Long-read capture with Twist target enrichment system

Tina Han, Holly Corbitt, Leonardo Arbiza, Esteban Toro, Chad Locklear.

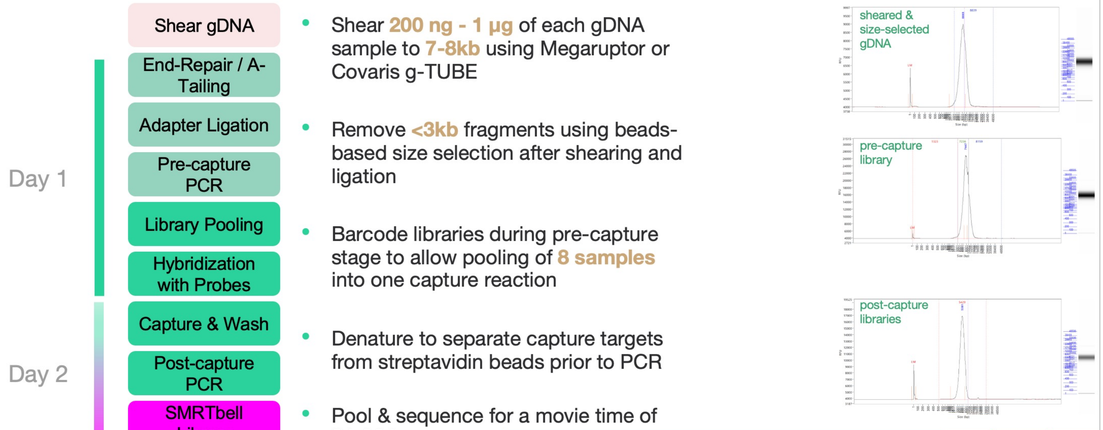
Twist Bioscience, South San Francisco, California, USA

1. Introduction

Targeted resequencing allows for high-resolution characterization of gene regions at a scale and cost that is more accessible than whole genome sequencing. While long-read PacBio HiFi sequencing has been shown to accurately and comprehensively interrogate complex clinically actionable loci, studies have been primarily focused on single genes using PCR amplicon-based methods. Here we describe a method to leverage Twist Bioscience target enrichment workflow for gene panels sequenced with HiFi reads. This poster presents the content and the performance of 2 alliance panels - a 50-gene pharmacogenomics panel and a nearly 400-gene panel of challenging, medically relevant "dark genes" - developed in collaboration with leading institutions.

2. Methods

We designed gene panels of various target sizes, ranging from 0.2-22 Mb. Probes were optimized using a proprietary algorithm to enable balanced capture of complex regions while at the same time reducing capture of off-target sequences. Our long-read hybrid capture protocol[1] starts with 200-1000 ng fragmented gDNA that were sheared using mechanical fragmentation, i.e. Diagenode Megaruptor or Covaris g-TUBE. After end-repair and a-tailing, truncated Y-shaped adapters were ligated to adapted gDNA. A pair of 10-bp unique dual indices (UDIs) for sample barcoding are added during PCR. 4-8 samples can be pooled in a single tube for overnight hybridization. The post-capture libraries then undergo SMRTbell library preparation using SMRTbell® prep kit 3.0 and sequencing on PacBio Sequel IIe with a 30-hour movie time. Depending on target size, up to 400 samples may be multiplexed and sequenced in one SMRT Cell with HiFi read length of 5-10 kb.



Name	Movie Time (h...	Status	Total Bases (Gb)	P0	P1	P2
Tina-SMRT-5 (CCS)	30	Complete	657.66	32.7	66.4	0.9
Tina-SMRT-6 (CCS)	30	Complete	741.05	28.5	70.5	1.1
Tina-SMRT-7 (CCS)	30	Complete	665.10	39.5	59.7	0.8

SMRT Link was used to generate HiFi reads, remove PCR duplicates, and demultiplex, and a PacBio pipeline was used to call variants for individual samples. The PacBio target enrichment workflow is publicly available on GitHub: https://github.com/PacificBiosciences/HiFiTargetEnrichment

3. Long-Read Pharmacogenomics (PGx) Panel

Pharmacogenes, including HLA genes and CYP2D6, are notoriously difficult to genotype via array or via short-read sequencing as they tend to have low sequence complexity and/or highly homologous pseudogenes. CYP2D6 is also subject to common partial- or whole-gene duplications and rearrangements. We set forth to develop a 50-gene pharmacogenomics panel. In addition to the nuclear content, the full mitochondrial genome is covered to enable simultaneous detection of herteroplasmy.

