Quality control of large genome datasets using genome fingerprints

Max Robinson and Gustavo Glusman*

Institute for Systems Biology, 401 Terry Ave N, Seattle, WA 98109, USA

* Correspondence: Gustavo@SystemsBiology.org

Gustavo Glusman
Institute for Systems Biology
401 Terry Ave N
Seattle, WA 98109, USA
Tel: 206 732-1273

Abstract

The 1000 Genomes Project is a foundational resource to modern human biomedicine, serving as a standard reference for human genetic variation. Recently, new versions of the 1000 Genomes Project dataset were released, expressed relative to the current version of the human reference sequence (GRCh38) and partially validated by benchmarking against reference truth sets from the Genome In A Bottle Consortium. We used our ultrafast genome comparison method (genome fingerprinting) to evaluate four versions of the 1000 Genomes Project datasets. These comparisons revealed several discrepancies in dataset membership, multiple cryptic relationships, overall changes in biallelic SNV counts, and more significant changes in SNV counts, heterozygosity and genotype concordance affecting a subset of the individuals. Based on these observations, we recommend performing global dataset comparisons, using genome fingerprints and other metrics, to supplement 'best practice' benchmarking relative to predefined truth sets.

Background

Since its initial release, the 1000 Genomes Project [1] has served as the standard reference for human genetic variation, with multiple applications including population structure analyses, genotype imputation, association studies, evaluation of gene annotation, improving the reference genome itself, and much more [2]. To date, most analyses have relied on the phase 3 dataset, including 2504 individuals, mapped onto version GRCh37 (hg19) of the human reference genome, and released in 2013. We hereafter refer to this dataset as TGP37. The 2504 individuals in TGP37 were sampled from 26 populations, themselves drawn from five regions (Africa, East Asia, South Asia, Europe and the Americas). Genotypes for all individuals were estimated based on a combination of whole-genome sequencing, targeted exome sequencing and high-density SNP microarrays. The resulting variant calls included biallelic and multiallelic SNPs, indels and structural variants.

Not long after the initial release of TGP37 data, a new and much improved version of the human reference sequence, GRCh38 (hg38), was released, prompting efforts to express TGP phase 3 variation data relative to this new reference. Initially, TGP37 variants were translated to GRCh38 coordinates (via liftOver, http://genome.ucsc.edu/cgi-bin/hgLiftOver), yielding a dataset we call TGP38L. More recently, the raw genomic sequence reads were remapped onto GRCh38 [3] to support 'native' variant calling on the new reference [4]. Two versions of these variant calls have been released to date. The first, released in late 2018, was restricted to just biallelic SNVs; we call this dataset TGP38S. The second, released in early 2019, included biallelic SNVs and indels; we refer to this dataset as TGP38.

Despite the desirability of including only unrelated individuals in the TGP cohort, a number of close and more distant relationships exist within TGP37, as reported by us and others [5]. TGP37 is supplemented with a small set of 31 related samples, which we call TGP37r. Likewise, a set of 150 related samples accompanies TGP38S and TGP38 - we refer to these in turn as TGP38Sr and TGP38r.

Here, we evaluate the four versions of the TGP (TGP37, TGP38L, TGP38S and TGP38) and their associated related samples (TGP37r and TGP38r), in terms of (a) their relative composition (shared samples), (b) findings of known and cryptic relatedness as evidenced by genome fingerprint comparisons, (c) number of SNVs and level of heterozygosity observed in each individual genome, and (d) patterns of SNV loss and genotype concordance when comparing pairs of datasets.

Results

Overview

We demonstrate the application of genome fingerprints [5] for rapid evaluation of large genome datasets relative to each other on the four reported versions of the 1000 Genomes phase 3 data (Fig. 1): the original release (GRCh37), these variants lifted to GRCh38 (GRCh37L), as well as direct mappings of the reads with (GRCh38) and without (GRCh38S) indels reported (see Methods). We used genome fingerprints and other metrics to compare the SNVs reported in these genomes. Based on these analyses, we identified a number of discrepancies and quality issues, including a missing individual, additional cryptic relations, and a set of genomes with significantly fewer SNV counts.

