# Clinical Validation of a Haplotyped Whole Genome Sequencing Assay Using PacBio HiFi Sequencing

**David Corney, Carrie Ziemniak, Wei Wang, Yu Qiu, Christopher Mozdzierz, Jeffrey Otto, Haythem Latif**

GENEWIZ from Azenta Life Sciences, South Plainfield, NJ 07080

## Abstract

Next generation sequencing (NGS) has revolutionized genomics and enabled unprecedented insights into the molecular basis of health and disease. Short-read NGS technologies have become integral to the diagnosis of genetic diseases, and in supporting clinical trials. Unfortunately, while current short-read technologies are more cost-effective and scalable than ever before, short read-lengths are limited in resolving complex genomic regions, detecting structural variations, and covering repetitive sequences. Additionally, short-reads may not be able to accurately differentiate between trans and cis compound heterozygosity. This is critical in human germline diploid studies, since whether variants are present on the same or different chromosomes may confer vastly different phenotypes. While such challenges can be mitigated by long-read sequencing approaches, widespread clinical adoption is limited due to challenges in sample handling, scalability, and cost.

To address these limitations, we report the CLIA/CAP validation of a clinical long-read WGS sequencing assay for the comprehensive genome-wide detection and phasing of SNVs, and INDELs. High molecular weight DNA was extracted from frozen EDTA whole blood, followed by DNA fragmentation to ~15-20 kb by automated pipette-mediated mechanical shearing, and SMRTbell® library preparation. After sequencing each sample on a PacBio® Revio™ SMRT® Cell, ~20-30x average genome-wide coverage was obtained. Variant calling accuracy was assessed by comparing variant calls in healthy donors to those obtained by Illumina® PCR-free WGS assay, as well as analysis of Genome In A Bottle (GIAB) well-characterized reference samples. Overall, the variant calling concordance F1 score was 0.99 and 0.98 for SNVs and INDELs, respectively. Phasing accuracy was performed by trio analysis on specific genes of interest.

This validated long-read WGS assay overcomes key limitations of short-read technologies, enabling comprehensive genome-wide detection and phasing of all variant classes. By leveraging automation and high-throughput workflows, this assay provides a powerful tool for precision genomic medicine, with the potential to improve diagnostic accuracy, treatment selection, and patient outcomes.

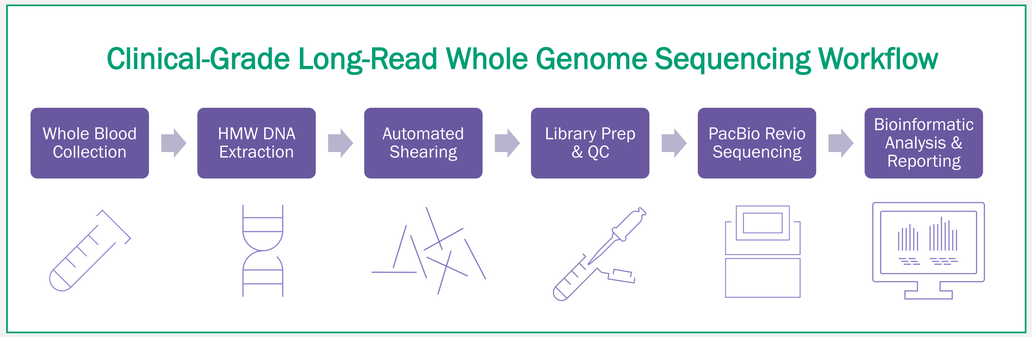



**Figure 1. Schematic of clinically validated PacBio WGS workflow.** High molecular weight (HMW) DNA is obtained from whole blood followed by automated DNA shearing using a Hamilton® STAR™ to target ~15-20 kb average insert size. Subsequently, libraries are generated and sequenced on a Revio sequencer followed by analysis for variant calling, and phasing of SNPs and INDELs.

A.

| SAMPLE ID | YIELD (ug) | A260/280 RATIO | $GQN_{10\ kb}$ |
|---|---|---|---|
| Sample 1 | 6.6 | 1.85 | 9.3 |
| Sample 2 | 7.5 | 1.87 | 9.6 |
| Sample 3 | 5.3 | 1.87 | 9.2 |
| Sample 4 | 8.4 | 1.84 | 9.7 |
| Sample 5 | 9.6 | 1.81 | 9.8 |

B.

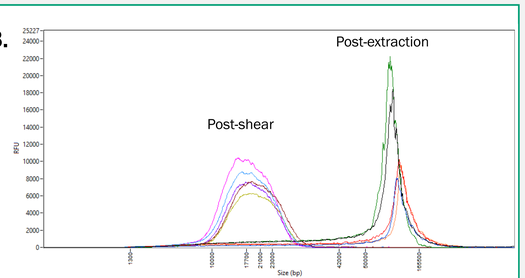



**Figure 2. HMW DNA QC and automated pipette shearing.** (A) Following extraction, nucleic acids are QC'd by NanoDrop™, Qubit™, and Femto Pulse™. QC results from five representative whole blood samples is shown - showing DNA with high yield, quality, and integrity obtained from 400 µl whole blood. (B) Femto Pulse plot showing HMW DNA with average ~100 kb sizing post-extraction ("Post-extraction"), and uniform ~15-20 kb sized DNA following automated pipette-mediated shearing using Hamilton Microlab STAR ("Post-shear"). A single shearing condition was used for all samples.

A.




