

# 1 StableLift: Optimized Germline and Somatic Variant Detection 2 Across Genome Builds

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

6

7

8 <sup>1</sup>Department of Human Genetics, University of California, Los Angeles

9 <sup>2</sup>Jonsson Comprehensive Cancer Center, University of California, Los Angeles

10 <sup>3</sup>Institute for Precision Health, University of California, Los Angeles

11 <sup>4</sup>Department of Urology, University of California, Los Angeles

12 # Corresponding author:

13 Center for Health Sciences

14 10833 Le Conte Avenue

15 Los Angeles, CA

16 United States of America

17 90095

18 Email: [pboutros@mednet.ucla.edu](mailto:pboutros@mednet.ucla.edu)

19

20

21

22

23

24

## 25 Abstract

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

## 37 Main

38 Since initial assembly of the human genome in 2001<sup>1,2</sup>, thousands of errors have been corrected,  
39 polymorphic regions have been defined and the diversity of included individuals has expanded<sup>3-5</sup>. These  
40 advances have led to a series of updated human reference genome “builds”, each with incompatible  
41 coordinate numbering. While new builds are more accurate representations, their adoption can be slow  
42 in both research and clinical settings<sup>6</sup>.

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

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

59 Most germline SNPs and structural variants identified were shared between the two builds (>93%;  
60 **Figure 1b-c**). Nevertheless, we detected  $166,704 \pm 14,829$  build-specific gSNPs and  $908 \pm 73$  build-  
61 specific gSVs per individual (mean  $\pm$  standard deviation; **Figure 1d**). Alignment to GRCh38 led to  
62 identification of more gSNPs and gSVs (**Figure 1e**). By contrast, somatic variant detection was  
63 dramatically more variable: only 82% of sSNVs and 53% of sSVs were identified in both builds (**Figure**  
64 **1f-g**). This led to  $3,611 \pm 2,025$  build-specific sSNVs and  $93 \pm 61$  build-specific sSVs (**Figure 1h**), with  
65 more somatic variants identified when aligning to GRCh38 (**Figure 1i**).

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

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

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

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

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

106 To quantify the cross-build stability of any individual variant, we created a machine-learning approach  
107 called StableLift. By leveraging features associated with build-discordance (**Supplementary Figures 1-**  
108 **7**), StableLift estimates the likelihood (“Stability Score”) that a given variant will be consistently  
109 represented across two genome builds (**Figure 2a**). We trained StableLift with variants detected from  
110 the same fifty tumour-normal WGS pairs using six variant callers spanning all four variant-types:  
111 HaplotypeCaller<sup>26</sup>, MuTect2<sup>27</sup>, Strelka2<sup>28</sup>, SomaticSniper<sup>29</sup>, MuSE2<sup>30</sup> and DELLY2<sup>31</sup>. We validated  
112 StableLift in 10 tumour-normal whole genomes<sup>32</sup> (**Supplementary Table 3**) and 60 tumour-normal  
113 exomes<sup>32</sup> (**Supplementary Table 4**) for area under the receiver operating characteristic curve (AUROC)  
114 and selected a default operating point to maximize F<sub>1</sub>-score in the whole genome validation set.

115 StableLift robustly identified build-discordant gSNP calls, with validation AUROCs of 0.958 for WGS and  
116 0.941 for exome sequencing (**Figure 2b**; **Supplementary Figure 8a-c**). At the F<sub>1</sub>-maximizing operating  
117 point,  $49.7 \pm 0.5\%$  of discordant gSNPs in WGS validation were discarded, corresponding to  $51,181 \pm$   
118  $4,884$  discordant variants removed per individual (**Figure 2c**). A variety of features contributed to the  
119 accuracy of these predictions, most notably TOPMed<sup>24</sup> population allele frequency (**Figure 2d**) driven

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

122 StableLift similarly identified build-discordant sSNVs, with validation AUROCs of 0.890 for WGS and  
123 0.851 for exome sequencing (MuTect2; **Figure 2e**; **Supplementary Figure 8d-f**) and a  $45.7 \pm 11.7\%$   
124 reduction of discordant calls ( $-209 \pm 56$  discordant sSNVs; **Figure 2f**). sSNV stability prediction was driven  
125 by a wide range of predictor features (**Figure 2g**). Models fit to three other sSNV callers achieved similar  
126 performance:  $AUROC_{WGS} = 0.932$  for Strelka2,  $AUROC_{WGS} = 0.964$  for SomaticSniper and  $AUROC_{WGS} =$   
127  $0.905$  for MuSE2 (**Supplementary Figure 9a-i**, **Supplementary Figure 10**). Different sSNV calling  
128 algorithms had similar but not identical patterns of feature importance, highlighting the interaction  
129 between genomic features and variant detection algorithms (**Supplementary Figure 9j**).

