Pipeline for improved genotype imputation using blended genome-exome sequencing and a diverse reference panel for large-scale population studies

Michael Gatzen, Christopher Kachulis, Candace Patterson, Matthew Coole, Edyta Malolepsza, Megan Shand, James C Meldrim, Matthew DeFelice, Niall J Lennon

Broad Institute of MIT and Harvard

BACKGROUND

- Whole genome sequencing (**WGS**) remains the **gold standard** for genetic studies, but even though it has become more affordable the relatively **high cost remains a barrier** to the feasibility of many population studies
- Whole exome sequencing (**WES**) is a **more affordable option**, however, the shortcomings of being **blind to significant portions of the genome** may be prohibitive for certain research questions
- **Imputation from genotyping arrays** provides a bridge between affordability and information about large regions of the genome, however the limitation of only being able to capture **predefined alleles** results in **reduced applicability to diverse populations and disease characteristics**
- **Blended Genome-Exome** combines **high-coverage exome** (40x) and **low-coverage whole genome** (1-3x) into one sequencing product[1]

HIGHLIGHTS

- **Imputation using blended genome-exome (BGE) achieves superior results to existing methods using GDA genotyping arrays**
- Cloud-native pipeline provides cost-effective imputation for large-scale cohorts
- **Accuracy of polygenic risk scores calculated from BGE data are on-par with or superior to existing technologies**, enabling both research and clinical applications

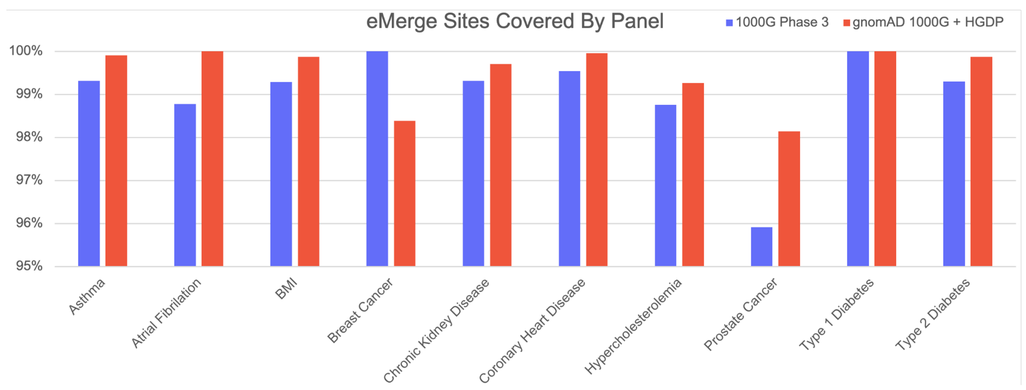
METHODS

Imputation Pipeline

- GLIMPSE2[2] is optimized for low-coverage whole genome imputation, scaling sub-linearly with number of samples and markers in reference panel
- Cost-optimized cloud-native pipeline for high throughput of samples

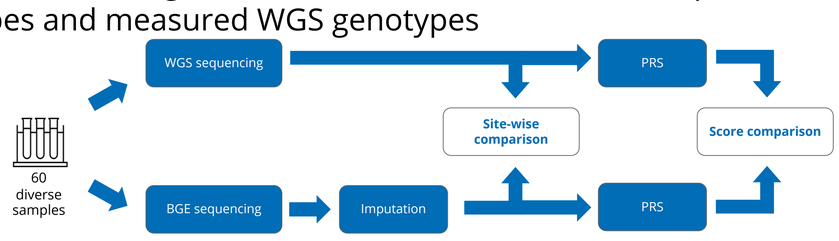
Reference Panel

- Imputation using gnomAD 1000 Genomes + Human Genome Diversity Project (HGDP) panel[3]
- 2,500 samples (1000G) + 780 samples (HGDP) from > 60 distinct populations from Africa, Europe, the Middle East, South and Central and South Asia, East Asia, Oceania, and the Americas, jointly phased with entirety of gnomAD
- 91% more sites than commonly-used 1000G Phase 3 panel after removing singletons
- Increase in covered sites for 10 eMerge PRS models[4] from 99.3% to 99.8%