B.

| SAMPLE ID | ACCURACY | | |
|---|---|---|---|
| | SNV | INDEL | SNV/INDEL COMBINED |
| Sensitivity | 99.7% | 91.4% | 98.5% |
| Specificity | 98.1% | 88.2% | 96.3% |
| Accuracy | 99.1% | 90.0% | 97.6% |
| Positive Predictive Value (PPV) | 98.7% | 90.4% | 97.5% |
| Negative Predictive Value (NPV) | 99.6% | 89.4% | 97.8% |
| F1 Score (GIAB) | 99.9% | 98.0% | N/A |

C.

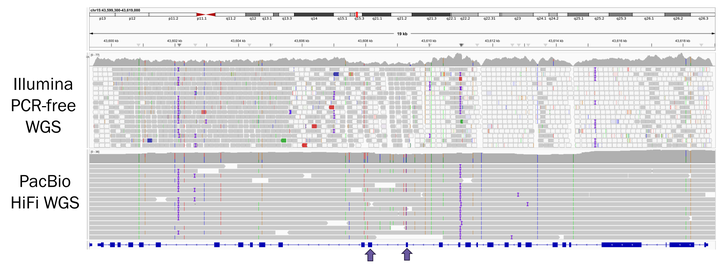
| SAMPLE ID | INTRA-BATCH CONCORDANCE | |
|---|---|---|
| | SNV | INDEL |
| NA24385 | 98.7% | 82.5% |
| NA24149 | 98.6% | 82.5% |
| Sample 1 | 98.8% | 86.8% |

D.

| SAMPLE ID | INTER-BATCH CONCORDANCE | |
|---|---|---|
| | SNV | INDEL |
| NA24385 | 98.7% | 83.7 |
| NA24149 | 98.6% | 83.0% |
| Sample 1 | 98.8% | 87.7% |

**Figure 3. Summary of analytical validation of HiFi WGS using PacBio Revio sequencer.** (A) Raw data output per Revio SMRT Cell. Values are shown for HMW DNA obtained from blood, as well as partially degraded DNA obtained from Coriell Genome In A Bottle (GIAB) samples. (B) Summary of analytic performance characteristics. All metrics were obtained by comparing SNV and INDEL calls by this assay to Azenta's clinical-grade PCR-Free Illumina WGS assay across 20 donor samples, except the F1 score which was calculated compared to public GIAB data. (C, D) Concordance between intra- and inter-batch triplicate reactions.

A.




B.

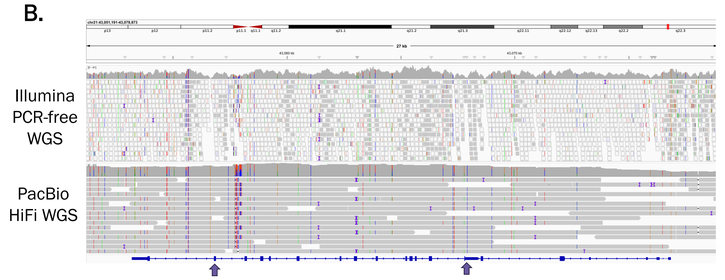



**Figure 4. PacBio HiFi data enables detection of variants within challenging regions of the genome compared to short-read WGS.** Two examples of medically-relevant genes which have been reported to be challenging to sequence by short-read platforms are shown. Arrows indicate exonic regions with low short-read coverage. (A) STRC, which is involved in deafness. (B) CBS, which has been associated with homocystinuria.

A.

| SAMPLE ID | PHASED VARIANTS IN TRUTHSET | CORRECTLY PHASED VARIANTS | PHASING CONCORDANCE |
|---|---|---|---|
| NA12878 | 190 | 190 | 100% |
| NA24385 | 138 | 138 | 100% |
| NA24631 | 72 | 72 | 100% |
| NA19238 | 229 | 227 | 99.1% |
| NA19239 | 178 | 178 | 100% |

B.

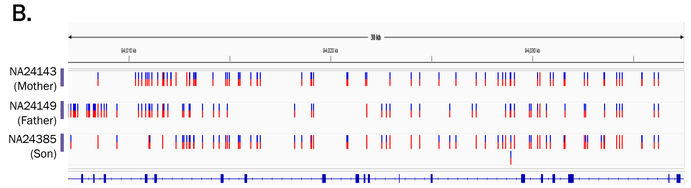



**Figure 5. The analysis pipeline accurately phases SNV and INDEL variants.** (A) Summary of phasing accuracy in a ~120 kb gene of interest compared to public data. (B) Example of variant phasing across a 30 kb region within a GIAB Trio (NA24143/NA24149/NA24385).

## Conclusions

- Long-reads enable detection of challenging variants and allows for haplotyping which may be especially critical in rare disease.
- Validation studies demonstrate up to 99% variant calling accuracy compared to reference datasets, and phasing concordance.
- Regulatory approval has been granted and the test is on Azenta's clinical test menu.

**Acknowledgements:** We are grateful to Nevil Patel, Ping Xu, Mia Sutton, and Ye Wang (Azenta) for automation and sequencing assistance, and Jay Morris, Gregory Young, and Andrew Mason (PacBio) for technical support.