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

135 Finally, we applied StableLift to structural variant calls made by DELLY2<sup>31</sup>. Despite only 28,350  
136 concordant cases and 734 discordant cases of gSV training data (**Figure 1c**), StableLift again accurately  
137 identified discordant calls, with a validation AUROC of 0.926 (**Figure 2h**) and a  $56.2 \pm 5.3\%$  reduction of  
138 discordant calls ( $-63 \pm 10$  discordant gSVs; **Figure 2i**). Length of variant was the most important single  
139 feature, with a range of predictive features differing from those driving the accuracy of the gSNP and  
140 sSNV models (**Figure 2j**). Accuracy in DELLY2 sSVs was equally high, achieving a validation AUROC of  
141 0.961 (**Figure 2k**) and removing 81.7% of discordant sSVs ( $-171 \pm 170$  discordant sSVs; **Figure 2l**). Only  
142 4,907 concordant and 1,845 discordant training cases were needed for this model, and its accuracy was  
143 driven by read count and SV length (**Figure 2m**).

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

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

## 157 **References**

- 158 1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921  
159 (2001).
- 160 2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of  
161 the human genome. *Nature* **431**, 931–945 (2004).
- 162 3. Church, D. M. *et al.* Modernizing Reference Genome Assemblies. *PLoS Biol.* **9**, e1001091 (2011).
- 163 4. Church, D. M. *et al.* Extending reference assembly models. *Genome Biol.* **16**, 13 (2015).
- 164 5. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies  
165 demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
- 166 6. Lansdon, L. A. *et al.* Factors Affecting Migration to GRCh38 in Laboratories Performing Clinical  
167 Next-Generation Sequencing. *J. Mol. Diagn. JMD* **23**, 651–657 (2021).
- 168 7. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief.*  
169 *Bioinform.* **14**, 144–161 (2013).
- 170 8. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome  
171 assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
- 172 9. Guo, Y. *et al.* Improvements and impacts of GRCh38 human reference on high throughput  
173 sequencing data analysis. *Genomics* **109**, 83–90 (2017).
- 174 10. Zheng-Bradley, X. *et al.* Alignment of 1000 Genomes Project reads to reference assembly  
175 GRCh38. *GigaScience* **6**, 1–8 (2017).
- 176 11. Gao, G. F. *et al.* Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data  
177 Commons' Data. *Cell Syst.* **9**, 24-34.e10 (2019).
- 178 12. Lowy-Gallego, E. *et al.* Variant calling on the GRCh38 assembly with the data from phase three  
179 of the 1000 Genomes Project. *Wellcome Open Res.* **4**, 50 (2019).
- 180 13. Pan, B. *et al.* Similarities and differences between variants called with human reference genome  
181 HG19 or HG38. *BMC Bioinformatics* **20**, 101 (2019).
- 182 14. Luu, P.-L., Ong, P.-T., Dinh, T.-P. & Clark, S. J. Benchmark study comparing liftover tools for genome  
183 conversion of epigenome sequencing data. *NAR Genomics Bioinforma.* **2**, lqaa054 (2020).
- 184 15. Li, H. *et al.* Exome variant discrepancies due to reference-genome differences. *Am. J. Hum.*  
185 *Genet.* **108**, 1239–1250 (2021).
- 186 16. Park, K.-J., Yoon, Y. A. & Park, J.-H. Evaluation of Liftover Tools for the Conversion of Genome  
187 Reference Consortium Human Build 37 to Build 38 Using ClinVar Variants. *Genes* **14**, 1875 (2023).
- 188 17. Ormond, C., Ryan, N. M., Corvin, A. & Heron, E. A. Converting single nucleotide variants between  
189 genome builds: from cautionary tale to solution. *Brief. Bioinform.* **22**, bbab069 (2021).
- 190 18. Fraser, M. *et al.* Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–  
191 364 (2017).
- 192 19. Patel, Y. *et al.* NFTest: automated testing of Nextflow pipelines. *Bioinformatics* **40**, btae081  
193 (2024).



