The UK10K Cohorts Project: Rare variant analysis by whole genome sequencing in 3,621 samples

Klaudia Walter

Wellcome Trust Sanger Institute

11-03-2016

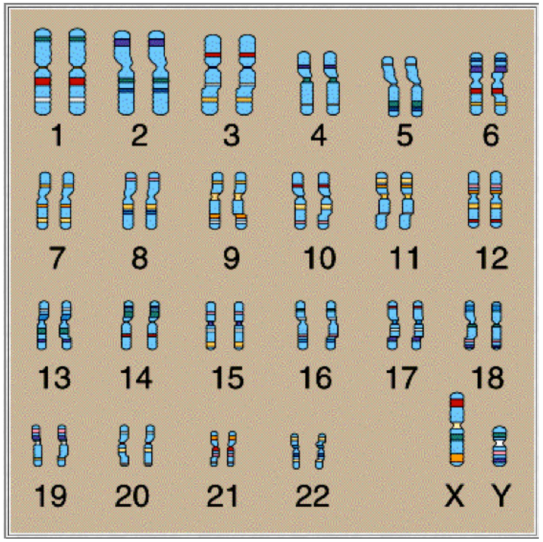
Outline

Overview

- Whole genome sequencing produces a lot of data
- High-coverage exome versus low-coverage whole genome sequencing
- Structure and aims of the UK10K Project

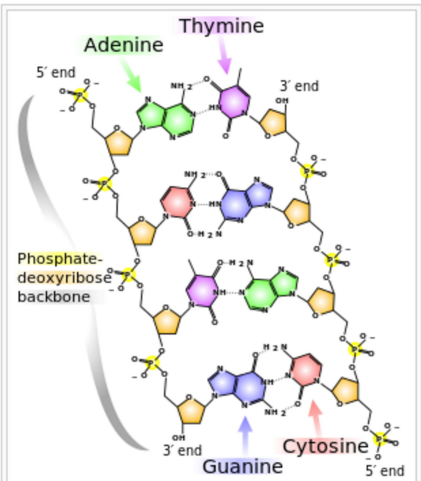
The Human Genome

Chromosomes



Double-Helix

Nucleotides
A, C, G, T

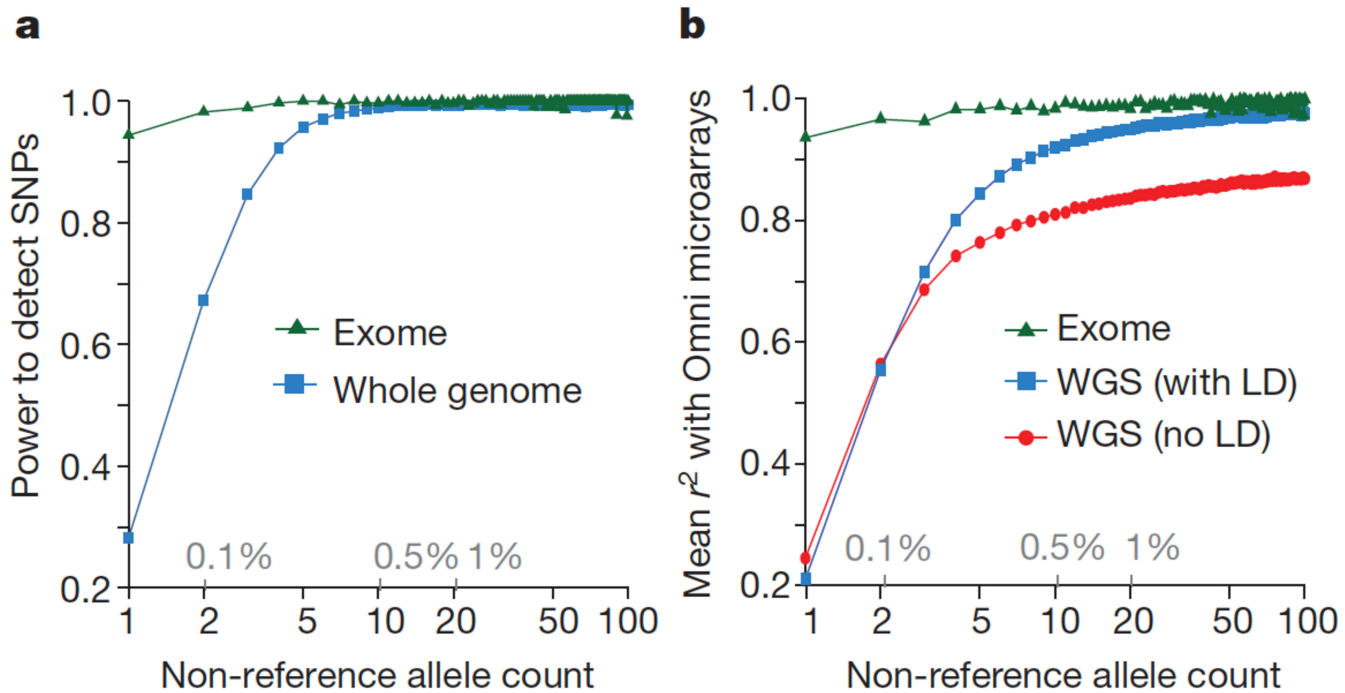


Chemical structure of DNA. Hydrogen bonds shown as dotted lines.

Large-scale re-sequencing in complex disease

Motivation

- Chip-based GWAS do not access low frequencies well
- 1000 Genomes Project is discovering most common and many low frequency/rare alleles but these are difficult to impute
- Evidence already exists that rare variants associate with disease



The 1000 Genomes Project Consortium, Nature 2012

UK10K: 10,000 UK Genomes (2010-2013)

Design

- 10.4M GBP strategic award grant by the Wellcome Trust
- 164 researchers from 51 institutions
- Sequence 10,000 samples from UK and Finland

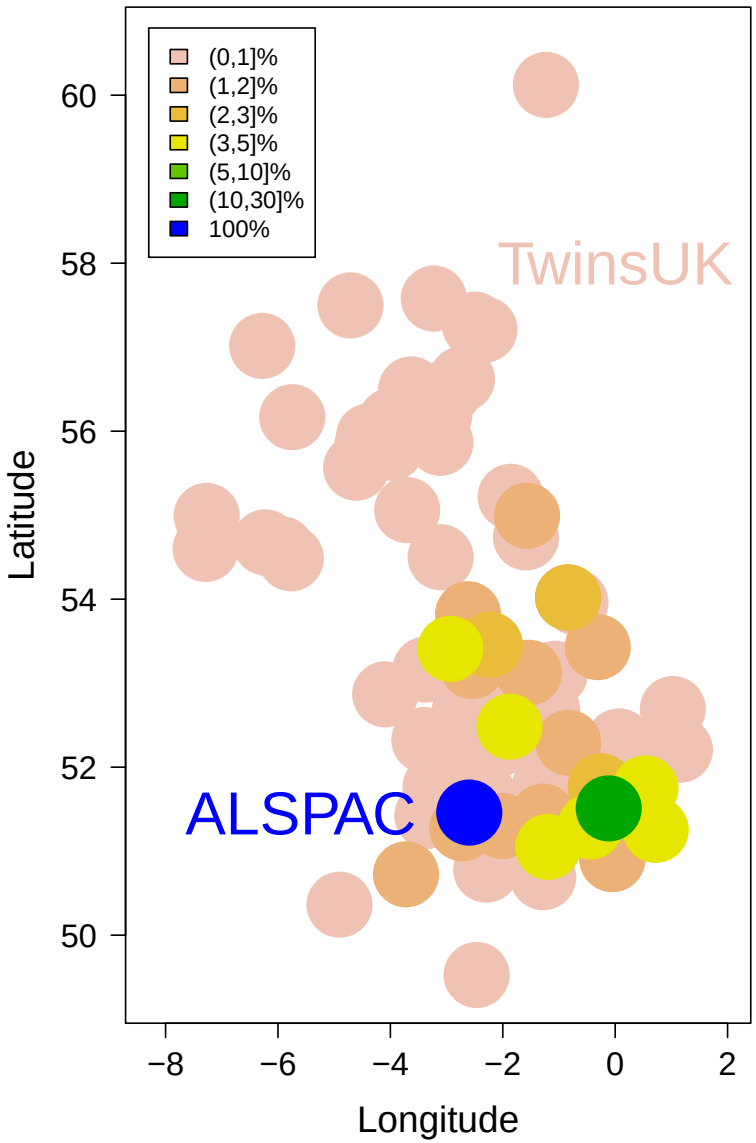
Goals

- Exhaustive discovery of rare and low frequency variants
- Direct association of sequenced samples
- Provide a sequence and phenotype variation resource for the community