# Direct Detection and Counting of Trinucleotide Repeat Expansion Using CRISPR-Cas9 and PacBio HiFi Sequencing

**Bhaven Mehta[1], David Corney[1], Wei Wang[1], Yu Qiu[1], Jian Chen[2], Hoda Bahgat[2], Faraz Zaidi[2], Christopher Mozdzierz[1], Haythem Latif[1]**

[1]GENEWIZ from Azenta Life Sciences, South Plainfield, NJ 07080

[2]CHDI Foundation, New York, NY 10001

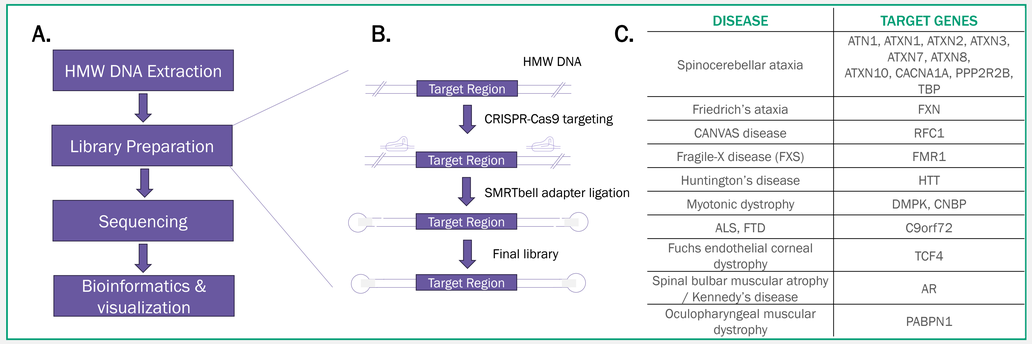
## Abstract

Trinucleotide repeat expansions are a major cause of neurological and developmental disorders, including Huntington's disease, fragile X syndrome, and various spinocerebellar ataxias. When the number of repeats exceeds a certain pathogenic threshold length, they can lead to toxic gain-of-function or loss-of-function effects. Historically, PCR and Southern blot approaches have been used to identify repeat expansions. However, there are significant technical challenges associated with these methods, including successfully amplifying through hundreds or thousands of base pairs of repetitive sequences and failed PCR due to high GC content. As a result, these approaches suffer from poor accuracy and resolution of the number of repeats. Consequently, the loci containing these genes are commonly referred to as “dark” regions of the genome and are poorly characterized.

To circumvent these challenges, we have tested a novel PCR-free CRISPR-Cas9 approach to enrich 20 genes commonly involved in repeat expansion disorders followed by long-read PacBio® sequencing to successfully sequence and align repetitive sequence. High molecular weight DNA was extracted from frozen EDTA whole blood or fresh frozen tissue, followed by PureTarget library preparation. In this workflow, DNA is first digested with Cas9 and guide RNAs to enrich 20 target genes, followed by ligation of SMRTbell® adapters on the cut end. After removal of non-ligated templates, libraries are pooled and sequenced on a PacBio sequencer.

Here, we describe the accuracy testing results of this novel approach. Accuracy of calling the number of trinucleotide repeats was established by processing previously characterized samples from Coriell with known expansions in DMPK, ATX1, ATXN3, FMR1, and FXN. Additionally, DNA extracted from blood of healthy donors without any known expansions, and from brain tissue of Huntington disease individuals was performed and compared to whole genome PacBio data.

The CRISPR-Cas9 approach combined with long-read PacBio sequencing offers a powerful method for characterizing trinucleotide repeat expansions, overcoming the limitations of traditional techniques. This novel approach paves the way for improved understanding of the genotype-phenotype and molecular basis of neurological and developmental disorders caused by repeat expansions.




C.

| DISEASE | TARGET GENES |
|---|---|
| Spinocerebellar ataxia | ATN1, ATXN1, ATXN2, ATXN3, ATXN7, ATXN8, ATXN10, CACNA1A, PPP2R2B, TBP |
| Friedrich’s ataxia | FXN |
| CANVAS disease | RFC1 |
| Fragile-X disease (FXS) | FMR1 |
| Huntington’s disease | HTT |
| Myotonic dystrophy | DMPK, CNBP |
| ALS, FTD | C9orf72 |
| Fuchs endothelial corneal dystrophy | TCF4 |
| Spinal bulbar muscular atrophy / Kennedy’s disease | AR |
| Oculopharyngeal muscular dystrophy | PABPN1 |

**Figure 1.** (A) Schematic of repeat expansion analysis using CRISPR-Cas9. (B) High molecular weight (HMW) DNA is dephosphorylated followed by CRISPR-Cas9 targeting with gRNAs against either end of the target regions. Subsequently, PacBio SMRTbell adapters are ligated to generate a final library which is sequenced on a PacBio Revio™ sequencer. Bioinformatic analysis and visualization is performed to identify repeat expansions. (C) Target genes contained within the PacBio PureTarget™ panel.

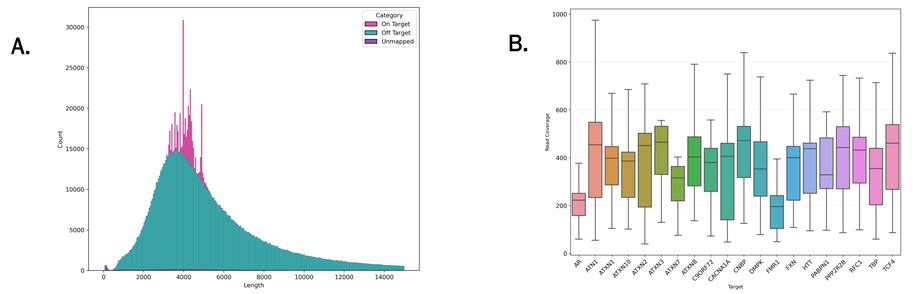
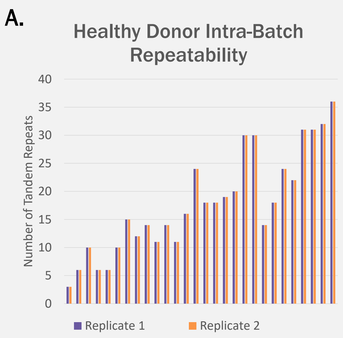



**Figure 2. Read alignment characteristics.** (A) Plot showing read mapping properties. A large proportion of reads are off-target. Nevertheless, a significant number of reads align to the 20 on-target regions included within the PureTarget panel. (B) Boxplot illustrating on-target coverage on a per-gene basis. Samples included a mixture of HMW and partially degraded samples which likely contributed to greater variability. A total of 22 samples were analyzed across 2 independent processing batches, with 99.8% of regions having at least 50x coverage.