- 194 20. Patel, Y. *et al.* Metapipeline-DNA: A comprehensive germline & somatic genomics Nextflow  
195 pipeline. *BioRxiv* (2024) doi:<https://doi.org/10.1101/2024.09.04.611267>.
- 196 21. Genovese, G. *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic  
197 variants across genome assemblies. *Bioinformatics* **40**, btae038 (2024).
- 198 22. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496  
199 (2004).
- 200 23. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**,  
201 D590–D598 (2006).
- 202 24. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.  
203 *Nature* **590**, 290–299 (2021).
- 204 25. Aaltonen, L. A. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
- 205 26. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-  
206 generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
- 207 27. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous  
208 cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
- 209 28. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*  
210 **15**, 591–594 (2018).
- 211 29. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome  
212 sequencing data. *Bioinformatics* **28**, 311–317 (2012).
- 213 30. Ji, S., Montierth, M. D. & Wang, W. MuSE: A Novel Approach to Mutation Calling with Sample-  
214 Specific Error Modeling. *Methods Mol. Biol. Clifton NJ* **2493**, 21–27 (2022).
- 215 31. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read  
216 analysis. *Bioinforma. Oxf. Engl.* **28**, i333–i339 (2012).
- 217 32. Abeshouse, A. *et al.* Comprehensive and Integrated Genomic Characterization of Adult Soft  
218 Tissue Sarcomas. *Cell* **171**, 950-965.e28 (2017).
- 219 33. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of  
220 genome inference. *Genome Res.* **27**, 665–676 (2017).
- 221 34. Rakocevic, G. *et al.* Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**,  
222 354–362 (2019).
- 223 35. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse  
224 genomes. *Science* **374**, abg8871 (2021).
- 225 36. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation.  
226 *Science* **376**, eabl3533 (2022).
- 227 37. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
- 228 38. Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-  
229 Cactus. *Nat. Biotechnol.* 1–11 (2023) doi:[10.1038/s41587-023-01793-w](https://doi.org/10.1038/s41587-023-01793-w).

- 230 39. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-  
231 MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium*  
232 (*IPDPS*) 314–324 (2019). doi:10.1109/IPDPS.2019.00041.
- 233 40. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*  
234 **35**, 316–319 (2017).
- 235 41. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome  
236 browsers. *Bioinformatics* **25**, 1841–1842 (2009).
- 237 42. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput.*  
238 *Biol.* **9**, e1003118 (2013).
- 239 43. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
- 240 44. Cingolani, P. *et al.* Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational  
241 Studies with a New Program, SnpSift. *Front. Genet.* **3**, 35 (2012).
- 242 45. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
- 243 46. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human  
244 genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
- 245 47. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311  
246 (2001).
- 247 48. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic*  
248 *Acids Res.* **47**, D766–D773 (2019).
- 249 49. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–  
250 D1009 (2023).
- 251 50. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.  
252 *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).
- 253 51. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).
- 254 52. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes.  
255 *Nature* **625**, 92–100 (2024).
- 256 53. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**,  
257 264–278 (2012).
- 258 54. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High  
259 Dimensional Data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).
- 260 55. P'ng, C. *et al.* BPG: Seamless, automated and interactive visualization of scientific data. *BMC*  
261 *Bioinformatics* **20**, 42 (2019).
- 262 56. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable  
263 Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, 35 (2011).
- 264 57. Hao, Z. *et al.* RIdiogram: drawing SVG graphics to visualize and map genome-wide data on the  
265 idiograms. *PeerJ Comput. Sci.* **6**, e251 (2020).

## 267 **Online Methods**

### 268 **Analysis cohort**

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

### 274 **Alignment and variant calling**

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

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

### 290 **LiftOver coordinate conversion**

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

### 296 **Variant concordance**

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

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



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

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

$$314 \quad NRD = \frac{(FP + FN)}{(FP + FN + TP)}$$

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

$$316 \quad FPR = \frac{FP}{(FP + TP)}$$

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

$$318 \quad FNR = \frac{FN}{(FN + TP)}$$

### 319 **Variant annotation**

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

### 326 **Targeted sequencing validation**

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

### 331 **Random forest stability prediction**

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

### 341 **Feature selection and hyperparameter optimization**

342 The set of features considered for each model included all variant fields provided by each variant caller,  
343 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)<sup>32</sup> 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<sup>25,53</sup> 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<sub>1</sub>-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)<sup>55</sup>,  
367 VennDiagram (v1.7.3)<sup>56</sup>, and RIdeogram (v0.2.2)<sup>57</sup> packages.

