poorly mapped by NGS, ctyper captures 96.5% of phased variants with ≥99.1% correctness of copy number on CNV genes and 94.8% of phased variants on CMR genes. Applying alignment-free algorithms, ctyper requires 1.5 hours per genome on a single CPU. The results largely improve predictions of gene expression compared to known expression quantitative trait loci (eQTL) variants. Allele-specific expression quantified divergent expression on 7.94% of paralogs and tissue-specific biases on 4.68% of paralogs. We found reduced expression of *SMN2* due to *SMN1* conversion, potentially affecting spinal muscular atrophy, and increased expression of translocated duplications of *AMY2B*. Overall, ctyper enables biobank-scale genotyping of CNV and CMR genes.

28 # Introduction

29 Human genomes are characterized by frequent duplications and deletions, leading
30 to copy number variation. Up to 10% of protein-coding genes are known to be copy-
31 number variable, showing distinct distributions across human populations[1,2] and
32 association with traits such as body mass index[3] and disease, including cancer[4],
33 cardiovascular diseases[5], and neurodevelopmental disorders[6,7]. While copy number
34 variants (CNVs) are infrequent genome-wide, regions of long, low-copy repeats called
35 segmental duplications (SDs) are enriched in genes and are catalysts for recurrent CNVs
36 due to non-allelic homologous recombination[8–10]. These regions include *TBC1D3*, *NPIP*,
37 and *NBPF*[11,12] with particular impact in humans through association with brain function
38 and adaptation[13–16]. SD regions not only undergo frequent copy number changes but also
39 experience more rapid structural alterations and exhibit an elevated rate of nucleotide
40 substitutions compared to nonrepetitive DNA[17]. As a result, studies that only consider the
41 aggregate copy number (aggreCN) of duplicated genes miss non-reference gene
42 duplicates and variation among multiallelic CNVs, which have both been shown to
43 influence phenotypes and disease susceptibility[18–20], including hypertension and type 2
44 diabetes[21].

45

46 There are limited studies on sequence variation between gene duplicates,
47 particularly in studies using short-read next-generation sequencing data. Existing CNV-
48 calling tools detect excess coverage on reference alignments, hiding variants among
49 copies[22]. Furthermore, alignments to a single reference contain ambiguity and bias[23] and
50 are unable to detect novel duplicates. Advances in single-molecule sequencing enabled

the assembly of pangenomes from diverse populations[24–26]. Almost all novel sequences found in the pangenome are CNV events[24], providing sequence-resolved CNVs and non-reference duplications. While graph-based pangenomes reduce reference bias[27], they merge similar sequences—including alternative alleles and paralogs from distinct genomic locations and functions—into shared graph paths, obscuring paralog-specific variants and novel duplications[28]. Furthermore, as pangenomes include more samples, the alternative sequences, gene conversion, and genome rearrangements that are prevalent in divergent genomic regions present an even greater challenge to be represented as graphs[17], motivating the need to develop tools to analyze duplicated genes using NGS reads with pangenomes.

Here, we present ctyper, a method that directly compares NGS reads to unmerged pangenome haplotypes to identify the most similar genomic segments corresponding to genes in an NGS sample and determine their copy number states. By avoiding sequence merging, ctyper preserves locally phased variants and captures complex variation, such as structural variation and gene conversion, that are often missed in NGS analyses. In this study, we focus on sequences of complex, duplicated genes that would otherwise be challenging to analyze using NGS data and canonical reference genomes. Leveraging alignment-free techniques, and introducing a novel mathematical model for genotyping haplotypes with a polynomial-time solution, ctyper achieves both high accuracy and computational efficiency, making it well-suited for large-scale biobank data analyses.

Results

Overview of the genotyping method

Our method, ctyper, genotypes sequence-resolved copy number variation in NGS data leveraging haplotype-resolved assemblies. Instead of representing genetic variation in the form of individual variants, we represent variation as haplotype segments that are short enough to minimize disruption by recombination, allowing precise sharing with an NGS sample through identity-by-descent[29], and long enough to capture structural information of genomic sequences. We refer to these haplotype segments as pangenome-derived alleles (PAs). PAs capture phased variation, representing structural variation, gene conversion events, and small variants. In this study, we set the boundaries of PAs to study the CNVs of protein-coding genes. Each PA includes consecutive exons separated by <20 kb, with an additional 5 kb of upstream and downstream sequences added as anchors (Methods). The 20 kb cutoff reflects both functional proximity, as most short-range transcription factors operate within this range[30], as well as population-level genomic linkage. PAs typically range from 10 to 100 kb, corresponding to the scale of linkage disequilibrium (LD) blocks[31], within which most single-nucleotide polymorphisms (SNPs) are in strong LD and can be treated as single units of inheritance. While PAs generally correspond to individual genes, they also may cover only a fraction of a gene flanked by long introns or, conversely, include paralogs within the 20 kb boundary.

For computational efficiency and to avoid alignment ambiguity in repetitive DNA, we use an alignment-free comparison of low-copy *k*-mers (DNA fragments of a fixed length *k*, here $k = 31$) measured in query NGS samples to genotype PAs. For each gene

97 of interest, we group all similar PAs in the pangenome, including orthologous genes,
98 paralogs, and homologous pseudogenes, to construct a matrix to be used in genotyping
99 that contains the *k*-mer composition of all similar PAs (Methods). The rows of a matrix
100 correspond to individual PAs, while columns correspond to *k*-mers found within the group
101 of PAs and not elsewhere in the genome. The values of the matrix reflect the multiplicity
102 of a *k*-mer in each PA (Methods; Figs. 1a,b).
103
104 The genotyping of an NGS sample is performed per matrix by identifying a
105 combination of PAs (rows) and their allele-specific copy numbers such that their total
106 corresponding *k*-mer counts are the least-squared distance to observed *k*-mer counts
107 from the sample. To do so, the *k*-mer counts from a sample are projected into the vector
108 space of each *k*-mer matrix, and then assigned integer copy numbers using recursive
109 rounding based on the phylogenetic tree of PA sequences (Methods; Figs. 1c-e). The
110 genotyping results are a list of PA-specific copy numbers (paCNs).
111
112 As an example, there are 178 PAs for *SMN* genes, the gene family associated with
113 spinal muscular atrophy. This includes copies of *SMN1* and *SMN2* as well as paralogs
114 that have undergone gene conversion[32], for example paralogs mapped to the *SMN2* locus
115 but containing the *SMN1* version of phe-280, the SNP responsible for dysfunctional exon
116 7 splicing of *SMN2*[33] (Fig. 1f).
117
118
119

120 **Pangenome-derived allele database construction and annotation**

121 We constructed a PA database for 3,351 genes previously reported as CNV[24,26]
122 (Supplementary Table 1), using 114 diploid PacBio-HiFi assemblies from the Human
123 Pangenome Reference Consortium (HPRC), Human Genome Structural Variation
124 Consortium (HGSVC), Chinese Pangenome Consortium (CPC), and two haplotype-
125 phased telomere-to-telomere assemblies[34,35], in addition to GRCh38 and CHM13[36]. PAs
126 were extracted from assemblies and annotated with other similar PAs within the same
127 matrices used for genotyping, for example, the amylase genes (Fig. 2a). In total, we
128 defined 1,408,209 PAs and organized them into 3,307 matrices (Fig. 2b,c).

129

130 Because the human population exhibits limited genetic diversity and LD is stronger
131 within a short distance, PAs are often highly similar or identical. To reduce genotyping
132 dimensionality and facilitate cohort analysis, particularly when sample sizes are limited,
133 we used the phylogenetic relationship between sequences to merge similar PAs into
134 highly similar subgroups (subgroups) to be treated as equal states (Methods). In total, we
135 defined 89,236 such subgroups, which were used to enumerate all PA states observed
136 in the pangenome, analogous to *HLA* nomenclature (Supplementary Figure 1).

