Statistical challenges of identifying the genetic determinants of rare diseases

Ernest Turro

Department of Haematology
University of Cambridge

MRC Biostatistics Unit

NHS
National Institute for
Health Research

MRC
Medical
Research
Council

16 Sep 2016

Outline

1. Overview of rare diseases projects in the UK
2. Phenotypic heterogeneity
3. Genetic heterogeneity and neutral variants

NIHR BioResource – Rare Diseases (BRIDGE)

BRIDGE is an international consortium aiming to discover the molecular basis of several classes of heritable rare diseases:

- PID - Primary Immune Disorders
- PAH - Pulmonary Arterial Hypertension
- BPD - Bleeding and Platelet Disorders
- SPEED - Specialist Pathology: Evaluating Exomes in Diagnostics (retinal and developmental disorders)
- SRNS - Steroid Resistant Nephrotic Syndrome
- EDS - Ehlers Danlos Syndrome
- . . .

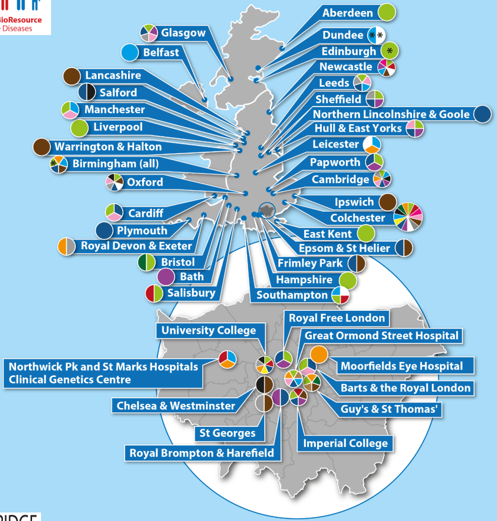
Uniform sequencing data generation across projects
500–2,000 genomes per project

Supporting institutions in the UK

NIHR BioResource
Rare Diseases

NIHRBR-RD Recruiting sites and projects

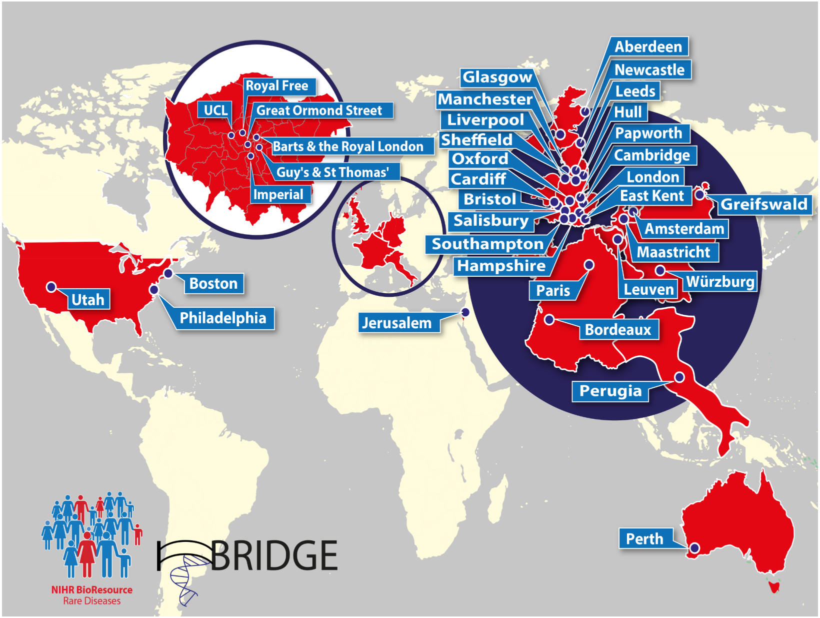
NHS
National Institute for
Health Research



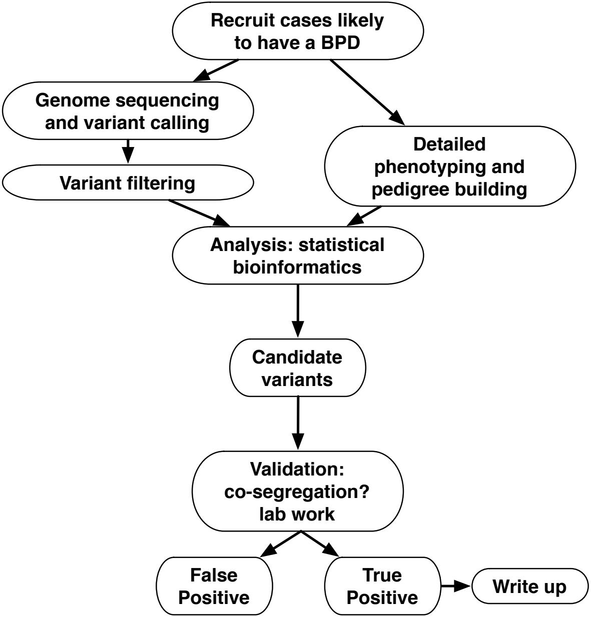
- ● Bleeding & Platelet Disorders
- ● Stem Cells & Myeloid Disorders
- ● Primary Immune Disorders
- ● Pulmonary Arterial Hypertension
- ● Cerebral Small Vessel Diseases
- ● Ehlers Danlos Syndrome
- ● Steroid Resistant Nephrotic Syndrome
- ● Primary Membranoproliferative Glomerulonephritis
- ● Specialist Pathology Evaluating Exomes in Diagnostics
- ● Neuropathic Pain Disorders
- ● Multiple Primary Malignant Tumours
- ● Intrahepatic cholestasis of pregnancy
- ● Leber Hereditary Optic Neuropathy
- ○ Myofilament-gene negative Hypertrophic Cardiomyopathy