Project arms

* Several distinct clinical phenotypes

UK10K cohorts design



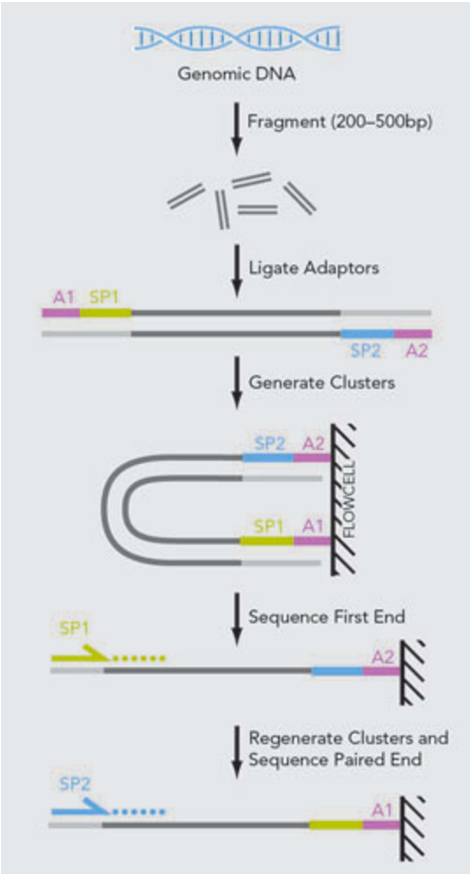
- **ALSPAC** (The Avon Longitudinal Study of Parents and Children, Bristol University)
  - Children/adolescents ($\sim$ 18 yrs)
  - Males and females
  - Geographically restricted
- **TwinsUK** (Identical and non-identical Twins, Department of Twin Research, Kings College London)
  - Adults (median age 46 yrs)
  - All females
  - UK-wide origin

Both with deep genetic and phenotype coverage (clinical, questionnaire, molecular)

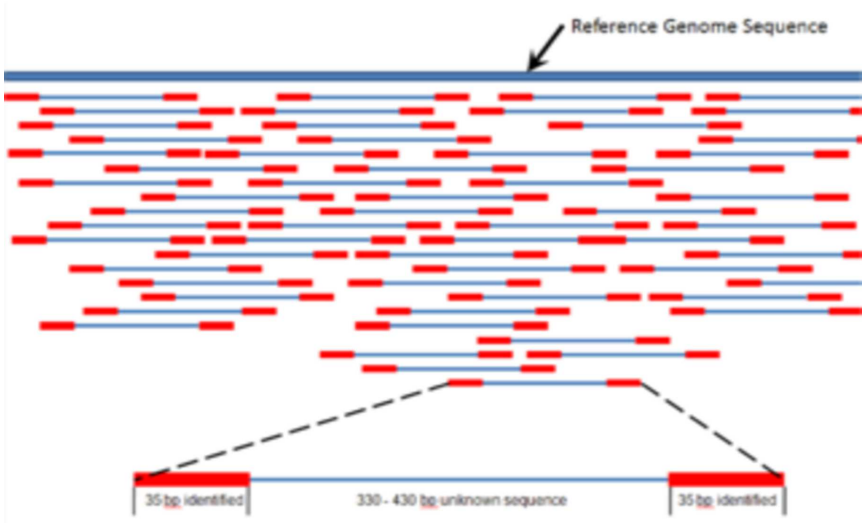
Data

- What does the sequencing data look like?
- Production pipeline
- Data formats

Short read sequencing and mapping

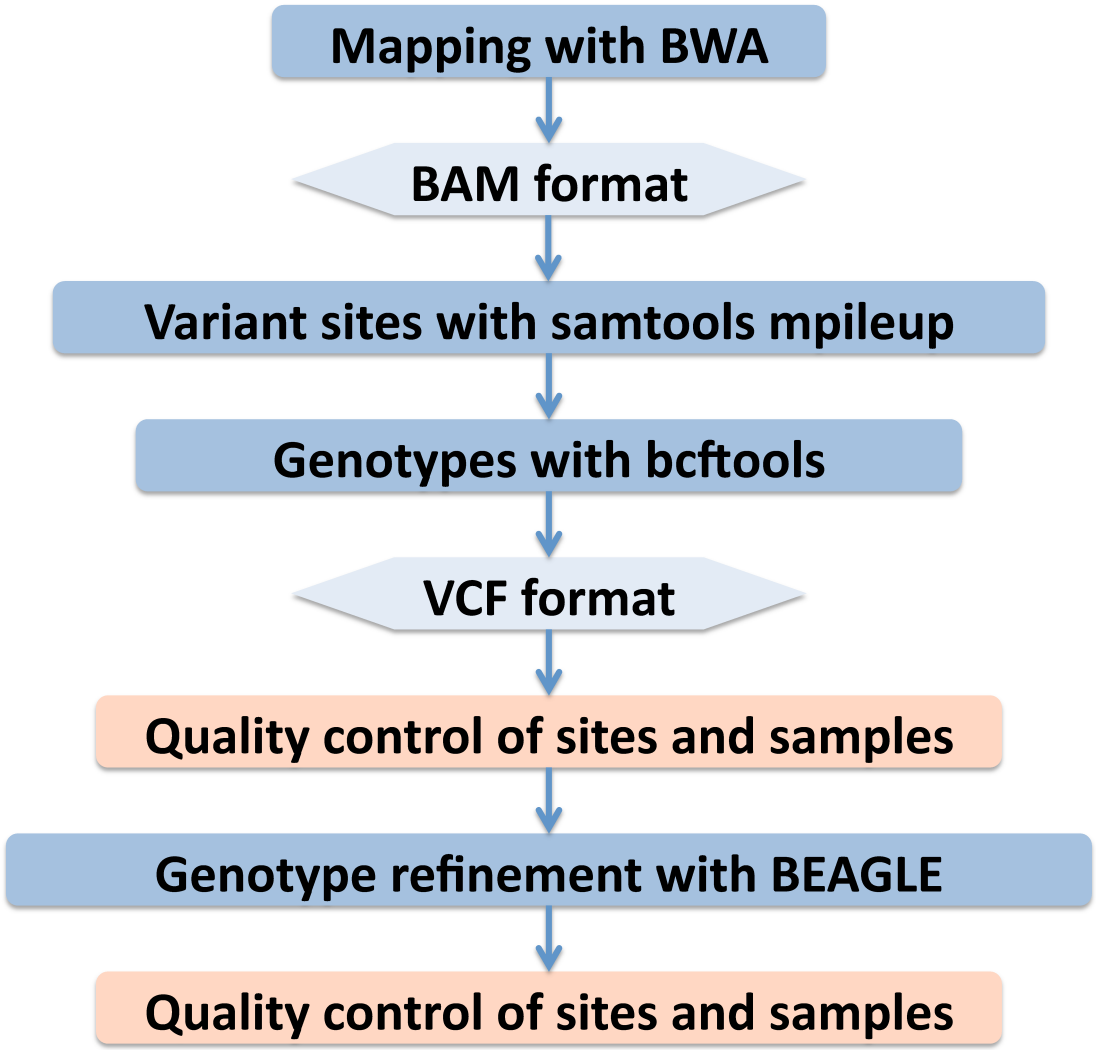


http://www.illumina.com



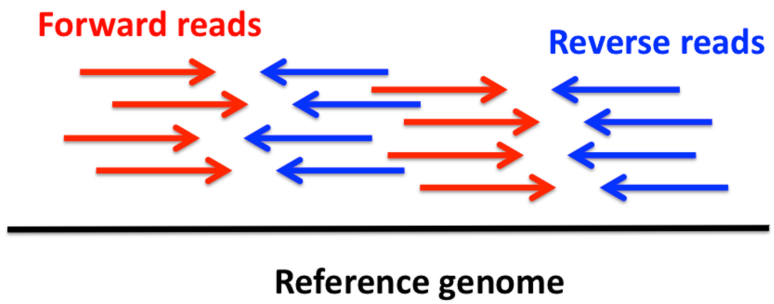
http://www.mn.uio.no/ifi/studier/masteroppgaver/-bio/benchmarking-system.html

Production pipeline



BAM format

Reads (= short sequences) are mapped against a reference genome.



