Using 1000 genomes data in disease studies

Jeff Barrett

ASHG, 3 November 2010

Two parallel goals in complex disease genetics

For a given disease, can we:

1. Explain heritability
2. Understand biology

Two parallel goals in complex disease genetics

For a given disease, can we:

1. Explain heritability (prediction/prognosis?)
2. Understand biology (treatment?)

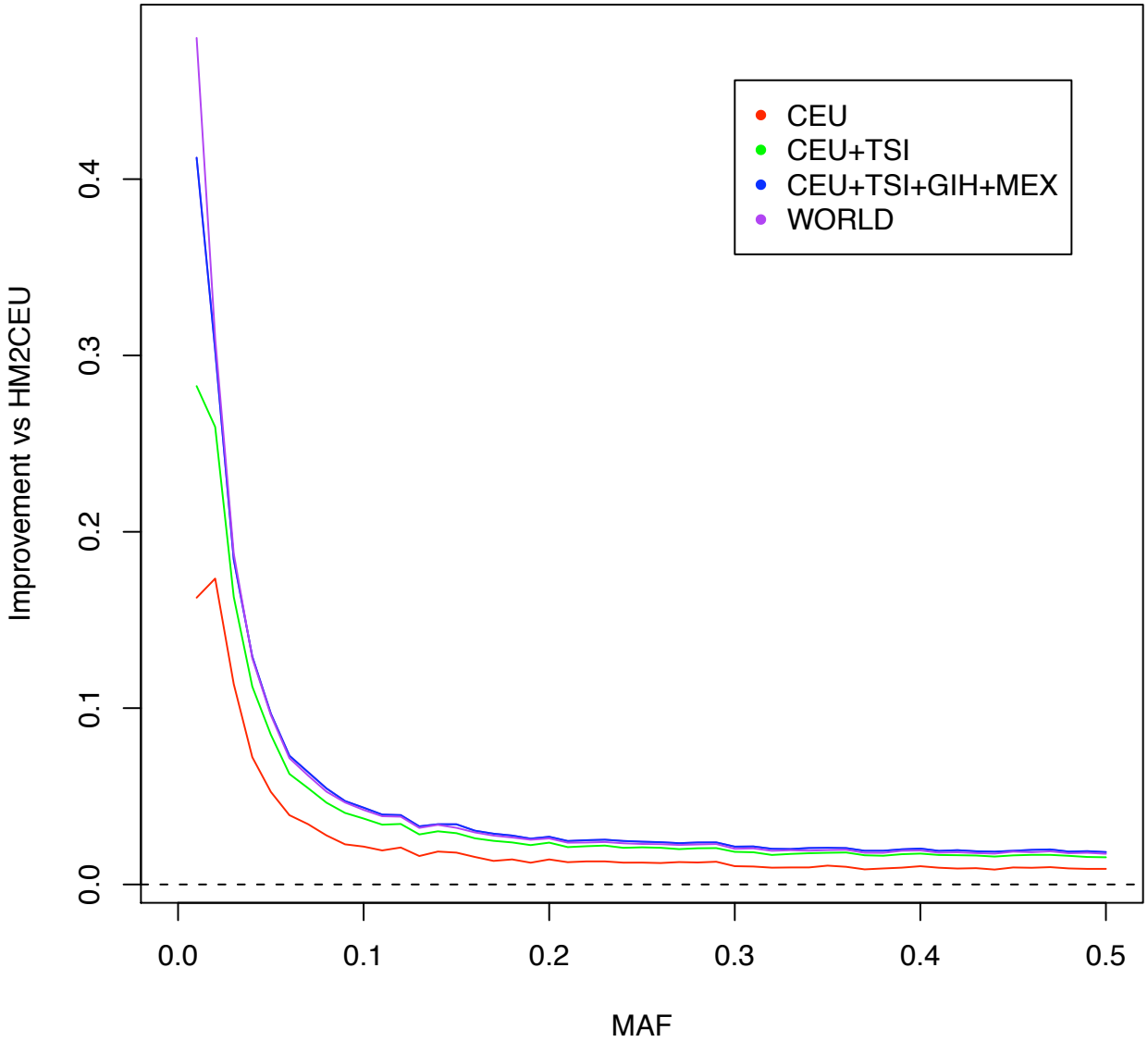
Two parallel goals in complex disease genetics

For a given disease, can we:

1. Explain heritability (imputation)
2. Understand biology (annotation)

1000 genomes data can play an important role in both these goals.