137

138 To annotate low-frequency variants not represented by subgroups and annotate
139 orthologous/paralogous relationships, we mapped PAs to GRCh38 genes
140 (Supplementary Methods). The alignments provide sequence variation at a greater
141 resolution than what is annotated by subgroups as well as the reference genomic
142 locations. In total, 164,237 paralogous PAs across 6,389 loci were determined. The

143 paralogous PAs that were similar to their corresponding reference locus (≥80% *k*-mer
144 similarity) were labeled as duplicative paralogs, and the remaining lower-identity
145 paralogous PAs were labeled as diverged paralogs. In total, 10,792 diverged paralogs
146 from 2,734 subgroups were identified across 333 matrices (Fig. 2d). The divergent
147 paralogs represent novel sequences recalcitrant to analysis using a single reference
148 genome. For example, some amylase PAs include paralogs for both *AMY1* and *AMY2B*
149 with low *k*-mer similarity to reference loci and were annotated as diverged paralogs, which
150 reflects an *AMY2B* translocation event (Fig. 2a).
151
152 While most duplication events happen on different loci from their original genes,
153 6,673 PAs reflected proximal (<20 kb) duplications, including 1,646 PAs across 36 genes
154 that exhibited "runway duplication"[37] with at least three proximal duplications
155 (Supplementary Figure 2). Proximally duplicated genes were grouped with their ortholog
156 genes in the same PAs as they are unlikely to be inherited independently. Since ctyper
157 can also genotype orthologous genes, we classified orthologous genes into two
158 categories: reference alleles, which belong to the same subgroup as the reference gene,
159 and alternative alleles, which fall outside of these subgroups (Fig. 2d).
160
161 **Pangenome-derived alleles and highly similar subgroups capture unique aspects**
162 **of population diversity**
163 Since other variant representations have been used in CNV studies, we assessed
164 whether PAs capture unique aspects of genomic information that cannot be replicated by
165 these methods. We compared paCNs to other CNV representations, including copy

166 numbers of reference genes[1,37], singly unique nucleotide *k*-mers[1,37,38] (SUNKs), and large
167 haplotype structures[39–42]. We found that PAs provide higher resolution of variation (e.g.,
168 SNVs), as 94.7% of variants are not reflected by the sequences of genes in GRCh38.
169 Additionally, both nearby SUNK markers (Fig. 2e) and large haplotype structures were
170 found to be poor proxies for PAs, and only a small proportion of paCNs were found to link
171 to SUNKs or large haplotype structures (Methods).
172
173 We evaluated the extent to which global population diversity could be represented
174 using subgroups compared to individual PAs. We found that even with largely reduced
175 dimensions, subgroups still capture more than 80% of the total population variation (Fig.
176 2f). Additionally, we performed saturation analysis[43,44] to estimate the percentage of the
177 unobserved subgroups in each new genome. The current cohort represents 98.7% of
178 subgroups in non-Africans and 94.9% in Africans (Methods), suggesting a near-saturated
179 database (Fig. 2g).
180
181 **Genotyping pangenome-derived alleles among NGS samples and benchmarking**
182 **results**
183 We applied ctyper to genotype NGS samples from the 1000 Genomes Project
184 (1kGP), including 2,504 unrelated individuals and 641 offspring. At the cohort level,
185 excluding intronic/decoy PAs, bias was measured using Hardy-Weinberg equilibrium
186 (Methods), and accuracy was measured using trio concordance (Methods;
187 Supplementary Table 2). Overall, 0.75% of subgroups had significant disequilibrium ($p <$
188 0.05) (Fig. 3a) and 27 matrices having >15% subgroups with significant disequilibrium,

189 which were mostly short genes (median = 4,564 bp) with fewer low-copy *k*-mers
190 (Supplementary Table 3). The average F-1 score for trio concordance was 97.58% (Fig.
191 3b), while 18 matrices had high discordance (>15%), primarily for subtelomeric genes or
192 on sex chromosomes (Supplementary Table 4).
193
194 We assessed copy number accuracy and bias in challenging and highly duplicated
195 gene families, such as amylase, *NBPF*, *NOTCH*, *GOLGA*, and *TBC1D3*. The copy
196 numbers derived from genotyping were compared to those in their corresponding
197 assemblies for 39 HPRC samples shared with the 1kGP, using a database inclusive of
198 these samples. For each sample, we benchmarked on all matrices where the
199 corresponding assembly was high in copy number (>10 PAs) and no low-confidence
200 sequences filtered during matrix construction (Methods). The results showed a high
201 correlation (ρ=0.996, Pearson correlation) and little bias (Fig. 3c), with a 0.2% rate of
202 missing copies (false negatives) and a 2.4% rate of additional copies (false positives),
203 with a higher false positive rate likely explained by unassembled genes in the assemblies.
204 When expanding benchmarking to include the remaining less challenging genes within
205 matrices without filtered sequences on all 2,504 1kGP samples, the copy number
206 between genotype and assemblies had an effectively identical correlation (ρ=0.996,
207 Pearson correlation).
208
209 To evaluate sequence accuracy at the individual PA level, we performed assembly
210 comparisons to assess the sequence similarity of the genotyped alleles to the sample
211 assembly for the 39 HPRC benchmarking genomes. Each sample was genotyped with

212 the full database (full-set) or the database excluding its corresponding PAs (leave-one-
213 out). We assigned the genotyped PAs to the corresponding assembly (Methods),
214 excluding introns/decoys and sequences with <1 kb non-repetitive bases, and measured
215 the similarity between the genotyped allele and assigned query using global
216 alignment[45,46]. We performed a similar analysis treating the closest neighbor to each
217 assembly PA from the database as the correctly genotyped locus. Due to mismatching
218 from database sampling or misassemblies, 2.9% of PAs from the leave-one-out and 1.0%
219 from the full-set could not be paired to truth copies for assessment. Using the full-set
220 database, paired PAs have 0.36 mismatches per 10 kb, with 93.0% having no
221 mismatches on non-repetitive regions. The leave-one-out tests had 2.7 mismatches per
222 10 kb on non-repetitive regions, which was 1.2 additional mismatches per 10 kb from the
223 optimal solutions (closest neighbors); 57.3% alleles had no mismatches, and 77.0% were
224 mapped to the optimal solution (Fig. 3d). The leave-one-out results were 96.5% closer to
225 the original PAs compared to the closest GRCh38 gene at 79.3 mismatches per 10 kb.
226
227 To isolate sources of errors in cases of misassembled duplications, we directly
228 compared leave-one-out genotyping results to a telomere-to-telomere assembly[34],
229 filtering out intronic/decoy sequences. The sample genotypes had 11,627 correctly
230 matched subgroups, 599 (4.8%) mistyped to other subgroups, 131 from singleton
231 subgroups removed from the database during the test (1.1%; out of reference), 127 false
232 positives (0.5% F-1), and 93 false negatives (0.4% F-1) for a total F-1 error of 6.7%
233 (Methods; Fig. 3e). This is a 3% increase in mistypes compared to trio discordance. The
234 copy number agreement was 99.1%.

235

236 The computational requirements are sufficient for biobank analysis. The average
237 runtime for genotyping 3,351 genes at 30× coverage was 80.2 minutes (1.0 min/1×
238 coverage for sample preprocessing, and 0.9 s/gene for genotyping) on a single core (Fig.
239 3f) using ~20GB RAM, with support for parallel processing.

240

241 We compared the benchmarking of *HLA*, *KIR*, and *CYP2D6* to the locus-specific
242 methods T1K[47] and Aldy[48]. For 31 *HLA* genes, ctyper had an F1-score of 98.9% in
243 predicting all four fields of *HLA* nomenclature[49,50] against the full-set and 86.3% among
244 the leave-one-out, while T1K had 70.8%. For protein-coding products (first two fields),
245 ctyper reached 99.98% against the full-set (with 99.9% copy number F1-correctness) and
246 96.5% (with 99.5% copy number F1-correctness) among the leave-one-out, and T1K had
247 97.2% (Fig. 3g) (Supplementary Table 5-6). For 14 *KIRs*, ctyper reached 98.5% of
248 predicting full fields against the full-set and 70.6% among leave-one-out, while T1K had
249 32.0% due to the limited database. For protein-coding products (first three digits), ctyper
250 reached 99.2% against the full-set (with 99.9% copy number F1-correctness) and 88.8%
251 among leave-one-out (with 99.2% copy number F1-correctness), while T1K had 79.6%
252 (Supplementary Figure 3). Benchmarking *CYP2D6* star annotations based on
253 assemblies[51], ctyper reached 100.0% against the full-set and 83.2% among leave-one-
254 out, compared to 80.0% using Aldy (Fig. 3h). There was perfect agreement of SNP
255 variants ctyper against the full-set and 95.7% among leave-one-out, compared to 85.2%
256 using Aldy.