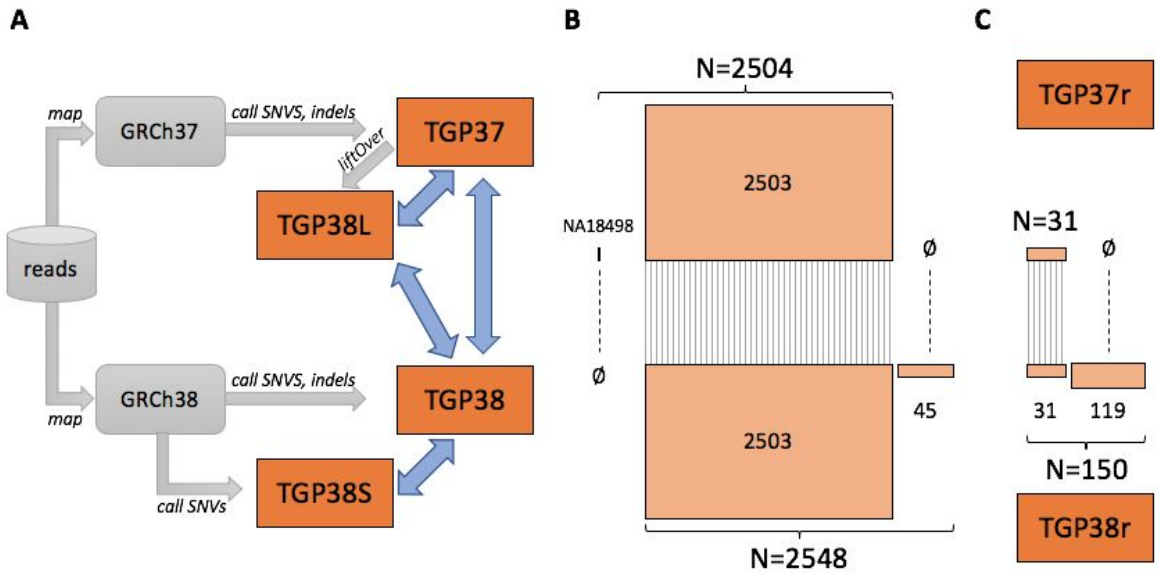

Figure 1. Overview of the datasets. A) Three methods were used to update the TGP genomes to GRCh38: liftOver (TGP38L) and remapping individual reads, followed by integrated variant calling resulting in SNVs (TGP38S) or SNV and indel calls (TGP38). These datasets were paired for comparison in four ways (blue double arrows) as discussed for each comparison. **B)** Cohort comparison. TGP37 and TGP38L contain the same 2504 genome identifiers; TGP38 and TGP38S contain the same 2548 identifiers, with 2503 identifiers in common. NA18498 is absent from TGP38/TGP38S, which contains 45 identifiers not in TGP37/TGP38L. **C)** The original set of 31 supplemental 'related samples' (TGP37r) has expanded to 150 (TGP38r).

QC evaluation: TGP37 vs. TGP38

Correlation between genome fingerprints provides a rapid means to estimate relatedness [5], and we used this tool to verify that identical genome identifiers corresponded to the same individual across datasets. TGP37 and TGP38 appear to be corresponding datasets (SNV and indels called from direct mapping of the same reads to different reference genomes), however TGP38 differs from TGP37 by omission of one genome identifier (NA18498, population YRI) and inclusion of 45 additional identifiers (Table 1).

The highest fingerprint correlation between NA18498 and any individual from TGP38 is 0.316 (HG03108, ESN); among the 150 supplemental individuals in TGP38r, the highest correlation is 0.312 (HG03373, ESN). These values are well below the 0.75 minimal correlation expected for versions of the same individual, confirming that NA18498 is indeed absent from TGP38.

We evaluated the relatedness of the 45 additional individuals in TGP38 by fingerprint comparison to TGP37 and TGP38, and observed that most (75%) seem to be related to other individuals, some with fingerprint correlations consistent with second-degree relations (Table 1).

ID	Population code	TGP38 & TGP37		TGP38, novel	
		Nearest	Correlation	Nearest	Correlation
NA18576	CHB	HG00530	0.521	NA18527	0.515
NA18527	CHB	HG00530	0.514	NA18576	0.515
HG02358	CDX	HG00530	0.506	HG02405	0.498
HG02170	CDX	HG02152	0.506	HG02173	0.498
HG02168	CDX	HG02396	0.505	NA18576	0.503
HG02169	CDX	HG01797	0.505	HG02173	0.496
HG02173	CDX	HG02389	0.502	HG02168	0.502
HG02176	CDX	HG01802	0.502	HG02170	0.497
HG02405	CDX	HG00844	0.501	HG02358	0.498
HG00134	GBR	HG00142	0.499	HG00249	0.460
NA18791	CHB	NA18960	0.498	NA18576	0.496
NA18955	JPT	NA19005	0.497	NA18576	0.493
HG00377	FIN	HG00329	0.481	NA20873	0.492
NA20873	GIH	HG00232	0.457	HG00377	0.492
NA20537	TSI	HG00232	0.469	HG00377	0.491
NA20816	TSI	NA20516	0.482	HG00377	0.472
HG00249	GBR	HG00232	0.481	HG00377	0.482
NA21121	GIH	HG04001	0.467	NA20883	0.480
NA20883	GIH	NA20854	0.462	NA21121	0.480
HG00152	GBR	NA12287	0.467	HG00377	0.470
HG00270	FIN	HG00269	0.466	HG00377	0.469
HG00104	GBR	HG00157	0.461	HG00377	0.468
HG00302	FIN	HG00373	0.466	HG00377	0.468