BRIDGE

Supporting institutions overseas



Bleeding and Platelet Disorders (BPD)



Recruit cases likely to have a BPD

Genome sequencing and variant calling

Variant filtering

Detailed phenotyping and pedigree building

Analysis: statistical bioinformatics

Candidate variants

Validation: co-segregation? lab work

False Positive

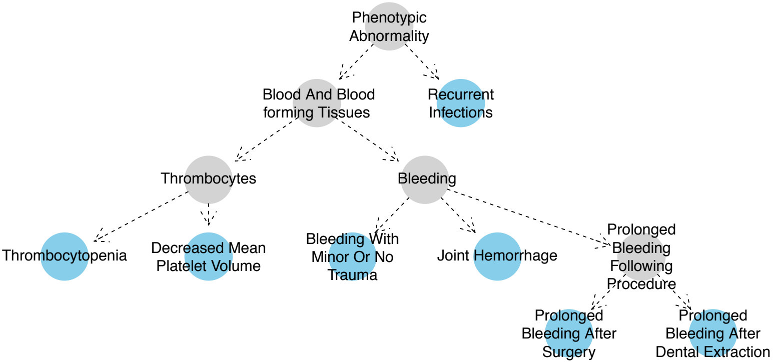
True Positive

Write up

Patient coding with Human Phenotype Ontology (HPO)

Thousands of patients with rare diseases are being sequenced

Important to standardise patient phenotypes

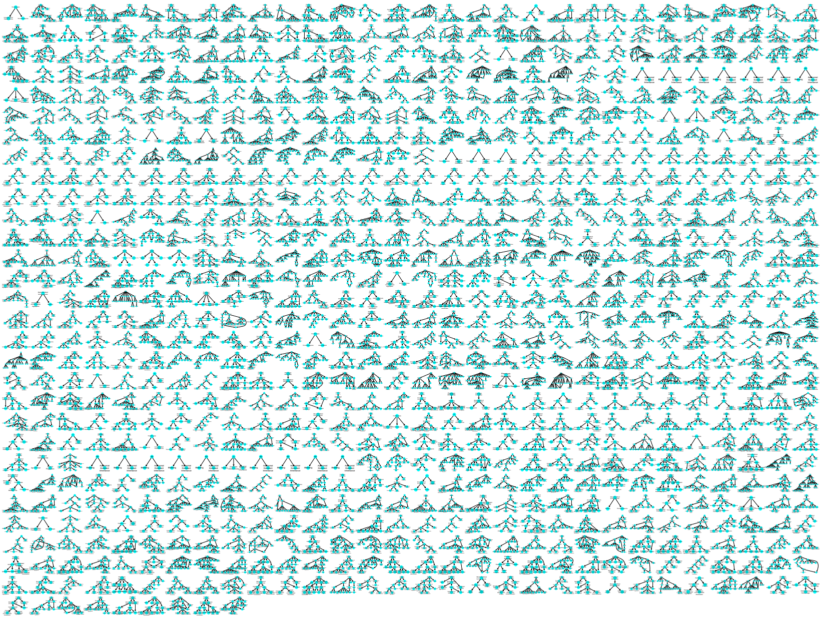

Phenotypic Abnormality
Blood And Blood forming Tissues
Recurrent Infections
Thrombocytes
Bleeding
Thrombocytopenia
Decreased Mean Platelet Volume
Bleeding With Minor Or No Trauma
Joint Hemorrhage
Prolonged Bleeding Following Procedure
Prolonged Bleeding After Surgery
Prolonged Bleeding After Dental Extraction

Terms in blue form a *minimal set* (non-redundant)

Terms in grey are implied by the terms in blue

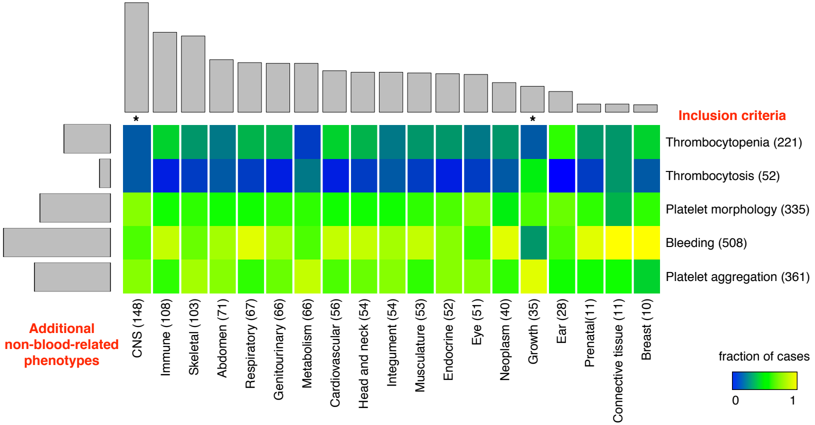
HPO has $> 10,000$ terms spanning all organ systems

