Rapporteurs: Melpomeni Kasapi and Taylorlyn Stephan, NHGRI

Missing Heritability, Ten Years On
May 1-2, 2018 - Silver Spring, MD
Executive Summary

Ten years ago the first Missing Heritability workshop was convened to examine the problem of missing heritability (h^2) in complex diseases and propose research strategies to identify its potential sources. Participants from the first workshop and other experts recently reconvened to review progress in missing h^2 and assess what has been learned since 2009, what has been and/or will be the value of identifying sources of missing h^2, and what research can be pursued to determine these sources.

Participants met to discuss issues in quantifying missing h^2, effects of rare variants and peripheral genes on gene expression, difficulties in complex traits and diseases, lessons learned from non-European populations, and early examples of missing h^2 in the clinic. The group agreed that large family studies, very large biobanks, and increased availability of complex data sets will expedite novel discoveries and facilitate bridging the gaps. This executive summary includes a brief description of the lessons learned and future directions for research to fully address the issue of missing heritability.

Lessons learned since 2009

Quantifying missing heritability

- Using all genotyped SNPs, irrespective of statistical significance, explains much more h^2 than genome-wide significant (GWS) SNPs alone: height 45% vs. 5%; LDL 80% vs. 20%.
- Imputation to now-available large sequenced reference samples has increased SNP-h^2 estimates further (height: 60%) and diminished the difference with family-based h^2 for quantitative traits.
- For disease traits, the difference between pedigree-h^2 and SNP-h^2 estimates is bigger than for quantitative traits despite the same genotyping and imputation technologies.
- There are few examples of dominance variance (1-3%) estimated in quantitative traits.
- Much of the genetic variance is captured by arrays.
- The few epistatic interactions that are currently seen are with very large effect loci like MHC.
- Lifetime risk has major impact on h^2 estimates, but we rarely know lifetime risks.
- There is value in having widely accessible datasets on vast numbers of people; imputation and identity by descent estimates are more powerful as datasets expand.
- Theoretically, epistasis could be pervasive and additive models would still fit the data. Estimates of h^2 and effect size can be biased upwards.

Missing heritability in the clinic

- Clinical diagnostic sequencing has proven to have high diagnostic yields, even with imprecise phenotypic characterization and not knowing *a priori* which gene to investigate.
- Sequencing in complex diseases identifies significant numbers of monogenic conditions which alter treatment. Examples include undetected Alport's and Wilson's diseases, MODY.
- The community had expected to find SNPs in cancer pathways that affect multiple cancer types, but to date 90% of the SNPs or even loci in cancer are not seen in another cancer type.
- Polygenic risk scores can identify 10-fold differences in risk; this will soon be useful clinically.

Rare variants, structural variants, and gene expression

- The “common vs rare variant” debate has largely disappeared.
- Most phenotypic variance is due to regulatory variation in genes expressed in the “right” tissues but without known roles in pathogenesis. The reasons behind this remains to be defined.
- SVs across the genome make very small contributions, because of the relatively few associations (~1000); however, at individual loci there is 2-4X more variance explained than by lead SNPs.
- Acquired mutations may contribute to h^2, but it is unclear whether vulnerability to mutation can be inherited, nor how important this is in non-hematologic tissue.
- In Mendelian disease peripheral genes outnumber core genes 100:1 (or more). For common disease, individual effects are very small, which may explain why a huge fraction of the genome contributes to a single trait.
- It is rare that a point mutation outside a gene will have strong effects on gene expression because of the built-in redundancy, but this remains to be proven.
- Regarding genetics of gene expression, there are large catalogs of *cis*-eQTLs, more diverse contexts, variants, and phenotypes available now for studying missing h^2 and complex traits.
- Rare variants have been shown to drive extreme expression levels in individuals; if confirmed, in aggregate this could explain a large proportion of h^2 of expression.
- There are more than 8,000 mosaic segmental mutations in at least 1% frequency in 150,000 UK Biobank participants. These cluster in genomic hotspots like fragile sites.
- *Trans* effects have important contributions to understanding disease heritability, but *trans*-eQTLs remain hard to identify.

Environment affects missing heritability

- There is genetic variation in sexual dimorphism as a context-dependent effect; we see massive gene-sex and gene-environment interactions in flies.
- Much of the interaction is antagonistic, which may explain the small effect sizes in flies. Whether this is similar in humans remains to be defined.
- True gene-environment and environment-environment interactions are rare, partly due to the need for large studies and accurate classification of exposure to detect them.
- Even if there’s little true multiplicative interaction, the increase in absolute numbers of cases attributed to “non-genetic” RF such as smoking and obesity at high genetic risk is much greater than at low genetic risk (“absolute risk”).
- Risks seem to multiply without synergism, so searching for multiplicative interactions is unlikely to improve prediction, which is good for the risk prediction algorithms.
- In Crohn’s disease, transcriptomics are more predictive of disease course than genetics, which could be due to environmental contributions or differences among subphenotypes, such as severity and treatment responses. Other diseases with this pattern should be defined.

Learning from diverse populations

- There is benefit of adding 50,000 non-European participants to large consortia such as GIANT.
- Integrated analysis across diverse populations is more powerful than stratified analysis.

- Controlling for global ancestry does not remove the effect of local ancestry, which can be controlled for by using a chromosomal segment as unit of analysis in admixture mapping.
- Specific populations are starting to reveal strong effect alleles and founder effects.

Genetic architecture of complex traits

- Family studies are valuable for identifying causal *de novo* mutations (and reducing false positives and negatives) and detection of shared genomic segments that contain disease-causing variants.
- Heterogeneity in Mendelian conditions is extensive, both in alleles and loci. “Multi-Mendels” occur in 3-5% of cases receiving a molecular diagnosis from exome sequencing.
- Two-locus models can explain incomplete penetrance, as in craniosynostosis.
- PheWAS have provided new insights since 2009, such as co-segregation of IBD with disorders of phosphorus metabolism.
- We are beginning to explain pleiotropic surprises like the *LRKK2* kinase domain variants in Crohn’s and Parkinson’s.

Future Directions

New or enhanced analyses

- Estimate genetic variation using large (>50K) WGS samples.
- Estimate variance due to non-SNP variation.
- Study the X chromosome, mitochondrial DNA, and potentially the Y chromosome.
- Generate true burden analyses rather than merely collapsing point mutations.
- Analyze large biobank genetic data systematically.

New or enhanced studies

- Enhance domain-specific sequence annotation.
- Explore *trans*-regulatory networks further and how they behave when perturbed.
- Explore VNTRs and clonal expansion of mosaic sites as they may be whole new frontier.
- Scale up expression studies overall with larger sample sizes, single cell analyses, and integrated analyses connecting epigenetic and expression data and GWAS to fill in mechanism further.
- Study expression during development and “de-development” (in cancer).
- Study large numbers of families to dissect within vs. between family effects.
- Study the genetics of disease progression/severity.
- Leverage phenotype risk scores to find hidden Mendelians and characterize phenotypic variation associated with genes that have not been studied yet.
- Test specifically core vs. peripheral regulation, conditioning on a set of core genes.
- Include African and other non-European populations and leverage subpopulation differences.
- Generate more explicit modeling with infectious agents.

Enormous progress has been made in the past decade. Heritability appears to have been over-estimated using traditional methods, leading to more apparent missing h^2 than is probably the case. GWAS as an experimental design is no longer questioned and has been highly successful in explaining h^2. Substantial proportions of h^2 are now captured from known variants, and nearly all traits appear to be polygenic.