1000 Genomes

A Deep Catalog of Human Genetic Variation

The 1000 Genomes Project

Gil McVean

Department of Statistics, Oxford

How can we achieve large-scale GWAS with genomic sequence data **now**?

Well Come

MARY RIDDELL

THIS BUDGET IS LABOUR'S LAST CHANCE

PAGE 20

TELEGRAF

PICK Y MORE THAN

PAGE S14

The Daily Tel

Tuesday, April 21, 2009 ★★★

BRITAIN'S BEST-SELLING QUALITY DAILY

'False hope' in hunt for genetic cures

By Richard Alleyne and Kate Devlin

A LEADING scientist has claimed that the hope that genetic research could provide a cure for a host of common illnesses has proved a "false dawn".

Prof Steve Jones, a geneticist, said the belief that a few genes held the key to ridding the world of conditions such as cancer and diabetes had proved to be "plain wrong".

into genes and that there was a danger it had become "largely unfounded". "Just a couple of years ago, there was real optimism that a new era of understanding was around the corner," he said. "That did not last long, for hubris has been replaced with concern."

Prof Jones added: "Of course there have been some successes, but it is the 'cure all' aspect of the work that has proved unfounded.

"It is the nature of the busi-

Misery for Slumdog star

The dark matter of genetics

Why can we explain only a fraction of the genetic risk?

- For most complex disease/phenotypes, the proportion of the variance explained by GWAS hits is less than 5%

- What explains the missing heritability?
  - **Common, but untagged SNPs?**
  - **Structural variation?**
  - **Rare variants?**
  - **GxG interactions?**
  - **GxE interactions?**

The 1000 Genomes Project

What is the 1000 Genomes Project?

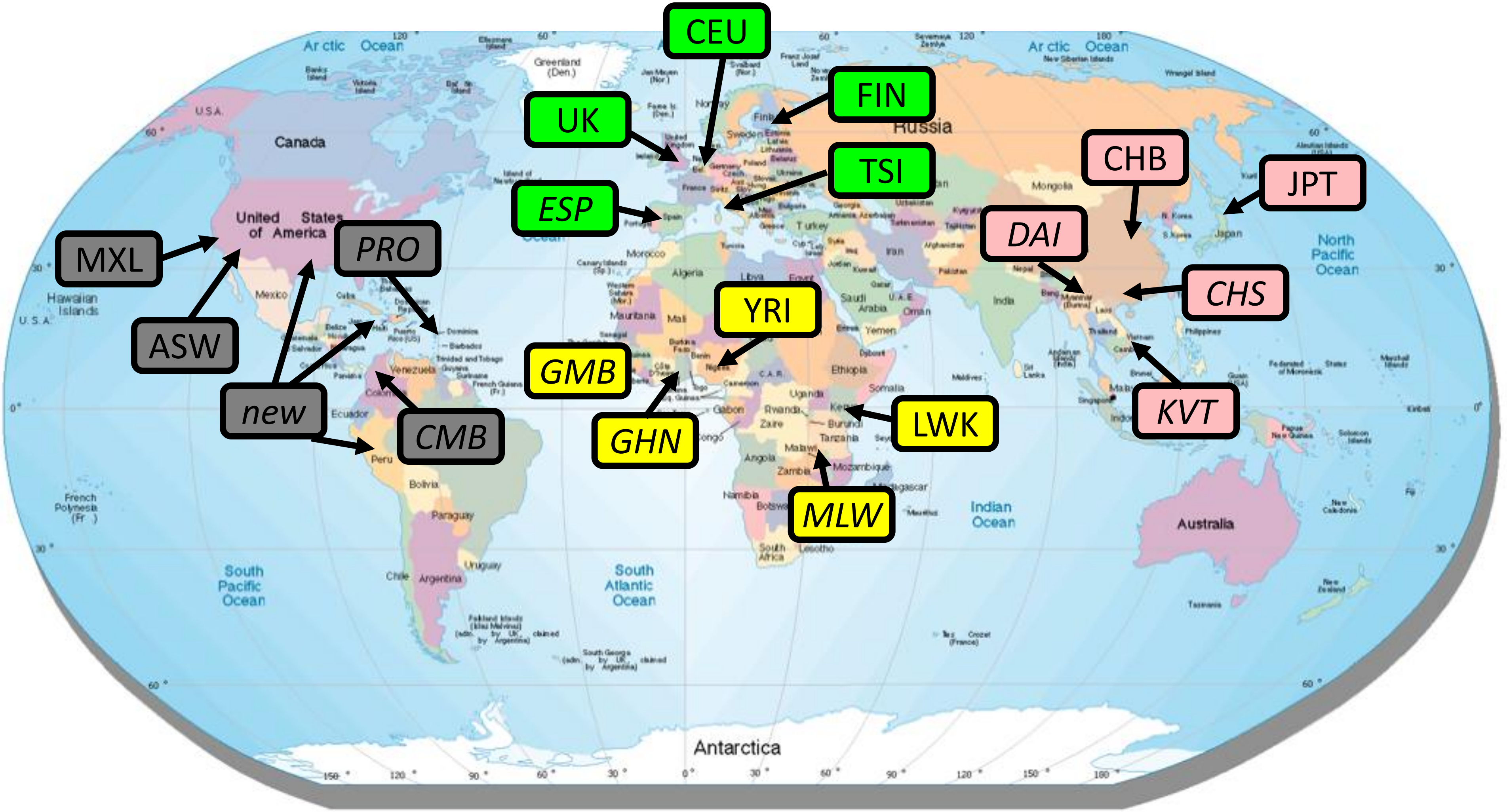
- A catalogue of **all** types of genetic variation, including **rare** variants (c. 1% frequency) obtained by sequencing at least 1000 individuals from geographic centres of major medical genetics interest

- A large international collaboration
  - UK, USA, China, Germany

- An exploration of the use of next-generation technologies for population-scale genome sequencing

