NEW YORK GENOME CENTER®

Germline and somatic variant calling with NovaSeq™ 6000 2x250bp reads

Minita Shah, Marta Byrska-Bishop, Wayne E. Clarke, Molly Johnson, Kanika Arora, Rashesh Sanghvi, Uday Evani, Kshithija Nagulapalli, Michael C. Zody, Soren Germer, Jade Carter, Giuseppe Narzisi, Nicolas Robine

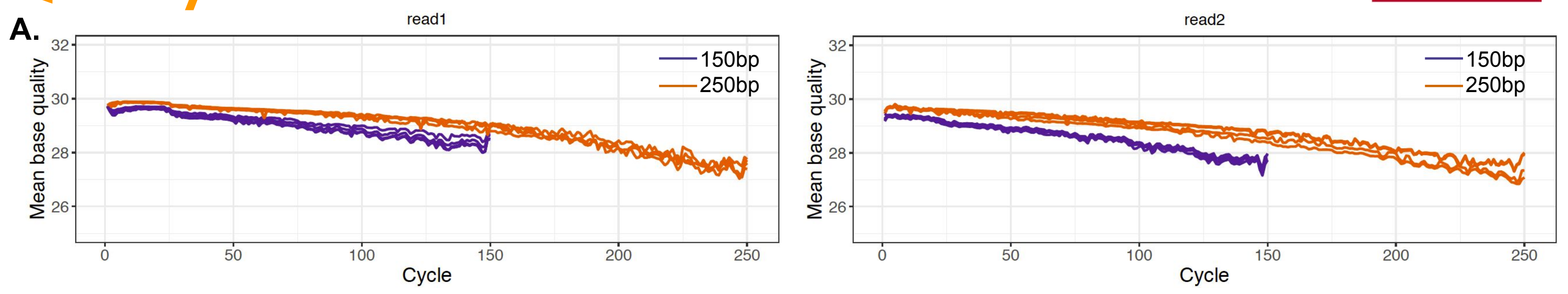
Abstract

Longer reads can substantially improve bioinformatics analysis. Despite the recent advances in ultra long read technologies (e.g. PacBio and Oxford Nanopore), Illumina's short reads have an unmatched throughput and error rate. The recently released NovaSeq™ 6000 system now supports a 500 cycle kit (so far only available on the SP flow cell) that allows the generation of 2x250bp paired reads, effectively increasing the read length by 66% over their previous 2x150bp kit. To evaluate the improvements provided by the longer Illumina reads, we sequenced 5 samples on the new NovaSeq™ 6000 SP 500 cycle kit (2x250bp) and evaluated the performance of the run, in comparison with the same samples sequenced on the standard S4 2x150bp kit. We employed the well-characterized CEU HapMap trio (NA12878, NA12891, and NA12892) and one breast cancer cell line (HCC-1143) along with its matched normal. We ran our standard germline and cancer pipelines, including alignment with BWA-MEM, variant calling with a variety of algorithms, filtering, and annotation. Here, we compared concordance of small and large variants between the two kits.

