HGG Advances

ARTICLE

Quality control of large genome datasets

Max Robinson,[1] Arpita Joshi,[1] Ansh Vidyarthi,[1] Mary Maccoun,[1] Sanjay Rangavajjhala,[1] and Gustavo Glusman[1,*]

Summary

The 1000 Genomes Project (TGP) is a foundational resource that serves the biomedical community as a standard reference cohort for human genetic variation. There are now seven public versions of these genomes. The TGP Consortium produced the first by mapping its final data release against human reference sequence GRCh37, then "lifted over" these genomes to the improved reference sequence (GRCh38) when it was released, and remapped the original data to GRCh38 with two similar pipelines. As best-practice quality validation, the pipelines that generated these versions were benchmarked against the Genome In A Bottle Consortium's "platinum quality" genome (NA12878). The New York Genome Center recently released the results of independently resequencing the cohort at greater depth (30×), a phased version informed by the inclusion of related individuals, and independently remapped the original variant calls to GRCh38. We performed a cross-comparison evaluation of all seven versions using genome fingerprinting, which supports ultrafast genome comparison even across reference versions. We noted multiple issues, including discrepancies in cohort membership, disagreement on the overall level of variation, evidence of substandard pipeline performance on specific genomes and in specific regions of the genome, cryptic relationships between individuals, inconsistent phasing, and annotation distortions caused by the history of the reference genome itself. We therefore recommend global quality assessment by rapid genome comparisons, alongside benchmarking as part of best-practice quality assessment of large genome datasets. Our observations also help inform the decision of which version to use, to support analyses by individual researchers.

Introduction

Since its initial release, the 1000 Genomes Project (TGP)[1] has served its intended purpose as the standard reference for human genetic variation in population structure analyses, genotype imputation efforts, association studies, evaluations of gene annotation, and efforts to improve the reference genome itself, among many other applications.[2] Since its release in 2013, most analyses of the TGP cohort have relied on the phase 3 final data release version, which includes estimated variation in 2,504 individuals mapped onto version GRCh37 (hg19) of the human reference genome. These individuals were sampled from 26 populations in five continent-level regions (Africa, East Asia, South Asia, Europe, and the Americas). Genotypes for all individuals were estimated based on a combination of low-coverage whole-genome sequencing (WGS), deep exome sequencing, and high-density single-nucleotide polymorphism (SNP) microarrays. The resulting variant calls included phased biallelic and multiallelic SNPs, indels, and structural variants, with high power to detect variants with alleles of at least 5% frequency (75%–95% depending on population)[1]. For the purpose of comparing variation in the same individuals relative to different reference sequences, we hereafter refer to this phase 3 dataset as TGP37.

Not long after the initial release of TGP37, a new and much improved version[3] of the human reference sequence, GRCh38 (hg38), was released, prompting efforts to express this standard set of variation data relative to this new reference. In 2015, the subset of TGP37 variants with reference SNP ID numbers (rsids)[4] were translated (via liftOver) to GRCh38 coordinates, yielding a version of the variation reference set we will call TGP38L (L for lifted). In 2017, the raw genomic sequence reads were mapped[5] onto GRCh38 to support "native" variant calling.[6] Two versions of these variant calls have been released to date via the International Genome Sample Resource (IGSR).[7] The first, released in late 2018, is restricted to biallelic single-nucleotide variants (SNVs); we call this version TGP38S (S for SNVs). The second, released in early 2019, includes both biallelic SNVs and indels; we refer to this extended dataset as TGP38X (X for extended). Also in early 2019, a new set of high-coverage whole-genome sequences were produced by the New York Genome Center (NYGC) and released via the European Nucleotide Archive (accession: PRJEB31736) and IGSR; we refer to this set as TGP38H (H for high coverage).[8] This new set was later (May 2020) enhanced by the addition of the genomes of 698 related individuals, leading to improved phasing; we refer to this improved set (limited to just the original 2,504 individuals in TGP) as TGP38N (N for new). The NYGC also remapped the original variant calls in TGP37 using CrossMap,[9] yielding the version we call TGP38C (C for crossmap). We anticipate that one of these versions will soon become the new standard reference version of the human genetic variation in the TGP cohort. As such, it is important to evaluate the quality of each of these genome sets, to enable researchers to choose the most appropriate version to support their specific research questions.

Despite the intent to include only unrelated individuals in the TGP cohort,[10] a number of close and more distant

[1]Institute for Systems Biology, 401 Terry Avenue N, Seattle, WA 98109, USA
*Correspondence: gustavo@isbscience.org
https://doi.org/10.1016/j.xhgg.2022.100123.

Table 1. Overview of datasets

TGP version	TGP37	TGP38L	TGP38C	TGP38S	TGP38X	TGP38H	TGP38N
Reference	GRCh37	GRCh38	GRCh38	GRCh38	GRCh38	GRCh38	GRCh38
Individuals	2,504	2,504	2,504	2,503 + 45	2,503 + 45	2,504	2,504
Relateds	31	-	-	150	150	-	698
Data type	ex+WGS	ex+WGS	ex+WGS	ex+WGS	ex+WGS	30× WGS	30× WGS
Method	map, call	liftOver	CrossMap	map, call	map, call	map, call	map, call
File format	VCFv4.1	VCFv4.1	VCFv4.1	VCFv4.3	VCFv4.3	VCFv4.2	VCFv4.1
Size (gigabytes)	15.8	16.2	16.3	12.9	14.2	1241.4	32.3
Autosomes	1-22	1-22	chr1-chr22	chr1-chr22	1-22	chr1-chr22	chr1-chr22
Sex chroms	X, Y	X, Y	chrX, chrY	chrX	X	chrX, chrY	chrX
Mitochondrial	MT	-	-	-	-	chrM	-
SNVs included	all	all	all	biallelic	biallelic	all	all
Indels included	all	all	all	none	biallelic	all	all
Multiallelic	single line	single line	multiline	-	-	single line	multiline
Variants with rsids	48.3%	100%	99.9%	0%	0%	0%	0%
Phased	yes	yes	yes	yes	yes	no	yes

ex+WGS, exome plus low-coverage WGS.

relationships exist within TGP37, as reported by us and others.[11] TGP37 is supplemented with a small set of 31 related individuals, which we call TGP37r (r for related). Likewise, a set of 150 related samples accompanies TGP38S and TGP38X, and we refer to these in turn as TGP38Sr and TGP38Xr. Finally, we refer to the aforementioned 698 genomes accompanying TGP38N as TGP38Nr.

Beyond the method by which they were produced, the seven versions of the TGP dataset differ in several ways (Table 1). The number of individuals included in each differs. All versions use the Variant Call Format (VCF) to represent the data, but they use different versions of VCF; represent chromosomes in different ways, include varying subsets of the sex and mitochondrial chromosomes, and include different combinations of SNVs, indels, and biallelic or multiallelic variants. The number of variants reported for each individual can vary significantly. Two of the versions (TGP38C and TGP38N) represent multiallelic variants in multiple lines, a format that may be unexpected to many researchers and may confuse analysis tools, requiring variant normalization for proper analysis. The various versions also differ in their level of inclusion of variant identifiers (rsids), and whether variants are phased or not. On a very practical level, they also differ substantially in the total size of the dataset to be downloaded for local analysis.

Given the importance of this high-impact reference set of human variation, validating the quality of each release is a crucial concern. TGP37 data were evaluated by comparing variant calls for one individual per population with calls based on 30× PCR-free sequence data, and by comparison with 47× sequence using Complete Genomics technology.[1] TGP38S data quality was evaluated prior to release by comparing variant calls for the NA12878 individual with "truth set" calls from the Genome In A Bottle (GIAB) Consortium.[6,12–14] At least two independent studies of TGP data quality have also been reported. A study of the array genotypes for 2,318 TGP samples found the data to be, in general, of high quality, although some discrepancies were found between reported and inferred sex or in degree of relationship for some samples.[15] More recently, a study found that phasing and imputation for rare variants are unreliable, based on comparison with haplotypes of 28 individuals, experimentally phased using a different sequencing technology.[16]

The availability of multiple versions of the same dataset enables a different type of quality control (QC): cross-comparison among the versions, which supports identification of version-specific results and data processing failures throughout the cohort, the majority of which is not assessed by benchmarking of results against a small subset of high-quality genomes. A cross-comparison between phase 1 and phase 3 data identified nearly 2 million SNPs present in the former but absent in the latter, which could be explained by differences in the samples included, the sequence data type, the variant-calling pipelines, and the reference used.[17] Cross-comparison was also used to evaluate SNV, indel, and structural variant (SV) calls in TGP38N against the original TGP phase 3 callset, which required lifting over variants from the GRCh37 reference (TGP37) to GRCh38 (TGP38C), as a direct comparison across references was not deemed possible.[8]

Cross-comparisons can be performed very efficiently using reduced genome representations, including genome fingerprints.[11] Here, we use genome fingerprints and

some additional statistics to compare the seven versions of the TGP (TGP37, TGP38L, TGP38C, TGP38S, TGP38X, TGP38H, and TGP38N) and their associated related samples (TGP37r, TGP38Sr, TGP38Xr, and TGP38Nr), in terms of (1) the set of genomes analyzed, (2) known and cryptic relatedness within each cohort, (3) patterns of SNV and genotype concordance between versions, and (4) phasing concordance. We then seek explanations for the results of these comparisons with a limited set of rapid follow-up studies.

Material and methods

Datasets

We obtained seven versions of the 1000 Genomes dataset, phase 3, from IGSR.[7]

TGP37

Variant calls relative to the GRCh37 (hg19) version of the human genome reference (N = 2,504).

TGP38L

Variant calls for the same set of genomes, lifted over to the GRCh38 (hg38) version of the human reference using dbSNP v. 149 (N = 2504). This version was later withdrawn (in March 2021).

TGP38C

Variant calls for the same set of genomes, crossmapped to the GRCh38 (hg38) version of the human reference using CrossMap (N = 2,504).

TGP38S

A set of integrated phased biallelic SNV calls, directly called against the GRCh38 (hg38) version of the human reference (N = 2,548).

TGP38X

An extended set of integrated phased biallelic SNV and indel calls, directly called against the GRCh38 (hg38) version of the human reference (N = 2,548).

TGP38H

A high-coverage (30×) set of recalibrated SNV and indel calls, produced by the NYGC, directly called against GRCh38 with alternative sequences and decoys (N = 2504) - European Network Archive Project PRJEB31736.

TGP38N

A high-coverage (30×) set of recalibrated SNP and indel calls, produced by the NYGC, directly called against GRCh38 with alternative sequences and decoys (N = 2504), and phased in combination with 698 related individuals: European Network Archive Study ERP114329, Project PRJEB31736. We downloaded VCF files including all 3,202 genomes, and split them into 2,504 main, and 698 relateds, for further analysis.

