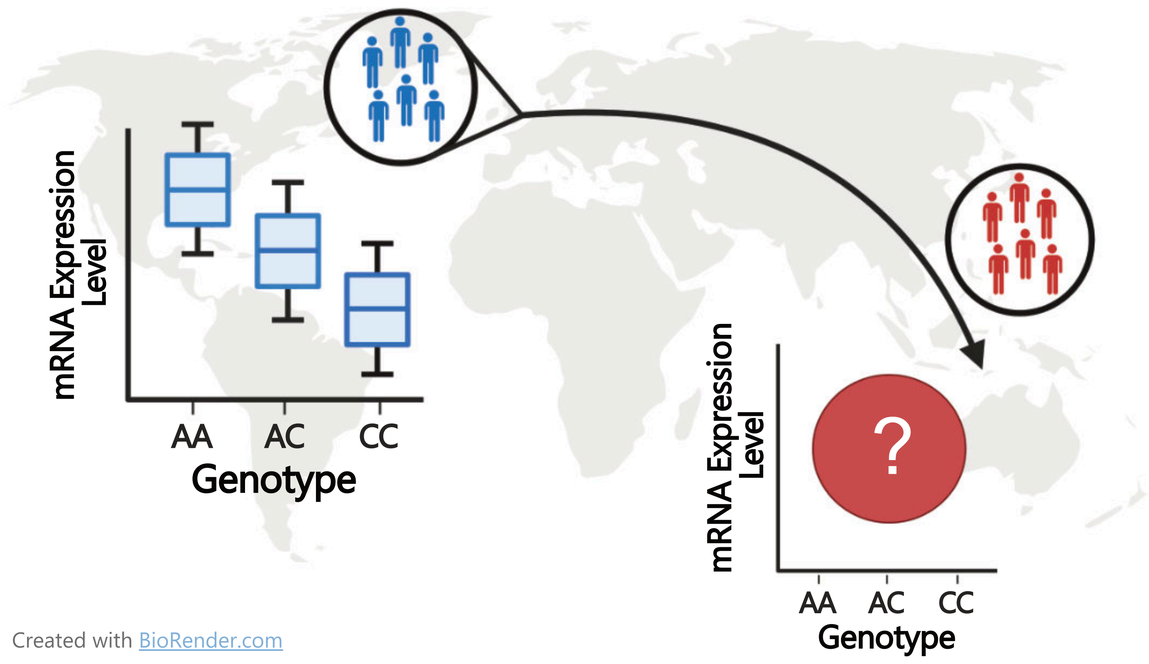
Predicting the cross-population portability of human expression quantitative trait loci (eQTLs)

Isobel J Beasley, Christina B Azodi, Irene Gallego Romero
@ijbeasley @cbazodi @ee_reh_neh

Motivation:



Created with BioRender.com

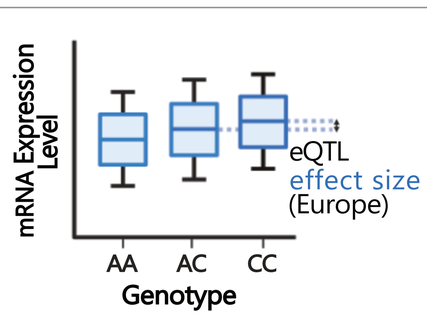
- The vast majority of participants in human genomics research are of European ancestry, to the detriment of scientific inquiry and the equitable translation of research (Martin et al., 2019, Nat. Genet., Landry et al., 2018, Health Aff.)
- Expression quantitative trait locus (eQTL) mapping identifies statistical associations between the genotype at a given locus and variation in gene expression
- eQTLs can be used to understand the gene regulatory consequences of disease-associated genetic variants (Gallagher and Chen Plotkin. 2018, Am. J. Hum. Genet., Umans et al., 2021, Trends Genet.)
- However, not all eQTLs are shared across populations
- Since Eurocentric biases are likely to persist for some time, **there is an urgent need to improve the transferability of European-derived eQTLs to underrepresentated populations** to ensure equitable research translation, including personalised medicine

