Twist Human Pangenome Panel

HUMAN PANGENOME SPIKE-IN DESIGN

As next-generation sequencing (NGS) becomes more prevalent, efforts in human genomics have focused on obtaining a deeper characterization of worldwide human variation. In parallel with this effort, the tools needed to support NGS assays have improved dramatically. At Twist Bioscience, we have focused our efforts on enabling NGS solutions, including demonstrating the effectiveness of NGS-based genotyping panels as an alternative to arrays[1] and enabling population-unbiased genome-wide imputation panels.[2]

One of the latest advances in this area is the human pangenome[3], a new human reference that currently incorporates 49 telomere-to-telomere (T2T) human genomes, with plans to grow into the hundreds. It presents complex variation, difficult regions that were previously unresolved, and diverse worldwide ancestries in a new graph format. In addition, novel alignment and analysis methods have been developed to handle pangenome data.

To support pangenome-based assays and developments, we have designed a human pangenome spike-in panel as an expansion to our Twist Exome 2.0 panel (based on hg38) which we evaluate here. We have leveraged information about the mutational tolerance of baits in our system and combined it with custom probe design strategies. In this way, we are able to target the vast majority of variant bases in the new pangenome reference that overlap coding sequences (See Figure 1A† for an example). Variants ≤5 bp will be covered by our existing baits with >90% efficiency (see Figure 1C). For variants >5 bp, we designed new probes effectively covering 94% of all pangenome variant bases with a total spike-in target set of 2.5 Mb and a bait footprint of 11.7 Mb (see Figure 1B).

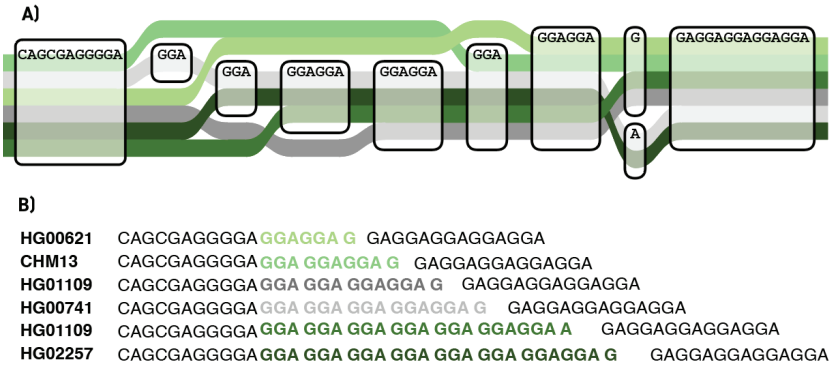
WORKFLOW

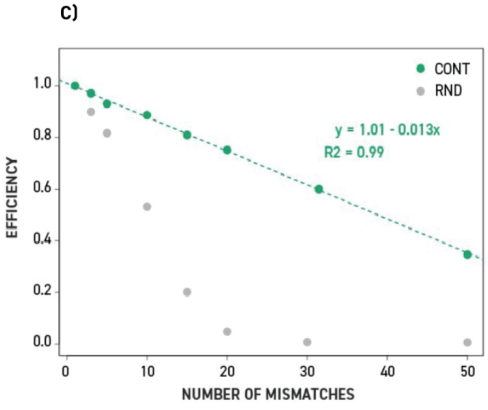
In order to evaluate spike-in performance and compare it with our exome, we performed replicate captures using the exome panel alone or the exome plus the pangenome spike-in (Exome + PanG) on two cell lines included in the pangenome reference set (HG02257 & HG00261). Post-capture libraries were sequenced on the NextSeq 500 with 75 bp paired-end reads and downsampled to 150X sequencing relative to the total target bases in each panel. We then compared the results of standard alignment with BWA[4] and the CHM13 reference[5] (v.2.0) against those based on Giraffe and the CHM13 minigraph cactus pangenome reference (vg v1.49.0 and hprc-chm13v.1.0).[6]

RESULTS ON OVERALL CAPTURE METRICS

Figure 2 shows that while pangenome variants are expected to be enriched for the type of difficult sequences that T2T genomes enable over the hg38 human reference, and many variants indeed have low complexity, the increase relative to the exome in MAPQ filtering (2%) and off bait for the exome + PanG (3%) were quite modest. Pangenome-based methods also slightly improved off bait over standard BWA, further reducing the modest difference in efficiency of the standard exome performance vs the exome + pangenome panel. Other metrics showed little difference with the exome or were improved, which validates the panel despite the difficult targets. The panel achieves excellent overall performance with minimal liability on capture efficiency over targets (e.g. mean target coverage of 62X vs 65X and % bases 30X of 95% vs 96% for exome + PanG vs the exome alone, respectively).

Figure 1. Pangenome Spike-In Design. Variation included in the pangenome graph (panel A†) and overlapping targets in our hg38 exome are directly used to generate baits against variant sequences not contained in hg38 (panel B). Bait coverage, content, and overlap with our exome are optimized based on experimental data regarding the effects of mismatches on capture efficiency (panel C, CONT: contiguous mutations, RND: dispersed mutations, as described in our whitepaper).





†Only a subset of all 49 haplotypes are shown for chr20 pos 46022976 in hg38, overlapping a variant of the SLC12A5 potassium-chloride transporter with OMIM links to developmental and neurological conditions. Figure generated with pangenome graph hprc-v1.1.-hg38, vg version 149.0, and SequenceTubemap.

Figure 2. Picard capture metrics showing effects of methods (standard BWA alignment vs. pangenome-based alignment with Giraffe, GRF) and panels (Exome vs. Exome + PanGenome spike-in).