ID	FLAG	CHR	POS	MAPQ	CIGAR	LEN	SEQ
HS11	99	20	2000094	60	100M	371	CCAAAAAATG
HS11	147	20	2000365	60	100M	-371	CAGAAATTGA

FLAG	
99	read paired, read mapped in proper pair, mate reverse strand, first in pair
147	read paired, read mapped in proper pair, read reverse strand, second in pair

VCF format

Variant calling format for SNVs, INDELs and structural variations

CHROM	POS	ID	REF	ALT	QUAL	FILTER
20	67184	rs189459753	C	T	999	PASS
20	67500	rs112142516	T	TTGGTATCTAG	999	PASS

INFO
DP=18784;AN=4864;AC=21;ICF=-0.00434;HWE=1.000000
DP=14657;INDEL;AN=4864;AC=3785;ICF=0.01506;HWE=0.445674

FORMAT	QTL190044
GT:DP:GL	0 | 0:6:0.00,-12.00,-12.00
GT:DP:GL	1/0:8:-12.00,0.00,-12.00

Final UK10K Cohorts data release (REL-2012-06-02)

	Allele frequency	REL-2012-06-02
Number of samples		3,781
TwinsUK		1,854
ALSPAC		1,927
Number of SNVs		42,001,210
Number of INDELs		3,490,825
SNVs by MAF	AF $<$ 1%	34,247,969
	1% $\leq$ AF $\leq$ 5%	2,298,220
	AF $>$ 5%	5,869,317
Number of large deletions		18,739
SNVs per sample		3,222,597
Singletons per sample		5,370
Read depth		7x
Data size		660Gb

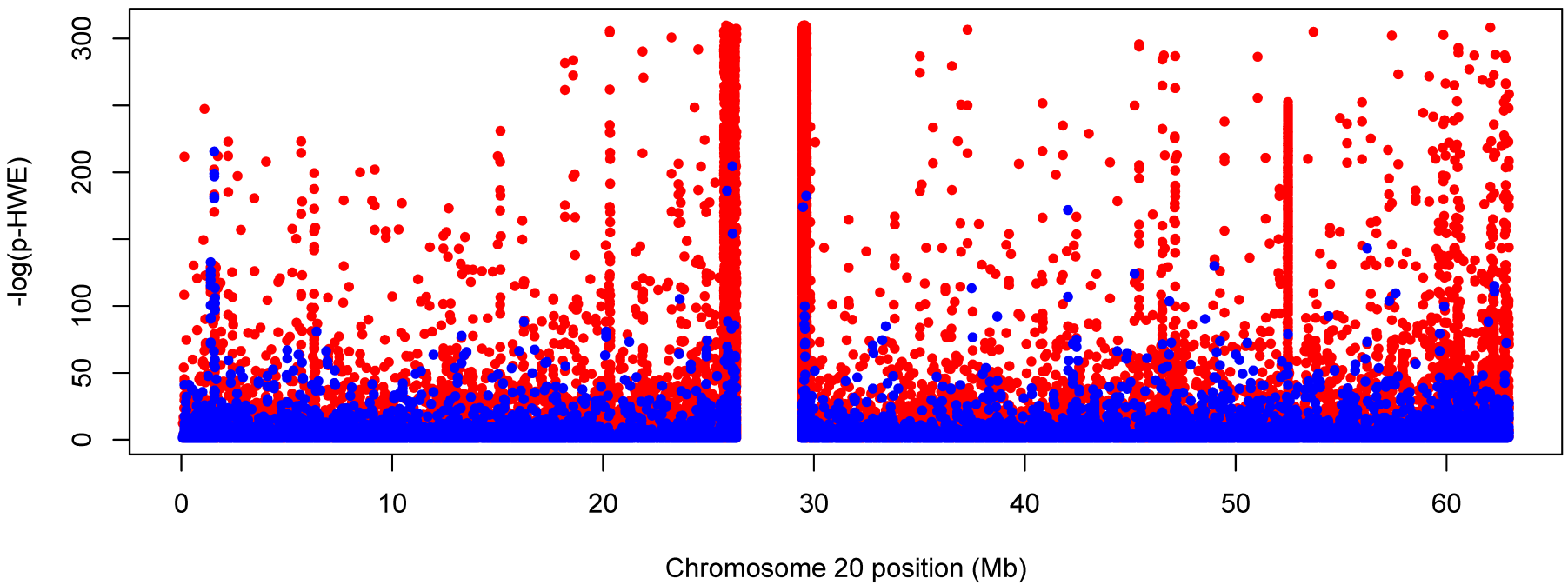
Quality control of sites

- Read depth and HWE (Hardy-Weinberg Equilibrium)
- VQSR - Variant Quality Score Recalibration
- Sites shared with 1000GP
- Batch effects

VQSR - Variant Quality Score Recalibration

- Assigns a well-calibrated probability to each variant call
- Uses SNV call annotations such as DP and MQ
- Trained against "true" sites such as HapMap 3
- VQSLOD in INFO field (log odds ratio)
- Filter based on this single estimate
- Developed at the Broad Institute

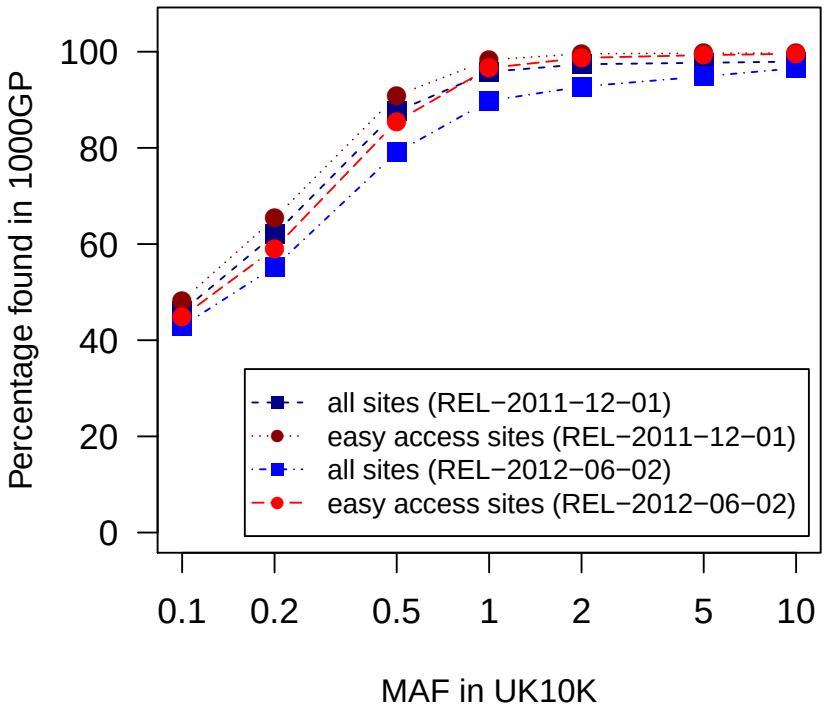
Filtering by VQSR versus by HWE *p*-values