ID	Population code	TGP38 & TGP37		TGP38, novel	
		Nearest	Correlation	Nearest	Correlation
HG00312	FIN	HG00308	0.467	NA20537	0.467
HG00303	FIN	HG00269	0.467	HG00249	0.466
HG04301	GBR	HG00306	0.463	HG00377	0.464
HG04303	GBR	NA11930	0.464	NA20537	0.464
HG01471	CLM	HG02266	0.463	HG00377	0.453
HG00359	FIN	HG00266	0.463	HG00377	0.463
NA20831	TSI	NA11881	0.457	HG00249	0.463
NA20829	TSI	HG00232	0.463	HG00249	0.462
HG00135	GBR	HG01516	0.462	HG00249	0.458
HG00156	GBR	HG00150	0.458	NA20537	0.461
HG04302	GBR	HG00337	0.458	NA20816	0.459
HG02436	ACB	HG02433	0.422	HG00377	0.404
NA19359	LWK	NA19309	0.389	NA19044	0.331
NA19044	LWK	HG01989	0.373	HG02436	0.365
HG03398	MSL	NA19189	0.349	NA19044	0.344
HG03393	MSL	HG03376	0.348	HG03398	0.330
HG03431	MSL	NA19189	0.336	HG03398	0.330
HG03171	ESN	NA18520	0.336	NA19044	0.323
HG03549	MSL	HG02643	0.334	HG03398	0.335
NA19371	LWK	NA19452	0.332	NA19044	0.333
NA19398	LWK	HG02442	0.330	NA19044	0.332
HG03462	MSL	HG03397	0.330	HG03398	0.329

Table 1: The 45 additional individuals in TGP38, absent from TGP37. None of these individuals have any annotated relationships in IGSR, but 34 have fingerprint correlations of at least 45% to other individuals in TGP38.

Fingerprint-based comparisons of the 2503 individuals shared between TGP37 and TGP38 confirms a one-to-one relationship: for each individual in TGP37 (excluding NA18498, discussed above), the highest correlation observed was to the TGP38 individual with the same identifier. For 2495 of these, the correlation is well above 0.75, as expected. On the other hand, the remaining eight individuals all from the ACB population (Table 2), have between-set correlations in the 0.55-0.60 range, which we previously found to be consistent with first-degree relationships [5]. An additional eight individuals are more minor outliers: their fingerprint correlations (ranging from 0.787 to 0.865) are above the 0.75 cutoff for recognizing them as the same individual, but are much lower than observed for other genomes in the dataset (0.885 +/- 0.0028).

Section	ID	Population code	Fingerprint correlation	SNV count		SNVs lost	Heterozygosity		Genotype concordance
				TGP37	TGP38		TGP37	TGP38	
missing	NA18498	YRI	NA	4281373	0	100.00%	65.8%	NA	NA
strongly affected	HG02325	ACB	0.550	4281515	3331950	22.18%	66.4%	57.3%	0.578
	HG02442	ACB	0.554	4261462	3349482	21.40%	65.9%	57.4%	0.587
	HG02433	ACB	0.570	4247360	3363226	20.82%	67.0%	58.4%	0.600
	HG01989	ACB	0.571	4204133	3334302	20.69%	66.6%	58.0%	0.602
	HG02445	ACB	0.571	4197022	3322449	20.84%	66.8%	57.9%	0.601
	HG02343	ACB	0.585	4192863	3347173	20.17%	67.2%	58.6%	0.615
	HG02420	ACB	0.595	4118119	3285587	20.22%	68.8%	59.1%	0.626
	HG01988	ACB	0.603	4044546	3261145	19.37%	67.6%	58.7%	0.632
mildly affected	NA19189	YRI	0.787	4259509	3975543	6.67%	65.1%	63.9%	0.871
	HG00232	GBR	0.814	3476261	3263125	6.13%	60.8%	59.0%	0.905
	HG00530	CHS	0.824	3492719	3301091	5.49%	56.1%	54.9%	0.919
	HG00542	CHS	0.827	3487836	3292402	5.60%	56.4%	55.2%	0.923
	HG00531	CHS	0.839	3488432	3319043	4.86%	56.3%	55.5%	0.938
	NA18960	JPT	0.840	3547651	3362291	5.22%	56.9%	56.0%	0.947
	NA18856	YRI	0.862	4307564	4157373	3.49%	66.3%	66.2%	0.965
	HG00116	GBR	0.865	3507492	3379020	3.66%	60.5%	60.4%	0.969
reference	NA12878	CEU	0.885	3516562	3437816	2.24%	60.5%	61.0%	0.989
	average	ACB (n=89)	0.887	4261005	4174998	2.02%	66.9%	67.3%	0.988
	st.dev.	ACB (n=89)	0.0024	33945	35464	0.13%	0.4%	0.4%	0.002
	average	all (n=2487)	0.885	3742994	3661305	2.19%	61.4%	61.9%	0.988
	st.dev.	all (n=2487)	0.0028	327070	324064	0.17%	3.6%	3.5%	0.002

Table 2. Observed statistics for the outlier individuals most affected by dataset recomputation from TGP37 to TGP38, in comparison to the 'platinum' NA12878 genome, the 89 ACB individuals unaffected by this bioinformatic difference, and the 2487 similarly unaffected individuals in the entire cohort.