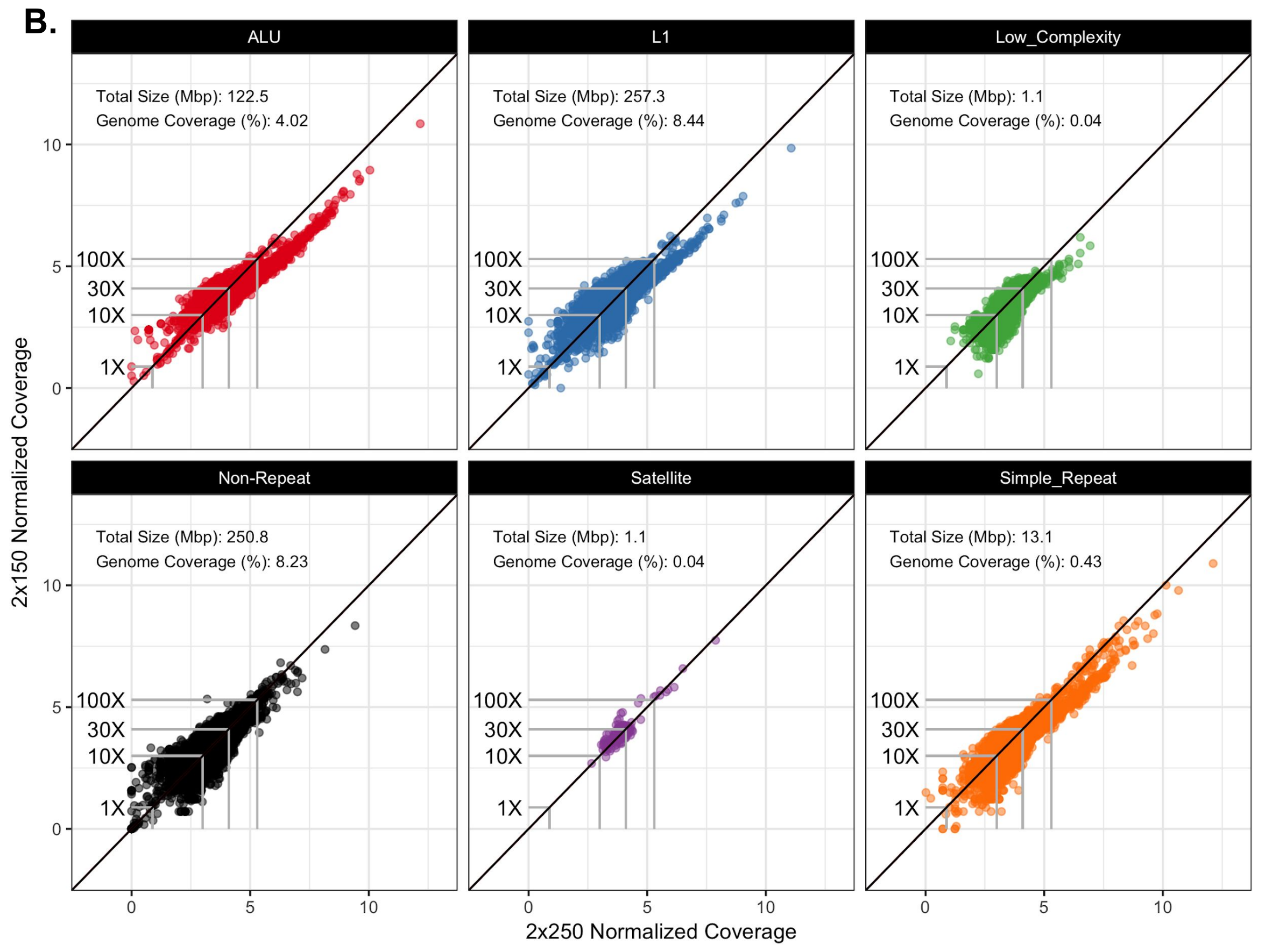
Recent Biorxiv link describing the somatic pipeline: Deep sequencing of 3 cancer cell lines on 2 sequencing platforms: https://doi.org/10.1101/623702