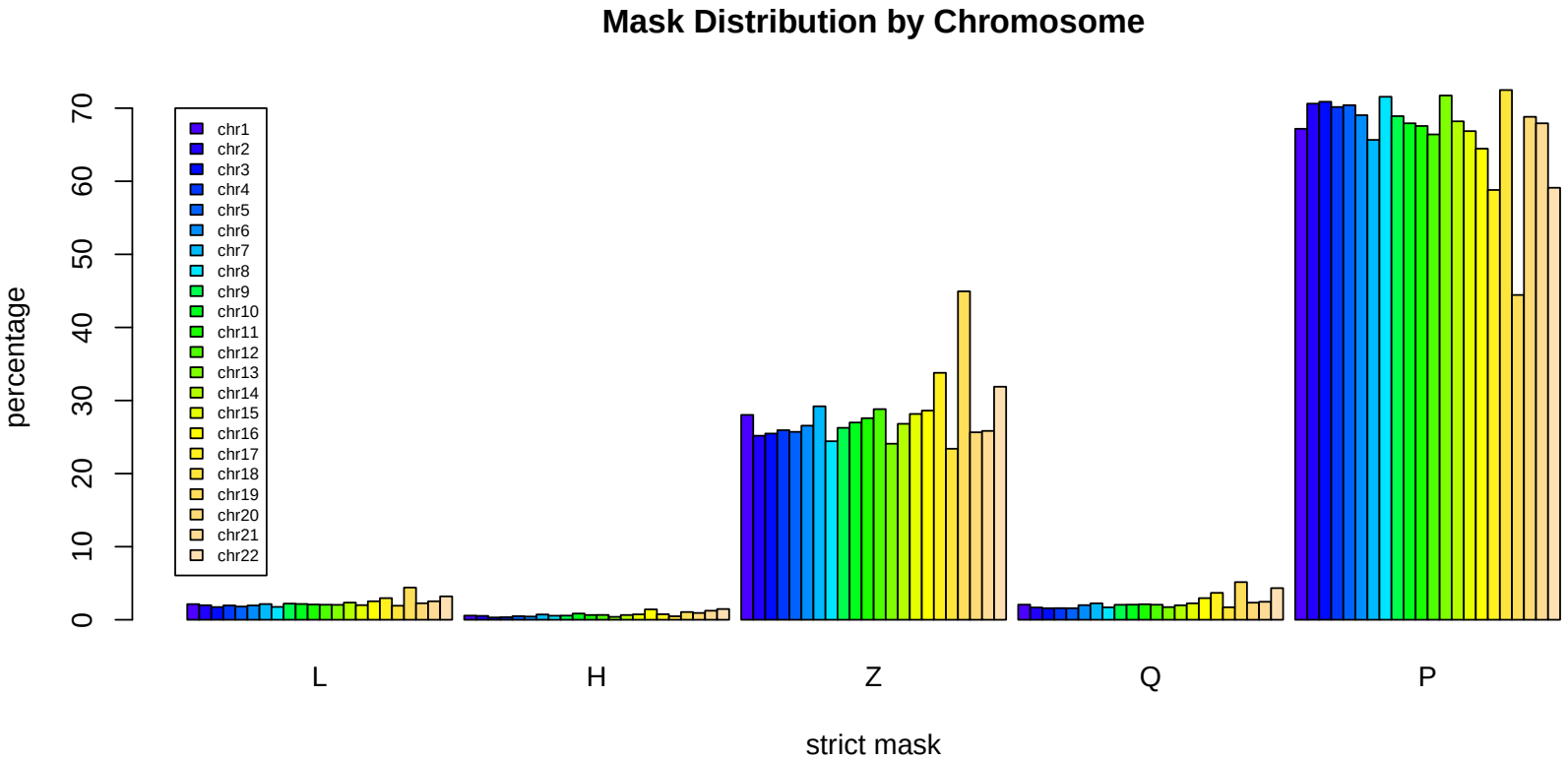
Filtering by VQSR removes most of the sites with extremely low HWE *p*-values.

Percentage of sites of UK10K in 1000GP

	2011-12-01	2012-06-02
MAF	overlap(%)	overlap(%)
0.1	46.0	42.9
0.2	62.0	55.2
0.5	87.6	79.1
1.0	95.8	89.8
2.0	97.4	92.7
5.0	97.7	94.9
10.0	98.0	96.6



Genome mask

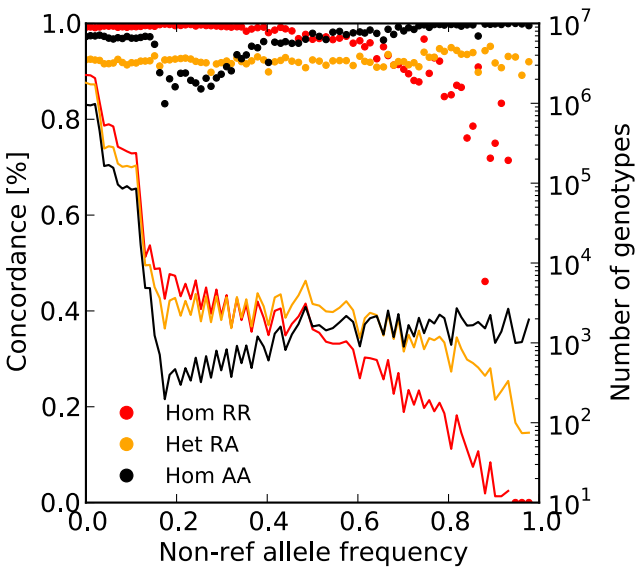


N	the base is an N in the reference genome GRCh37
L	depth of coverage is much lower than average
H	depth of coverage is much higher than average
Z	too many reads with zero mapping quality overlap this position
Q	the average mapping quality at the position is too low
P	the base passed all filters

Phase-aware genotype refinement

Genotype discordance by AF

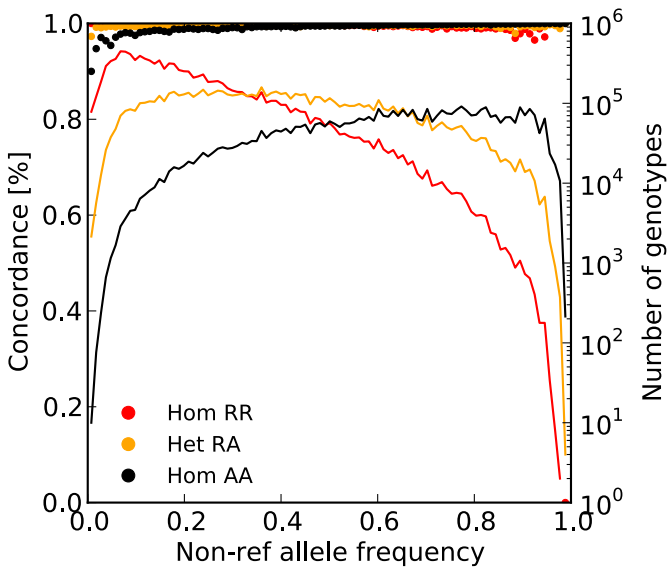
REF/REF	REF/ALT	ALT/ALT	NRD
0.76%	7.73%	2.69%	6.55%



Calculate genotype likelihoods per sample from sequence data

Genotype discordance by AF

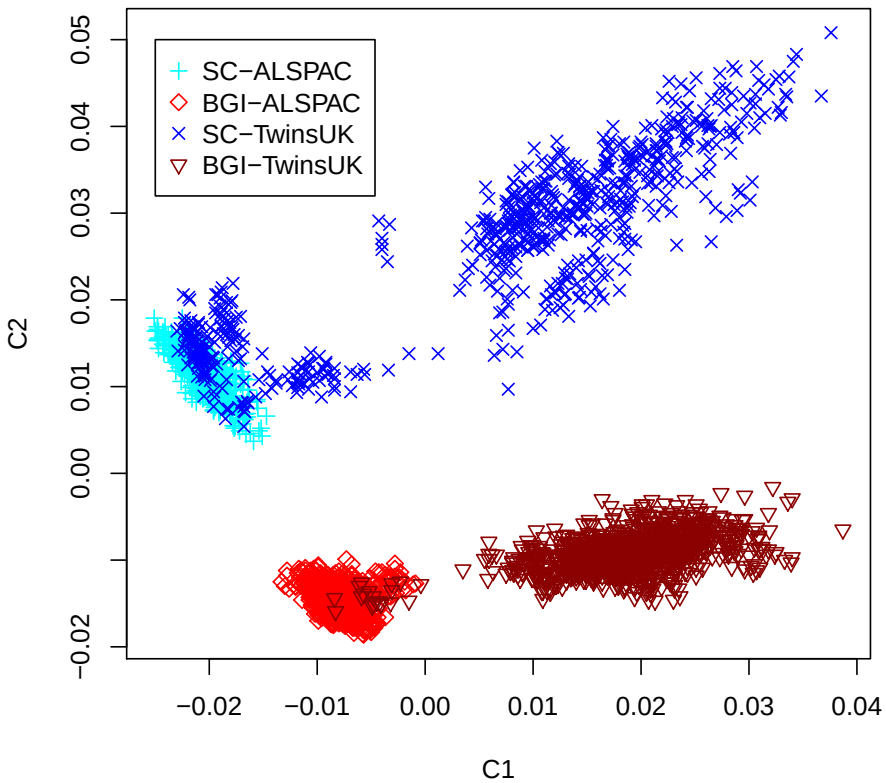
REF/REF	REF/ALT	ALT/ALT	NRD
0.14%	0.49%	0.41%	0.59%



