StableLift: Optimized Germline and Somatic Variant Detection Across Genome Builds

Nicholas K. Wang[1,2,3], Nicholas Wiltsie[1,2,3], Helena K. Winata[1,2,3], Sorel Fitz-Gibbon[1,2,3], Alfredo E. Gonzalez[1,2,3], Nicole Zeltser[1,2,3], Raag Agrawal[1,2,3], Jieun Oh[1,2,3], Jaron Arbet[1,2,3,4], Yash Patel[1,2,3], Takafumi N. Yamaguchi[1,2,3], Paul C. Boutros[1,2,3,4,#]

[1]Department of Human Genetics, University of California, Los Angeles
[2]Jonsson Comprehensive Cancer Center, University of California, Los Angeles
[3]Institute for Precision Health, University of California, Los Angeles
[4]Department of Urology, University of California, Los Angeles
Corresponding author:
Center for Health Sciences
10833 Le Conte Avenue
Los Angeles, CA
United States of America
90095
Email: pboutros@mednet.ucla.edu

Abstract

Reference genomes are foundational to modern genomics. Our growing understanding of genome structure leads to continual improvements in reference genomes and new genome "builds" with incompatible coordinate systems. We quantified the impact of genome build on germline and somatic variant calling by analyzing tumour-normal whole-genome pairs against the two most widely used human genome builds. The average individual had a build-discordance of 3.8% for germline SNPs, 8.6% for germline SVs, 25.9% for somatic SNVs and 49.6% for somatic SVs. Build-discordant variants are not simply false-positives: 47% were verified by targeted resequencing. Build-discordant variants were associated with specific genomic and technical features in variant- and algorithm-specific patterns. We leveraged these patterns to create StableLift, an algorithm that predicts cross-build stability with AUROCs of 0.934 ± 0.029. These results call for significant caution in cross-build analyses and for use of StableLift as a computationally efficient solution to mitigate inter-build artifacts.

Main

Since initial assembly of the human genome in 2001[1,2], thousands of errors have been corrected, polymorphic regions have been defined and the diversity of included individuals has expanded[3–5]. These advances have led to a series of updated human reference genome "builds", each with incompatible coordinate numbering. While new builds are more accurate representations, their adoption can be slow in both research and clinical settings[6].

One key factor slowing adoption of new genome builds is computational cost: re-aligning sequencing data requires local storage of raw reads and investment of substantial compute time. To avoid these time and financial costs, tools have been created to convert or "liftover" genomic coordinates between builds[7,8]. Despite widespread use, coordinate conversion using these tools was designed for larger intervals and can introduce artifacts when applied to individual variant calls[9–17]. It remains unclear whether and what biases are introduced by coordinate conversion, especially in the context of structural and somatic variant detection.

To fill this gap, we compared DNA whole genome sequencing (WGS) alignment and variant detection on the two most widely used reference genomes: GRCh37 and GRCh38 (**Figure 1a**). Fifty human tumour-normal WGS pairs were analyzed on both builds using identical tools and software versions *via* standardized Nextflow pipelines (**Supplementary Figure 1a; Supplementary Table 1**)[18–20]. Variants detected from sequencing data aligned to GRCh37 were converted to GRCh38 coordinates using BCFtools/liftover[21] with UCSC chain files[22,23]. Converted GRCh37 variants were compared to variants detected from sequencing data directly aligned to GRCh38. We evaluated four variant classes: germline single nucleotide polymorphisms (gSNPs, including indels), germline structural variants (gSVs), somatic single nucleotide variants (sSNVs, including indels) and somatic structural variants (sSVs).

Most germline SNPs and structural variants identified were shared between the two builds (>93%; **Figure 1b-c**). Nevertheless, we detected 166,704 ± 14,829 build-specific gSNPs and 908 ± 73 build-specific gSVs per individual (mean ± standard deviation; **Figure 1d**). Alignment to GRCh38 led to identification of more gSNPs and gSVs (**Figure 1e**). By contrast, somatic variant detection was dramatically more variable: only 82% of sSNVs and 53% of sSVs were identified in both builds (**Figure 1f-g**). This led to 3,611 ± 2,025 build-specific sSNVs and 93 ± 61 build-specific sSVs (**Figure 1h**), with more somatic variants identified when aligning to GRCh38 (**Figure 1i**).