To evaluate the nature of these discrepancies, we tabulated the number of biallelic autosomal SNVs observed for each individual genome in each of the four datasets. We observed a reduction of ~2% of total SNV count for most individuals (Table 2). This reduction could be explained by changes in the reference, including reference/alternate allele switches and improved variant calling leading to fewer false positives. The 8 'mildly affected' individuals lost 3.5%-6.7% of SNVs. In contrast, the 8 ACB individuals described above lost 20-22% of SNVs - a very large reduction, not easily accountable for. While this could be a correction of variant miscalls in TGP37, it could also reflect false negative calls in TGP38. We compared the SNV counts of these eight individuals to those of the remaining 89 ACB individuals (Fig. 2) and observed that the TGP37 SNV counts of the eight strongly affected ACB individuals are consistent with the rest of this population, but they are outliers in terms of TGP38 SNV counts. We further tabulated the heterozygosity fraction for each individual and observed, again, that the eight strongly affected ACB individuals become low-heterozygosity outliers relative to their population, when transitioning from TGP37 to TGP38 (Table 2). Finally, we computed the concordance of the reported genotypes for each individual, i.e., in what fraction of SNVs the individual is deemed heterozygous in both datasets (ignoring phasing information), or homozygous for the alternate allele, out of the total number of SNVs in which the individual is not homozygous for the reference allele. We again observed markedly reduced genotype concordance for these individuals.

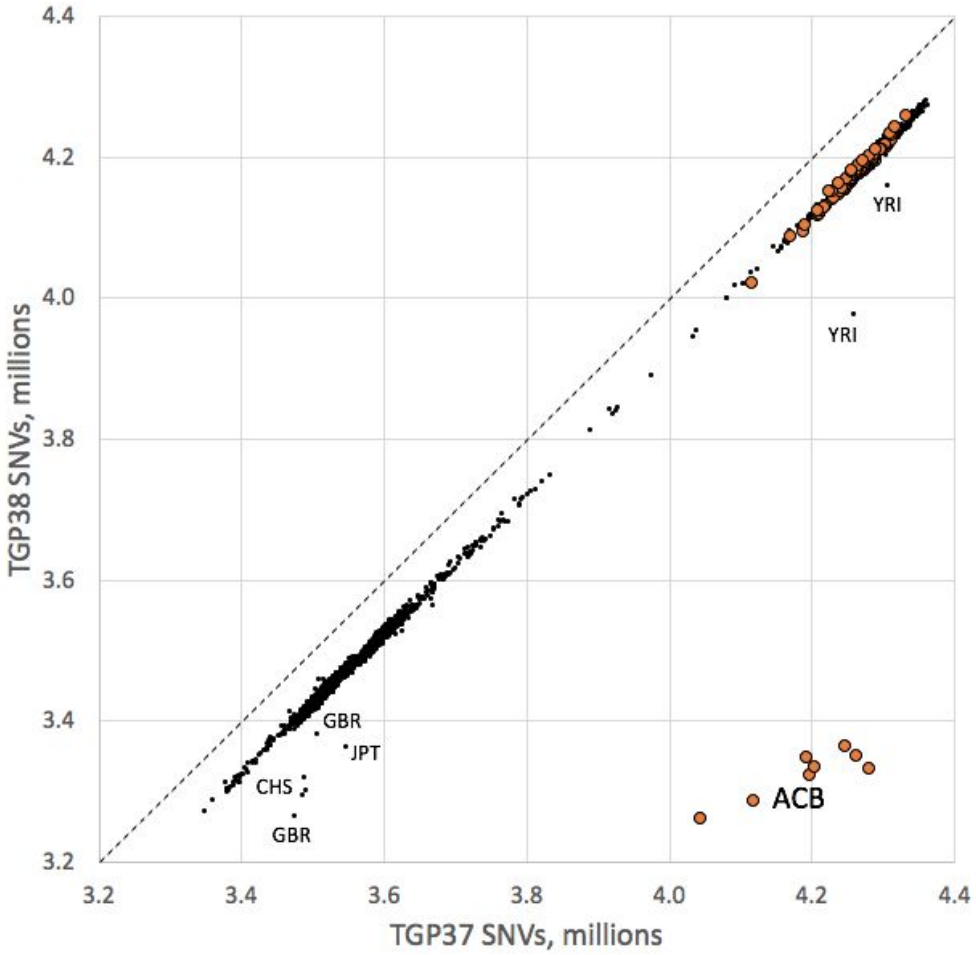

Figure 2. Biallelic autosomal SNV counts in TGP37 and TGP38. ACB individuals are highlighted with orange circles; the eight strongly affected ACB individuals are clearly most different from other ACB individuals in terms of TGP38 SNV counts. Other outliers are labeled with their assigned populations.