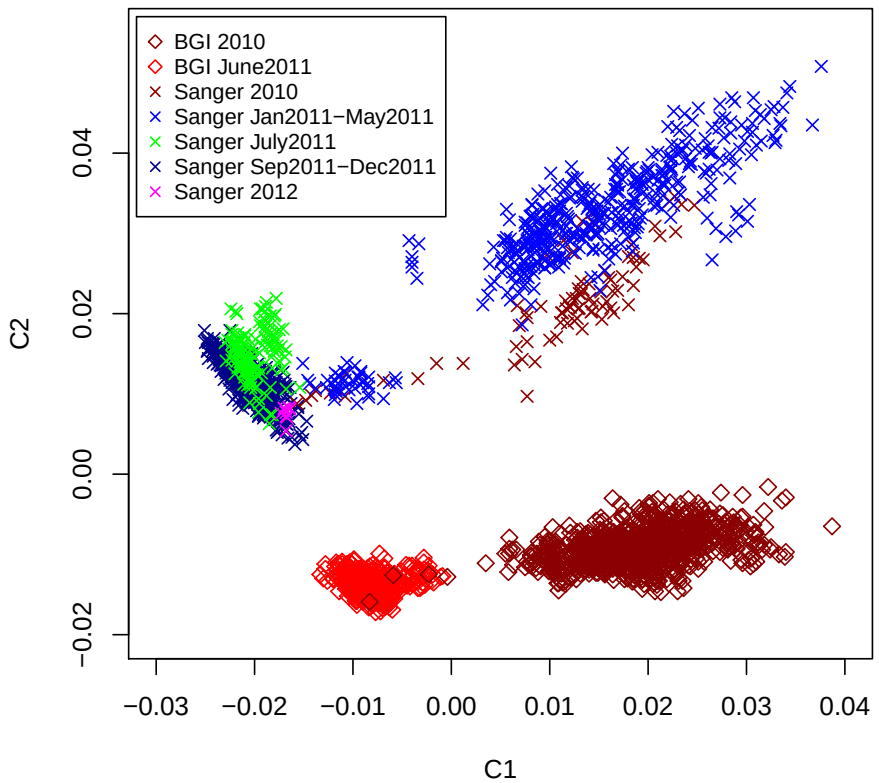
Use imputation based methods (BEAGLE, IMPUTE2) to implicitly share data across samples which share haplotypes

Batch effect Sanger versus BGI

By centre and cohort

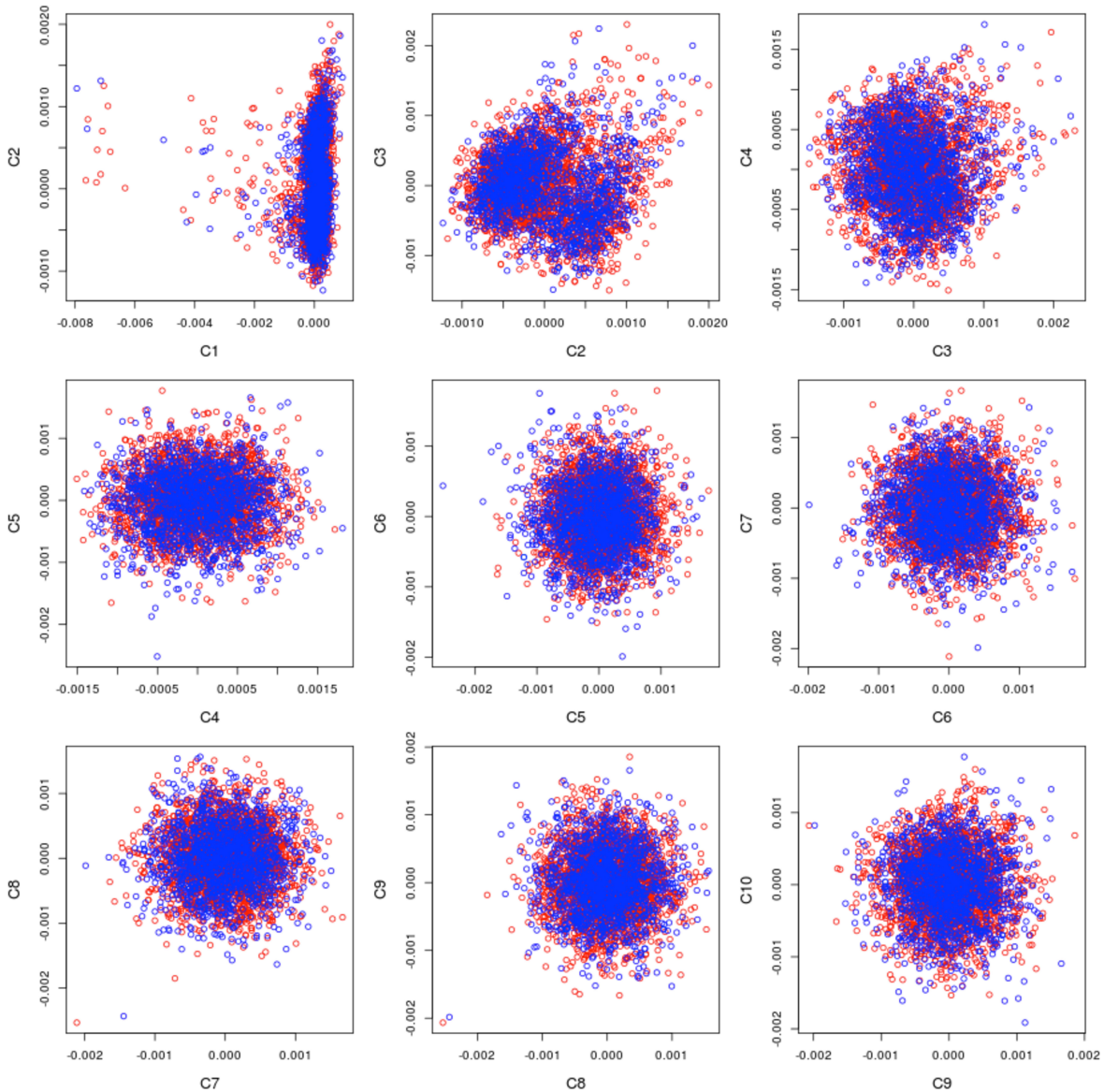


By centre and date



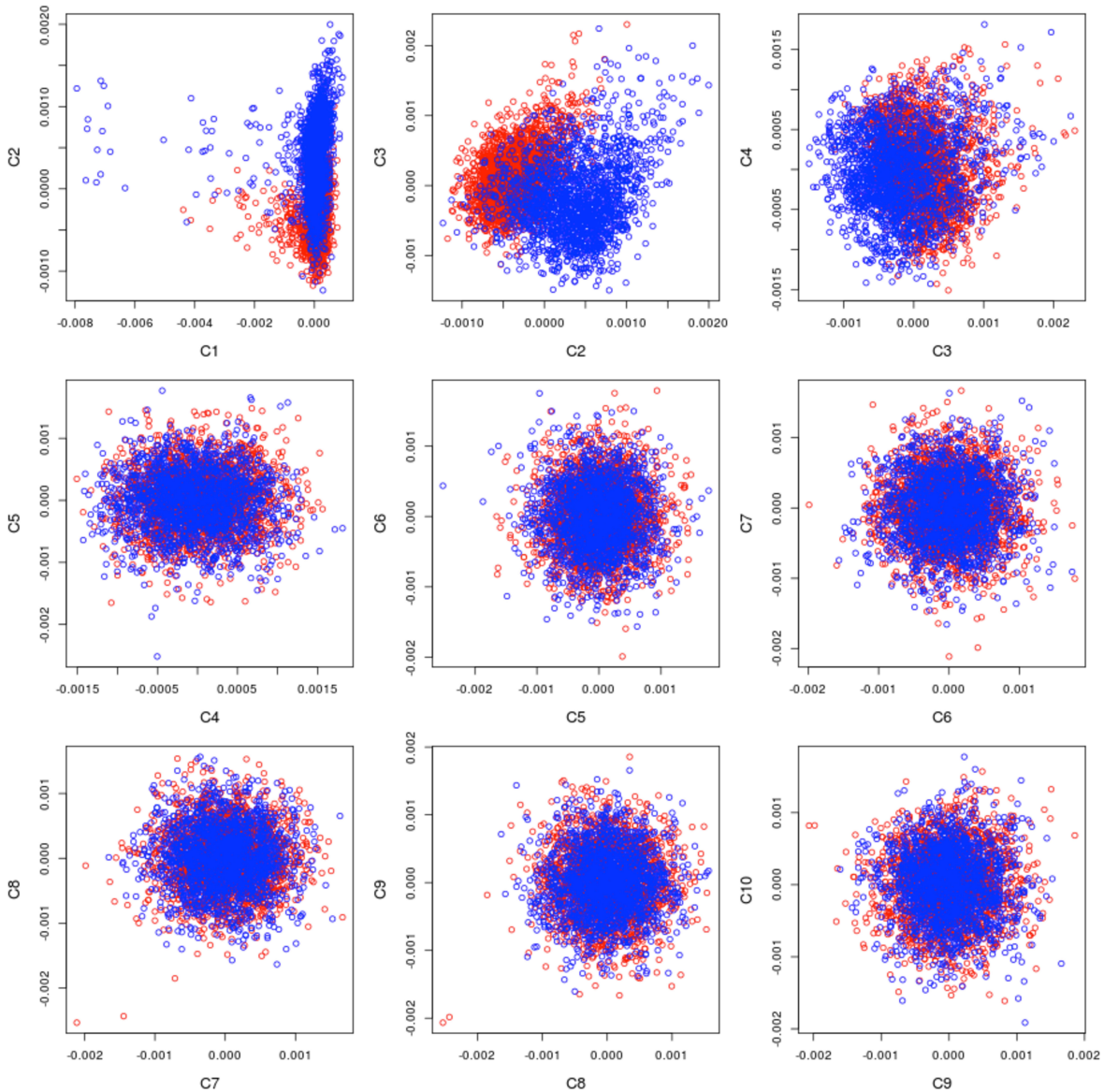
After correcting for Sanger/BGI batch effect