257

258 We compared genotyping F1-scores on *HLA* to a contemporary method also
259 based on pangenomes called Locityper[46]. Locityper performs slightly better at predicting
260 all four fields, including noncoding variants (leave-one-out: Locityper 87.9% vs. ctyper
261 86.3%; full-set: Locityper 99.5% vs. ctyper 98.9%), while ctyper is slightly better on the
262 first two fields of protein-coding products (leave-one-out: Locityper 94.0% vs. ctyper
263 96.5%; full-set: Locityper 99.5% vs. ctyper 99.98%). If only run on 28 *HLA* genes, using
264 one thread, ctyper takes 35 seconds (1.1s for genotyping) on genotyping all *HLA*s. As a
265 comparison, the Locityper genotyping of *HLA* genes required about 8.5 minutes using 8
266 threads (4 minutes spent genotyping). Separating out the genotyping runtime from read
267 preprocessing (which may be amortized across multiple loci), ctyper gives a 218×
268 speedup in genotyping.
269
270 Finally, we used ctyper to genotype and repeat the previous assembly
271 comparisons on 273 challenging medically relevant (CMR) genes[52], of which 212 are not
272 reported as CNV. Unrepetitive (unmasked) regions had on average 0.29 mismatches per
273 10 kb against the full-set; these mismatches are 99.7% less than the total variants when
274 aligning the assembly to corresponding GRCh38 sequences (reference-based
275 divergence). The genotypes using leave-one-out databases had 4.9 mismatches per 10
276 kb, with 94.8% fewer than reference variants (Supplementary Figures 4-6). Including
277 repeat-masked low-complexity sequences (e.g., variable-number tandem repeats), there
278 were 10.5 mismatches per 10 kb against the full-set, which is 97.6% fewer than reference-
279 based divergence, and 74.7 mismatches per 10 kb among leave-one-out, which is 82.7%
280 fewer than reference-based divergence (Supplementary Figures 7-9). Compared with

281 Locityper, ctyper works on more CMR genes (273 vs. 259) and has a lower average
282 mismatch rate in the full-set (10.5 vs. 13.2 mismatches per 10 kb), but has a higher
283 average mismatch rate in the leave-one-out analysis (74.7 vs. 45.4 mismatches per 10
284 kb). When limiting comparisons to genomic regions in common with Locityper, ctyper has
285 improved performance (3.0 mismatches in the full-set and 22.3 mismatches in the leave-
286 one-out analysis per 10 kb; Supplementary Table 7 and Supplementary Figure 10).
287

288 **Sequence level diversity of CNVs in global populations**

289 We used principal component analysis (PCA) to examine the population structure
290 of PA genotypes on 2,504 unrelated 1kGP samples, 879 Genotype-Tissue Expression
291 (GTEx) samples, and 105 diploid assemblies (excluding HGSVC due to lower coverage
292 after filtration) (Fig. 4a,b). Following standard population analysis, rare subgroups (<0.05
293 allele frequency) were excluded, and copy numbers were limited to 10 to balance the
294 weights of principal components (PCs). The 1kGP, GTEx, and genome assemblies were
295 clustered by population as opposed to the data source, suggesting little bias between
296 genotyping and assembly, or across NGS cohorts.
297

298 The top 0.1% highest weighted subgroups on PC1 have an average aggreCN
299 variance of 26.33, significantly larger compared to an overall of 4.00 (p-value=1.11e-16,
300 F-test). Similarly, PC2 and PC3 have mean aggreCN variance of 19.73 and 7.20,
301 suggesting CNVs are weakly associated with sequence variants. Furthermore, PC1 is the
302 only PC that clustered all samples into the same sign with a geographic center away from
303 0, suggesting it corresponds to modulus variance (hence aggreCN) if treating samples as

304 vectors of paCNs. Meanwhile, PC2 and PC3 were similar to the PCA plots based on SNP
305 data on global samples[53], suggesting they are associated with the sequence diversity of
306 CNV genes. The total number of duplications is elevated in African populations (Fig. 4c),
307 reflected in the order of PC1 (Fig 4a).
308
309 We examined ctyper genotypes to measure the extent to which duplications show
310 population specificity. We used the F-statistic, a generalization of the F_{st} that
311 accommodates more than two genotypes (Methods), to test the differences in
312 distributions across five continental populations (Fig. 4d). In total, 4.4% (223/5,065) of
313 duplicated subgroups showed population specificity (F-statistic > 0.2; Supplementary
314 Table 8). The subgroups of PAs with the highest F-statistic (0.48) contain duplications of
315 *HERC2P9* which is known to have population differentiation[9,54]. Another example is a
316 converted copy of *SMN2* annotated as a duplication of *SMN1* that is enriched in African
317 populations (F-statistic=0.43).
318
319 We then measured whether duplicated genes were similar or diverged from
320 reference copies, indicating recent or ancient duplications and providing a measure of
321 reference bias from missing paralogs. We constructed multiple sequence alignments
322 (MSAs; Methods) for sequences of each matrix and measured the pairwise differences in
323 non-repetitive sequences for each pair of PAs. We determined the average paralog
324 sequence divergence relative to ortholog divergence (Methods), which we refer to as
325 relative paralog divergence (RPD). We also measured copy number diversity using the
326 mean absolute error (MAE) of genes in the populations, indicating the CNV level among

327 populations (Fig. 4e). Based on RPD, using Density-Based Spatial Clustering of
328 Applications with Noise[55], we identified two peaks at 0.71 and 3.2, with MAE centers at
329 0.18 and 0.93. The first peak indicates genes with rare and recent CNVs, while the second
330 peak indicates more divergent and common CNVs, often CNVs that may be inherited as
331 different structural haplotypes that cannot be analyzed using a single reference genome.
332 For example, *AMY1A* has a high RPD at 3.10 because of the truncated duplications of
333 *AMY1A* (blue gene annotations in Fig. 2a). If we consider the sequence variation indicates
334 duplication age, these results are consistent with ancient bursts of duplications in humans
335 and primate ancestors[56].
336
337 We next studied the haplotype linkage of PAs to investigate the levels of
338 recombination at different CNV loci. We determined multiallelic linkage disequilibrium[57]
339 (mLD; Methods) between PAs using the unrelated 1kGP samples for 989 subgroups that
340 were adjacent and less than 100 kb apart on GRCh38 (Fig. 4f), and reported the average
341 mLD within each matrix. There was a strong negative rank correlation between MAE of
342 copy number and mLD ($\boldsymbol{\rho}$=-0.24, p-value=3.4e-15, Spearman's rank), which was stronger
343 than the rank correlation between mLDs and total locus length ($\boldsymbol{\rho}$=-0.21, p-value = 1.5e-
344 11, Spearman's rank), suggesting a reduced haplotype linkage on genes with frequent
345 CNVs. The lowest mLD=0.013 was found on *FAM90*, a gene with frequent duplications
346 and rearrangements[58]. Not surprisingly, the 29 loci with highest mLD (mLD > 0.7) are
347 enriched in the sex chromosomes (N=19). Furthermore, *HLA-B* and *HLA-DRB* had
348 mLD >0.7 and only deletion CNV (Supplementary Methods). In agreement with our prior

349 diversity analysis based on long-read data, the amylase locus has a value of 0.293 due
350 to recombination.
351
352 **Expression quantitative trait loci (eQTLs) on pangenome-derived alleles**
353 To investigate the impact of paCNs on expression, we performed eQTL analysis
354 using the Genetic European Variation in Disease[59] (GEUVADIS) and GTEx[60] cohorts.
355 There were 4,512 genes that could be uniquely mapped in RNA-seq alignments. An
356 additional 44 genes, such as *SMN1/2* and *AMY1A/1B/1C*, were analyzed by pooling
357 among all copies because they have indistinguishable transcription products. We
358 assigned PAs to these transcripts based on exonic sequences (Methods; Supplementary
359 Table 9).
360
361 We corrected expression bias using PEER[61] with the first three PCs from SNP
362 genotypes[62] and performed association analyses with paCNs. After merging paCNs to
363 aggreCNs, 5.5% (178/3,224) of transcripts showed significance (corrected P = 1.6e-0.5,
364 Pearson correlation) as previously observed[37]. We then tested whether using paCNs
365 would provide a stronger fit by updating the aggreCNs with individual paCNs and
366 performing multivariable linear regression on expression (Methods). There were
367 significant improvements in fit for 27.6% (890/3,224) of transcripts (corrected P = 1.6e-
368 05, one-tailed F-test; Fig. 5a).
369
370 The improved fit could be explained by the nonuniform expression of different
371 alleles of the same reference gene. To test this, we used a linear mixed model (LMM;