Evaluation of the related individuals: TGP38 vs. TGP38r

We evaluated the degree of relatedness of the 150 'related individuals' in TGP38r, expecting all of them to show some degree of relatedness to at least one of the individuals in TGP38. We computed fingerprints for all TGP38r individuals, then compared them to all TGP38 individuals and to each other (Table 3). Over two thirds of the TGP38r individuals can indeed be recognized as closely related to TGP38 individuals, with fingerprint correlations above 0.45. On the other hand, at least 28 of the TGP38r individuals seem not to be related to anyone else in TGP38 or to each other (by correlation < 0.4) and thus could have been included in the TGP38 set.

ID	Population code	TGP38r to TGP38		TGP38r	
		Nearest	Correlation	Nearest	Correlation
HG03982	STU	HG03858	0.868	HG03761	0.430
HG00512	CHS	HG00524	0.710	HG00501	0.687
HG00578	CHS	HG00581	0.707	HG00635	0.689
HG00501	CHS	HG00524	0.701	HG00512	0.687
HG02046	KHV	HG02067	0.701	HG00866	0.490
HG00577	CHS	HG00584	0.697	HG00418	0.501
HG02381	CDX	HG02373	0.694	HG00427	0.489
HG00635	CHS	HG00581	0.676	HG00578	0.689
NA20526	TSI	NA20792	0.678	NA07346	0.453
HG03948	STU	HG03673	0.674	HG04053	0.441
HG00983	CDX	HG00978	0.665	HG00866	0.511
HG02372	CDX	HG02371	0.659	HG00866	0.502
HG02024	KHV	HG02026	0.659	HG00427	0.495
HG00702	CHS	HG00657	0.656	HG00866	0.496
HG00866	CDX	HG00867	0.645	HG00427	0.515
NA20898	GIH	NA20886	0.635	HG03723	0.444
NA20871	GIH	NA20868	0.631	HG03723	0.452
NA12891	CEU	NA12878	0.626	NA12892	0.464
HG00153	GBR	HG00158	0.624	NA11993	0.465
HG00733	PUR	HG00732	0.623	NA11993	0.443
NA19685	MXL	NA19661	0.623	NA19660	0.611
NA19675	MXL	NA19678	0.616	NA11993	0.446
HG03715	ITU	HG03713	0.616	HG03723	0.448
NA12892	CEU	NA12878	0.612	NA11993	0.474
NA19660	MXL	NA19664	0.513	NA19685	0.611
NA20336	ASW	NA20334	0.603	NA11993	0.350
NA19373	LWK	NA19374	0.600	NA19150	0.334
NA20341	ASW	NA20289	0.597	HG01274	0.331
NA19470	LWK	NA19443	0.596	NA19469	0.538
NA19396	LWK	NA19397	0.588	NA19444	0.330

ID	Population code	TGP38r to TGP38		TGP38r	
		Nearest	Correlation	Nearest	Correlation
HG03383	MSL	HG03391	0.432	HG03508	0.323
HG01452	CLM	HG01961	0.430	HG01983	0.428
HG03454	MSL	HG03457	0.389	NA19150	0.335
NA20344	ASW	NA20340	0.371	HG01274	0.340
HG03574	MSL	HG03556	0.370	HG03373	0.330
HG03566	MSL	HG03547	0.364	HG03454	0.329
NA19150	YRI	HG02325	0.363	NA20361	0.347
HG03034	GWD	HG02464	0.334	HG03033	0.349
HG03033	GWD	HG02861	0.342	HG03034	0.349
HG03339	ESN	HG03366	0.342	HG03361	0.347
HG03361	ESN	HG03342	0.334	HG03339	0.347
HG02762	GWD	HG02643	0.345	HG03373	0.335
HG03250	GWD	HG03246	0.345	HG03249	0.340
HG03493	ESN	HG03511	0.345	HG03373	0.335
HG03373	ESN	NA19130	0.344	NA19150	0.342
HG03306	ESN	HG03514	0.344	HG03307	0.332
HG03307	ESN	HG03398	0.343	NA19150	0.342
HG03312	ESN	NA19147	0.343	HG03249	0.336
HG03309	ESN	HG03193	0.343	NA18487	0.341
HG03249	GWD	HG02678	0.343	HG03250	0.340
NA18487	YRI	NA18934	0.339	HG03309	0.341
HG02869	GWD	HG02896	0.341	NA20361	0.336
HG02965	ESN	HG02981	0.341	HG03249	0.335
HG03508	ESN	HG03199	0.339	HG02964	0.334
HG03076	MSL	NA19129	0.339	HG03569	0.335
HG02964	ESN	HG03193	0.339	HG03508	0.334
HG02478	ACB	NA18858	0.338	NA19150	0.337
HG03582	MSL	HG03398	0.333	NA19150	0.337
HG03569	MSL	NA18915	0.336	HG03076	0.335
HG03408	MSL	HG03115	0.334	NA19150	0.330

Table 3. Sixty of 150 related individuals in TGP38r. Left side: top 30 by similarity to TGP38 or to other TGP38r individuals, starting with the HG03982-HG03858 pair discussed in the text. Right side: bottom 30 by similarity to TGP38 or to other TGP38r individuals, showing correlations consistent with no close family relationships.

