Genetic Discovery in the 23andMe Participant Cohort

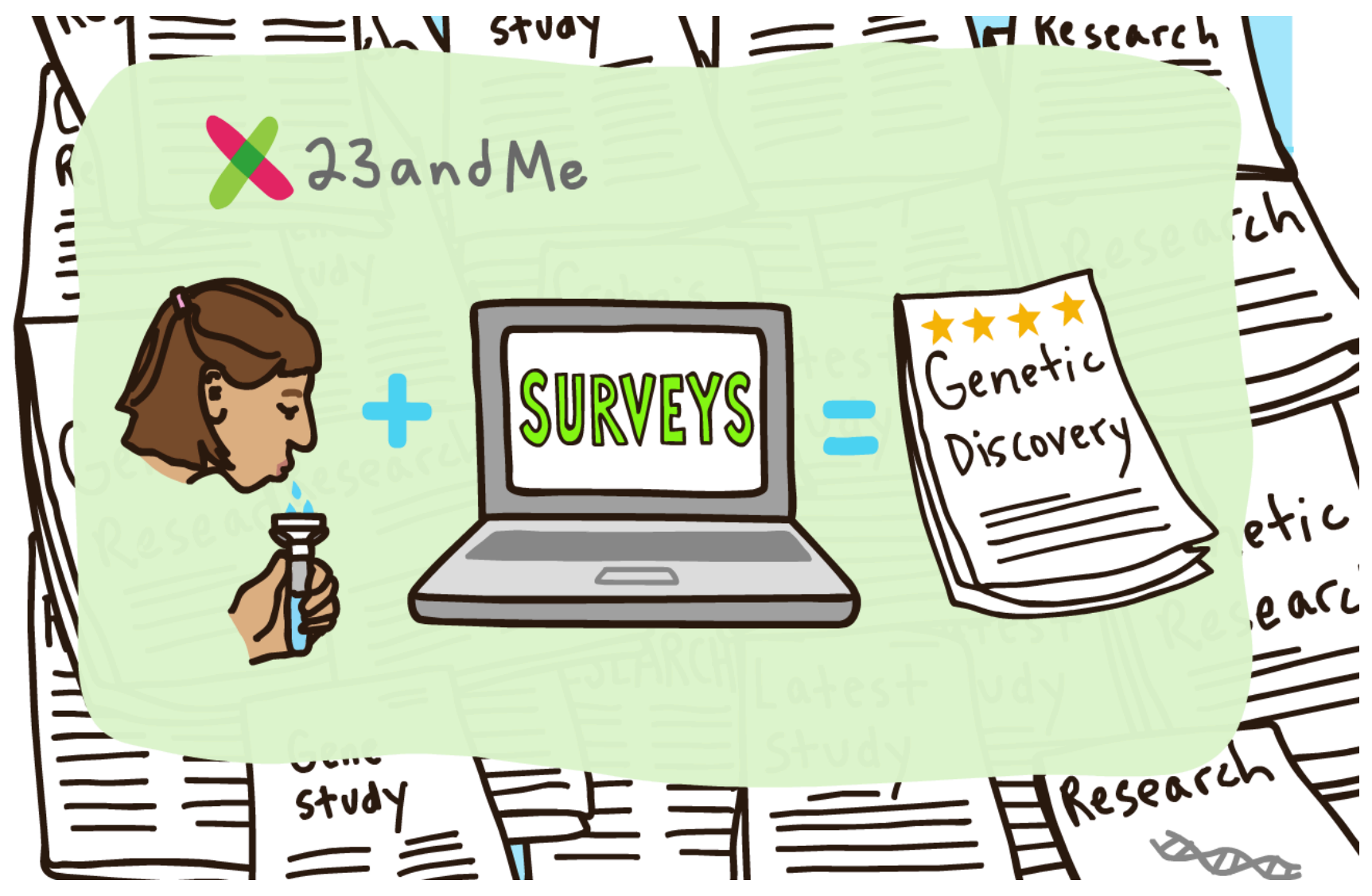
David A. Hinds[1], Carrie A.M. Northover[1], Matthew H. McIntyre[1], Catherine Wilson[1], Karen E. Huber[1], Aaron Kleinman[1], Fah Sathirapongsasuti[1], Robert K. Bell[1], Emma Pierson[1], Katarzyna Bryc[1], Alena S. Shmygelska[1], Nicholas A. Furlotte[1], Youna Hu[1], Chao Tian[1], Eric Y. Durand[1], Cory Y. McLean[1], Brian T. Naughton[1], Joanna L. Mountain[1], Nicholas Eriksson[1], and Joyce Y. Tung[1]

[1]23andMe, Inc, Mountain View, CA.

Introduction

The 23andMe participant cohort now includes more than 600,000 genotyped subjects who have consented to participate in research, including more than 135,000 individuals with substantial non-European ancestry. Most participants provide phenotypic information through web based surveys covering a wide range of topics. The scale of this cohort has required us to develop specialized solutions for many tasks where conventional approaches have inconvenient scaling properties. These include methods for ancestry inference and detection of identity by descent; phasing and imputation; and genome-wide association testing. While most of our data is collected on the sample of convenience of 23andMe customers, we have recruited cohorts in several disease areas, including Parkinson's disease, myeloproliferative neoplasms, and inflammatory bowel disease.