We also obtained four sets of genomes of samples related to those in the main TGP dataset.

TGP37r

Integrated phased biallelic SNV and indel calls, relative to GRCh37 (N = 31).

TGP38Sr

Integrated phased biallelic SNV calls, relative to GRCh38 (N = 150).

TGP38Xr

Integrated phased biallelic SNV and indel calls, relative to GRCh38 (N = 150).

TGP38Nr

Integrated phased SNV and indel calls, relative to GRCh38 (N = 698): European Network Archive Study ERP120144, Project PRJEB36890.

Variant normalization

We normalized the variants in TGP38C and TGP38N using the bcftools norm -m+ command.[18]

Whole-genome fingerprinting

We previously described an algorithm for computing "fingerprints" from genome data.[11] In brief, fingerprints are locality-sensitive hashes that only need to be computed once per genome, not once per comparison, and can be rapidly compared with determine whether two genome sequences are derived from the same individual, closely related individuals, or unrelated individuals. Fingerprints represent collections of local microhaplotypes by encoding consecutive pairs of SNVs in terms of their alleles and relative distance, mitigating most issues arising from differences in sequencing technology, representation format, reference version, and analysis pipeline used. The main parameter to the method (L) determines the size of the fingerprint hash and thus the quality of the comparisons. We computed fingerprints for all genomes in all sets as described, with L = 200. Unless otherwise specified, all genome fingerprints include only biallelic autosomal SNVs. This computation does not include the pseudoautosomal regions (PARs) of chromosomes X and Y.

Fingerprint comparison for relatedness assessment

Genome fingerprints are compared using a simple Spearman correlation; the resulting correlation coefficient (ρ) decreases with increasing degree of relationship and should not be interpreted as a kinship coefficient. Fingerprint correlations above 0.75 indicate the two samples being compared are of the same individual, or identical twins.[11] First-degree relatives typically show correlations between 0.55 and 0.65; progressively higher-degree relatives show lower correlations, and unrelated individuals from the same population show correlations around 0.41, and lower for individuals from different populations. We thus use a conservative cutoff of $\rho < 0.4$ to confidently identify completely unrelated individuals, and a cutoff of $\rho > 0.45$ to identify closely related individuals (typically, degree of relationship 4, or less).

Chromosome fingerprinting

To compute single-chromosome fingerprints, we restricted SNV pair collection to each chromosome and normalized the single-chromosome raw fingerprints separately, yielding single-chromosome normalized fingerprints. Other than restricting the range to the individual chromosome, the procedure is identical to that used for computing whole-genome fingerprints. We applied this procedure also to the PARs.

Other metrics

We computed the SNV count of an individual as the number of biallelic SNVs observed in their genome in either heterozygous state or homozygous for the alternate allele. We computed the genotype concordance of an individual between two datasets as the number of biallelic SNVs (with same coordinates, reference (REF) and alternate (ALT) alleles in the two datasets) in which the individual is heterozygous in both datasets (ignoring phasing of heterozygous sites) or homozygous alternate allele in both datasets, divided by the total number of biallelic SNVs in which the individual was not homozygous reference in both datasets.

Evaluation of sex-chromosome coverage

We downloaded compressed reference-oriented alignment map (CRAM) index files for all samples from the Sri Lankan Tamil

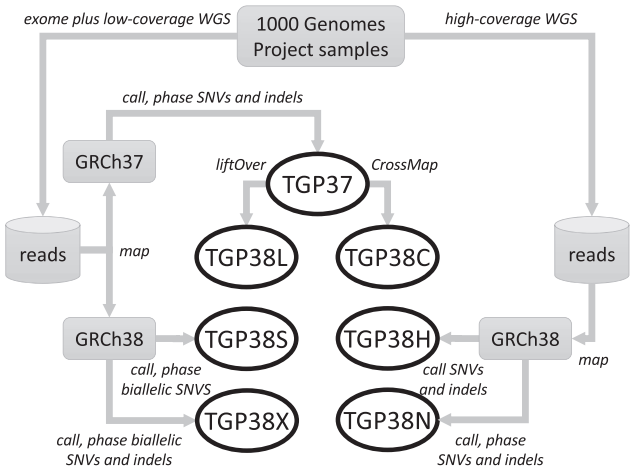

Figure 1. Overview of the seven available versions of sequence variations within the TGP phase 3 cohort

Phase 3 of the 1000 Genomes mapped exome sequences and low-coverage WGS onto GRCh37 to identify sequence variations, yielding version TGP37. Three approaches have since been used to represent the human genomic sequence variation against the GRCh38 reference sequence: liftOver of variants called against GRCh37 (yielding version TGP38L) and similarly using CrossMap (yielding version TGP38C); remapping the individual reads from the original data directly against GRCh38, followed by integrated variant calling resulting in only SNV calls (TGP38S) or both SNV and indel calls (TGP38X); and independent whole-genome resequencing of new reads, for these 2,504 sequences at the modern standard sequencing depth of 30×, (yielding versions TGP38H and TGP38N, which is also phased).

population in the UK (STU) population (*.alt_bwa-mem_GRCh38DH.20150718.STU.low_coverage.cram.crai) and used these to estimate total coverage in chrX and chrY (and hence estimate chrX and chrY copies) by normalizing to the observed coverage in autosomes. We similarly obtained and analyzed CRAM index files for the 698 samples in TGP38Nr.

Identification of genomic regions discrepant between TGP versions

We aligned the VCF files for each pair of TGP versions, to identify variants shared between the two versions (same chromosome, position, reference, and alternate alleles) and to enumerate variants unique to one version (i.e., absent from the other). When encountering variants in which all individuals were reported as homozygous reference (i.e., zero counts of the alternate allele, allele count [AC] = 0), we considered them as absent. We then identified segments enriched in (or depleted of) unique variants, using this procedure: for each pair of consecutive shared variants, we counted how many unique variants are present between them in each of the two TGP versions being compared, and retained for subsequent analysis segments at least 1 kb long. We identified a subset of these segments as significant, based on p values returned by the corresponding cumulative distribution function of the Poisson distribution, modeled using the length of the segment and the number of unique variants in it (separately for each TGP version being compared). The Poisson distribution is used to estimate the number of events in a given time or space interval. It uses a shaping parameter lambda (λ), which denotes the expected number of events in a unit of time or space. We use the genome-wide density of variants to estimate λ for a given segment of length *l*.

$$\lambda = \frac{\textit{total number of variants unique to a TGP version}}{\textit{total length of the genome}} \times l$$

The p value to determine the extremity of occurrence of *k* variants in a segment is then evaluated as:

if $(k > \lambda)$:

$$1 - \sum_{k=0}^{k-1} \frac{1}{(k-1)!} \times e^{-\lambda}\lambda^{k-1}$$

else:

$$\sum_{k=0}^{k} \frac{1}{k!} \times e^{-\lambda}\lambda^{k}$$

Using Bonferroni correction, we considered significant any segment with $p\text{–}value \leq 0.05/2N$, where N = 104,572 is the total number of segments being evaluated in all pairwise comparisons (two p values are computed per segment). We then used bedtools[19] to subtract from the list of unique segments those that overlap with large genomic gaps or satellite sequences (yielding a list of 95,843 segments) and finally to merge them into a set of 12,307 non-overlapping regions.

Identification and classification of discrepant alternate alleles

We compared all six TGP versions on GRCh38 (TGP38L, TGP38S, TGP38X, TGP38H, and TGP38N) to identify autosomal variants in which the alternate allele (ALT) differs between TGP versions. To maximize interpretability, we restricted this analysis to the 2,503 individuals represented in all TGP versions. We tabulated the reported ALT alleles and their ACs in each TGP version, and classified the ALT-discrepant variants into three non-overlapping classes: (1) multiallelic, including variants with more than two ALT alleles in any TGP version; (2) indel-related, including variants in which the REF allele or any ALT allele has length different from 1, or one of the alleles is a "star allele" (the * character), in any TGP version; and (3) biallelic SNV, including all remaining ALT-discrepant biallelic single-nucleotide variants not obviously associated with indels in any version of TGP. Sites that are both multiallelic and indel related are counted among the multiallelic.

Phasing-aware genome fingerprints

We modified the fingerprint-computation method to capture local phasing information by changing the third step of raw fingerprint computation.[11] Specifically, when joining the keys of consecutive SNVs, we reverse the key of the second SNV (i.e., the ALT allele is assigned to the first phased chromosome and the REF allele is assigned to the second chromosome) if only one of the two SNVs matches the genotype pattern "1|0." We computed phasing-aware genome fingerprints for all TGP versions except for TGP38H, since this version is not phased. Finally, we compared the datasets in pairwise fashion using standard and phasing-aware fingerprints.

Results

Overview of datasets and quality assessment strategies

We demonstrate the application of genome fingerprints[11] for rapid cross-comparison of large genome datasets on the seven reported versions of the 1000 Genomes phase

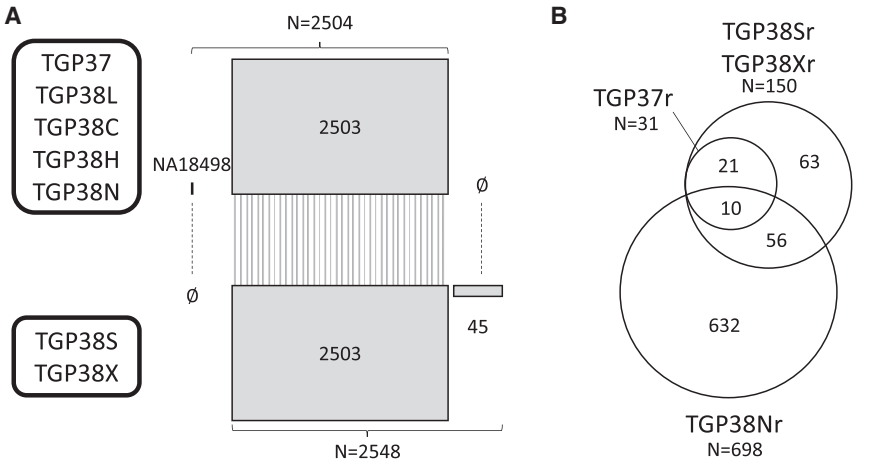

Figure 2. Comparison of cohort structure
We counted sequence identifiers in each TGP set, and came to the following conclusions.
(A) While TGP38L, TGP38H, and TGP38N include the same 2,504 genome identifiers as TGP37, as expected, TGP38X and TGP38S include 2,548 identifiers, only 2,503 of which are in common with TGP37. NA18498 is absent from versions TGP38S and TGP38X, which also include 45 identifiers not in TGP37.
(B) The original set of 31 supplemental related samples (TGP37r) has been expanded to 150 in the versions TGP38Sr and TGP38Xr. TGP38Nr includes 698 genomes with partial overlap with the other sets.