372 Methods)[63,64] to regress total expression to individual subgroups and estimate allele-
373 specific expression, then compared these values to other subgroups of the same matrix
374 that were assigned to the same reference gene (Supplementary Table 10). For subgroups
375 within solvable matrices with >10 samples, we found that 7.94% (150/1,890) of paralogs
376 and 3.28% (546/16,628) of orthologs had significantly different expression levels
377 (corrected with sample size = number of paralogs + orthologs, corrected P = 2.7e-06, chi-
378 squared test, Fig. 5b). Overall, paralogs were found to have reduced expression (Fig. 5c),
379 consistent with previous findings on duplicated genes[65].
380
381 The granularity of PA genotypes enabled testing for preferential expression of
382 paralogs among different tissues. After correcting raw transcript per million values using
383 DESeq2[66], we compared expression in 57 tissues in the GTEx samples using LMMs to
384 estimate the expression levels in each tissue (Methods; Supplementary Table 11). There
385 was alternative tissue specificity for 132 of 2,820 paralogs (4.68%) and 225 of 19,197
386 orthologs (1.17%) (corrected P = 6.4e-08, union of two chi-squared tests; Methods; Fig.
387 5d).
388
389 Additionally, we used analysis of variance (ANOVA) to estimate the proportion of
390 expression variance (R^2) explained by paCNs using PEER-corrected GEUVADIS, and
391 compared it to a model based on known SNPs, indels, and eQTL structural variants[67]
392 (Methods). As expected, the highly granular paCNs explained the most variance: on
393 average, 10.3% (14.3% including baseline). In contrast, 58.0% of transcripts are genes
394 with known eQTL variants that explained valid variance by 2.14% (1.60% considering

395 experimental noise, in agreement with a previous estimate of 1.97%[68]). On average,
396 1.98% of the variance was explained by aggreCNs, and 8.58% by subgroup information.
397 When combining both paCNs and known eQTL sites, 10.4% (19.0% including baseline)
398 of the valid variance was explained (Fig. 5e).
399
400 We examined the *SMN* and *AMY2B* genes as case studies due to their importance
401 in disease and evolution[33,69]. The *SMN* genes were classified into three categories:
402 *SMN1*, *SMN2*, and *SMN-converted*. We estimated the total expression of all transcripts
403 and the expression of only isoforms with valid exon 7 splicing junctions. For total
404 expression, no significant difference was found between *SMN1* and *SMN2* (0.281 ± 0.008
405 vs. 0.309 ± 0.009; P = 0.078, chi-squared test). However, significant differences were
406 found between *SMN-converted* and *SMN1/2* (0.226 ± 0.012 vs. 0.294 ± 0.002; P = 1.75e-
407 07, chi-squared test), with a 23.0% reduction in expression of *SMN-converted*. In contrast,
408 despite having lower overall expression, *SMN-converted* had 5.93× the expression of
409 *SMN2* (P = 2.2e-16, chi-squared test) for valid exon 7 splicing[70], indicating that while
410 *SMN-converted* has full functional splicing[71], its overall expression level is lower (Fig. 5f).
411
412 We studied the expression of *AMY2B* duplications when they are translocated
413 proximally to other *AMY* genes, such as the PAs containing *AMY1* and *AMY2B* in Fig. 2a.
414 Using PEER-corrected GTEx pancreas data, we estimated their expression as well as for
415 other duplications. We found that these translocated *AMY2B* genes had significantly
416 higher expression than other duplications (1.384 ± 0.233 vs. -0.275 ± 0.183, P = 7.87e-
417 09, chi-squared test) (Fig. 5g).

418

419 # Discussion

420 New pangenomes present both opportunities and challenges for the study of
421 complex genetic variation (e.g., CNVs, recurrent SVs, translocations and gene
422 conversions): while they reveal the landscape of complex variation, such variation
423 requires new models for representation and analysis. We represent genomic variation as
424 PAs: haplotype segments that capture genomic structural information and phased
425 variation. To support large NGS cohort analyses, we developed an alignment-free
426 genotyping tool, ctyper, to genotype PAs among NGS data, providing allele-specific
427 sequence information as well as copy number state. The genotyping approach is based
428 on a new mathematical model that relaxes an NP-hard problem into a more efficient
429 polynomial semi-analytic solution while maintaining accuracy, with increased robustness
430 and copy number sensitivity. It is also important to note that although the analysis in this
431 study focuses on CNV genes, ctyper can also genotype non-CNV genes and provide a
432 calling for complex genetic variation and local phasing.

433

434 The use of ctyper genotypes increases the scope of NGS studies to profile
435 variation in challenging, medically relevant and CNV genes that are otherwise
436 unmappable. For example, our finding that CNVs reflect two modes of variation– highly
437 similar (and likely recent) and low-identity (ancient and polymorphic) duplications– is
438 based on the 1kGP genotypes rather than assembly annotation. As another example, the
439 ctyper genotypes yield tissue-specific expression of paralogs as well as relative
440 contributions to the expression of different forms of duplications, such as *SMN*.

441

442 We investigated the significant improvement of ANOVA on PAs, whose genotypes
443 are multiallelic and reflect different combinations of variants in contrast to known biallelic
444 eQTL variants from conventional analysis. First, compared to PAs, there were either very
445 few or very many eQTL variants per gene, indicating LD (Supplementary Figure 11) as
446 addressed by fine-mapping[72], and increasing multiple testing burden[73]. Indirect
447 association due to LD also explains why there was a greater proportion of variance
448 explained among genes with more CNVs by conventional eQTL variants (for example,
449 the *HPR* genes, Supplementary Figure 2). However, as the frequency of CNVs increases,
450 the explained variance by eQTL variants increases (t= 3.80, p-value = 1.6e-04, Pearson's
451 correlation), and the number of eQTL variants decreases (t = -4.79, p-value = 2.1e-06,
452 Pearson's correlation), suggesting that larger effects like CNVs might overshadow the
453 discovery of other eQTL variants not in LD (the increase of total variance reduces
454 significance in association analysis using a Gaussian-like model). Because PAs already
455 incorporate LD information, they will suffer less from such LD-based problems.
456 Furthermore, gene expression might not be a linear additive effect of all variants[74]. For
457 example, although *SMN*-converted contains variants that are either from *SMN1* or *SMN2*,
458 its overall expression is lower than both. In this manner, using a genetic model with linked
459 variants such as PAs would improve upon the linear additive model in predicting gene
460 expression. Because those limitations also apply to non-CNV genes, the concept of PAs
461 may have a wider potential for future genome-wide association analysis.

462

Our analysis has some limitations. Due to the limited sample size, our associations are based on subgroups rather than individual PAs. Different cohort sizes may require different levels of granularity when defining subgroups. Our current classification of subgroups was designed for biobank cohorts; association tests in smaller cohorts may need to aggregate more PAs into subgroups. For example, the three subtypes of *SMN-converted* showed little difference in our eQTL analysis, so they were merged in our case study, but larger cohort studies may find their differences. The granularity of genotyping is additionally defined by the length of PA sequences; genotypes using shorter PAs will more accurately reflect small variants, while longer sequences can preserve more structural information and may be preferable in regions with low recombination, such as *HLA-DRB*.

Ctyper has limitations. First, while it is possible to detect CNVs smaller than PA units using ctyper (Supplementary Methods), full support requires additional benchmarking data. Additionally, ctyper currently does not provide confidence values for genotypes.

Data availability

Software: https://github.com/ChaissonLab/Ctyper.

Allele database and annotations: https://doi.org/10.5281/zenodo.13381931.

Benchmarking and analysis code: https://github.com/Walfred-MA/CNVAnalyze.

Acknowledgments

This work was supported by NHGRI R01HG011649 and NHGRI U01HG010973. W.M. conceived the method, performed the analysis, and wrote the manuscript. M.J.P.C. conceived the method and wrote the manuscript. We thank Dr. Matthew Pennell for the constrictive critiques of our manuscript.

521 1. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy
522 genes. *Science* **330**, 641–646 (2010).
523 2. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature*
524 **444**, 444–454 (2006).
525 3. Macé, A. *et al.* CNV-association meta-analysis in 191,161 European adults reveals
526 new loci associated with anthropometric traits. *Nat. Commun.* **8**, 744 (2017).
527 4. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature*
528 **606**, 984–991 (2022).
529 5. Soemedi, R. *et al.* Contribution of global rare copy-number variants to the risk of
530 sporadic congenital heart disease. *Am. J. Hum. Genet.* **91**, 489–501 (2012).
531 6. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a
532 genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).
533 7. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism
534 spectrum disorders. *Nature* **466**, 368–372 (2010).
535 8. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution,
536 diversity and disease. *Nat. Rev. Genet.* **7**, 552–564 (2006).
537 9. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human
538 genomes. *Nature* **526**, 75–81 (2015).
539 10. Collins, R. L. *et al.* A structural variation reference for medical and population
540 genetics. *Nature* **581**, 444–451 (2020).
541 11. Searles Quick, V. B., Davis, J. M., Olincy, A. & Sikela, J. M. DUF1220 copy number
542 is associated with schizophrenia risk and severity: implications for understanding
543 autism and schizophrenia as related diseases. *Transl. Psychiatry* **5**, e697 (2015).

