“Singleton Variants Dominate the Genetic Architecture of Human Gene Expression” and its application

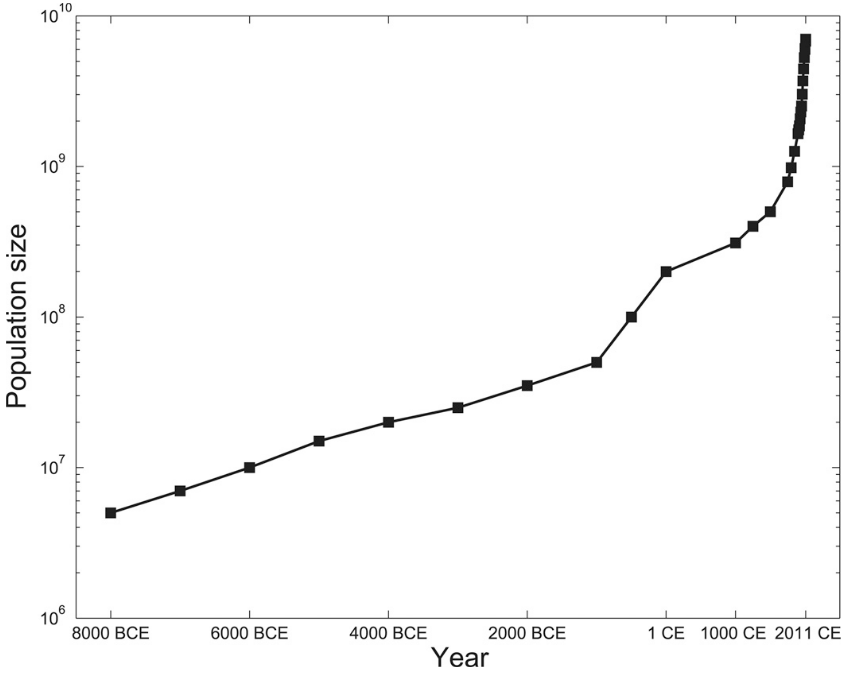
Mar 7, 2019

Wonji Kim

Introduction

- Recent explosive growth of human populations
  - Abundance of genetic variants with MAF < 1%
- Role of rare variants
  - Mendelian diseases vs complex diseases
- Improvement in imputation services
  - Imputation quality of rare variants

Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *science*, *336*(6082), 740-743.



- However, these studies excluded the rarest variants or included only well-imputed variants

Introduction

- Goal
  - Development of an approach for inferring the relative phenotypic contributions of all variants, from **singletons** to high frequency
- Application
  - Narrow-sense heritability of gene expression
- Evaluation of robustness to
  - Genotyping errors
  - Read mapping errors
  - Population structure
  - Rare variant stratification
  - Wide range of possible genetic architecture

Partitioning heritability by MAF

- Overview of model and method
  - *M* SNPs and *N* individuals,

$$y_i = \sum_{j=1}^{M} g_{ij}\beta_j + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma_e^2)$$

where g_{ij} is the genotype of individual *i* at SNP *j*

β_j is the effect size of SNP *j*

ϵ_i is the residual for individual *i*

  - They partition the SNPs into *K* disjoint sets determined by the MAF and heritability of *k*th SNP set is

$$h_k^2 = \sigma_k^2 / \sigma_y^2$$

$$\sigma_g^2 = \sum_{k=1}^{K} \sigma_k^2 \;\&\; \sigma_y^2 = \sigma_g^2 + \sigma_e^2 = 1$$

Partitioning heritability by MAF

- Haseman-Elston (H-E) regression
  - Phenotypic covariance (*P*) : for a single gene, the outer product of quantile-normalized FPKM across individuals
  - Genotypic covariance (R_k) : for *k*th partition, a kinship matrix generate from all SNPs in the partition

$$R_k = G_k G_k{}'/M_k$$

  where G_k is a column-standardized genotype matrix of SNPs in the *k*th partition (*N* rows and M_k coulmns)

