VIRAL RESEARCH
GRANT SUPPORT DOCUMENT

The study of viruses and the host response to infection can encompass a wide range of applications, many of which benefit from the highly accurate long reads produced using HiFi sequencing on the Sequel II System.

PacBio is the only technology offering HiFi sequencing, which provides the long read lengths (>15 kb) and high accuracy (>99%) necessary to fully characterize all the mutations that can arise as a virus replicates within a host to produce a complex population of closely related variants. Similarly, HiFi reads allow you to characterize the diversity of adaptive immune response with full-length BCR repertoire sequencing. Finally, HiFi reads enable the haplotyping of 'hard to sequence' regions of the human genome involved in the immune response, including HLA genes and the IGH locus. (Table 1)

Table 1. Comparison of the precision and recall for single nucleotide variants (SNVs), indels, and structural variants obtained by sequencing technologies.

Precision | Recall (%)

	Illumina	ONT	PacBio HiFi Reads
SNVs	99.9 | 99.9	99.1 | 93.3	**99.9 | 99.9**
Indels	99.6 | 99.4	-	**96.9 | 96.0**
SVs	85.3 | 55.9	83.2 | 87.5	**96.1 | 96.0**

Illumina, HiFi Reads: Wenger, et al. (2019) Nature Biotechnology. ONT SNVs: Edge and Bansal, (2019) Nature Communications. ONT SVs: pbsv 2.1.0 run on GIAB HG002 reads, evaluated against GIAB v0.6 SV benchmark using Truvari

Virus Sequencing

- HiFi sequencing facilitates SNP phasing along entire viral genes or genomes, giving a comprehensive view of viral evolution in spillover events and species jumps, within a host during infection, or over time within a community, or across geographic regions during an epidemic.
- Single-molecule, deep sequencing allows quasispecies resolution and the detection of rare variants linked to immune evasion or drug resistance.

Experiment planning guidance

- Reliable detection of a variant present at 1% requires 6,000 reads per sample.
- One SMRT Cell 8M will yield the following estimate of HiFi reads (2.0 Chemistry, Sequel II System Software v8.0):
    - 15 kb: ~2 million HiFi reads
    - 10 kb: ~2.4 million HiFi reads
    - 2 kb: ~3 million HiFi reads
- Sample multiplexing will depend upon the number of amplicons in your virus sequencing protocol and the desired limit of detection for rare variants. With multi-amplicon protocols, it is prudent to initially allow for high amplicon-to-amplicon variance in coverage until the multiplex PCR or

amplicon pooling is optimized. An allowance of ~20% should similarly be allowed for sample-to-sample variance in coverage. A general formula for calculating multiplex level is:

- Read budget / amplicons per sample/ cvg requirement x 2 = multiplexing level
- Example 1: Using the 2.5 kb amplicon Eden protocol for SARS-CoV-2 sequencing to detect the major variant
  - 3 M reads / 14 amplicons per sample / 500x median coverage = ~400 samples per SMRT Cell 8M
- Example 2: Using a single 5 kb amplicon to detect variants in the spike protein present down to 1% relative abundance
  - 2.6 M reads / 1 amplicon per sample / 7200x coverage = ~350 samples per SMRT Cell 8M

Key resources

Website: Resolve viral populations

Review: Djiking, A., and Spiro, D. (2009) Advancing full length genome sequencing for human RNA viral pathogens. *Future Virology*. 4(1): 47–53.

Zoonoses: Tao, Y., et al. (2017) Surveillance of bat coronaviruses in Kenya identifies relatives of human coronaviruses NL63 and 229E and their recombination history. *Journal of Virology*, *91*(5), e01953–16.

Drug resistance / HCV: Betz-Stablein, B. D., et al. (2016) Single-molecule sequencing reveals complex genomic variation of hepatitis B virus during 15 years of chronic infection following liver transplantation. *Journal of Virology*, *90*, 7171–7183.

Quasispecies / HIV: Brese, R. L., et al. (2018) Ultradeep single-molecule real-time sequencing of HIV envelope reveals complete compartmentalization of highly macrophage-tropic R5 proviral variants in brain and CXCR4-using variants in immune and peripheral tissues. *Journal of Neurovirology*, *24*(4), 439–453.