Accurate and deep reference sets are key to imputing low frequency variants



Imputation software

- IMPUTE v2 (Howie & Marchini)
  `mathgen.stats.ox.ac.uk/impute/impute_v2.html`
- BEAGLE (Browning & Browning)
  `faculty.washington.edu/browning/beagle/beagle.html`
- MACH & Minimac (Li, Fuchsberger & Abecasis)
  `www.sph.umich.edu/csg/abecasis/MACH/`
  `genome.sph.umich.edu/wiki/Minimac`

Imputation is computationally heavy duty

- Imputing into $\approx 16,000$ WTCCC samples using combined SNP/indel pilot data
- Formatting files, aligning strands, etc. can be fiddly
- IMPUTE v2 'factory default' settings
- Genome split into ≈ 600 chunks, each chunk submitted as a job to Sanger farm, each job requiring 4–6 GB memory
- Total processing time > 2 CPU years

Imputation is computationally heavy duty

- Imputing into $\approx 16,000$ WTCCC samples using combined SNP/indel pilot data
- Formatting files, aligning strands, etc. can be fiddly
- IMPUTE v2 'factory default' settings
- Genome split into ≈ 600 chunks, each chunk submitted as a job to Sanger farm, each job requiring 4–6 GB memory
- Total processing time > 2 CPU years
- 1–2 CPU hours per sample (scales approx linearly with sample size)

Pre-phasing can save a great deal of time

- Simplistically, imputation aims to match skeletal target haplotypes to more complete (in terms of variation) reference haplotypes.
- In the past, target datasets have been unphased genotype data (e.g. basic GWAS output). This requires a combination of phasing and matching, which underlies much of the computational burden.
- Phasing target data in advance (and saving the result) means imputation, and re-imputation with other references, is much faster (comparing haplotypes to each other, rather than genotypes to pairs of haplotypes).
- Implemented via flags in IMPUTE v2, BEAGLE and via Minimac for MACH.

Reference data, past, present & future

- Past: HapMap2 and HapMap3 (270–1000 samples, 2 million SNPs)

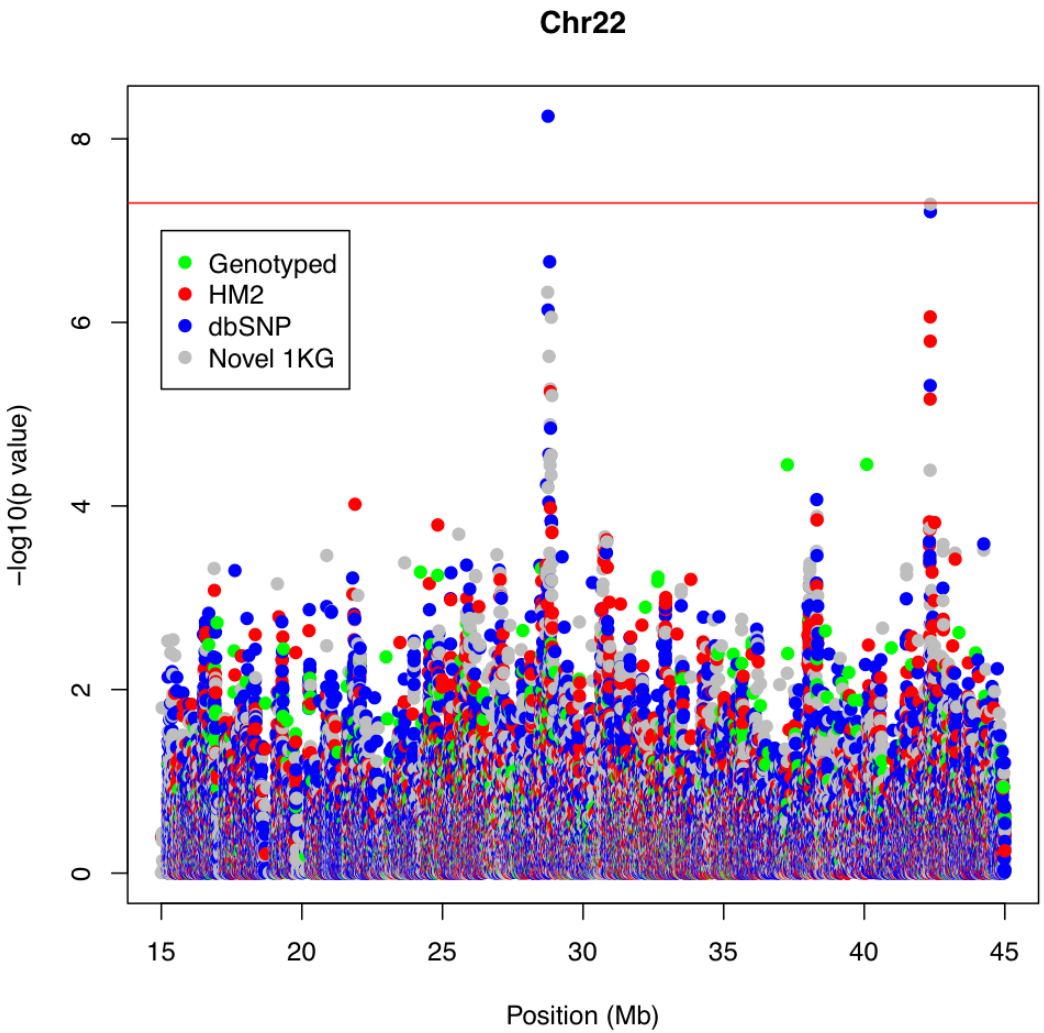
Reference data, past, present & future

- Past: HapMap2 and HapMap3 (270–1000 samples, 2 million SNPs)
- Present: 1000 genomes pilot (179 samples, >10 million SNPs & small indels, SV coming)
  `www.1000genomes.org`
  `mathgen.stats.ox.ac.uk/impute/impute_v2.html`

