Low-cost enrichment of datasets for genetic analysis: imputation and electronic health record linkage

The human genome consists of 3 billion base-pairs built from the nucleotides adenine (A), thymine (T), cytosine (C) or guanine (G). DNA sequencing is the process of reading and recording the nucleotide at each base in the genome. The DNA sequences of any two people are very similar, but each human has, on average, 10 million genetic variants in their genome that create a DNA pattern unique to that person. These genetic differences (along with environmental influences) are responsible for the variation observed in traits such as height, blood pressure or blood glucose levels.

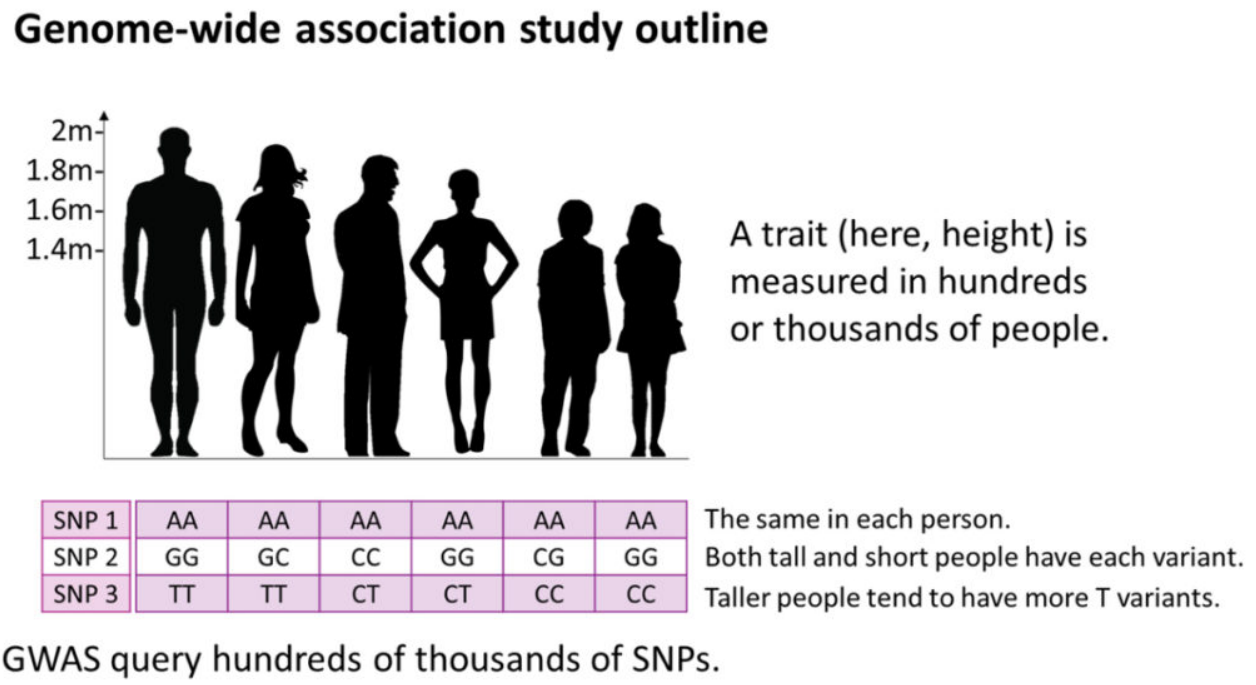

Fig. 1.

Genome-wide association studies (GWAS) aim to identify genetic variants that influence a trait of interest, helping us to understand the genetic ‘wiring diagram’ (Fig. 1). This improves our ability to predict a trait based on genetic information alone, which can lead to earlier disease diagnosis and the discovery of new drug targets, paving the way for the development of novel treatments.

GWAS generally study single-nucleotide polymorphisms (SNPs), which are a type of genetic variation where at a single base, different people have different nucleotides. Due to the high cost of

sequencing a whole genome, these SNPs are usually ascertained by genotyping, which queries a small subset of the genome known to contain SNPs. Genotyping reveals only a fraction of all SNPs, however, and misses many rare variants that might be important drivers of trait variation. Using a statistical approach called imputation, we can fill in some of these blanks, drawing on the fact that variants that are near each other tend to be inherited together (Fig. 2).

In this article, we present the Generation Scotland: Scottish Family Health Study (GS:SFHS), which is a family-based population cohort with DNA, biological samples, socio-demographic, psychological and clinical data from approximately 24,000 adult volunteers across Scotland. Genotyping was performed on just over 20,000 individuals, measuring 605,000 SNPs. Imputation was performed with the help of the Haplotype Reference Consortium dataset, yielding 24.1 million high-confidence genetic variants for analysis, a 40-fold increase in the number of queriable variants.

