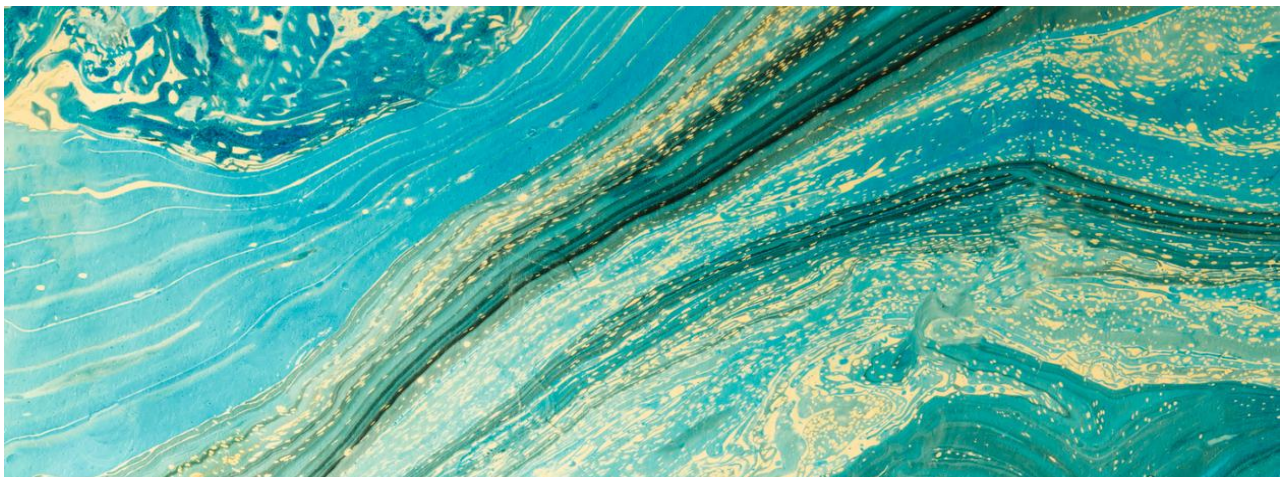
THE IMPUTE Module

Fast and scalable genome imputation of Low-Coverage WGS and Microarray data in minutes

Allelica's PRS pipeline is built around a set of interconnected modules. Each module is designed to give the user enough control to run it as needed whilst leaving all the hard computational work behind the scenes. Which modules a user needs will depend on their specific needs, but the complete PRS pipeline has been developed to be able to run a full workflow on a starting dataset containing just two things: genomic data from your sample population and a set of summary statistics from a genome-wide association study (GWAS) for a disease or trait of interest. (Users interested in computing a PRS for an individual using already available PRS can use the PREDICT module with input genomic data only.)



In this piece, we'll run over what the IMPUTE module does. The main aim of this module is to align the variants in the genotype data with those of the summary statistics from the GWAS. These rarely match completely, so this ensures that as much of the information about how these variants influence disease risk or trait value can be used.

Our IMPUTE module is based on the very latest imputation methodology which we've optimised further to run in a powerful and scalable cloud architecture. The result is the fastest solution available to the market with state of the art imputation accuracy. IMPUTE can be used as a stand alone module but is also currently integrated within the DISCOVER and PREDICT modules.

Why do we need to impute genetic data?

Because genomic data input into a PRS pipeline will often not contain all the genetic variants listed in GWAS summary statistics, the first step of many analyses will be to impute missing genetic variants. But why do we need to do this?

To understand why this is necessary, we need to first look at how genetic data is generated at scale in the first place.

DNA is extracted from a sample and is then 'sequenced'. Sequencing is the process that turns the physical molecule of DNA into a digital representation of its sequence of component nucleotide bases. There are just four of these bases which are represented by four letters: A,T,C and G. So sequencing gives us the raw genetic data that can be used to analyse its relationship with disease.

Scientists typically use one of two main methods to generate a set of genetic data for the samples in their analysis.