B.

| GENE | REPEAT UNIT | SAMPLE ID | BATCH | ALLELE 1 | | ALLELE 2 | |
|---|---|---|---|---|---|---|---|
| | | | | Expected Repeats | Observed Repeats | Expected Repeats | Observed Repeats |
| DMPK | CAG | NA03697 | 1 | N/A | 12 | < 500 | 383 |
| | | | 2 | | 12 | | 384 |
| | | | | | 12 | | 383 |
| | | | | | 12 | | 383 |
| ATXN1 | CAG | NA06926 | 1 | 29 | 30 | 52 | 54 |
| | | | 2 | | 30 | | 54 |
| | | | | | 30 | | 54 |
| | | | | | 30 | | 54 |

**Figure 3. Assay repeatability.** (A) HMW DNA from a healthy donor was analyzed twice within the same processing batch. Identical number of tandem repeats was observed for each replicate in every gene. (B) Two Coriell samples were analyzed in triplicate in one batch and independently in another batch with largely identical results consistent with the expected number of repeats.

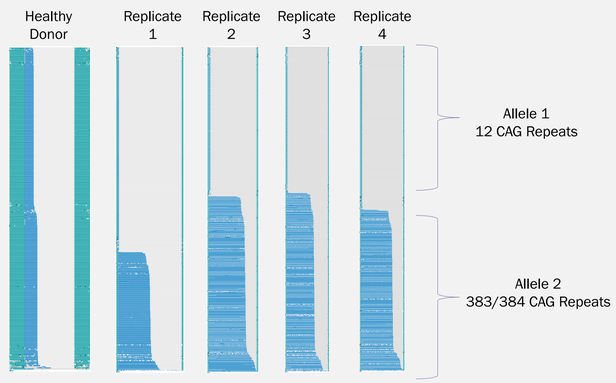



**Figure 4. Example illustrating DMPK repeat expansion.** Results are shown in a waterfall plot for each of four independent technical replicates of Coriell sample NA03697 alongside a healthy donor. Heterogeneity is observed with regards to repeat number of allele 2 of NA03697. Teal: flanking sequence; Blue: CAG repeat.

A.

| GENE | REPEAT UNIT | SAMPLE ID | ALLELE 1 | | ALLELE 2 | |
|---|---|---|---|---|---|---|
| | | | Expected Repeats | Observed Repeats | Expected Repeats | Observed Repeats |
| ATXN3 | GCT | NA06153 | 23 | 20 | 71 | 69 |
| FMR1 | CGG | NA06905 | 23 | 23 | 70 | 79 |
| FXN | GAA | NA15850 | 650 | 634 | 1030 | 705 |

B.

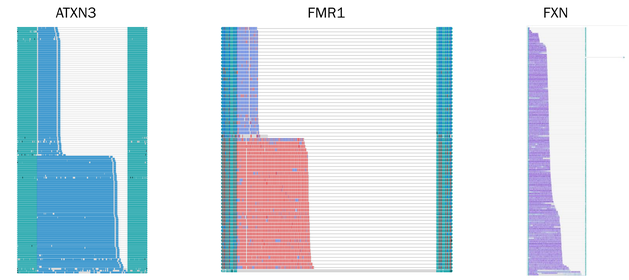



**Figure 5. Concordance of PureTarget in Coriell reference samples.** (A) Summary of tandem repeats in three Coriell samples. (B) Waterfall plots of ATXN3, FMR1, and FXN repeats identified in part A, including CpG methylation status in FMR1. Teal: flanking sequence; Blue/purple: repeat unit; Red/blue (FMR1): methylated/unmethylated CpG.

A.

| GENE | REPEAT UNIT | DISEASE STATUS | ALLELE 1 | | ALLELE 2 | |
|---|---|---|---|---|---|---|
| | | | HiFi WGS Repeats | PureTarget Repeats | HiFi WGS Repeats | PureTarget Repeats |
| HTT | CAG : CCG | HD | 19 : 10 | 19 : 10 | 44 : 9 | 44 : 9 |
| | | HD | 22 : 9 | 22 : 9 | 49 : 9 | 50 : 9 |
| | | Control | 17 : 9 | 17 : 9 | 20 : 12 | 20 : 12 |
| | | Control | 17 : 12 | 17 : 12 | 21 : 12 | 21 : 12 |
| | | Control | 19 : 12 | 19 : 12 | 19 : 12 | 19 : 12 |

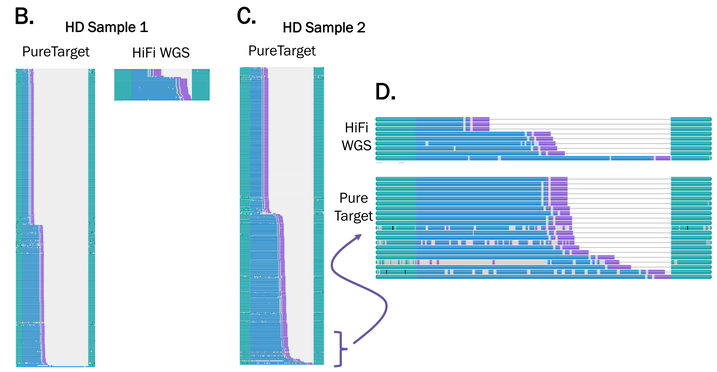



**Figure 6. Concordance of PureTarget compared to HiFi WGS.** HMW DNA obtained from five brain tissue samples was analyzed by PureTarget and HiFi WGS. (A) Summary table of HTT genotype. (B,C) HTT TR plots within two Huntington's Disease brain tissues identified by PureTarget and HiFi WGS. (D) Magnified view of panel C highlighting rare reads with > 100 CAG repeat units. Teal: flanking sequence; Blue: CAG repeat; Purple: CCG repeat.