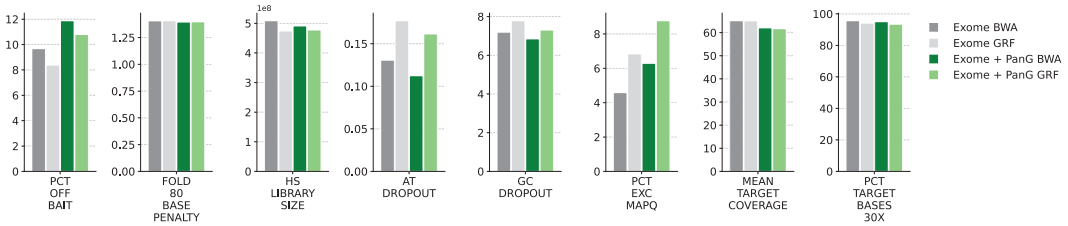
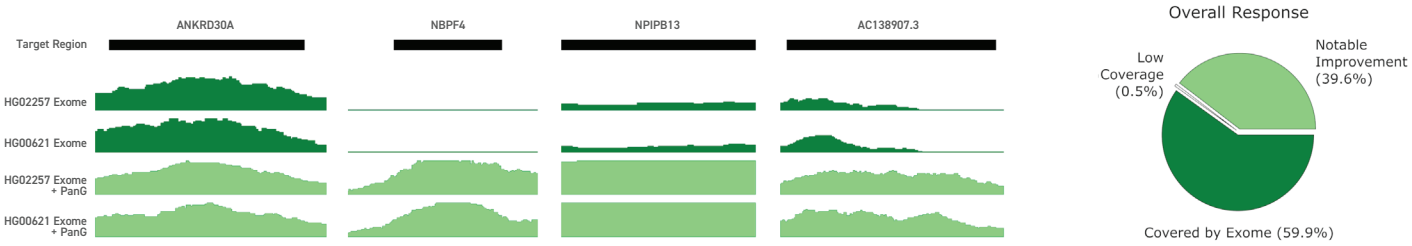


Figure 3. Spike-in response over 207 coding sequences in CHM13 unique regions. Examples of coverage patterns and improvements observed with exome alone (Exome) and exome with pangenome spike-in (Exome + PanG) over captures in different cell lines. Some regions were well covered by both panels (leftmost), while others showed a marked improvement with the pangenome spike-in. The pie chart to the right shows the proportions of the main patterns. Only 1 region showed mean coverage <20X (5-10X among bases).



COVERAGE IMPROVEMENTS IN PANGENOME PANEL REGIONS

The number of sizable variants present over coding sequences in any given genome is small. This makes statistical comparisons of exome vs pangenome capture differences difficult without the use of various cell lines. Indeed, genotyping results over regions expanded with PanG variants (using giraffe + deep variant[7,8]) yielded only 309 variant calls, all of which agreed between exome and exome + PanG panels even where cell lines had variants of 50bp or more.

However, there is a set of 207 coding regions, brought about mainly by improvements in T2T genomes, that are not present in hg38 when compared with CHM13, but are assayed for all samples in the pangenome. These CHM13 unique regions[5] thus serve as an excellent model of sizable variants in the pangenome not present in hg38.

Figure 3 shows that the vast majority of CHM13 unique coding regions are either well covered by the Twist Exome 2.0 panel (~60%, providing evidence the exome alone already captures many pangenome variants, see introduction), or are noticeably improved by the spike-in (~40%). Only 1 region showed mean coverage <20X in both panels. With the pangenome spike-in, an additional 1.4% and 6.5% of regions in all coding sequences in the genome (beyond those in CHM13 unique regions) saw a statistically significantly improved mean coverage >1.5X and >3X the interquartile range for exome alone, respectively. There were no statistically significant drops with the pangenome spike-in and the more complex mapping landscape handled by pangenome alignment methods.

CONCLUSION

Despite the difficult nature of variants targeted in the pangenome-aware exome, target enrichment performance is markedly improved in pangenome variant regions without significantly degrading the performance of the panel elsewhere.

The pangenome spike-in provided clear improvements in capture for large, complex, population-informed variation.

The methods and approaches developed and tested here, including the full set of pangenome variant types, can easily be applied to related problems such as the careful characterization of a pangenome-informed variant landscape for focused sub-collections of challenging targets, other species (where pangenomes can be particularly powerful for encapsulating complex genomes and varieties of agricultural importance), other sequencing technologies (e.g., targeted long-read sequencing), and a host of other applications.

Interested in learning more about Pangenome Target Enrichment Panels? Contact us for more information.

twistbioscience.com/ngs
sales@twistbioscience.com

REFERENCES

1. Capture-based SNP Genotyping with Twist Target Enrichment Panels. Twist Bioscience https://www.twistbioscience.com/resources/application-note/capture-based-snp-genotyping-twist-target-enrichment-panels (2020).

2. Abecasis, G. No More Arrays: Genotyping by Sequencing Enables Economical and Improved Association Studies. Twist Bioscience https://www.twistbioscience.com/resources/webinar/no-more-arrays-genotyping-sequencing-enables-economical-and-improved-association (2021)

3. Liao, WW., Asri, M., Ebler, J. et al. A draft human pangenome reference. Nature 617, 312–324 (2023). https://doi.org/10.1038/s41586-023-05896-x

4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009). https://doi.org/10.1093/bioinformatics/btp324

5. Nurk, S. et al. The complete sequence of a human genome. Science 376, 44–53 (2022). https://doi.org/10.1126/science.abj6987

6. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science 374, abg8871 (2021).https://doi.org/10.1126/science.abg8871

7. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 36, 983–987 (2018). https://doi.org/10.1038/nbt.4235

8. Rakocevic, G. et al. Fast and accurate genomic analyses using genome graphs. Nat. Genet. 51, 354–362 (2019). https://doi.org/10.1038/s41588-018-0316-4