HPO terms for BPD index cases



Overview of HPO coding of BPD probands

~ 7 HPO codes per person; many are non-haematological



Westbury *et al.*, *Genome Medicine*, 2015

Take into account phenotypes outside primary area of interest

Difficulties of grouping patients into groups

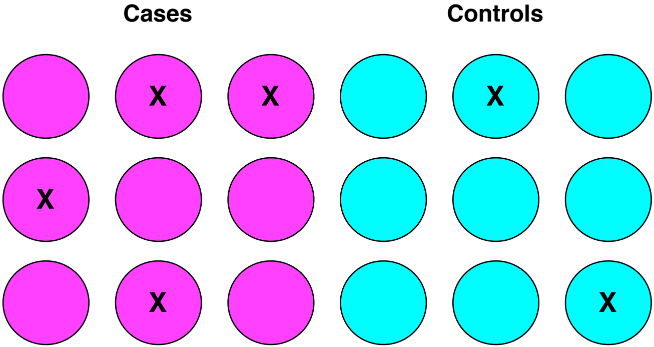
Phenotypic heterogeneity (variable expressivity of traits)

MYH9 mut.	Pheno.	Hearing Loss	Urine abn.	Renal dysf.	Cataracts
S96L	Epstein	–	+	ND	–
S96L	Epstein	+	+	CKD5	–
S96L	Epstein	+	–	–	–
S96L	Epstein	–	–	–	–
S96L	Epstein	+	+	ND	–
S96L	Epstein	+	+	ND	–
S96L	Epstein	+	+	ND	–
S96L	MYH9-RD	ND	ND	ND	ND
S96L	MYH9-RD	ND	ND	ND	ND
S96L	Fechtner	+	+	CKD5	+
S96L	Epstein	+	+	CKD4	–
S96L	Epstein	+	+	CKD5	–
S96L	MCTP	–	–	–	–

Murayama *et al.*

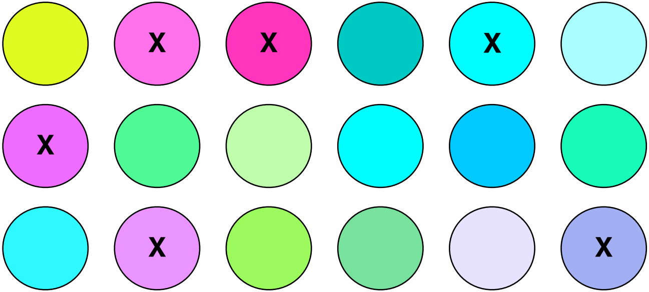
A priori grouping into clusters with shared (though unknown) genetic aetiology using phenotypes alone is challenging

Power of HPO-based analysis



X: has variant in gene

Power of HPO-based analysis



X: has variant in gene

Regression methods summarising phenotypes with unstructured binary or quantitative variables may be underpowered

Bayesian phenotype similarity regression

Compare "inverse regression" models (*genotype* is the response):

$$y_i \text{ ("rare genotype") } \sim \text{Bernoulli}(p_i)$$

Baseline model ($\gamma = 0$):

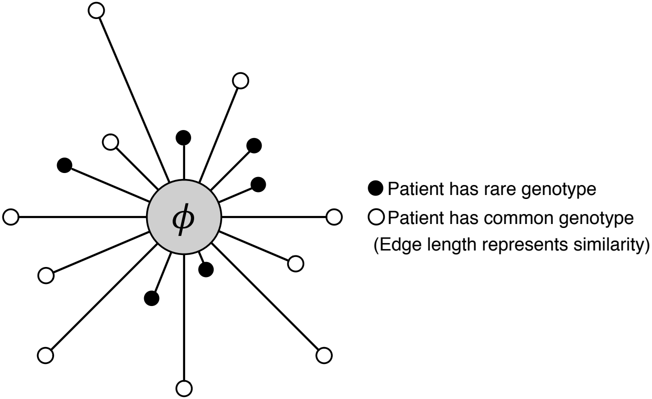
$$\log\left(\frac{p_i}{1 - p_i}\right) = \alpha$$

Alternate model ($\gamma = 1$):

$$\log\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta \cdot S(x_i, \phi)$$

- $y_i \in (0, 1)$: "genotype" of patient *i* (e.g. has > 0 rare variants)
- x_i: HPO phenotype of patient *i*
- ϕ: latent HPO-coded *characteristic phenotype* of disease
- α: background rate of rare genotype
- β: effect of phenotypic similarity on log odds of rare genotype
- $S(x_i, \phi)$: similarity of patient *i* to characteristic phenotype ϕ