One of the 'related individuals' in TGP38r (HG03982, STU) has fingerprint correlation of 0.868 to an individual in TGP38 (HG03858, STU). This fingerprint correlation would suggest these are the same individual, and yet HG03858 is annotated as female in IGSR, but HG03982 is annotated as male in IGSR. There is no annotation that either of these individuals having any relatives in either dataset, nor can 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. We evaluated whether HG03982 could indeed be a male sample as annotated. No chrY data were released for TGP38r, and chrX data are available only in the pseudoautosomal regions (PARs, which combine data from chrX and chrY). We compared the genotype calls in the PARs of HG03858 and HG03982 and observed 91.8% genotype concordance, consistent with these being the same person. We evaluated the coverage levels along chrY for both samples (from low-coverage data) and found that both are consistent with the absence of chrY. We further computed chromosome-specific fingerprints, including for the PARs. The resulting 0.928 correlation of PAR-specific fingerprints suggests these two samples have the same karyotype

(XX) and the same chrX haplotypes, consistent with being sisters or the same individual. For comparison, we observe 0.954 correlation of PAR-specific fingerprints of samples HG00578 and HG00635 (both 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).
We conclude that these two samples are both genetically female. Lacking further information about the individual(s), we hypothesize that HG03982 may have been annotated as male as a result of a clerical error.

Other dataset comparisons

We extended the genome fingerprint correlation analysis of the final datasets (TGP37 vs. TGP38, described above) to evaluate (1) the effect of lifting over variants from one reference version to another (TGP37 vs. TGP38L), (2) the concordance of such lifting with native mapping and variant calling on the new reference (TGP38L vs. TGP38) and (3) the effect of variant calling retaining only biallelic SNVs or both biallelic SNVs and indels (TGP38S vs. TGP38). These four comparisons yield quite distinct distributions of correlations, with variable numbers of outliers (Fig. 3).

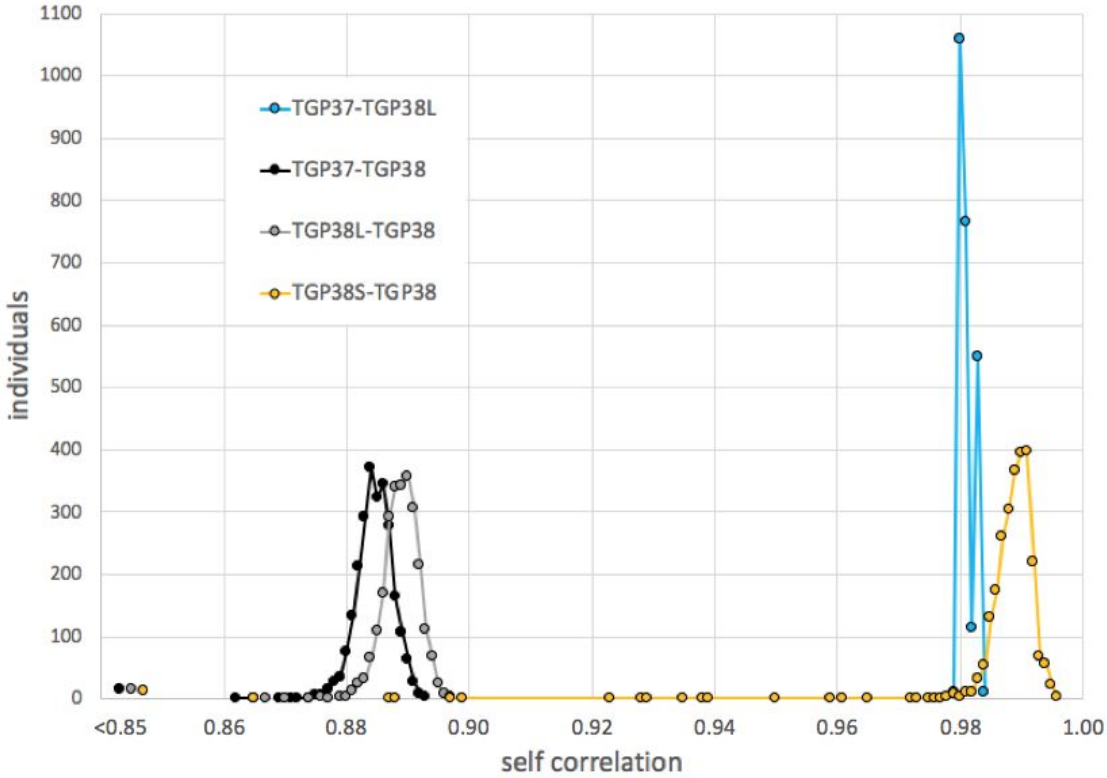

Figure 3. Distributions of self-correlations across datasets.