  - H-E regression is then performed using the *lm*() function in R:

$$P \sim R_1 + \cdots + R_K$$

Partitioning heritability by MAF

- Haseman-Elston (H-E) regression
  - The effect size for the *k*th term represents the genetic variance explained by the *k*th SNP partition ($\beta_k = \sigma_k^2$)
  - Total genetic variance explained by all SNPs given by $\sigma_g^2 = \sum_{k=1}^{K} \sigma_k^2$.
  - Heritability

$$h^2 = \sigma_g^2$$

Partitioning heritability by MAF

- Singleton heritability
  - *N* individuals and *M* SNPs, the linear mixed model (LMM) for phenotype vector $y \in R^{N\times 1}$ and an $N\times M$ SNP genotype matrix $G \in \{0,1,2\}^{N\times M}$*:*

$$y = G\beta + \epsilon,$$

$$\beta_j \sim N\left(0, \frac{1}{M}\sigma_g^2\right),\ \epsilon_i \sim N(0, \sigma_e^2)$$

  - If we define $u = G\beta$, then heritability is given by

$$h^2 = \frac{Var(u)}{Var(y)}$$

Partitioning heritability by MAF

- Singleton heritability
  - Assume that *G* consists of only **singletons**. Then, u_i simplifies:

$$u_i = (G\beta)_i = \sum_{j=1}^{M} G_{ij}\beta_j = \sum_{j:G_{ij}=1}^{M} N\left(0, \frac{1}{M}\sigma_g^2\right) \sim N\left(0, x_i\sigma_g^2\right)$$

where $x_i = \frac{\text{\# singletons for person } i}{\text{\# singletons total}} = \frac{\sum_j G_{ij}}{M}$

  - The phenotype vector *y* simplifies to marginal models on each observation:

$$y_i \sim N(0, x_i\sigma_g^2 + \sigma_e^2)$$

  - The heritability is simple to evaluate:

$$h^2 = \frac{E(Var(u|x)) + Var(E(u|x))}{E(Var(y|x)) + Var(E(y|x))} = \frac{E\left(x\sigma_g^2\right)}{E\left(x\sigma_g^2 + \sigma_e^2\right)} = \frac{\frac{1}{N}\sigma_g^2}{\frac{1}{N}\sigma_g^2 + \sigma_e^2}$$

Simulation studies

- Simulation data
  - Real genotype data by randomly sampling genes
  - All genetic variants within 1 Mb of transcription start and end sites of genes
- Simulation parameters

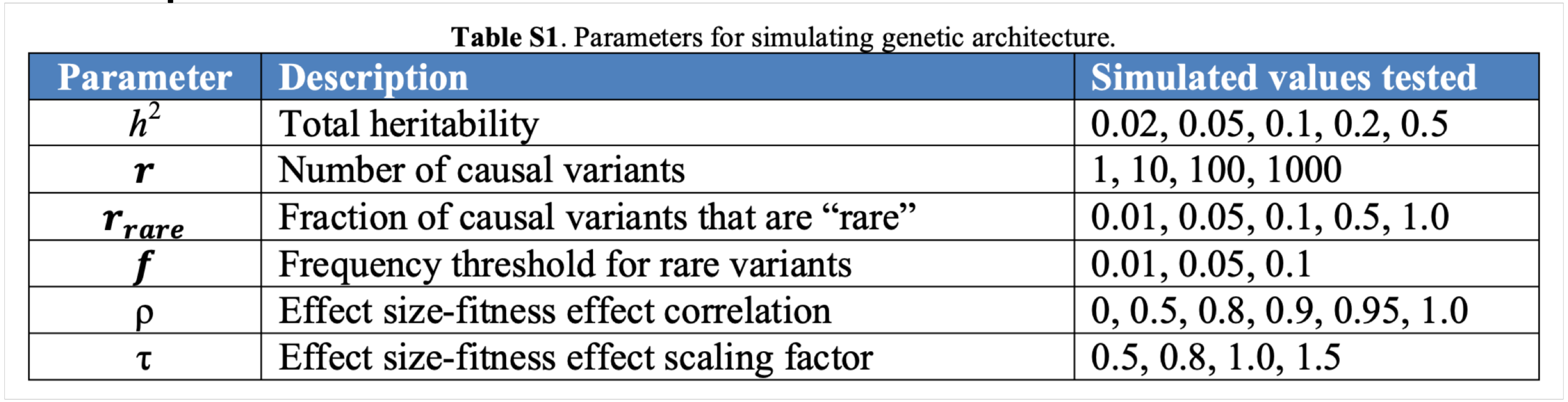
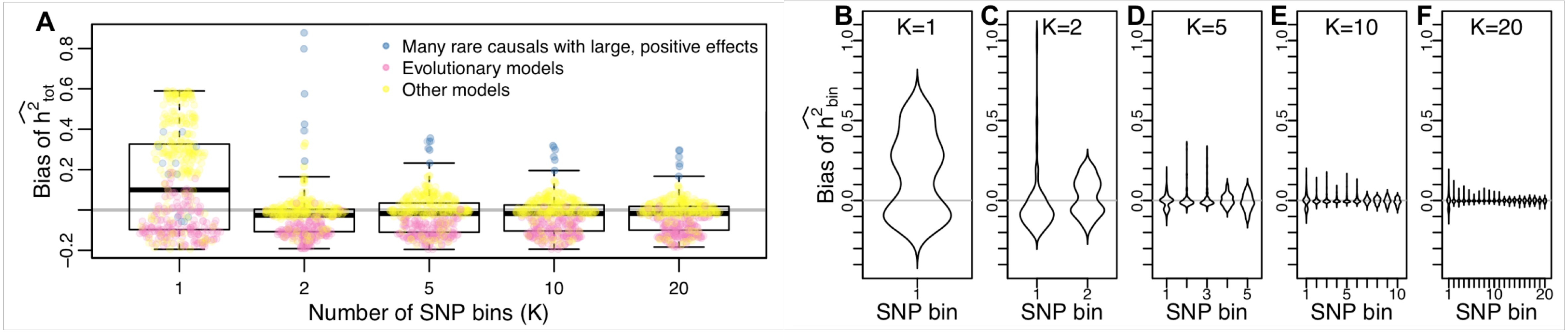