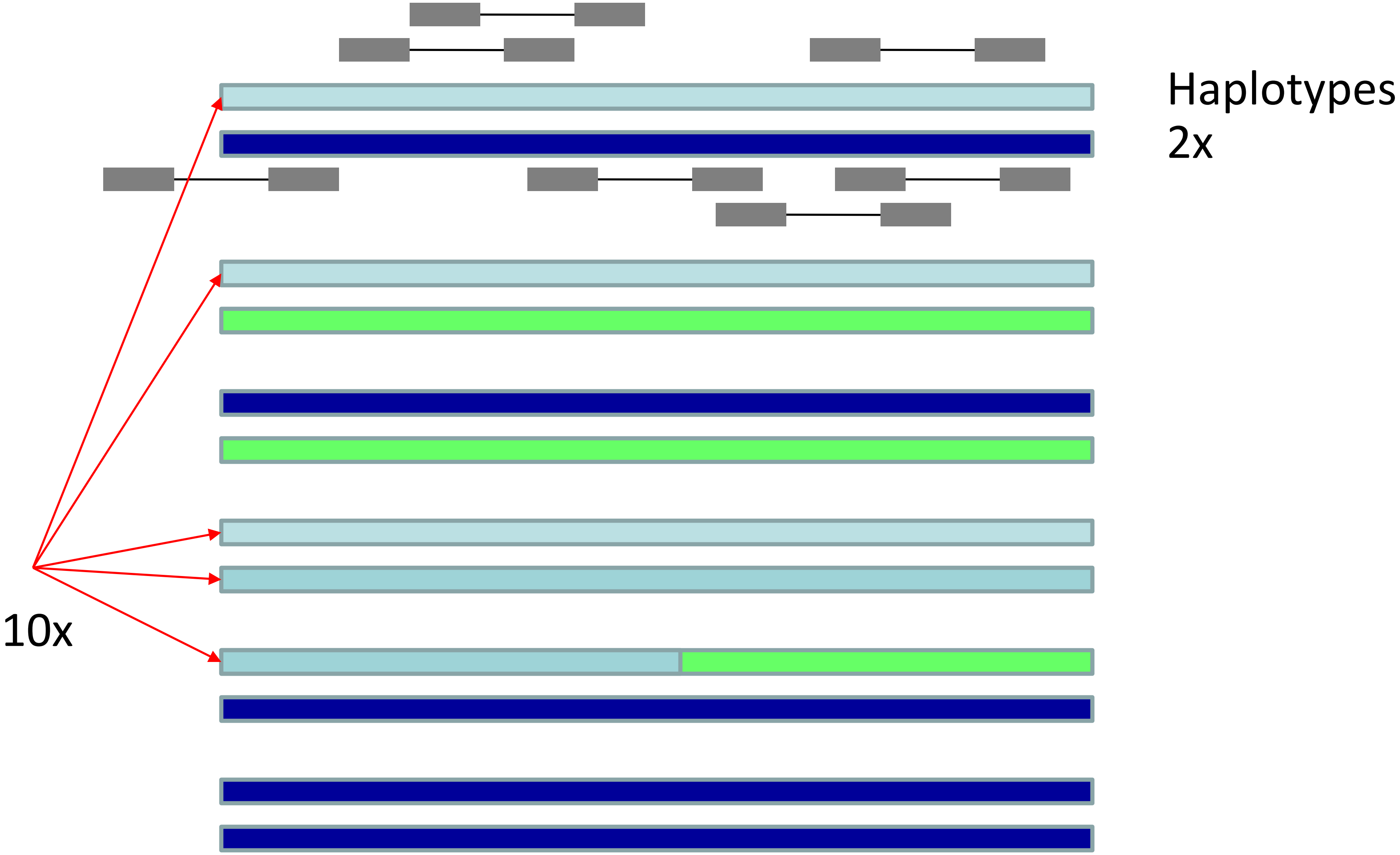
- A resource for accelerating the rate of identifying disease mechanisms in the follow-up to disease-association studies

Samples for the main project



Major population groups comprised of subpopulations of c. 100 each

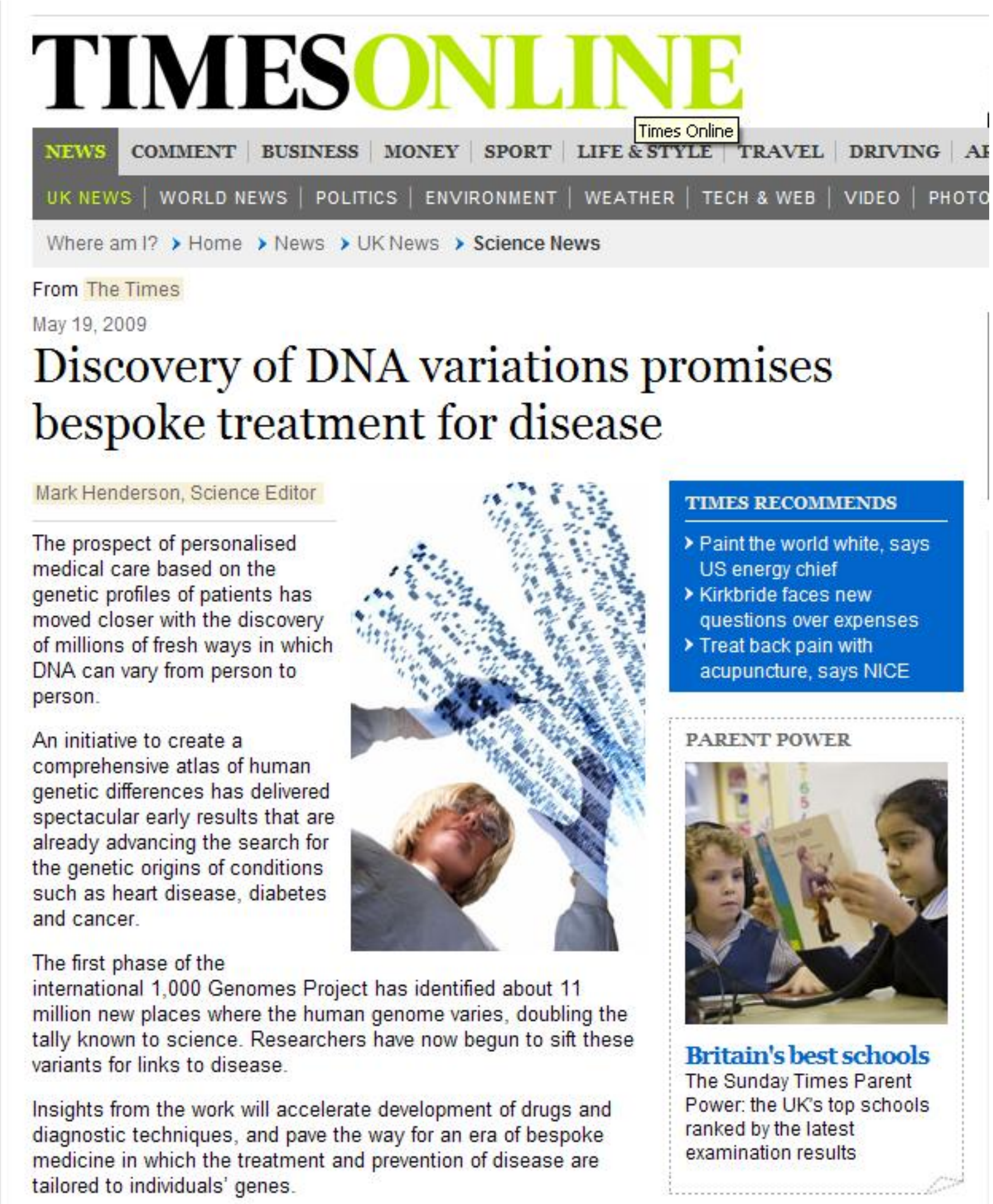
Population-scale sequencing



Pilot experiments

- Pilot 1
  - Low-coverage (4x-8x) on 60 unrelated individuals from each of CEU, YRI and CHB+JPT
- Pilot 2
  - High-coverage (20x diploid) on 2 trios (one from CEU, one from YRI)
- Pilot 3
  - Exons from 1000 genes to 20x in c. 1000 samples (largely European)

Complete!