As expected, lifting variants over from one reference to the other yields the most uniform results (blue curve in Fig. 3), with no outliers. The ~2% loss in correlation is largely due to some degree of variant loss in regions that could not be lifted over, and a small rate of ‘reference switches’ in which the alternate allele in GRCh37 becomes the reference allele in GRCh38.

The concordance between variant lifting and native mapping (gray curve in Fig. 3) is similar to (and slightly higher than) that of the final datasets (black curve in Fig. 3), with the same set of outliers. The slightly higher concordance can again be attributed to a more consistent set of variants included, and fewer reference/alternate allele discrepancies.
When comparing the two versions of native variant calling on the new reference, with or without indels, the correlations are highest as expected (orange curve in Fig. 3). Surprisingly, this comparison yields the largest number of outliers (32 individuals), including all outliers observed when comparing TGP37 and TGP38 (16 individuals). These 32 outliers have significantly reduced genome fingerprint correlations and genotype concordances (Table 4).

Section	ID	Population code	Fingerprint correlation	SNV count		SNVs lost	Heterozygosity		Genotype concordance
				TGP38S	TGP38		TGP38S	TGP38	
strongly affected	HG02436	ACB	0.727	3496464	3349895	4.19%	58.8%	57.5%	0.689
	HG02325	ACB	0.730	3481154	3331950	4.29%	58.6%	57.3%	0.683
	HG02442	ACB	0.731	3499931	3349482	4.30%	58.7%	57.4%	0.686
	HG01989	ACB	0.739	3481183	3334302	4.22%	59.2%	58.0%	0.696
	HG02433	ACB	0.740	3508051	3363226	4.13%	59.6%	58.4%	0.696
	HG02445	ACB	0.745	3459846	3322449	3.97%	59.1%	57.9%	0.697
	HG02343	ACB	0.749	3489501	3347173	4.08%	59.9%	58.6%	0.704
	HG01988	ACB	0.763	3389432	3261145	3.78%	60.2%	58.7%	0.715
	HG02420	ACB	0.764	3409560	3285587	3.64%	60.7%	59.1%	0.713
	HG00377	FIN	0.807	3083904	3014303	2.26%	54.6%	53.7%	0.747
	NA19044	LWK	0.815	3919241	3796612	3.13%	63.9%	63.0%	0.799
	NA20873	GIH	0.825	3193470	3112562	2.53%	56.7%	55.5%	0.777
	NA20537	TSI	0.865	3272135	3192729	2.43%	58.0%	56.5%	0.838
	NA19189	YRI	0.887	4055238	3975543	1.97%	64.5%	63.9%	0.885
	NA18527	CHB	0.888	3286549	3217439	2.10%	55.0%	53.8%	0.863
	NA20883	GIH	0.897	3374736	3310653	1.90%	59.3%	58.2%	0.885
	NA21121	GIH	0.899	3368293	3303995	1.91%	59.4%	58.2%	0.885
mildly affected	HG00232	GBR	0.923	3308075	3263125	1.36%	59.8%	59.0%	0.919
	HG03398	MSL	0.928	4178239	4135437	1.02%	66.0%	65.6%	0.933
	HG00249	GBR	0.929	3314289	3272920	1.25%	59.8%	59.1%	0.923
	HG00530	CHS	0.935	3332144	3301091	0.93%	55.4%	54.9%	0.932
	HG00542	CHS	0.938	3325805	3292402	1.00%	55.7%	55.2%	0.937
	NA18576	CHB	0.939	3357895	3323353	1.03%	56.4%	55.9%	0.936
	HG00531	CHS	0.950	3343457	3319043	0.73%	55.8%	55.5%	0.950
	NA18960	JPT	0.959	3383189	3362291	0.62%	56.2%	56.0%	0.961
	NA20816	TSI	0.961	3390396	3372127	0.54%	60.6%	60.3%	0.965
	NA18856	YRI	0.965	4178400	4157373	0.50%	66.3%	66.2%	0.969
	HG00116	GBR	0.972	3394446	3379020	0.45%	60.5%	60.4%	0.975
	NA19468	LWK	0.973	4177198	4170318	0.16%	66.6%	66.5%	0.980
	NA19436	LWK	0.975	4177601	4170177	0.18%	66.8%	66.8%	0.980
	NA19456	LWK	0.976	4182394	4176973	0.13%	66.8%	66.8%	0.982
	NA19431	LWK	0.977	4191962	4185700	0.15%	66.5%	66.5%	0.982
reference	NA12878	CEU	0.991	3439095	3437816	0.04%	61.1%	61.0%	0.992
	average	ACB (n=88)	0.989	4176746	4174998	0.04%	67.3%	67.3%	0.991
	st.dev.	ACB (n=88)	0.0028	35037	35464	0.03%	0.4%	0.4%	0.002
	average	all (n=2516)	0.989	3662368	3660580	0.05%	61.9%	61.9%	0.991
	st.dev.	all (n=2516)	0.0027	324263	324333	0.04%	3.5%	3.5%	0.002

Table 4. Observed statistics for the outlier individuals most affected by variant calling including or excluding indels (TGP38S vs. TGP38), in comparison to the 'platinum' NA12878 genome, the 88 ACB individuals unaffected by this bioinformatic difference, and the 2516 similarly unaffected individuals in the entire cohort.