3 cohort (Table 1 and Figure 1), which included the variant set as originally released, mapped to GRCh37 (TGP37); the same variants lifted over to GRCh38 (TGP38L and TGP38C); a reanalysis of the same raw sequence reads directly mapped onto GRCh38, first limited to biallelic SNVs (TGP38S) and then also including biallelic indels (TGP38X); as well as a recent independent high-coverage resequencing (TGP38H and TGP38N) of the original cell lines (sequenced to a nominal target depth of 30× coverage; see section "materials and methods"). We also evaluate smaller sets of related individuals.

We used genome fingerprints as well as simple summary statistics to compare the SNVs reported in these genomes. Based on these analyses, we identified a number of discrepancies and quality issues, including an unexpectedly missing individual, cryptic relations, and a set of genomes with significantly fewer SNV counts. We further identified regions of the genome with excess variation, or significantly lacking variation, in some TGP versions and discrepancies in the identity of the alternate allele (ALT), sometimes at high population frequency. Finally, we observed that phasing is very inconsistent between the TGP versions.

Cohort composition: Metadata versus genome comparison

We compared the identifiers of the genomes included in the seven TGP versions and observed that there are differences in cohort membership (Figure 2A), with four versions (TGP38L, TGP38C, TGP38H, and TGP38N) retaining the cohort used in TGP37, but the two versions directly mapping the original data to GRCh38 (TGP38S, TGP38X) omitting genome NA18498 (Yoruba in Ibadan, Nigeria [YRI]), and including 45 additional genomes. The sets of related genomes also differ (Figure 2B), with 119 additional genomes (TGP38Sr, TGP38Xr) directly mapped to GRCh38 that were not in the original cohort (TGP37r), and 698 genomes in TGP38Nr, partially overlapping TGP37r and TGP38Xr. We found no overlap between the additional 45 genomes in TGP38X, and the 698 related genomes in TGP38Nr.

According to the metadata in the IGSR data portal,[20] the omitted genome NA18498 should have been present in TGP38X. We used genome fingerprint comparisons to evaluate whether NA18498 might have been mislabeled, but found this not to be the case: the highest fingerprint correlation between NA18498 (from TGP38L) and any individual from TGP38X is 0.316 (to HG03108, Esan in Nigeria [ESN]), which also exceeds correlation with any genome among the 150 supplemental individuals in TGP38Xr. These values are well below the 0.75 minimal correlation expected for versions of the same individual,[11] confirming that NA18498 was not included in TGP38X or TGP38Xr under a different identifier. Of the 45 additional individuals in TGP38X, most (75%) seem to be related to individuals in the TGP37 cohort, some with fingerprint correlations consistent with second-degree relationships (>0.55; Table S1), suggesting these genomes should not be included as part of the main TGP cohort.

Fingerprint-based comparisons of the 2,503 individuals shared between TGP37 and TGP38X or TGP38S (Figure 2A) confirm a one-to-one relationship: for each individual in TGP37 (excluding NA18498, discussed above), the highest fingerprint correlation observed was to the TGP38X/TGP38S individual with the same identifier. For all but eight, the correlation is well above 0.75, as expected; the remaining eight individuals ("strongly affected;" Table 2), all from the African Caribbean in Barbados (ACB) population, have between-set correlations that would be erroneously consistent with first-degree relationships.[11] An additional eight individuals ("mildly affected;" Table 2) have between-set fingerprint correlations ranging from 0.787 to 0.865, indicating the same individual (>0.75) but much lower than observed for the other between-set self-comparisons (0.885 ± 0.0028).

To evaluate the nature of these discrepancies, we tabulated the number of biallelic autosomal SNVs observed for each individual genome in each of the seven datasets. For most individuals, we observed about 2% fewer SNVs in TGP38X than in TGP37 (Table 2 and Figure 3). Possible explanations for this reduction might include changes in the reference (including reference/alternate allele

Table 2. Observed statistics for the outlier individuals most affected by dataset recomputation from TGP37 to TGP38X, in comparison with the platinum NA12878 genome, the 89 ACB individuals unaffected by this bioinformatic difference, and the 2487 similarly unaffected individuals in the entire cohort

				SNV count (millions)		SNVs lost	
Section	ID	Population code	Fingerprint correlation	TGP37	TGP38X	(%)	Genotype concordance
Missing	NA18498	YRI	NA	4.28	0.00	100.00	NA
Strongly affected	HG02325	ACB	0.550	4.28	3.33	22.18	0.578
	HG02442	ACB	0.554	4.26	3.35	21.40	0.587
	HG02433	ACB	0.570	4.25	3.36	20.82	0.600
	HG01989	ACB	0.571	4.20	3.33	20.69	0.602
	HG02445	ACB	0.571	4.20	3.32	20.84	0.601
	HG02343	ACB	0.585	4.19	3.35	20.17	0.615
	HG02420	ACB	0.595	4.12	3.29	20.22	0.626
	HG01988	ACB	0.603	4.04	3.26	19.37	0.632
Mildly affected	NA19189	YRI	0.787	4.26	3.98	6.67	0.871
	HG00232	GBR	0.814	3.48	3.26	6.13	0.905
	HG00530	CHS	0.824	3.49	3.30	5.49	0.919
	HG00542	CHS	0.827	3.49	3.29	5.60	0.923
	HG00531	CHS	0.839	3.49	3.32	4.86	0.938
	NA18960	JPT	0.840	3.55	3.36	5.22	0.947
	NA18856	YRI	0.862	4.31	4.16	3.49	0.965
	HG00116	GBR	0.865	3.51	3.38	3.66	0.969
Reference	NA12878	CEU	0.885	3.52	3.44	2.24	0.989
	average	ACB (n = 89)	0.887	4.26	4.17	2.02	0.988
	SD	ACB (n = 89)	0.0024	0.03	0.04	0.13	0.002
	average	all (n = 2487)	0.885	3.74	3.66	2.19	0.988
	SD	all (n = 2487)	0.0028	0.33	0.32	0.17	0.002

Genotype concordance is between TGP38L and TGP38X, as TGP37 and TGP38X are not on the same reference.

switches), improved variant calling leading to fewer false-positives, and stricter coverage requirements leading to more false-negatives. The eight mildly affected genomes have 3.5%–6.7% fewer SNVs in TGP38X than TGP37. In contrast, the eight strongly affected genomes have 20%–22% fewer SNVs in TGP38X than in TGP37. These high missingness values are well beyond the range observed for the mildly affected genomes (5.14% ± 1.11%), which are themselves well beyond the range of the majority of genomes (2.19% ± 0.17%). While the strongly affected ACB genomes have similar SNV counts in TGP37 to the remaining 89 ACB individuals, they have markedly fewer SNVs in TGP38X. We also observed reduced genotype concordance for the strongly affected genomes relative to the other ACB individuals (Table 2). We evaluated the sizes of CRAM files and their associated CRAM index (CRAI) files as indicators of the amount of data in each of these remapped genomes, and found that those of the affected individuals are sometimes significantly smaller than for other genomes, likely leading to undercalling of variants (Figure S1). Finally, comparison with the TGP38H and TGP38N versions, which were constructed using an independent set of sequence reads, shows that the SNV counts in these individuals are concordant between TGP37, TGP38H, and TGP38N (Figure 3). Taken together, these differences suggest that TGP38X erroneously excludes a set of 0.8–0.9 million SNVs from the strongly affected ACB genomes; these SNVs are included by independent analyses (e.g., TGP37 and TGP38N). TGP38S is similarly affected.

Evaluation of the associated samples of related individuals

We evaluated the 698 related individuals in TGP38Nr and the 150 related individuals in TGP38Xr (Figure 2B), expecting all of them to show some degree of relatedness to at least one of the individuals in the main TGP cohort. (Due to the similarity between TGP38Sr and TGP38Xr, we elected to only report results for the latter.) We computed fingerprints for all TGP38Nr individuals, then compared them among themselves, and with TGP38N. All TGP38Nr individuals are indeed closely related to at least one other individual in TGP38N or TGP38Nr

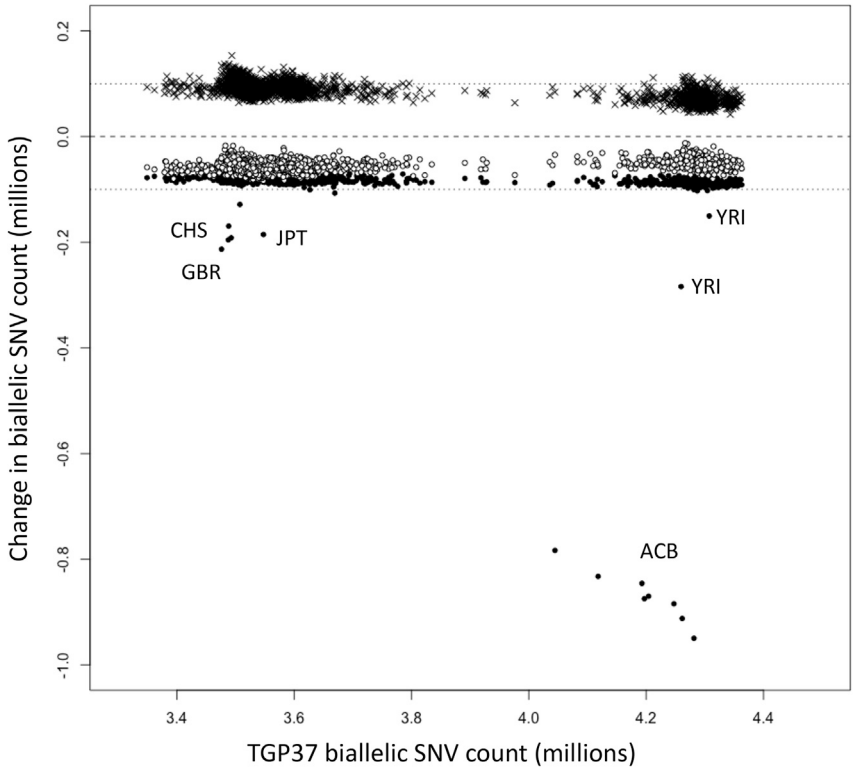

Figure 3. SNV count comparisons
Change in biallelic SNV counts (in millions) relative to TGP37, as observed in TGP38H (crosses), TGP38N (open circles) and TGP38X (filled circles). TGP38X and TGP38N universally report fewer SNVs than TGP37; 16 genomes are visibly exceptional in TGP38X. In particular, eight individuals from the ACB population have significantly fewer SNVs in TGP38X relative to TGP37. TGP38H universally reports more SNVs than TGP37, but without obvious outliers. Dashed line, no SNV count change; dotted lines, addition (or loss) of 100,000 SNVs.

