ASHG 2014 meeting highlights: From Bytes to Phenotypes

Mohsen Hosseini
Session’s theme

• Phenome
  – The set of all phenotypes in an organism originally used by Davis in 1949.

• PheWAS
  – Analyzing many phenotypes compared to a single genetic variant, originally described in the context of electronic medical records (EMR)
  – Using ICD-9 codes
  – Exploring other ways to do PheWAS
1. Investigation of Synthetic Association (SA) in GWAS using PheWAS and Exome Sequencing

- Causality of GWAS findings is unknown
- SA Hypothesis: one (or more) rare causal alleles on a haplotype containing common neutral allele explain the GWAS findings
Simplified view of genetic variation at the NOD2 locus, a well-documented example of a synthetic association.

Simplified view of genetic variation at the NOD2 locus, a well-documented example of a synthetic association. The left-hand side shows a genealogical tree representing six SNPs in this region after discarding rare recombinant events. The right-hand side shows the resulting haplotypes and their population frequencies (48), with coloured circles representing common GWAS SNPs, and starbursts representing previously identified low-frequency coding variants responsible for association between NOD2 and CD. While none of the GWAS SNPs is strongly correlated with any individual causal allele, the three coding variants create a synthetic association because they cluster by chance on the side of the tree marked by the green GWAS SNP (rs2076756).

Synthetic association (cont.)

- 29.7k European individuals genotyped on Illumina HumanExome in a PheWAS
  - Genome-wide significant SNPs in GWAS catalogue
  - Mapped the trait to PheWAS phenotype (n=104)
  - Tested all missense and nonsense SNPs with 1% < MAF < 5% in the gene reported in NHGRI catalogue (1743 genotype-phenotype pairs)
- replicated 66% of NHGRI signals that they had power for:
  - 84 assoc with p < 0.01
  - 59 with OR > 2 or < 0.5 for 38 unique phenotypes
  - 22 assoc. passing Bonferroni correction
- Conditional GWAS to see whether common could be explain by rare.
- Only replicated NOD2 and Crohn’s example.
2. Problem of circularity in deleteriousness predictions of missense variants

- many *in silico* predictors of deleteriousness
- Circularity:
  - Overlapping training and evaluation sets (often not public)
  - Gene-level confounding: most SNPs belong to pure proteins and are annotated the same way
- One cannot properly compare prediction tools
- Polyphen2 seems to be superior to others especially if retrained on Varibench.
3. Application of Clinical Text Data in PheWAS

• Majority of PheWAS used ICD9 diagnostic codes to define case-control status
  – Primarily for billing
  – Limited phenotype granularity
  – Do not allow for other clinically relevant information

• Alt: text-based phenome
Application of Clinical Text Data in PheWAS

• Text based phenome from Marshfield Clinic EMR:
  – 4200 patients
  – 1.5 M clinical notes
  – 423 M words
  – 23k clinically relevant word medical dictionary
  – Five SNPs genotyped
Application of Clinical Text Data in PheWAS

• Performed equally well with ICD9
• AMD SNP
  – Macular degeneration
  – Nonexudative (type)
  – Exudative (type)
  – Visudyne (drug)
• Hashimoto’s thyroiditis
  – ICD9 non-significant because term is specific
  – Text based: Hashimoto (1E-12)
4. Warped Linear Mixed Models

- Linear mixed models commonly used in genetics
  - GWAS
  - Heritability estimation
  - Phenotype prediction
- Fundamental assumption is normality of noise (error)
- If violated:
  - Power loss
  - Biased estimate
  - Reduced accuracy
- Current solution:
  application of appropriate transformation pre-processing, which may be challenging (manual, distribution of residual and not phenotype)
- *Fusi et al, Nature communications, 2014*
Warped Linear Mixed Models (WLMM)

- Automatically learns the suitable phenotype transformation
- Tests “infinite” number of transformations to find the best
- Increases power, reduced bias, increases accuracy
- Permits back transformation to the original scale unlike rank based transformations
Summary of method

• $y_n$: observed non-normal phenotype; $z_n$: unobserved normal distributed phenotype; $f$: transformation function

$$z_n = f(y_n; \psi) .$$  \hspace{1cm} (1)

$$z_n = x_n \beta + u_n^g + \epsilon_n$$  \hspace{1cm} (3)

• Likelihood for vector $z$ for $N$ individuals:

$$z \sim N \left( X\beta, \sigma_g^2 K + \sigma_e^2 I \right) .$$  \hspace{1cm} (4)

• $K$ is the genetic relatedness matrix for $S$ SNPs

• WLMM identifies the most probable transformation ($\mathcal{f}$) by maximizing likelihood (4) using a warping function proposed by Snelson et al.
Simulation of heritability estimate

(a) Simulated heritability vs. $\hat{h}^2 - h^2$

(b) Number of causal variants vs. $\hat{h}^2 - h^2$

(c) Number of samples vs. $\hat{h}^2 - h^2$

(d) Linearity of transformation vs. $\hat{h}^2 - h^2$
Heritability study in mouse

- 58 phenotypes in mice, manually transformed
- Compared heritability estimates in the original paper using LMM with WLMM.
- 18 of 47 phenotypes the two estimates were significantly different and in 17/18 of these WLMM yielded higher heritability estimates
- Because no gold standard: used training and validation in 90% vs 10% of data which WLMM was superior
(a) Heritability estimates using a LMM on the untransformed phenotype versus the heritability estimates obtained by WarpedLMM. Empirical error bars were obtained from ten bootstrap replicates, using 90% of the data in each replicate. Significant differences are coloured in red (paired $t$-test, $\alpha=0.05$). (b) Out-of-sample prediction accuracy assessed by the squared correlation coefficient $r^2$, considering either a LMM on the untransformed data or a WarpedLMM. Prediction accuracies were assessed from ten random train-test splits. Phenotypes with significant deviations in prediction accuracy of the LMM and the WarpedLMM are highlighted in red (paired $t$-test, $P$-value$\leq0.05$).
WLMM for GWAS

• Analyzed for metabolic traits
  – HDL (linear vs WLMM)
  – LDL (linear vs WLMM)
  – TG (log transformed vs WLMM)
  – CRP (log transformed vs WLMM)

• Overall WLMM increased GWAS power

• An implementation of WarpedLMM in python is available at
  http://github.com/pmbio/warpedLMM.
(a) The GWAS results for C-reactive protein, and (b) the GWAS results for low-density lipoprotein. Red circles denote significant associations ($\alpha<5 \times 10^{-8}$, marked on the plots with a dashed line). The two rightmost panels show an enlarged view of interesting regions in chromosomes 1 and 19, with black arrows highlighting loci that were identified only when using WarpedLMM.