White paper | 00000

Predictive genomics

Navigating the technology landscape for population-scale predictive genomics

Strategies for impact and return of results in disease risk and drug response research

Keywords

Predictive genomics research, whole-genome sequencing, low-pass sequencing, genome-wide microarrays, variant calling, genome-wide association study, monogenic and polygenic risk scores, pharmacogenomics, imputation, genotype accuracy, common variants, low-frequency variants, rare variants.

Introduction

As healthcare costs rise, outpacing inflation and wages [1], providers and payers are looking to predictive genomics research to shape future stratified medicine programs that can focus health resources for maximum impact. At the same time, the economics and capabilities of genomics technologies are steadily improving. Researchers and clinicians are therefore exploring their potential to yield useful clinical-grade data for studies of disease risk and drug response in cohorts representing >20% of a country's population. This ambition is extending globally to studies in under-served populations including in Africa [2] and Asia [3] as health disparities and the importance of genetic diversity are recognized.

This white paper scrutinizes the complementary capabilities of whole-genome sequencing (WGS), low-pass sequencing (LPS), and genome-wide microarrays. It discusses the advantages and trade-offs of each technology in the context of genomic discovery and variant calling in populations and supports researchers to make holistic technology choices.

Therapeutic area defines required variants, frequencies, and genotype accuracy

Comprehensive interpretation of genetic risk in any population requires assessment of monogenic and polygenic components [4]. For studies of a specific pathology or phenotype, the science may indicate that discrete allele frequency (MAF) ranges are important or that analysis across the MAF spectrum is required (Fig. 1). It may also suggest that analysis should extend beyond the ubiquitous single nucleotide polymorphism (SNP) to include other types of variants such as copy number variants (CNVs) and insertion/deletions (indels).

It will also be clear if a genome-wide analysis is required or if targeted analysis of specific genomic biomarkers is the best strategy; this is likely to influence the genotyping accuracy that will be required for the study.

Genome-wide association study (GWAS) analyses, e.g., to calculate polygenic risk scores (PRS), typically accept some error in return for performing genome-wide imputation to infer millions of additional common (MAF>0.05) and low-frequency (MAF 0.01 to 0.05) variants. The benefit of expanding the data set far outweighs the imputed genotype errors, assuming that a robust reference genome and imputation reference panel are available. This is achieved by prior WGS at a recommended ≥30x theoretical coverage in the study population. Imputation accuracy also depends on the size of the reference panel and how well it covers the population diversity [5] and genetic background; accuracy is known to be lower in non-European populations [6].

Conversely, if a study targets a specific set of high-value genomic biomarkers, each variant is usually individually important and therefore requires very high genotype accuracy. In the common to low-frequency MAF range, imputation would introduce too much error and so direct genotyping of the targeted markers is required to achieve very high accuracy. This is often important in clinical research applications, for example, pharmacogenomics (PGx) or blood-typing research. High-accuracy genotyping is critical in PGx because even one missed or incorrect call can fundamentally change metabolizer status [7]. Several important PGx genes are also technically challenging due to difficult sequences (e.g., UGT1a1), gene deletions, or pseudogenes (e.g., CYP2D6). Such genes are difficult to sequence with WGS, including whole-exome sequencing, making probe-based testing more favorable for pre-emptive PGx [8]. In blood-typing research, direct genotyping is essential to deliver the extremely high accuracy needed to identify genomic biomarkers that can improve safety of allogenic blood donations [9].

Direct genotyping is also necessary for rare variants because imputation becomes less effective at lower MAFs and most rare variants cannot be imputed accurately [10]. As the genetics community has moved away from the “common disease, common variant” hypothesis, rare alleles have been recognized as an important component of risk stratification (Fig. 1) that must be genotyped accurately especially when they have relatively large effect sizes [11].

