Download

Oxford NANOPORE Technologies APPLICATIONS

Nanopore Adaptive Sampling combines both targeted and low-pass whole genome sequencing in a single assay

Eoghan D. Harrington[1], Sergey Aganezov[1], Carly Tyer[1], Hayley Greenfield[2], Scott Hickey[1], Phillip James[2] & **Sissel Juul**[1]

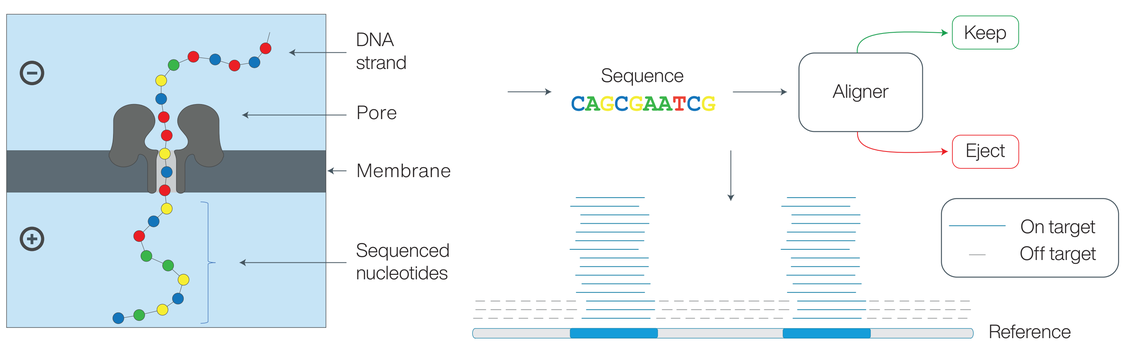
[1]Oxford Nanopore Technologies Inc, New York, NY, [2]Oxford Nanopore Technologies plc, Oxford, UK,

Contact: eoghan.harrington@nanoporetech.com

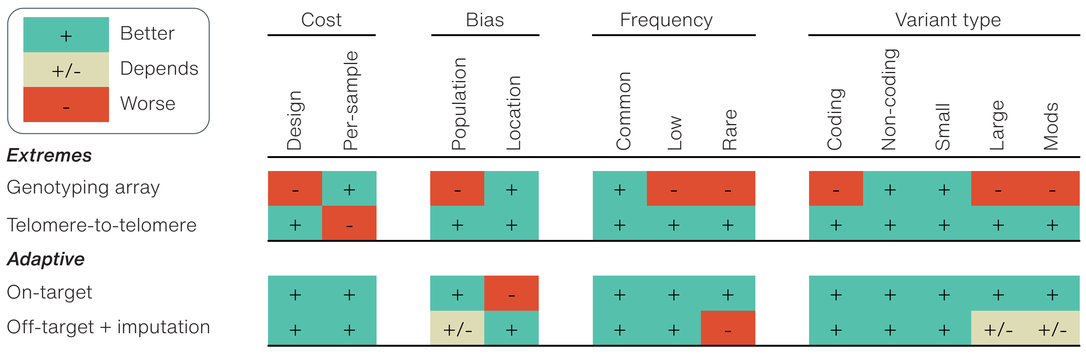
Abstract

In the 15 years since the first large-scale genome-wide association study (GWAS), there has been an explosion in the number of diseases and traits with significant associations[1]. These discoveries not only shed light on the biology underlying complex traits but also enable the use of polygenic scores (PGS) to predict phenotypes. Genotyping arrays have played a crucial role in this by providing a cost-effective method to assay common SNVs in large cohorts. More recently low-pass whole-genome sequencing (lpWGS) in combination with genotype imputation has emerged as an alternative method for GWAS and PGS calculation[2]. However unlike higher-coverage WGS and targeted sequencing, these technologies are not well-suited to the detection of rare or novel variation, limiting their ability to explore the effects of rare variants and make phenotype predictions based on the full allele-frequency spectrum. Here we demonstrate that Adaptive Sampling (AS) can be combined with genotype imputation to interrogate both rare and common variation in a single assay.