bioRxiv

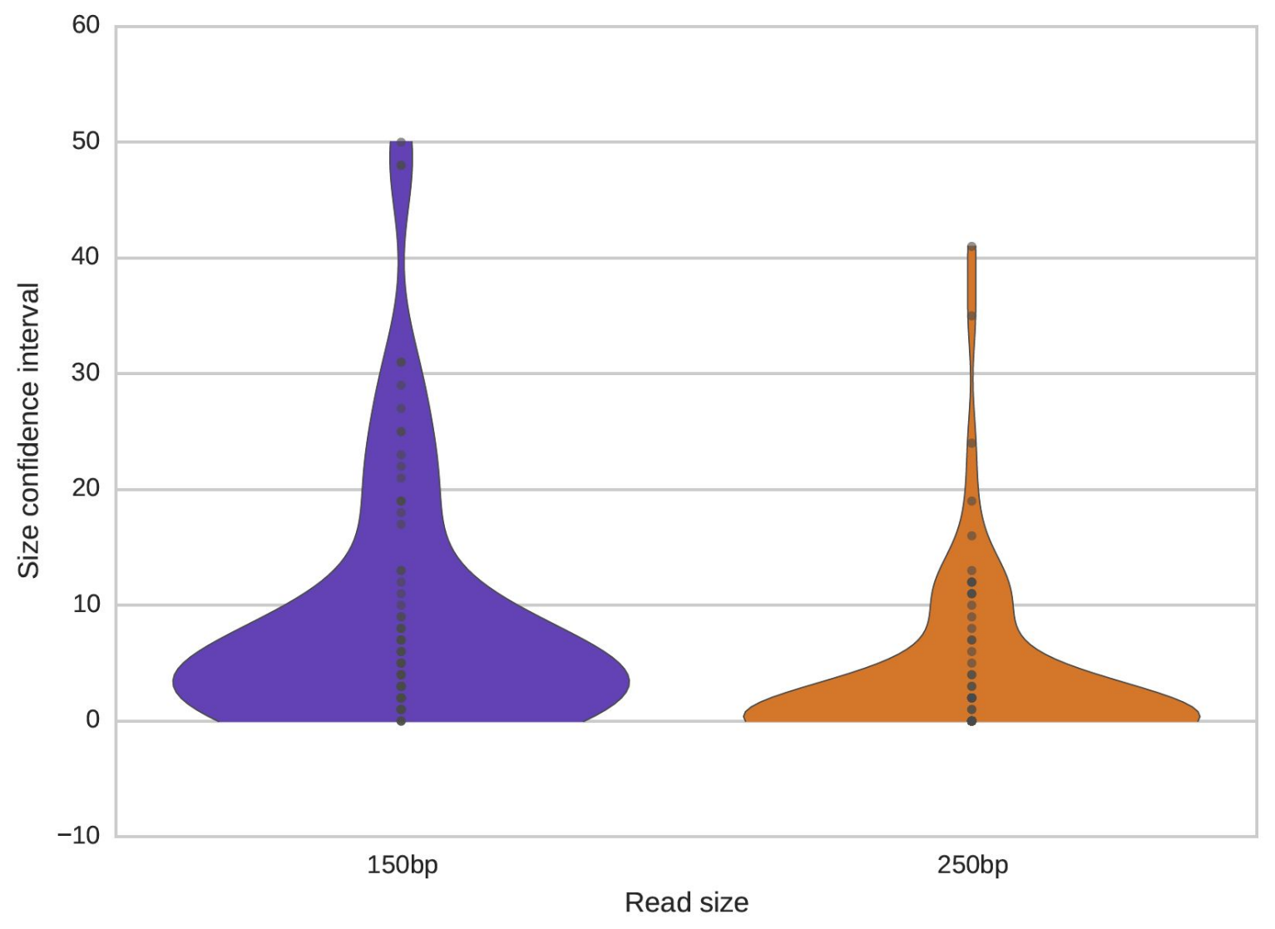
Quality metrics



Quality metrics. **A)** Quality by cycle for 2x150bp and 2x250bp reads. **B)** Coverage of six categories of low confidence genomic regions normalized to 30X coverage and transformed using the Inverse Hyperbolic Sine (IHS) transformation comparing 2x150bp and 2x250bp reads.