To better characterize build-specific calls, we calculated three complementary metrics of genotype concordance. First, we assessed non-reference discordance (NRD), which is the fraction of all non-reference genotypes that disagree between builds. Next, we considered direct variant calling on GRCh38 as ground truth and calculated false positive rate (FPR) and false negative rate (FNR). Consistent with variant detection numbers, all three metrics of genotype concordance were substantially better for germline than somatic variants: 3.8 ± 0.0% NRD for gSNPs and 8.6 ± 0.1% for gSVs *vs.* 25.9 ± 11.0% for sSNVs and 49.6 ± 11.2% for sSVs (per individual mean ± standard deviation; **Figure 1j**). The high FNR of somatic variant detection on GRCh37 (20.4 ± 9.5% sSNVs, 38.1 ± 11.0% sSVs; **Figure 1j**) suggests that the many published studies aligning to GRCh37 may systematically underestimate somatic mutation burden (or alternatively those aligning to GRCh38 may overestimate it).

To understand whether these discordances are randomly distributed, we first evaluated different classes of gSVs. Deletions and insertions were less discordant between builds than duplications,

inversion and translocations (**Figure 1k**). The high FNR of duplications (35.2 ± 7.7%) suggested increased sensitivity in GRCh38 potentially due to improved resolution of duplicated or homologous regions. This led us to investigate whether discordance in germline SNPs also varied spatially across the genome. Consistent with the gSV results, we observed significant heterogeneity in build-specific differences within and across chromosomes (**Figure 1l**). For example, a one Mbp region of 6p21.3 in the HLA region contained 16,784 gSNPs with mean 8.5% NRD, while a neighboring one Mbp region had 8,626 gSNPs with mean 1.2% NRD.

A wide range of other features are associated with discordance across builds (**Figure 1m**; **Supplementary Figure 1-7**). As an example, discordant sSNVs were more likely to have lower quality scores but higher GC content (**Figure 1m**; **Supplementary Figure 2a,c**). Discordant sSNVs also exhibited a non-monotonic association with coverage: both atypically-low and atypically-high coverage was associated with increased discordance, possibly due to erroneous mapping to homologous or repetitive regions (**Supplementary Figure 2b**). sSNVs with higher somatic allele frequencies tended to be less discordant, while variants seen at higher allele frequencies in TOPMed[24] were more likely to be discordant (**Figure 1m**; **Supplementary Figure 2d-e**). Discordance rates varied significantly across chromosomes (mean NRD ranging from 6.3% on chromosome 13 to 47.8% on chromosome Y; **Supplementary Figure 6a**) and trinucleotide contexts (mean NRD ranging from 4.7% to 17.3%; **Supplementary Figure 6d**). sSNVs in satellite repeat regions were particularly discordant (mean 59.8% NRD; **Supplementary Figure 6e**), supportive of repetitive regions as a major source of discordance.

One natural explanation of these results is that almost all build-discordant genetic variation results from false-positive predictions from variant-detection algorithms. To quantify this, we exploited targeted deep-sequencing validation (mean 653x coverage) on sSNV calls from five tumour-normal, whole genome pairs (**Supplementary Table 2**)[25]. Build-concordant variants had a validation rate of 93.3% (**Figure 1n**). Nevertheless, 34.6% of GRCh37-specific variants and 51.3% of GRCh38-specific variants were validated by targeted deep-sequencing. This is a clear enrichment of false-positives relative to build-concordant variants, but demonstrates that build-specific variants are a balance of false-positives and false-negative predictions. As a result, simply using the latest genome build is insufficient: one third of variants detected on GRCh37 but not in GRCh38 are false-negatives.

To quantify the cross-build stability of any individual variant, we created a machine-learning approach called StableLift. By leveraging features associated with build-discordance (**Supplementary Figures 1-7**), StableLift estimates the likelihood ("Stability Score") that a given variant will be consistently represented across two genome builds (**Figure 2a**). We trained StableLift with variants detected from the same fifty tumour-normal WGS pairs using six variant callers spanning all four variant-types: HaplotypeCaller[26], MuTect2[27], Strelka2[28], SomaticSniper[29], MuSE2[30] and DELLY2[31]. We validated StableLift in 10 tumour-normal whole genomes[32] (**Supplementary Table 3**) and 60 tumour-normal exomes[32] (**Supplementary Table 4**) for area under the receiver operating characteristic curve (AUROC) and selected a default operating point to maximize F_1-score in the whole genome validation set.