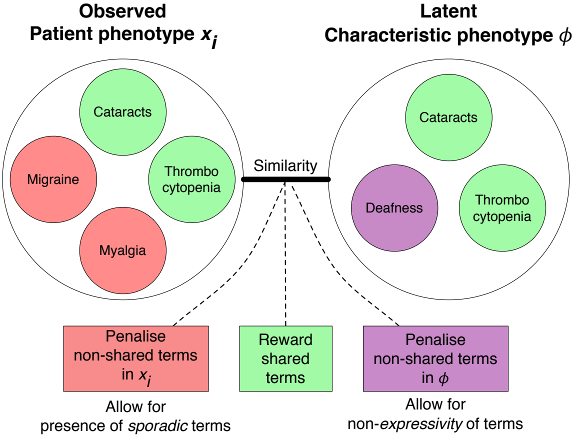
$\gamma = 1$ model


ϕ
● Patient has rare genotype
○ Patient has common genotype
(Edge length represents similarity)

- ϕ: latent HPO-coded *characteristic phenotype* of disease
- Edge length: similarity between patient *i* and ϕ
- Prior on ϕ informed by terms in human/mouse databases

Similarity measure

- Sharing of HPO terms → increases similarity
- Non-sharing of HPO terms → decreases similarity
- Rare terms carry more weight



Observed
Patient phenotype x_i

Cataracts

Migraine

Thrombo cytopenia

Myalgia

Similarity

Latent
Characteristic phenotype ϕ

Cataracts

Deafness

Thrombo cytopenia

Penalise non-shared terms in x_i

Allow for presence of *sporadic* terms

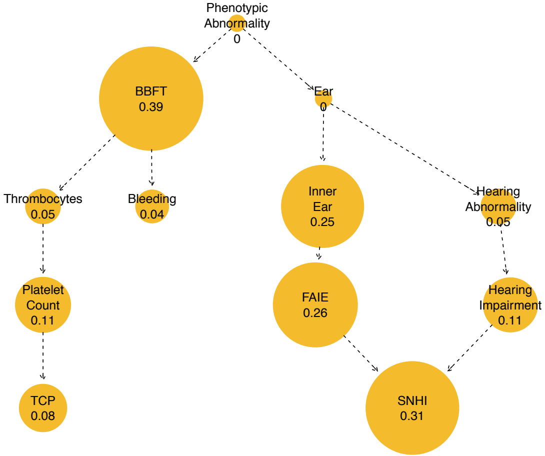
Reward shared terms

Penalise non-shared terms in ϕ

Allow for non-*expressivity* of terms

Penalise *flexibly* on each side to obtain good model fit

New cause of macroTCP: *DIAPH1* ($\mathbb{P}(\gamma = 1|y) = 0.81$)

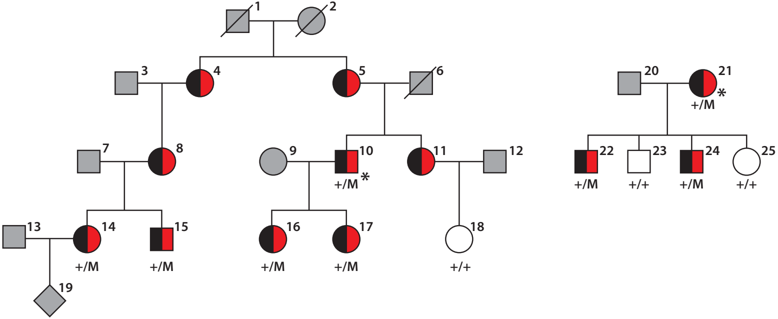


TCP: thrombocytopenia

SNHI: sensorineural hearing impairment

Stritt S*, Nurden P*, Turro E*, *et al.*, *Blood*, 2016 Jun; **127**(23)

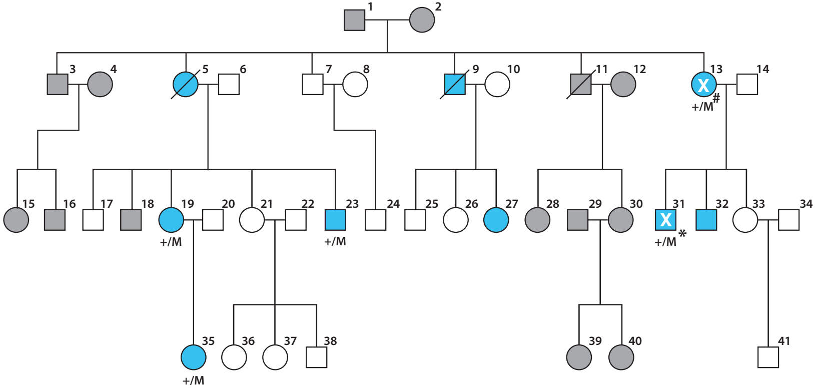
New cause of macroTCP: *DIAPH1* ($\mathbb{P}(\gamma = 1|y) = 0.81$)