Repeat expansion analysis

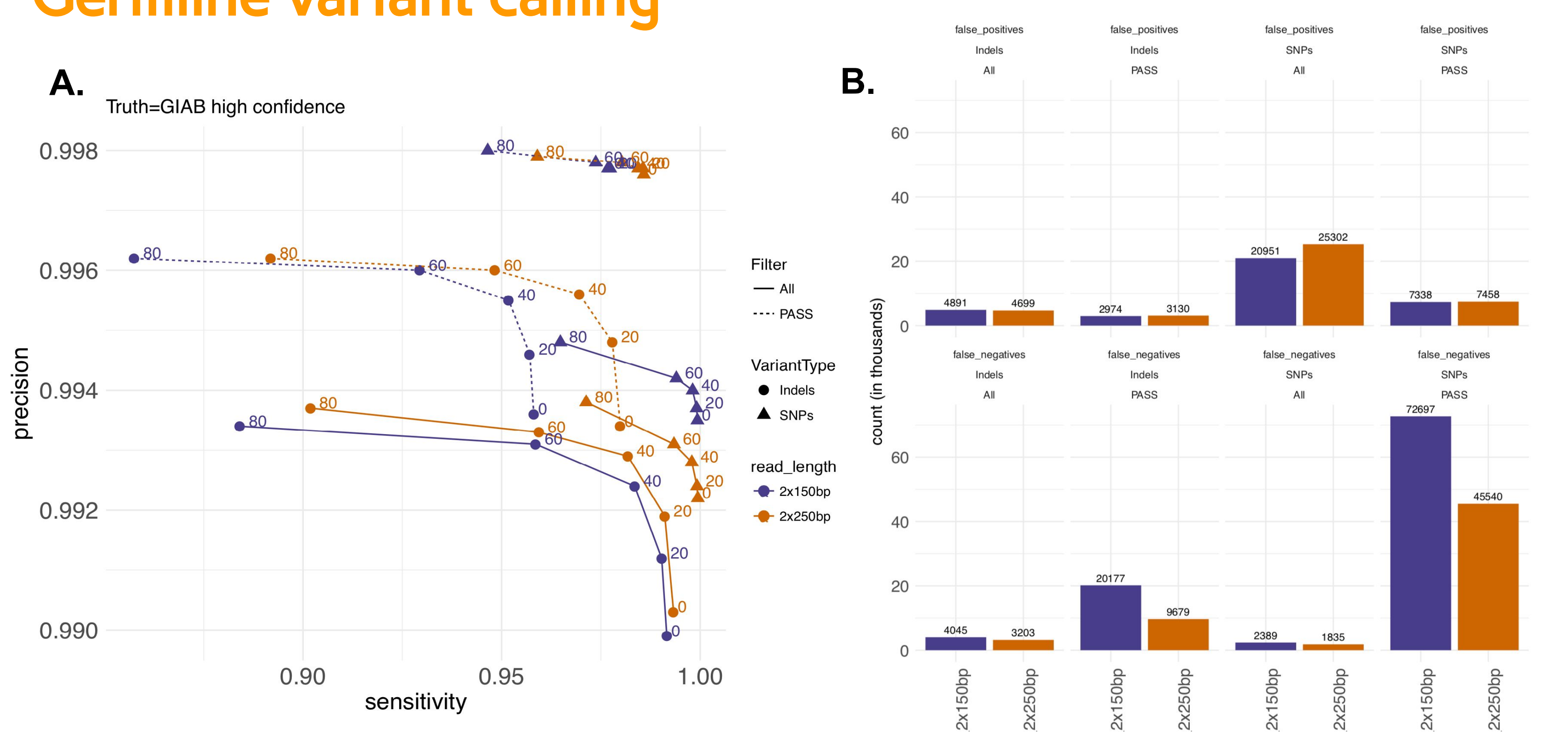


Repeat expansion analysis. STRs which have a reference allele greater than 75bp in the human genome (909 sites) were selected and genotyped with ExpansionHunter [1] using both 2x150bp and 2x250bp reads for NA12878. Only informative sites are plotted skipping over loci where the confidence intervals is 0 for both read lengths. The distribution of the confidence interval is tighter using longer reads, with the center of the distribution shifting towards 0 for the 2x250bp reads indicating that single base pair resolution is achievable on substantially more sites using the 2x250bp reads compare to the 2x150bp reads.

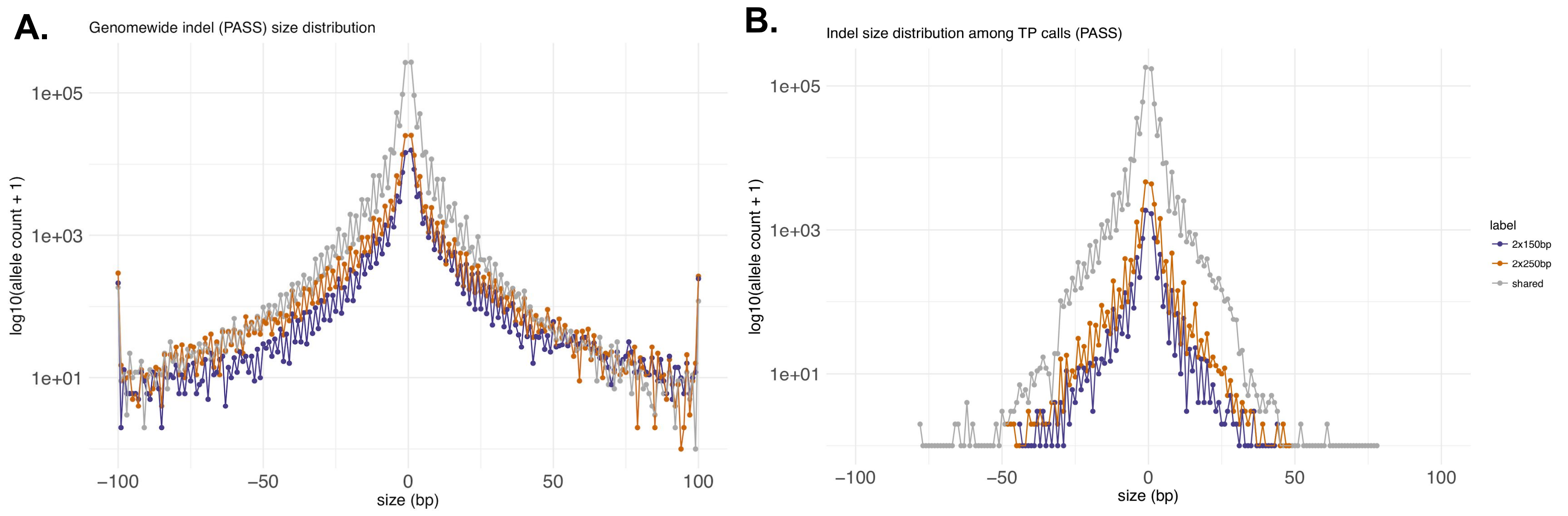
References

1. Dolzhenko *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.,* 2017. 27:1895-1903
2. Zook JM *et al*. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, 2014. 32(3), 246–251.
3. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 2015.
4. Parikh H *et al*. svclassify: a method to establish benchmark structural variant calls. *BMC Genomics*, 2016.
5. Roller E *et al*. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics*, 2016.
6. Arora K *et al.* Deep sequencing of 3 cancer cell lines on 2 sequencing platforms. *bioRxiv*, 2019.