The first method is to perform Whole Genome Sequencing (WGS) which aims to sequence the whole of an individual's genome, usually to a very high coverage. High coverage here means that each base of the genome is sequenced a large number of times. The gold standard is 30X coverage, which means that -- on average -- each base in the genome is sequenced 30 times. At this level of coverage, you can be pretty confident that the quality of the data produced is high. Whilst WGS generates sequence data of the highest quality, it also generates a number of huge files which need to be processed and is currently the most expensive way to generate a human genome's worth of genetic data.

A derivative method, called low-coverage WGS, is a cheaper sequencing technique than full WGS. As the name implies, this approach is identical to WGS but samples are sequenced at a very low coverage, typically between 0.2X and 2X coverage. This means that whilst you get some data from across the genome, the depth is patchy. Nevertheless, this technique has the advantage of being scalable to hundreds or thousands of individuals, and can potentially identify rare genetic variants.

A different approach to sequencing DNA leverages the fact that people have very similar DNA. On average two individuals' DNA will only differ by about 1 base in every thousand, so it's about 99.9% similar, on average. If we were to WGS many individuals, you'd end up sequencing a lot of DNA that is essentially the same. So whilst you'd be getting high quality, precise data for a lot of individuals, this level of detail is not always required. Alternative approaches to WGS sequence a subset of specific places in the genome and then use what we know about the human genetics to guess at the variants across the rest of their genome. Using a so-called genotyping array (or chip), researchers can assay many specific places of the human genome at a predefined set of places in the genome that are known to differ among individuals. These are cheap and easy to run (with the correct equipment) and can scale to sequence hundreds or thousands of individuals in a short amount of time.

Typically, genotyping chips look at anywhere between 50,000 and 5 million different genetic places in the genome, rather than the full 3 billion bases of the human genome and can be used to impute most of the common variation between individuals.

Imputation maximises the power of genetic data

Despite the availability of WGS technology, the vast majority of large genetics projects generate data on individuals using genotyping chips or low-coverage WGS. However, even if these approaches only generate data at a few hundred to a few million genetic variants per individual, we can use imputation to fill in the gaps between these variants.

People inherit their DNA from their parents as 23 pairs of chromosomes. The order of the bases in chromosomes is the same in everyone, so when we sequence DNA we can compare it to a universal human genome map and understand where in the human genome the sequence belongs.

The key to imputation is to compare the subset of an individual's genetic variants assayed on a genotyping chip to a set of WGS genomes from a reference dataset. An individual's genotypes will not match the whole of one of the genomes in a dataset, but will match parts, or chunks, of the genome in a number of different reference genomes. The more closely related an individual is to the reference genome, the more likely it is that chunks will match, so a key consideration of imputation is to use a reference dataset containing individuals with close genetic ancestry to the individuals you're imputing.

Imputation leverages the fact that if two or more individuals share the same letter at parts of the genome that are physically nearby, then it's likely that the letters in between will be the same as well. By way of example, if an individual has the same base at positions 100 and 500 on chromosome 1 as an individual in the reference dataset, then they'll likely have the same bases at positions 101 through 499 too. Our individual's full genome can now be reconstructed by matching chunks of it to the closest match in the reference dataset.

There is a growing collection of WGS data from diverse human populations that can be used in imputation. This means that if we have a subset of places in the genome where we know an individual's DNA letters, then we can predict what the intervening letters are. So, if you have data on a set of individuals that have been genotyped using Illumina's Global Screening Array, which allows 610,000 genetic variants to be genotyped, then imputation can be used to turn this data into ~60 million variants for an individual.

Allelica's IMPUTE module

In addition to the diversity of the reference panel, the quality of imputation is influenced by several factors. These include the frequency of the variants that are present in the populations, and the method that is used to impute, amongst other things.

At Allelica, we use as standard a large and diverse set of reference genomes from the 1000 Genomes Project. We use two imputation methodologies depending on the input data: GLIMPSE is used to impute in low coverage WGS and BEAGLE for genotype array input data. These are implemented in an elastic cloud architecture allowing the imputation of hundreds of thousands of samples in parallel.

Please get in touch for further details about our IMPUTE module or to request a demo of our whole PRS pipeline.