Journal Club
Barton et al. 2021

2021-07-14

Background: data

- UK Biobank contains genetic and phenotypic data of 500k **healthy** British, most of which are of European ancestry (focus in this study)
- Microarray data before March 2019
- 50k exome sequencing data released in March 2019
- Preprint submitted on 01 September 2020
- 200k exome sequencing data released in Oct/Nov 2020

Background: method

- Phasing:
  - Figure out which of the two haplotypes the variant is from
  - **Eagle 2**
- Imputation:
  - Figure out the genotype of a variant that is not assayed (often relying on the phased haplotypes)
  - *Minimac4*
- Association
  - **BOLT-LMM**
- Fine-mapping
  - Given a list of associated variants, to figure out which variant/variant set is **statistically** mostly likely to be causal
  - Custom method + FINEMAP

Research question:

Given 50k WES data, 500k array data and 500k phenotype data, how can we

- impute to produce 500k WES-like data,
- find its associations with the phenotype data
- and find the causal variants

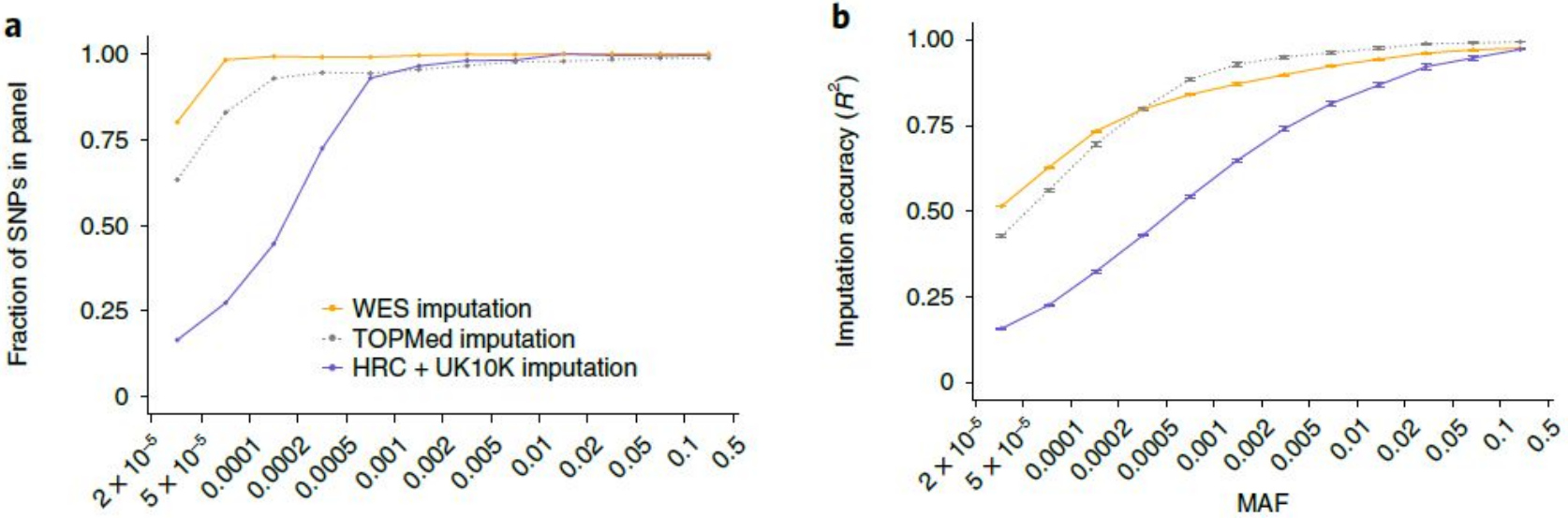
Why impute:

- Better power (even larger than using 200k exome associations alone)

Why focus on rare variants:

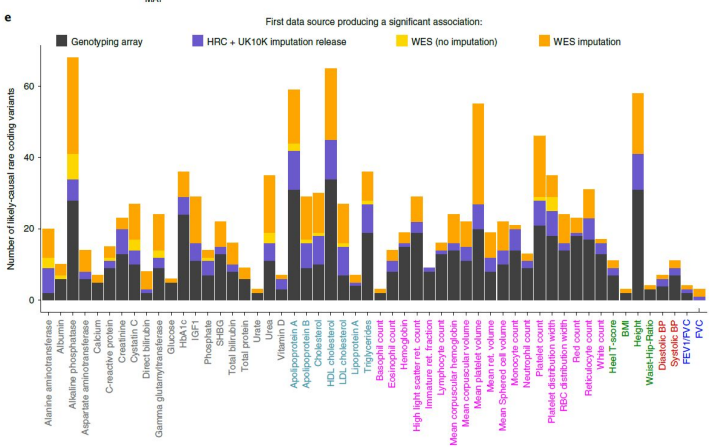
- Array already capture most of the common variants
- Common variants are more likely to be causal