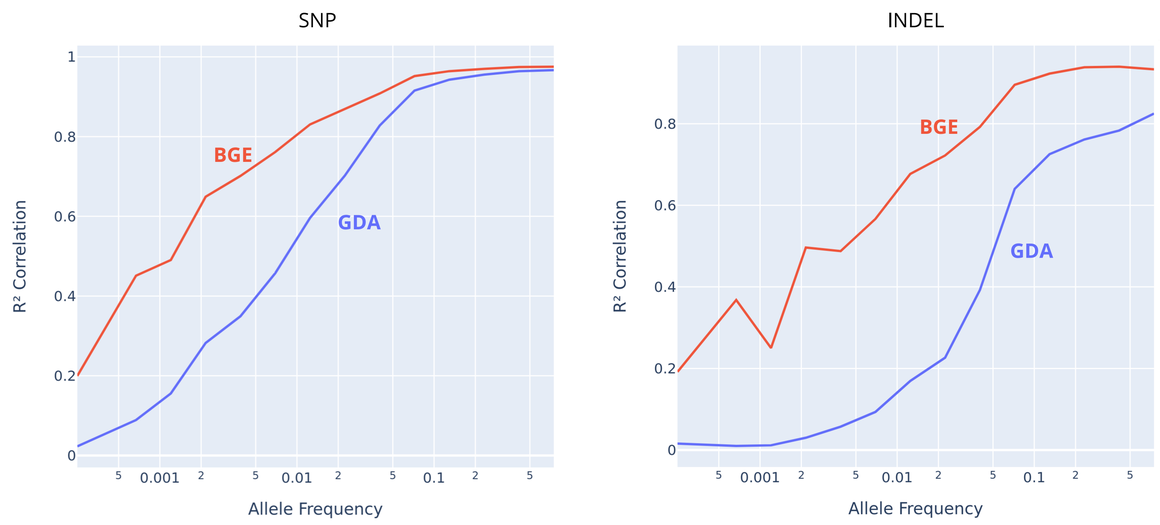
Evaluation

- 60 samples of diverse ancestries with matched WGS, BGE, and GDA genotyping data
- Site-wise comparison of imputed (BGE/GDA) genotypes to measured (WGS) genotypes
- Calculation of eMerge Prostate Cancer PRS scores[5] on imputed (BGE/GDA) genotypes and measured WGS genotypes



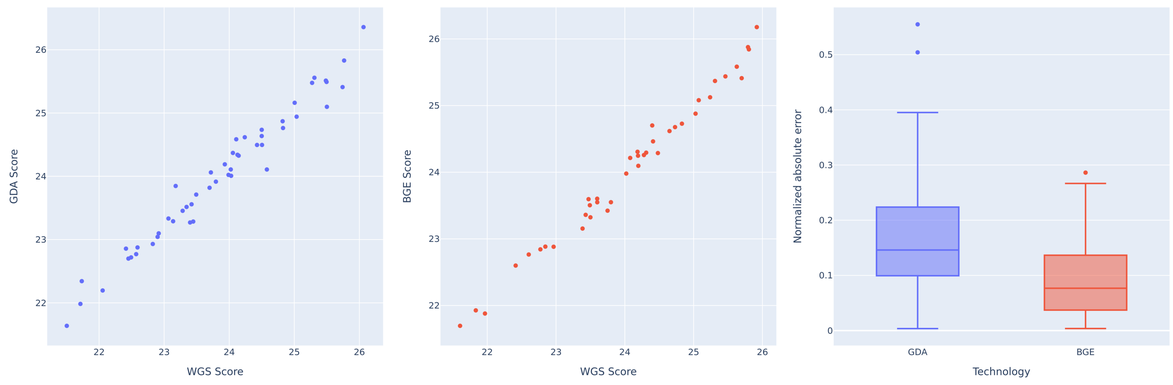
RESULTS

Correlation between imputed sites and WGS sequencing



Site-wise correlation between WGS sequencing data and imputed genotypes from BGE and GDA data (chr20).

Effect on Polygenic Risk Scores



PRS scores calculated from WGS data compared to PRS scores calculated from imputed BGE and GDA data, as well as the normalized absolute error between GDA/WGS scores and BGE/WGS scores.

Validation of predictive power of different technologies

We calculated PRS scores (eMerge Prostate Cancer[5]) based on All of Us[6] WGS genotype data and simulated BGE and GDA imputed data by adding the noise determined above, and validated the predictive power using corresponding phenotype data. PRS scores calculated from BGE data have better predictive power than PRS scores calculated from GDA data, and is comparable in accuracy to WGS.

CONCLUSIONS

- The combination of Blended Genome-Exome data as an input for imputation with an improved, more diverse reference panel significantly improves the accuracy of results as compared with current approaches
- Combined with high-confidence over the exome calls for rare variants, Blended Genome-Exome provides a cost-effective and accurate solution for population genetics studies without the need for multiple analysis modalities
- The scalable cloud-native imputation pipeline enables a high throughput of samples for both research and clinical applications

References

1. Broad Clinical Labs. Blended Genome-Exome (BGE) Product Datasheet. https://broadclinicallabs.org (2023)
2. Rubinacci, S., Hofmeister, R.J., Sousa da Mota, B. et al. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. Nat Genet 55, 1088–1090 (2023)
3. Siwei Chen, Laurent C. Francioli, Julia K. et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv 2022.03.20.485034 (2022)
4. Lennon, N.J., Kottyan, L.C., Kachulis, C. et al. Selection, optimization, and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse populations. medRxiv 2023.05.25.23290535 (2023)
5. Conti, D.V., Darst, B.F., Moss, L.C. et al. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. Nat Genet 53, 65–75 (2021)
6. The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants.