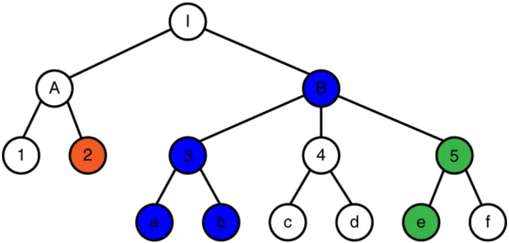
TreeWAS

Exploring hierarchical phenotypic data in genomic datasets

Lino Ferreira

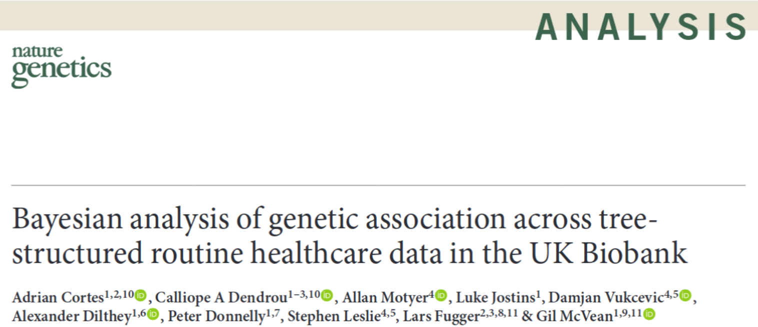
6th December 2019



Some genomic datasets organise phenotypic information in a **tree of diseases**.

*How can we use this greater resolution to **estimate associations more accurately** without sacrificing statistical power?*

- Model the **correlation structure** of the genetic effects across different phenotypes

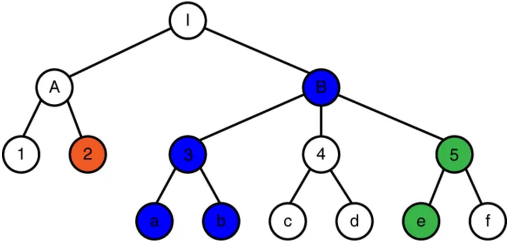


ANALYSIS

nature genetics

Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank

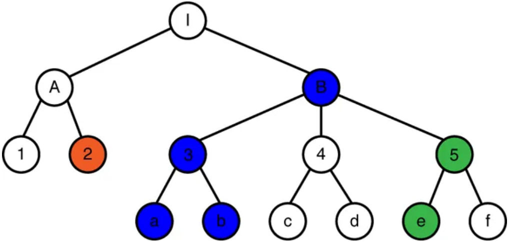
Adrian Cortes[1,2,10], Calliope A Dendrou[1–3,10], Allan Motyer[4], Luke Jostins[1], Damjan Vukcevic[4,5], Alexander Dilthey[1,6], Peter Donnelly[1,7], Stephen Leslie[4,5], Lars Fugger[2,3,8,11] & Gil McVean[1,9,11]



Each node j is a binary indicator of disease modelled through **logistic regression**:

$$\operatorname{logit}\left(\mathbb{P}(Z_j = 1)\right) = \beta_j^0 + \beta_j^1\,\mathbb{I}(\text{heterozygous}) + \beta_j^2\,\mathbb{I}(\text{homozygous})$$

The β coefficients evolve down the tree in a **Markov process**.



The β coefficients evolve down the tree in a Markov process:

- Parent coefficients inherited with probability $e^{-\theta}$
- Otherwise drawn from mixture prior:
    - Null with probability π_1
    - Otherwise drawn from joint mean-zero normal

Use dynamic programming to determine the **marginal posterior probability** that each coefficient is non-zero and estimate its effect size.

Achieve an **increase in power** of more than 20%.