368

## 369 Online Methods References

- 370 39. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-  
371 MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium*  
372 *(IPDPS)* 314–324 (2019). doi:10.1109/IPDPS.2019.00041.
- 373 40. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*  
374 **35**, 316–319 (2017).
- 375 41. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome  
376 browsers. *Bioinformatics* **25**, 1841–1842 (2009).
- 377 42. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput.*  
378 *Biol.* **9**, e1003118 (2013).
- 379 43. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
- 380 44. Cingolani, P. *et al.* Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational  
381 Studies with a New Program, SnpSift. *Front. Genet.* **3**, 35 (2012).
- 382 45. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
- 383 46. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human  
384 genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
- 385 47. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311  
386 (2001).
- 387 48. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic*  
388 *Acids Res.* **47**, D766–D773 (2019).
- 389 49. Seal, R. L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, D1003–  
390 D1009 (2023).
- 391 50. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.  
392 *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).
- 393 51. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).
- 394 52. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes.  
395 *Nature* **625**, 92–100 (2024).
- 396 53. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High  
397 Dimensional Data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).
- 398 54. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**,  
399 264–278 (2012).
- 400 55. P'ng, C. *et al.* BPG: Seamless, automated and interactive visualization of scientific data. *BMC*  
401 *Bioinformatics* **20**, 42 (2019).
- 402 56. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable  
403 Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, 35 (2011).
- 404 57. Hao, Z. *et al.* Rldeogram: drawing SVG graphics to visualize and map genome-wide data on the  
405 idiograms. *PeerJ Comput. Sci.* **6**, e251 (2020).

## 406 **Data availability**

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

## 414 **Code availability**

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

## 419 **Acknowledgments**

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

## 425 **Author Contributions**

426 **Conceptualization:** NKW, PCB

427 **Data Curation:** NKW, SF, AEG, RA, JO, YP, TNY

428 **Formal Analysis:** NKW, NZ

429 **Funding Acquisition:** PCB

430 **Methodology:** NKW, HKW, JA, PCB

431 **Software:** NKW, NW, YP

432 **Supervision:** PCB

433 **Writing – Original Draft:** NKW, PCB

434 **Writing – Review & Editing:** NKW, NW, HKW, SF, AEG, NZ, RA, JO, JA, YP, TNY, PCB

## 435 **Funding Sources**

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

## 443 **Conflicts of Interest**

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

## 446 **Figure Legends**

### 447 **Figure 1: Overview of differences between GRCh37 and GRCh38 variant calls.**

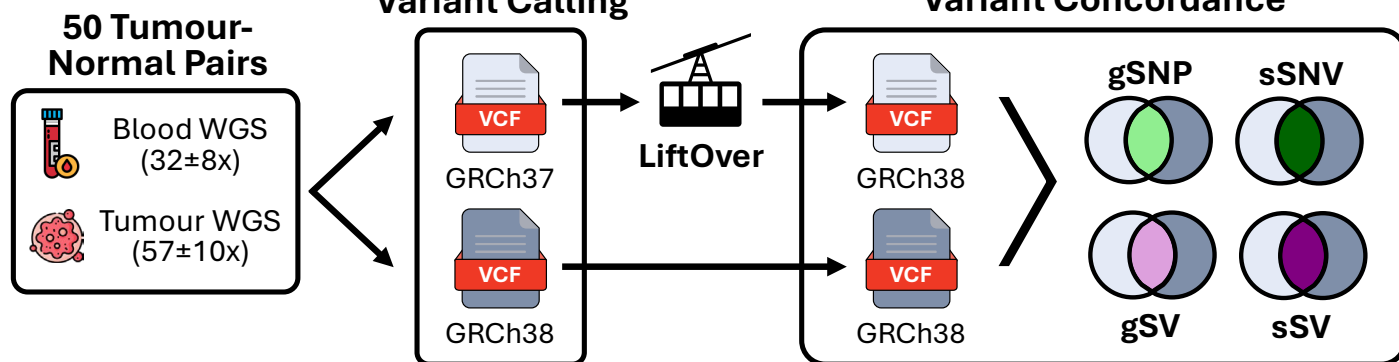
448 **a)** Experimental design for matched comparison of germline and somatic variants in a representative  
449 cancer genomics workflow. **b-c)** Cohort level overlap of converted GRCh37 vs. GRCh38 germline variants  
450 (gSNP, gSV). **d)** Number of build-specific germline variants per sample. **e)**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