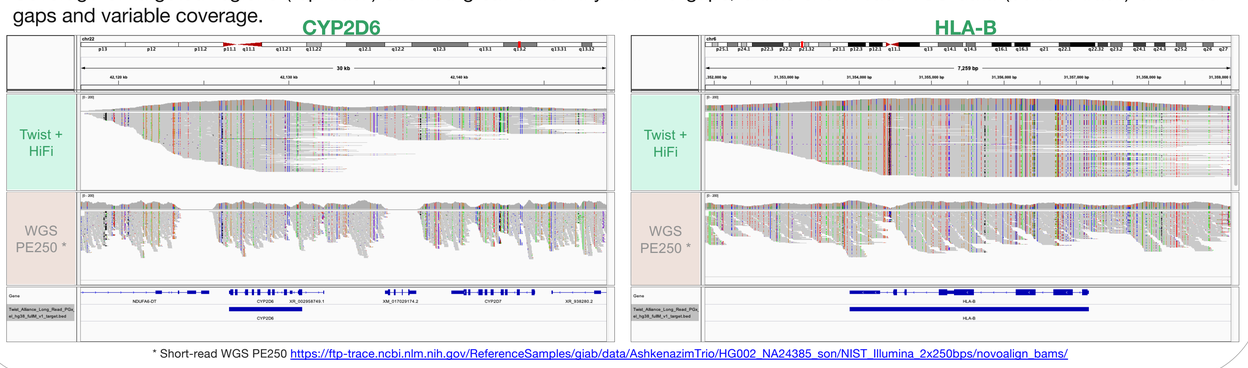
Panel of 50 Clinically Significant Genes *(full gene coverage in bold)*

CYP	HLA	Mitochondrial	Others		
CYP1A2	**HLA-A**	Entire mitochondrial genome is targeted to cover variants in mt-RNR1	**ABCB1**	F2	**OPRD1**
CYP2B6	**HLA-B**		**ABCG2**	F5	**OPRK1**
CYP2C19	**HLA-DQA1**		ADD1	**G6PD**	**OPRM1**
CYP2C8	**HLA-DRB1**		**ADRA2A**	**GBA**	POLG
CYP2C9			ANKK1	GRIK4	**RYR1**
CYP2D6			**APOL1**	HTR2C	**SLC6A4**
CYP3A4			**BCHE**	**IFNL3**	**SLCO1B1**
CYP3A5			CACNA1S	**MTHFR**	**TPMT**
CYP4F2			CFTR	**NAGS**	**UGT1A1**
			COMT	**NAT2**	**UGT2B15**
			CTBP2P2	**NUDT15**	**VKORC1**
			DPYD		YEATS4
			DRD2		

24 GeT-RM Coriell samples were sequenced on 1 SMRT Cell 8M on the Sequel IIe system. Samples had on average 150k HiFi reads, with a mean read length of ~5.3 kb. Only 2% of duplicates were removed from downstream analysis. Across all targets, a mean target coverage of 190x was achieved. Across all samples, 96% of target regions exceeded 20x coverage and 93% of target regions exceeded 30x coverage.

Panel Size	Number of Genes	Samples per SMRT Cell 8M	HiFi Reads per Sample	Mean Target Coverage	Depth of Coverage	Fold Enrichment	Duplicate Rate	Mean HiFi Read Length
2 Mb	50 genes	24	150k	190x	96% ≥20x 93% ≥30x	784	2%	5.3 kb

Coverage of targeted regions (top track) showed great uniformity with no gaps, unlike short-read WGS data (bottom track) exhibited gaps and variable coverage.



* Short-read WGS PE250 https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/novoalign_bams/

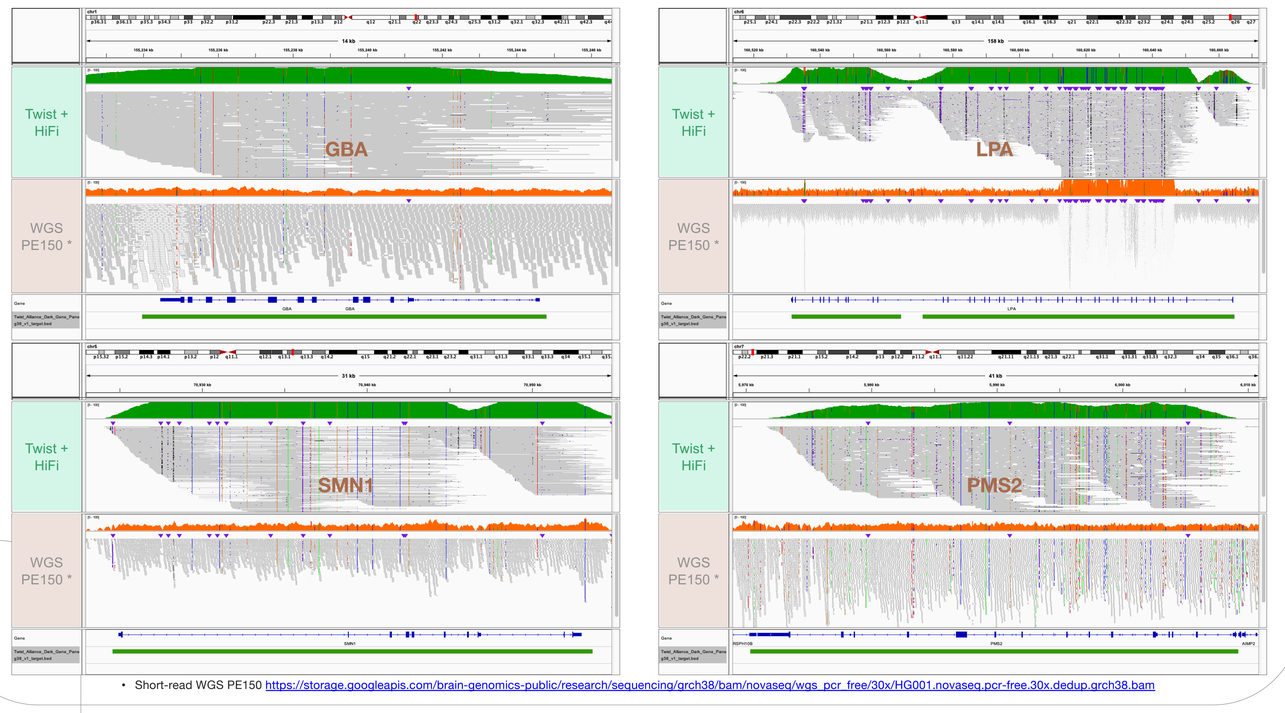
4. Long-Read Dark Genes Panel

The repetitive nature and polymorphic complexity of 389 medically relevant genes poses a challenge for their accurate analysis in a clinical setting, but ~70% can be resolved over HiFi assemblies[2]. This panel provides full gene coverage for 389 difficult-to-call genes, including many genes in "NGS dead zone" that are difficult to sequence or map with short-reads[3,4]. These genes are reported to impact a range of diseases including cardiovascular, neuropathies, immunodeficiencies, vision related diseases, also included are cancer driver genes (e.g. PTEN).

*A4GALT, ABCG8, ABO, ABR, ADAMTS10, ADAMTSL2, AFP, AGL, AGRN, ALOXE3, ANKRD11, ANO7, APOBEC1, APOBEC3H, APOC1, APOC2, APOC4, ARHGEF10, ASIP, ATPAF2, AXIN1, B3GAT3, BAX, BFSP2, BLOC1S3, BRAF, BSG, BTRC, C1R, C3, CABIN1, CALR3, CANT1, CASP10, CBR3, CBS, CCL3L1, CD247, CD320, CD4, CD55, CDH15, CDH17, CEL, CFC1, CFC1B, CFD, CFHR1, CFHR3, CHL1, CHMP1A, CHRNA4, CLCN7, CLIP2, CNR2, COL18A1, COL6A1, COL6A2, COX14, COX6B1, CR1, CREB3L3, CRYAA, CTDP1, CYB5R3, **CYP2D6**, CYP2G1P, CYP4F12, CYP4F3, D2HGDH, DAXX, DAZL, DCLRE1C, DEAF1, DGCR6, DIP2C, DLGAP2, **DMPK**, DNMT3L, DOK7, DPP6, DPY19L2, DRD4, DSPP, DUX4, DUX4L1, ECHS1, EEF1A2, EHMT1, EIF2B5, EIF4E, ELANE, ENO3, ESPN, ESRRA, ETFB, ETHE1, EXTL2, F7, FAM20C, FAT1, FCGR1A, FCGR2B, FCGR3A, FGF3, FGFRL1, FKBP8, FLAD1, FLG, FLT4, FOXN1, FSCN2, FTCD, FUT1, FUT3, **FXN**, G6PC3, GAK, GALNT9, GALR1, GALT, **GBA**, GCGR, GCSH, GDF3, GIP, GIPC3, GNPTG, GOLGA3, GP1BA, GP6, GPI, GPIHBP1, GRIN1, GRK1, GSTM1, GTF2I, GTF2IRD2, GUSB, GYPA, GYPB, GYPE, H19, HBG1, HBM, HCN2, HCN3, HES7, **HLA-B**, **HLA-DQB1**, **HLA-DRB1**, HMGCL, HMX1, HNF1A, HOMER2, HOXB8, HPD, HSD11B2, HYAL1, HYDIN, IFITM3, IFNL3, IGHA1, IGHG1, IGHG2, IGHM, IGHV3-21, IGKC, IGKV1-5, IKBKB, IKZF1, IMPA1, INPP5D, INPP5E, INSL3, INSR, JAG2, KANSL1, KATNAL2, KCNE1, KCNJ18, KCNV2, KDM2B, KIR2DL1, KIR2DL3, KIR3DL1, KISS1, KISS1R, KLF11, KLF14, KLK4, KMT2C, KNG1, KRTAP1-1, LAMB1, LBR, LCE3B, LHFPL5, LIPN, LIX1, LMF1, LMNB2, **LPA**, LRIG2, LRPAP1, LZTFL1, MAFA, MAN1B1, MAP2K3, MARVELD2, MASP2, MBOAT7, MC1R, MDK, MEST, MLC1, MLPH, MOGS, MPG, MRC1, MST1R, MUC1, MUC16, MUC3A, MUC4, MUC5B, MUSK, MYO9B, MYOT, MYT1, NACA, NAIP, NAPRT, NBEAP1, **NCF1**, NCF1C, NCR3, NDUFA6, NDUFAF1, NDUFB1, NDUFV3, NFKBIL1, NLRP12, NLRP2, NLRP7, NOD1, NOTCH2, NPM1, NPPA, NSMF, NUTM2B, NUTM2D, OCLN, OPRL1, OR12D2, OR4F5, OR51A2, ORC6, P2RX2, P2RX5, PADI4, PAPSS2, PCBP1, PCCB, PCDHA10, PCMT1, PDE4DIP, PDE6B, PDLIM3, PDPK1, PDSS1, PEX5, PGAM5, PHKG2, PIGV, PKD1, PKN3, PLA2G10, PLTP, **PMS2**, PNKP, POLG2, PPIA, PPIP5K1, PRG4, PRKCG, PRODH, PROZ, PRSS2, PSPH, PTEN, PTK6, PTPRC, PTPRN2, PTPRQ, PXDN, RFX2, RGPD3, RHCE, RHOA, RNF212, RNF213, RPIA, RPL22, RPN1, RPS17, SAR1B, SBDS, SBK3, SDHA, SEC63, SEMG1, SERPINF2, SH2B1, SHANK2, SHANK3, SIGLEC16, SIRT3, SLC17A5, SLC22A1, SLC22A12, SLC26A9, SLC27A4, SLC27A5, SLC29A4, SLC5A11, SLC6A18, SLC6A3, SMG1, **SMN1**, **SMN2**, SMOC2, SNORD64, SNTG2, SOHLH1, SPATA31C1, SPI1, SPRN, SRGAP2, SRR, SSTR5, STK11, STXBP2, SULT1A1, SUZ12, TAPBP, TAS2R45, TAS2R46, TBXA2R, TCF3, TERT, TFPT, THBS2, TJP2, TM4SF19, TMC6, TMEM114, TNNI3, TNNT1, TNNT3, TPCN2, TPO, TRAPPC10, TRBV9, TRMT1, TRPM4, TTC37, TTLL1, TUBGCP6, TWIST2, TYK2, TYMS, U2AF1, UGT2A1, UGT2A2, UGT2B17, UGT2B28, UNKL, USP8, UVSSA, VANGL1, VKORC1, VPS53, ZAN, ZNF141, ZNF407, ZNF419, ZNF469, ZNF479*