1. Background



Adaptive sampling (AS) is a software-based approach to enrich regions of interest[3]. During sequencing, strands are base-called and mapped to a reference genome in real time. Strands that align outside of the target regions within the first ~400 base pairs are ejected as they are sequenced, while strands that are on target are allowed to sequence completely. Sequencing reads are distributed uniformly across the genome, with the on-target reads having a length distribution determined by the fragment sizes of the library, while the off-target reads have a shorter and tighter length distribution determined by the decision time of the software.



A fundamental trade-off in genomics is between genome completeness and cohort size. At one end of this trade-off are genotyping arrays which tag blocks of linkage disequilibrium (LD) by capturing a subset of common variants spaced uniformly across the genome. The low per-sample cost has allowed them to be used at a vast scale, with cohort sizes reaching millions of individuals. At the other end of the scale is the telomere-to-telomere (T2T) approach which uses long reads to both span repetitive regions and resolve variants into haplotypes to create a complete representation of an individual's genome, albeit at a relatively high cost.

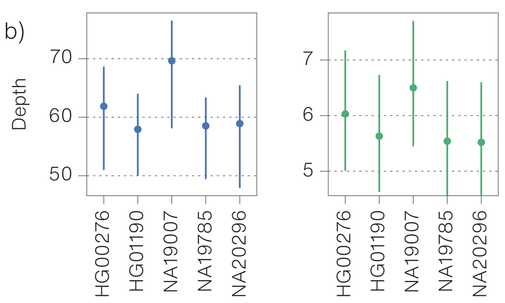
AS offers a unique solution to the trade-off between genome completeness and cohort size by combining the strengths of both approaches. With AS, long, on-target reads enable accurate variant calls that are resolved into haplotypes which can be used to discover novel, high-effect size variants. AS does not require special library-preparation steps, allowing for precise experimental design, and can even be used for epigenetic studies. Additionally, AS reads are uniformly sampled from the genome, making them compatible with methods that exploit LD-structure of common variants, such as imputation, GWAS and PGS. In this poster we focus on this latter set of applications. For examples of other applications please visit https://nanoporetech.com/resource-centre.

2. Sequencing and alignment

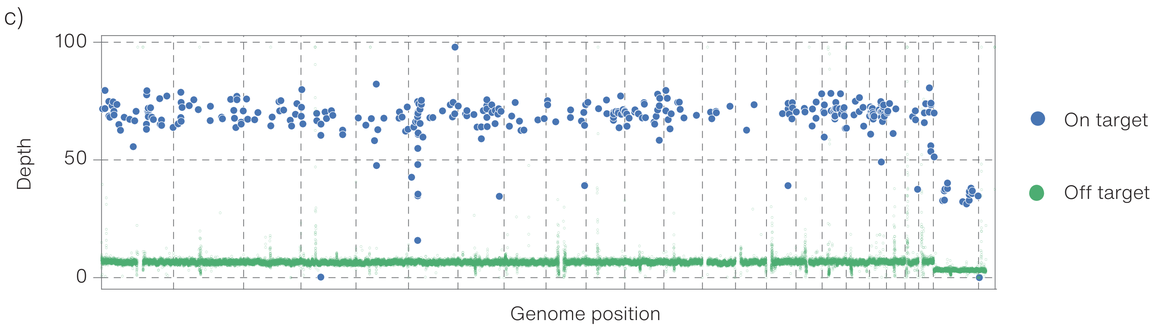
a)

		Yield (Gb)		Reads (M)		N50 (bp)	
		On-target	Off-target	On-target	Off-target	On-target	Off-target
Adaptive sampling (5 barcodes)	HG00276	3.1	20.5	0.48	28.5	7,996	648
	HG01190	2.7	18.9	0.43	26.4	7,948	648
	NA19007	3.3	21.9	0.50	30.7	8,167	646
	NA19785	2.9	18.9	0.44	26.2	8,090	648
	NA20296	2.9	18.8	0.44	26.1	8,269	647

b)