Discussion

We presented here the application of genome fingerprints [5] for quick and simple comparison of four versions of the TGP, expressed relative to two versions of the reference genome (GRCh37 and GRCh38). In addition to the overall comparison of the full datasets (TGP37 vs. TGP38, and related samples), other pairwise comparisons of these versions provided insights into the effects of lifting over variants from one reference version to the other (TGP37 vs. TGP38L), of lifting over vs. native mapping and variant calling (TGP38L vs. TGP38), and of different variant calling procedures (TGP38 vs. TGP38S). Through these comparisons, we identified some discrepancies between the datasets, pointing at changes in the list of included genomes, some additional cryptic relationships, overall changes in biallelic SNV counts, and more significant changes in SNV counts, heterozygosity and genotype concordance affecting a subset of the individuals.

Best practices for benchmarking variant calls are largely based on the use of 'truth set' resources of the Genome In A Bottle (GIAB) Consortium [6–8]. Specifically, TGP38 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 [4]. 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 [9] 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.

Materials and Methods

Datasets. We obtained four versions of the 1000 genomes dataset, phase 3:

1. **TGP37**: Variant calls relative to the GRCh37 (hg19) version of the human genome reference (N=2504). ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
2. **TGP38L**: Variant calls for the same set of genomes, "lifted over" to the GRCh38 (hg38) version of the human reference and using dbSNP v. 149 (N=2504). http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/
3. **TGP38**: A set of integrated phased biallelic SNP and indel calls, directly called against the GRCh38 (hg38) version of the human reference (N=2548). http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/
4. **TGP38S**: Integrated phased biallelic SNP calls, directly called against the GRCh38 (hg38) version of the human reference (N=2548). http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/

We also downloaded three sets of genomes of samples related to those in the main 1000 genomes dataset:

1. **TGP37r**: Integrated phased biallelic SNP and indel calls, relative to GRCh37 (N=31). http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/supporting/related_samples/
2. **TGP38r**: Integrated phased biallelic SNP and indel calls, relative to GRCh38 (N=150). http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/supporting/related_samples/
3. **TGP38Sr**: Integrated phased biallelic SNP calls, relative to GRCh38 (N=150). http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/supporting/related_samples/

Whole-genome fingerprinting. We computed fingerprints for all genomes in all sets as described [5], 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.

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 *heterozygosity* of an individual as the number of biallelic SNVs observed in their genome in heterozygous state, divided by their SNV count. We computed the *genotype concordance* of an individual between two datasets as the number of biallelic SNVs in which the individual is heterozygous in both two 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.

Availability. Genome fingerprints (L=200) for all datasets are available through the genome fingerprints project website, db.systemsbiology.net/gestalt/genome_fingerprints. Code for computing genome fingerprints is available from github.com/gglusman/genome-fingerprints.

Abbreviations used

GIAB: Genome In A Bottle Consortium
GRCh37, GRCh38: Genome Reference Consortium, human reference versions 37 and 38
IGSR: International Genome Sample Resource, http://www.internationalgenome.org
SNV: Single-nucleotide variant
TGP: Thousand Genomes Project
TGP37, TGP37r, TGP38, TGP38L, TGP38r, TGP38S: The TGP datasets studied in this work

Author contributions

GG conceived of the study. GG, MR performed analyses. GG, MR wrote the manuscript. All authors edited and approved its final version.

Conflict of Interest Statement

GG and MR hold a patent application (WO2017210102A1) on the method used to generate and compare reduced genome datasets. GG holds stock options in Arivale, Inc. Arivale, Inc. did not fund the study and was not involved in its design, implementation, or reporting.

References

1. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

2. Zheng-Bradley X, Flicek P. Applications of the 1000 Genomes Project resources. Brief Funct Genomics. 2017;16:163–70.

3. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, et al. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. Gigascience. 2017;6:1–8.

4. Lowy-Gallego E, Fairley S, Zheng-Bradley X, Ruffier M, Clarke L, Flicek P, et al. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. Wellcome Open Res. 2019;4:50.

5. Glusman G, Mauldin DE, Hood LE, Robinson M. Ultrafast Comparison of Personal Genomes via Precomputed Genome Fingerprints. Front Genet. 2017;8:136.

6. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;32:246–51.

7. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol [Internet]. 2019; Available from: http://dx.doi.org/10.1038/s41587-019-0054-x

8. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, et al. An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol [Internet]. 2019; Available from: http://dx.doi.org/10.1038/s41587-019-0074-6

9. Deutsch EW, Kramer R, Ames J, Bauman A, Campbell DS, Chard K, et al. BDQC: a general-purpose analytics tool for domain-blind validation of Big Data [Internet]. bioRxiv. 2018 [cited 2019 Apr 4]. p. 258822. Available from: https://www.biorxiv.org/content/10.1101/258822v1