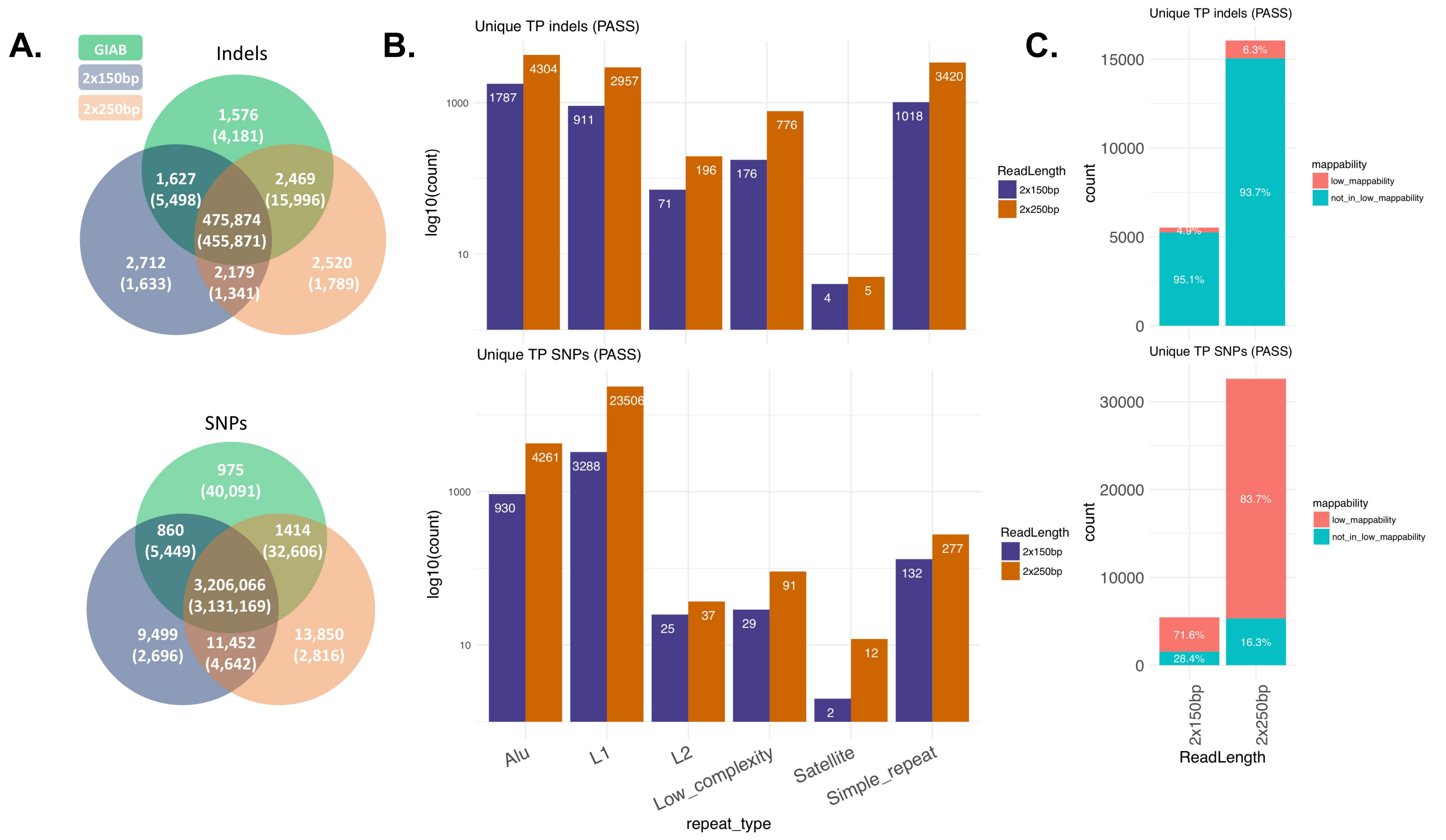
Germline variant calling



Germline SNV and indel calling performance. **A)** Precision and sensitivity of SNV (triangles) and indel (circles) calls in NA12878, sequenced using 2x150bp (blue) and 2x250bp (orange) paired reads, as compared to GIAB truth set [2] within high confidence regions. Solid lines correspond to all variants, whereas dashed lines represent PASS variants only. Numbers along precision-sensitivity curves correspond to genotype quality (GQ) thresholds from 0 to 80. **B)** Counts of false positive and false negative SNV and indel (all and PASS only) calls in 2x150bp and 2x250bp.