Sequencing centre effect



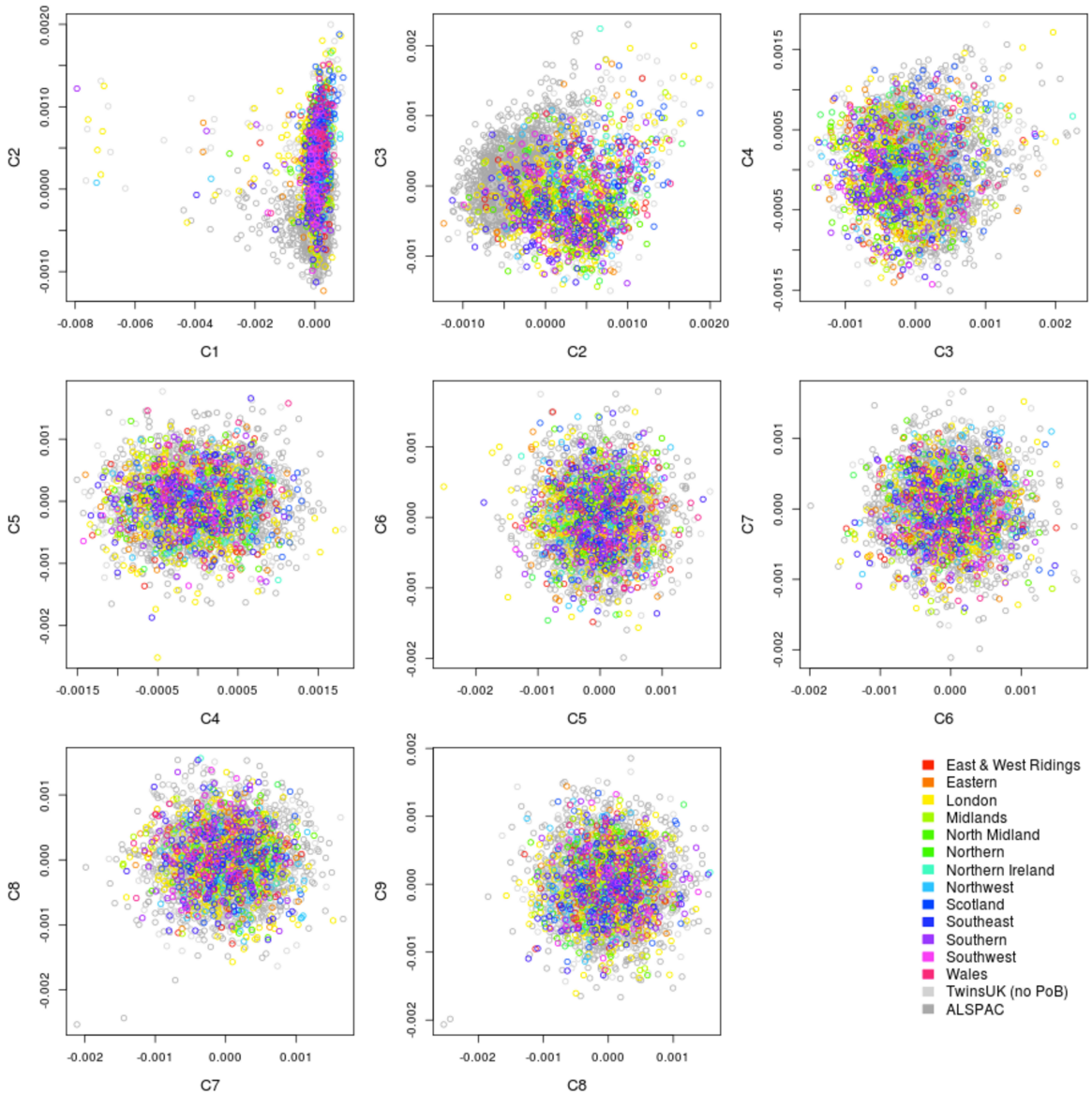
After correcting for Sanger/BGI batch effect