Rare variants

MAF <0.01

Adiponectin; age-related macular degeneration; Alzheimer’s disease; blood typing; bone mineral density; breast cancer; cardiovascular disease (e.g., CAD); cystic fibrosis; familial hypercholesterolemia; HDL cholesterol; height; idiopathic pulmonary fibrosis; pharmacogenomics; platelet count; respiratory volume (FEV1, FVC); rheumatoid arthritis; schizophrenia; sickle cell; triglycerides; type 1 and 2 diabetes; VLDL; Von Willebrand factor antigen.

Low frequency variants

MAF 0.01-0.05

Alzheimer’s disease; ApoB; adiponectin; bone mineral density; cardiovascular disease (e.g., hypertension); cholesterol (LDL, HDL, total); chronic kidney disease; height; fasting glucose; mean cell hemoglobin; mean cell volume; obesity; pharmacogenomics; platelet count; respiratory volume (FVC); triglycerides; TSH (thyrotropin); type 1 diabetes; type 2 diabetes; FT4 (free thyroxine).

Common variants

MAF > 0.05

Cardiovascular (e.g., hypertension, ischemic stroke, hemorrhagic stroke, CAD, CHF, atherosclerosis); infectious disease (host immune response, susceptibility, severity); metabolic (e.g., type 1 diabetes, type 2 diabetes, obesity); neurodegenerative (e.g., Alzheimer’s disease, Parkinson’s disease); neuropsychiatric (e.g., schizophrenia, pharmacogenomics.

Figure 1: For comprehensive risk stratification, studies must capture rare, low, and, common frequency variants [10,12].

Genomic technology choice must enable science and execution

Within all predictive genomics research studies, there is a common set of scientific and operational execution factors to consider (Fig. 2), but the relative weighting of these is specific to each study. Currently, three primary and complementary genomic technologies can serve these needs: whole-genome sequencing (WGS), low-pass sequencing (LPS), and genome-wide genotyping microarrays.

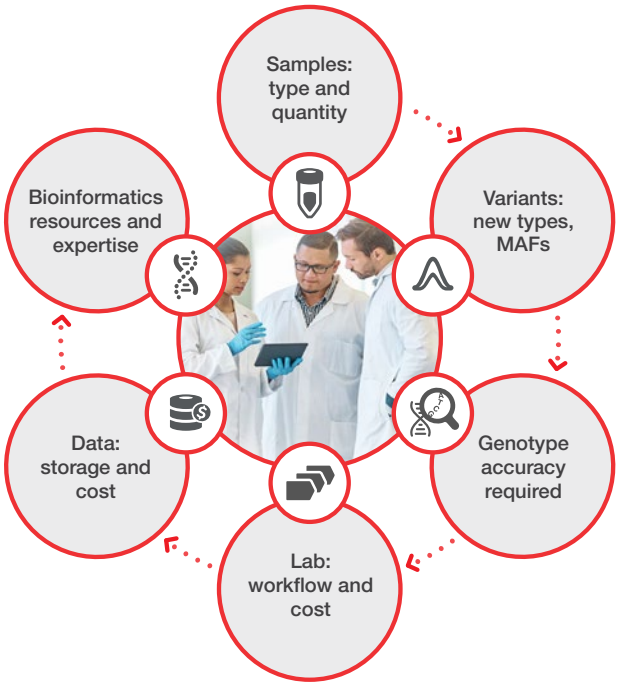

Fig 2: Study design and cost. Key factors that influence technology choice in predictive genomics studies

WGS and LPS both use the same next-generation sequencing (NGS) technology to generate millions of short-sequence reads that are assembled into contigs *in silico*. However, they balance cost vs. data quality very differently (Fig. 3).

Box 1: Overview of key whole-genome technologies for predictive genomics research

Whole-Genome Sequencing (WGS)

Prioritizes data quality at the expense of high cost per sample and low throughput. A minimum ≥20x (recommended 30x-50x) theoretical average read depth[a] enables *de novo* variant discovery and direct calling of genotypes from sequence data but scalability is limited by cost and throughput. Data and analysis are high complexity.

Low Pass Sequencing (LPS)[b]

An emerging method typically at 0.4x to 4x theoretical average read depth[a] which prioritizes low cost per sample at the expense of data quality and rare variant genotyping. Some *de novo* variant discovery[c] is possible but all genotypes must be imputed and there is no direct genotyping option for high accuracy calling of rare variants or critical biomarkers. Data and analysis are high complexity.