544 12. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human
545 genome. *Science* **376**, eabj6965 (2022).
546 13. Pollen, A. A., Kilik, U., Lowe, C. B. & Camp, J. G. Human-specific genetics: new
547 tools to explore the molecular and cellular basis of human evolution. *Nat. Rev.*
548 *Genet.* **24**, 687–711 (2023).
549 14. Fiddes, I. T. *et al.* Human-specific NOTCH2NL genes affect Notch signaling and
550 cortical neurogenesis. *Cell* **173**, 1356–1369.e22 (2018).
551 15. Fossati, M. *et al.* SRGAP2 and its human-specific paralog co-regulate the
552 development of excitatory and inhibitory synapses. *Neuron* **91**, 356–369 (2016).
553 16. Mangan, R. J. *et al.* Adaptive sequence divergence forged new
554 neurodevelopmental enhancers in humans. *Cell* **185**, 4587–4603.e23 (2022).
555 17. Vollger, M. R. *et al.* Increased mutation and gene conversion within human
556 segmental duplications. *Nature* **617**, 325–334 (2023).
557 18. Blair, I. P., Nash, J., Gordon, M. J. & Nicholson, G. A. Prevalence and origin of de
558 novo duplications in Charcot-Marie-Tooth disease type 1A: first report of a de novo
559 duplication with a maternal origin. *Am. J. Hum. Genet.* **58**, 472–476 (1996).
560 19. Payen, C., Koszul, R., Dujon, B. & Fischer, G. Segmental duplications arise from
561 Pol32-dependent repair of broken forks through two alternative replication-based
562 mechanisms. *PLoS Genet.* **4**, e1000175 (2008).
563 20. Hallast, P., Nagirnaja, L., Margus, T. & Laan, M. Segmental duplications and gene
564 conversion: Human luteinizing hormone/chorionic gonadotropin beta gene cluster.
565 *Genome Res.* **15**, 1535–1546 (2005).
566 21. Hujoel, M. L. A. *et al.* Protein-altering variants at copy number-variable regions

567 influence diverse human phenotypes. *Nat. Genet.* **56**, 569–578 (2024).
568 22. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat.*
569 *Rev. Genet.* **21**, 171–189 (2020).
570 23. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing:
571 computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).
572 24. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324
573 (2023).
574 25. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated
575 analysis of structural variation. *Science* **372**, (2021).
576 26. Gao, Y. *et al.* A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–
577 121 (2023).
578 27. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic
579 diversity. *Nature* **604**, 437–446 (2022).
580 28. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**,
581 139–162 (2020).
582 29. Browning, S. R. & Browning, B. L. Identity by descent between distant relatives:
583 detection and applications. *Annu Rev Genet* **46**, 617–633 (2012).
584 30. Chen, C.-H. *et al.* Determinants of transcription factor regulatory range. *Nat.*
585 *Commun.* **11**, 2472 (2020).
586 31. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the
587 human genome. *Nat. Rev. Genet.* **4**, 587–597 (2003).
588 32. Ogino, S., Gao, S., Leonard, D. G. B., Paessler, M. & Wilson, R. B. Inverse
589 correlation between SMN1 and SMN2 copy numbers: evidence for gene conversion

590 from SMN2 to SMN1. *Eur. J. Hum. Genet.* **11**, 275–277 (2003).
591 33. Lorson, C. L., Hahnen, E., Androphy, E. J. & Wirth, B. A single nucleotide in the
592 SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc.*
593 *Natl. Acad. Sci. U. S. A.* **96**, 6307–6311 (1999).
594 34. He, Y. *et al.* T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference
595 Genome for Han Chinese. *Genomics Proteomics Bioinformatics* **21**, 1085–1100
596 (2023).
597 35. Yang, C. *et al.* The complete and fully-phased diploid genome of a male Han
598 Chinese. *Cell Res.* **33**, 745–761 (2023).
599 36. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53
600 (2022).
601 37. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat.*
602 *Genet.* **47**, 296–303 (2015).
603 38. Prodanov, T. & Bansal, V. Robust and accurate estimation of paralog-specific copy
604 number for duplicated genes using whole-genome sequencing. *Nat. Commun.* **13**,
605 3221 (2022).
606 39. McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nat.*
607 *Genet.* **38**, 86–92 (2006).
608 40. Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural
609 haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44**,
610 881–885 (2012).
611 41. Yilmaz, F. *et al.* Paleolithic Gene Duplications Primed Adaptive Evolution of Human
612 Amylase Locus Upon Agriculture. *bioRxiv* (2024) doi:10.1101/2023.11.27.568916.

613 42. Ebler, J. *et al.* Pangenome-based genome inference allows efficient and accurate
614 genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525
615 (2022).
616 43. Creators Wangfei MA1 Show affiliations 1. UCSF. *WalfredMA/NUI_Projection:*
617 *Primary Release for NUI_projection*. doi:10.5281/zenodo.3496215.
618 44. Wong, K. H. Y. *et al.* Towards a reference genome that captures global genetic
619 diversity. *Nat. Commun.* **11**, 5482 (2020).
620 45. Myers, E. W. & Miller, W. Approximate matching of regular expressions. *Bull. Math.*
621 *Biol.* **51**, 5–37 (1989).
622 46. Prodanov, T. *et al.* Locityper: targeted genotyping of complex polymorphic genes.
623 *bioRxiv* (2024) doi:10.1101/2024.05.03.592358.
624 47. Song, L., Bai, G., Liu, X. S., Li, B. & Li, H. T1K: efficient and accurate KIR and HLA
625 genotyping with next-generation sequencing data. (2022)
626 doi:10.1101/2022.10.26.513955.
627 48. Numanagić, I. *et al.* Allelic decomposition and exact genotyping of highly
628 polymorphic and structurally variant genes. *Nat. Commun.* **9**, 828 (2018).
629 49. Robinson, J. *et al.* IPD-IMGT/HLA Database. *Nucleic Acids Res.* **48**, D948–D955
630 (2020).
631 50. Lefranc, M. P. IMGT, the international ImMunoGeneTics database. *Nucleic Acids*
632 *Res.* **29**, 207–209 (2001).
633 51. PacificBiosciences. *Pangu, https://github.com/PacificBiosciences/pangu*. (Github).
634 52. Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant
635 autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).

636 53. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142
637 diverse populations. *Nature* **538**, 201–206 (2016).
638 54. Saitou, M. & Gokcumen, O. Resolving the Insertion Sites of Polymorphic
639 Duplications Reveals a HERC2 Haplotype under Selection. *Genome Biol. Evol.* **11**,
640 1679–1690 (2019).
641 55. Ester, M., Kriegel, H., Sander, J. & Xu, X. A density-based algorithm for discovering
642 clusters in large spatial databases with noise. *KDD* 226–231 (1996).
643 56. Dennis, M. Y. *et al.* The evolution and population diversity of human-specific
644 segmental duplications. *Nat Ecol Evol* **1**, 69 (2017).
645 57. Okada, Y. eLD: entropy-based linkage disequilibrium index between multiallelic
646 sites. *Hum Genome Var* **5**, 29 (2018).
647 58. Bosch, N. *et al.* Characterization and evolution of the novel gene family FAM90A in
648 primates originated by multiple duplication and rearrangement events. *Hum. Mol.*
649 *Genet.* **16**, 2572–2582 (2007).
650 59. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional
651 variation in humans. *Nature* **501**, 506–511 (2013).
652 60. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across
653 human tissues. *Science* **369**, 1318–1330 (2020).
654 61. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation
655 of expression residuals (PEER) to obtain increased power and interpretability of
656 gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
657 62. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the
658 regulatory effect size of cis-acting genetic variation using allelic fold change.

659 *Genome Res.* **27**, 1872–1884 (2017).
660 63. Henderson, C. R. Estimation of Variance and Covariance Components. *Biometrics*
661 **9**, 226–252 (1953).
662 64. Fox, J. & Weisberg, S. Mixed-effects models in R. *An R Companion to Applied*
663 *Regression; SAGE: Thousand Oaks, CA, USA* (2002).
664 65. Lan, X. & Pritchard, J. K. Coregulation of tandem duplicate genes slows evolution
665 of subfunctionalization in mammals. *Science* **352**, 1009–1013 (2016).
666 66. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and
667 dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
668 67. Keys, K. L. *et al.* On the cross-population generalizability of gene expression
669 prediction models. *PLoS Genet.* **16**, e1008927 (2020).
670 68. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using
671 reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
672 69. Pajic, P. *et al.* Independent amylase gene copy number bursts correlate with
673 dietary preferences in mammals. *Elife* **8**, (2019).
674 70. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–
675 21 (2013).
676 71. Lorson, C. L., Rindt, H. & Shababi, M. Spinal muscular atrophy: mechanisms and
677 therapeutic strategies. *Hum. Mol. Genet.* **19**, R111–R118 (2010).
678 72. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B. & Eskin, E. Identification of
679 causal genes for complex traits. *Bioinformatics* **31**, i206–13 (2015).
680 73. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat.*
681 *Rev. Genet.* **20**, 467–484 (2019).