Cohorts effect

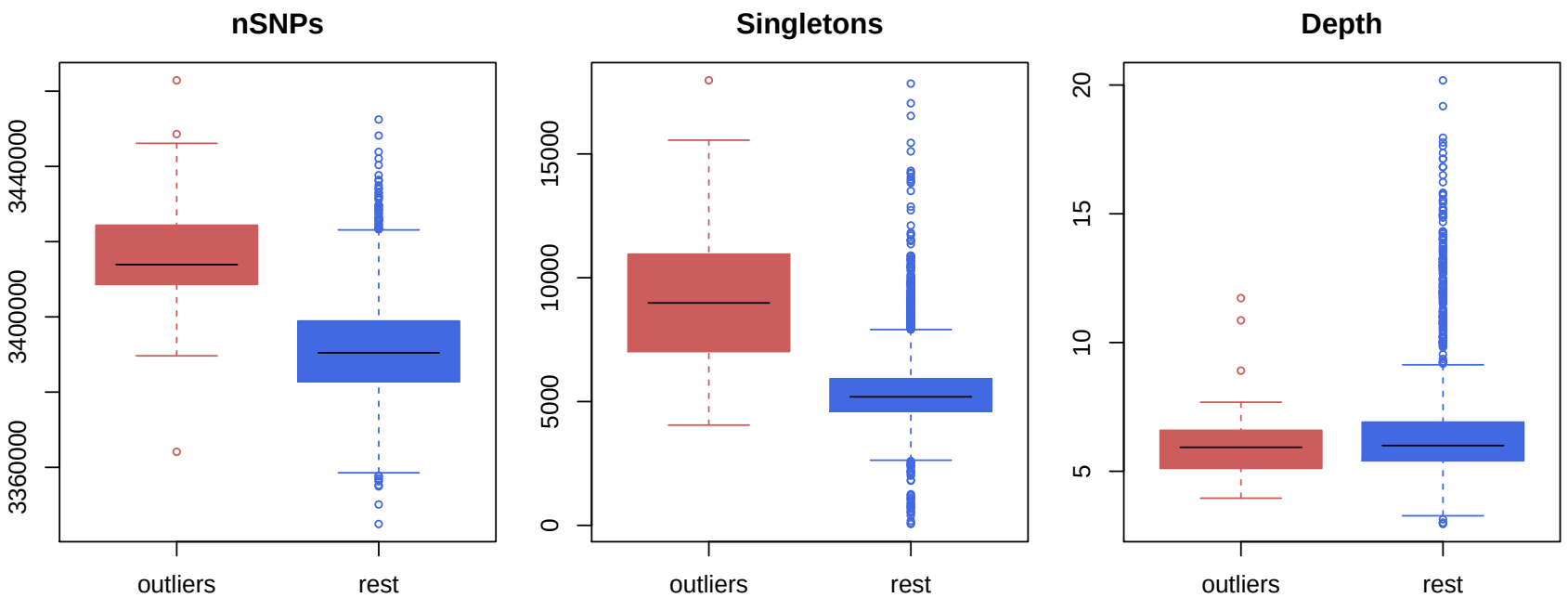


After correcting for Sanger/BGI batch effect

Population structure

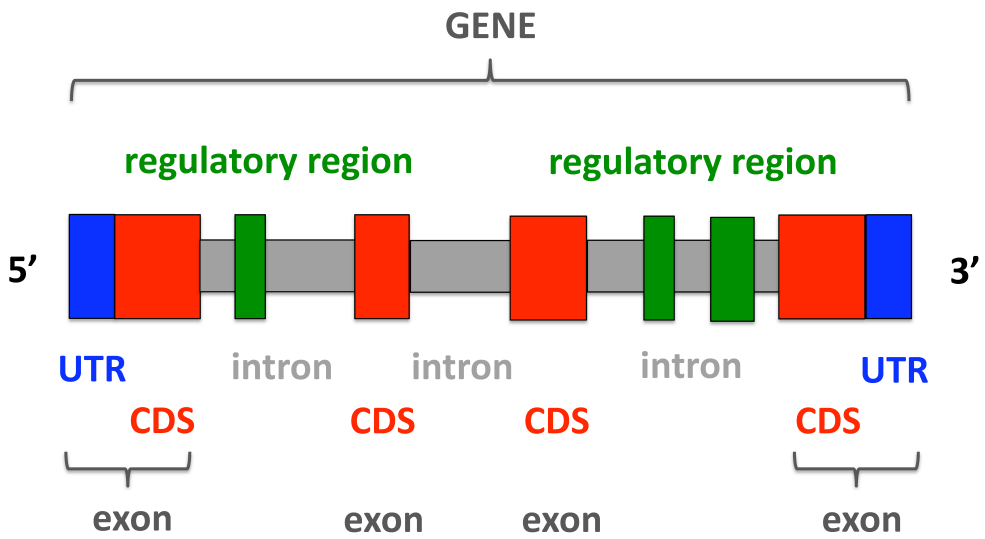


Outlier characteristics



Comparison with exome data

- There are 142 samples with exome and low-coverage genotypes (REL-2011-12-01)
- Chr20 was selected for genotype comparison
- 3433 sites and 61 samples in common with low-coverage
- Overall genotype discordance is 0.5%



CDS = coding sequence
UTR = untranslated region

Comparison of low-coverage with exome genotypes

Discovery		Exome			
		HomRef	Het	HomAlt	N/A
LC	HomRef	166936	660	26	3915
	Het	151	22910	196	881
	HomAlt	0	68	13037	633

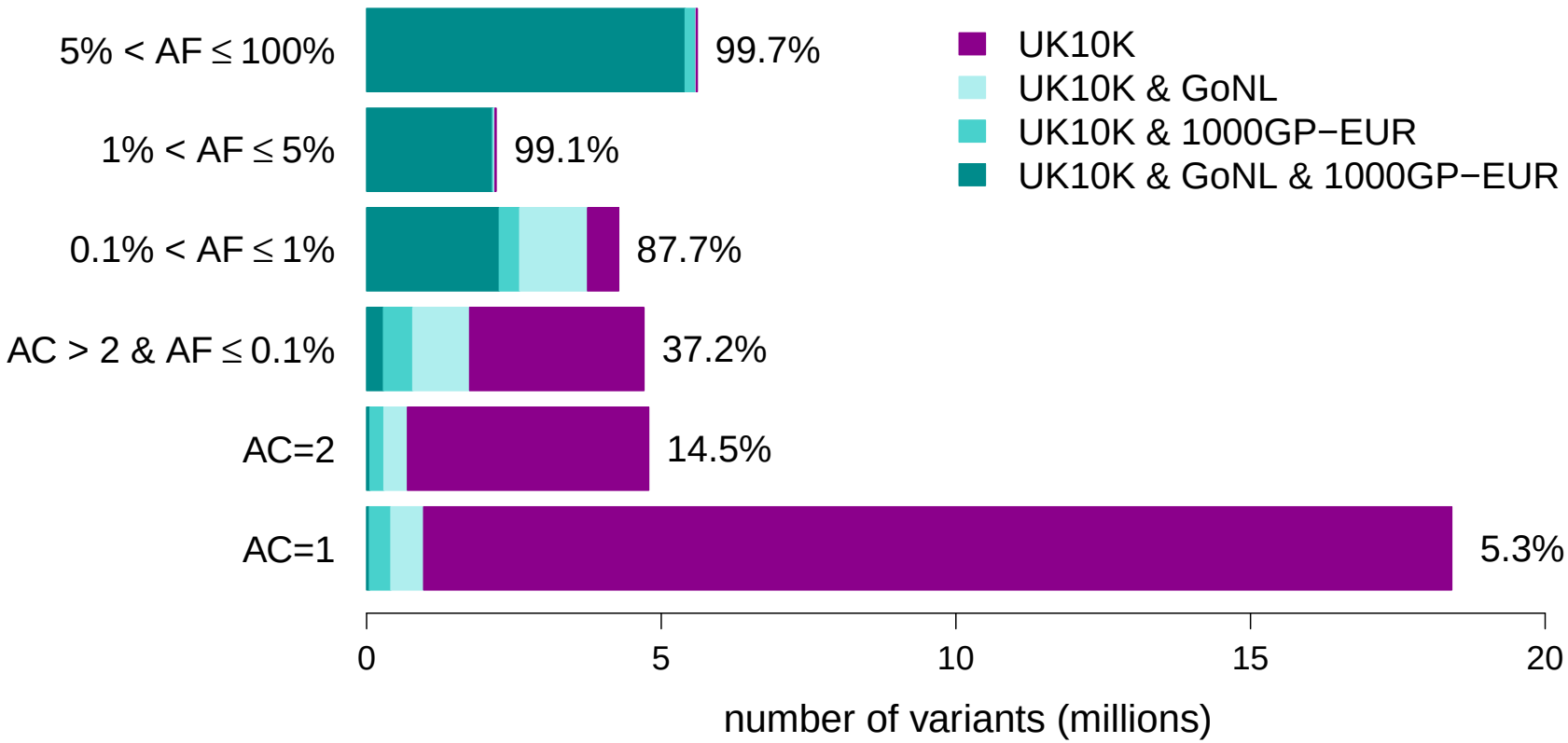
Overall genotype concordance = 99.5%

Non-reference discordance rate = 2.97%

For variant sites with MAF > 5% the NRD = 0.6%

HomRef = homozygous reference
Het = heterozygous
HomAlt = homozygous alternative

Variants shared with 1000GP-EUR and GoNL

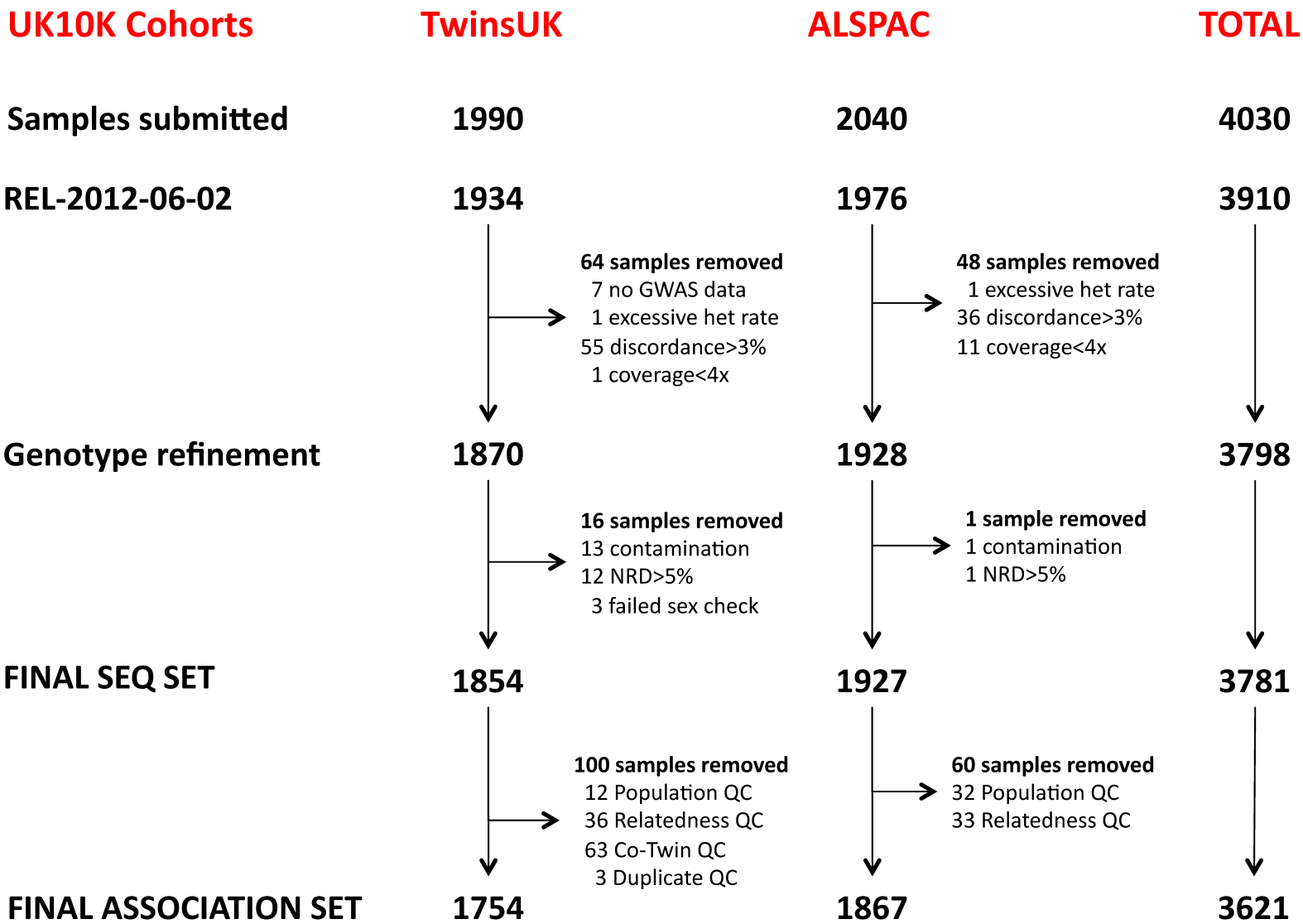


UK10K Consortium, Walter *et al.* (Nature 2015)