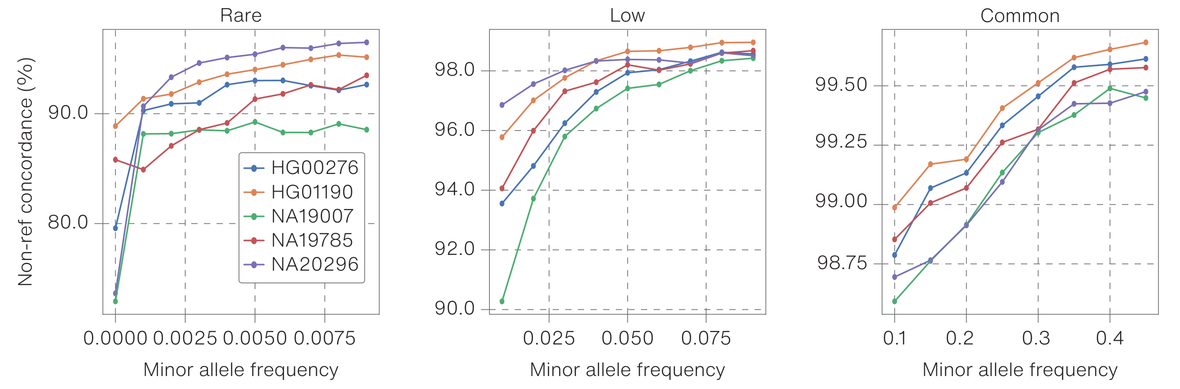


c)



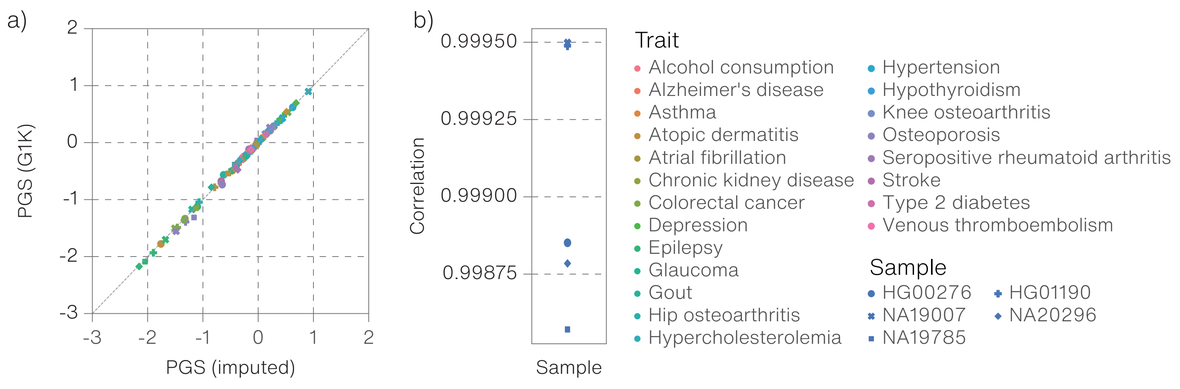
We performed multiplexed AS with five samples from the 1,000 Genomes (G1K) Project[4] on a PromethION™ 2 Solo using the R10 Native Barcoding Sequencing Kit. Reads were aligned using minimap2 and sequencing metrics were calculated for the on- and off-target regions (a). Sequencing depth was calculated for both on- and off-target regions, the latter divided into 50 kb bins (b). A plot of genome-wide coverage for NA19007 shows the uniformity of coverage (c).