Phasing and imputation



- Phase WES cohort to 4.9 million variants with **minor allele count** >= 2
- SNP array data were previously phased
- Impute with R2 > 0.5 for variants with MAF ~0.00005
- TOPMED: 97k genomes vs UKB 50k WES panel: similar results in panel coverage (left) and accuracy (right)

Fine-mapping



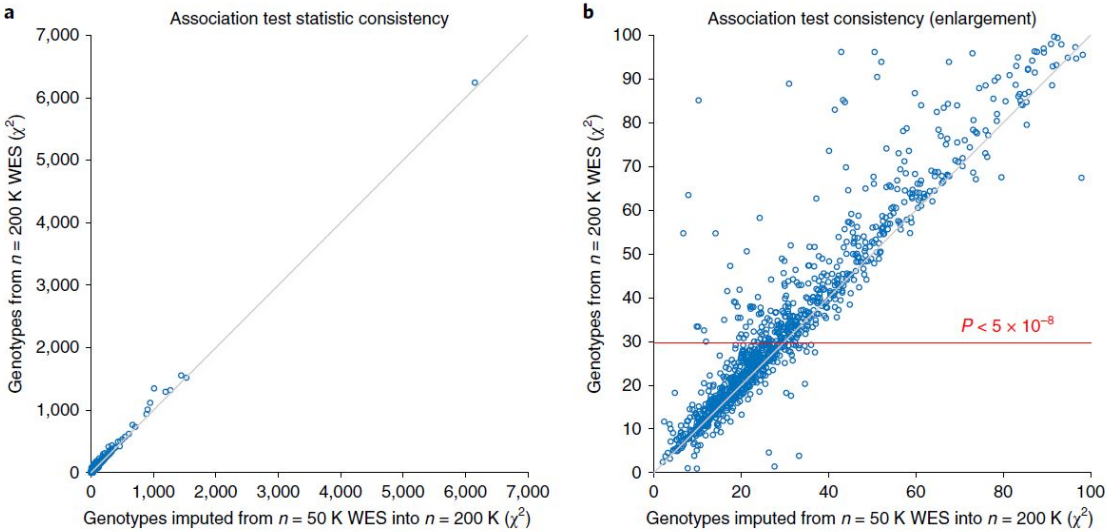
- Initial association yields tens of thousands of associations
- Filtering:
  - Criteria
    - MAF < 0.01 ∩
    - imputed MAF > 1e-5
    - ∩ INFO > 0.5 ∩
    - (CADD > 20 ∪ SpliceAI >= 0.5)
  - 529k rare coding variants, of which 1.6k have associations of $p < 5e\text{-}8$
  - 2.7k associations in total with 1.4 false discoveries expected
- Fine-mapping:
  - Custom method + FINEMAP (multiple causal variants model)
  - 1.1k associations involving 675 variants
- 30% discovered only using imputation method
- Compared with the most recent literature
  - 26% new gene-trait pairs found for blood cell traits
  - 45% found for height

Imputed vs 200k WES

Surveying the contribution of rare variants to the genetic architecture of human disease through exome sequencing of 177,882 UK Biobank participants

Quanli Wang, Ryan S. Dhindsa, Keren Carss, Andrew Harper, Abhishek Nag, Ioanna Tachmazidou, Dimitrios Vitsios, Sri VV Deevi, Alex Mackay, Daniel Muthas, Michael Hühn, Susan Monkley, Henric Olsson, Wasilewski, Katherine R. Smith, Ruth March, Adam Platt, łaefliger, Slavé Petrovski, AstraZeneca Genomics Initiative