TIMESONLINE
Times Online
NEWS | COMMENT | BUSINESS | MONEY | SPORT | LIFE & STYLE | TRAVEL | DRIVING | AF
UK NEWS | WORLD NEWS | POLITICS | ENVIRONMENT | WEATHER | TECH & WEB | VIDEO | PHOTO
Where am I? › Home › News › UK News › Science News

From The Times

May 19, 2009

Discovery of DNA variations promises bespoke treatment for disease

Mark Henderson, Science Editor

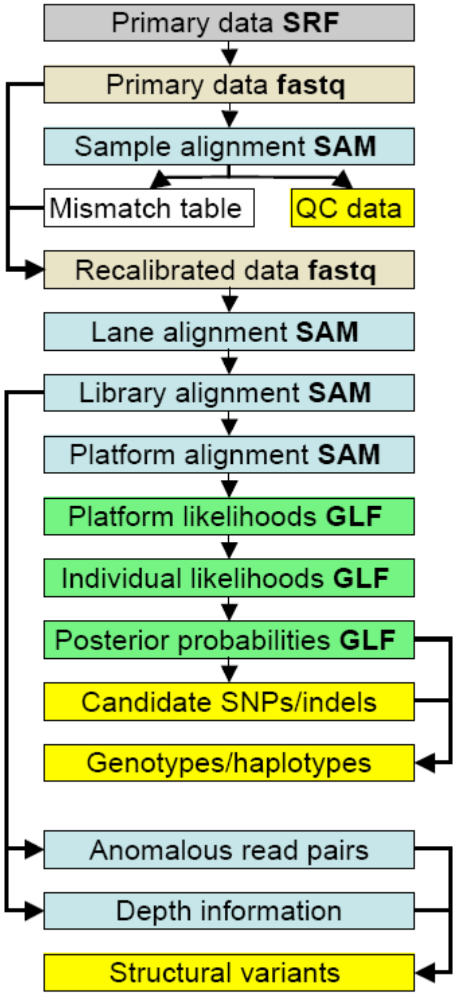
The prospect of personalised medical care based on the genetic profiles of patients has moved closer with the discovery of millions of fresh ways in which DNA can vary from person to person.

An initiative to create a comprehensive atlas of human genetic differences has delivered spectacular early results that are already advancing the search for the genetic origins of conditions such as heart disease, diabetes and cancer.

The first phase of the international 1,000 Genomes Project has identified about 11 million new places where the human genome varies, doubling the tally known to science. Researchers have now begun to sift these variants for links to disease.

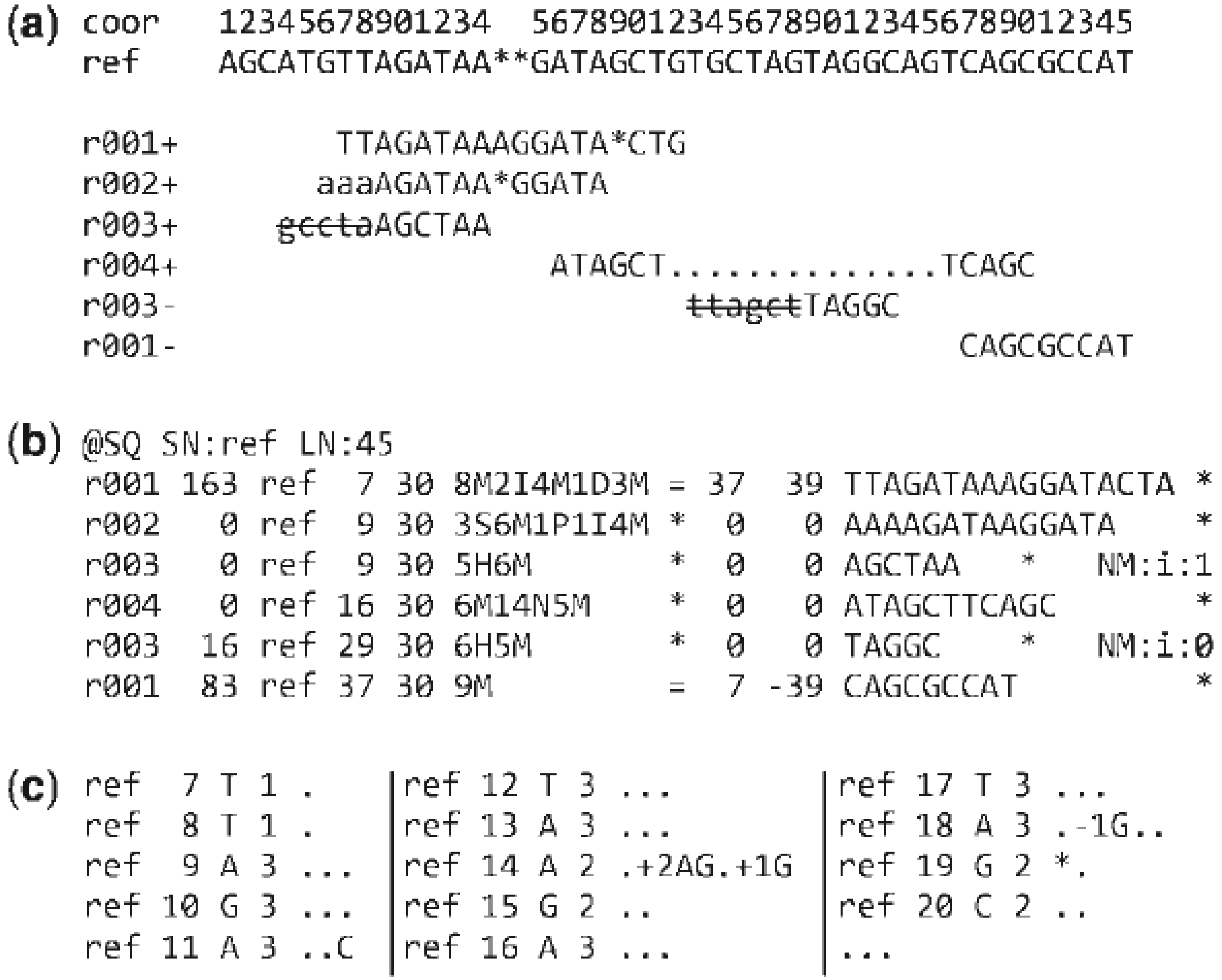
Insights from the work will accelerate development of drugs and diagnostic techniques, and pave the way for an era of bespoke medicine in which the treatment and prevention of disease are tailored to individuals' genes.