Macrothrombocytopenia; hearing impairment.
Segregation ($p = 3.66 \times 10^{-4}$, conditional on the genotypes of the index cases)

Stritt S*, Nurden P*, Turro E*, *et al.*, *Blood*, 2016 Jun; **127**(23)

Variant prioritisation

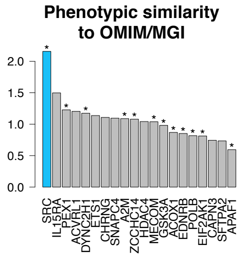


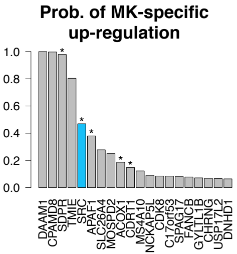
Myelofibrosis (early in life); tooth fractures; osteoporosis; macrothrombocytopenia; abnormal platelet granules; bleeding.

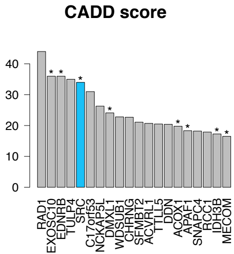
67 rare variants in 67 genes shared by cases 13 and 31.

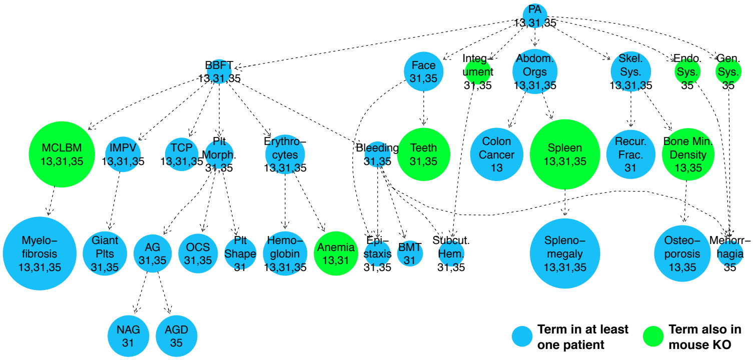
How best to prioritise these variants?

Variant prioritisation


Phenotypic similarity to OMIM/MGI


Prob. of MK-specific up-regulation

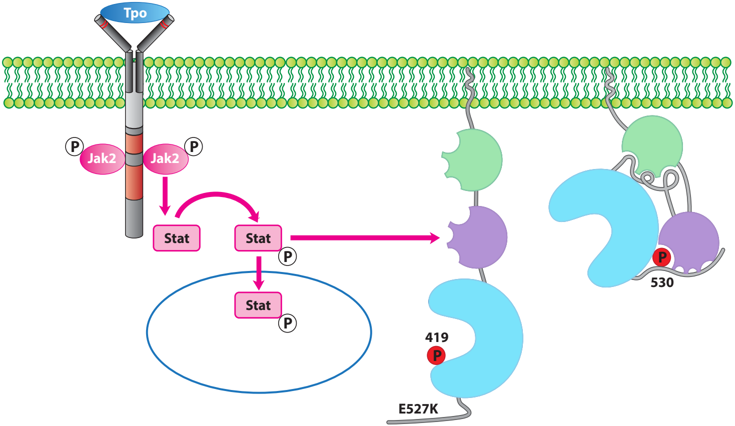

CADD score



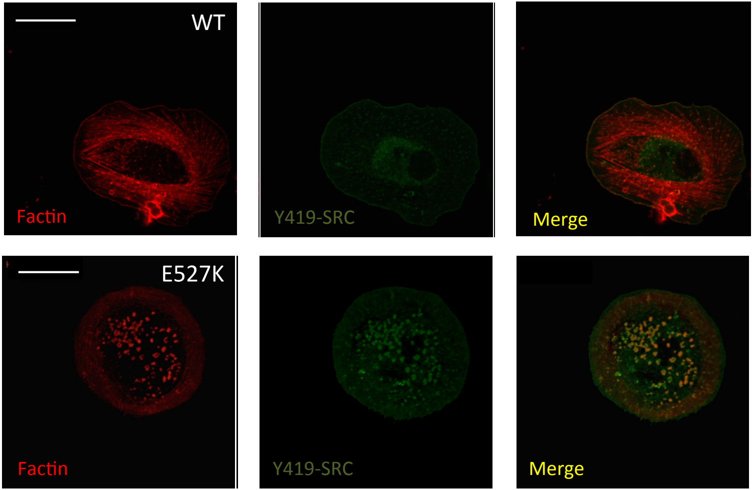
SRC

- First-ever discovered oncogene (Rous sarcoma virus, 1911)
- Src Family Kinase (SFK) inhibitors undergoing clinical trials for cancer treatment
- Thrombocytopenia and bleeding a frequent, unexplained side-effect of SFK inhibitors
- Mouse KO platelets normal
- No published germline pathogenic mutations