Microarrays

The established method for direct genotyping of 1,000s to 100,000s of variants per sample at highly scalable cost and throughput but at the expense of *de novo* variant discovery and potential for ascertainment bias of markers selected and designed onto the array. Millions of additional genotypes can be imputed from genome-wide array data. The best performance is achieved with "imputation-aware" microarrays whose content is specifically selected to power optimal imputation. Direct genotypes are highly accurate enabling calling of rare variants and critical biomarkers. Data and analysis are low complexity.

[a] Theoretical average read depth indicates the number of sequence reads at any locus if reads were evenly distributed across the genome assembly. In practice, an approximate Poisson distribution can be expected meaning some regions have more than the average number of reads while others have fewer or zero reads. The number of gaps in the assembled sequence will increase as average read depth decreases.

[b] LPS is also known as low-coverage sequencing (LCS) or low coverage whole-genome sequencing (lcWGS)

[c] Studies indicate that LCS can detect *de novo* common variants, but low-frequency and rare variant discovery is confounded by sequence errors and high false positive rates.

NOTE: Whole-exome sequencing (WES) is also used for genotyping in specific applications but is excluded here because of the importance of regulatory, intronic and intergenic variation in genetic risk stratification. Challenges with meta-analysis of WES data also limits its use in population-scale genomic studies.

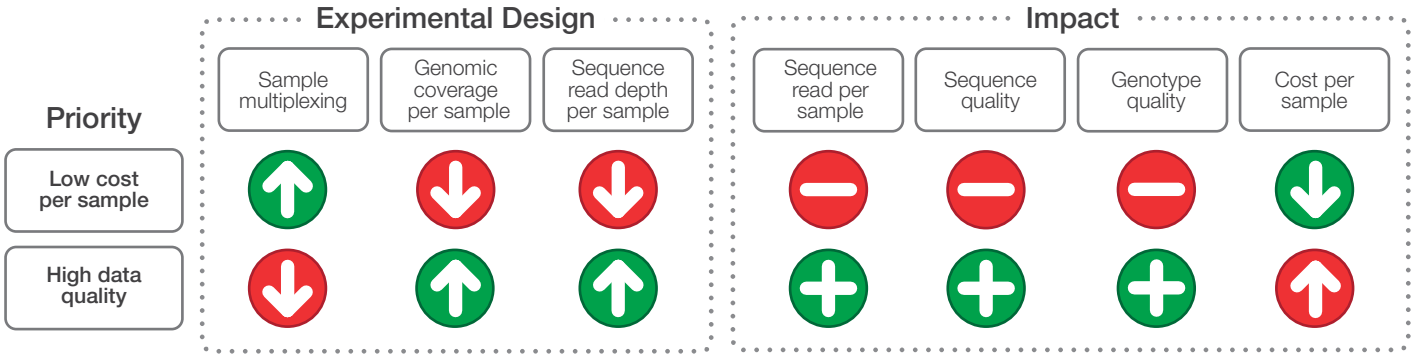

Fig. 3: NGS experiments can prioritize cost per sample or data quality, but not both. NGS systems have a finite total sequence read capacity that can be generated during one run. This total read capacity must be invested differently depending on the priority of a study.

Microarray uses hybridization probe technology to capture millions of target sequences that contain known polymorphisms. By fluorescent staining and laser scanning of the captured sample DNA, it can call genotypes with very high accuracy.

Choice of technology platform is a critical step to achieving study success. To assist you in holistically assessing your study's needs, Table 1 (page 5) examines key scientific and operational criteria that will be important to consider.

Summary

Clear study priorities are key to developing an effective technology strategy

Genome-wide microarray, WGS, and LPS are the three primary whole-genome technologies that offer complementary capabilities and execution options for predictive genomics studies at population scale.