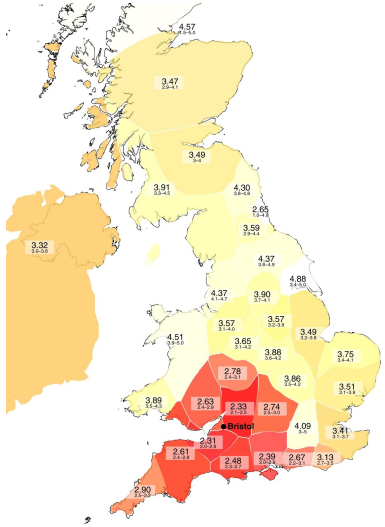
://doi.org/10.1101/2020.12.13.422582

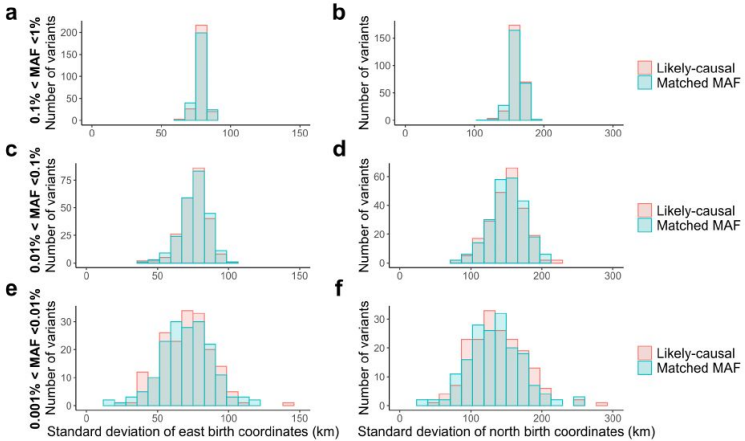


- Strong association between the test stats using imputed and 200k WES data
- Roughly half were not discoverable in WES data -> imputation R^2 of 0.5 gives power equivalent to 250k WES data

Confirmatory analysis

- Replicated associations from large-scale exome array studies
- Replicated when subsetting to unrelated UKB subjects
- Similar geographical distributions of likely causal alleles to MAF-matched background variants
  - Expect to be more localised if caused by population stratification
- Indicate that
  - Although common variant associations can be confounded by geographical stratification (Haworth 2019, Top Figure)
  - Rare variant association analysis in UKBB is not affected by highly localised stratification (Matheison 2012 based on simulated data)



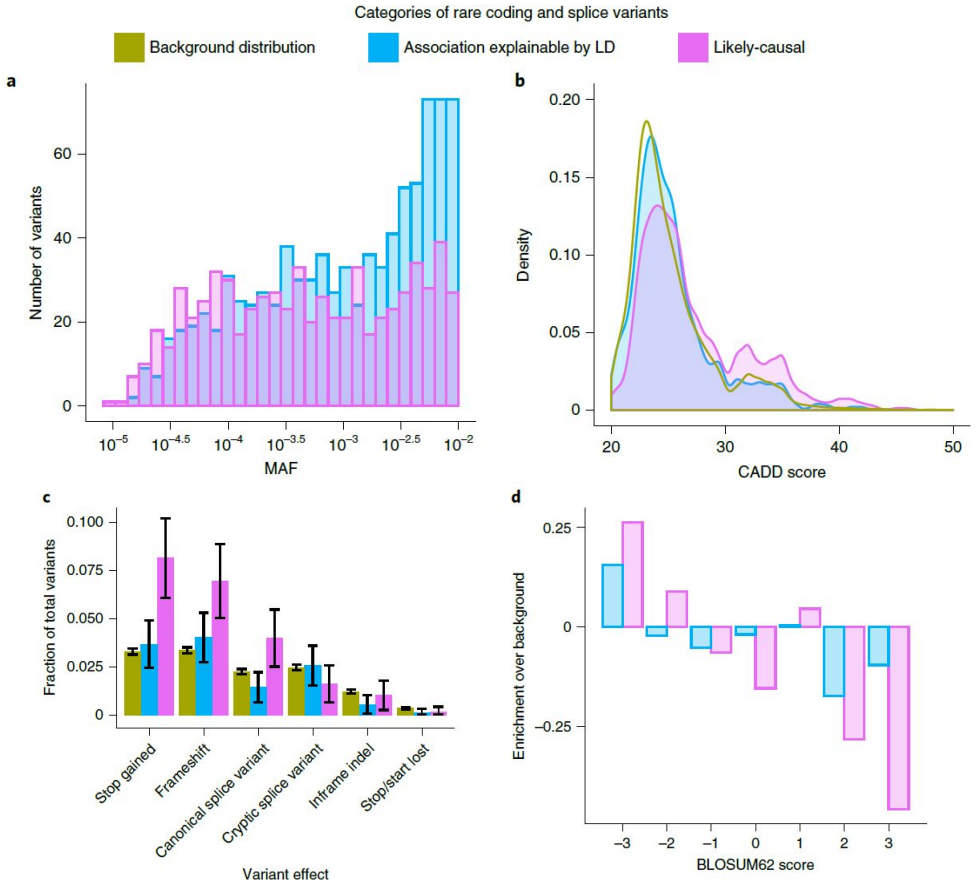


Likely causal: rare and deleterious

a. More common: more likely explained by LD (less likely causa
b. Likely causal: higher CADD
c. Predicted LOF effect
d. missense : more negative BLOSUM62 scores than background and explained by LD

