ACMG
American College of Medical Genetics and Genomics

Genetics in Medicine
An Official Journal of the ACMG

Genetics in Medicine OPEN

ACMG/GIM/GIMO
April 24, 2025
The Big Data Problem in Genomic Medicine: Acquisition, Processing, Interpretation, Storage, and Retrieval.

The realization of genomics-first medicine depends on effectively harnessing population-scale genomic data within the clinical setting and integrating it with rich clinical and phenomic information.

This talk explored the critical infrastructure required to support such ambitious initiatives, examining the challenges of building scalable systems for secure storage, efficient annotation, and rapid querying of these massive datasets.

Crucially, the presenter discussed the role of integrating genomic data with detailed phenomic information to unlock clinical utility, enabling real-time insights, as well as the challenges presented by genomics-first implementation, including potentially under-appreciated variability in the penetrance and phenotypic spectrum of genetic conditions.

Also examined was the innovative application of large language models (LLMs) to extract granular phenotypic insights from large-scale observational data, enabling a deeper understanding of genotype-phenotype relationships and paving the way for personalized, data-driven health care.

Supported by Illumina, the webinar featured Kyle Retterer, MS, Chief Data Science Officer, Geisinger. It was moderated by Marc S. Williams, MD, FACMG
Professor and Director Emeritus, Department of Genomic Health, Geisinger.

Below are questions and responses from the webinar's Q&A session.

1. **Is it possible to divide the genome into bins and identify the parts that are more conserved (less likely to be variable across individuals/populations)? Then, try to only save the variations of each genome, and for the rest of their genome, refer to those filtered regions/bins. This way, each genome's data would take up much less storage.**

   As described in the talk, only the differences (i.e., variants) would typically need to be readily available for the purposes of Genomic Medicine. This strategy of only storing variants is already widely used in flat file formats (like VCF or gVCF) and typically propagated downstream. Some raw sequencing data compression techniques also take advantage of this idea to reduce their storage footprint.

2. **Are the tools shared here, like PheNominal, PhenoTagger, etc., available for use as open source globally?**

   Some but not all the tools are open source. Of those you mention, PheNominal is not open source, but PhenoTagger is available on GitHub with an MIT license (https://github.com/ncbi-nlp/PhenoTagger). OntoGPT, PhenoGPT, and txt2hpo are also open source on GitHub.

3. **I would think multiple companies will offer WGS. Finding markers will be faster if the places offering sequencing would share their data. I see that as a challenge.**

   Laboratories may also have commercial incentives and regulatory restrictions that limit their willingness or ability to share such data.

4. **What is your opinion on the likely readiness of different use cases (for research or clinical purposes)?**

   If the ultimate goal is to have a "genome on file" for a patient and to be able to re-use that data at any point in time, then I would expect the earliest use cases to be those that are both the simplest and offer the highest value. Pharmacogenomics would likely be at the top of that list since we are already seeing that implemented in fragments.