4 Coriell samples were sequenced on 1 SMRT Cell 8M on the Sequel IIe system. Samples had on average 893k HiFi reads, with a mean read length of ~5.2 kb. Only 3% of duplicates were removed from downstream analysis. Across all targets, a mean target coverage of 75x was achieved. Across all samples, 93% of target regions exceeded 10x coverage and 90% of target regions exceeded 20x coverage.

Panel Size	Number of Genes	Samples per SMRT Cell 8M	HiFi Reads per Sample	Mean Target Coverage	Depth of Coverage	Fold Enrichment	Duplicate Rate	Mean HiFi Read Length
22 Mb	389	4	893k	75x	93% ≥10x 90% ≥20x	54	3%	5.2 kb



- Short-read WGS PE150 https://storage.googleapis.com/brain-genomics-public/research/sequencing/grch38/bam/novaseq/wgs_pcr_free/30x/HG001.novaseq.pcr-free.30x.dedup.grch38.bam

5. Conclusions

We demonstrate that our long-read capture method efficiently enables comprehensive coverage of gene targets using Coriell samples run with multiple gene panels of varying sizes, two of which we highlight here include complex regions like CYP2D6, HLA, SMN1, and LPA. This long-read hybrid capture protocol can be utilized with Twist custom or fixed gene panels to efficiently capture genes of interest using long-read sequencing. Optional secondary panels (spike-ins) can also be easily added during hybridization for additional content. The demonstrated method allows for scalable and cost-efficient hybrid capture with long read lengths, minimizing coverage bias, and maximizing accuracy to fully capture all variant types. This includes structural variation and haplotype phasing information which are inaccessible to short-read and Sanger sequencing.

6. Reference

1. Twist Long Read Library Preparation and Standard Hyb v2 Enrichment https://www.twistbioscience.com/resources/protocol/long-read-library-preparation-and-standard-hyb-v2-enrichment
2. Sedlazeck et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nature Biotech* (2022)
3. Mandelker et al Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genetics in Medicine* (2016)
4. Wenger et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotech* (2019)

7. Supplementary Materials

1. Target bed file of **Twist Alliance Long-Read PGx Panel**: https://www.twistbioscience.com/resources/data-files/twist-alliance-long-read-pgx-panel-bed-file
2. Target bed file of **Twist Alliance Dark Genes Panel**: https://www.twistbioscience.com/resources/data-files/twist-alliance-dark-genes-panel-bed-file
3. Demo datasets: https://www.pacb.com/connect/datasets/#targeted-datasets

8. Acknowledgements

The authors would like to thank Stuart A. Scott and his group at Stanford University for the content curation of Pharmacogenomics Panel and Fritz Sedlazeck and his group at Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) for Dark Genes Panel. The authors would also like to thank the technical assistance of Sarah Kingan, Nina Gonzaludo, John Harting, Xiao Chen, Christine Lambert, and Primo Baybayan at PacBio.