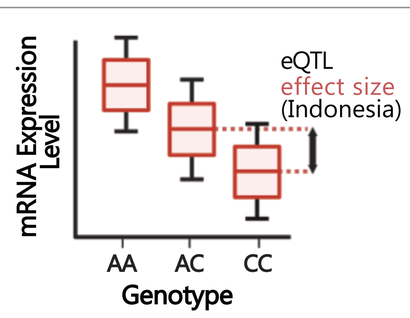
Could we use the features of eQTLs to predict their portability from European discovery cohorts to understudied populations?

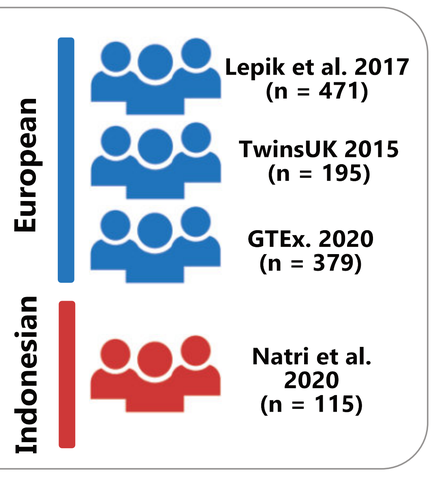
Method:

1. Classify eQTLs

- Combine summary statistics from eQTL studies in the same tissue
- Apply multiadaptive shrinkage method (mashr, Urbut et al. 2019 Nat. Genet.)
- Label eQTLs as population-specific or shared between Europe and Indonesia







Definition (inspired by Urbut et al., 2019, Nat. Genet.): An eQTL is **population-specific to Indonesia** if it is **statistically significant** (lfsr<0.01) **in Indonesia**, and **its' effect in Indonesia is inconsistent** (in direction and/or magnitude) with **European datasets**

2. Integrate Features

- Extract publicly available information on the evolutionary, functional, and expression properties of these eQTLs

gnomAD GTEx GENEONTOLOGY Unifying Biology

3. Train and Test Models

- Train random forest machine learning models to label eQTLs as Indonesian-specific or shared between Indonesia and Europe using these features
- Test performance on a held-out test set of eQTLs (chromosome 8,16)
- Each model uses a different definition of 'shared', so the proportion of test set eQTLs declared population-specific is different for each model

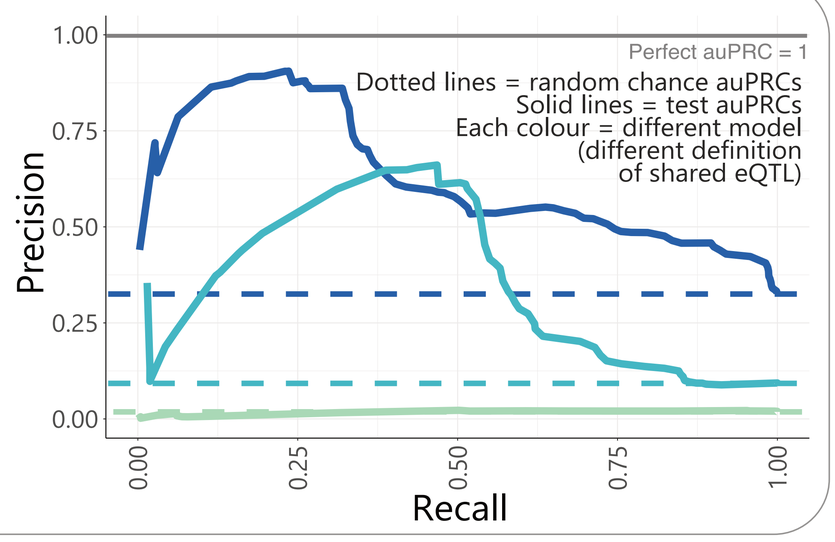
Results:

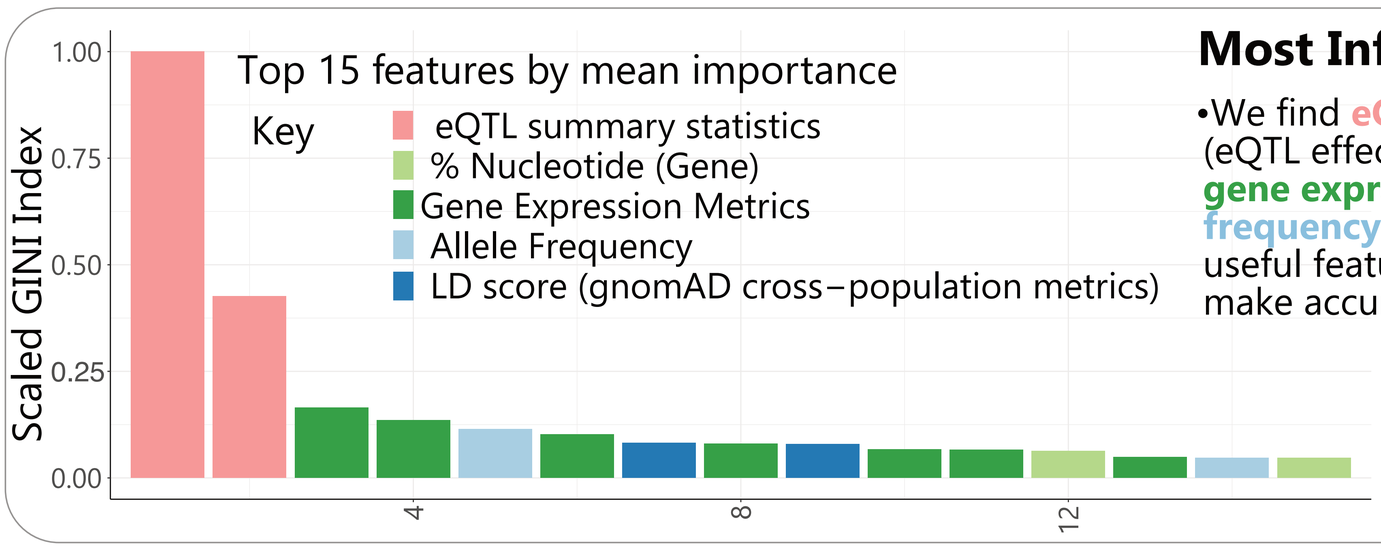
Prediction Performance

More relaxed definition of 'Indonesian-specific'

Definition of 'shared' eQTL effect:	Random chance	Test auPRC	Delta auPRC
Same direction	0.02	0.017	-0.003
Within a factor of 0.2	0.09	0.35	0.26
Within a factor of 0.5	0.33	0.63	0.30

- Used auPRC to measure how good our models were at detecting and correctly labelling eQTLs as Indonesian-specific (see Precision-Recall curves, right)
- Most trained models improved upon random chance (Delta auPRC > 0)





Most Informative Features

- We find **eQTL summary statistics** (eQTL effect size and standard error), **gene expression metrics** and **allele frequency** are some of the most useful features for our models to make accurate predictions

Main takeaways:

- The incomplete portability of eQTLs from European to underserved populations precludes the equitable translation of genomics research
- **Population-shared and specific eQTLs have different patterns of features** (Allele frequency, gene expression, eQTL effect size etc.)
- Hence, **machine learning models could be used to improve the transferability of eQTLs from Europeans to underrepresented populations**
- More work is needed to evaluate model performance across different contexts (populations, tissues etc.), to interpret trained models and to develop biologically meaningful and useful definitions of 'shared' eQTL effects

Acknowledgements:
Thanks for all the support and advice: Davis McCarthy, the BioCellGen Group at SVI and the Gallego Romero Group at the University of Melbourne!