E527K places SRC in a constitutively active state

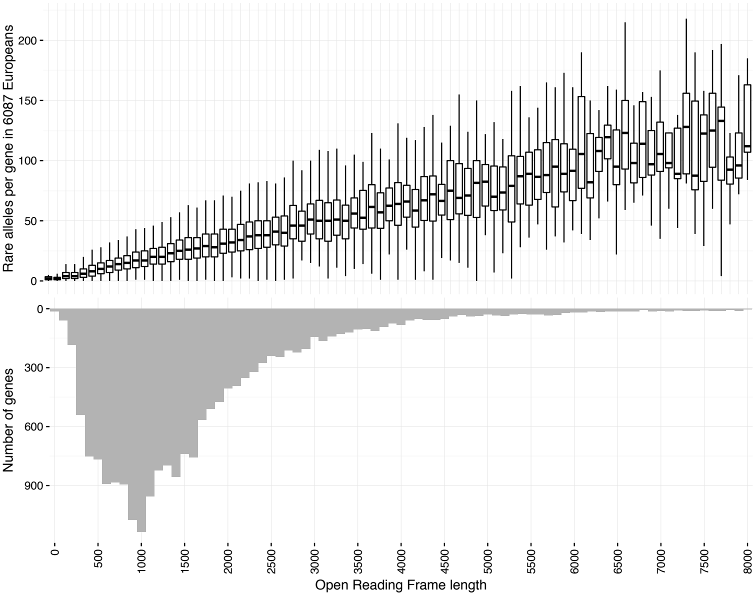


E527K results in enhanced podosome formation



Podosomes (actin-based protrusions on the plasma membrane) linked to osteoclast formation and cancer, and now platelet formation (Freson lab).

Most rare variants neutral with respect to severe disease



Motivation

- Most rare variant methods test association between genotypes and complex traits.
- Some can be adapted to work for rare Mendelian disease, however these either:
  - lose power by aggregating variants within genes.
  - don’t explicitly model a mixture of pathogenic and non-pathogenic variants
  - don’t model true Mendelian inheritance
  - tend to be very slow
- We address these issues in our method ‘BeviMed’.

No association

No assoc
Dominant
Recessive

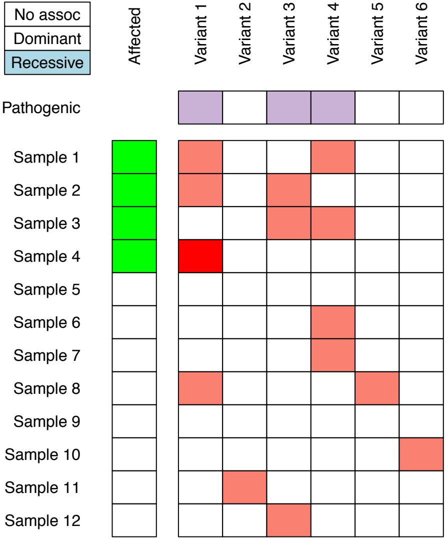
	Affected	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5	Variant 6
Pathogenic		▨	▨	▨	▨	▨	▨
Sample 1	🟩	🟧		🟧			
Sample 2	🟩						
Sample 3	🟩			🟧			
Sample 4	🟩				🟧		🟧
Sample 5						🟧	🟧
Sample 6						🟧	
Sample 7				🟧			🟧
Sample 8		🟧	🟧				🟧
Sample 9			🟧				
Sample 10							
Sample 11							🟧
Sample 12			🟥	🟧			

Dominant inheritance

No assoc
Dominant
Recessive

	Affected	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5	Variant 6
Pathogenic		🟪		🟪			
Sample 1	🟩	🟧					
Sample 2	🟩			🟧			
Sample 3	🟩			🟧			
Sample 4	🟩	🟧					
Sample 5							🟧
Sample 6			🟥				
Sample 7							
Sample 8							
Sample 9					🟧		
Sample 10							
Sample 11							
Sample 12						🟧	

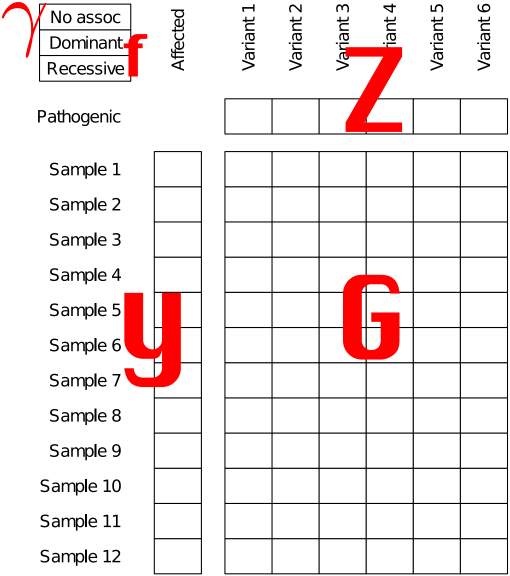
Recessive inheritance



No assoc
Dominant
Recessive

	Affected	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5	Variant 6
Pathogenic		🟪		🟪	🟪		
Sample 1	🟩	🟧			🟧		
Sample 2	🟩	🟧		🟧			
Sample 3	🟩			🟧	🟧		
Sample 4	🟩	🟥					
Sample 5							
Sample 6					🟧		
Sample 7					🟧		
Sample 8		🟧				🟧	
Sample 9							
Sample 10							🟧
Sample 11			🟧				
Sample 12				🟧			