In this white paper we have compared the capabilities of these technologies and identified how each can best be used depending on the priority aims of a specific study. The choice depends on the weighting of multiple experimental and operational factors and may require a practical trade-off of less important needs against the most important, for example:

1. Does budget cover all WGS variant discovery and downstream genotyping? Or must discovery and imputation accuracy be compromised to ensure genotyping is funded?
2. What level of genotype accuracy is needed? Is imputation error acceptable or do you need higher accuracy for rare or highly predictive variants?
3. How important is genotype ascertainment bias in your (genetically diverse) population and how will you mitigate this?
4. Is whole-genome genotyping more important to enable discovery research or do ethical rules mean that targeted data is better to avoid unwanted, reportable secondary findings (e.g., *BRCA* mutations)?
5. How do time, number of samples, statistical power, and budget influence throughput, scalability, and cost requirements?
6. What is the bioinformatics burden and what resources are available to address technical and biological variances?

Technologies can be integrated across study phases

The trade-off analysis of important and less important needs ultimately leads to the best fit technology strategy across the key phases of a predictive genomics program:

Phase 1: *de novo* variant discovery:

WGS is currently the only viable method for comprehensive discovery of genomic variation across the MAF spectrum. The 1,000 Genomes Project demonstrated that for MAF ≥0.01, sequencing many samples at moderate depth (~8x) is most effective [26]. However, detection of lower MAFs and rare variants requires 20x–30x sequencing [13]. If WGS is not possible, perhaps due to cost, complexity, or time-to-results, LPS could provide a low-cost option for combined *de novo* discovery and genotyping. However, truly effective *de novo* discovery is limited to common variants. Low-frequency and rare allele discovery is confounded by sequence errors and high false-positive rates.

Phase 2: creating a haplotype and imputation reference panel in the study population:

WGS data at ≥20x is again the gold standard to create reference panels direct from the sequence data, if sufficient samples are sequenced [6]. LPS, itself, relies on imputation to call all genotypes from the data so cannot be used to create an imputation reference.

Phase 3: genotyping:

Here, the key questions are 1) what MAFs and 2) what genotype accuracy?

Imputation methods are only effective for MAF ≥0.01 and, even in this MAF range, imputation introduces genotype error. There may be times when this error is unacceptable and very high genotype accuracy is required, for example when a highly predictive marker is being genotyped. In this case, direct genotyping will be required. At lower MAFs, imputation error becomes large and direct genotyping must be deployed to achieve accurate calls.

In targeted studies of specially selected genomic biomarkers, for example, PGx or blood typing, each marker is individually important and requires a very high genotype accuracy. This indicates for direct genotyping with microarray or deep WGS as summarized in Table 3. LPS can only offer imputed research-grade genotype accuracy for an individual marker [23,25].

In genome-wide studies, number of samples is the key driver of statistical power and imputation is used routinely to reduce costs and maximize sample number. Imputation error is minimized by use of population-specific imputation reference panels derived from WGS data. It is important to note that if LPS has been used instead of WGS for Phase 1 *de novo* variant discovery, a generic imputation panel must be used, and this will impact imputation accuracy [27].

If a genome-wide study is focused only on research-grade genotyping in variants with MAF ≥ 0.01, imputation may be all that's needed. In this case, LPS may have potential (if the library prep costs, scalability, and bioinformatic complexity can be managed), because it offers unbiased variant ascertainment