We performed GWAS on a range of quantitative traits measured in all participants during recruitment. We replicate known associations, but also reveal novel findings, predominantly with imputed, rare variants with minor allele frequencies in the 0.08-1% range.

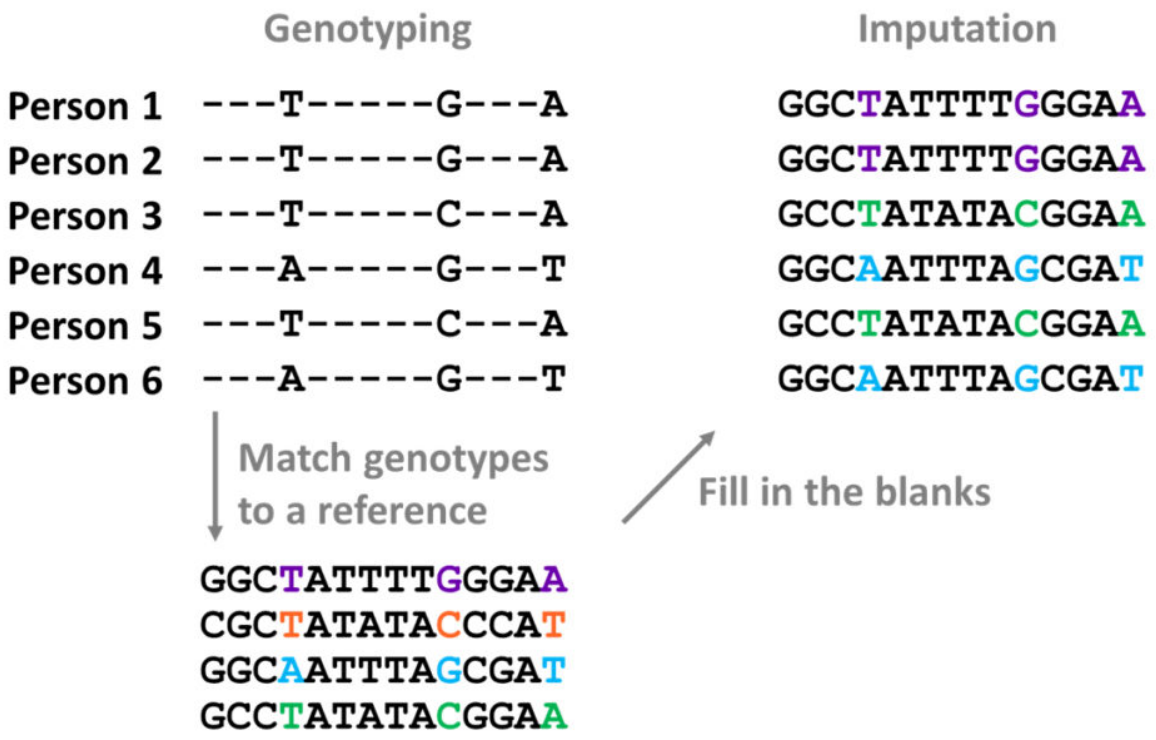

Fig. 2.

We also present a small (2,077-person) proof of principle GWAS of serum urate levels that were measured by the Scottish National Health Service for clinical purposes, by linking to the electronic health records (EHRs) of GS:SFHS participants. In this GWAS, we replicate an association at the *SLC2A9* gene locus, which encodes a well-established urate transporter. This is an encouraging result, and efforts are ongoing to obtain and curate phenotype data from electronic health records for a range of other clinically relevant traits, for all GS:SFHS participants.

In conclusion, this study demonstrates the value of leveraging statistical methods and EHRs to boost the power of genetic studies with no additional costly laboratory experiments. It also reveals novel genetic associations that may be useful for predicting individuals' risks for developing certain conditions such as hypertension or diabetes.

Réka Nagy
MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, United Kingdom

Publication

Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants.
Nagy R, Boutin TS, Marten J, Huffman JE, Kerr SM, Campbell A, Evenden L, Gibson J, Amador C, Howard DM, Navarro P, Morris A, Deary IJ, Hocking LJ, Padmanabhan S, Smith BH, Joshi P, Wilson JF, Hastie ND, Wright AF, McIntosh AM, Porteous DJ, Haley CS, Vitart V, Hayward C
Genome Med. 2017 Mar 7