(fingerprint correlations, $\rho > 0.53$). We similarly computed fingerprints for all TGP38Xr individuals, then compared them with all TGP38X individuals and with each other (Table S2). Over two-thirds of the TGP38Xr individuals can indeed be recognized as closely related to TGP38X individuals, with fingerprint correlations $\rho > 0.45$ (see section "materials and methods"). On the other hand, at least 28 of the TGP38Xr individuals seem not to be related to anyone else in TGP38X or TGP38Xr (fingerprint correlation, $\rho < 0.4$).

One of the related individuals in TGP38Xr (HG03982, from the STU population) has fingerprint correlation of 0.868 to an individual in TGP38X (HG03858, also STU). This fingerprint correlation level would suggest these are the same individual, and yet, in IGSR, HG03858 is annotated as female, while HG03982 is annotated as male. Neither individual has any relatives annotated in either TGP38X or TGP38Xr, nor could we identify any relatives by fingerprint comparison. We considered various hypotheses, including whether these individuals could be sex-discordant monozygotic twins (as a result of sex change, through differential resolution of XXY karyotype, mosaicism, etc.), the result of mislabeling of twin samples, or mislabeled, redundant samples of the same individual. TGP37 data support HG03858 being genetically female, with two copies of chrX and no chrY. To test these hypotheses, we evaluated whether HG03982 could indeed be a male sample, as annotated.

1. No chrY variant calls were released for TGP38Xr, and chrX variant calls are available only in the pseudoautosomal regions (PARs, which combine data from chrX and chrY). We computed chromosome-specific fingerprints, including for the PARs. The resulting 0.928 correlation of PAR-specific fingerprints of HG03858 and HG03982 suggests these two samples have the same sex-chromosome karyotype (XX) and the same chrX haplotypes, consistent with being the same individual, or identical twins. For comparison, we observe 0.954 correlation of PAR-specific fingerprints of samples HG00578 and HG00635 (female siblings, with overall autosomal fingerprint correlation of 0.689), and 0.622 correlation of samples HG00512 and HG00501 (male and female siblings, respectively, with overall autosomal fingerprint correlation of 0.687).
2. We observed 91.8% genotype concordance in the PARs of HG03858 and HG03982, consistent with these being the same person or female siblings. For comparison, the PAR genotype concordance of female siblings HG00578 and HG00635 was 94.1%, and the PAR genotype concordance of male and female siblings HG00512 and HG00501 was 45.1%.
3. Based on the coverage levels along chrX and chrY (from low-coverage data), we estimated the number of chrX and chrY copies for all STU samples and found that HG03982 clusters with the female samples (Figure S2A).

We conclude that HG03858 and HG03982 are both genetically female. Lacking further information about the individuals, we hypothesize that these two samples derive from the same individual or from identical twins, and that HG03982 may have been annotated as male as a result of a clerical error.

When comparing, through fingerprinting, the partially overlapping sets of related individuals in TGP38Xr and TGP38Nr (Figure 2B), we identified pairs of individuals with different identifiers but with genome fingerprint correlations consistent with them being the same individual,

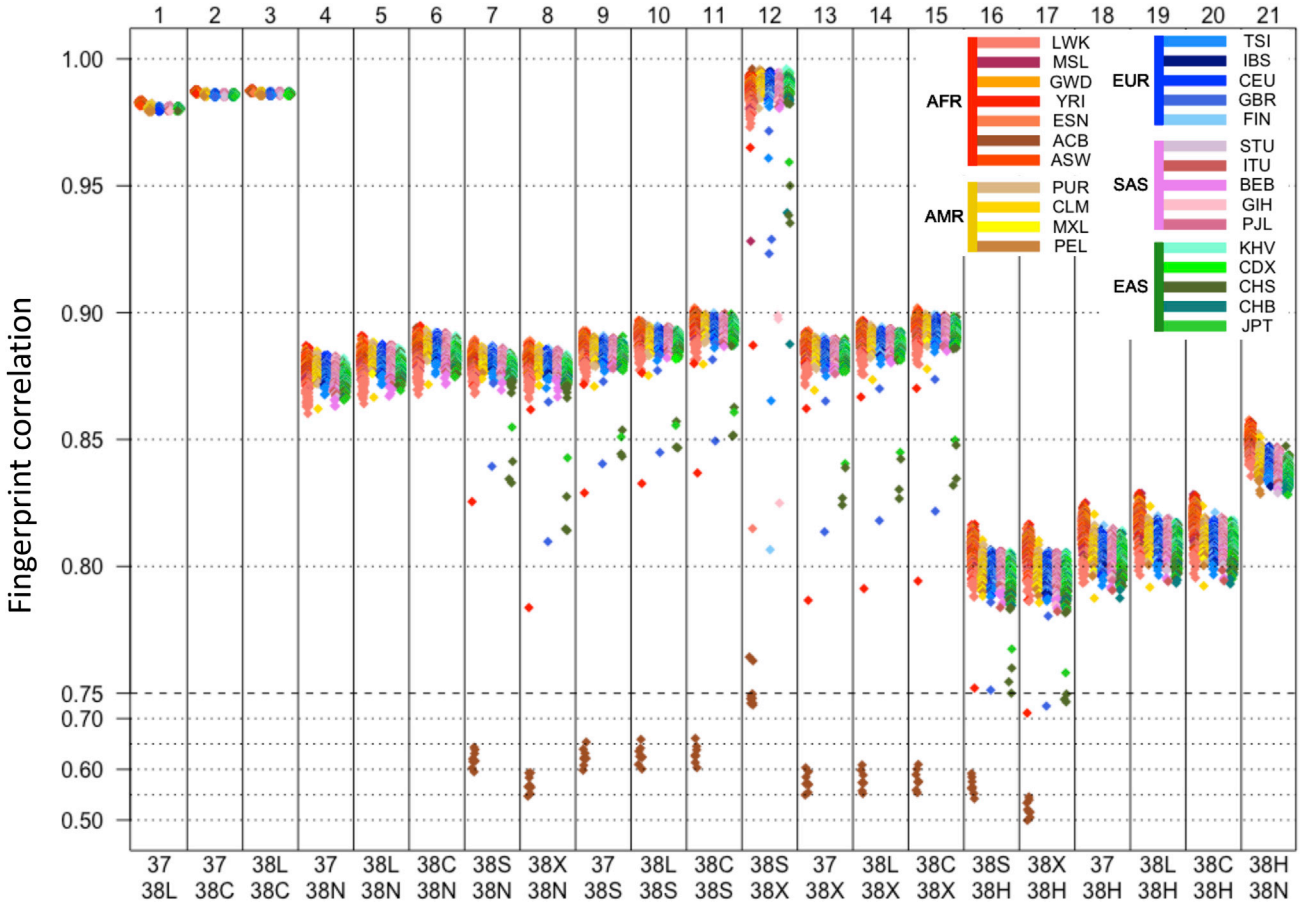

Figure 4. Distributions of fingerprint self-correlations in pairwise dataset comparisons
We compared fingerprints across TGP versions for each individual (color coded by population), resulting in a distribution of Spearman's ρ (y axis) for each pair of versions (columns, labeled with numbers above, and with the versions being compared, below). Self-correlations of fingerprints for the same individual are expected to be high ($\rho > 0.75$) irrespective of genome version or analysis pipeline; lower self-correlations across versions show some genomes are unrecognizable as the same individual. Low outliers ($\rho < 0.75$) are shown on a compressed scale under the dashed line.

suggesting identifier swaps: genome HG03797 in TGP38Xr is equivalent to genome HG03799 in TGP38Nr, and conversely genome HG03799 in TGP38Xr is equivalent to genome HG03797 in TGP38Nr. Since HG03797 and HG03799 are, respectively, the father and mother in different trios (Bengali families BD16 and BD17), we were able to determine, through fingerprint comparisons with the offspring in these trios, that the genome identifiers were inadvertently swapped in TGP38Xr (and similarly in TGP38Sr). We similarly observed a potential identifier swap involving genomes HG03699 and HG03700, which are respectively the father and mother in Punjabi family trio PK59. In this case, fingerprint comparisons with the offspring could not elucidate the nature of the discrepancy, since both parents are equally related to their child. We instead estimated chrX/chrY copy numbers in the TGP38Nr data, as described above, to determine that HG03699 is male and HG03700 is female, consistent with the metadata on these individuals (Figure S2B). Thus, we conclude that the identifiers were again mistakenly swapped in TGP38Xr (and TGP38Sr).

Evaluation of fingerprint differences across TGP versions

While comparison of results pertaining to a high-quality genome (NA12878) can validate that a data processing pipeline has the capacity to perform its intended task well, it cannot assess whether the pipeline performs consistently across different input data. Since in-depth comparison of every genome would be low throughput and expensive, we used genome fingerprints for their primary purpose: to perform rapid self-comparison of every genome across the different processing pipelines (Figure 4).

The comparisons among TGP37, TGP38L, and TGP38C (Figure 4, columns 1–3) corresponded well for every genome ($\rho > 0.96$). In addition, the low variability in fingerprint correlation, both within and across TGP populations (colors), suggests that the liftOver and CrossMap pipelines perform very consistently across individuals.