StableLift robustly identified build-discordant gSNP calls, with validation AUROCs of 0.958 for WGS and 0.941 for exome sequencing (**Figure 2b; Supplementary Figure 8a-c**). At the F_1-maximizing operating point, 49.7 ± 0.5% of discordant gSNPs in WGS validation were discarded, corresponding to 51,181 ± 4,884 discordant variants removed per individual (**Figure 2c**). A variety of features contributed to the accuracy of these predictions, most notably TOPMed[24] population allele frequency (**Figure 2d**) driven

by elevated discordance of variants with allele frequencies near zero (rare variants/singletons) or one (reference artifacts; **Supplementary Figure 1e**).

StableLift similarly identified build-discordant sSNVs, with validation AUROCs of 0.890 for WGS and 0.851 for exome sequencing (MuTect2; **Figure 2e; Supplementary Figure 8d-f**) and a 45.7 ± 11.7% reduction of discordant calls (-209 ± 56 discordant sSNVs; **Figure 2f**). sSNV stability prediction was driven by a wide range of predictor features (**Figure 2g**). Models fit to three other sSNV callers achieved similar performance: $AUROC_{WGS}$ = 0.932 for Strelka2, $AUROC_{WGS}$ = 0.964 for SomaticSniper and $AUROC_{WGS}$ = 0.905 for MuSE2 (**Supplementary Figure 9a-i, Supplementary Figure 10**). Different sSNV calling algorithms had similar but not identical patterns of feature importance, highlighting the interaction between genomic features and variant detection algorithms (**Supplementary Figure 9j**).

To understand how predicted variant stability relates to variant validation status, we ran StableLift on the previously described five whole genome pairs with targeted deep-sequencing validation (**Supplementary Figure 11a**). sSNVs predicted to be "Stable" were 1.3-9.6x more likely to validate than those predicted to be "Unstable" (**Supplementary Figure 11b-c**). Similarly, the Stability Score distribution was higher for validated *vs.* unvalidated variants (**Supplementary Figure 11d-g**).

Finally, we applied StableLift to structural variant calls made by DELLY2[31]. Despite only 28,350 concordant cases and 734 discordant cases of gSV training data (**Figure 1c**), StableLift again accurately identified discordant calls, with a validation AUROC of 0.926 (**Figure 2h**) and a 56.2 ± 5.3% reduction of discordant calls (-63 ± 10 discordant gSVs; **Figure 2i**). Length of variant was the most important single feature, with a range of predictive features differing from those driving the accuracy of the gSNP and sSNV models (**Figure 2j**). Accuracy in DELLY2 sSVs was equally high, achieving a validation AUROC of 0.961 (**Figure 2k**) and removing 81.7% of discordant sSVs (-171 ± 170 discordant sSVs; **Figure 2l**). Only 4,907 concordant and 1,845 discordant training cases were needed for this model, and its accuracy was driven by read count and SV length (**Figure 2m**).

This work calls for significant caution in cross-build analyses. GRCh37 remains in routine use and while re-alignment to GRCh38 is preferable, this is computationally expensive. In many cases realignment may not be possible: raw data or software pipelines may no longer be available, particularly for older technologies. Similarly, variant databases created with GRCh37 coordinates can introduce challenges in annotating newer GRCh38-derived results. StableLift can create models to convert between any two genome builds. While our results focused on converting GRCh37 results to GRCh38, we provide models of similar accuracy for the inverse conversion of GRCh38 to GRCh37 (**Supplementary Figure 12-16**).

StableLift provides an attractive approach to mitigate bias in many cases, but the build-sensitivity of somatic and structural variant calling warrants increased attention from algorithm developers. Some biases appear to be systematic, and while GRCh38 calls are generally more accurate, we identified apparent false-negatives with both genome builds. As genetic analyses gradually transition from linear reference genomes to graph-based pangenomes[33–38], quantifying build-specific variation and efficiently minimizing error rates in cross-build conversion will become increasingly important.

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

3. Church, D. M. *et al.* Modernizing Reference Genome Assemblies. *PLoS Biol.* **9**, e1001091 (2011).

4. Church, D. M. *et al.* Extending reference assembly models. *Genome Biol.* **16**, 13 (2015).

5. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

6. Lansdon, L. A. *et al.* Factors Affecting Migration to GRCh38 in Laboratories Performing Clinical Next-Generation Sequencing. *J. Mol. Diagn. JMD* **23**, 651–657 (2021).

7. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).

8. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).

9. Guo, Y. *et al.* Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **109**, 83–90 (2017).

10. Zheng-Bradley, X. *et al.* Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *GigaScience* **6**, 1–8 (2017).

11. Gao, G. F. *et al.* Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst.* **9**, 24-34.e10 (2019).

12. Lowy-Gallego, E. *et al.* Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res.* **4**, 50 (2019).

13. Pan, B. *et al.* Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics* **20**, 101 (2019).

14. Luu, P.-L., Ong, P.-T., Dinh, T.-P. & Clark, S. J. Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data. *NAR Genomics Bioinforma.* **2**, lqaa054 (2020).

15. Li, H. *et al.* Exome variant discrepancies due to reference-genome differences. *Am. J. Hum. Genet.* **108**, 1239–1250 (2021).

16. Park, K.-J., Yoon, Y. A. & Park, J.-H. Evaluation of Liftover Tools for the Conversion of Genome Reference Consortium Human Build 37 to Build 38 Using ClinVar Variants. *Genes* **14**, 1875 (2023).

17. Ormond, C., Ryan, N. M., Corvin, A. & Heron, E. A. Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Brief. Bioinform.* **22**, bbab069 (2021).

18. Fraser, M. *et al.* Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364 (2017).

19. Patel, Y. *et al.* NFTest: automated testing of Nextflow pipelines. *Bioinformatics* **40**, btae081 (2024).

20. Patel, Y. *et al.* Metapipeline-DNA: A comprehensive germline & somatic genomics Nextflow pipeline. *BioRxiv* (2024) doi:https://doi.org/10.1101/2024.09.04.611267.

21. Genovese, G. *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics* **40**, btae038 (2024).

22. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).

23. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).

24. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

25. Aaltonen, L. A. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

26. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

27. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

28. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).

29. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).

30. Ji, S., Montierth, M. D. & Wang, W. MuSE: A Novel Approach to Mutation Calling with Sample-Specific Error Modeling. *Methods Mol. Biol. Clifton NJ* **2493**, 21–27 (2022).

31. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinforma. Oxf. Engl.* **28**, i333–i339 (2012).

32. Abeshouse, A. *et al.* Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* **171**, 950-965.e28 (2017).

33. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).

34. Rakocevic, G. *et al.* Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**, 354–362 (2019).

35. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).

36. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).

37. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

38. Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* 1–11 (2023) doi:10.1038/s41587-023-01793-w.

39. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (2019). doi:10.1109/IPDPS.2019.00041.

40. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).

41. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

42. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).

43. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

44. Cingolani, P. *et al.* Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* **3**, 35 (2012).

45. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

46. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).

47. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

48. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

49. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).

50. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).

51. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).

52. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

53. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**, 264–278 (2012).

54. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).

55. P'ng, C. *et al.* BPG: Seamless, automated and interactive visualization of scientific data. *BMC Bioinformatics* **20**, 42 (2019).

56. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, 35 (2011).

57. Hao, Z. *et al.* RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* **6**, e251 (2020).

Online Methods

Analysis cohort

To assess LiftOver concordance in a representative cancer genomics workflow, we chose to evaluate a cohort of 50 patients spanning eight cancer types from the International Cancer Genome Consortium (ICGC PRAD-CA)[18] and the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium[25] (**Supplementary Table 1**). All patients had paired tumour-normal whole-genome sequencing with germline and somatic coverage of 32±8x and 57±10x, respectively.

Alignment and variant calling

Sequencing reads were aligned to the GRCh37 (hs37d5) and GRCh38 (hg38) reference builds using BWA-MEM2 (v2.2.1)[39] in paired-end, alt-aware mode followed by GATK's `MarkDuplicatesSpark` (v4.2.4.1)[26] (**Supplementary Figure 1a**). Indel realignment and base quality score recalibration were performed using GATK's `IndelRealigner` (v3.7.0), `BaseRecalibrator` (v4.2.4.1), and `ApplyBQSR` (v4.2.4.1)[26]. Germline SNPs were called using GATK's `HaplotypeCaller` (v4.2.4.1) in GVCF mode followed by variant quality score recalibration using `VariantRecalibrator` (v4.2.4.1) and 'ApplyVQSR' (v4.2.4.1) and joint genotyping across all normal samples using `GenotypeGVCFs` (v4.2.4.1)[26]. Somatic SNVs were called using MuTect2 (v4.2.4.1)[27] in tumour-normal mode with default parameters. Germline and somatic SVs were called using DELLY2 (v1.2.6)[31] with default parameters and a more stringent minimum paired-end mapping quality threshold of 20. Germline SVs were regenotyped using the output of `delly merge` and filtered with `delly filter -f germline` (v1.2.6)[31].