682 74. Li, P., Guo, M., Wang, C., Liu, X. & Zou, Q. An overview of SNP interactions in
683 genome-wide association studies. *Brief. Funct. Genomics* **14**, 143–155 (2015).

684

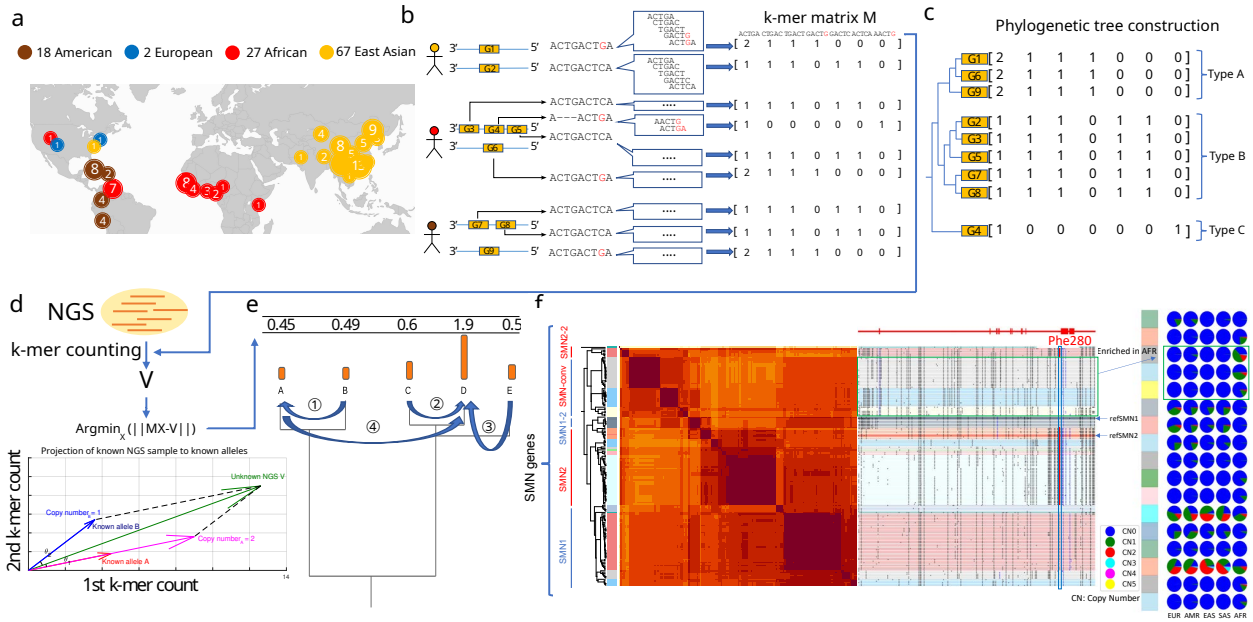

Figure 1. a, Demography of the reference pangenome assemblies. **b,** Construction of pangenome *k*-mer matrices for CNV genes. Each individual gene is represented as a vector of counts of *k*-mers exclusively found among homologous sequences. All similar sequences including paralogs and orthologs are included and integrated as a *k*-mer matrix. **c,** Construction of phylogenetic trees based on *k*-mer matrices. **d,** Schematic of approach to estimate genotypes of alleles using NGS data. The *k*-mers from each matrix are counted in NGS data and normalized by sequencing depth. The normalized *k*-mer counts are projected to all pangenome genes. **e,** Reprojection of the raw results in the last step to integer solutions recursively based on the phylogenetic tree. **f,** An illustrative annotation and genotyping results on *SMN1/2* genes using HPRC samples. All *SMN* genes are categorized into five major types and 17 subgroups. *SMN1*/*SMN2* correspond to the most common types of each paralog; *SMN1-2*: *SMN1* partially converted to *SMN2*; *SMN*-conv: additional converted SMN genes *SMN2-2*: a rare outgroup of *SMN2*. *SMN*-conv is predominantly mapped to the *SMN2* locus and is found to be enriched in African populations. The GRCh38 assembly includes *SMN1-2* and *SMN2*. On the right side of the classification, the phylogenetic tree and heatmap of pairwise similarities are shown along with a mutant plot based on an MSA highlighting point differences to *SMN1* in CHM13. Phe-280, the variant found to disrupt the splicing of *SMN2* transcripts, is highlighted in red. The genotyping results of 1kGP continental populations are shown on the right. Rows correspond to subgroups, columns correspond to continental populations, and the colors of pie charts give distributions of copy numbers among each continental population.

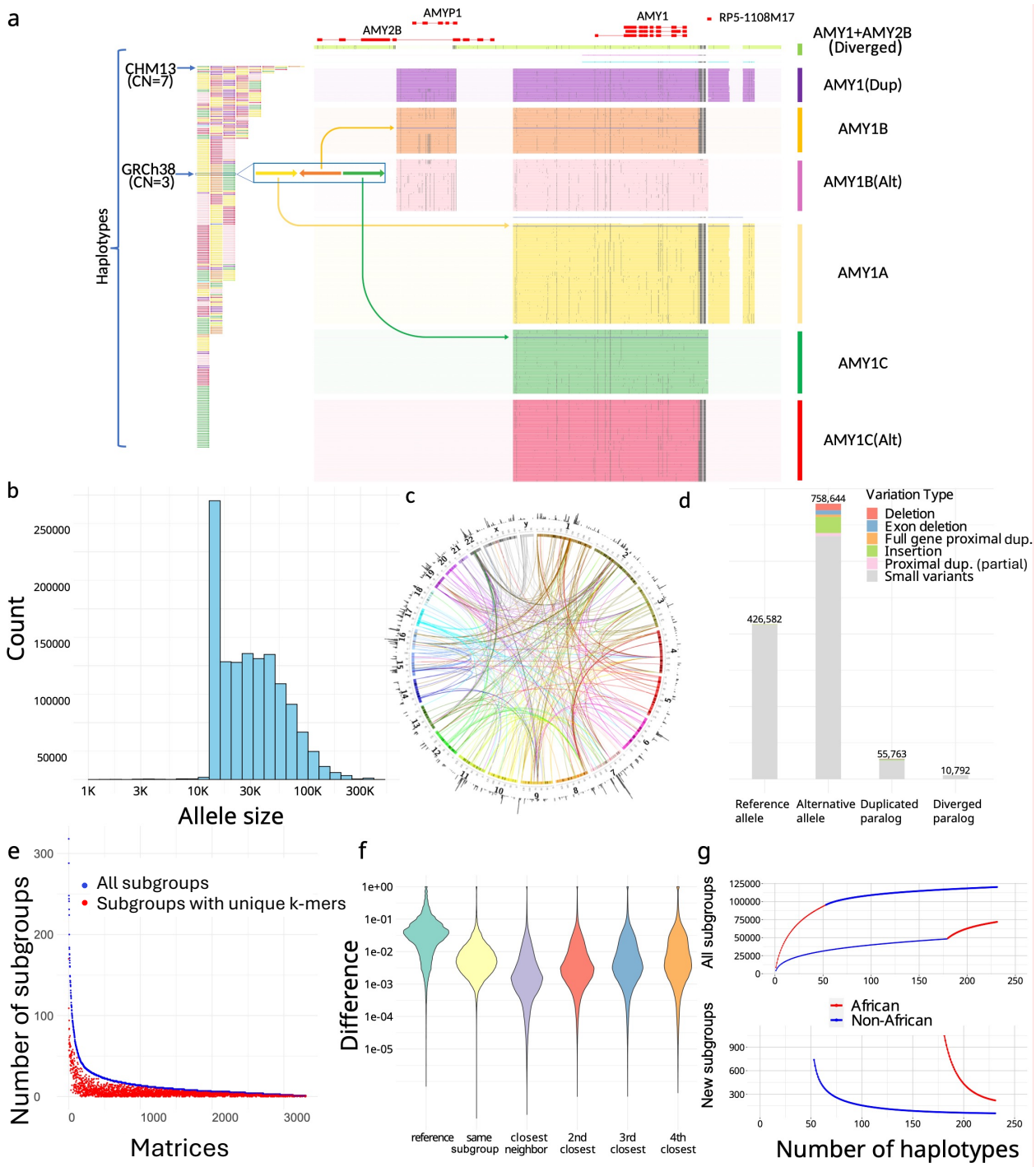

Figure 2. **a,** An overview of amylase 1 pangenome-derived alleles (PAs). (left) The corresponding order of all *AMY1* PAs on assemblies, which are colored based on their major groups (unions of subgroups devoid of SVs >300 bp between members). (right) *AMY1* genes are extracted as PAs as well as their flanking genes and sequences, including an *AMY2B* translocated proximally to *AMY1*, and two pseudogenes: *AMYP1* and *RP5-1108M17*. All PAs are vertically ordered according to the phylogenetic tree and aligned via graphic MSAs (Supplementary Methods). Homologous sequences are

