Sarah Hunt [1], Matthias Wielscher[2], Loukas Moutsianas[2], Irina Armean[1], Patrick Short[3], Paul Flicek[1], Matthew Hurles[3], Augusto Rendon[2], Anna Need[2], Ewan Birney[1],the Genomics England Research Consortium

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK
[2] Genomics England, Dawson Hall, Queen Mary University of London, Charterhouse Square, London, UK
[3] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

Creation of a dense, public allele frequency resource

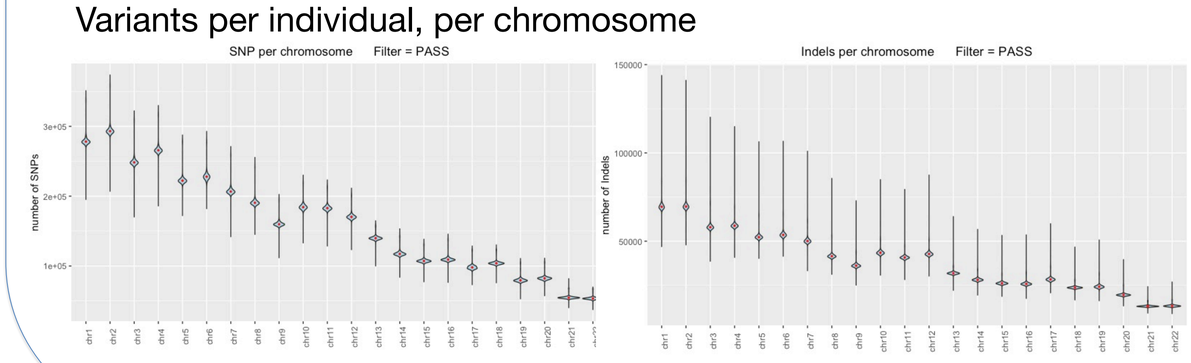
Data Available

Sequence from 59,464 individuals (release 5.1) has been analysed and 677,003,512 variants have been called on GRCh38.
Basic quality control has been applied:

Sample filtering:
- cross-contamination < 5%
- mapping rate > 75%
- mean sample coverage > 20
- insert size < 250

The variant PASS flag is set if:
- missingness < 5%
- coverage >= 10
- genotype quality >=15
- genotypes passing allelic imbalance test >= 0.25

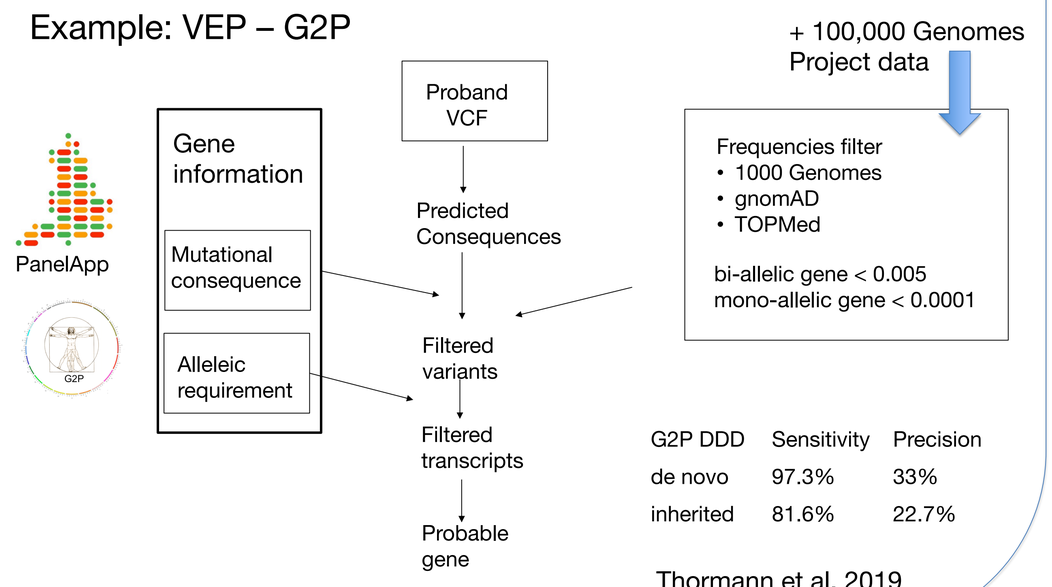
Variants per individual, per chromosome



Data Importance

The 100,000 Genomes Project will create a deep collection of population frequency data
- These data will enable an improved understanding of constraint in the genome
- Allele frequency distributions at this scale can accelerate diagnosis of rare disease