Methods



Phenotypic data collection

- Web based surveys covering a very wide range of topics:
  - Basic health profile covering >100 conditions
  - Detailed surveys covering specific topics in depth
  - Research snippets: stand-alone single questions

Genotyping platforms

- V1,V2 platforms: humanhap550 + custom (OMIM, HGMD, PharmGKB, HLA, Y+Mito), 580,000 working SNPs
- V3 platform: OmniExpress + custom, plus V2 compatibility: 960,000 working SNPs
- V4 platform: full custom, ~500K SNPs chosen for V2/V3 continuity with less redundancy; more custom content: rare coding variation

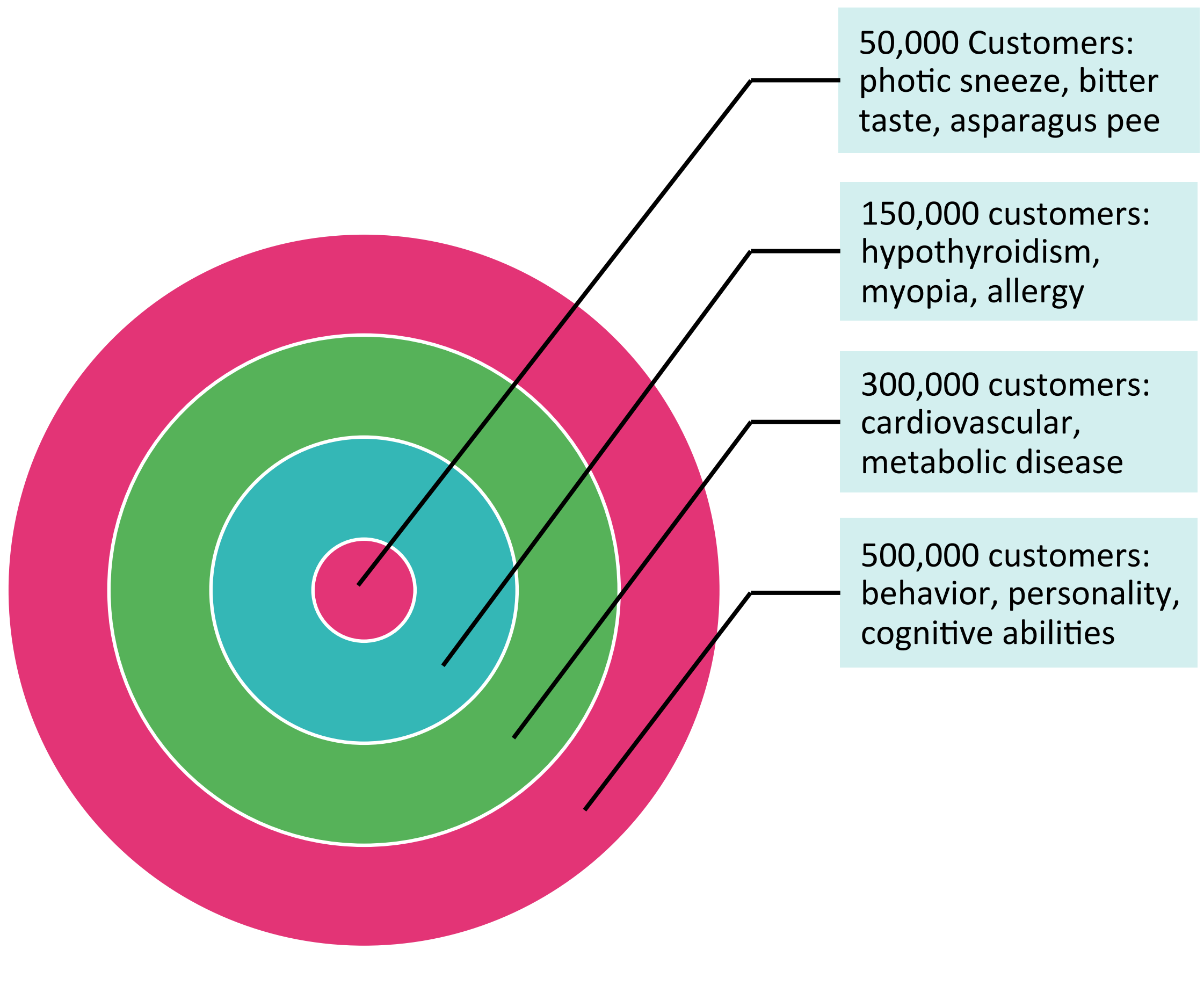
Imputation pipeline

- Out-of-sample implementation of Beagle haplotype-graph-based phasing to enable fast pre-phasing without batch effects
- Imputation against March 2012 release of 1000 Genomes reference haplotypes, all populations, pre-filtered to ~13M imputable SNPs
- Performance improvements for minimac:
  - Reorganize data to improve memory access patterns
  - Use high performance BLAS library where possible
  - Rewrite all inner loops to be vectorizable by gcc
  - Net performance improvement: ~20x faster