vertically aligned. Mutations are visualized as dots, and large gaps (deletions) are visualized as spaces. Seven major groups are categorized including five paralogs and two orthologs. There are no pseudogenes around *AMY1C*, while *AMY1A* has *RP5-1108M17* nearby and *AMY1B* has *AMYP1* nearby. There are alternative versions of *AMY1B and AMY1C*, with sequence substitutions. A new paralog called *AMY1(Dup)* is found primarily on haplotypes with duplications and has both pseudogenes nearby. Another paralog of *AMY1* found with translocated *AMY2B* is called *AMY1+AMY2B*. There are also two rare paralogs (blue and violet) and one singleton ortholog (steel-blue). **b,** The size distribution of PAs on a log density. The minimum size of PAs is 15 kb, though smaller alleles may be annotated on alternative haplotypes on GRCh38 and as partial loci when dividing large genes into alleles without recombination. **c**. CIRCOS plot of all PAs. (outer ring) The density of PAs in each Mb on GRCh38. (arcs) Interchromosomal PAs are included in the same group. **d,** Annotation of PAs according to orthology and variants with respect to GRCh38. **e,** Identifiability of highly similar subgroups by unique *k*-mers. The total number of subgroups (blue) and the number of subgroups that may be identified by paralog-specific *k*-mers (red) are shown for each matrix with a size of at least three. **f,** The distribution of logistic pairwise divergence of PAs depending on orthology and phylogenetic relationship. The values shown are average values from each of the matrices. Small neighbor distances are an indicator of the strong representativeness of the current cohort. **g,** Saturation analysis for all subgroups using a recapture mode according to two sorted orders: African genomes considered first, and non-African genomes considered first. The former order has a smoother curve than the latter order, indicating there are more African-specific subgroups.

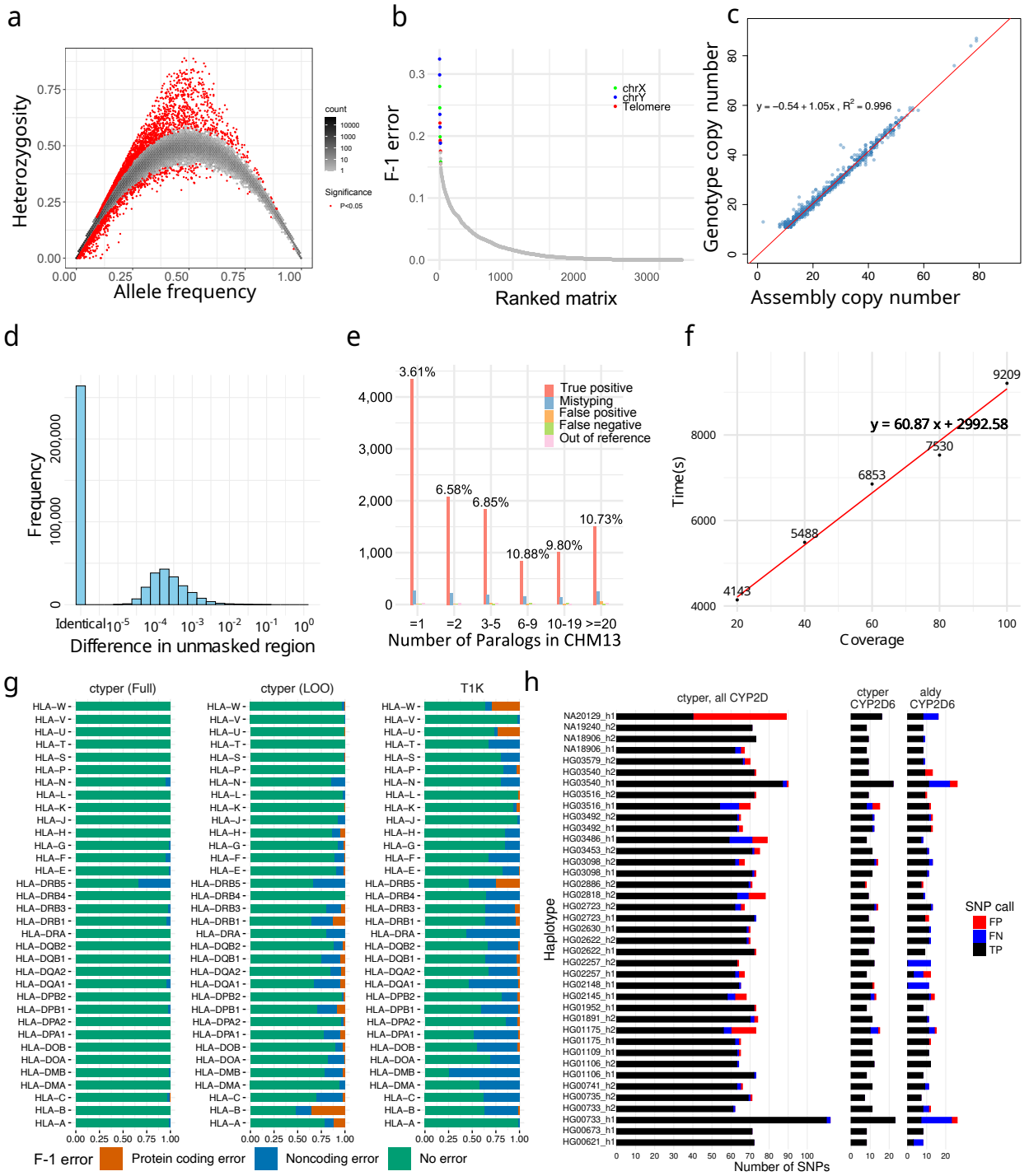

Figure 3. a, Hardy-Weinberg equilibrium of genotyping results on 2,504 1kGP unrelated samples. **b,** Genotype concordance of genotyping results on 641 1kGP trios, ordered by F-1 error. The matrices with an F-1 error of more than 15% were labeled by genomic location. **c,** Copy number comparison between assemblies and genotyping results on 39 HPRC samples shared with 1kGP. **d,** Sequence differences between genotyped and original alleles during leave-one-out test using pairwise alignment of nonrepetitive

sequences. **e,** Detailed leave-one-out comparison in the diploid telomere-to-telomere genome CN1. The results are categorized regarding the number of paralogs in CHM13 to show performances on different levels of genome complexity and the main sources of errors. CN1 NGS sample was about 40 coverage. **f,** Runtime of ctyper on CN1 on all loci for varying coverage. **g,** Benchmarking of *HLA* genotyping using ctyper on full, and leave-one-out (LOO) databases, compared with T1K on 31 *HLA* genes. **h,** Benchmarking of *CYP2D* annotation on all *CYP2D* genes and *CYP2D6* exclusively.