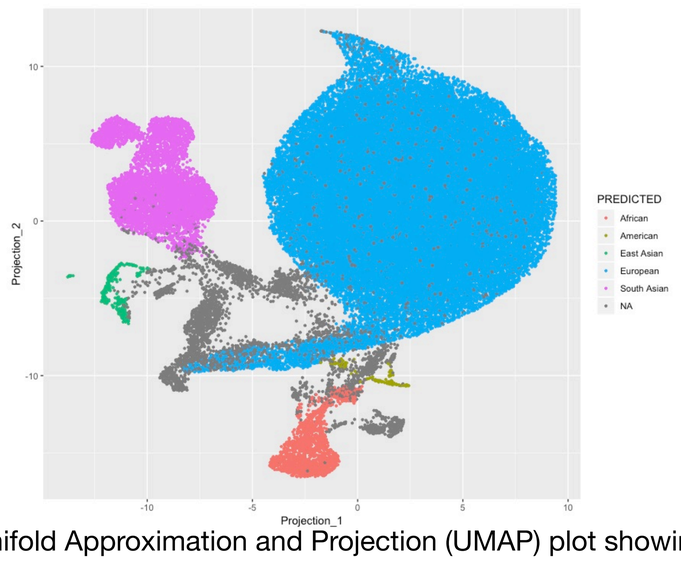
Example: VEP – G2P



G2P DDD	Sensitivity	Precision
de novo	97.3%	33%
inherited	81.6%	22.7%

Thormann et al. 2019

Ancestry prediction

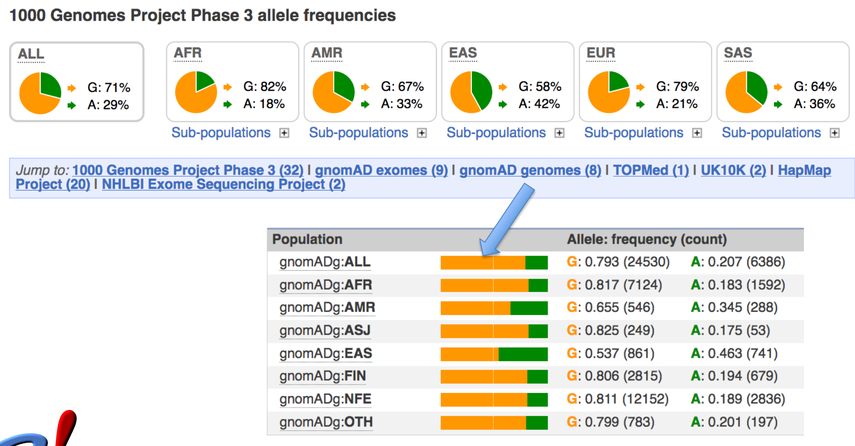


Uniform Manifold Approximation and Projection (UMAP) plot showing inferred ancestries.

Ancestries were inferred using the R-package randomForest, and data from the 1000 Genomes project (1kGP3) as the truth. The 1kGP3 study participants were split into broad ancestries (African, American, East Asian, European, and South Asian). Based on the similarity of the first 6 PCs between the 1kGP3 dataset and the 100,000 Genomes dataset, probabilities of each sample in the 100,000 Genomes dataset belonging to any one of the broad ancestries were calculated. A threshold of T=0.9 was applied for assignment. Participants below this threshold are shown in grey. For further details, see data freeze folder:
/gel_data_resources/main_programme/aggregated_illumina_gvcf/GRCH38/20190228