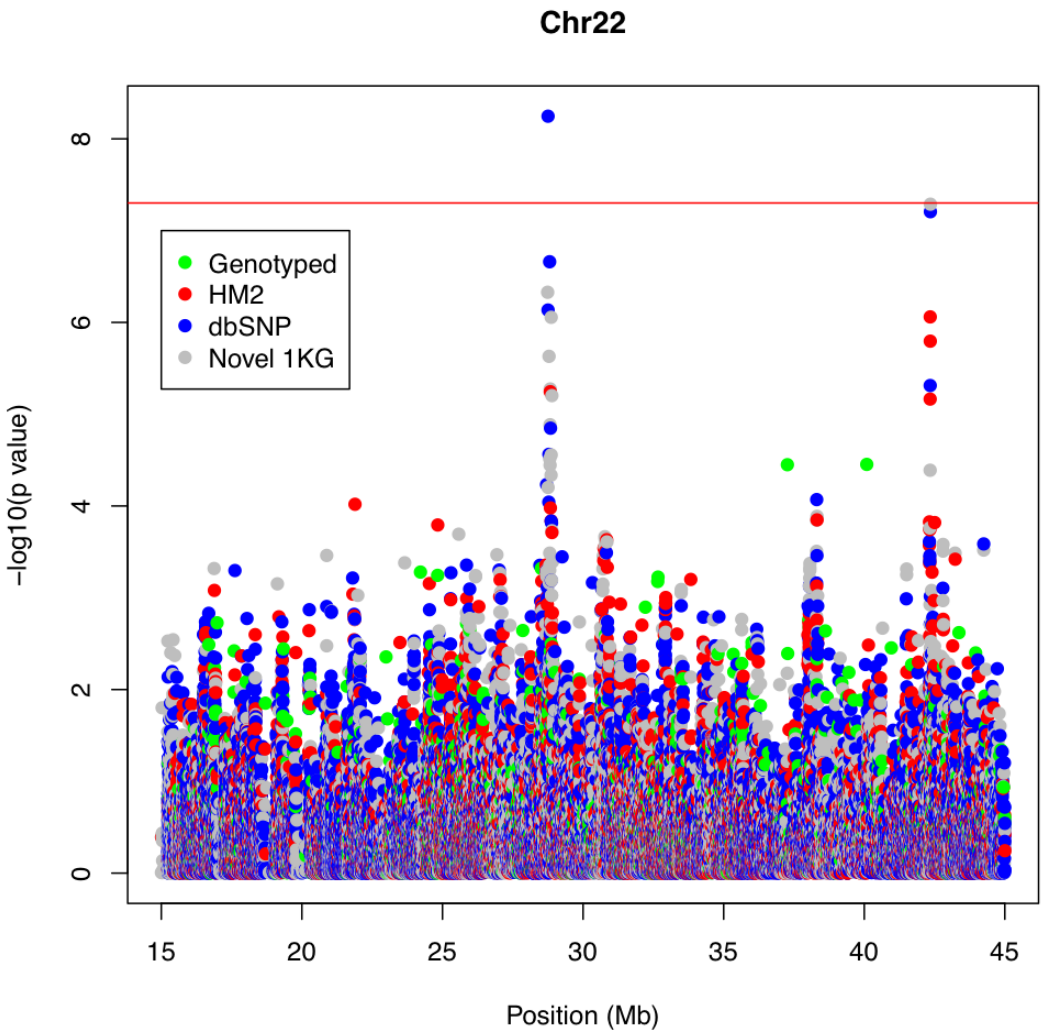
Reference data, past, present & future

- Past: HapMap2 and HapMap3 (270–1000 samples, 2 million SNPs)
- Present: 1000 genomes pilot (179 samples, >10 million SNPs & small indels, SV coming)
  `www.1000genomes.org`
  `mathgen.stats.ox.ac.uk/impute/impute_v2.html`
- Future: 1000 genomes complete data (2,500 samples, 30(?) million SNPs, indels, SVs). Phased releases of data integrated from all platforms (low coverage sequence, high coverage exomes, genotyping arrays, arrayCGH. . .)

Example: Crohn's disease

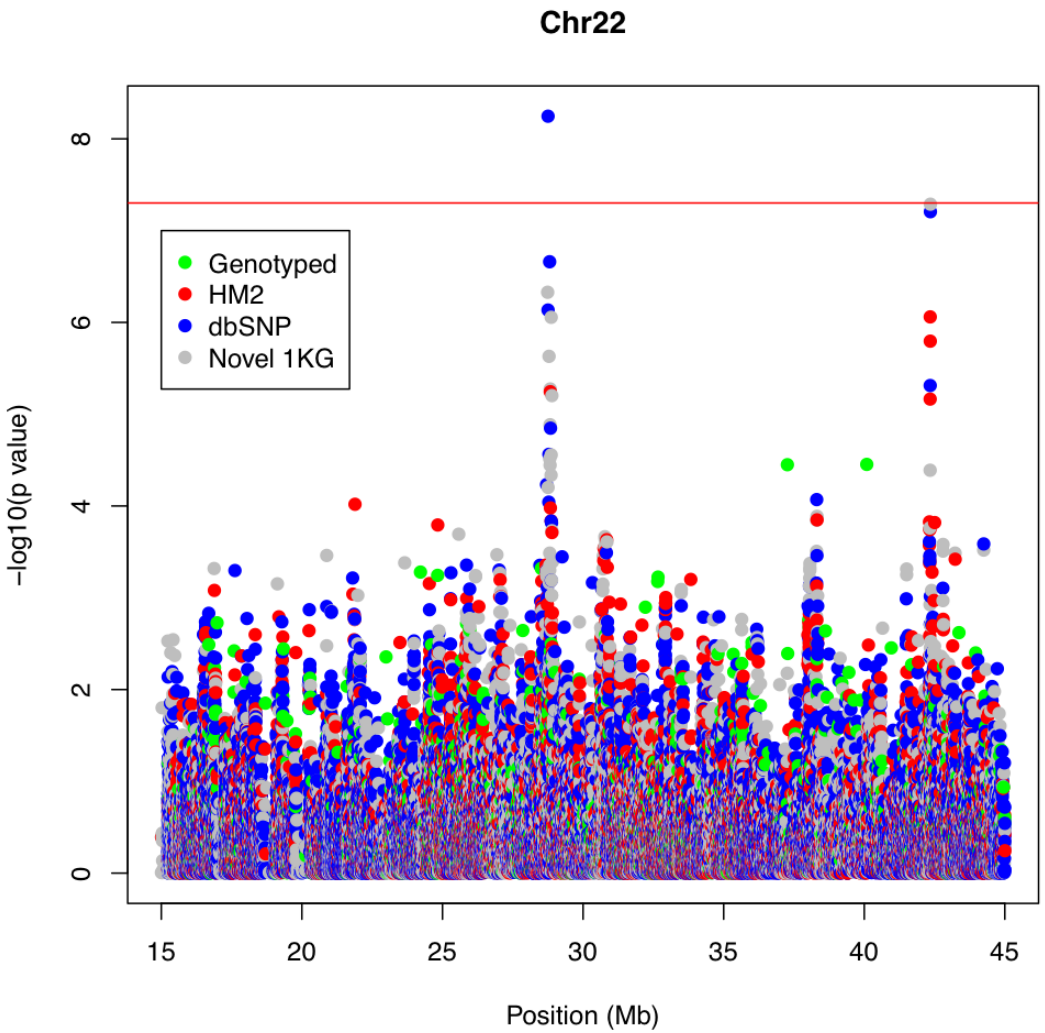


Example: Crohn's disease



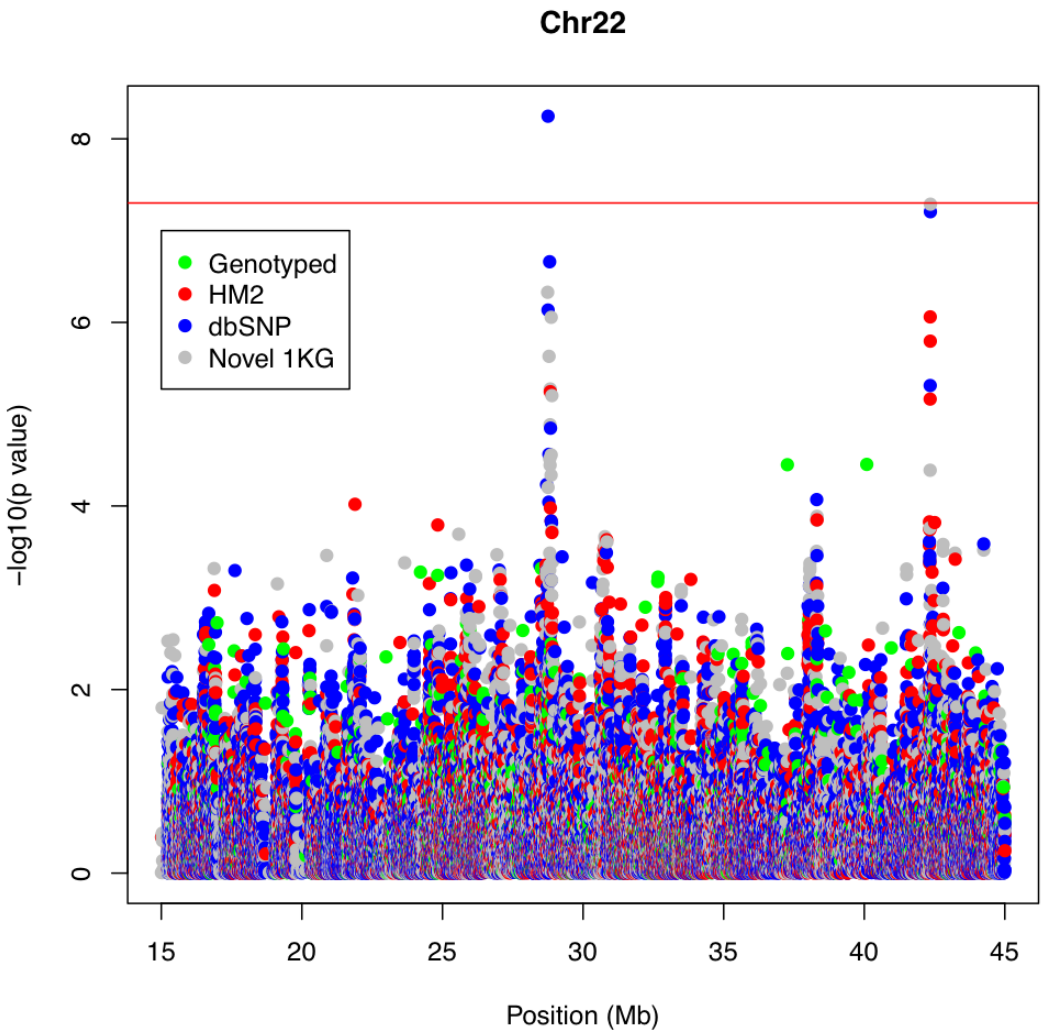
- Hit at 28 MB missed in WTCCC and 2008 meta-analysis ($p > 10^{-4}$).
  Hit SNP MAF: 3%

Example: Crohn's disease



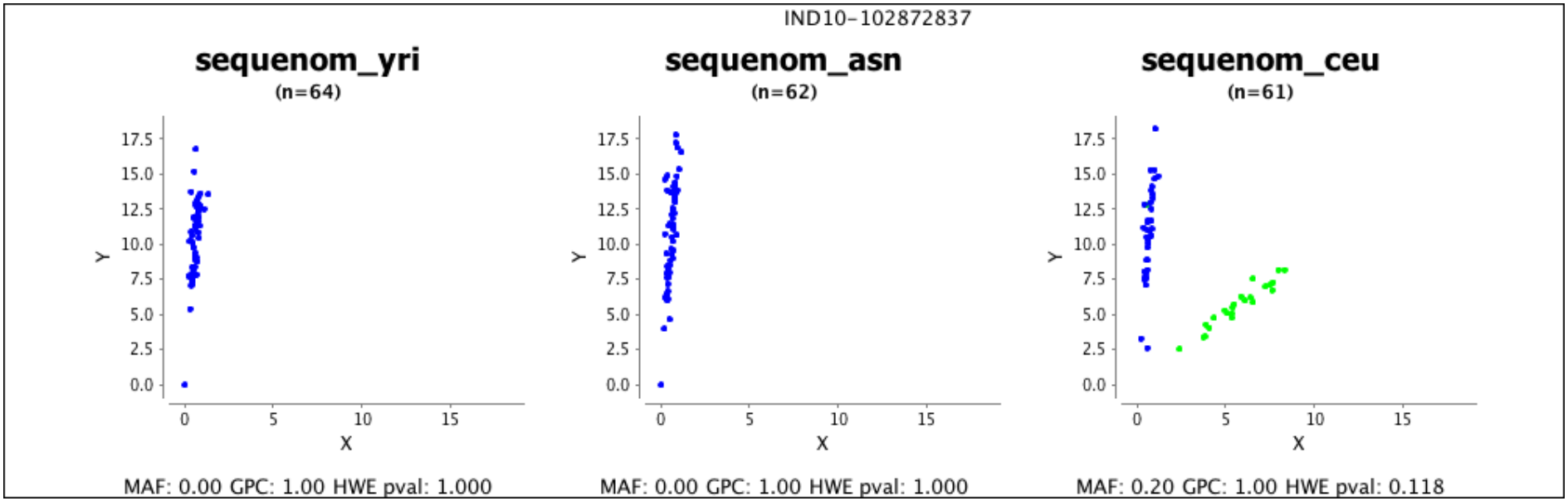
- Hit at 28 MB missed in WTCCC and 2008 meta-analysis ($p > 10^{-4}$). Hit SNP MAF: 3%
- Picked up in 2010 meta-analysis ($> 20,000$ total samples). Hit SNP MAF: 13%

Example: Crohn's disease

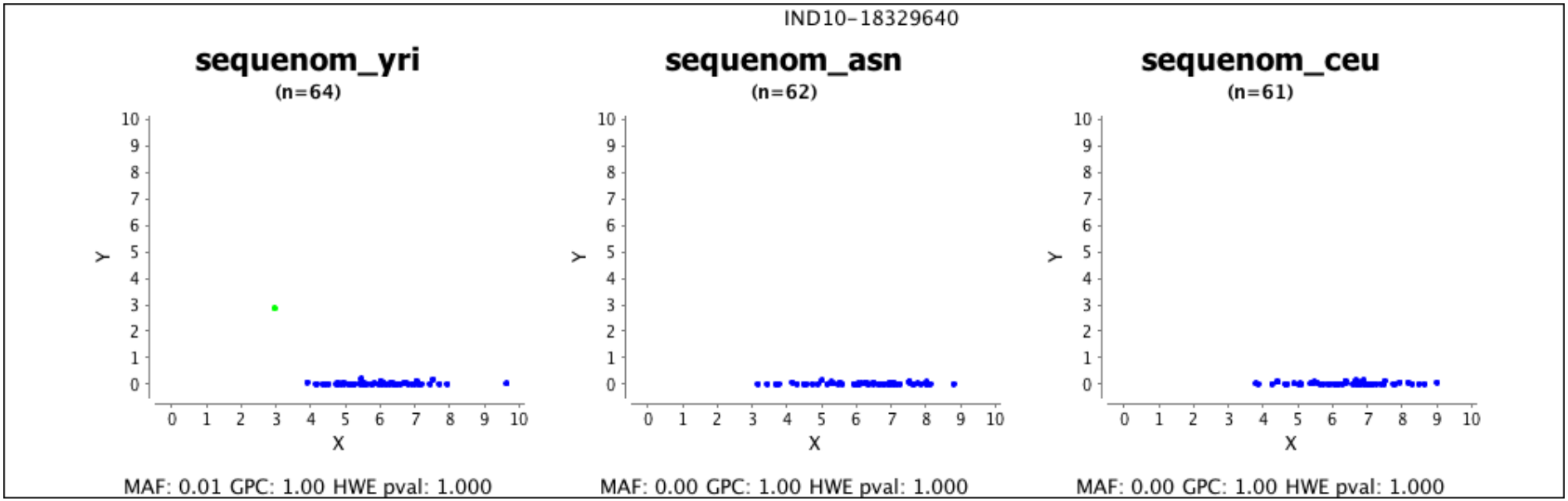


- Hit at 28 MB missed in WTCCC and 2008 meta-analysis ($p > 10^{-4}$). Hit SNP MAF: 3%
- Picked up in 2010 meta-analysis ($> 20,000$ total samples). Hit SNP MAF: 13%
- Hit at 42 MB not supported at all in meta-analysis...

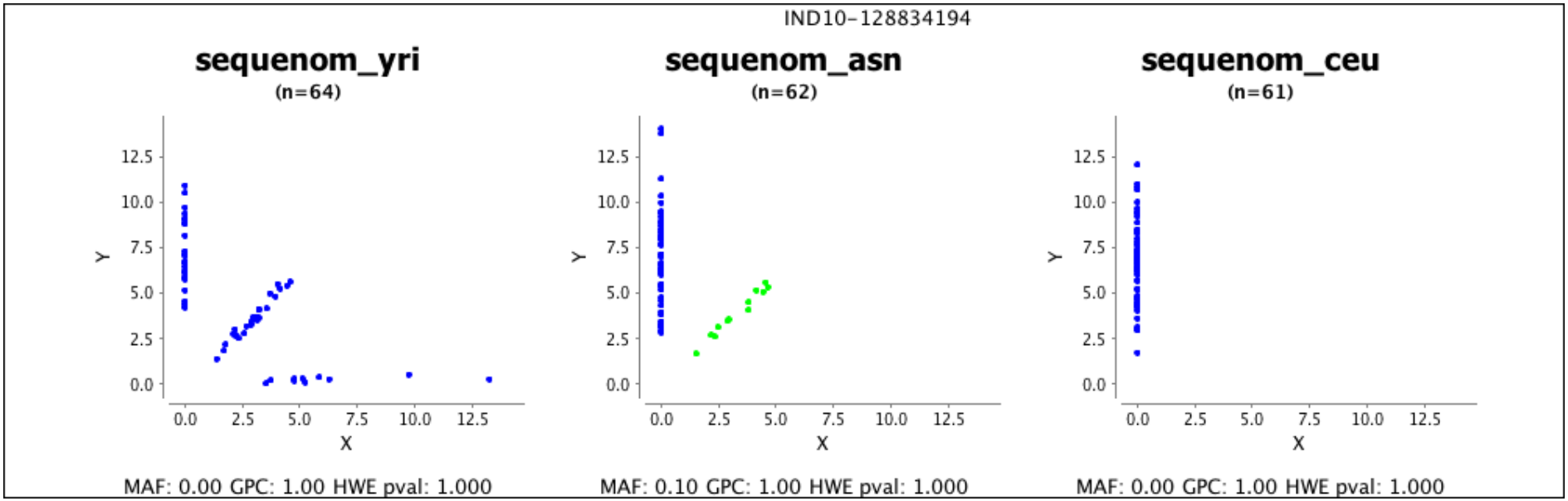
Validation and 'gold standards'



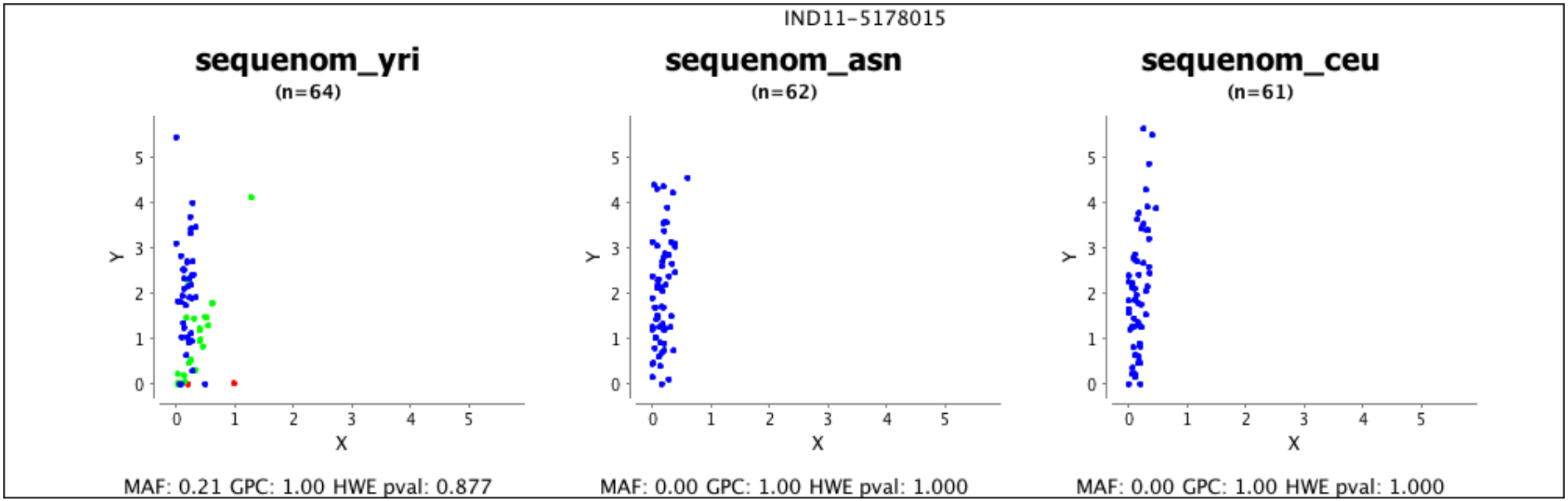
Validation and 'gold standards'



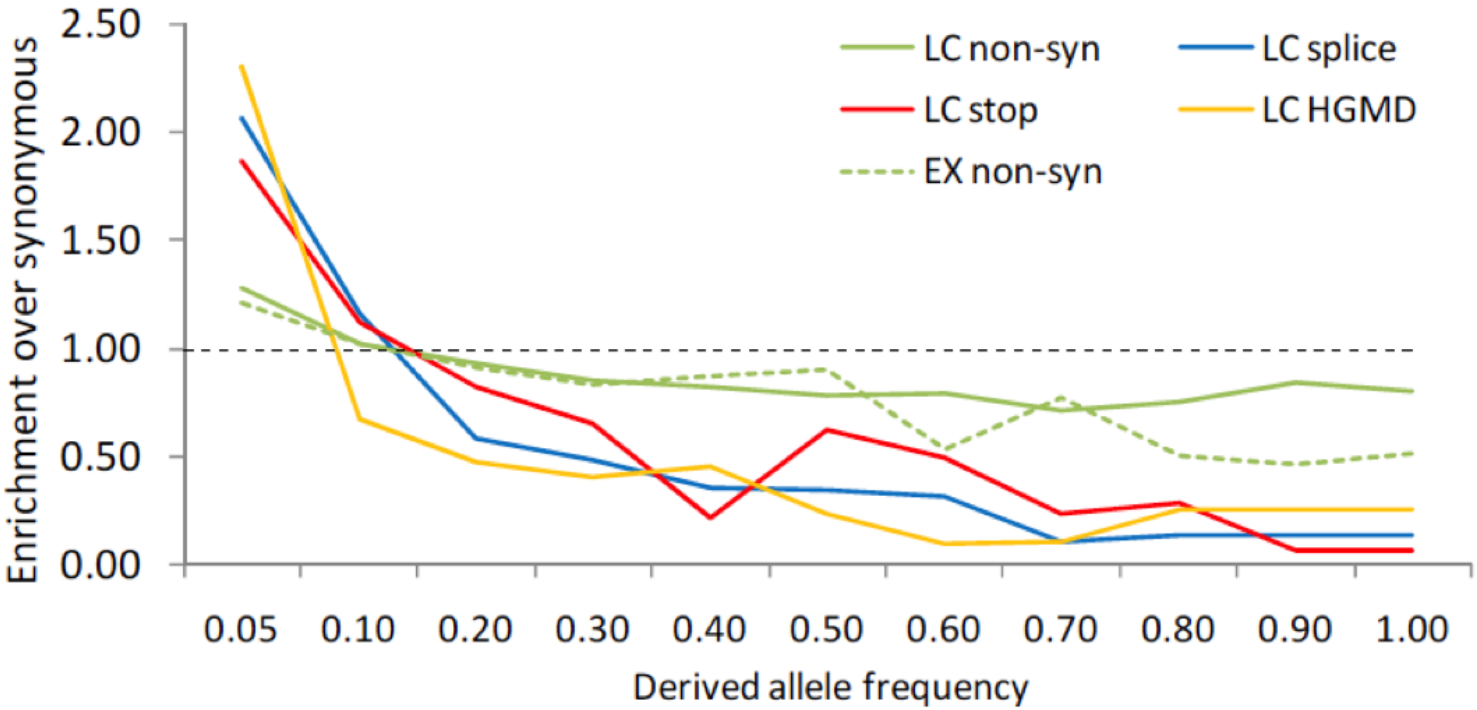
Validation and 'gold standards'



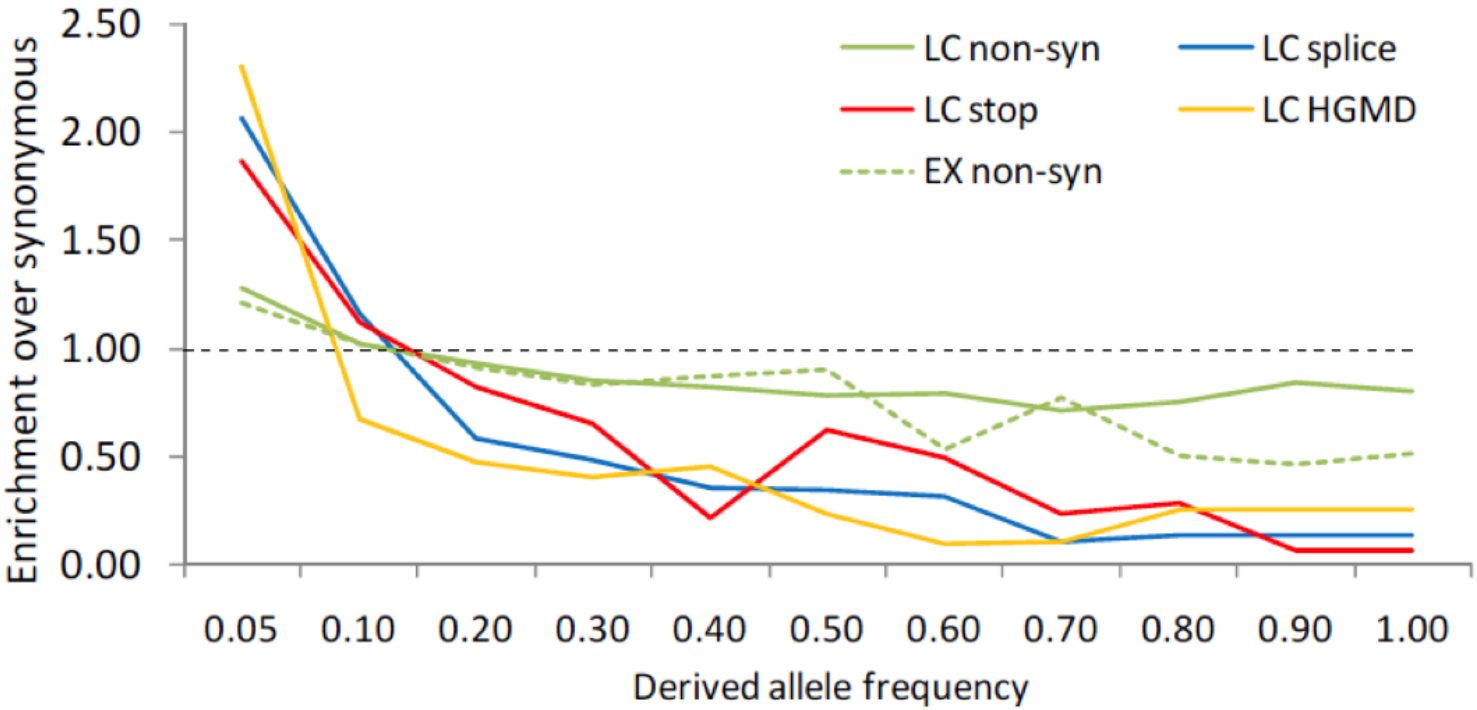
Validation and 'gold standards'



Functional annotation



Functional annotation



. . . $<$ 10% of GWAS hit SNPs have $r^2 > 0.9$ with a coding SNP

Conclusions

- 1000 genomes data will increasingly become the default imputation reference panel.

Conclusions

- 1000 genomes data will increasingly become the default imputation reference panel.
- The project also enables the discovery and annotation of variants, standardization of files and development of genotyping products.

Conclusions

- 1000 genomes data will increasingly become the default imputation reference panel.
- The project also enables the discovery and annotation of variants, standardization of files and development of genotyping products.
- Coming to grips with the subtleties of the data will take time and continue to evolve.

Thanks

1000 Genomes Project
Gonçalo Abecasis
Brian Browning
Bryan Howie
Jonathan Marchini
Daniel Macarthur
James Morris

Luke Jostins