3. Imputation and genotype concordance



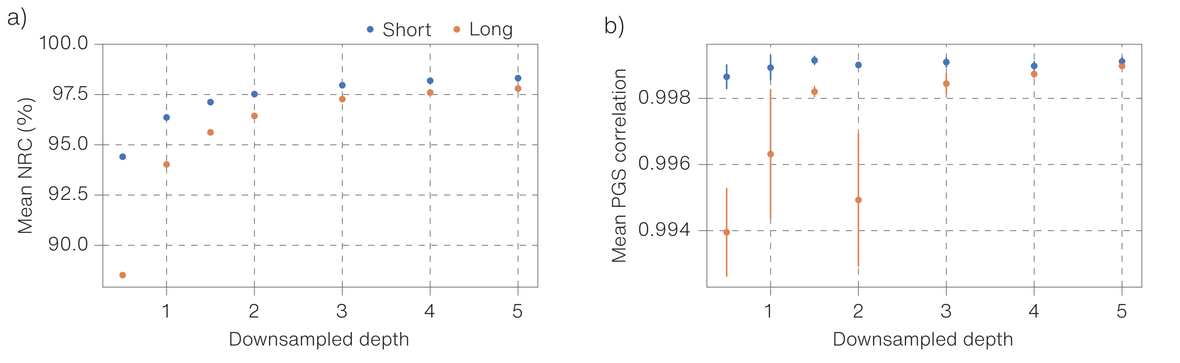
Imputation on the aligned BAM file was carried out using the QUILT tool[5] with the G1K[4] reference panel. A leave-one-out (LOO) strategy was used to measure imputation accuracy - the genotype for each of the 5 samples was removed from the reference panel before imputation. The held-out genotype could then be used as the 'truth' when measuring Non-Reference Concordance (NRC) - the percentage of concordant genotype calls at non-reference positions[2]. Genotypes were binned by minor allele frequency (MAF) before calculating NRC. Concordance is plotted separately for rare (0.0 - 0.01), low (0.01-0.1) and common (0.1-0.5) MAFs.

4. Polygenic scores



We imputed genotypes using the leave-one-out approach and computed several polygenic scores for analysed samples. Polygenic scores for 24 common diseases[6] were calculated for both the imputed and G1K genotypes for each sample using pgs_calc[7] (a). The Pearson correlation between truth and imputed polygenic scores across all 24 diseases was calculated for each sample (b).

5. Effect of read length and sequencing depth



The lpWGS approach to imputation involves sampling haplotypes from the genome using each read. Increasing the read count improves coverage of reference panel positions, resulting in better imputation accuracy. Typically, lpWGS studies target an average sequencing depth of 1-2x. However, this assumption is based on a short, constant read length. In the case of nanopore sequencing, read lengths can vary significantly, so two libraries with the same sequencing depth can have vastly different read counts. To investigate the combined impact of read length and sequencing depth on imputation accuracy, we computationally fragmented a long-read WGS library (N50 14 kb) for the G1K sample NA12878 to a constant size of 500 bp. We then downsampled both the original (long) and fragmented (short) libraries to average depths ranging from 0.5x to 5.0x, creating three replicates for each depth. Imputation was performed using the leave-one-out (LOO) strategy, and genotype and PGS concordance were measured for each replicate. The results, including means and standard errors, are presented in figures (a) and (b).

Conclusions

- The combination of Adaptive Sampling with multiplexed sequencing is a simple, flexible method to carry out targeted sequencing and genotype imputation with a low cost per-sample.
- Using a 5-plex design for gene panel yields i) on-target depth suitable for sensitive variant discovery and, ii) genome-wide depth suitable for accurate genotype imputation.
- Downsampling analysis shows that imputation accuracy is stable at lower sequencing depths, suggesting that a higher level of multiplexing is possible.
- It is important to account for read length when designing lpWGS experiments for Nanopore sequencing. For a given depth, libraries with longer read lengths have fewer reads and therefore are have higher sampling variability.

References

1. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. The American Journal of Human Genetics 110, 179–194 (2023).
2. Li, J. H., Mazur, C. A., Berisa, T. & Pickrell, J. K. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Research 31, 529–537 (2021).
3. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. Nature Methods 13, 751–754 (2016).
4. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell 185, 3426–3440.e19 (2022).
5. Davies, R. W. et al. Rapid genotype imputation from sequence with reference panels. Nature Genetics 53, 1104–1111 (2021).
6. Mars, N. et al. Systematic comparison of family history and polygenic risk across 24 common diseases. The American Journal of Human Genetics 109, 2152–2162 (2022).
7. Lambert, S. A. et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nature Genetics 53, 420–425 (2021).