GWAS pipeline

- Multithreaded, distributed multi-phenotype GWAS pipeline
  - Each cluster node processes a slice of the genome
  - Phenotypes distributed across threads within each node
  - IO is overlapped with computation
- High performance logistic regression implementation:
  - Intel MKL vectorized math library, optimized BLAS

The Power of Large Numbers



50,000 Customers: photic sneeze, bitter taste, asparagus pee

150,000 customers: hypothyroidism, myopia, allergy

300,000 customers: cardiovascular, metabolic disease

500,000 customers: behavior, personality, cognitive abilities

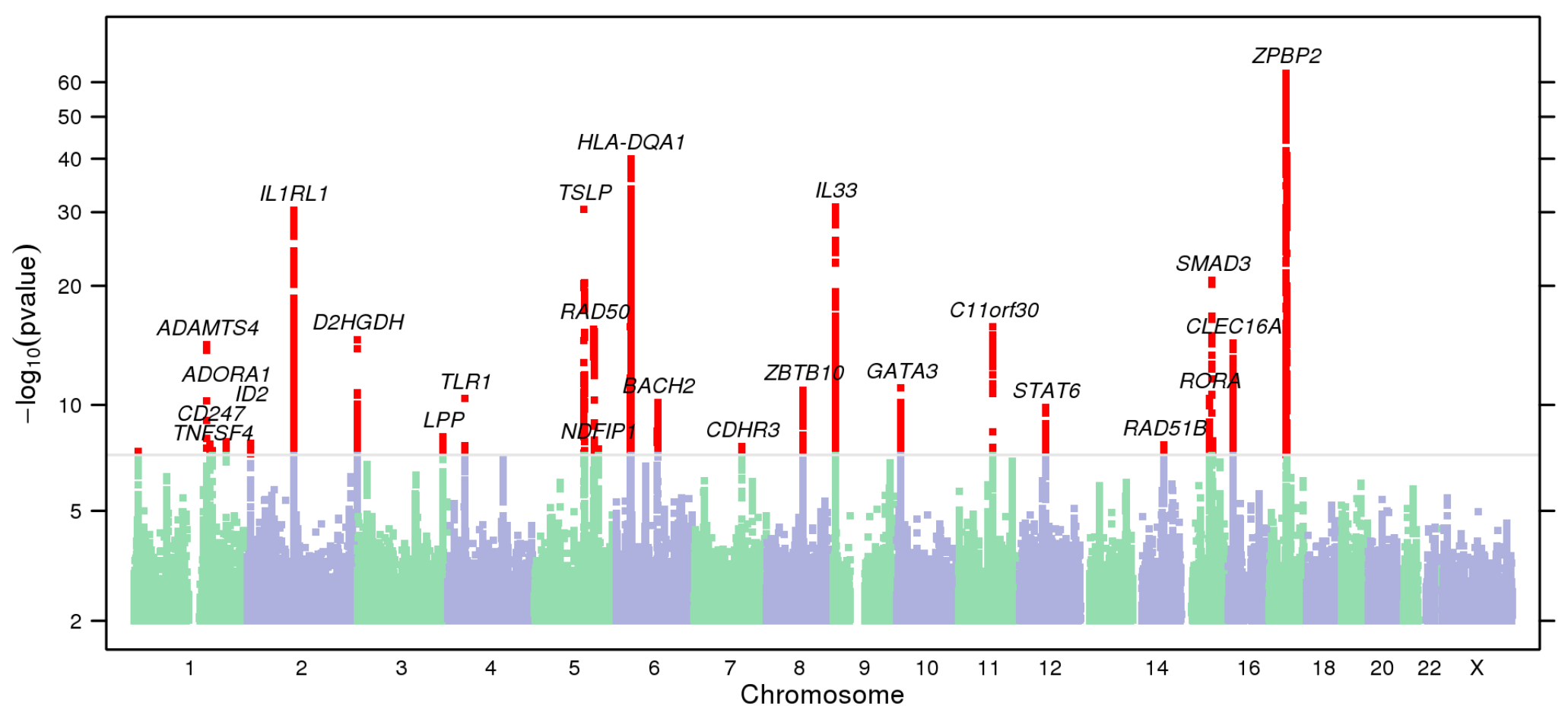


Figure 1. Asthma: 28399 cases, 128843 controls, 26 associations

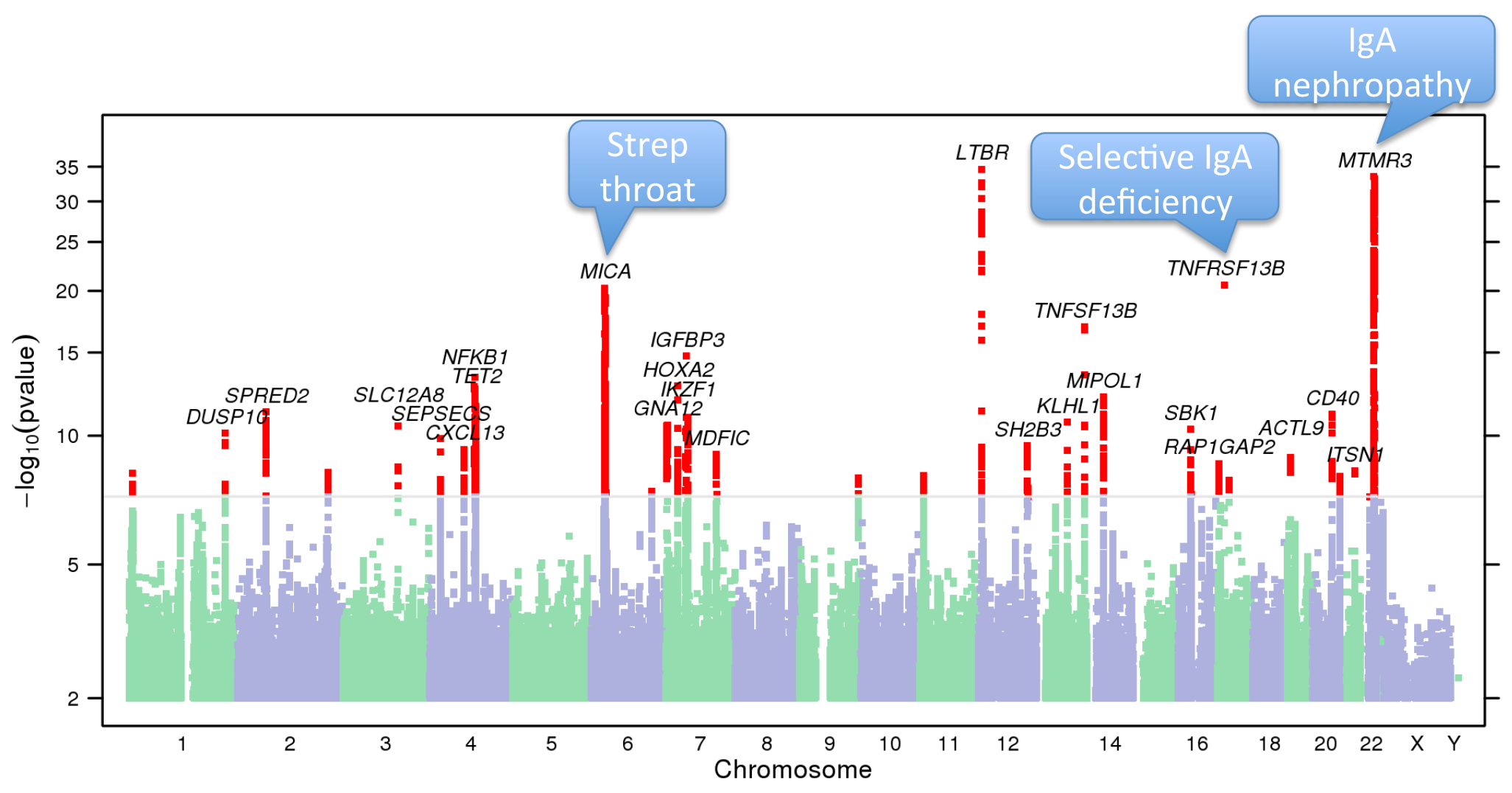