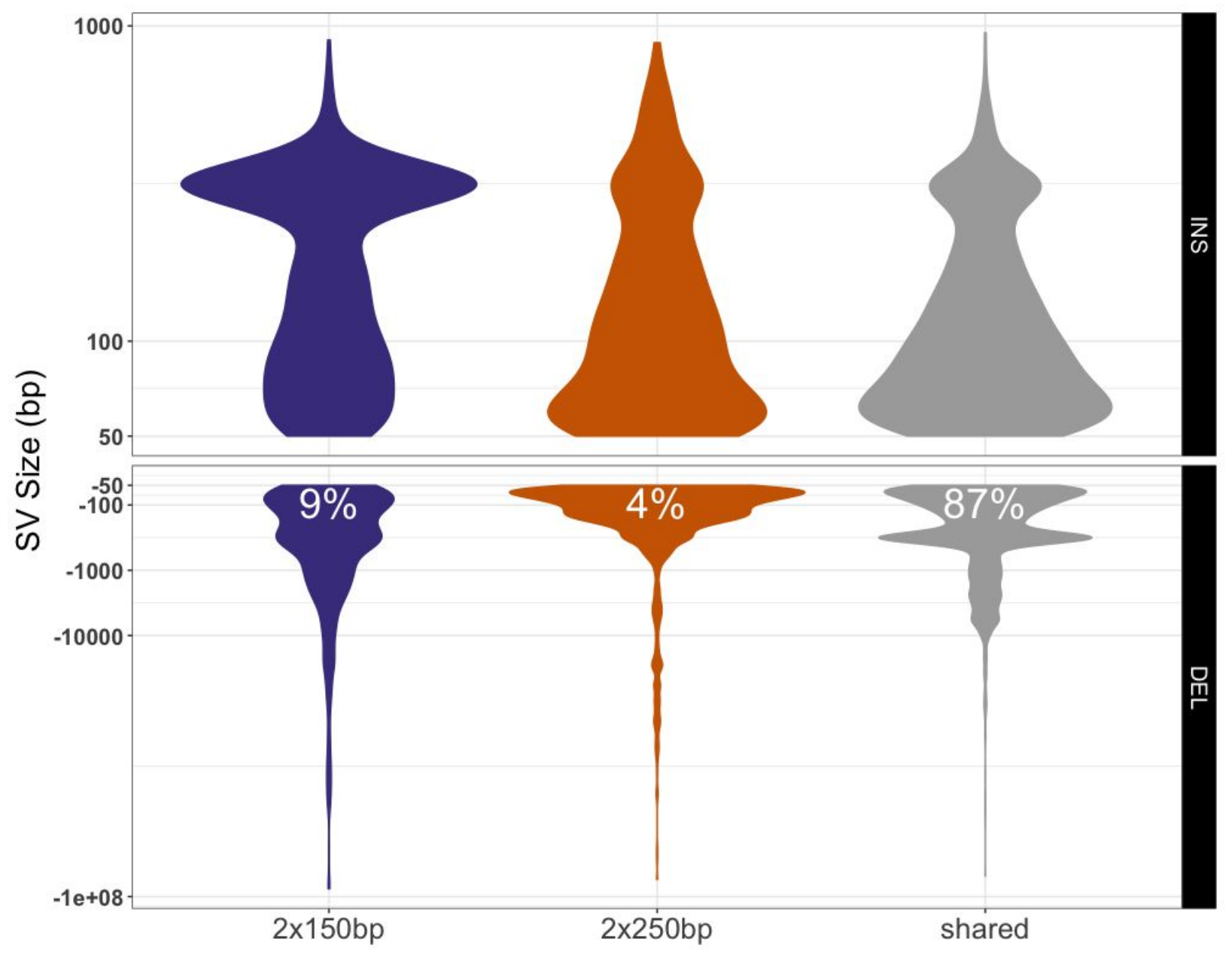


Indel size distribution for shared and unique calls in 150bp vs. 250bp. **A)** Genome-wide distribution of indel sizes among NA12878 indel calls (PASS) that are shared (grey) between 2x150bp and 2x250bp and those that are unique to either 2x150bp (blue) and 2x250bp (orange). **B)** Distribution of indel sizes among true positive calls identified based on the comparison to GIAB truth set [2]. 2x250bp results in more indel calls in the longer size range as compared to 2x150bp, both genome-wide and among true positive calls.