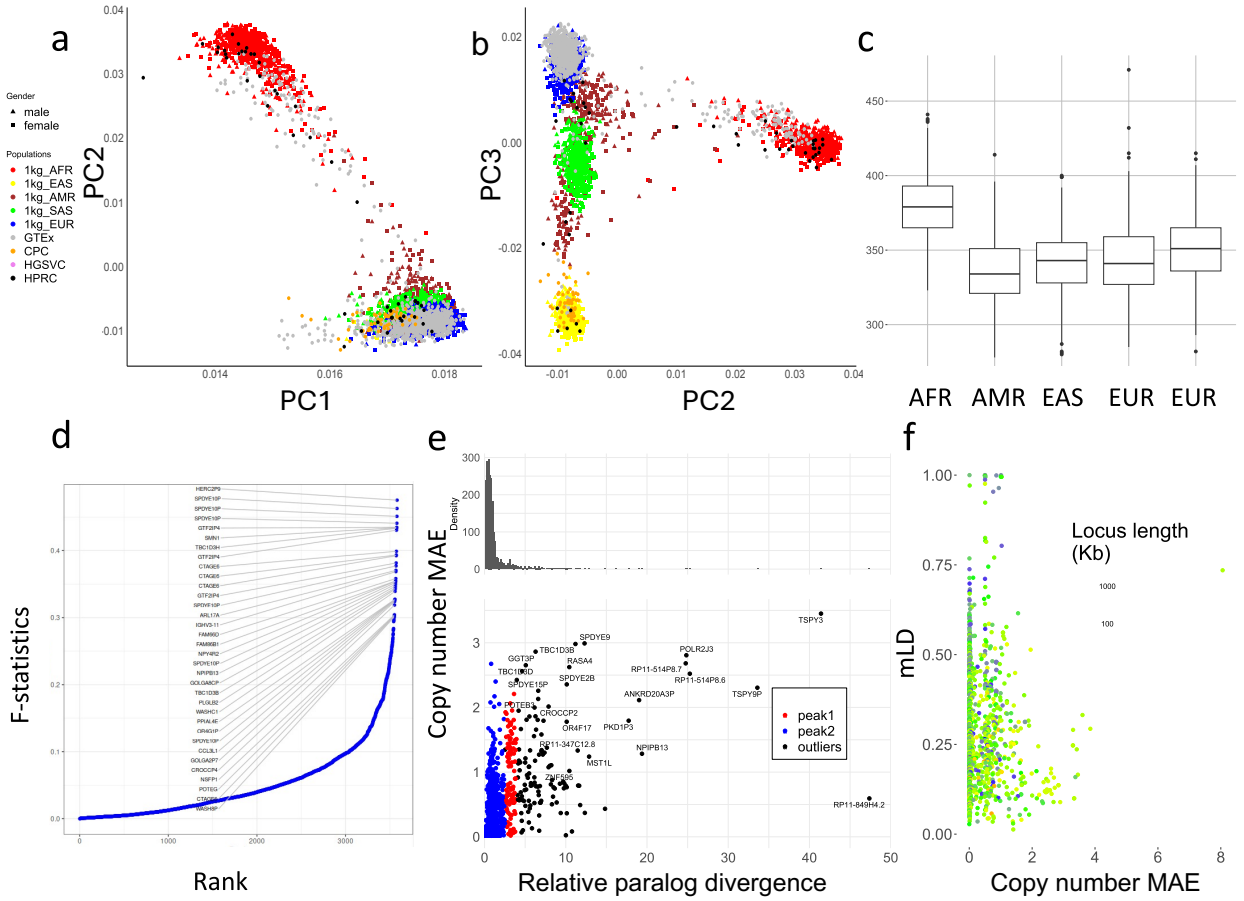

Figure 4.a,b. PCAs of allele-specific copy numbers on the union of PA genotyping results and assembly annotations. **c,** Distribution of total autosomal gene copy numbers among unrelated 1kGP samples, including pseudogenes. **d,** Population differentiation measured by F-statistics of duplications among different continental populations. The genes with a paralogous subgroup with an F-statistic of more than 0.3 are labeled. **e,** Mean absolute variation in copy numbers and relative paralog divergence in sequences. Based on our genotyping results on unrelated 1kGP genomes, for genes found to be CNV to the population median in more than 20 samples, we determined the average aggregate copy number difference (mean absolute error [MAE]) between individuals

and estimated the average paralog difference in sequences relative to ortholog difference. **f,** Multi-allelic linkage disequilibrium (mLD) between pairs of CNV genes less than 100kb apart. The larger MAE value of each pair is used for the x-axis values. The total locus length denotes the length from the beginning of the first gene to the end of the last gene.

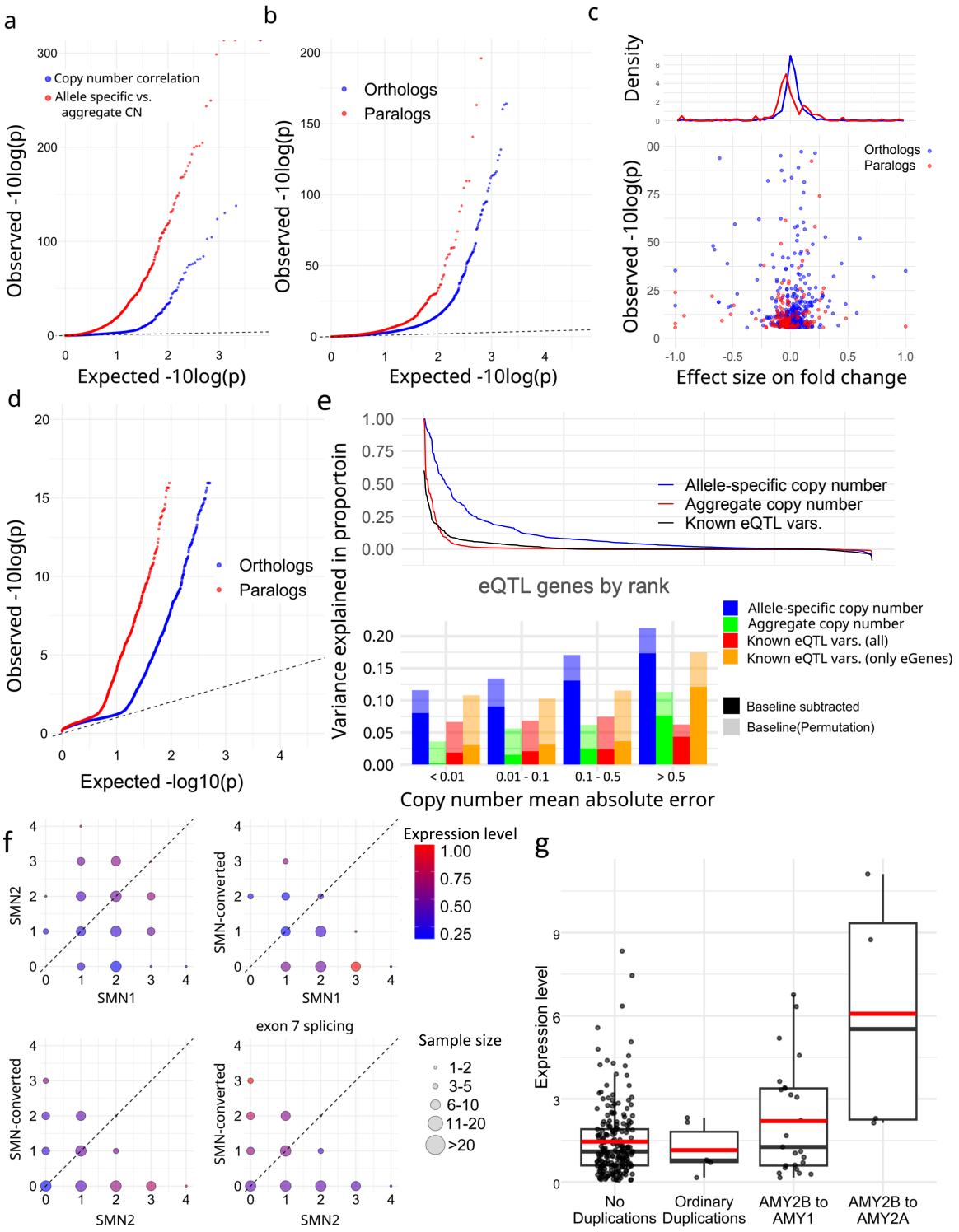

Figure 5. a. Q-Q plot of associations of (blue) aggregate copy numbers to gene expression in GEUVADIS samples (red) and the improvement of allele-specific copy number over aggreCN. **b,** Comparative gene expression Q-Q plots of orthologs (blue) and paralogs (red). **c,** Fold change effect size of all significant alternative expressions in b. The fold changes as well as p-values are shown. **d,** Preferential tissue expression of

orthologs and paralogs. **e,** (top) Comparison of different models on explained expression variance (R^2). (bottom) Quantification of variance explained by different representations at different levels of CNV frequencies: full paCN genotypes, aggregate copy number, and known eQTL variants. **f,** Case study on *SMN* genes showing decreased gene expression on *SMN*-converted. The average expression level in PEER-corrected GEUVADIS samples is shown under different copy numbers of *SMN1*, *SMN2*, and *SMN*-converted. Transcript levels are the total coverage of all isoforms, and the exon 7 splicing level is measured by counting isoforms with a valid exon 7 splicing junction. **g,** Case study on amylase genes showing increased gene expression on translocated *AMY2B* using PEER-corrected GTEx pancreas data.