The TGP38S and TGP38X versions of each genome differ by whether indels were reported as variants (TGP38X only) along with the SNVs. For the majority of genomes, self-comparisons between these versions (Figure 4, column 12) corresponded well and varied little across genomes

Table 3. Statistics of pairwise comparisons of the six GRCh38 callsets

			Unique SNVs		% Unique in hard regions		
Set 1	Set 2	Shared SNVs	To set 1	To set 2	In set 1	In set 2	Genotype concordance
TGP38L	TGP38C	77,311,452	82,520	260,502	11.2	11.0	0.9999
TGP38L	TGP38S	68,021,487	9,344,300	3,407,548	5.2	24.5	0.9594
TGP38L	TGP38X	68,077,971	9,287,798	3,413,699	5.2	24.5	0.9593
TGP38L	TGP38H	74,767,408	2,605,724	29,944,460	15.6	18.6	0.9435
TGP38L	TGP38N	50,748,581	26,569,818	8,785,284	2.9	16.8	0.9264
TGP38C	TGP38S	68,143,522	9,400,227	3,285,493	5.2	25.0	0.9593
TGP38C	TGP38X	68,200,090	9,343,642	3,291,561	5.2	25.0	0.9592
TGP38C	TGP38H	74,862,849	2,687,911	29,848,665	15.7	18.6	0.9435
TGP38C	TGP38N	50,853,713	26,642,369	8,679,853	2.9	16.8	0.9263
TGP38S	TGP38X	71,334,231	123,182	185,835	5.3	5.5	0.9711
TGP38S	TGP38H	68,970,797	2,473,081	35,748,569	27.9	15.4	0.9535
TGP38S	TGP38N	48,313,920	23,069,356	11,221,574	4.3	12.0	0.9367
TGP38X	TGP38H	69,028,474	2,478,010	35,690,845	27.9	15.4	0.9507
TGP38X	TGP38N	48,331,682	23,114,098	11,203,663	4.3	12.0	0.9322
TGP38H	TGP38N	57,792,440	46,932,927	1,809,657	10.7	32.0	0.9993

Number of shared SNVs, number of SNVs unique to each set, percentage of unique SNVs in hard genomic regions (spanning 8.76% of the autosomal sequence), and average genotype concordance over the 2,503 individuals included.

($\rho > 0.97$). However, a subset of 17 genomes ("strongly affected" in Table S3) show very reduced fingerprint correlations, very reduced SNV counts, and low genotype concordances. These include nine genomes from the ACB population, adding HG02436 (which is absent from the TGP37 set) to the eight ACB genomes described above. A further set of 15 individuals ("mildly affected" in Table S3) showed an intermediate, outlying level of degradation in self-correlations, SNV counts, and genotype concordance. Since these same 32 individuals include the outliers in all other comparisons involving versions TGP38S or TGP38X (Figure 4, columns 7–17), while the comparisons not involving these TGP versions do not produce comparable outliers (Figure 4, columns 1–6 and 18–21), we deduce that the pipelines used to create the TGP38S and TGP38X versions were less successful in processing these genomes than the rest of the cohort.

While TGP38H and TGP38N are based on the same high-coverage data, these two versions differ in many ways (Table 1), not just in TGP38N being phased. In particular, variants are represented in very different formats. Furthermore, the TGP38H versions of each genome include many additional variants, particularly around centromeric regions (see next section). This leads to reduced self-correlations in comparisons between TGP38H and other TGP versions (Figure 4, columns 16–21).

Identification of genomic regions with excess variation or missing variation

The various TGP versions differ significantly in the total number of SNVs reported. We compared in pairwise fashion the four versions of the TGP expressed relative to GRCh38, by tallying (from the VCFs) how many variants are shared (i.e., both datasets have an SNV at a given coordinate, with the same REF and ALT alleles), and how many are unique to each set. We based this analysis on the 2,503 genomes shared by all TGP versions (i.e., all genomes in the original TGP37 set, minus NA18498). We found that each TGP version has a large number of SNVs absent from the other versions (Table 3). We evaluated whether these mutually unique SNVs are located in "hard" regions of the genome, namely centromeres, microsatellite repeats, and low-complexity regions.[21] While in some comparisons, and particularly in comparisons involving TGP38X and TGP38S, there is an enrichment of unique variants in hard regions, most of the unique variants are nevertheless outside these more difficult-to-sequence regions. In fact, the chromosomal distribution of these discrepancies is a combination of (1) a nearly uniform background level characteristic to each dataset, and (2) multiple clusters of variants unique to one dataset (Figure 5).

A particularly noteworthy signal involves a very large number of SNVs in and around centromeric regions, unique to TGP38H (Figure 5 lower inset, and Figure S3). Their absence in TGP38L and TGP38C is not surprising, since GRCh37 lacked centromeric sequence models and thus there were no centromeric variants in TGP37 to be lifted over; as for TGP38S and TGP38X, the method used for generating these versions filtered out variants in centromeres.[6] The density of centromeric variants in TGP38H also significantly exceeds the density observed elsewhere along the chromosomes (Figure S3).

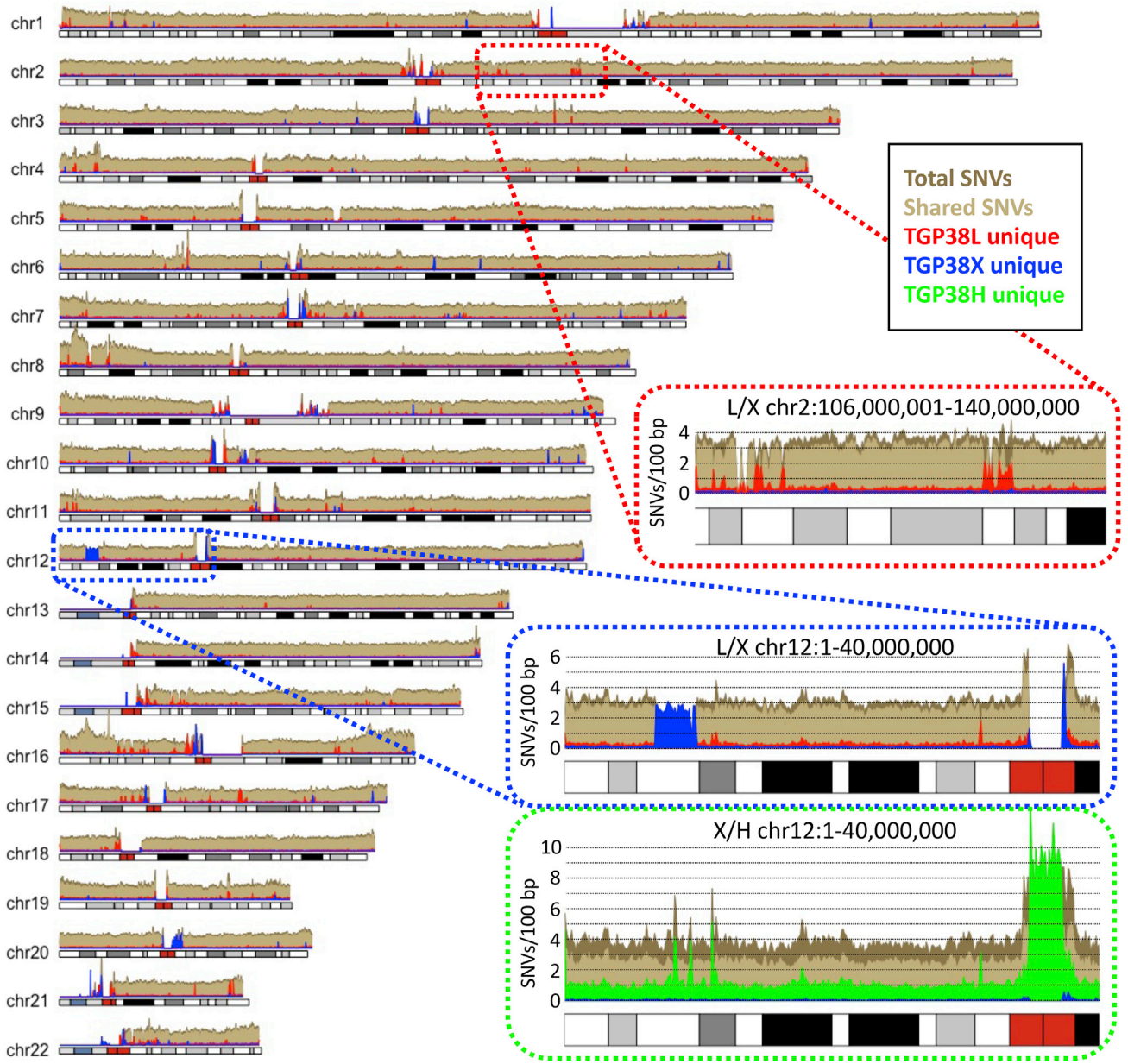

Figure 5. Density map of shared and unique variants along the chromosomes
The main plot shows the comparison between TGP38L and TGP38X, with shared variants in light brown, variants unique to TGP38L in red, those unique to TGP38X in blue, and the union of all variants in dark brown. Top inset (red outline): detail of a 36-Mb region on chr2 showing patches enriched in variants unique to TGP38L (absent from TGP38X). Middle inset (blue outline): detail of the first 40 Mb of chr12 showing a segment with only variants unique to TGP38X and no variation at all in TGP38L. Neither version has variants in the centromeric region. Lower inset (green outline): the same chr12 region when comparing TGP38X and TGP38H, highlighting the overall higher density of TGP38H-unique variants (in green), and the excess variation unique to TGP38H in centromeric regions.

Through pairwise comparison of the TGP versions, we identified 104,572 genomic segments significantly enriched (or depleted) in SNVs unique to one TGP version relative to the other (see examples in Figure 5 insets) and evaluated them for overlap with genomic features. We found that the segments with the largest counts of unique variants correspond to centromeric, telomeric, and other heterochromatin and satellite sequences. After excluding these segments, we merged the 95,843 remaining ones by overlap into 12,307 regions.

The longest of these regions was a gene-dense 2.96-Mb segment on chromosome 12 erroneously lacking variation in TGP38L (Figure 5, middle inset). This segment (chr12:6889106-9847458, between dbsnp:rs143703503 and dbsnp:rs111979444) has 71,946 variants unique to TGP38X (112,750 in TGP38H) but no variants shared with TGP38L, and no SNVs unique to TGP38L. In fact, the only variation reported in this region in TGP38L includes two indels (dbsnp:rs11394173 and dbsnp:rs56027063) with alternate allele frequency of 1, suggesting an error in the reference

Table 4. Discrepant ALT alleles of pairwise comparisons of the six GRCh38 callsets

		ALT change					Observed/expected	
Set 1	Set 2	All	AC>0	AC>9	Exonic	Coding	Exonic	Coding
TGP38L	TGP38C	184	184	103	16	6	1.70	2.31
TGP38L	TGP38N	75,729	74,471	8,122	3,828	1,174	0.97	1.10
TGP38C	TGP38N	75,942	74,696	8,154	3,852	1,169	0.97	1.09
TGP38S	TGP38N	74,137	71,845	7,126	4,268	1,571	1.13	1.53
TGP38X	TGP38N	73,690	71,691	6,952	4,260	1,565	1.13	1.53
TGP38L	TGP38S	28,351	28,323	3,484	1,614	549	1.08	1.35
TGP38C	TGP38S	28,397	28,227	3,418	1,605	539	1.08	1.33
TGP38S	TGP38X	0	0	0	0	0	NA	NA
TGP38L	TGP38X	28,358	27,886	3,105	1,598	547	1.09	1.37
TGP38C	TGP38X	28,328	27,807	3,047	1,590	537	1.08	1.35
TGP38S	TGP38H	13,535	9,111	3,899	459	190	0.95	1.46
TGP38X	TGP38H	13,852	9,012	3,871	451	190	0.94	1.47
TGP38L	TGP38H	21,106	17,693	4,989	876	307	0.93	1.21
TGP38C	TGP38H	21,484	18,114	5,059	932	312	0.97	1.20
TGP38H	TGP38N	7,818	6,998	2,004	220	102	0.58	1.01