Data processing innovation and standards



		Unit (one file per …)	Who? (italics if not done yet)
1. Submit	Primary data **SRF**	*lane*	production centres
2. Extract	Primary data **fastq**	*lane*	DCC
3. Map sample	Sample alignment **SAM**	*lane*	Sanger *to DCC*
4. Recalibrate	Mismatch table QC data	*lane*	Sanger *to DCC*
	Recalibrated data **fastq**	*lane*	Sanger *to DCC*
5. Map	Lane alignment **SAM**	*lane*	Data Processing
6. Merge and remove dups	Library alignment **SAM**	*library*	Data Processing
7. Merge	Platform alignment **SAM**	*platform/individual*	Data Processing
8. Calc likelihoods	Platform likelihoods **GLF**	*platform/individual*	Data Processing
9. Combine l'hoods	Individual likelihoods **GLF**	*individual*	Data Processing
10. Apply priors	Posterior probabilities **GLF**	*individual*	Data Processing
11. Call SNPs/indels	Candidate SNPs/indels	*experiment/population*	Data Processing
12. Call genotypes	Genotypes/haplotypes	*individual*	Data Processing
13. Collect read pair info	Anomalous read pairs	*library*	Structural Variation
13. Collect depth info	Depth information	*library*	Structural Variation
14. Call SVs	Structural variants	*experiment and individual*	Structural Variation

Sequence AlignMent format



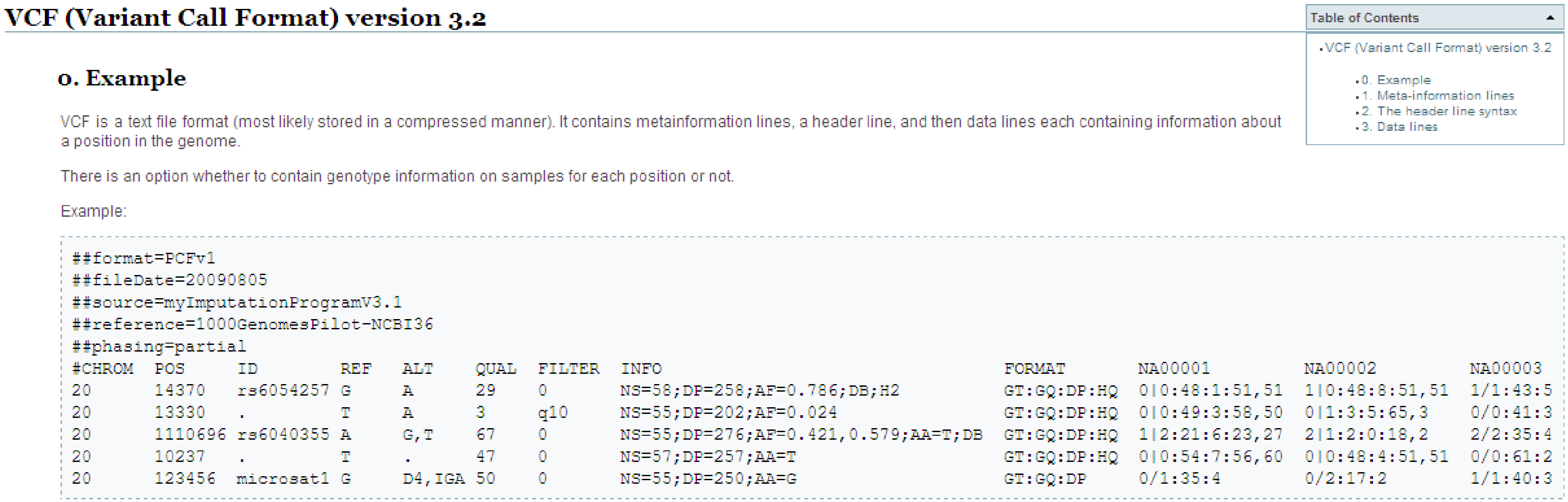
```
(a) coor   12345678901234  5678901234567890123456789012345
    ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

    r001+        TTAGATAAAGGATA*CTG
    r002+       aaaAGATAA*GGATA
    r003+     ~~gccta~~AGCTAA
    r004+                    ATAGCT..............TCAGC
    r003-                           ~~ttagct~~TAGGC
    r001-                                         CAGCGCCAT
```

```
(b) @SQ SN:ref LN:45
    r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
    r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
    r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
    r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
    r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
    r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