Table S1. Parameters for simulating genetic architecture.

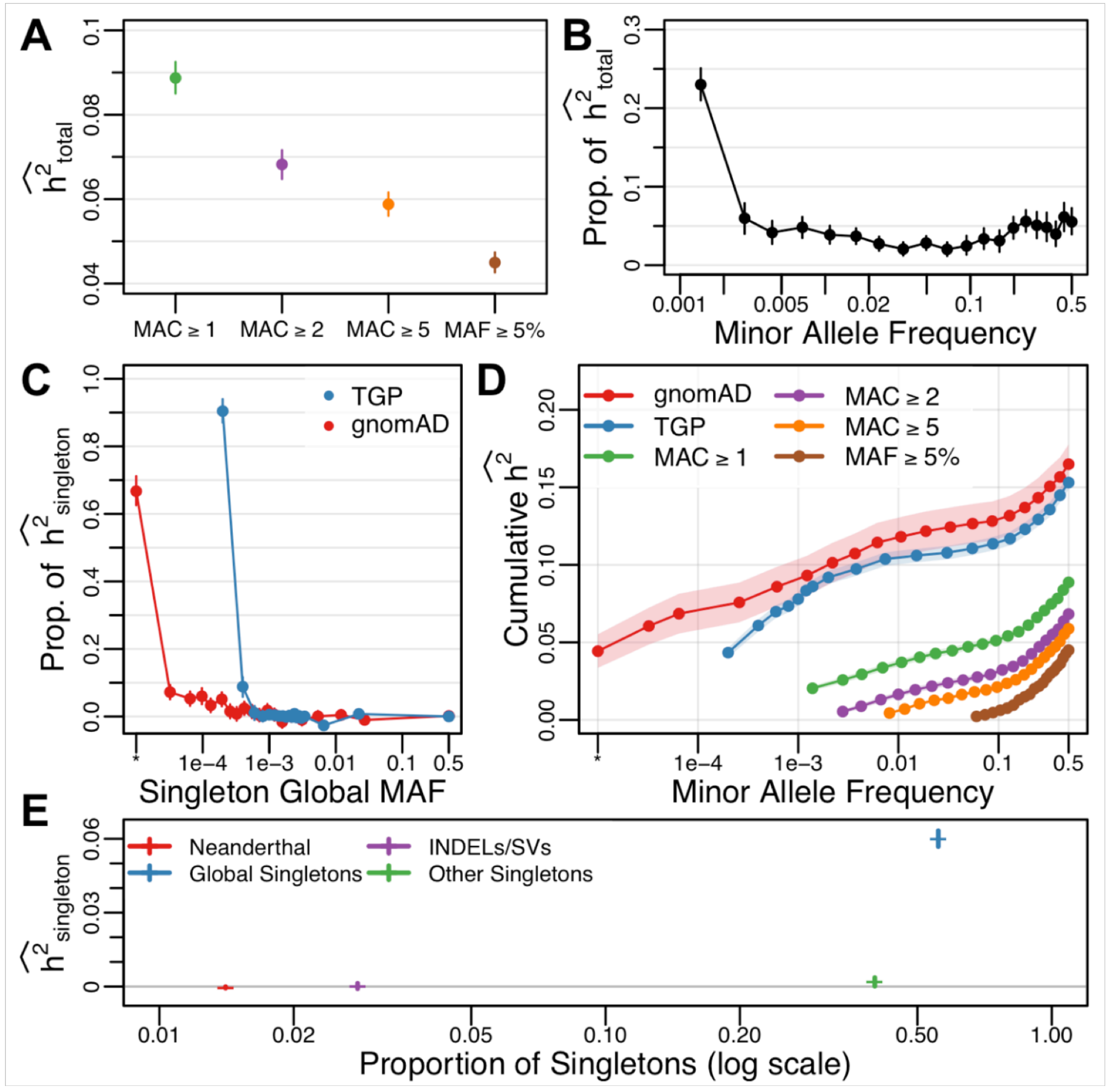
Parameter	Description	Simulated values tested
h^2	Total heritability	0.02, 0.05, 0.1, 0.2, 0.5
$\boldsymbol{r}$	Number of causal variants	1, 10, 100, 1000
$\boldsymbol{r_{rare}}$	Fraction of causal variants that are "rare"	0.01, 0.05, 0.1, 0.5, 1.0
$\boldsymbol{f}$	Frequency threshold for rare variants	0.01, 0.05, 0.1
ρ	Effect size-fitness effect correlation	0, 0.5, 0.8, 0.9, 0.95, 1.0
τ	Effect size-fitness effect scaling factor	0.5, 0.8, 1.0, 1.5