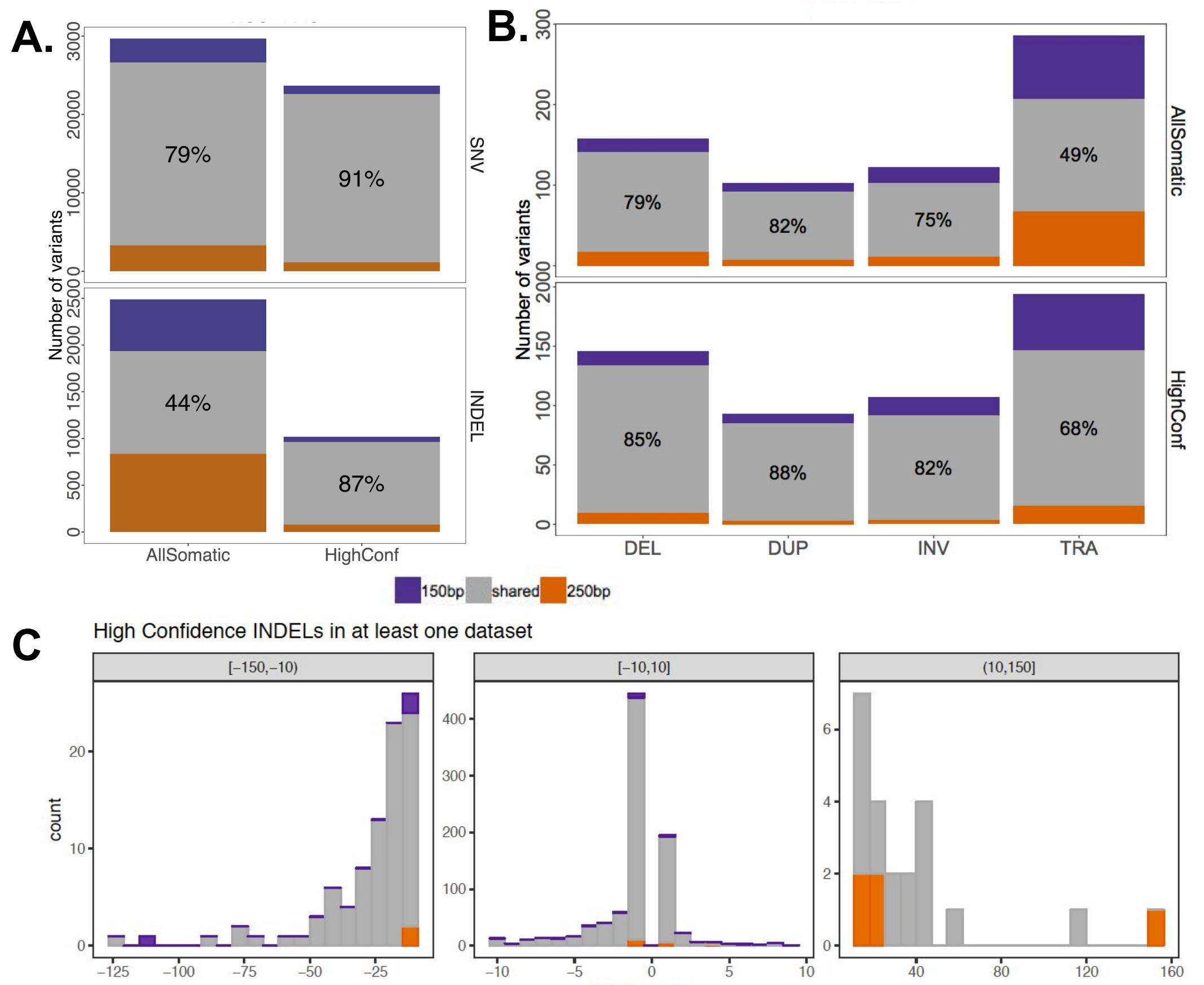


Overlaps with repeats and low mappability regions. **A)** Venn diagrams showing overlaps between indel (top) and SNV (bottom) calls in NA12878 sequenced using 2x150bp and 2x250bp reads and GIAB truth set (high confidence regions). Numbers in parenthesis correspond to comparison of PASS only variants. **B)** Barplots showing numbers of unique true positive PASS indels (top) and SNVs (bottom) that overlap different classes of repeats in NA12878 sequenced using 2x150bp (blue) and 2x250bp (orange) kit; **C)** Same as B) but showing overlaps with low mappability regions of the genome.