```
(c) ref  7 T 1 .    |ref 12 T 3 ...      |ref 17 T 3 ...
    ref  8 T 1 .    |ref 13 A 3 ...      |ref 18 A 3 .-1G..
    ref  9 A 3 ...  |ref 14 A 2 .+2AG.+1G|ref 19 G 2 *.
    ref 10 G 3 ...  |ref 15 G 2 ..       |ref 20 C 2 ..
    ref 11 A 3 ..C  |ref 16 A 3 ...      |...
```

Bioinformatics (2009) http://samtools.sourceforge.net

Variant Call Format



VCF (Variant Call Format) version 3.2

Table of Contents

0. Example

VCF is a text file format (most likely stored in a compressed manner). It contains metainformation lines, a header line, and then data lines each containing information about a position in the genome.

There is an option whether to contain genotype information on samples for each position or not.

Example:

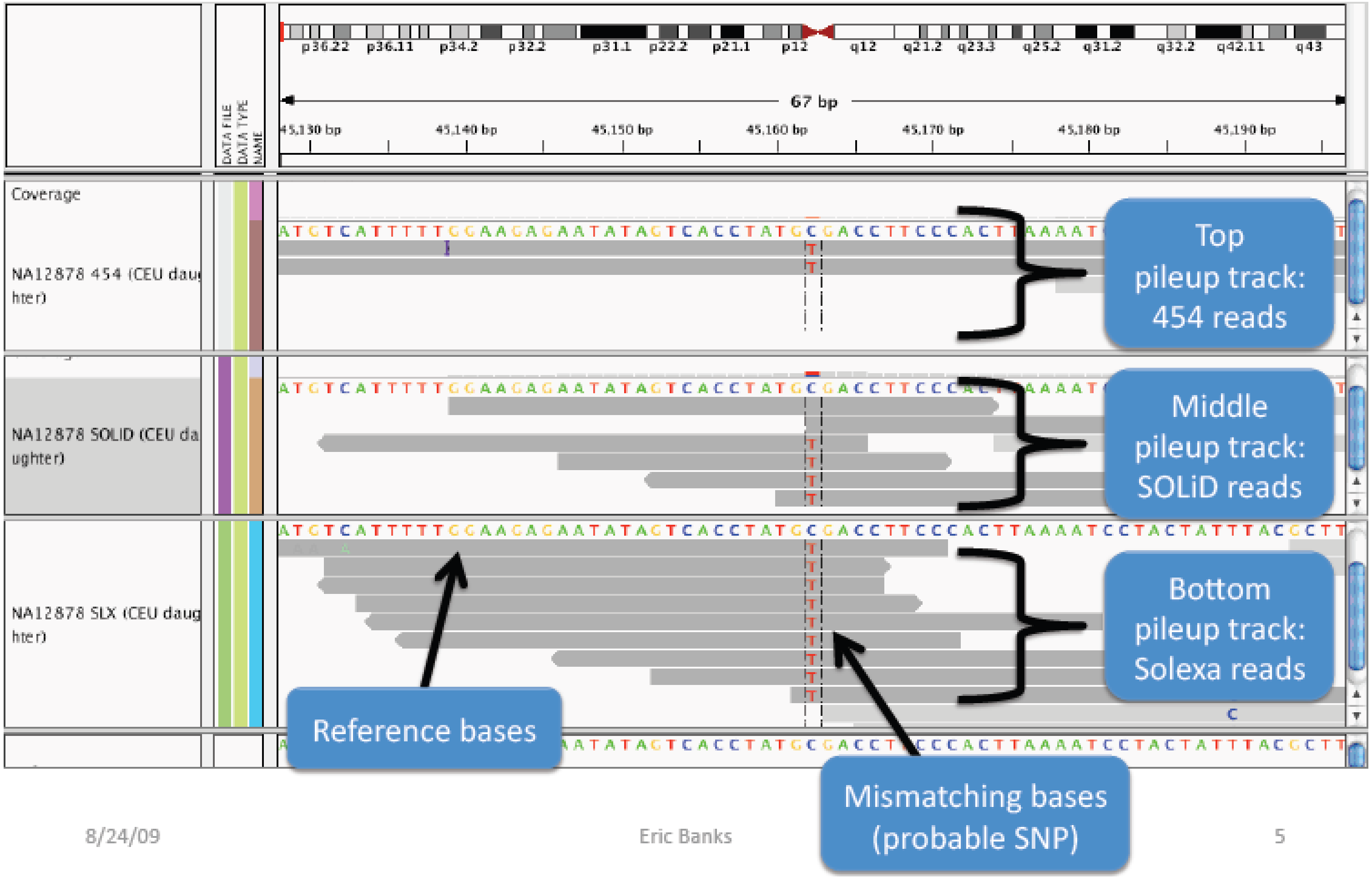
```
##format=PCFv1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
#CHROM  POS     ID        REF   ALT    QUAL  FILTER  INFO                                FORMAT       NA00001         NA00002         NA00003
20      14370   rs6054257 G     A      29    0       NS=58;DP=258;AF=0.786;DB;H2         GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5
20      13330   .         T     A      3     q10     NS=55;DP=202;AF=0.024               GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3    0/0:41:3