Table 1: Study selection criteria and best fit genotyping technology

Study needs or goal	Technology	Score	Rationale
Variant types and MAF ranges			
New variant discovery	Microarray	–	Not possible. Known variants only.
	WGS	+++	All MAFs (at ≥20x minimum coverage; 30x–50x recommended).
	LPS	+	Discovery power varies by MAF; @1x: common ≤95%; low-frequency ~90%; rare ≥9% with high false positive rates [13,14,15,16]. Variants that are not present in the imputation reference panel will not be genotyped [17].
Variant categories	Microarray	+++	Direct or imputed genotyping of SNPs, CNVs, and indels.
	WGS	+++	Direct genotyping of SNPs, CNVs, and indels.
	LPS	++	Imputed SNPs; direct detection of CNVs at ≥1x coverage [18]; low quality imputation of indels [16].
GWAS and PRS	Microarray	+++	Integrates genome-wide imputation of MAFs ≥0.01 at >90% genotype accuracy [19] (ascertainment bias in diverse populations can impact this [4]) with direct genotyping of rare (>80% accuracy [20]) and highly predictive (>99.6% or >99.9% accuracy, depending on assay) markers [19].
	WGS	-	Genome-wide direct genotyping is possible but cost and scalability are usually impractical [13,21].
	LPS	++	Improved imputation/GWAS/PRS for MAFs ≥0.01 @1x depth (>95% genotype accuracy [16]) but no direct genotype option to integrate rare variants (50-75% imputed accuracy [4,16]) or highly predictive markers [16].
Targeted genotyping (common to low MAFs)	Microarray	+++	Direct genotyping enables >99.6% concordance with standard assay; >99.9% with Axiom™ Plus Assay [19].
	WGS	++	Direct genotyping enables ≥99% concordance (20x theoretical coverage [13].
	LPS	+	No direct genotyping. Research-grade concordance only (~95-98% for MAF ≥1% @1x [13,16]).
Targeted genotyping (rare MAFs)	Microarray	++	Direct genotyping >80% PPV at MAF >0.00001 (0.001%) with the latest calling algorithms [20].
	WGS	+++	Direct genotyping with ≥20x minimum coverage (30x-50x recommended).
	LPS	-	Not recommended due to need for and performance of imputation (50-75% concordance @1x) [4,16].
Operational factors			
Assay optimization	Microarray	+++	Robust protocols (e.g., blood, saliva buccal sample). End-to-end design support available for Axiom™ Arrays [22].
	WGS	++	Optimization required e.g., library prep.
	LPS	+	Emerging technique—significant optimization e.g., library prep, sequence coverage depth.
Sample throughput scalability	Microarray	+++	Scalable from 10s to 100,000s of samples.
	WGS	+	Low-throughput only.
	LPS	+	High throughput sample multiplexing (≤1,500 per run) required to achieve low run costs.
Data storage and analysis	Microarray	+++	Low data storage (<0.5 Gb per sample); robust, standardized data analysis tools.
	WGS	+	Very high data storage (~60-90 Gb per sample) and computing resource; analysis requires complex bioinformatics.
	LPS	+	High data storage (>3x-6x vs. array); 50x computer processing time vs array; very complex bioinformatics [16].
Cost per sample	Microarray	+++	Per genome-wide sample: ~$32 USD consumables; ~$50 USD including lab service fees. Data storage and computing costs: low.
	WGS	–	Per sample @30x: ~$1,000 USD including lab service fees [23]. Data storage and computing costs: highest.
	LPS	++	Per sample @1x: ~$50 USD consumables; ~$90 USD including service fees [14,17,23,24,25]. Data storage and computing costs: high.

which improves imputation accuracy, GWAS signal detection, and PRS power if sufficient sequencing depth is used.

However, the current state-of-the-art is to integrate genome-wide imputed common and low-frequency alleles with direct genotyping of rare variants and high-value (e.g., PGx) genomic biomarkers [28,29,30]. This hybrid "genome-wide plus targeted" approach offers the opportunity to develop PRS that combine polygenic and monogenic risk components or to study both disease genetics and pharmacogenetics in one cohort. In this case, microarray remains the established technology of choice because it enables genome-wide imputation plus direct genotyping to be scaled from small to very large cohorts with well-established and low complexity assays and data analysis.