## Conclusions

- The PureTarget workflow enables detection of tandem repeat (TRs) in 20 genes with high inter- and intra-batch consistency.
- Expected TRs are identified in all genes tested, including DMPK, ATXN1, FMR1, and HTT.
- PureTarget is cost-effective to identify HTT expansions with greater sensitivity and scalability compared to HiFi WGS.

**Acknowledgements:** We are grateful to Ping Xu, Mia Sutton, and Ye Wang (Azenta) for sequencing assistance, and Sarah Kingan (PacBio) for discussions and technical support on the PureTarget workflow.

# Cohort-based Integrated Multiomics from Single Sample Blood Draws Provides Comprehensive Holistic Phenotypic Insights

**Andrea O'Hara[1], Tom Cohen[2], Ethan Stancliffe[2], Adam Richardson[2], David Corney[1], Laure Turner[1], Chris Mozdzierz[1], Haythem Latif[1], Ginger Zhou[1]**

[1]GENEWIZ from Azenta Life Sciences, South Plainfield, NJ 07080

[2]Panome Bio, St Louis, MO 63310

## Abstract

The omics era has greatly expanded the repertoire of approaches available for researchers and clinicians to unravel the complexity underpinning human health: next generation sequencing (NGS) and mass spectrometry (MS) approaches can characterize genomes, epigenomes, transcriptomes, proteomes and metabolomes. Advanced DNA barcoding and automated microfluidics can take this to the next level, enabling multiomic characterization of single cells. Peripheral blood mononuclear cells (PBMCs) offer a minimally invasive approach that, when combined with omics tools, can provide a near holistic view of immune processes across patient cohorts. Here we detail a proof-of-principle workflow which utilizes a single blood draw to rapidly produce a diverse set of multiomics results including genomics, epigenomics, transcriptomics, proteomics and metabolomics. Blood draws using heparin tubes were collected and processed within 24 hours of the primary blood draw to ensure high viability and yield of PBMCs, along with simultaneous plasma separation and collection. These PBMC and plasma samples are then processed such that whole exome sequencing, whole genome methylation sequencing, single-cell RNA sequencing, proteomic analysis and metabolomic analysis can all be collected from patient blood draw. Integrated analysis across multiple data modalities allows for holistic views of pathways and processes that are highly impacted. Genomics, alongside proteomics and transcriptomics, offers insights into genetic and functional potential. When combined with metabolomics, it enables a deeper understanding of phenotypic variations both within and across individuals in a patient cohort. All these datatypes can be produced within days of primary sample collection using minimal sample amounts from a single blood draw, to address the biological questions at hand. With increased accessibility of omics approaches, integrative workflows such as the one described continue to gain broader adoption and drive greater insights and innovation in human health applications.

## Methods

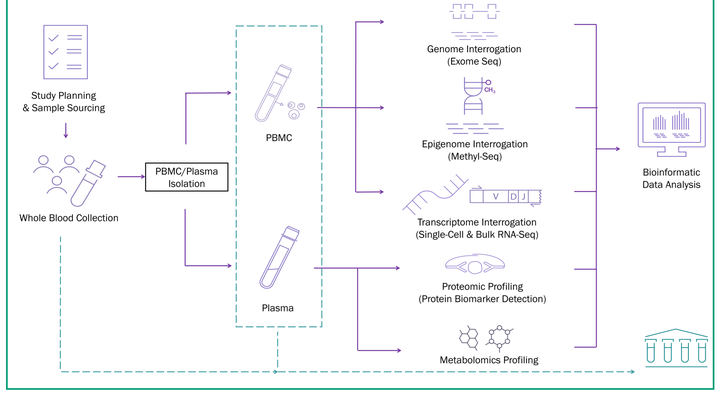



**Figure 1. Schematic of liquid biopsy multiomics sample processing workflow.** A 10mL blood draw yields ~4–5mL of plasma and ~0.5–3 x $10^6$ cells per mL, or roughly 5+ aliquots of plasma >500ml and 5+ aliquots of >1M cryopreserved cell pellets. For the workflow above, a single cell pellet is used for DNA isolation, with enough DNA (>2ug) for multiple genome/epigenome assays. One cryopreserved cell pellet was used for single-cell RNA-seq (~1M cells), and one for bulk RNA-Seq, with remainder PBMC aliquots in long-term storage. A single aliquot of plasma can be split between Olink® (~15ul) and SEER® (~250ul) proteomics and LC/MS metabolomics (~50ul) assays, done simultaneously, with extra aliquots in long-term storage.