All: number of autosomal variants (SNVs and indels) with different stated ALT alleles. AC > 0: number of ALT-discrepant autosomal SNVs with ACs of at least 1 in both datasets. AC > 9: number of ALT-discrepant autosomal SNVs, further requiring the sum of the ACs in both datasets to be at least 10. Exonic: number of AC > 0 ALT-discrepant autosomal SNVs mapped to any exons (coding and non-coding). Coding: number of AC > 0 ALT-discrepant autosomal SNVs mapped to coding exons. Observed/expected (exonic, coding): ratio of AC > 0 ALT-discrepant SNVs in exons and coding regions, respectively, relative to their corresponding expected values from non-exonic and non-coding densities.

sequence. On the other hand, the corresponding range in GRCh37 (chr12:6999223-10000058) includes 43,406 variants (SNVs and indels) in TGP37. We verified that it is possible, using the liftOver tool, to transform coordinates in this region from GRCh37 to GRCh38. We conclude that a bioinformatic pipeline error likely caused this extensive set of variants to fail to lift over from TGP37 to TGP38L, yielding a significant gap in the reference variation. We note that this is a large and gene-dense region, including 100–150 transcript clusters, 75 of which are not pseudogenic. Bioinformatic analyses in this region, using annotated variation in TGP38L as reference, may have been confounded by this issue.

We evaluated the remaining regions over 250 kb in length and found them to largely correspond to genomic loci that were missing (gaps) or misassembled in previous versions of the reference genome (Table S4). Most of these regions have since been corrected or partially improved in GRCh38, and some still include gaps. Of note, since much knowledge on population variation was annotated (and submitted to dbSNP) based on older genome reference versions, and then lifted over to the current version, many of these regions show significantly reduced variation compared with the regions surrounding them (e.g., when visualized via genome browsers). For example, when comparing TGP38L with TGP38X, we identified a 135-kb segment on chromosome 6 with 4,588 SNVs unique to TGP38X, but no variation in TGP38L. These 135 kb of sequence in GRCh38 replace a 50-kb clone gap in GRCh37 (GRC issue HG-564). As a result, this region is significantly depleted in variants in dbSNP (Figure S4).

Identification of alternate allele discrepancies

One of several improvements in GRCh38 relative to GRCh37 was the correction of several thousand reference alleles from a minor/rare allele to the major allele.[22] Since TGP38L, TGP38C, TGP38S, TGP38X, TGP38H, and TGP38N are all represented relative to the same reference sequence, all variants shared between any two of these TGP versions have the same reference allele. On the other hand, while tallying shared and unique variants among the TGP versions (previous section), we identified SNVs with discrepant alternate (ALT) alleles. While TGP38S and TGP38X had the same ALT allele throughout the genome, and there were few ALT differences between TGP38L and TGP38C (both being directly derived from TGP37), we observed (tens of) thousands of discrepant ALT alleles genome-wide between every other pair of TGP versions (Table 4). These discrepancies are enriched in exons and in coding regions, relative to the number expected from the densities observed in non-exonic and non-coding regions, respectively, perhaps due to the inclusion of exome-derived data in all TGP versions except TGP38H and TGP38N. The enrichment in coding regions is most prominent in comparisons involving TGP38S or TGP38X.

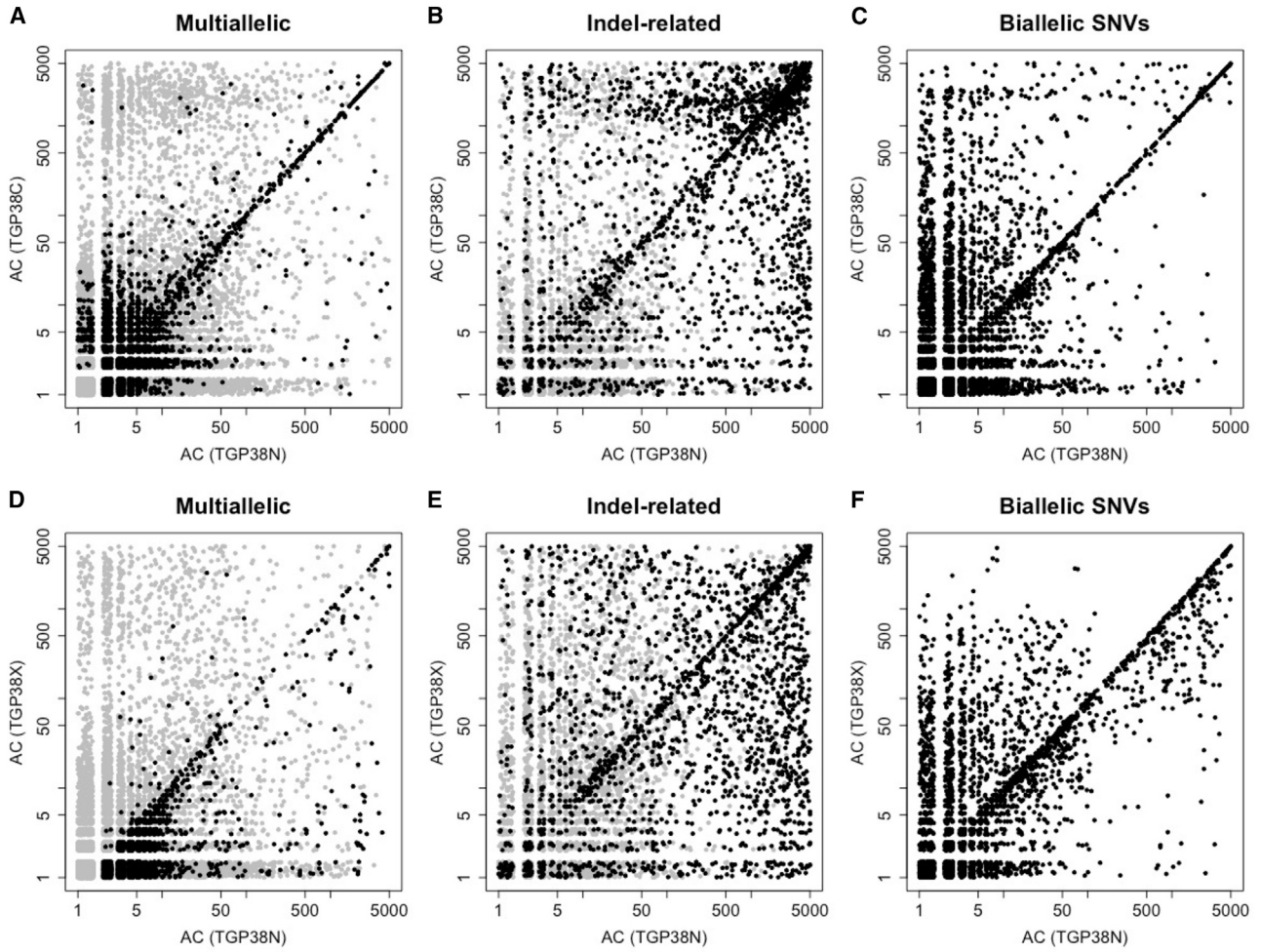

Figure 6. ALT-discrepant sites span the frequency range
Allele counts (ACs) in log scale, jittered horizontally and vertically to denote density, for ALT-discrepant sites between TGP38N and the TGP version on the y axis. TGP38C (A–C); TGP38X (D–F).
(A and D) Sites that are multiallelic in either version being compared (black) or in any of the other versions (gray).
(B and E) Indel-related sites in either version being compared (black) or in any of the other versions (gray).
(C and F) Biallelic SNVs not affected by indels.

Several reasons can contribute to the identification of discrepant alternate alleles. From observational noise alone, one can expect different alternate alleles to be observed at low frequencies, particularly for independently sequenced samples. Some ALT discrepancies can be partially explained by the fact that TGP38S and TGP38X report only biallelic sites, while TGP38L, TGP38C, TGP38H, and TGP38N also include multiallelic sites. For example, at chr1:44,819,962 (chr1:45,285,634 in GRCh37 coordinates), TGP38L reports three individuals with ALT = A, while both TGP38S and TGP38X report one individual with ALT = C. In contrast, TGP38H reports a triallelic site with three individuals with ALT = A, and one with ALT = C. In some other cases, the discrepancy can be attributed to different representations of variation in the vicinity of insertions, deletions, and multi-nucleotide variants (Table S5). We thus tallied all ALT discrepancies, classified them based on these features (see section "materials and methods"), and visualized the three resulting classes relative to their ACs (Figure 6).

While most of the ALT-discrepant variants are rare (a fraction have AC = 0 in one or more of the TGP versions; see Table 4), some are high frequency and thus cannot be explained by observational noise. In fact, for all classes of discrepant ALT allele sites, we observed loci where every individual is reported to be homozygous for a non-reference allele, but for different alleles in the two TGP versions.

Multiallelic sites in TGP38C discrepant with TGP38N span the frequency spectrum; discrepant sites that are biallelic in both versions but multiallelic in some other version (TGP38L, TGP38S, TGP38X, or TGP38H) display generally higher frequencies in TGP38C than in TGP38N (Figure 6A). Indel-related variants are enriched in high frequencies and tend to have higher reported frequency in TGP38C relative to TGP38N (Figure 6B). Similarly, we observed a tendency for higher reported frequencies in TGP38C for simple biallelic SNVs, unaffected by indels, that are ALT discrepant with TGP38N (Figure 6C).