Modelling


γ
No assoc
Dominant
Recessive
f
Affected
Variant 1
Variant 2
Variant 3
Variant 4
Variant 5
Variant 6
Z
Pathogenic
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
Sample 11
Sample 12
y
G

Modelling

Baseline model, $(\gamma = 0)$:

$$\mathbb{P}(y_i = 1) = q$$

Association model, $(\gamma = 1)$, with mode of inheritance f:

$$\mathbb{P}(y_i = 1) = \begin{cases} q & f(G_{i\cdot}, Z) = 0 \\ p & f(G_{i\cdot}, Z) = 1 \end{cases}$$

Priors

- Probability of association model

$$\gamma \sim \text{Bern}(0.01)$$

- Probability of dominant inheritance given association

$$\mathbb{P}(f = f_{dom} | \gamma = 1) \sim \text{Bern}(0.5)$$

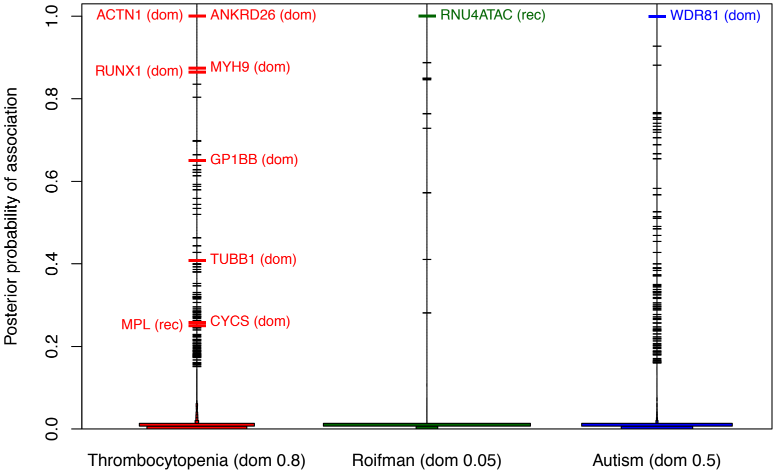
- Probability of pathogenicity for variant *i*

$$\mathbb{E}(Z_j = 1 | \gamma = 1, f = f_{dom}) = 0.25$$

$$\mathbb{E}(Z_j = 1 | \gamma = 1, f = f_{rec}) = 0.45$$

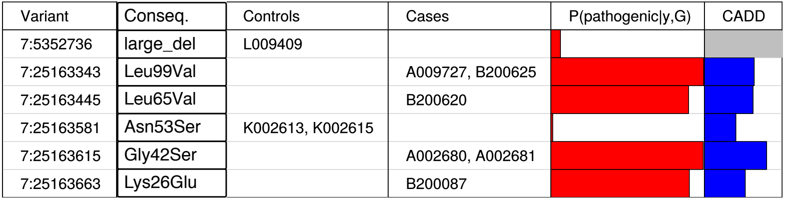
Demo

Results of gene-by-gene analysis



Mixture of pathogenic & non-pathogenic variants

- *CYCS* (variants enhance the intrinsic apoptotic pathway but causes only thrombocytopenia)
- Do we also observe an association between rare variants in *CYCS* and thrombocytopenia? (Yes, log BF $>$ 11)
- If so, which rare variants are likely pathogenic? Evidence supporting these four:

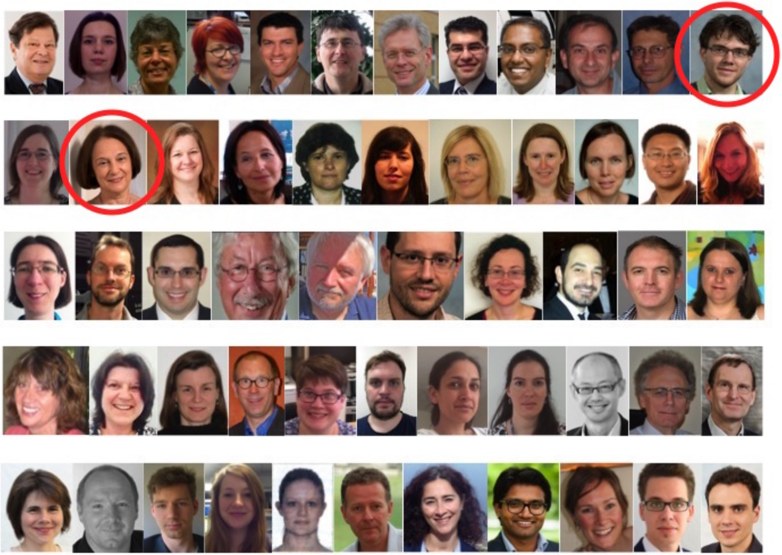


Variant	Conseq.	Controls	Cases	P(pathogenic|y,G)	CADD
7:5352736	large_del	L009409			
7:25163343	Leu99Val		A009727, B200625		
7:25163445	Leu65Val		B200620		
7:25163581	Asn53Ser	K002613, K002615			
7:25163615	Gly42Ser		A002680, A002681		
7:25163663	Lys26Glu		B200087		