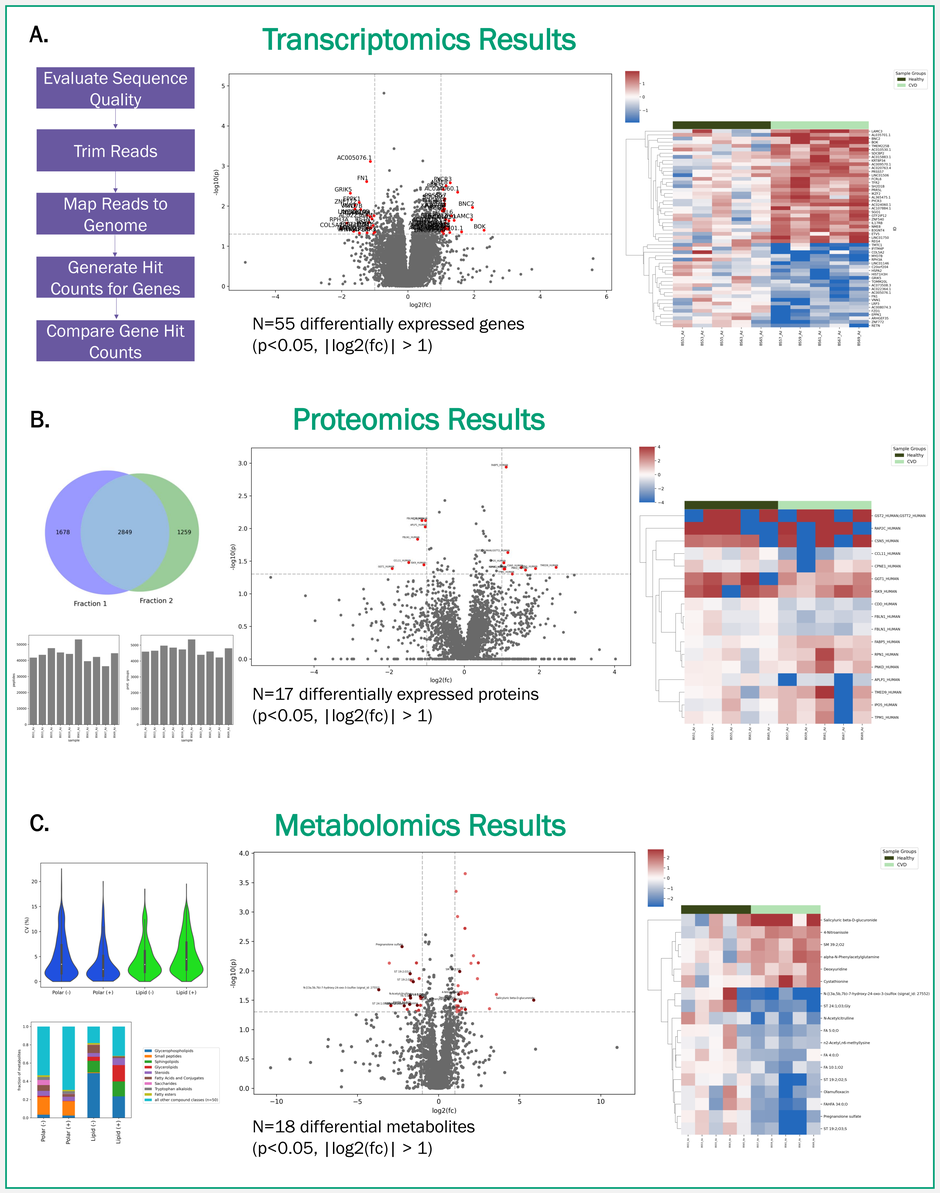



**Figure 2.** (A) RNA-Sequencing analysis identified 55 differentially expressed genes between groups. Bulk RNA-seq performed on the PBMCs collected from each sample. Results were normalized on a per-sample level and analyzed across all 10 samples. A total of 16,071 genes were profiled. Note: N=0 differentially expressed genes with (q<0.05). (B) Proteomics results identified 17 differentially express proteins between groups. Plasma samples were prepared with Seer Proteograph™ XT assay with two fractions produced. LC/MS/MS was profiled on each fraction and data was combined for peptide and protein analysis. Of the 5639 proteins detected, 3239 were found in 2 or more samples. (C) Untargeted metabolomics analysis identified 18 differentially expressed metabolites between groups. Four LC/MS (liquid chromatography mass spectrometry) assays were employed to profile the metabolome of the plasma samples. Of the 2607 metabolites profiled, 1925 metabolites were positively identified of the >50 polar metabolite and lipid classes surveyed.