Simulation studies



- Across a broad range of parameters, the accuracy of heritability interference improves as the number of SNP bins increases.

Simulation studies



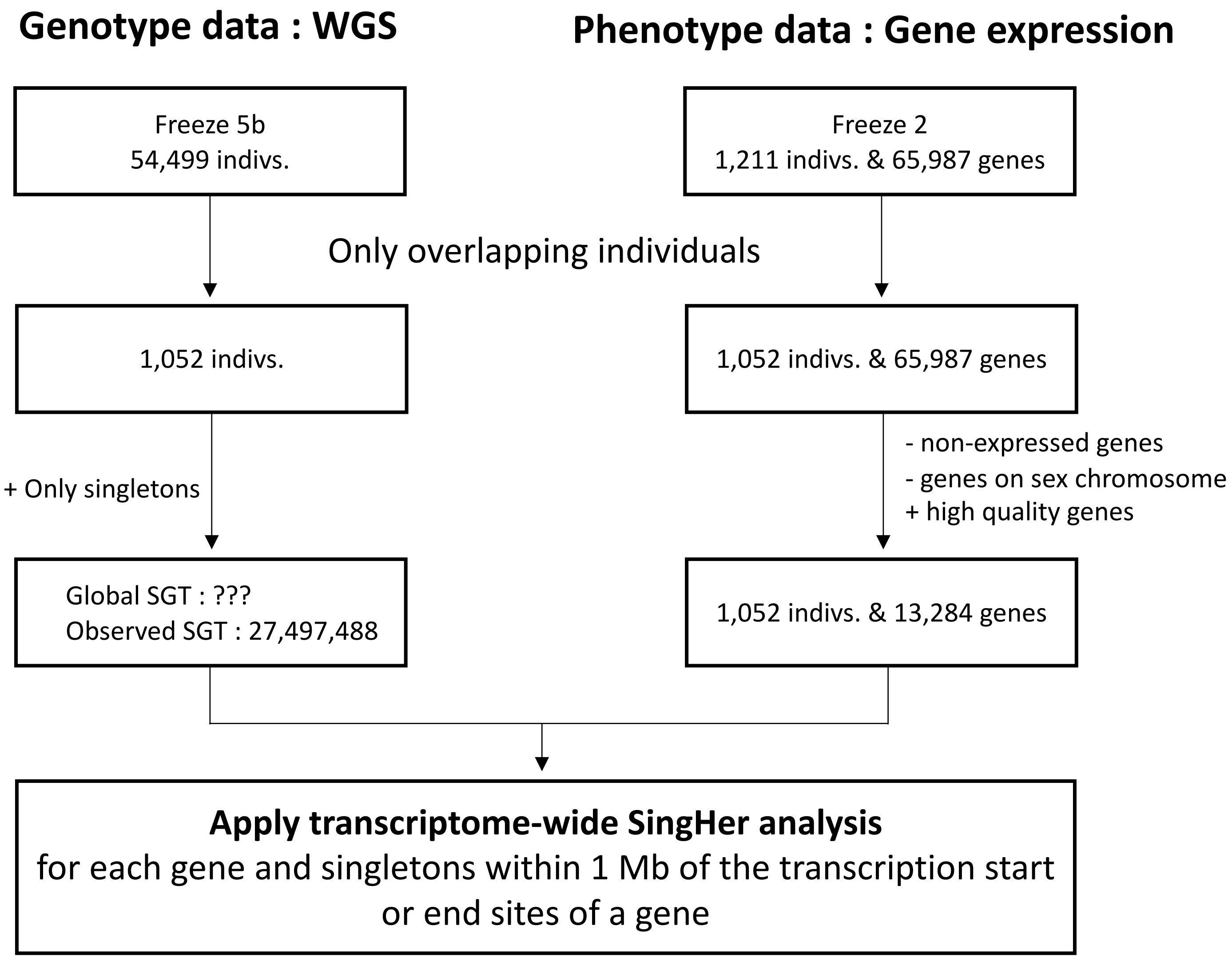
- Characterizing the genetic architecture of human gene expression
  - A. Average total heritability inferred across genes for different frequency filters
  - B. The proportion of heritability attributed to each MAF bin
  - C. Partitioning singletons by global MAF based on TGP and gnomAD
  - D. Cummulative heritability
  - E. Singleton heritability for type of singletons