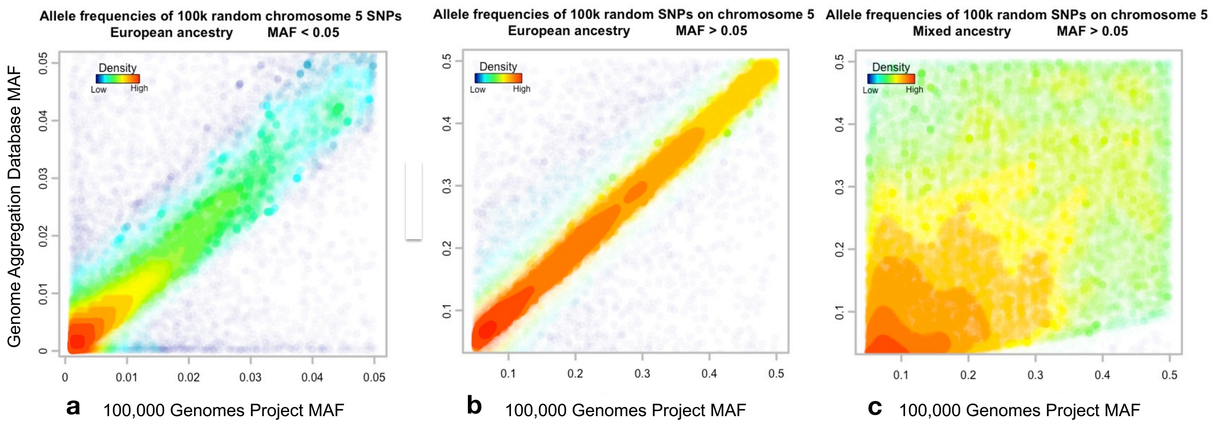
Data Distribution

- A 100,0000 Genomes Project data browser is planned
- Frequencies will be also displayed alongside other large scale data in the Ensembl browser and available in Ensembl VEP



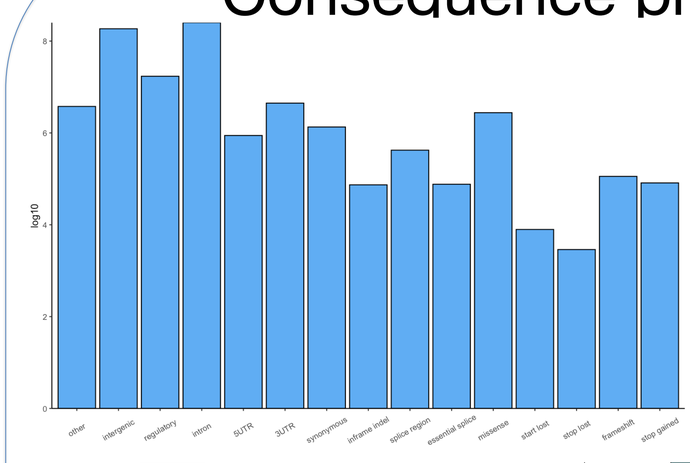
Example of current Ensembl frequency distribution views

Minor allele frequency comparison

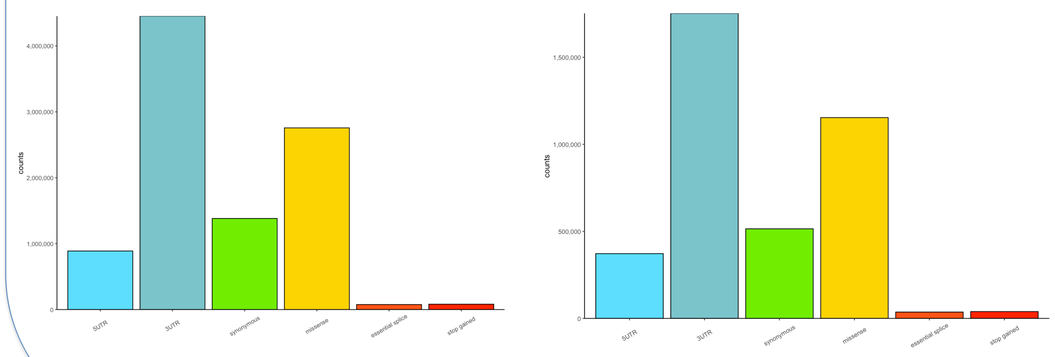


Minor allele frequencies (MAF) across cohorts, for variants with MAF below (**a**) or above (**b**) 5%, for individuals of European ancestry or between individuals of European and African ancestry **(c)**.
Every dot represents a pair of minor allele frequencies.
Colour indicates the proportion of SNPs with a pair of MAFs (red: common, blue: rare).
Data from unrelated participants was stratified by inferred ancestry
When allele frequencies across different ancestries are compared **(c)**, we see reduced concordance between the datasets as expected.

Consequence prediction



Most severe consequence per variant in the most supported Ensembl transcript per gene, as calculated with Ensembl VEP, version 98.
Passed variants only



All passed variants, major transcript consequences only.

Only singletons, major transcript consequences only.

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

Ensembl receives majority funding from the Wellcome Trust (grant number WT108749/Z/15/Z) with additional funding for specific project components from the National Human Genome Research Institute (U41HG007823 and 2U41HG007234), the Biotechnology and Biological Sciences Research Council (BB/L024225/1 and BB/M011615/1), Open Targets, the Wellcome Trust (WT104947/Z/14/Z, WT200990/Z/16/Z, and WT201535/Z/16/Z) and the European Molecular Biology Laboratory. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 634143 (MedBioinformatics). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 733161 (MultipleMS).