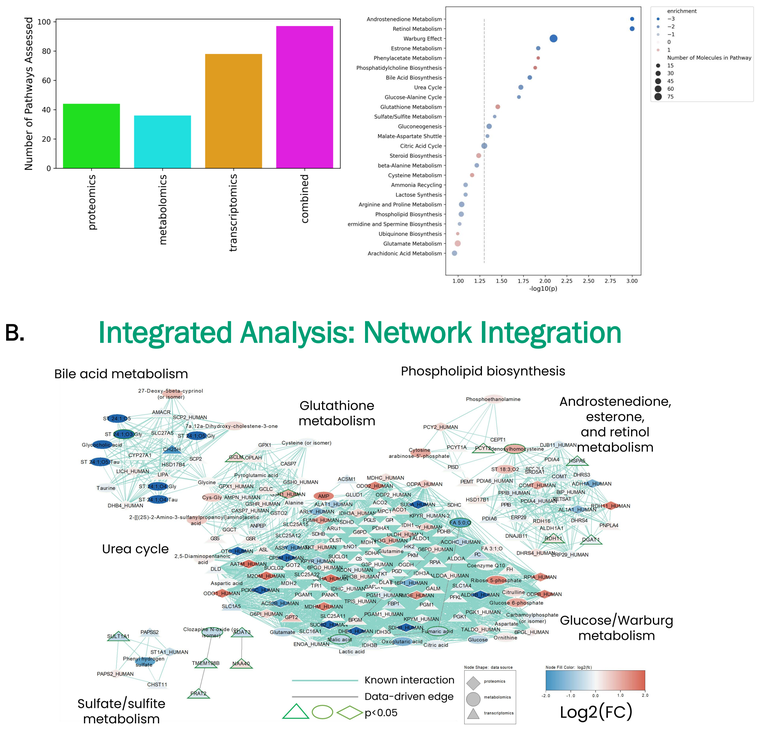



**Figure 3.** (A) Integrated pathway analysis. Thirteen metabolic pathways enriched in CVD when all three data modalities are combined. (B) Integrated network analysis. After correlating metabolite, protein, and gene expression levels across all samples to form a data-driven network with known interactions (Metabolite-protein, Metabolite-gene, Gene-protein), the network is pruned to the most tightly connected metabolites/genes/proteins. The net integration of >20k molecular measurements yields 225 interconnected metabolites (42), proteins (73), and genes (110).

## Conclusions

- Complex multiomics information can be analyzed from a single sample, including transcriptomics, proteomics and metabolomics. This approach reduces the need for repeated patient collections and lowers bio-storage requirements, making it a more patient-friendly and cost-effective solution.
- Multiomics workflows can be modified to accommodate a variety of input sample types, allowing for flexible sample processing.
- Integrated analysis allows for deeper insights into key pathways and networks for future discovery.
- Metabolomics analysis identified 18 differentially abundant metabolites; proteomics analysis identified 17 differentially expressed proteins; transcriptomics analysis identified 54 differentially expressed genes.
- Integrated analysis identified dysregulation of 13 pathways. Network analysis revealed these changes can be grouped into 6 modules, all of which have previously been correlated with CVD risk: Glucose metabolism (down), Sulfur metabolism (down), Bile acid metabolism (down), Phospholipid metabolism (up), Glutathione metabolism (up), Steroid/hormone metabolism (down).

# A Novel NGS-based Sample Tracking System for Robust Sample Chain of Custody and Quality Control in Population-Scale Genomics Studies

**Katherine Milligan, Yu Qiu, Andrea O'Hara, Carrie Ziemniak, Chirayu Goswami, Wei Wang, David Corney, Haythem Latif**

GENEWIZ from Azenta Life Sciences, South Plainfield, NJ 07080

## Abstract

When processing large cohorts of samples, sample identity and functional quality are important characteristics to ensure the derived data is accurate and attributed to the appropriate sample. This is especially critical in the era of population-scale genomics where a sample swap or poor sample quality may result in incorrect data or delays due to re-processing. While traditional methods such as qPCR can be used for functional quality control, they may not be an effective predictor of sample success in next generation sequencing (NGS) based assays due to differential susceptibility to inhibitors and interfering substances. Performing QC using the same methodology for data generation allows for true functional analysis for interfering substances, poor sample quality and/or sample degradation. We have created a scalable, cost-effective sample tracking and enhanced quality control assay that may be incorporated prior to performing whole exome sequencing (WES) or whole genome sequencing (WGS). A range of different sample types can be processed on the amplicon-based sequencing platform, targeting regions in the exome in addition to sex-based targets to allow tracking of sample identity.

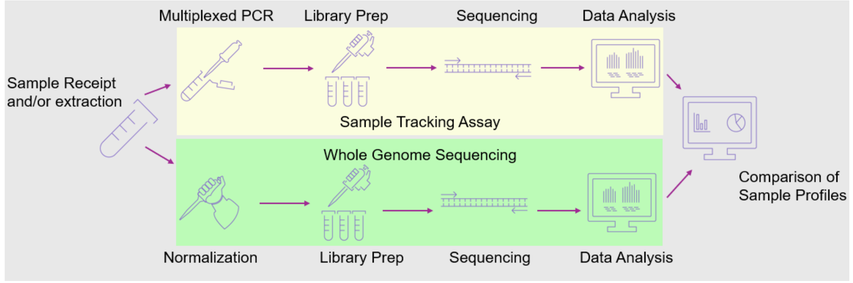
This amplicon panel was designed to target autosomal SNPs with varying minor allele frequencies in different ethnic populations, ensuring that there is no bias towards variation in European ethnic populations. The average minor allele frequency for the panel developed for the general population is 0.451, 0.459 for European background, 0.436 for African background and 0.463 for Asian background based on data from ALFA and gnomAD. Sex determination targets were selected for SNPs on the X chromosome and short unique fragments on the Y chromosome. Targets on the Y chromosome were selected to be less than 150 base pairs to reduce the effects of sample degradation and increase the likelihood of successful identification. This assay was designed and assessed to confirm consistency in identity for samples processed between WGS and WES platforms, and samples submitted for the same subject across different cohorts. In addition to identifying poor-performing samples, as low as 10% contamination levels can be detected in contrived samples.