Figure 2. Tonsillectomy: 60098 cases, 113323 controls, 36 associations

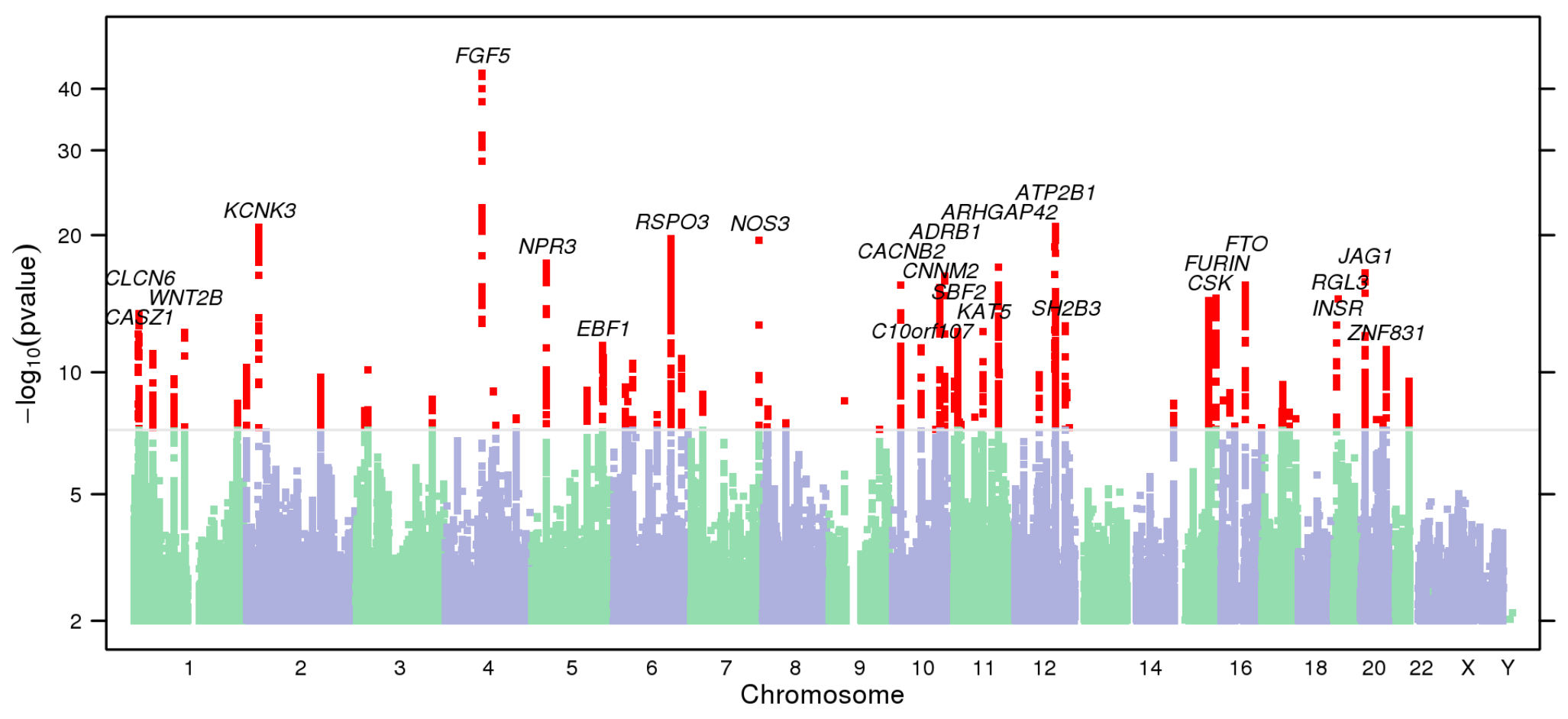


Figure 3. High Blood Pressure: 67844 cases, 188884 controls, 70 associations

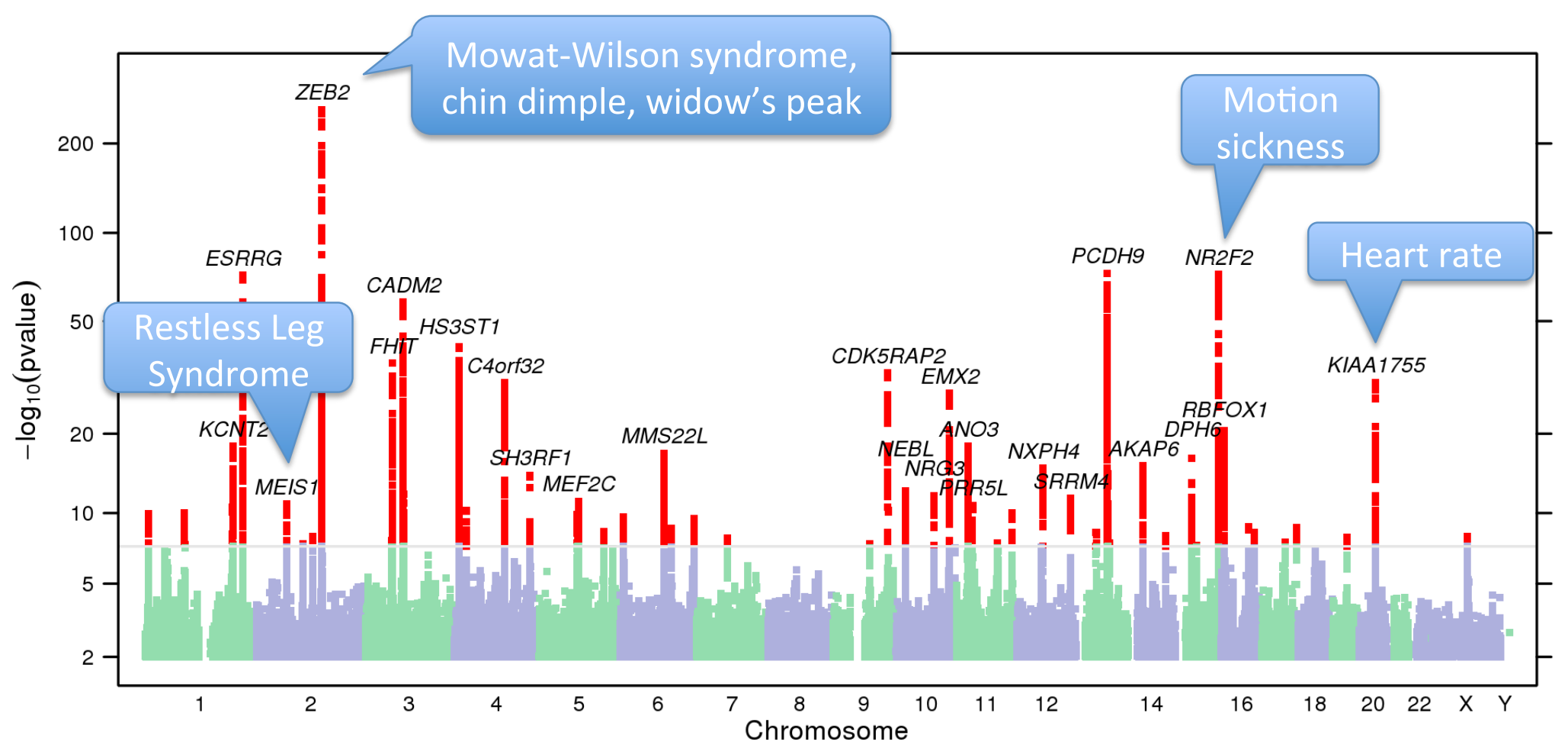


Figure 4. Photic Sneeze: 32446 cases, 67249 controls, 51 associations

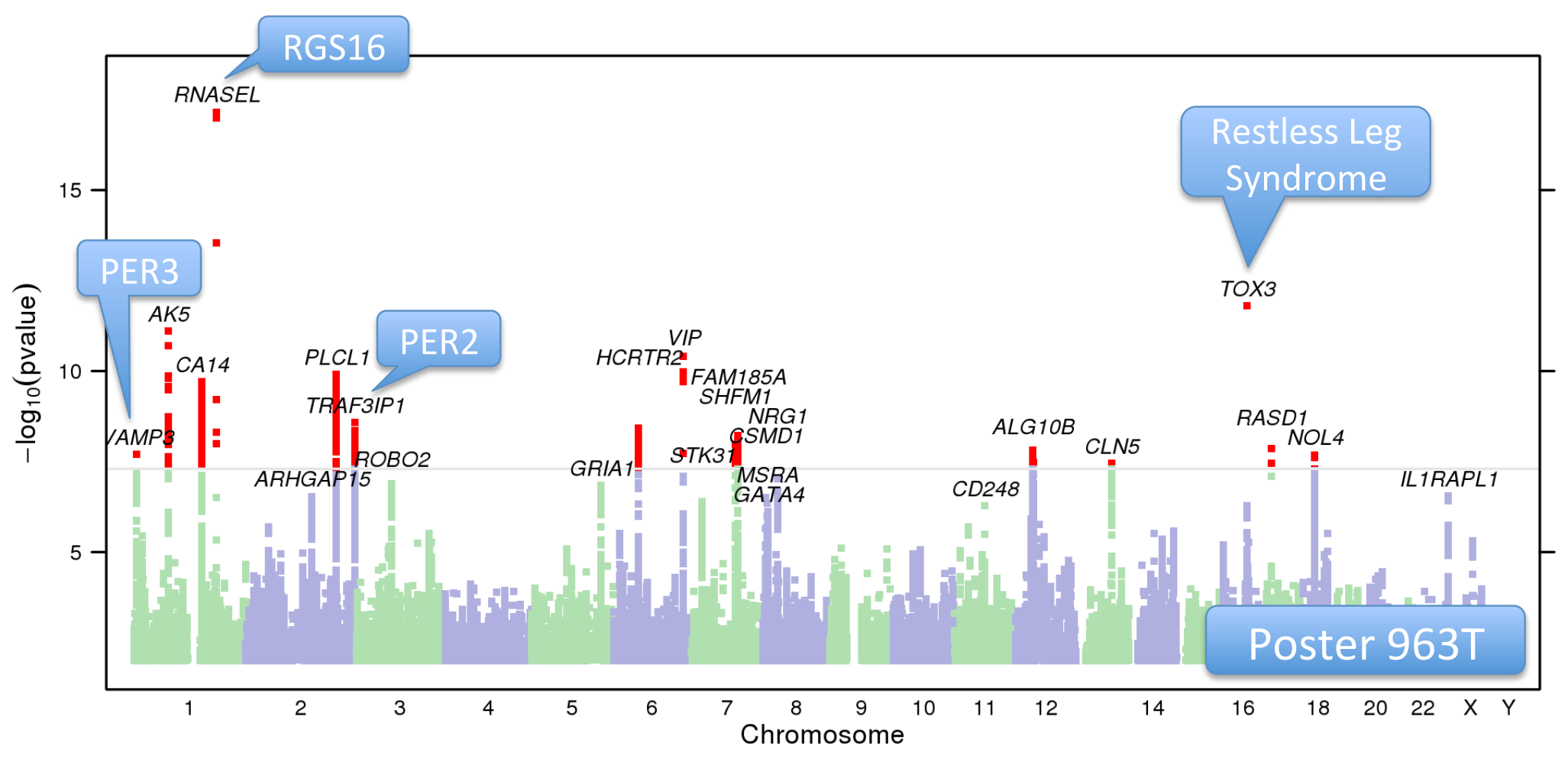


Figure 5. Morning Person: 38937 cases, 50346 controls, 15 associations