References

1. PwC Health Research Institute (2021) Medical cost trend: Behind the numbers 2022. *PwC*. ePub:1-45.
2. Harlemon M, Ajayi O, Kachambwa P et al. (2020) A custom genotyping array reveals population-level heterogeneity for the genetic risks of prostate cancer and other cancers in Africa. *Cancer Res* 80(13):2956-2966.
3. Taiwan Precision Medicine Initiative tpmi.ibms.sinica.edu.tw/www/en/precision-medicine/
4. Homburger JR, Neben CL, Mishne G et al. (2019) Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome Med* 11(74).
5. Graham SE, Clarke SL, Wu KH H et al. (2021) The power of genetic diversity in genome-wide association studies of lipids. *Nature* 600, 675–679.
6. McCarthy S, Das S, Kretzschmar W et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 48(10):1279–1283.
7. Hoshitsuki K, Crews KR, Yang W et al. (2019) Challenges in clinical implementation of CYP2D6 genotyping: choice of variants to test affects phenotype determination. *Genet Med* 22:232–233.
8. Crisamore K, Adams S, Empey P (2019) Comparison of genetic testing approaches for implementation of preemptive pharmacogenomics. Poster presented at 69th Annual Meeting of ASHG; Oct 15-19; Houston, USA.
9. Gleadall NS, Veldhuisen B, Gollub J et al. (2020) Development and validation of a universal blood donor genotyping platform: a multinational prospective study. *Blood Adv* 4(15)3495-3506.
10. Momozawa Y, Mizukami K (2021) Unique roles of rare variants in the genetics of complex diseases in humans. *J Hum Genet* 66:11-23.
11. Forgetta V, Manousaki D, Istomine R et al. (2020) Rare genetic variants of large effect influence risk of type 1 diabetes. *Diabetes* (69):784-795.
12. Bomba L, Walter K, Soranzo, N (2017) The impact of rare and low-frequency genetic variants in common disease. *Genome Biol* 18(77).
13. Li Y, Sidore C, Kang HM et al. (2011) Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res* 21:940-951.
14. Martin AR, Atkinson EG, Chapman SB et al. (2021) Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am J Hum Genet* 108:656-668.
15. Pasaniuc B, Rohland N, McLaren PJ et al. (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet* 44:631–635.
16. Gilly A, Southam L, Suveges D et al. (2019) Very low-depth whole-genome sequencing in complex trait association studies. *Bioinformatics* 35(15):2555-2561.
17. Li JH, Mazur CA, Berisa T et al. (2021) Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res* 31:529-537.
18. Zhou B, Ho SS, Zhang X et al. (2018) Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J Med Genet* 55:735-743.
19. Thermo Fisher Scientific (2019) Axiom™ Precision Medicine Diversity Research Array. Data Sheet COL33285 1019. www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FGSD%2FReference-Materials%2Faxiom-microarray-pmda-datasheet.pdf
20. Mizrahi Man O, Woehrmann MH, Webster TA et al. (2021) Novel genotyping algorithms for rare variants significantly improve the accuracy of Applied Biosystems™ Axiom™ array genotyping calls. *bioRxiv* 2021.09.13.459984.
21. Tam V, Patel N, Turcotte M et al. (2019) Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 20:467–484.
22. Axiom™ myDesign Custom Genotyping Arrays. Thermo Fisher Scientific. www.thermofisher.com/uk/en/home/life-science/microarray-analysis/agrigenomics-solutions-microarrays-gbs/axiom-genotyping-solution-agrigenomics/axiom-mydesign-genotyping-arrays.html
23. Wasik K, Berisa T, Pickrell JK et al. (2021) Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genom* 22(197).
24. Dong Z, Zhang J, Hu P et al. (2016) Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet Med* 18:940–948.
25. Lou RN, Jacobs A, Wilder AP et al. (2021) A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol Ecol* 30(23):5966-5993.
26. The 1000 Genome Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74.
27. Schurz H, Müller SJ, van Helden PD et al. (2019) Evaluating the accuracy of imputation methods in a five-way admixed population. *Front Genet* 10:34.
28. Bycroft C, Freeman C, Petkova D et al. (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562:203-209.
29. Mars N, Koskela JT, Ripatti P et al. (2020) Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* 26:549-557.
30. Hunter-Zinck H, Shi Y, Li M et al. (2020) Genotyping array design and data quality control in the Million Veteran Program. *Am J Hum Genet* 106(4):535-548.

Learn more at **thermofisher.com/predictive-genomics**

applied biosystems