20      1110696 rs6040355 A     G,T    67    0       NS=55;DP=276;AF=0.421,0.579;AA=T;DB GT:GQ:DP:HQ  1|2:21:6:23,27  2|1:2:0:18,2    2/2:35:4
20      10237   .         T     .      47    0       NS=57;DP=257;AA=T                   GT:GQ:DP:HQ  0|0:54:7:56,60  0|0:48:4:51,51  0/0:61:2
20      123456  microsat1 G     D4,IGA 50    0       NS=55;DP=250;AA=G                   GT:GQ:DP     0/1:35:4        0/2:17:2        1/1:40:3
```

www.1000genomes.org/wiki/doku.php?id=1000_genomes:analysis:vcfv3.2

www.1000genomes.org

ftp.1000genomes.ebi.ac.uk

Read-scale view


p36.22 p36.11 p34.2 p32.2 p31.1 p22.2 p21.1 p12 q12 q21.2 q23.3 q25.2 q31.2 q32.2 q42.11 q43
67 bp
45,130 bp 45,140 bp 45,150 bp 45,160 bp 45,170 bp 45,180 bp 45,190 bp
DATA FILE
DATA TYPE
NAME
Coverage
NA12878 454 (CEU daughter)
NA12878 SOLID (CEU daughter)
NA12878 SLX (CEU daughter)
ATGTCATTTTTGGAAGAGAATATAGTCACCTATGCGACCTTCCCACTTAAAATCCTACTATTTACGCTT
T

Top pileup track: 454 reads

Middle pileup track: SOLiD reads

Bottom pileup track: Solexa reads

Reference bases

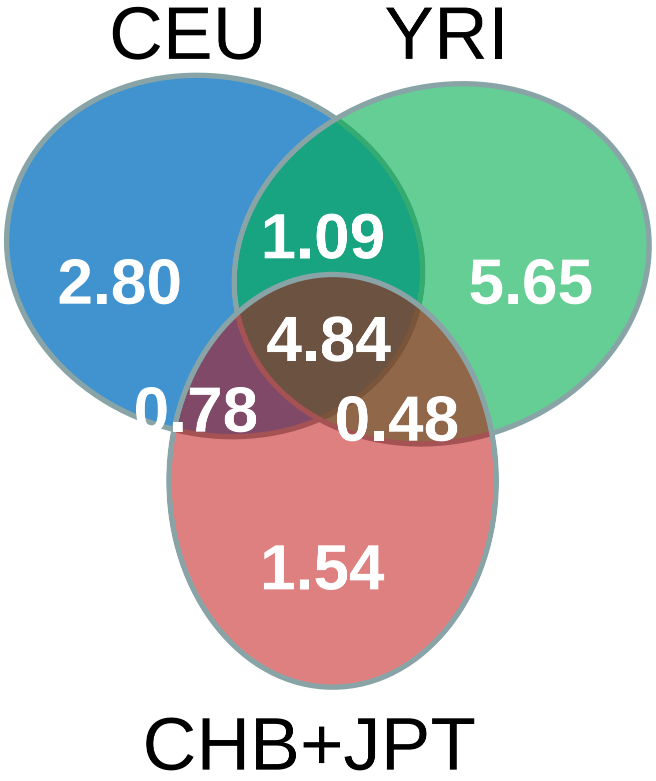
Mismatching bases (probable SNP)

Eric Banks (Broad)

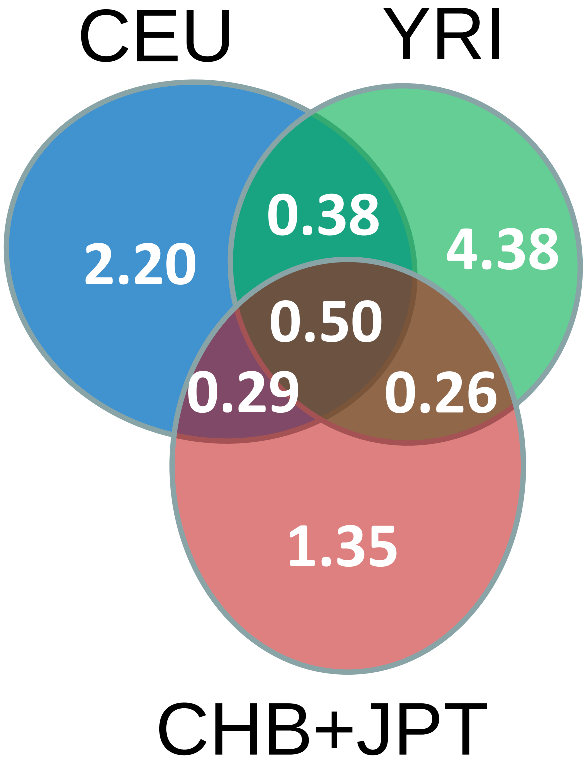
Genome wide SNP discovery

- Total 17.2 M SNPs called
- Previously ~12M SNPs "known" (dbSNP 129)