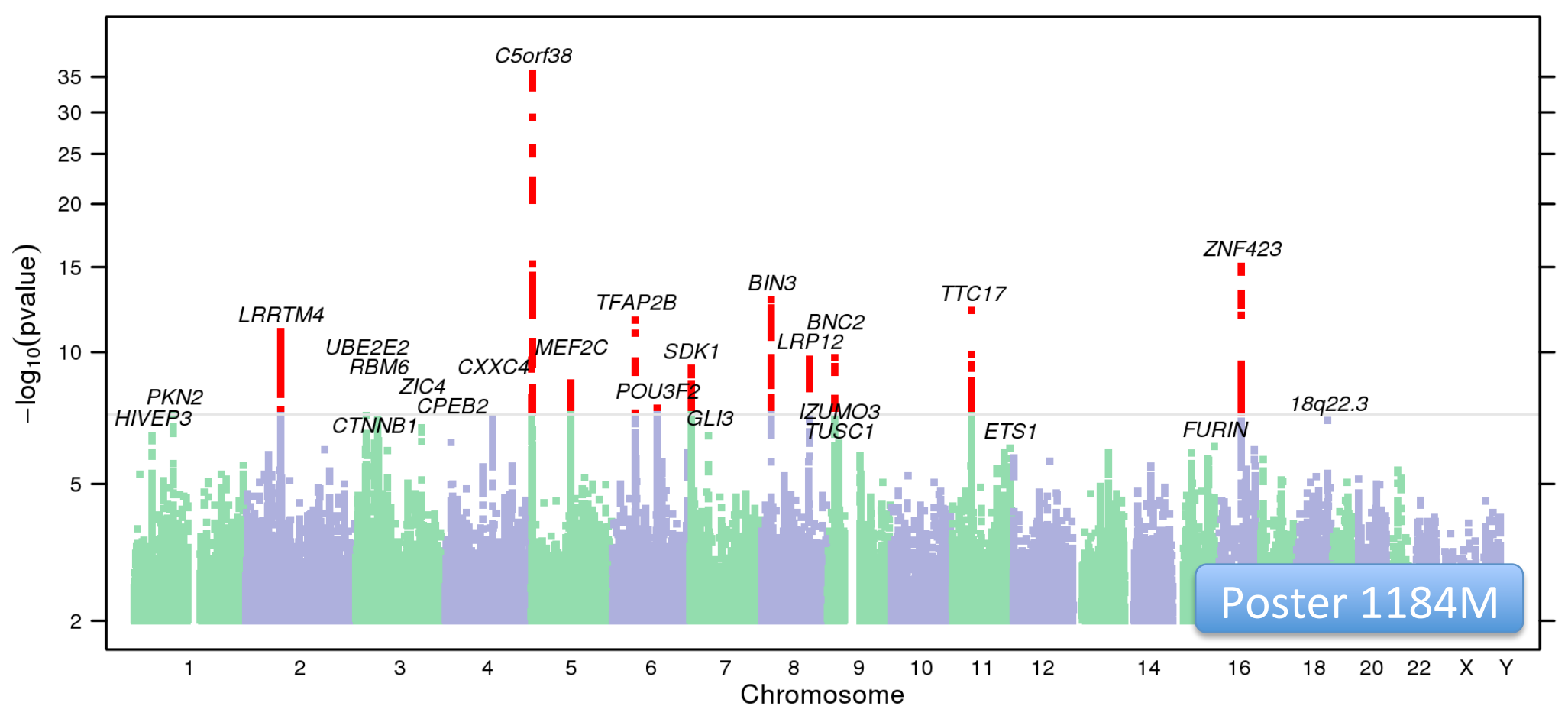


Figure 6. "Do you cry easily?": 72841 cases, 112368 controls, 12 associations

Results

- In nearly every phenotypic category, we find examples of phenotypes that seem to be particularly amenable to GWAS (Fig. 1 to 6). This is probably due to a combination of genetic architecture and ease of web-based self report.
- At a high level, the genetic architecture of many behavioral and cognitive traits does not appear to be fundamentally different from that of other complex traits, and we should expect robust GWAS results from well powered studies.
- We find numerous examples of pleiotropy across traits and in some cases across unexpected phenotypic categories. Sometimes this can be used to more precisely map association signals (Fig. 7).

