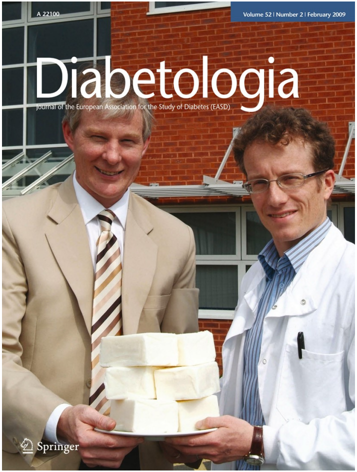
Haplotypes and Imputation

Jeff Barrett

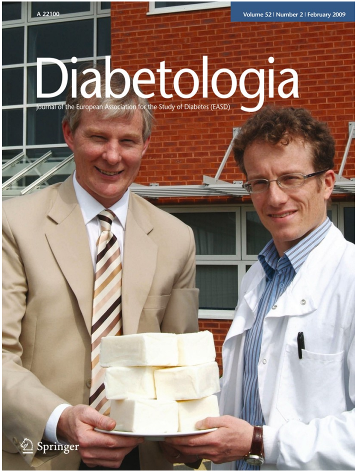
Boulder Workshop, 2011

Finding a gene is exciting. . .

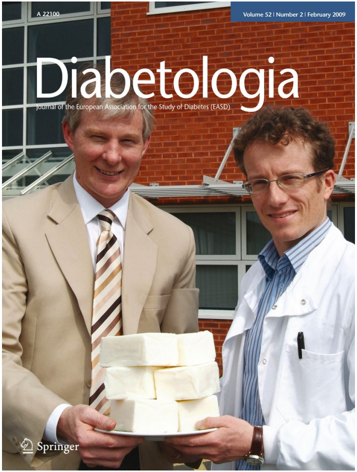
Finding a gene is exciting. . .


A 22100
Volume 52 | Number 2 | February 2009
Diabetologia
Journal of the European Association for the Study of Diabetes (EASD)
Springer

Finding a gene is exciting. . .


A 22100
Volume 52 | Number 2 | February 2009
Diabetologia
Journal of the European Association for the Study of Diabetes (EASD)
Springer

Finding a gene is exciting. . .



A 22100

Volume 52 | Number 2 | February 2009

Diabetologia

Journal of the European Association for the Study of Diabetes (EASD)

Springer

. . . but what to do next?

Stages of genetic mapping

My work focuses on developing statistical and computational tools for the design of genetic studies, the detection of gene variants influencing complex human traits and the dissection of these effects in the larger context of other genetic and environmental factors.

—Shaun “Prof. PLINK” Purcell

Stages of genetic mapping

My work focuses on developing statistical and computational tools for the design of genetic studies, the detection of gene variants influencing complex human traits and the dissection of these effects in the larger context of other genetic and environmental factors.

—Shaun "Prof. PLINK" Purcell

Stages of genetic mapping

My work focuses on developing statistical and computational tools for the design of genetic studies, the detection of gene variants influencing complex human traits and the dissection of these effects in the larger context of other genetic and environmental factors.

—Shaun "Prof. PLINK" Purcell

Stages of genetic mapping

My work focuses on developing statistical and computational tools for the design of genetic studies, the detection of gene variants influencing complex human traits and the dissection of these effects in the larger context of other genetic and environmental factors.

—Shaun "Prof. PLINK" Purcell

Returning to LD

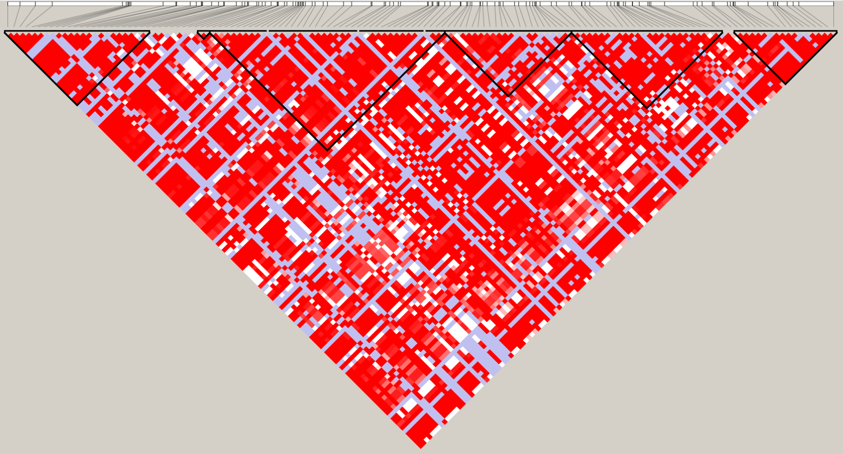
		SNP 1	
		p	1-p
SNP 2	q	π_{11}	π_{12}
	1-q	π_{21}	π_{22}

$$D = \pi_{11} - pq$$

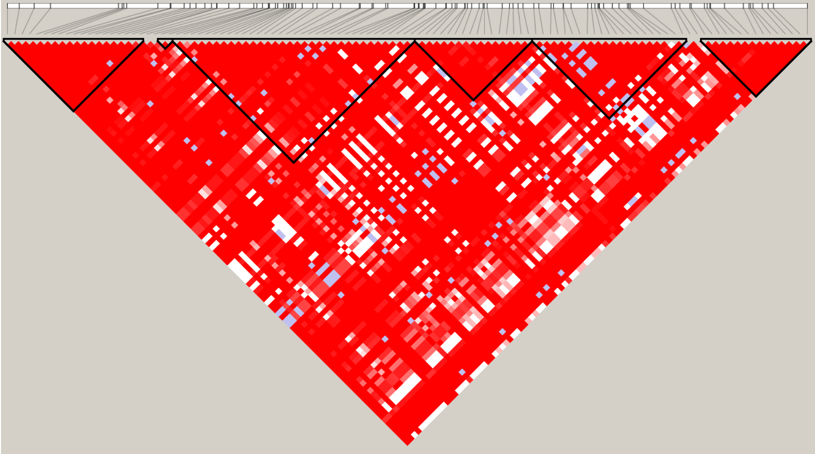
$$D' = D/D_{\mathrm{max}}$$

$$r^2 = D/p(1-p)q(1-q)$$

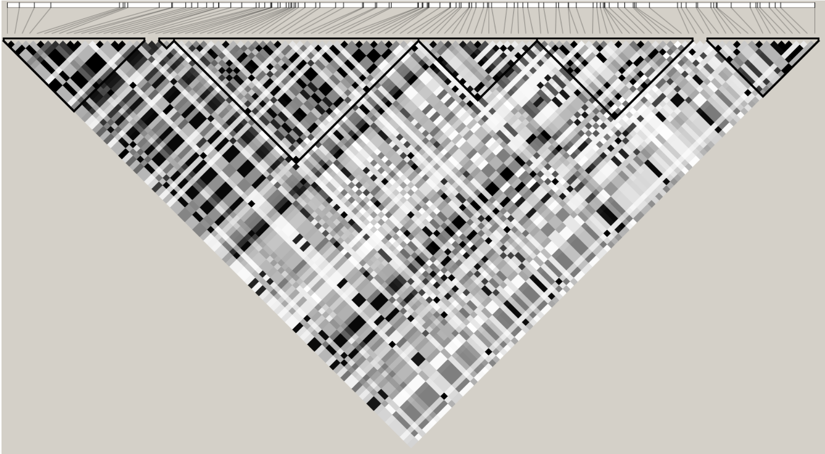
D' in a region of 100kb



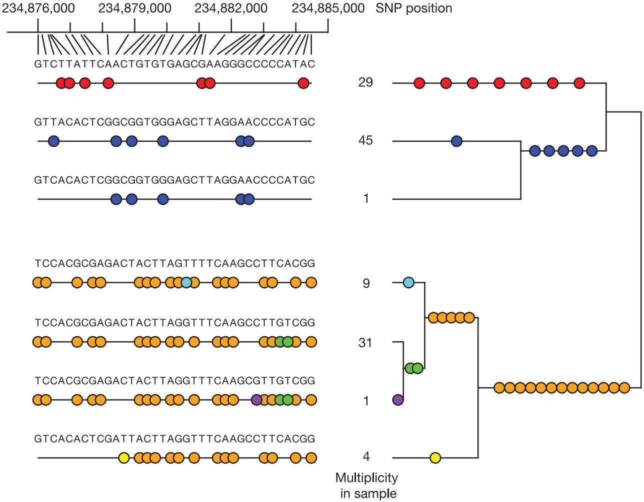
D' for common SNPs in a region of 100kb



r^2 for common SNPs in a region of 100kb

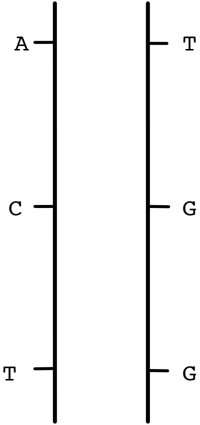


D' and r^2 in a haplotypic context


234,876,000
234,879,000
234,882,000
234,885,000
SNP position
GTCTTATTCAACTGTGTGAGCGAAGGGCCCCCATAC
29
GTTACACTCGGCGGTGGGAGCTTAGGAACCCCATGC
45
GTCACACTCGGCGGTGGGAGCTTAGGAACCCCATGC
1
TCCACGCGAGACTACTTAGTTTTCAAGCCTTCACGG
9
TCCACGCGAGACTACTTAGGTTTCAAGCCTTGTCGG
31
TCCACGCGAGACTACTTAGGTTTCAAGCGTTGTCGG
1
GTCACACTCGATTACTTAGGTTTCAAGCCTTCACGG
4
Multiplicity
in sample

Haplotypes and phase

The truth...



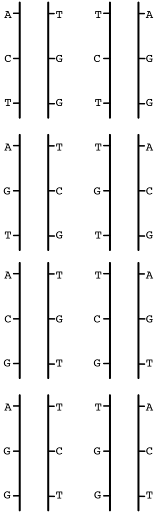
What we observe...

AT

CG

GT

Haplotypes and phase



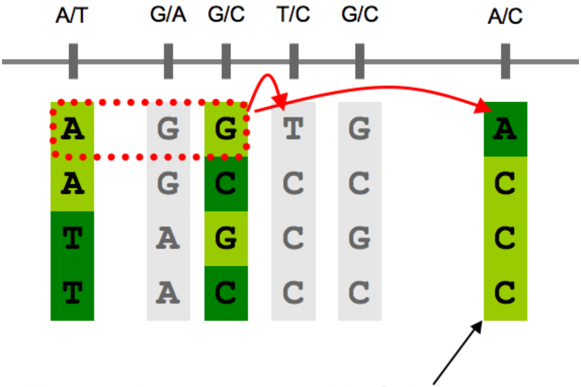
What we observe...

AT

CG

GT

Haplotype analysis



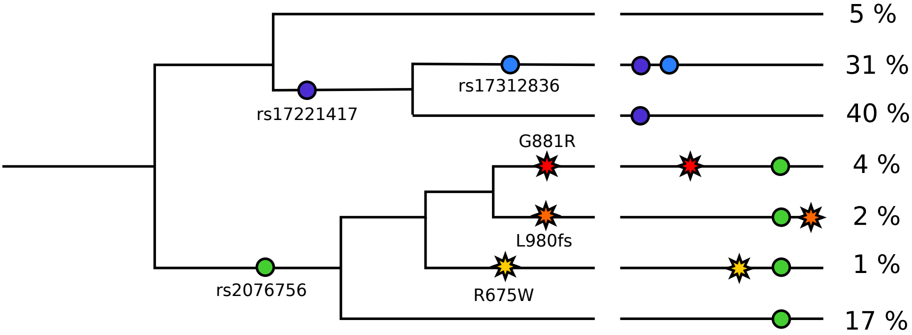
A/T	G/A	G/C	T/C	G/C	A/C
A	G	G	T	G	A
A	G	C	C	C	C
T	A	G	C	G	C
T	A	C	C	C	C

No need to genotype this SNP

Rare Variants Create Synthetic Genome-Wide Associations

Samuel P. Dickson[1,2], Kai Wang[3], Ian Krantz[3,4,5], Hakon Hakonarson[3,4,5], David B. Goldstein[1]*

1 Institute for Genome Sciences and Policy, Center for Human Genome Variation, Duke University, Durham, North Carolina, United States of America, **2** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Center for Applied Genomics, Children's Hospital of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **4** Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **5** Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America



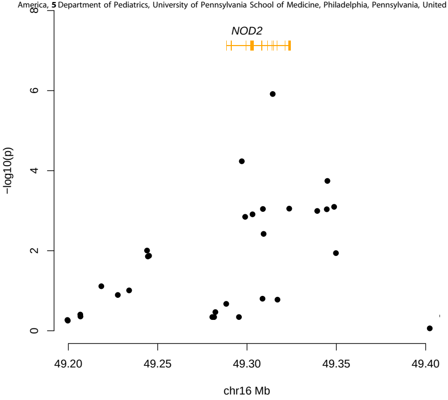
OPEN ACCESS Freely available online

PLoS BIOLOGY

Rare Variants Create Synthetic Genome-Wide Associations

Samuel P. Dickson[1,2], Kai Wang[3], Ian Krantz[3,4,5], Hakon Hakonarson[3,4,5], David B. Goldstein[1]*

1 Institute for Genome Sciences and Policy, Center for Human Genome Variation, Duke University, Durham, North Carolina, United States of America, **2** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Center for Applied Genomics, Children's Hospital of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **4** Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **5** Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America



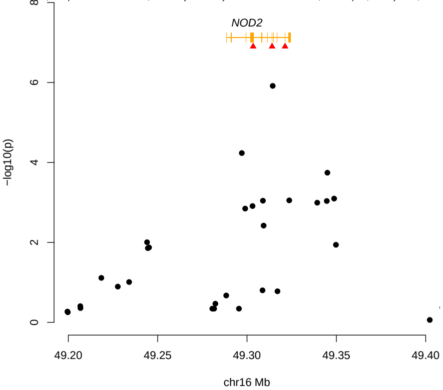
OPEN ACCESS Freely available online

PLoS BIOLOGY

Rare Variants Create Synthetic Genome-Wide Associations

Samuel P. Dickson[1,2], Kai Wang[3], Ian Krantz[3,4,5], Hakon Hakonarson[3,4,5], David B. Goldstein[1]*

1 Institute for Genome Sciences and Policy, Center for Human Genome Variation, Duke University, Durham, North Carolina, United States of America, **2** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Center for Applied Genomics, Children's Hospital of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **4** Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **5** Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America



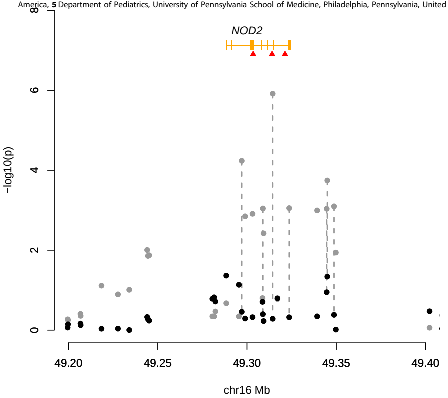
OPEN ACCESS Freely available online

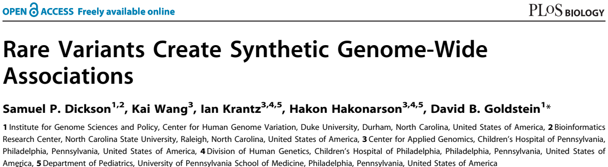
PLOS BIOLOGY

Rare Variants Create Synthetic Genome-Wide Associations

Samuel P. Dickson[1,2], Kai Wang[3], Ian Krantz[3,4,5], Hakon Hakonarson[3,4,5], David B. Goldstein[1]*

1 Institute for Genome Sciences and Policy, Center for Human Genome Variation, Duke University, Durham, North Carolina, United States of America, **2** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Center for Applied Genomics, Children's Hospital of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **4** Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **5** Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America





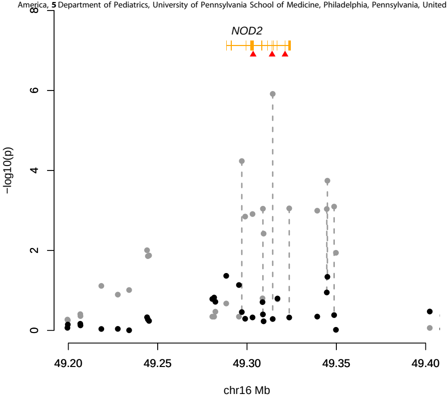
OPEN ACCESS Freely available online

PLOS BIOLOGY

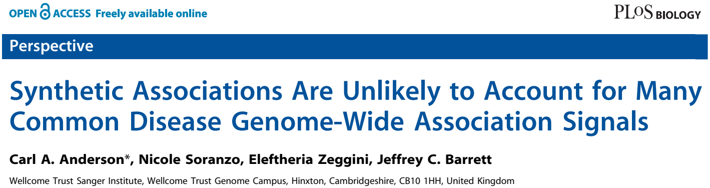
Rare Variants Create Synthetic Genome-Wide Associations

Samuel P. Dickson[1,2], Kai Wang[3], Ian Krantz[3,4,5], Hakon Hakonarson[3,4,5], David B. Goldstein[1]*

1 Institute for Genome Sciences and Policy, Center for Human Genome Variation, Duke University, Durham, North Carolina, United States of America, **2** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Center for Applied Genomics, Children's Hospital of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **4** Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **5** Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America



Tag SNP explains 0.8% of h^2, but causal alleles explain 5%



OPEN ACCESS Freely available online

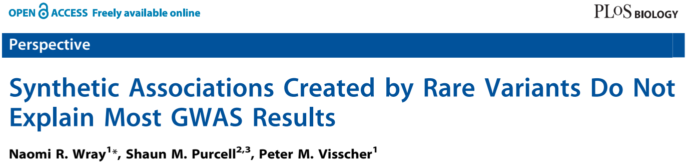
PLoS BIOLOGY

Perspective

Synthetic Associations Are Unlikely to Account for Many Common Disease Genome-Wide Association Signals

Carl A. Anderson*, Nicole Soranzo, Eleftheria Zeggini, Jeffrey C. Barrett

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1HH, United Kingdom



OPEN ACCESS Freely available online

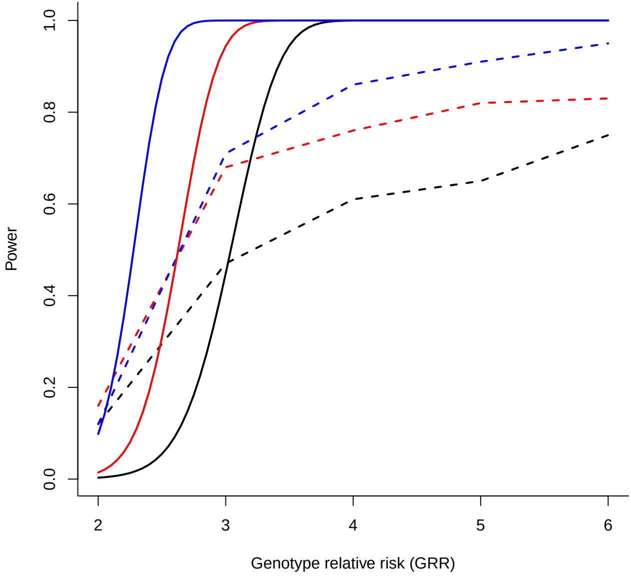
PLoS BIOLOGY

Perspective

Synthetic Associations Created by Rare Variants Do Not Explain Most GWAS Results

Naomi R. Wray[1]*, Shaun M. Purcell[2,3], Peter M. Visscher[1]

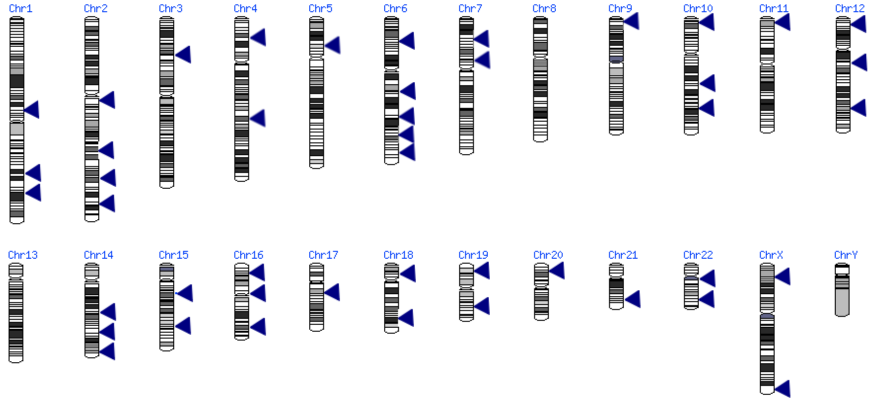
Linkage is well powered to detect these models



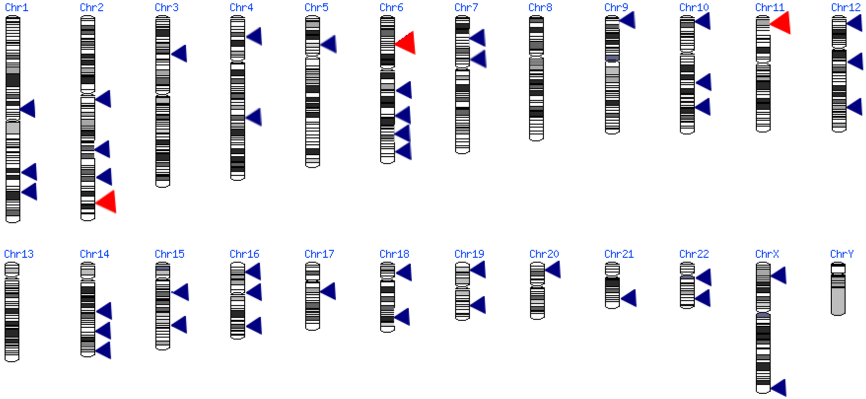
Linkage:
2,657 ASPs (T1DGC)

GWAS:
2,000/2,000

Overlap of T1DGC GWAS and linkage results



Overlap of T1DGC GWAS and linkage results



Imputation cartoon

Observed Genotypes

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

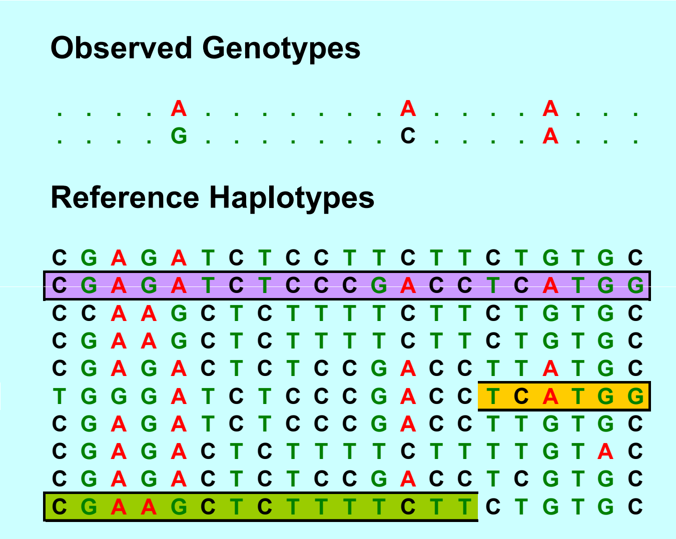
Study
Sample

Reference Haplotypes

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

HapMap

Imputation cartoon



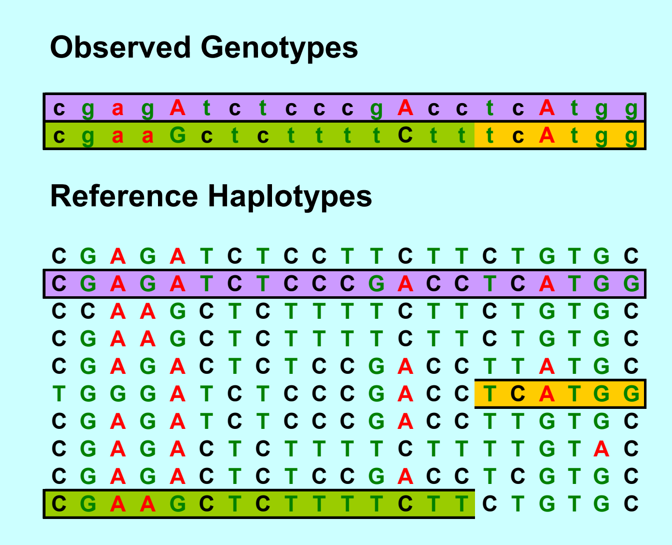
Observed Genotypes

. . . . A . . . . A . . . . A . . .
. . . . G . . . . C . . . . A . . .

Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C

Imputation cartoon



Observed Genotypes

c g a g A t c t c c c g A c c t c A t g g

c g a a G c t c t t t t C t t t c A t g g

Reference Haplotypes

C G A G A T C T C C T T C T T C T G T G C

C G A G A T C T C C C G A C C T C A T G G

C C A A G C T C T T T T C T T C T G T G C

C G A A G C T C T T T T C T T C T G T G C

C G A G A C T C T C C G A C C T T A T G C

T G G G A T C T C C C G A C C T C A T G G

C G A G A T C T C C C G A C C T T G T G C

C G A G A C T C T T T T C T T T T G T A C

C G A G A C T C T C C G A C C T C G T G C

C G A A G C T C T T T T C T T C T G T G C

Imputation implementation (MACH, IMPUTE)

- Markov model used to model each haplotype conditional on all others
- Markov chain Monte Carlo (e.g. Gibbs sampler) is used to estimate parameters, and update predicted (imputed) haplotypes
  - Each individual is updated conditional on all the others
  - In parallel to updating haplotypes, estimate “error rates” and “crossover” probabilities
- Simpler models (e.g. BEAGLE, no parameters to estimate) appropriate in some circumstances

Imputation is computationally heavy-duty

- A GWAS of N samples typed on M SNPs yields a very large $N \times M$ matrix of input data into imputation
- 'Chunking' can be done both along the *sample* and *SNP* axes
- Sample chunks should be mixed case/control in the same ratio as the overall sample (on the order of hundreds of samples per chunk)
- SNP chunks should be at least several Mb, with overlapping buffers at chunk breakpoints to avoid edge effects.

Pre-phasing can save a great deal of time

- 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 and requires less memory.
- 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. . .)

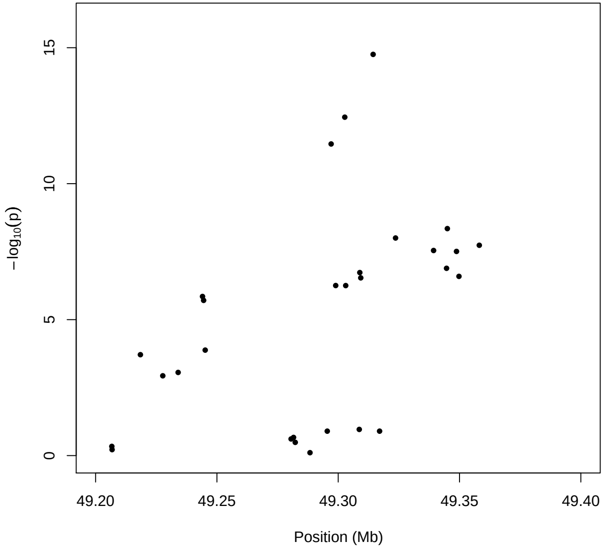
Example: WTCCC & 1000 Genomes pilot reference

- Imputing into $\approx 16,000$ WTCCC samples using combined SNP/indel 1000 genomes pilot data
- IMPUTE v2 'factory default' settings (N.B. formatting files, aligning strands, etc. can be fiddly)
- Total processing time > 2 CPU years

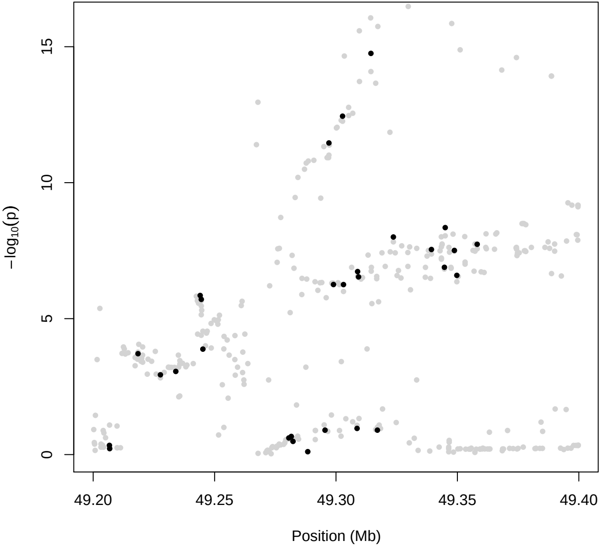
Example: WTCCC & 1000 Genomes pilot reference

- Imputing into $\approx 16,000$ WTCCC samples using combined SNP/indel 1000 genomes pilot data
- IMPUTE v2 'factory default' settings (N.B. formatting files, aligning strands, etc. can be fiddly)
- Total processing time > 2 CPU years
- Genome split into $\approx$ 600 chunks (5+1 Mb), runs of 1600 samples
- Each chunk submitted as a job (6000 total) to Sanger farm, each job requiring 4–6 GB memory
- 1–2 CPU hours per sample (scales approx linearly with sample size)

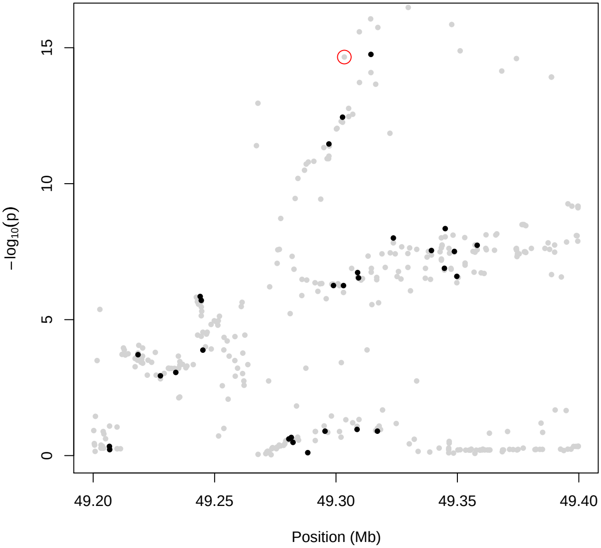
Imputation of rare alleles can identify causal variants



Imputation of rare alleles can identify causal variants



Imputation of rare alleles can identify causal variants



Analyzing imputed data

- Can transform probabilistic outputs of imputation into “best guess” genotypes, which has some advantages in interpretability, but is not advisable except when confidence is very high

Analyzing imputed data

- Can transform probabilistic outputs of imputation into "best guess" genotypes, which has some advantages in interpretability, but is not advisable except when confidence is very high
- Straightforward to analyze using a logistic regression on "dosage" $e_{ij} = 0p_{ij0} + 1p_{ij1} + 2p_{ij2}$ (e.g. PLINK)

Analyzing imputed data

- Can transform probabilistic outputs of imputation into “best guess” genotypes, which has some advantages in interpretability, but is not advisable except when confidence is very high
- Straightforward to analyze using a logistic regression on “dosage” $e_{ij} = 0p_{ij0} + 1p_{ij1} + 2p_{ij2}$ (e.g. PLINK)
- Other tests (either frequentist or Bayesian) possible, but not much difference except when uncertainty is very high (e.g. SNPTEST)

Analyzing imputed data

- Can transform probabilistic outputs of imputation into "best guess" genotypes, which has some advantages in interpretability, but is not advisable except when confidence is very high
- Straightforward to analyze using a logistic regression on "dosage" $e_{ij} = 0p_{ij0} + 1p_{ij1} + 2p_{ij2}$ (e.g. PLINK)
- Other tests (either frequentist or Bayesian) possible, but not much difference except when uncertainty is very high (e.g. SNPTEST)
- All programs produce a confidence metric of imputed data (IMPUTE: info; MACH, BEAGLE: r^2). Filtering recommendations vary slightly, and represent a trade-off of power

Imputation resources

MACH
`http://www.sph.umich.edu/csg/abecasis/MACH/`
`http://genome.sph.umich.edu/wiki/Minimac`

IMPUTE
`http://mathgen.stats.ox.ac.uk/impute/impute_v2.html`

BEAGLE
`http://faculty.washington.edu/browning/beagle/beagle.html`

Marchini & Howie. *Nat Rev Genet*. 2010.