All alignment and variant calling operations were run on a Slurm high-performance computing cluster using Nextflow (v23.04.2) pipelines[19,20,40] to ensure reproducibility and compatibility across computing environments. The GRCh37 and GRCh38 analysis pipelines used identical parameters except for the reference genome input and associated resource files.

LiftOver coordinate conversion

GRCh37 SNV calls were converted to GRCh38 coordinates using the BCFtools/liftover plugin (v1.20)[21] with UCSC chain files[22,23]. For SVs, a custom R script was used to convert variants by breakpoint (CHROM, POS, END for DEL, DUP, INS, INV variants; CHROM, POS, END, CHR2, POS2 for BND variants) using the UCSC chain files along with the rtracklayer (v1.62.0)[41] and GenomicRanges (v1.54.1)[42] R packages.

Variant concordance

SNV concordance was evaluated at the cohort level using `vcf-compare` from VCFtools (v0.1.16)[43] and at the sample level using `SnpSift concordance` (v5.2.0)[44]. Per variant SNV concordance was quantified using `bcftools stats --verbose` (v1.20)[45]. SV concordance was evaluated using `SVConcordance` (v4.4.0.0) from GATK.

To accurately assess the practical impacts of LiftOver operations on variant calling, performance metrics need to be carefully chosen[46]. Metrics including true negative counts should be used with caution. In the case of SNVs, the number of sites matching the reference far outnumber variant sites and can lead to inflated estimates of accuracy. Furthermore, standard SNV calling pipelines typically only report sites which differ from the reference sequence. Outside of targeted re-genotyping, the absence of a variant cannot be assumed to be a reference match as the missing call could be attributed to a lack of coverage or insufficient evidence. This issue is even more pronounced with structural variants.

We utilized the following three metrics to i) characterize the concordance and error profiles of LiftOver operations and ii) provide guidance for when and where these errors are the most relevant. True positive (TP), false positive (FP), true negative (TN) and false negative (FN) calls are computed for converted GRCh37 variant calls relative to GRCh38.

Non-reference discordance (NRD) measures the overall disagreement between the two variant sets and is equivalent to overall accuracy with true negatives excluded from the denominator:

$$NRD = \frac{(FP + FN)}{(FP + FN + TP)}$$

False positive rate (FPR) represents the fraction of variants identified in GRCh37, but not in GRCh38:

$$FPR = \frac{FP}{(FP + TP)}$$

False negative rate (FNR) represents the fraction of variants identified in GRCh38, but not in GRCh37:

$$FNR = \frac{FN}{(FN + TP)}$$

Variant annotation

For SNVs, dbSNP (build 151)[47], GENCODE (v34)[48], and HGNC (Nov302017)[49] annotations were added using GATK's `Funcotator` (v4.6.0.0)[26] with pre-packaged data source v1.7.20200521s. Trinucleotide context was determined using `bedtools getfasta` (v2.31.0)[50]. RepeatMasker (v3.0.1)[51] intervals were obtained from the UCSC Table Browser[22] and intersected with variant calls using `bedtools intersect` (v2.31.0)[50]. SVs were intersected with the gnomAD-SV (v4)[52] database (FILTER == "PASS") using a custom R script and annotated with population allele frequency.

Targeted sequencing validation

Additional targeted deep-sequencing data from five patients in the analysis cohort[25,53] (653x mean coverage; **Supplementary Table 2**) was used to validate a subset of sSNV calls. sSNVs identified in the whole genome data within targeted validation regions were considered validated if they were also identified in the targeted deep-sequencing data (**Supplementary Figure 11a**).