### 461 **Figure 2: Machine-learning approach to predicting variant stability across genome builds.**

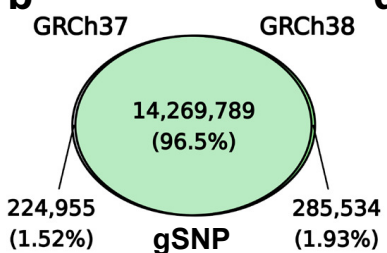
462 **a)** Overview of StableLift as a multi-purpose genomics utility performing LiftOver coordinate conversion,  
463 variant annotation, and cross-build stability prediction. **b)** Random forest model performance for gSNPs  
464 (HaplotypeCaller) shown as ROC curves and AUC measures for out-of-bag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



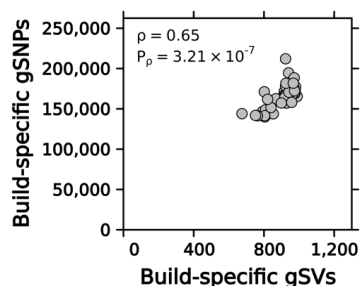
**a**



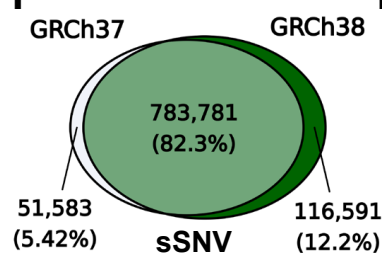
**b**



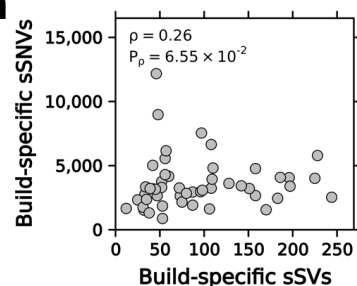
**d**



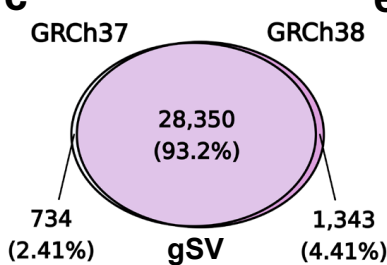
**f**



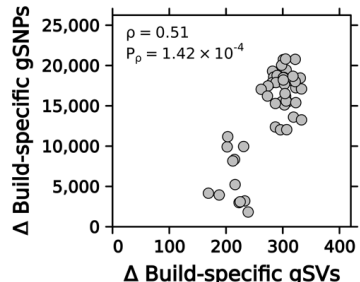
**h**



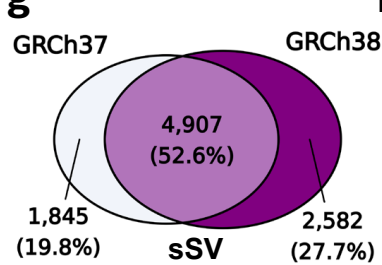
**c**



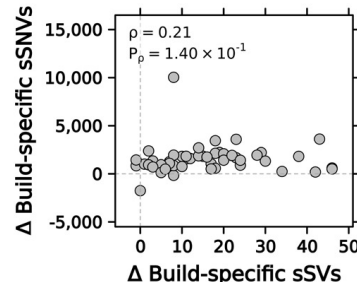
**e**



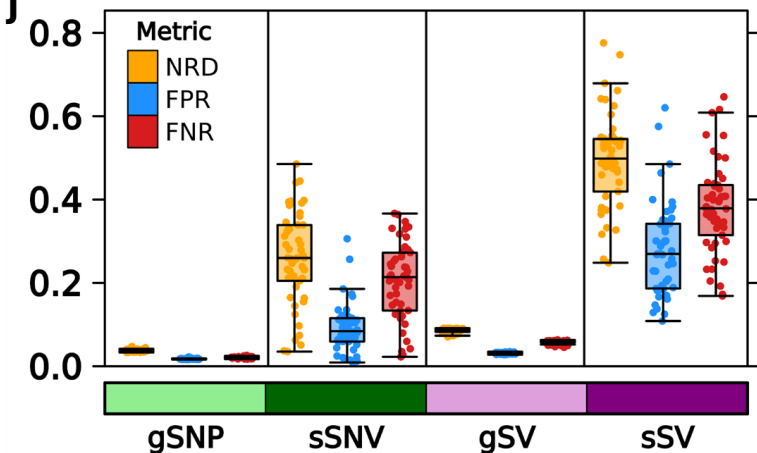
**g**



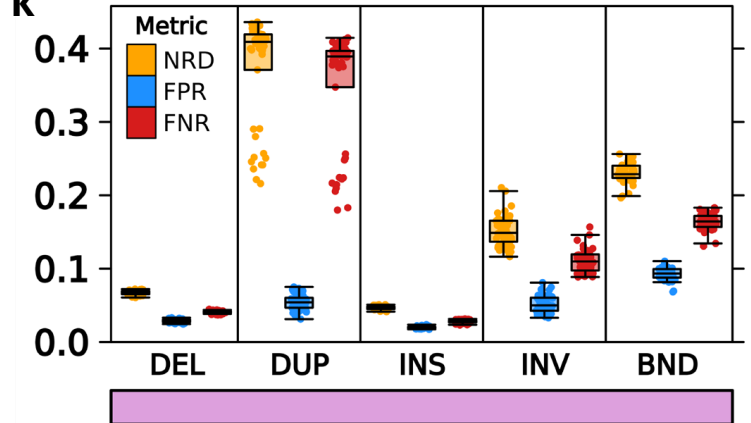
**i**



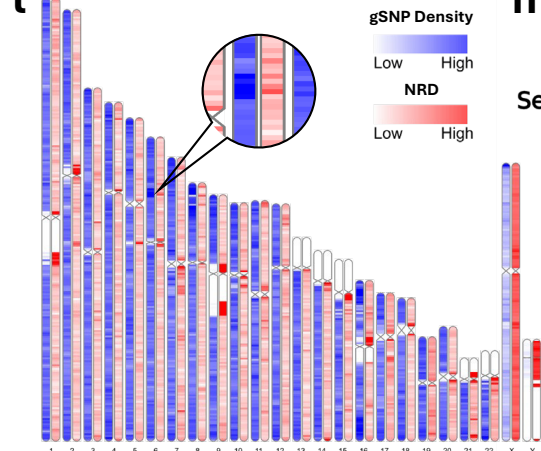
**j**



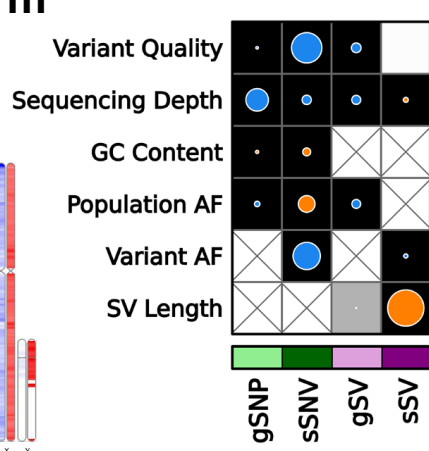
**k**



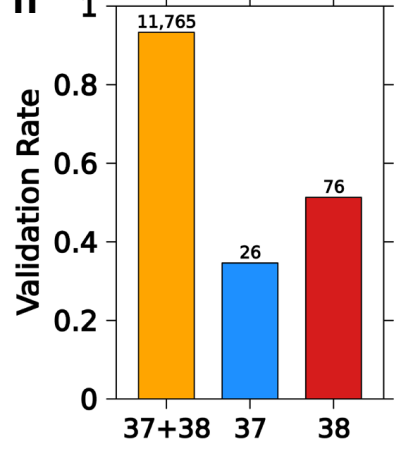
**l**



**m**



**n**



# StableLift