Software availability

- Three open source software tools are available by request to the authors
  - SingHer.R – Singleton Heritability inference with REML implementation in R of the unbiased singleton-based LMM
  - HEplay.R – H-E regression simulation in R that implements all the genotype-phenotype maps
  - HEh2.R – H-E regression implementation in R that performs all H-E analyses

Preliminary SingHer analysis for COPDGene

COPDGene dataset and QC



Genotype data : WGS

Phenotype data : Gene expression

Freeze 5b
54,499 indivs.

Freeze 2
1,211 indivs. & 65,987 genes

Only overlapping individuals

1,052 indivs.

1,052 indivs. & 65,987 genes

+ Only singletons

- non-expressed genes
- genes on sex chromosome
+ high quality genes

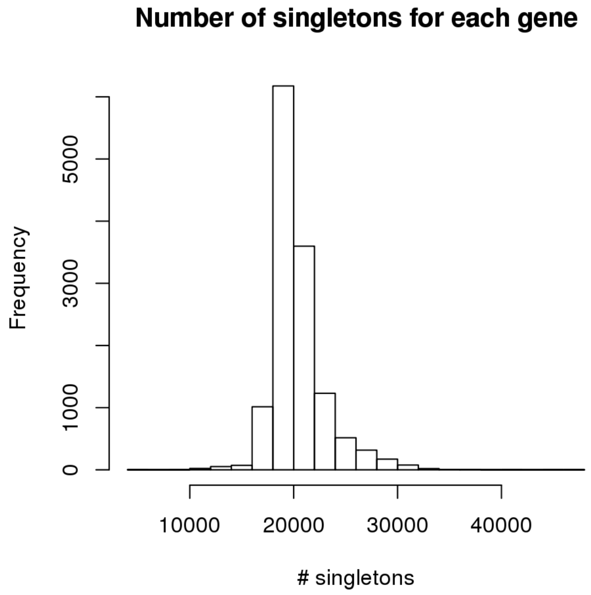
Global SGT : ???
Observed SGT : 27,497,488