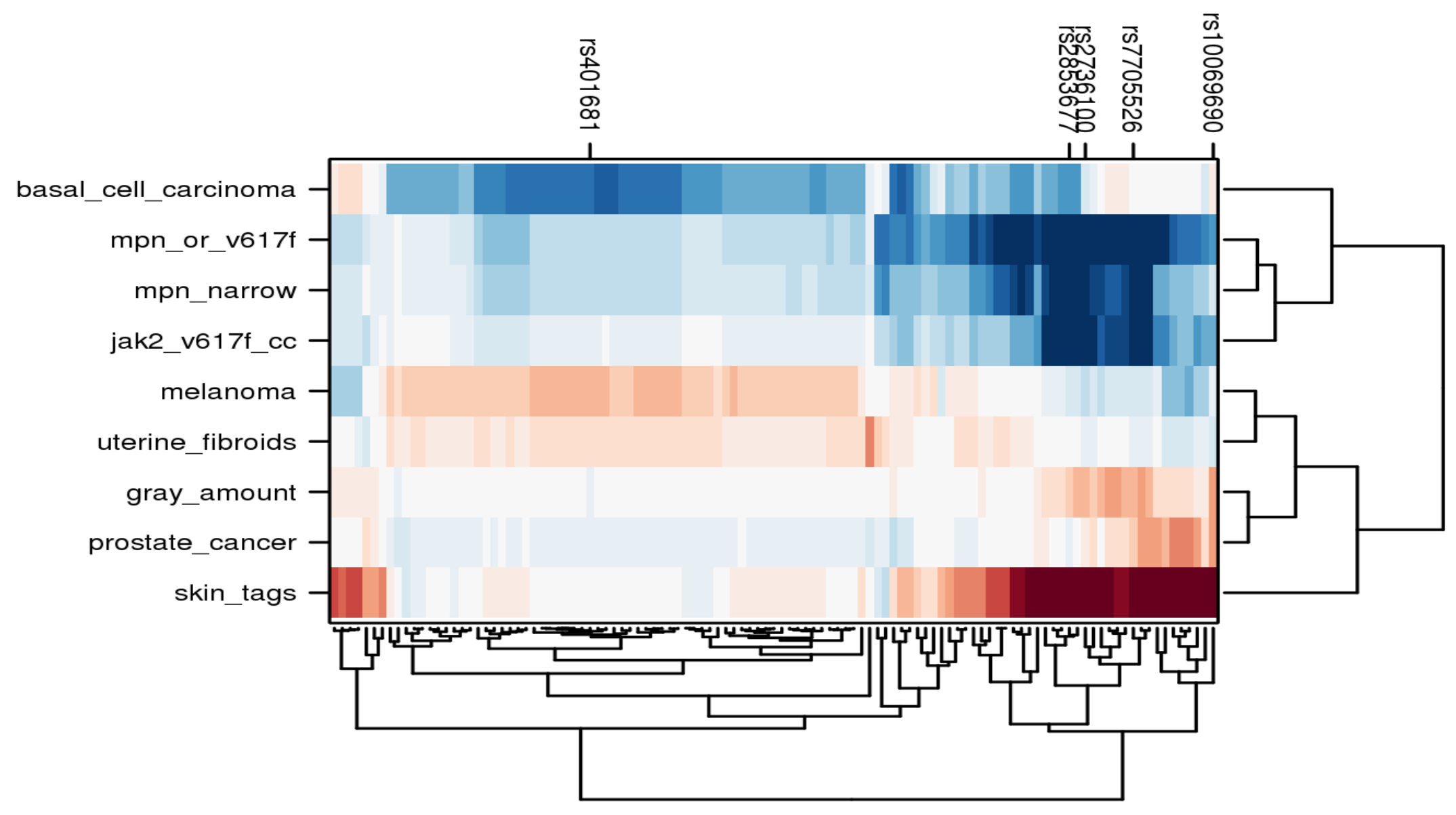


Figure 7. Phenotypic profiling at the *TERT* locus. We selected SNPs associated with any phenotype, and phenotypes associated with any SNP, at P<1e-4, from a curated set of 300 GWAS, and bi-clustered the association test results.

Current Collaborations

- Asthma, rhinitis, eczema, infectious and autoimmune disease
- Cardiovascular disease
- Myloproliferative neoplasms
- Morphology: baldness, pigmentation, facial features
- Neurological: Parkinson's disease, migraine, restless legs
- Personality, depression and bipolar disorders
- Recombination rates

What about sequencing?

- Cost of sequencing has fallen fast but not fast enough to displace SNP arrays
- There is very little that sequence data can add to what we could say to a healthy individual
- For complex disease genetics, the benefits of sequencing at scale have been somewhat limited
- For now, the best application of sequencing to what we do is to do deeper imputation, which can leverage all the SNP data we've already collected

Why do more, larger GWAS?

- Genetic association data is inherently sparse: filtered by availability of functional variation
- More and larger GWAS allow us to more fully populate and annotate biological networks using genetic information
- To do that, we would like to have hundreds of associations for thousands of phenotypes in millions of samples
- Larger sample sizes, deeper imputation expands the set of gene targets with functional variation that can be effectively interrogated
- Broad phenotyping can give biological context that can help with interpretation of disease associations

Acknowledgments

We thank the customers of 23andMe who answered surveys, as well as the employees of 23andMe, who together made this research possible. This work was supported in part by the National Human Genome Research Institute of the National Institutes of Health under grant number 2R44HG006981-02.