Closing remarks

- Highly heterogeneous groups of patients
- Careful, detailed phenotyping using HPO
- Phenotype “similarity”-based methods for prioritising variants and association testing
- Genetic heterogeneity with many neutral variants
- Methods that model explicit mixture of variant types under Mendelian inheritance
- Data from cell type specific genomic assays to focus search in non-coding regions of the genome

Acknowledgements



Acknowledgements

UNIVERSITY OF CAMBRIDGE

British Heart Foundation

BRIDGE

MRC Medical Research Council

NIHR BioResource
Rare Diseases

NHS
National Institute for
Health Research

KU LEUVEN

wellcometrust

NHS
Blood and Transplant

The Children's Hospital of Philadelphia

CRPP

University of BRISTOL

UNIVERSITY OF UTAH SCHOOL OF MEDICINE

Universitätsmedizin GREIFSWALD

Imperial College London

THE UNIVERSITY OF WESTERN AUSTRALIA

MRC Biostatistics Unit

UNIVERSITY OF LIVERPOOL

UMC Utrecht

Beth Israel Deaconess Medical Center
A teaching hospital of Harvard Medical School

Royal Free London NHS
NHS Foundation Trust

CHARITÉ

EMBL-EBI

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Salisbury NHS
NHS Foundation Trust

Hampshire Hospitals NHS
NHS Foundation Trust

Maastricht UMC+
azM | Maastricht University

UK

Rudolf Virchow Zentrum
DFG Research Center for Experimental Biomedicine
University of Würzburg

University Hospital Southampton NHS
NHS Foundation Trust

southern haemophilia network

ThromboGenomics

Oxford University Hospitals NHS
NHS Trust

KENT HAEMOPHILIA CENTRE
A COMPREHENSIVE CARE CENTRE

East Kent Hospitals University NHS
NHS Trust

Cambridge University Hospitals NHS
NHS Foundation Trust

GIG CYMRU NHS WALES
Bwrdd Iechyd Prifysgol
Abertawe Bro Morgannwg
University Health Board

Great Ormond Street NHS
Hospital for Children
NHS Trust

Central Manchester NHS
University Hospitals
NHS Foundation Trust

Hôpitaux Universitaires Est Parisien
TROUSSEAU
LA ROCHE-GUYON

CATGO
CAmbridge Translational GenOmics Laboratory

NHS
Guy's and St Thomas'
NHS Foundation Trust

The Newcastle upon Tyne Hospitals NHS
NHS Foundation Trust

Barts Health NHS
NHS Trust

References

1. Westbury S*, Turro E*, Greene D*, Kelly AM*, Lentaigne C*, Bariana T*, Simeoni I, Pillois X, . . . , Furie B, Robinson PN, Van Geet C, Rendon A, Gomez K, Laffan M, Lambert M, Nurden P, Ouwehand WH§, Richardson S§, Mumford AD§, Freson K§ on behalf of the BRIDGE-BPD Consortium.
   Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders.
   Genome Medicine, 2015; **7**:36.

2. Greene D, NIHR BioResource, Richardson S*, Turro E*.
   Phenotype similarity regression for identifying the genetic determinants of rare diseases.
   American Journal of Human Genetics, 2016 Mar; **98**:1–10.

3. Stritt S*, Nurden P*, Turro E*, Greene D, Jansen SBG, Westbury SK, Petersen R, Astle WJ, Marlin S, Bariana TK, Kostadima M, Lentaigne C, Maiwald S, Papadia S, Kelly AM, Stephens JC, Penkett CJ, . . . , BRIDGE-BPD Consortium, Gomez K, Erber WN, Stirrups K, Rendon A, Bradley JR, Van Geet C, Raymond FL, Laffan MA, Nurden A, Nieswandt B, Richardson S, Freson K§, Ouwehand WH§, Mumford A§.
   A gain-of-function variant in *DIAPH1* causes dominant macrothrombocytopenia and hearing loss.
   Blood, 2016; **127**(23) 2903–2914.

4. Turro E, Greene D, Wijgaerts A, Thys C, Lentaigne C, Bariana TK, Westbury SK, Kelly AM, Selleslag D, Stephens JC, Papadia S, Simeoni I, Penkett C, . . . , BRIDGE-BPD Consortium, De Maeyer M, Rendon A, Gomez K, Erber WN, Mumford AD, Nurden P, Stirrups K, Bradley J, Raymond FL, Laffan MA, Van Geet C, Richardson S, Freson K*, Ouwehand WH*.
   A dominant gain-of-function mutation in universal tyrosine kinase *SRC* causes thrombocytopenia, myelofibrosis, bleeding and bone pathologies
   Science Translational Medicine, 2016 Mar; **8**:328.

5. Greene D. BeviMed and SimReg R packages: https://cran.r-project.org/web/packages/SimReg, https://cran.r-project.org/web/packages/BeviMed.