1,052 indivs. & 13,284 genes

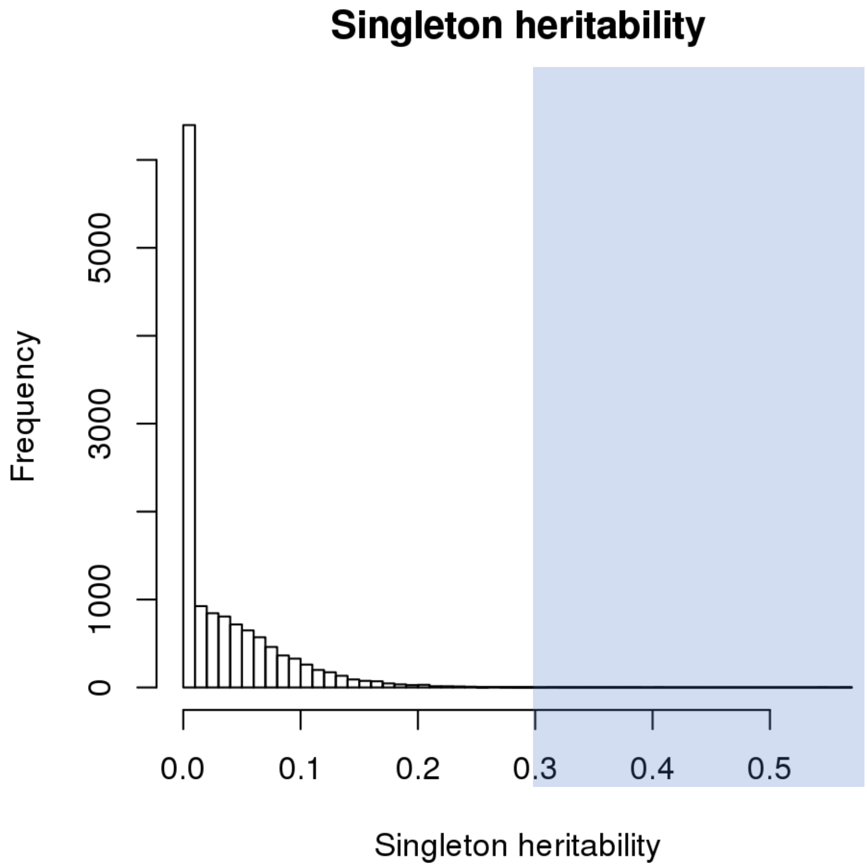
Apply transcriptome-wide SingHer analysis
for each gene and singletons within 1 Mb of the transcription start or end sites of a gene

The number of genes for which the proportion of individuals (x) has log(CPM) >Y. (Row : X, Column : Y)

	1	1.5	2	2.5	3	4
0	27805	24088	20993	18430	16318	12936
0.1	19167	17030	15203	13650	12223	9507
0.2	17873	15946	14346	12867	11524	8892
0.3	16951	15239	13714	12335	11028	8480
0.4	16179	14560	13130	11818	10553	8054
0.5	15306	**13820**	12431	11188	9986	7603
0.6	14714	13271	11972	10732	9568	7278
0.7	14097	12747	11511	10320	9167	6899
0.8	13447	12179	10998	9848	8741	6528
0.9	12583	11429	10314	9240	8198	6036



SingHer analysis



Gene_Name	CHR	Start_bp	End_bp	Gene_Type	h2
MYOM1	18	3066807	3220108	protein_coding	0.5690
MTCO1P12	1	631074	632616	unprocessed_pseudogene	0.5475
ABCA5	17	69244311	69327244	protein_coding	0.4044
AL008721.2	22	25476218	25479971	sense_intronic	0.3857
HEBP2	6	138403531	138422197	protein_coding	0.3820
LINC00937	12	8295986	8396803	lincRNA	0.3813
RNF182	6	13924446	13980302	protein_coding	0.3731
HERC2P9	15	28589492	28685264	transcribed_unprocessed_pseudogene	0.3634
ST6GALNAC2	17	76565379	76586956	protein_coding	0.3586
MIR646HG	20	60087840	60527458	lincRNA	0.3523
VWDE	7	12330885	12403941	protein_coding	0.3442
CDC27	17	47117703	47189422	protein_coding	0.3330
1-Mar	1	220786759	220819657	protein_coding	0.3320
CNTNAP3	9	39072767	39288315	protein_coding	0.3278
LRRC6	8	132571953	132675617	protein_coding	0.3222
AC011472.2	19	11300777	11324441	3prime_overlapping_ncRNA	0.3185
CRYBB2P1	22	25448105	25520854	transcribed_unprocessed_pseudogene	0.3135
FCAR	19	54874248	54890472	protein_coding	0.3134
RBP7	1	9997206	10016020	protein_coding	0.3117

Further works…

- Using global singletons using TOPMed WGS data
- Considering missingness rate for quality control of genotype data (and comparing the results)
- Applying to other quantitative traits such as FEV_1, FEV_1/FVC …

Thank you