Random forest stability prediction

Using the variant calls from our analysis cohort and their corresponding NRD labels, we trained a random forest model to predict variant concordance for each of six variant callers – HaplotypeCaller (v4.2.4.1)[26], MuTect2 (v4.2.4.1)[27], Strelka2 (v2.9.10)[28], SomaticSniper (v1.0.5.0)[29], MuSE2 (v2.0.4)[30], DELLY2 (v1.2.6)[31] – across four variant types (gSNP, sSNV, gSV, sSV; **Supplementary Figure 1a**). Variants were dichotomized based on a 20% NRD threshold and a probability forest (`num.trees` = 500 for gSNPs and 1,000 for sSNVs, gSVs, sSVs) was trained using the ranger (v0.16.0)[54] R package to predict concordant *vs.* discordant variants. Variants failing LiftOver coordinate conversion were excluded. The model outputs a "Stability Score" for each variant indicating the fraction of trees predicting concordant status.

Feature selection and hyperparameter optimization

The set of features considered for each model included all variant fields provided by each variant caller, along with external annotations and site information. Feature inclusion and normalization were

344 determined by optimizing for AUROC in the validation sets for each respective model. Hyperparameters
345 were tuned using a grid search over `mtry` and `min.node.size`.

346 **Model validation datasets**

347 For gSNPs and sSNVs, 10 sarcoma tumour-normal whole genome pairs (**Supplementary Table 3**) and 60
348 sarcoma tumour-normal exome pairs (**Supplementary Table 4**) from The Cancer Genome Atlas (TCGA-
349 SARC)[32] were used as validation sets to demonstrate generalizability across sequencing methods (whole
350 genome *vs.* exome) and cancer types (sarcoma not represented in the training set). Raw sequencing
351 data was downloaded and reprocessed with the same pipelines used for the comparative analysis. For
352 gSVs and sSVs, only the 10 whole genome pairs were used for validation as exome data provides
353 insufficient coverage for comprehensive SV calling.

354 Five whole genomes from the targeted sequencing validation cohort[25,53] were used to evaluate
355 StableLift predictions against an independent truth set of validated *vs.* unvalidated sSNVs
356 (**Supplementary Table 2; Supplementary Figure 11a**).

357 **StableLift**

358 We incorporated these pre-trained and validated models into a standardized workflow accepting either
359 GRCh37 or GRCh38 input VCFs from six variant callers (HaplotypeCaller, MuTect2, Strelka2,
360 SomaticSniper, MuSE2, DELLY2) spanning four variant types (gSNP, sSNV, gSV, sSV). Input variants are
361 converted and annotated as described above and output with a predicted "Stability Score" for filtering
362 based on user-specified thresholds. Performance in the TCGA-SARC whole genome validation set is
363 included with each model to define the default F_1-maximizing operating point and allow for custom
364 filtering based on pre-calibrated sensitivity and specificity estimates.

365 **Data visualization**

366 Figures were generated in R (v4.3.3) using the lattice (v0.22-6), latticeExtra (v0.6-30), BPG (v7.1.0)[55],
367 VennDiagram (v1.7.3)[56], and RIdeogram (v0.2.2)[57] packages.

368

Online Methods References

39. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (2019). doi:10.1109/IPDPS.2019.00041.

40. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).

41. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

42. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).

43. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

44. Cingolani, P. *et al.* Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* **3**, 35 (2012).

45. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

46. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).

47. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

48. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

49. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–D1009 (2023).

50. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).

51. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).

52. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

53. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).

54. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**, 264–278 (2012).

55. P'ng, C. *et al.* BPG: Seamless, automated and interactive visualization of scientific data. *BMC Bioinformatics* **20**, 42 (2019).

56. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, 35 (2011).

57. Hao, Z. *et al.* RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* **6**, e251 (2020).

Data availability

Somatic VCFs, resource files for variant annotation, and pre-trained random forest models for GRCh37→GRCh38 and GRCh38→GRCh37 conversions are available on GitHub as release attachments (https://github.com/uclahs-cds/pipeline-StableLift/releases). The tumour-normal whole genome pairs used for analysis and training StableLift can be accessed through the European Genome-Phenome Archive (https://ega-archive.org/studies/EGAS00001000900) and the Bionimbus Protected Data Cloud (https://icgc.bionimbus.org/). TCGA-SARC exome and whole genome datasets used for validation can be accessed from the GDC Data Portal (portal.gdc.cancer.gov/projects/TCGA-SARC).