  - 7.9M confirmed
  - 9.2M novel

Total SNPs

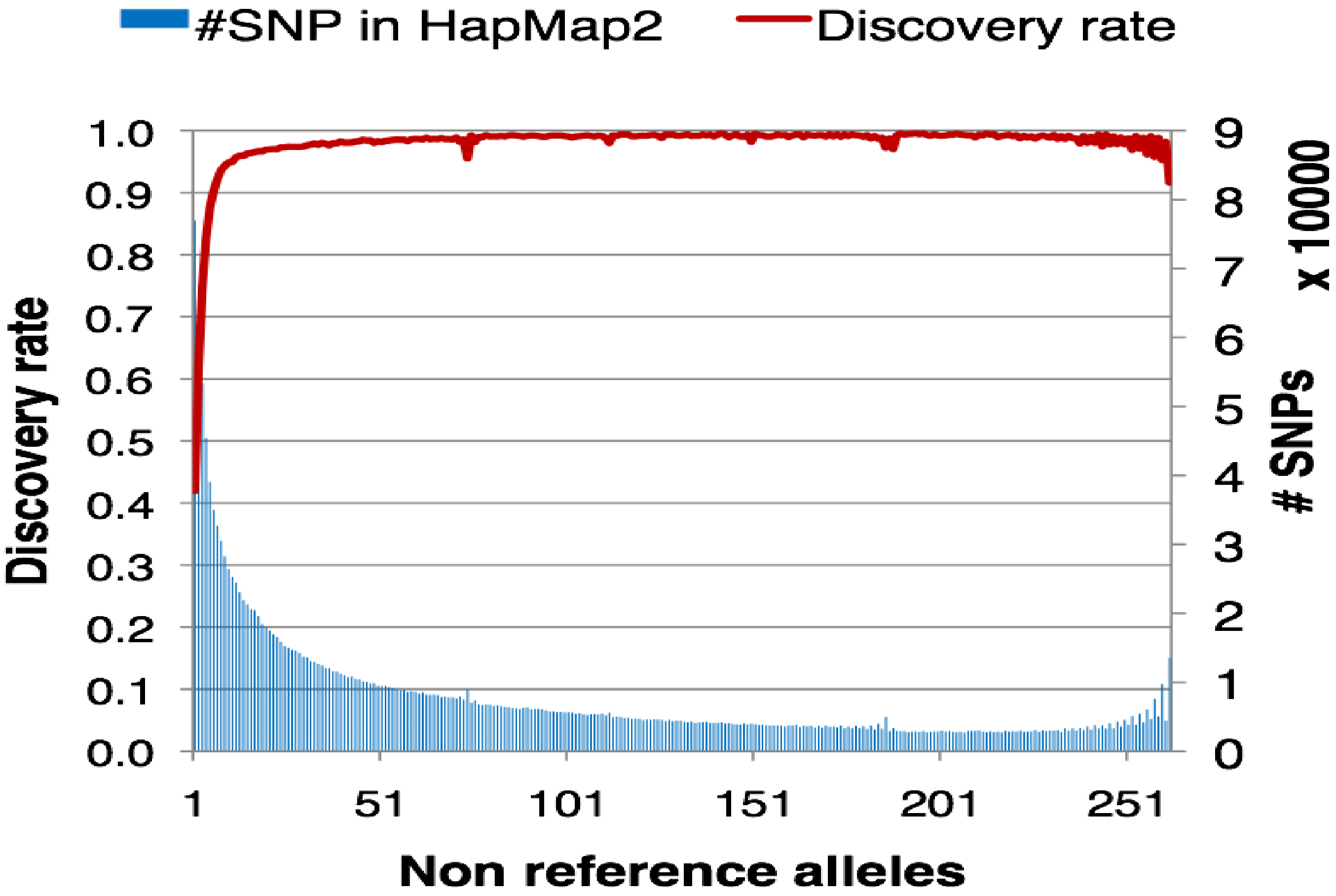


Novel SNPs



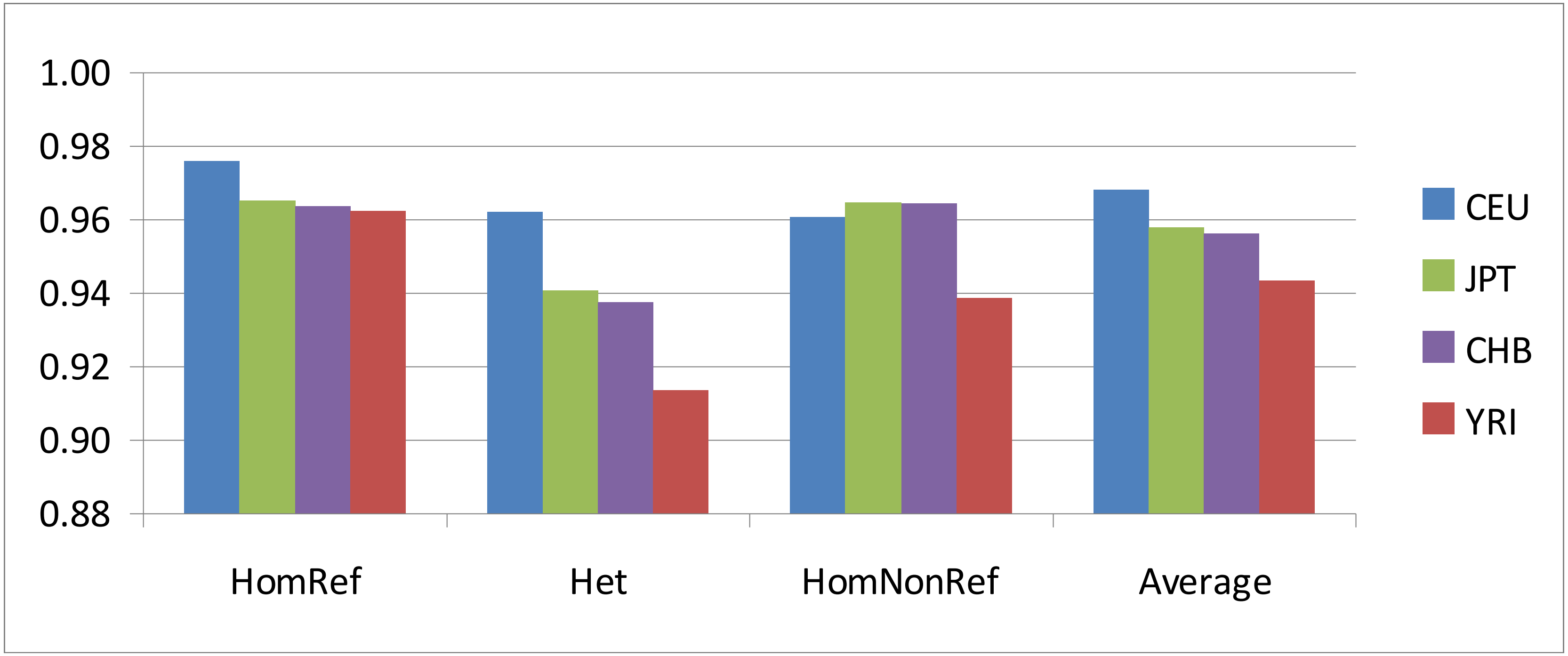
Le Quang

Completeness



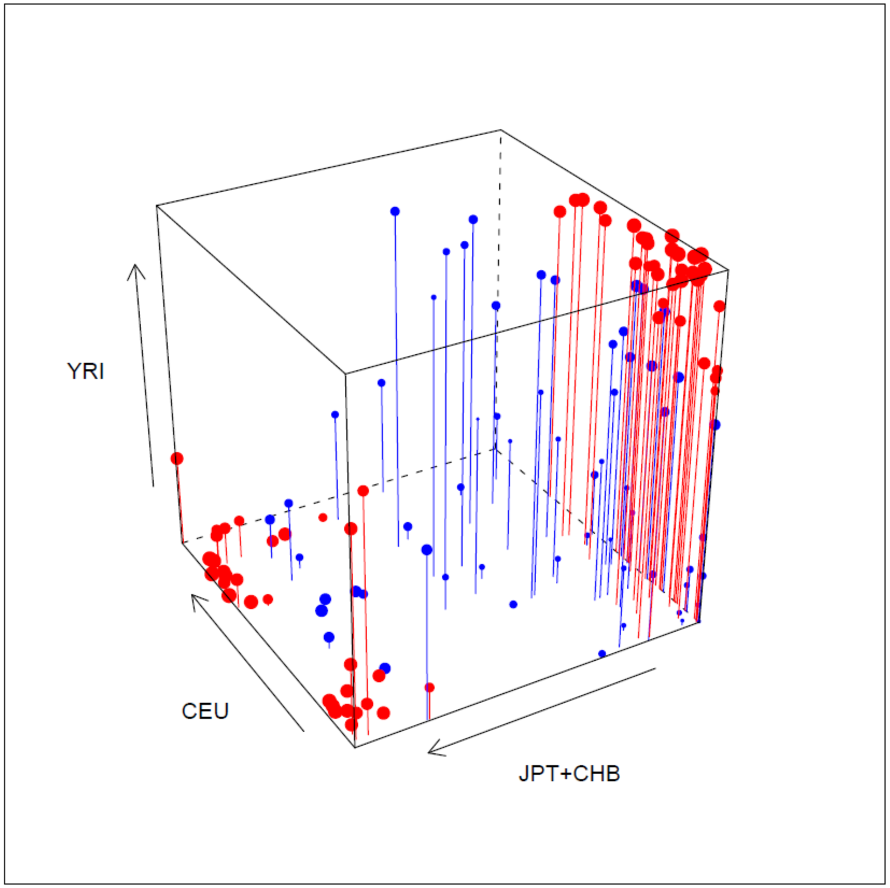
Durbin, Le Quang

Genotype accuracy on HapMap2



- This is about where simulations suggest we should be with 2-4x on 60 samples
- Much higher than independent calls

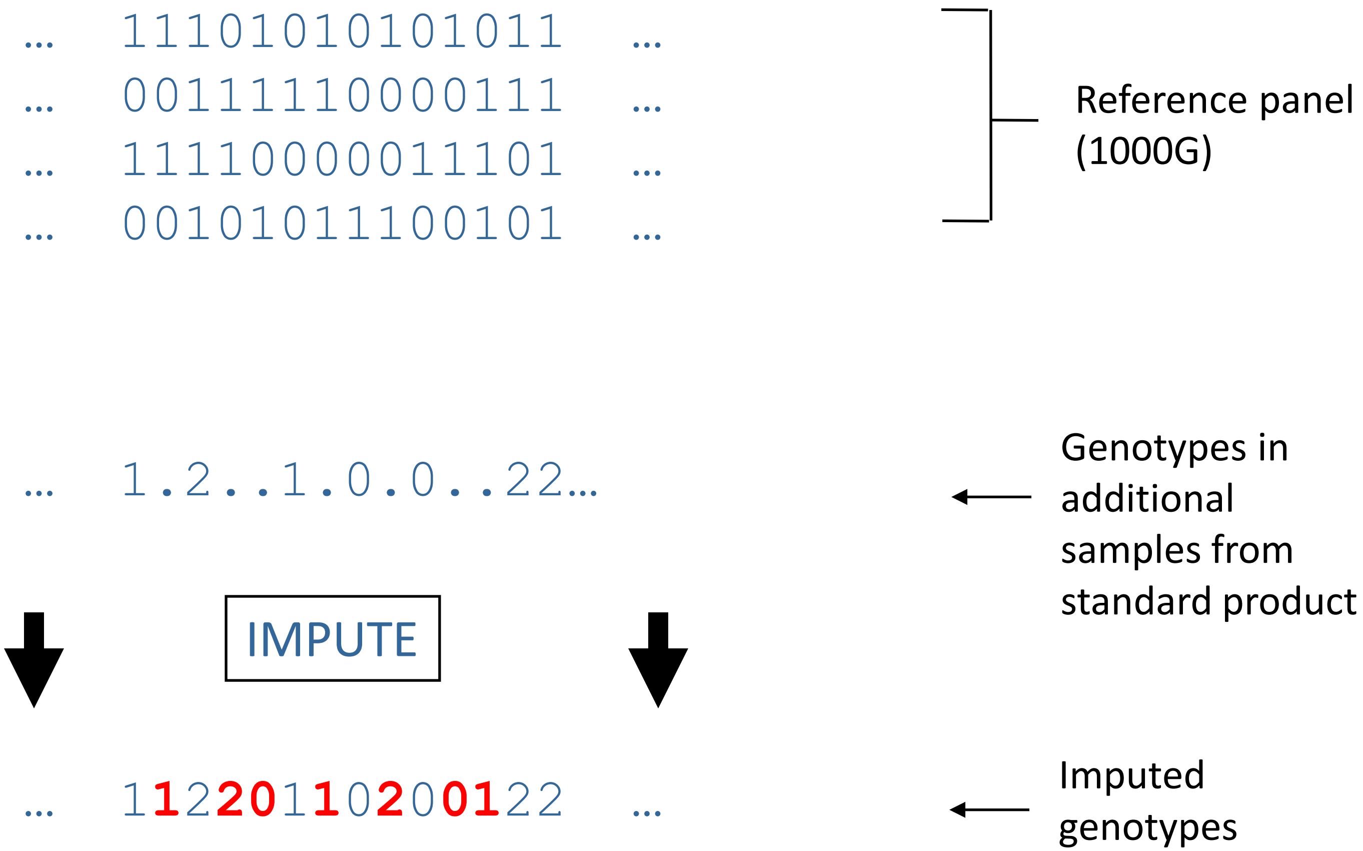
Some surprises – high Fst SNPs



Ryan Hernandez, Adam Auton

Using the 1000G data now

Imputation



… 11101010101011 …

… 00111110000111 …

… 11110000011101 …

… 00101011100101 …

Reference panel (1000G)

… 1.2..1.0.0..22…

Genotypes in additional samples from standard product

IMPUTE

… 1**1**2**20**1**1**0**2**0**01**22 …

Imputed genotypes

What can we use imputation for in GWAS/fine-mapping?

- To
  - Help define genomic regions likely to contain causal variants
  - Define a small set of SNPs to take through to additional genotyping?
  - Fine-map?

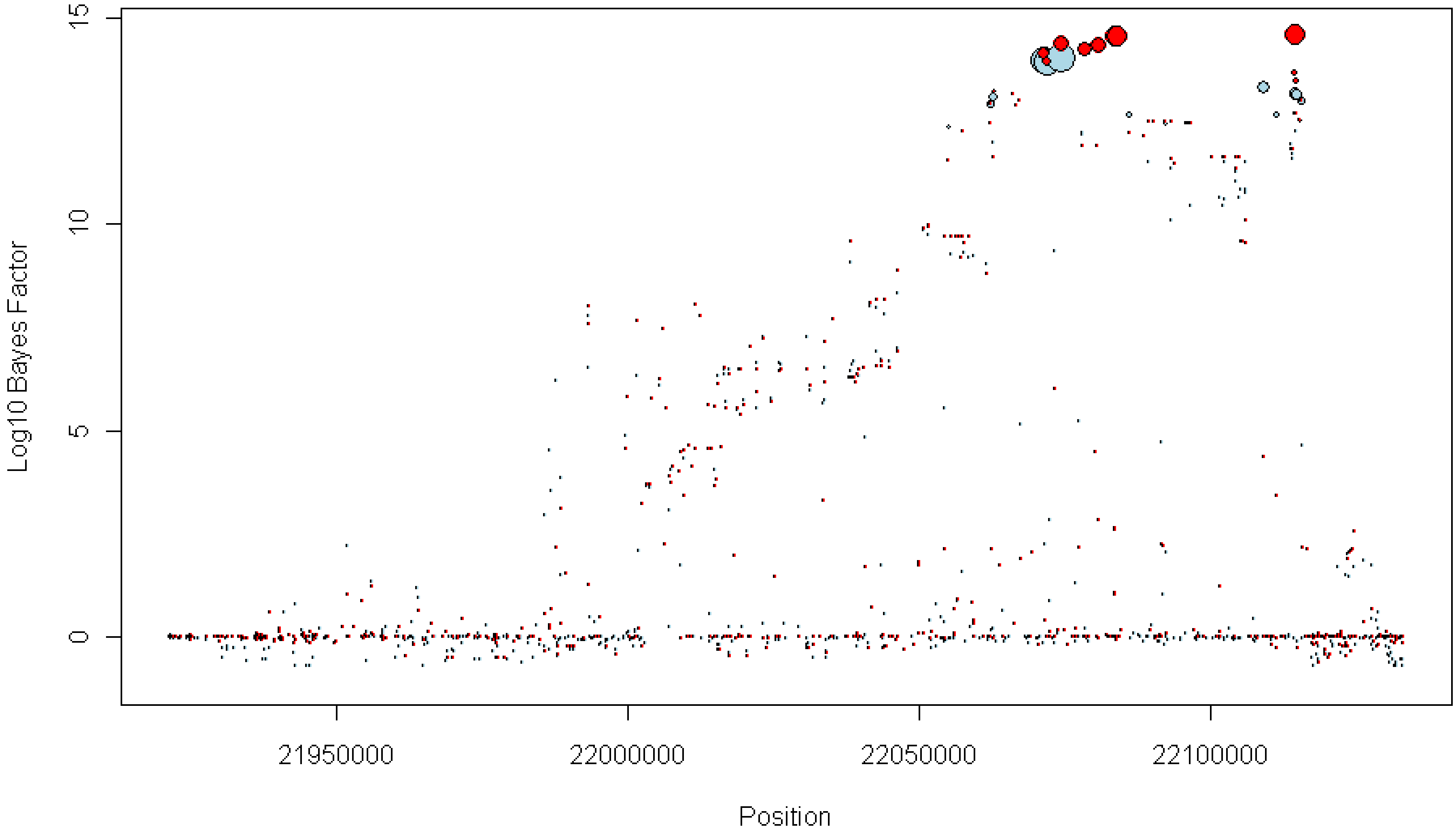
- Accuracy depends on completeness of imputation source and accuracy of imputed genotypes

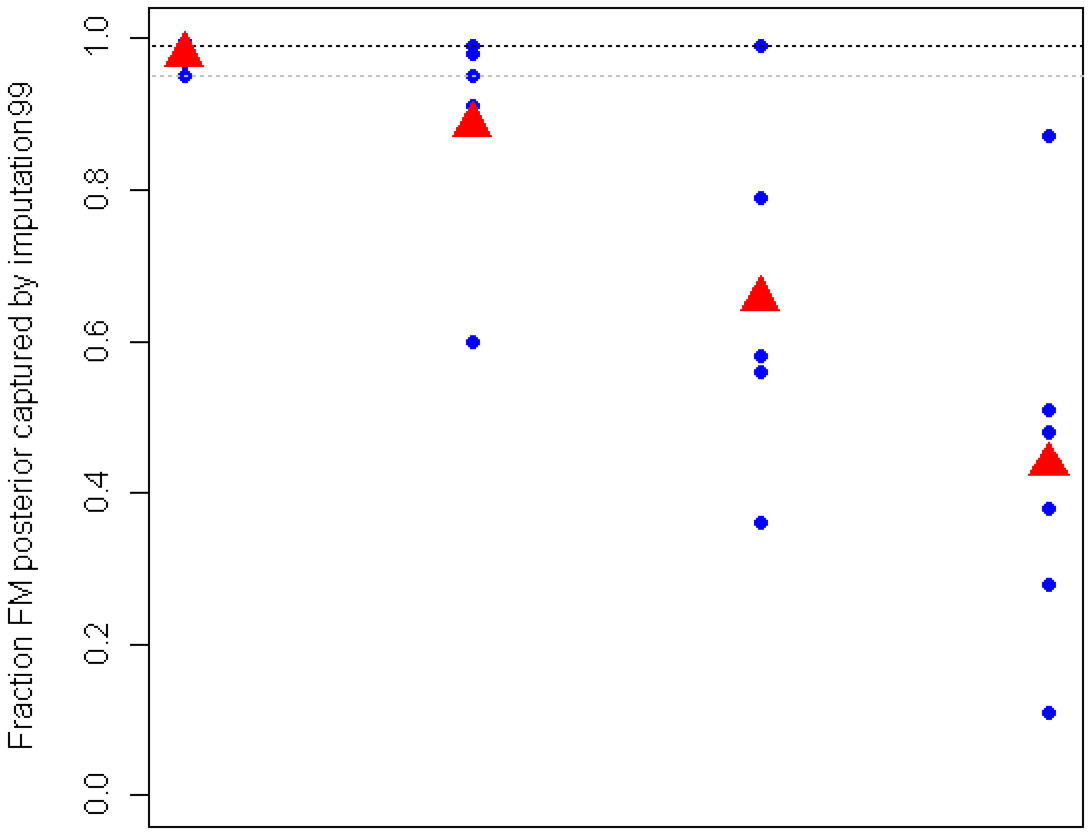
Imputation resources

Data source	Type	Number haplotypes	Number SNPs	Includes
HapMap2	GT	120	671	Mainly common SNPs, phased from trios
Reseq	RS	64	1543	SNPs found from resequencing. Genotypes called independently
FM panel	GT	64	1699	All SNPs from RS pilot and dbSNP for which design possible
1000 Genomes pilot (CEU)	RS	114	2561	SNPs found from RS, integrated into haplotype structure of HM3 SNPs

How can we measure imputation?

CAD: CDKN2A/2B





FM panel

1000 Genomes

Reseq panel

HapMap2

How good is imputation?

- Imputation is never going to be as convincing as genotyping

- BUT it is sufficiently accurate, at least for common variants, to
  - Define sets of SNPs of interest
  - Exclude SNPs
  - Indicate whether the signal is likely to be localised through additional genotyping (a function of power and haplotype structure)

- The completeness of the 1000G project data is extremely valuable
  - The WTCCC decided to stop any further sequencing of controls or cases

Open questions

- How reliably can you detect structural variation from low-coverage data?
- How do you combine information across populations?
- Can we generate reliable de novo assemblies from the data?

Acknowledgements

- Oxford
  - Adam Auton, Zam Iqbal, Gerton Lunter, Jules Maller, Simon Myers, Jonathan Marchini, Peter Donnelly
- The Wellcome Trust Case Control Consortium
- The 1000 Genomes Project