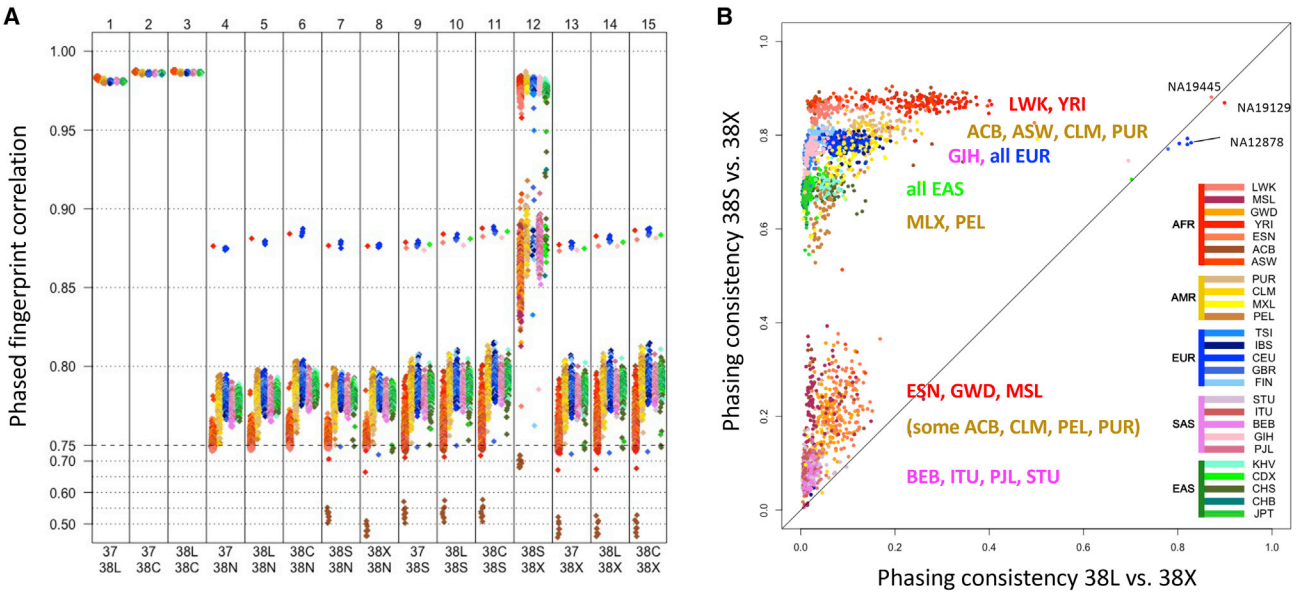

Figure 7. Consistency of phasing between TGP versions

(A) We compared phase-aware fingerprints across TGP versions for each individual (color coded by population), resulting in a distribution of Spearman's ρ (y axis) for each pair of versions (columns, labeled with numbers above, and with the versions being compared, below). Low outliers ($\rho < 0.75$) are shown on a compressed scale under the dashed line.

(B) Using a phasing accuracy metric to compare TGP38L versus TGP38X (x axis) and TGP38S versus TGP38X (y axis) confirms the low phasing concordance for most genomes, and the differential phasing concordance by population, as observed in (A), column 12.

When comparing TGP38X with TGP38N, we observed a strong tendency for higher allele frequencies in TGP38N relative to TGP38X for multiallelic SNVs (Figure 6D) and those affected by indels (Figure 6E). We observed discrepant biallelic SNVs unaffected by indels at all frequencies (Figure 6F). We also noted an enrichment in discrepant biallelic SNVs with high frequencies in TGP38N but lower frequencies in TGP38X, potentially as a result of undercalling in the latter version.

Analysis of phasing concordance

We evaluated the concordance among the six TGP versions that are phased (TGP37, TGP38L, TGP38C, TGP38S, TGP38X, and TGP38N) in two ways: through phasing-aware fingerprinting, and through a phasing accuracy metric we described previously, representing the likelihood that any two heterozygous loci in a diploid chromosome are correctly phased with respect to each other.[23] The metric ranges from zero (for fully random phasing) to one (for perfect phasing). Since mutual phasing between two loci is unperturbed by any intervening incorrectly phased loci, this metric is robust to high-frequency switch errors. On the other hand, a single phasing error in the middle of the sequence could drop phasing accuracy to zero, effectively penalizing gross phasing errors.

First, we modified the procedure for computing genome fingerprints to capture phasing information. Genome fingerprints normally capture the presence of alternate alleles but ignore whether the alternate allele was phased relative to other variants. We modified the method to consider whether the alternate alleles in two consecutive heterozygous variants are phased to be on the same haplotype (e.g., both maternal) or on opposite haplotypes (i.e., one maternal and one paternal). We then encode the latter scenario (opposite haplotypes) by swapping the reference and the alternate alleles for the second variant. In this way, phasing-aware genome fingerprints can capture local phasing information (microhaplotypes), leading to decreased similarity between genomes that differ in such local phasing.

We computed phasing-aware genome fingerprints for the six phased TGP versions and evaluated the change in self-correlation for each individual. When comparing versions TGP37 and its liftover derivatives TGP38L and TGP38C (Figure 7A, columns 1–3), we observed that adding phasing information leads to slightly higher self-correlations compared with standard fingerprints (Figure 4), as expected from consistent phasing. In contrast, in all other comparisons adding phasing information strongly reduces correlations, indicating phasing discrepancies between the TGP versions. This reduction in correlations is generally stronger for genomes of African individuals. In comparisons of either TGP38S or TGP38X with other versions, we again observed several outliers with reduced correlations (including the degraded ACB genomes). In contrast, five genomes have much higher correlations than the rest, suggesting their phasing was much less inconsistent across versions: NA12878, NA10851, NA10847, and NA07048 from the CEPH Utah Residents (CEU) population, and NA19129 (YRI). An additional four genomes also have increased correlations when comparing TGP38S or TGP38X with TGP37, TGP38L, or TGP38C: NA20910 (Gujarati Indian from Houston, Texas [GIH]),

NA19445 (Luhya in Webuye, Kenya [LWK]), HG00867 (Chinese Dai in Xishuanagbanna, China [CDX]), and HG00155 (British in England and Scotland [GBR]).

When comparing TGP38S and TGP38X (Figure 7A, column 12), we observed overall degradation indicating phasing discrepancies between these two related TGP versions: there were many outliers (including the degraded ACB genomes), and one-third of the genomes had much stronger discrepancies in phasing. Surprisingly, this analysis revealed a differential effect by population: the most affected genomes include all the individuals in the South Asian (SAS) populations Bengali from Bangladesh (BEB), Indian Telugu from the UK (ITU), Punjabi from Lahore, Pakistan (PJL), and STU, but none from GIH; all the individuals in the African (AFR) populations ESN, GWD, and Mende in Sierra Leone (MSL), but none from LWK and YRI; sizable subsets of Admixed American (AMR) populations ACB, CLM, PEL, and PUR; and a few individuals from East Asian (EAS) populations CDX and CHS, European (EUR) populations FIN, GBR, and IBS, and AMR populations ASW and MXL.

Second, we studied the VCFs of phased TGP versions on GRCh38 in pairwise fashion to identify phase switches (one version relative to the other) in each individual. Using a phasing accuracy metric,[23] we computed for each individual the level of phasing agreement between TGP38L and TGP38X, and separately between TGP38S and TGP38X (Figure 7B). We observed overall very low phasing concordance between TGP38L and TGP38X except for the same nine individuals highlighted in the fingerprint-based results. Similarly, the TGP38S-TGP38X comparison shows differential phasing consistency per population (see population annotations in Figure 7B), again corroborating the fingerprint-based results.

Discussion

We presented here a cross-comparison of seven versions of the TGP dataset, as an effective way to perform QC. The TGP versions are expressed relative to two versions of the reference genome (GRCh37 and GRCh38); the ability to compare genomes across different versions of the reference sequence was a specific goal in the design of genome fingerprints, which we used for this study.[11] In addition to the overall comparison of the full datasets of phased SNVs and indels (TGP37 versus TGP38X, and related samples), other pairwise comparisons of TGP versions provided insights into the effects of lifting over variants from one reference version to the other (TGP37 versus TGP38L/C), of lifting over versus native mapping and variant calling (TGP38L/C versus TGP38X), of different variant-calling procedures (TGP38X versus TGP38S), and of high-coverage sequencing (TGP38H/N). Through these comparisons, we identified multiple discrepancies between the datasets, pointing at changes in the list of included genomes, some additional cryptic relationships, overall changes in biallelic SNV counts, more significant changes in SNV counts and reduced genotype concordance affecting a subset of the individuals, and low concordance of variant phasing. These observations illustrate the value of employing data reduction techniques, such as the genome fingerprints used here, to enable quality evaluation of the output from every input to an automated pipeline, and not just a gold-standard subset of the individuals.

The utility to perform QC of large genome datasets through cross-comparison is not unique to genome fingerprinting. Multiple methods and tools have been developed for comparing genomes, aiming to establish whether two samples represent the same individual, or to robustly estimate their degree of relatedness. Notable examples include HYSYS,[24] which uses the concordance of homozygous germ-line variants as a metric of relatedness; NGSCheckMate,[25] which compares allele read fractions of known exonic SNPs and can take multiple data types as input; and the recently published Somalier,[26] which extracts "sketches" of variant information for fast comparison. To achieve robustness against diversity of technologies, pipelines, and noise, these methods focus on a few thousand well-studied, high-quality marker SNPs that are typically exonic, sometimes with additional population frequency requirements, and avoiding difficult genomic regions. Such methods would be applicable for performing some of the aspects of cross-comparison QC we described; e.g., the cross-matching of individuals among TGP versions, the analysis of the additional 45 individuals in TGP38X/S, and the study of related individuals. On the other hand, since these methods are by design limited to exonic marker SNPs (a gold-standard subset of the genome), they are unable to assess the quality of the reported genome variation elsewhere (i.e., in the vast majority of the genome). Thus, in the context of QC, this specific method of achieving robustness to representational differences may be detrimental. In contrast, genome fingerprinting achieves robustness to representational differences by locality-sensitive hashing of most reported SNVs, regardless of their location in the genome and their population frequency. In this way, genome fingerprint correlations are a more sensitive and informative QC tool than robust and precise measures of degree of relatedness, enabling our analysis of the overall degraded quality of a subset of the genomes and, by adding phasing information into the fingerprints, to easily quantify phasing concordance.

Minor reference alleles are a problem for identifying and reporting clinical variants,[27,28] and discrepant ALT alleles could produce similar issues. Cross-comparison of TGP versions identifies many discrepancies of alternate alleles observed. While some reflect infrequent observations (low ACs), many discrepantly observed alleles seem to be very frequent in the population. There can be multiple causes for ALT allele discrepancies, including the presence of multiple alleles, and overlap or vicinity with indels and more

complex variants. In many cases, these can only be identified through comparison with an additional dataset. For example, an ALT-discrepant biallelic SNV between TGP38C and TGP38X may overlap by coordinate with an indel reported in TGP38N. This suggests that there is still room for improvement in variant-calling and variant-reporting pipelines, particularly in the presence of multiple alleles, indels, and low-complexity sequences. Differences in reference sequences were shown to lead to discrepancies in variant calls in the exome.[29] We observed these discrepancies to be genome-wide, with an enrichment in ALT allele discrepancies in coding regions, particularly when comparing TGP versions based on a mixture of exome and low-pass WGS versus versions based solely on 30× WGS. This suggests there may still be some difficulty integrating the two data types, and that less biased results may be obtained from comprehensive WGS, particularly using long reads.[30]

We evaluated the consistency of phasing among TGP versions using two methods: phasing-aware fingerprinting and a phasing accuracy metric, which respectively emphasize high-frequency or low-frequency switch errors.[23] These comparisons revealed significant discrepancies in phasing among the different TGP versions for most (but not all) genomes. While pairwise comparisons can reveal discrepancies, it is not always evident which one is the correct result. In the context of phasing, it is reasonable to assume that TGP38N is the better phased version, since its phasing method included multiple related individuals.