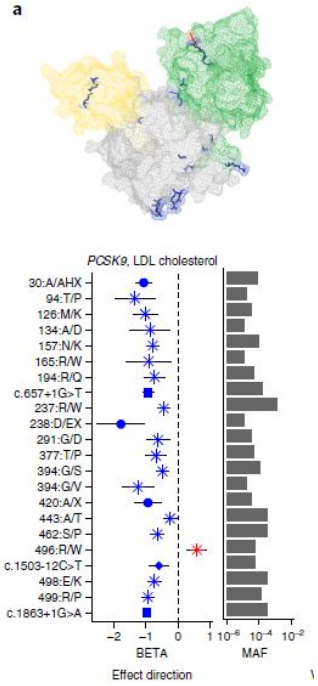
$$LogOddRatio = 2\log_2\left(\frac{P(O)}{P(E)}\right)$$

Mo… than expected



Rare coding variants form long allelic series

- Allelic series: variants in the same gene that produce a range of phenotypic effects
- To increase the power in detecting allelic series (i.e. additional independently associated variants), previous filters are relaxed
- 56 gene-trait pairs have allelic series of 10 or more variants on distinct haplotypes
- Distributed throughout protein structures and tend to have effect size of the same direction



Cell
Volume 169, Issue 7, 15 June 2017, Pages 1177-1186
CellPress

Perspective

An Expanded View of Complex Traits: From Polygenic to Omnigenic

Evan A. Boyle [1], Yang I. Li [1], Jonathan K. Pritchard [1, 2, 3]

Pleiotropic effects by rare variants

- one gene influences two or more seemingly unrelated phenotypic traits
  - “In PDE3B, the stop gain variant rs150090666 associated likely-causally with ten distinct traits, including expected associations with waist–hip ratio and lipid measurements, but also associations with red blood cell traits, sex hormone-binding globulin levels and height.”
- Omnigenic model:
  - “We propose that gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways.”

Large effects and disease risk

- Large effects (:
  - 10 novel variants with >0.5 sd effect sizes (previously identified was 0.3 sd)
- Disease risk):
  - 11/12 were genotyped/imputed before; 1/12 was implicated in Familial Hypercholesterolemia
  - “This behavior was consistent with the greater difficulty of identifying robust statistical associations with **disease** traits (for which **causal variants tend to have low penetrance**) as compared to **molecular or cellular** traits (for which causal variants can have much more **direct effects**).”

Single variant tests vs burden tests

- 32% gene-trait pairs in single-variant association tests were not detected by burden tests
- Most gene-trait associations from burden analysis were significant in single-variant analysis
- 37% of burden analysis significant results failed LD filters
  - Many could be false positives tagging a nearby gene
- 51% of burden-test associations were dominated by one variant
  - Collapsed genotype share LD
  - “Highlight the need to account for LD even in the context of burden analysis”
  - **How?**
- Burden test more likely to be helpful if
  - Variants are directly assayed, not genotyped
  - Whether a gene is strongly haploinsufficient

Related papers

- IBD-based association analysis
- Focus on fine-scale population structure and positive selection
- Only burden-style approach

Article | Open Access | Published: 30 November 2020

Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations

Juba Nait Saada, Georgios Kalantzis, Derek Shyr, Fergus Cooper, Martin Robinson, Alexander Gusev & Pier Francesco Palamara

Nature Communications **11**, Article number: 6130 (2020) | Cite this article

Discussion

- Would it be valid to impute the UKBB data with 100k Genome?
  - I guess array genotyping and exome sequencing are both investigating mostly coding regions so the LD is better captured, but neither really inform very well for the LD of the non-coding regions
- After the authors have validated the 50k imputation quality with the 100k release, why did they not just use the 100k data to have better imputation quality?
  - (could it be a project? Not enough novelty?)
- How applicable is the method for ELGH?