Code availability

StableLift is available on GitHub (https://github.com/uclahs-cds/pipeline-StableLift) as a Nextflow pipeline featuring LiftOver coordinate conversion, variant annotation with external databases and prediction of cross-build variant stability. Nextflow pipelines for alignment and variant calling are on GitHub (https://github.com/uclahs-cds/metapipeline-DNA) and described elsewhere[20].

Acknowledgments

The authors thank all members of the Boutros lab and the Office of Health Informatics and Analytics (OHIA) at UCLA. The comparative analysis and training of StableLift were based upon data generated by the International Cancer Genome Consortium (ICGC) and the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium. Validation of StableLift was based upon data generated by The Cancer Genome Atlas (TCGA) Research Network and the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium.

Author Contributions

Conceptualization: NKW, PCB
Data Curation: NKW, SF, AEG, RA, JO, YP, TNY
Formal Analysis: NKW, NZ
Funding Acquisition: PCB
Methodology: NKW, HKW, JA, PCB
Software: NKW, NW, YP
Supervision: PCB
Writing – Original Draft: NKW, PCB
Writing – Review & Editing: NKW, NW, HKW, SF, AEG, NZ, RA, JO, JA, YP, TNY, PCB

Funding Sources

This work was supported by the NIH through grants P30CA016042, U2CCA271894, U24CA248265 and R01CA270108, by the DOD through grant W81XWH2210247 and by a Prostate Cancer Foundation Special Challenge Award to PCB (Award ID #: 20CHAS01) made possible by the generosity of Mr. Larry Ruvo. NKW, HKW, JO were supported by a Jonsson Comprehensive Cancer Center Fellowship. AEG was supported by a Howard Hughes Medical Institute Gilliam Fellowship. NZ was supported by the NIH through grants T32HG002536 and F31CA281168. RA was supported by NIGMS grants T32GM008042 and T32GM152342 and a Jonsson Comprehensive Cancer Center Fellowship.

Conflicts of Interest

PCB sits on the Scientific Advisory Boards of Intersect Diagnostics Inc., BioSymetrics Inc. and previously sat on that of Sage Bionetworks. All other authors declare no conflicts of interest.

Figure Legends

Figure 1: Overview of differences between GRCh37 and GRCh38 variant calls.
a) Experimental design for matched comparison of germline and somatic variants in a representative cancer genomics workflow. **b-c)** Cohort level overlap of converted GRCh37 *vs.* GRCh38 germline variants (gSNP, gSV). **d)** Number of build-specific germline variants per sample. **e)** Difference in per sample germline variant counts found in GRCh38 relative to GRCh37. **f-g)** Cohort level overlap of converted GRCh37 *vs.* GRCh38 somatic variants (sSNV, sSV). **h)** Number of build-specific somatic variants per sample. **i)** Difference in per sample somatic variant counts found in GRCh38 relative to GRCh37. **j-k)** Variant discordance per sample stratified by variant type and gSV subtype. (NRD = non-reference discordance, FPR = false positive rate, FNR = false negative rate; DEL = deletion, DUP = duplication, INS = insertion, INV = inversion, BND = breakend/translocation) **l**) Distribution of gSNP density and NRD across the genome. **m)** Correlation between continuous covariates and NRD per variant type. Spearman's correlation indicated by dot size and color; statistical significance with false discovery rate correction indicated by background shading. **n)** Validation rate of build-concordant, GRCh37-specific, and GRCh38-specific sSNVs by targeted deep-sequencing.

Figure 2: Machine-learning approach to predicting variant stability across genome builds.
a) Overview of StableLift as a multi-purpose genomics utility performing LiftOver coordinate conversion, variant annotation, and cross-build stability prediction. **b)** Random forest model performance for gSNPs (HaplotypeCaller) shown as ROC curves and AUC measures for out-of-bag whole genome training (OOB, solid black), whole genome validation (WGS, solid red), and whole exome validation (WXS, dashed red) sets. Default operating point maximizing F_1-score highlighted (blue diamond) with corresponding sensitivity and specificity in the whole genome validation set. **c)** Comparison of concordant (TP) and discordant (FP) gSNP counts before and after default StableLift filtering. **d)** Random forest feature importance colored by caller-specific metrics, variant annotations and site information. Normalized features indicated by *. **e-g)** Same as b-d for sSNVs (MuTect2). **h-j)** Same as b-d for gSVs (DELLY2). **k-m)** Same as b-d for sSVs (DELLY2).