Notably, a small subset of genomes showed more consistent phasing; these genomes include those most commonly studied and used for evaluating methodologies. Current best practices for benchmarking variant calls are largely based on the use of truth set resources of the GIAB Consortium.[12–14] Specifically, TGP38X was evaluated by comparing the variant call sets observed for the platinum NA12878 genome, and computing false-positive and false-negative call rates in regions for which the GIAB considers calls to be high confidence.[6] Strathern's generalization of Goodharts Law, "When a measure becomes a target, it ceases to be a good measure,"[31] would seem to apply; since the Platinum NA12878 genome and other GIAB truth sets are used to train experimental and computational technologies, it seems inappropriate to also use them for QC purposes. We observe that such verification may be insufficient for global evaluation of large genome datasets including samples from diverse population backgrounds, which may be differentially affected by reference and software changes.

As a partial way to mitigate this deficiency, we recommend performing global dataset comparisons using genome fingerprints, and other general-purpose[32] or domain-specific metrics. Such relative benchmarking, in which each individual genome can serve as its own reference, can supplement absolute benchmarking relative to truth sets. As a result of such relative benchmarking, multiple discrepancies may become evident that cannot be immediately resolved in the absence of a truth set; resolving such discrepancies would certainly necessitate further computational analyses and, in some cases, experimental testing.

For most researchers, though, the existence of multiple versions of the TGP dataset raises an important question: what version should be used? The answer will naturally depend on the specific analysis goals and available data. For studying genomes expressed relative to GRCh37, TGP37 is the obvious choice. For genomes expressed relative to GRCh38, though, there are multiple options available, differing from each other in multiple ways (Table 1). Since the beginning of this comparison work, version TGP38L was formally withdrawn, leaving TGP38C as the version most similar to TGP37 in content, and thus perhaps yielding the most comparable results. Versions TGP38S and TGP38X are limited to biallelic content; we also found multiple quality issues in these versions. TGP38H is the only version that includes variants in centromeric regions, which may be an important benefit, or a negative, depending on the study. It is also by far the most difficult version to work with due to its size. Finally, TGP38N seems to have the most advantages, being based on comprehensive WGS resequencing of the original samples and phasing using families. We noted, though, that both TGP38C and TGP38N require some preprocessing due to their unusual representation of multiallelic variants and, in the case of TGP38N, the inclusion of 698 additional genomes that may not be needed for many studies.

Data and code availability

Genome fingerprints for all datasets are available through the genome fingerprints project website: db.systemsbiology.net/gestalt/tgpqc. Code for computing genome fingerprints is available from Github: github.com/gglusman/genome-fingerprints.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100123.

Acknowledgments

This work was supported by NIBIB grant U54 EB020406 and NIA grant U19 AG023122.

Author contributions

G.G. conceived of the study. G.G., M.R., A.J., A.V., M.M., and S.R. performed analyses. G.G. and M.R. wrote the manuscript. All authors approved the final version.

Declaration of interests

G.G. and M.R. hold a patent application (WO2017210102A1) on the method used to generate and compare reduced genome datasets.

Received: December 8, 2021
Accepted: June 2, 2022

Web resources

Project website: db.systemsbiology.net/gestalt/tgpqc

Genome fingerprints project website: db.systemsbiology.net/gestalt/genome_fingerprints

Genome fingerprints Github page: github.com/gglusman/genome-fingerprints

IGSR: International Genome Sample Resource, http://www.internationalgenome.org

liftOver tool: http://genome.ucsc.edu/cgi-bin/hgLiftOver

TGP37: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

TGP38L: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/

TGP38C: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/phase3_liftover_nygc_dir/

TGP38S: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/

TGP38X: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/

TGP38H: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/

TGP38N and TGP38Nr: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/

TGP37r: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/related_samples_vcf/

TGP38Sr: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/supporting/related_samples/

TGP38Xr: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/supporting/related_samples/

References

1. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
2. Zheng-Bradley, X., and Flicek, P. (2017). Applications of the 1000 genomes project resources. Brief. Funct. Genomics *16*, 163–170. https://doi.org/10.1093/bfgp/elw027.
3. Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D.C., and Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. Genomics *109*, 83–90. https://doi.org/10.1016/j.ygeno.2017.01.005.
4. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311. https://doi.org/10.1093/nar/29.1.308.
5. Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., and Flicek, P. (2017). Alignment of 1000 Genomes Project reads to reference assembly GRCh38. Gigascience *6*, 1–8. https://doi.org/10.1093/gigascience/gix038.
6. Lowy-Gallego, E., Fairley, S., Zheng-Bradley, X., Ruffier, M., Clarke, L., and Flicek, P. (2019). Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. Wellcome Open Res. *4*, 50. https://doi.org/10.12688/wellcomeopenres.15126.2.
7. Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. *48*, D941–D947. https://doi.org/10.1093/nar/gkz836.
8. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Preprint at bioRxiv. https://doi.org/10.1101/2021.02.06.430068.
9. Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.P., and Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics *30*, 1006–1007. https://doi.org/10.1093/bioinformatics/btt730.
10. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.Y., et al. (2015). An integrated map of structural variation in 2, 504 human genomes. Nature *526*, 75–81. https://doi.org/10.1038/nature15394.
11. Glusman, G., Mauldin, D.E., Hood, L.E., and Robinson, M. (2017). Ultrafast comparison of personal genomes via precomputed genome fingerprints. Front. Genet. *8*, 136. https://doi.org/10.3389/fgene.2017.00136.
12. Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. *32*, 246–251. https://doi.org/10.1038/nbt.2835.
13. Krusche, P., Trigg, L., Boutros, P.C., Mason, C.E., De La Vega, F.M., Moore, B.L., Gonzalez-Porta, M., Eberle, M.A., Tezak, Z., Lababidi, S., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. Nat. Biotechnol. *37*, 555–560. https://doi.org/10.1038/s41587-019-0054-x.
14. Zook, J.M., McDaniel, J., Olson, N.D., Wagner, J., Parikh, H., Heaton, H., Irvine, S.A., Trigg, L., Truty, R., McLean, C.Y., et al. (2019). An open resource for accurately benchmarking small variant and reference calls. Nat. Biotechnol. *37*, 561–566. https://doi.org/10.1038/s41587-019-0074-6.
15. Roslin, N.M., Weili, L., Paterson, A.D., and Strug, L.J. (2016). Quality control analysis of the 1000 genomes project Omni2.5 genotypes. Preprint at bioRxiv. https://doi.org/10.1101/078600.
16. Belsare, S., Levy-Sakin, M., Mostovoy, Y., Durinck, S., Chaudhuri, S., Xiao, M., Peterson, A.S., Kwok, P.Y., Seshagiri, S., and Wall, J.D. (2019). Evaluating the quality of the 1000 genomes project data. BMC Genom. *20*, 620. https://doi.org/10.1186/s12864-019-5957-x.
17. Why are there differences between the different analyses of the 1000 genomes samples?. https://www.internationalgenome.org/faq/why-are-there-differences-between-the-different-analyses-of-the-1000-genomes-samples/.

18. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.
19. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
20. Data Portal | 1000 Genomes. https://www.internationalgenome.org/data-portal/sample/NA18498.
21. Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.J., Kher, M., Banks, E., Ames, D.C., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun. *9*, 4038. https://doi.org/10.1038/s41467-018-06159-4.
22. Grc & View My Complete Profile. GenomeRef. http://genomeref.blogspot.com/2013/12/announcing-grch38.html.
23. Glusman, G., Cox, H.C., and Roach, J.C. (2014). Whole-genome haplotyping approaches and genomic medicine. Genome Med. *6*, 73. https://doi.org/10.1186/s13073-014-0073-7.
24. Schröder, J., Corbin, V., and Papenfuss, A.T. (2017). HYSYS: have you swapped your samples? Bioinformatics *33*, 596–598. https://doi.org/10.1093/bioinformatics/btw685.
25. Lee, S., Lee, S., Ouellette, S., Park, W.Y., Lee, E.A., and Park, P.J. (2017). NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. Nucleic Acids Res. *45*, e103. https://doi.org/10.1093/nar/gkx193.
26. Pedersen, B.S., Bhetariya, P.J., Brown, J., Kravitz, S.N., Marth, G., Jensen, R.L., Bronner, M.P., Underhill, H.R., and Quinlan, A.R. (2020). Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. Genome Med. *12*, 62. https://doi.org/10.1186/s13073-020-00761-2.
27. Koko, M., Abdallah, M.O.E., Amin, M., and Ibrahim, M. (2018). Challenges imposed by minor reference alleles on the identification and reporting of clinical variants from exome data. BMC Genom. *19*, 46. https://doi.org/10.1186/s12864-018-4433-3.
28. Barbitoff, Y.A., Bezdvornykh, I.V., Polev, D.E., Serebryakova, E.A., Glotov, A.S., Glotov, O.S., and Predeus, A.V. (2018). Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling. Genet. Med. *20*, 360–364. https://doi.org/10.1038/gim.2017.168.
29. Li, H., Dawood, M., Khayat, M.M., Farek, J.R., Jhangiani, S.N., Khan, Z.M., Mitani, T., Coban-Akdemir, Z., Lupski, J.R., Venner, E., et al. (2021). Exome variant discrepancies due to reference-genome differences. Am. J. Hum. Genet. *108*, 1239–1250. https://doi.org/10.1016/j.ajhg.2021.05.011.
30. Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., and Zarate, S. (2021). A complete reference genome improves analysis of human genetic variation. Preprint at bioRxiv.
31. Strathern, M. (1997). 'Improving ratings': audit in the British University system. Eur. Rev. *5*, 305. https://doi.org/10.1017/s1062798700002660.
32. Deutsch, E.W., Kramer, R., Ames, J., Bauman, A., Campbell, D.S., Chard, K., Clark, K., Arcy, M.D., Dinov, I.D., Donovan, R., et al. (2018). BDQC: a general-purpose analytics tool for domain-blind validation of big data. Preprint at bioRxiv. https://doi.org/10.1101/258822.