A strong quality control assay enables NGS laboratories to identify poor sample quality or degradation prior to processing, allowing for re-extraction or new sample collection before continuing on to processing on a more complex or expensive assay, ultimately saving time and budget. This newly developed sample tracking assay ensures both sample integrity and identity, leading to an overall increase in data quality for research studies.

| ASSAY | ONBOARDING COST | PER SAMPLE COST | FUNCTIONAL ASSAY FOR NGS | ABILITY TO DETECT CONTAMINATION | ABILITY TO ID SAMPLES | FLEXIBILITY TO UPDATE PANEL |
|---|---|---|---|---|---|---|
| NGS | N/A | $ | ✓✓✓ | ✓✓✓ | ✓✓✓ | ✓✓✓ |
| Sanger Sequencing | $$ | $$$ | ✓✓ | ✓ | ✓✓✓ | ✓✓ |
| Fragment Analysis | $$ | $$$ | ✓✓ | ✓✓ | ✓✓✓ | ✓✓ |
| High Throughput qPCR | $$$ | $$ | ✓ | ✓ | ✓✓✓ | ✓✓ |
| Commercial Panel | $$$ | $$ | ✓ | ✓ | ✓✓✓ | X |

**Figure 1. There are different considerations when choosing an appropriate sample identification panel for a high throughput sequencing laboratory, in addition to its ability to identify samples.** The cost of onboarding an assay, such as equipment or infrastructure, will vary depending on what equipment is in the laboratory and the cost to run it will vary based on the kit used and the design. A functional assessment of a sample to perform on a different assay is best done by a scaled down version of the same assay. NGS provides a high level of multiplexing and quantitative data that can allow for thorough contamination detection. There is also a high degree of flexibility with an amplicon-based NGS panel to update the panel to include additional targets.

A.




B.

| PANEL CHARACTERISTICS | |
|---|---|
| Number of Autosomal SNPs Targeted | 42 |
| Number of Sex Chromosome Regions Targeted | 14 |
| Chromosomes Included | 21 |
| Average Overall SNP MAF | 0.451 |
| Average European SNP MAF | 0.459 |
| Average African SNP MAF | 0.436 |
| Average Asian SNP MAF | 0.463 |

C.

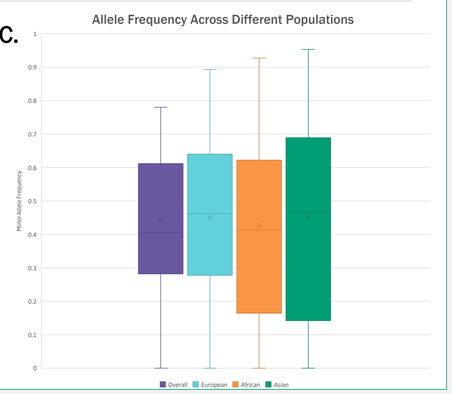
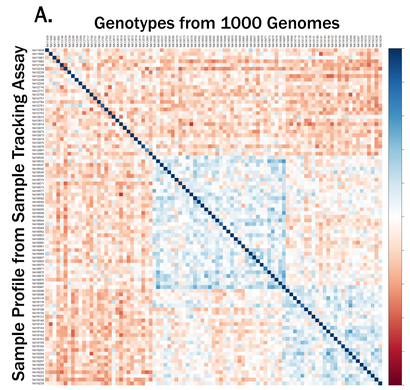



**Figure 2. Sample tracking workflow and panel details.** (A) A schematic of the NGS amplicon-based sample tracking workflow. Samples are received and DNA is extracted. The DNA is tested on the sample tracking workflow, including a multiplexed PCR that targets the amplicons, library prep, sequencing and data analysis to generate a variant file. Concurrently, samples are processed on the workflow of interest, such as WGS or WES. Finally, the profiles are compared between the assays. (B) The characteristics of the panel include SNPs from both the autosomal and sex chromosomes, targeting most of them to avoid potential linkage disequilibrium. The average MAF is also identified for the overall population as well as different ethnic groups. (C) The distribution of the MAF across different ethnic groups is displayed.

| METRIC | CORIELL SAMPLES: SAMPLE TRACKING VS 1000 GENOMES PUBLIC DATA | EXTRACTED DONOR DNA: SAMPLE TRACKING VS LP-WGS |
|---|---|---|
| Accuracy | 99.40% | 97.04% |
| Sensitivity | 99.40% | 97.01% |
| Specificity | 99.39% | 97.06% |
| Positive Predictive Value | 99.40% | 96.81% |
| Negative Predictive Value | 99.39% | 97.24% |

**Figure 3. Summary of analytical performance of the sample tracking assay.** All metrics were obtained by comparing SNV calls obtained from the assay to 1000 genome data or to Azenta's Low-Pass Whole Genome Sequencing (LP-WGS) data.

A.




B.

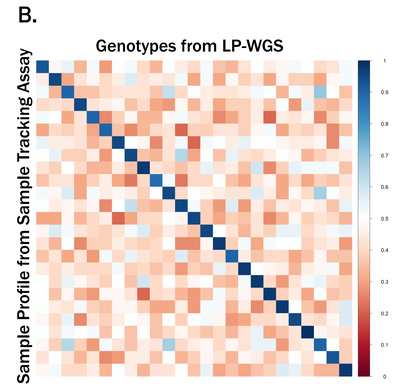



**Figure 4. Graphic display of concordance between known matching samples and all other samples in the cohort.** The color of the square is linked to the percent concordance of the SNPs between the associated data set. A: Concordance of Coriell samples to 1000 Genomes data set. B: Concordance of donor samples to LP-WGS internally derived data.

## Conclusions

- GENEWIZ from Azenta Life Sciences has created a robust, cost-effective sample tracking assay that allows functional quality control, as well as sample identity tracking.
- The sample tracking panel includes varied SNPs across the majority of chromosomes with minor allele frequencies that are varied between different ethnic groups.
- The sample tracking assay is robust with consistently high accuracy and precision.
- The sample tracking assay provides profiles that are unique to the sample of interest to allow for identification of swapped samples.

**Acknowledgements:** We are grateful to Tina Cheung and Mila Paiva for library preparation support and Alex Guzzo, Drasti Patel, and Rupal Jain for sequencing assistance.