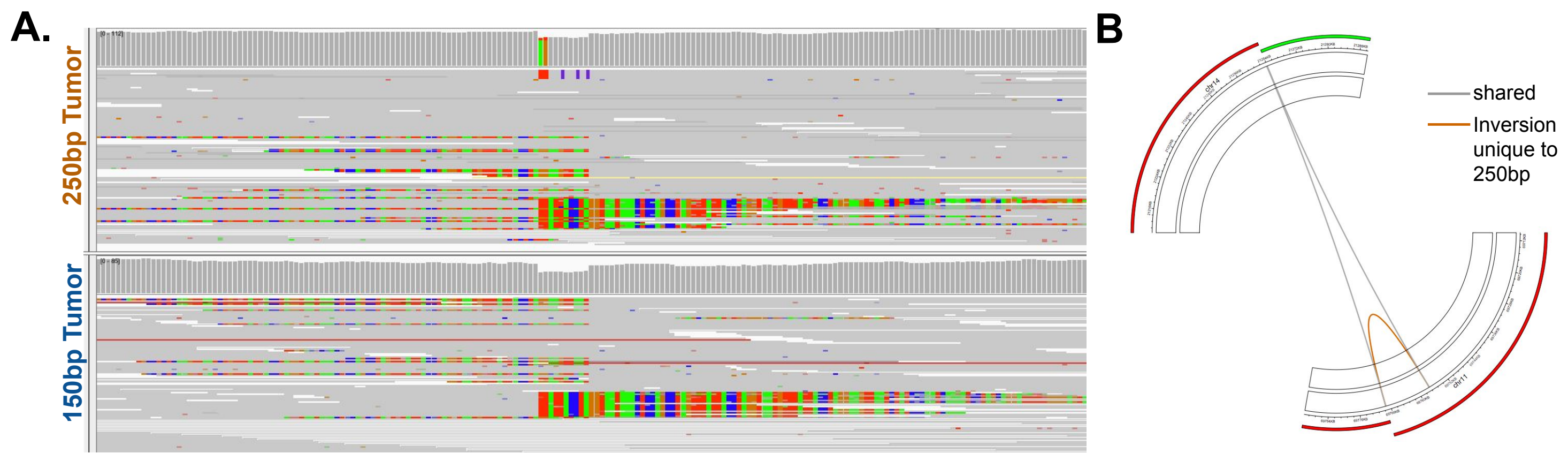
Germline structural variant (SV) size distribution and overlap with truth set. Distribution of SV sizes among NA12878 (PASS) variants that are shared (grey) between 2x150bp and 2x250bp and unique to either 2x150bp (blue) and 2x250bp (orange). For deletions (bottom), the numbers indicate percentage of true positive calls (as determined by 80% reciprocal overlap) when comparing to an in-house union of 3 previously published NA12878 deletions truth sets [3,4,5].



Somatic variant calling



Somatic variants concordance summary and indel size distribution. **A)** Somatic SNV, indel and **B)** SV concordance between 2x150bp and 2x250bp. We have previously defined "AllSomatic" and "HighConf" variants in [6] **C)** Distribution of indel sizes for variants shared(grey) and unique to 2x150bp (blue) and 2x250bp (orange). Some of the unique indels in one category were called by SV callers in the other category.



Examples of the advantages of 2x250bp reads. **A)** Evidence for a complex (replacement) event was found in the alignment of 2x250bp reads (top), but not in the 2x150bp reads (bottom), thus enabling indel callers to correctly identify the event. **B)** Evidence of an inverted translocation event; the inversion was not called in the 2x150bp dataset because average mapping quality of discordant read pairs was <30.

Discussion

Overall, the 2x250bp reads show comparable accuracy to the 2x150bp reads under a variety of different variant calling experiments, while also boosting additional power in specific applications. For example, comparison of the SNP and indel calls to the "Genome In A Bottle" (GIAB) truth set for NA12878 shows increased recall for indels using the 2x250bp reads while preserving the same high standards of precision and recall for SNPs achievable with the 2x150bp data. Specifically, there were approximately half as many false negative indel calls in the 2x250bp as compared to the 2x150bp data. For the cancer cell lines, the somatic SNVs, indels and structural variants from the 2x250bp data show good concordance with the 2x150bp data. For repeat expansion detection, base pair resolution is achievable for a greater number of STRs genome-wide. For SV detection, 2x250bp data often provided better resolution of breakpoints. Taken all together, our extensive experimental comparison demonstrates the benefit of using longer and high-quality reads across the whole spectrum of genomics analysis. As part of future work, we would like to test different alignment and variant calling parameters for longer reads and also look at indels and SVs together.

Cold Spring Harbor Laboratory | Columbia University in the City of New York | Weill Cornell Medical College | The Jackson Laboratory | Memorial Sloan Kettering Cancer Center | Icahn School of Medicine at Mount Sinai | NewYork-Presbyterian | NYU School of Medicine NYU Langone Medical Center | North Shore LIJ | The Rockefeller University | Stony Brook University | Einstein Albert Einstein College of Medicine