B Cell Receptor Sequencing

HiFi reads allow you to sequence the full-length BCR receptor with high accuracy, enabling:

- Priming from the most conserved region of the constant domain, reducing bias against variants with primer mismatches (unknown alleles, somatic hypermutation) and PCR bias arising from highly multiplexed amplification reactions.
- Detection of all variants resulting from VDJ recombination and affinity maturation, including long CDR3 regions that thwart paired-end assembly and mutations arising outside the CDR3 region during hypermutation. Somatic hypermutation is known to be critical for developing broadly neutralizing antibodies against refractory pathogens.
- Determination of not only isotype but also IgG subclass. Since each subclass has different effector functions, this information can be important in understanding the immune response.

Experiment planning guidance

- Artisan PCR method was developed by a PacBio customer (*publication*)
- The SMARTer Human BCR IgG IgM H/K/L Profiling Kit (*commercial kit*) is available from Takara. To take full advantage of PacBio technology, Takara can supply an alternative IgG primer, which produces a longer amplicon suitable for isotype and subclass identification.
- The typically collected number of BCR sequences per sample depends on the application:
  - Vaccine response: as few as 10,000 sequences
  - Immune response: 30-60,000 sequences
  - Naïve B cell profiling: up to 1,000,000 sequences
- One SMRT Cell 8M can deliver ~3 million HiFi reads when sequencing a ~600 bp BCR amplicon.

Key Resources

BCR sequencing: Koning, M. T., et al. (2016) ARTISAN PCR: rapid identification of full-length immunoglobulin rearrangements without primer binding bias. *British Journal of Haematology*, *178*, 979–994.

Paired H/L chain sequencing*:* DeKosky, B. J., et al. (2015) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nature Medicine*, *21*(1), 86–91.

Website: Understanding coronavirus with PacBio sequencing

Host Genomics and the Response to Infection

IGH Loci

With HiFi reads you can accurately detect insertions, deletions, and rearrangements, phasing complete haplotypes to connect variation at the IGH locus to differences in health and disease.

- The IGH loci encompasses V, D, and J genes, which rearrange to form B cell receptors. The intrinsically repetitive nature of the loci makes it highly prone to insertion and deletion events on an evolutionary time scale while also making it impossible to assemble with short reads. In fact, only a handful of complete human haplotypes has been reported for this locus. Its plausible that rearrangements at this locus may have an impact on a wide range of traits linked to immune function, but because it is an NGS blind spot, to date no high-throughput sequencing studies have linked any disease to the IGH locus.
- A newly published method [Rodriguez, 2020] combines target capture with HiFi sequencing to yield fully resolved haplotypes of the IGH locus, finally enabling high throughput studies that aim to understand how allelic variation at this locus correlates with patient outcomes.
- Using this method, 8 patient samples have been successfully multiplexed per SMRT Cell 1M on the Sequel System. On the Sequel II System, it is estimated that 40 patient samples can be multiplexed per SMRT Cell 8M.

Key Resources

Review: Watson, C.T. and Breden, F. (2012). The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes & Immunity 13:363–373.

Haplotypes from WGS*:* Watson, C. T., et. al. (2013) Complete Haplotype Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic and Copy-Number Variation. Am J Hum Genet. 92(4): 530–546.

***Method*:** Rodriguez, O. L., et al. (2020) A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. bioRxiv Preprint.

Analysis tool: Ford, M., et al (2020) Genotyping and Copy Number Analysis of Immunoglobin Heavy Chain Variable Genes Using Long Reads. *iScience 23(3),* 100883*.*

HLA Loci

- HLA haplotype is known to predict risk or protection from many infectious diseases, including viruses such as dengue fever, HCV, HBV, and HIV. Ultra-high-resolution HLA typing with HiFi sequencing allows you to explore these correlations without relying on imputation.https://www.pacb.com/applications/complex-populations/viral/

- Using the GenDx NGSgo-AmpX kit, one can multiplex up to 96 samples per SMRT Cell 8M when sequencing HLA Classes I and II genes (HLA-A, B, C, D, E, F, DP, DQ, DR).
- Anthony Nolan, Histogenetics, and LabCorp are three PacBio certified service providers that offer their own PacBio-enabled HLA sequencing solutions.

Key Resources

HLA sequencing: Mayor, N. P., et. al. (2019) Recipients Receiving Better HLA-Matched Hematopoietic Cell Transplantation Grafts, Uncovered by a Novel HLA Typing Method, Have Superior Survival: A Retrospective Study. *Biol of Blood and Marrow Transplantation*. 25(3):443-450.

HLA-E: Lucas, J. A. M., et al. (2020) Single Molecule Real-Time DNA sequencing of the full *HLA-E* gene for 212 reference cell lines. HLA.