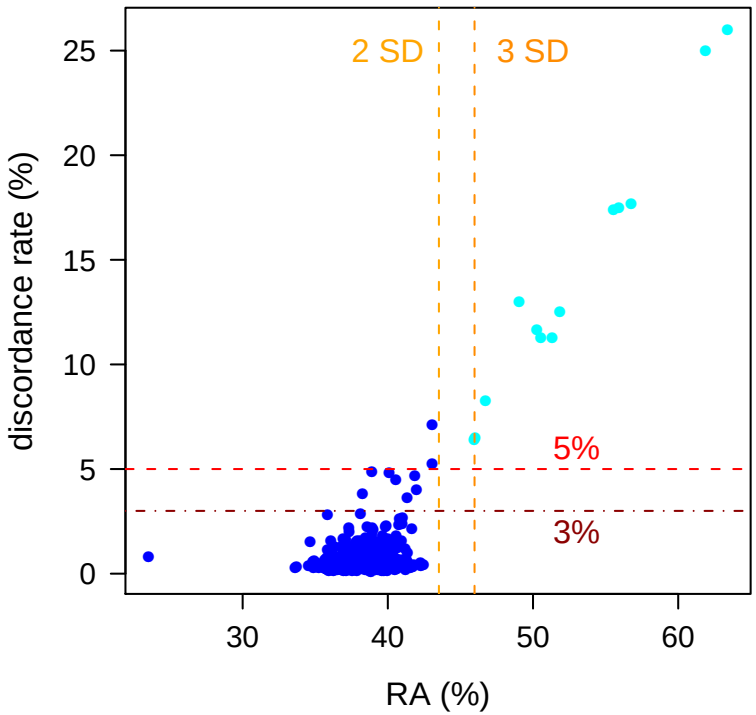
Quality control of samples

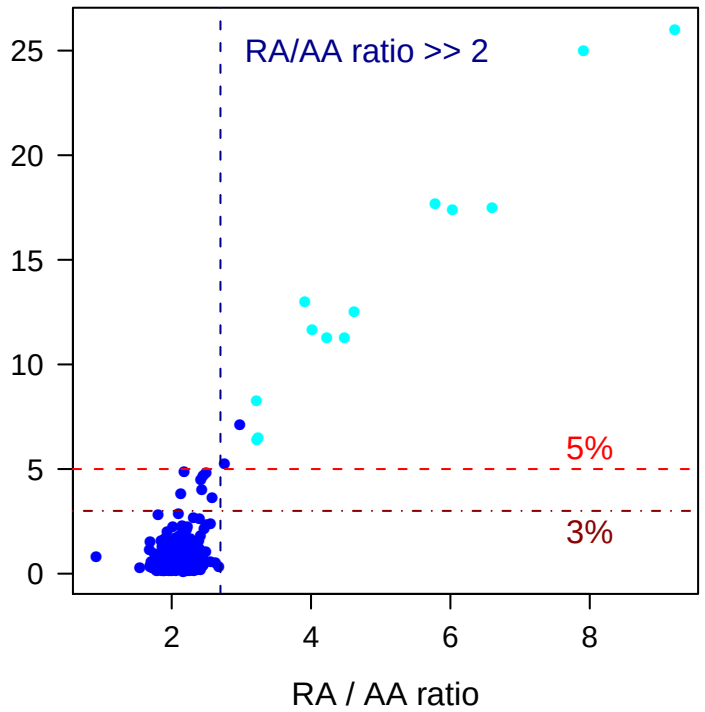
- Discordance with GWAS genotype
- Excess heterozygosity
- CHIPMIX and FREEMIX

Sample QC workflow



Heterozygous rate versus discordance

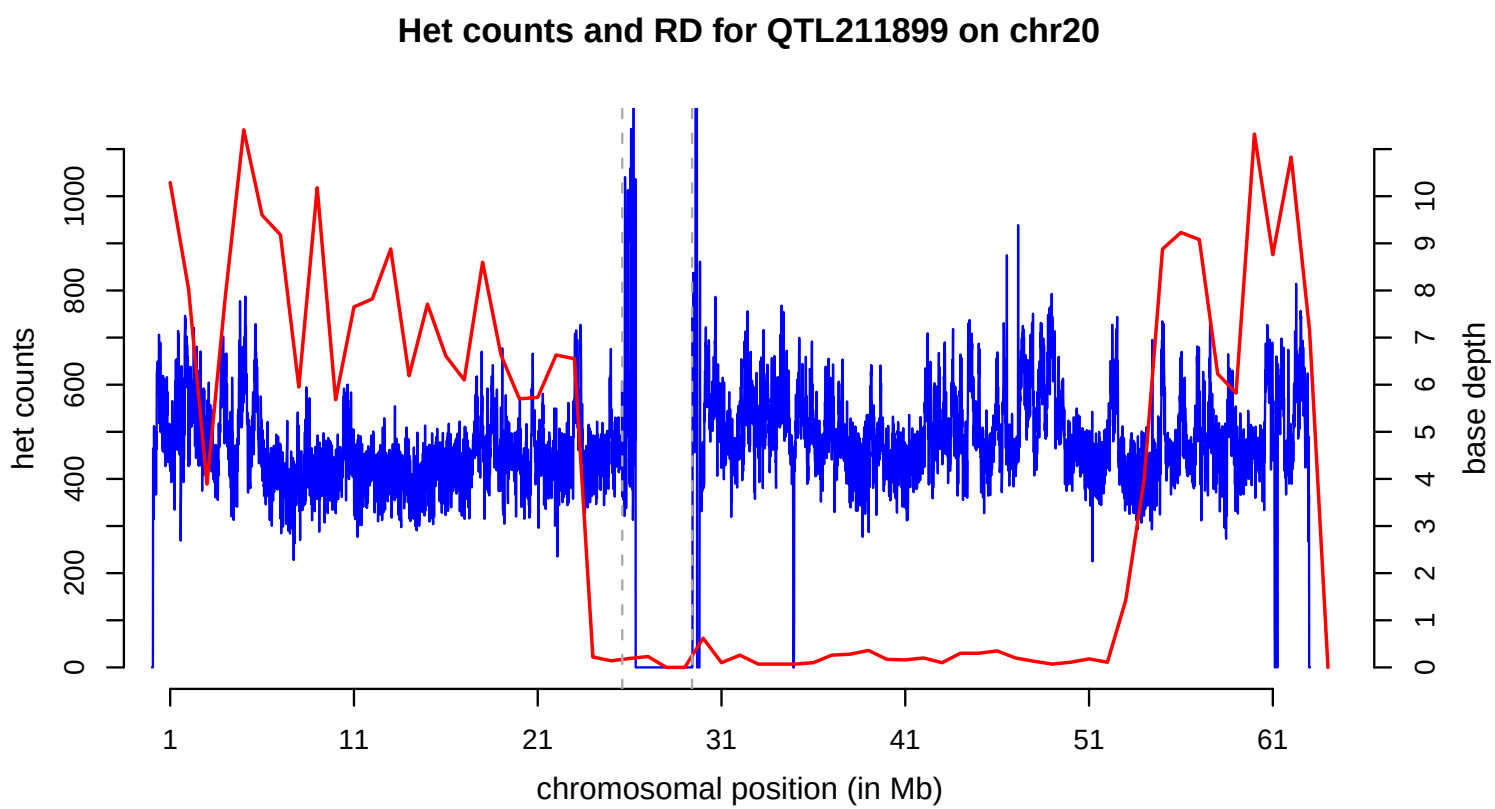




Quality control of samples:

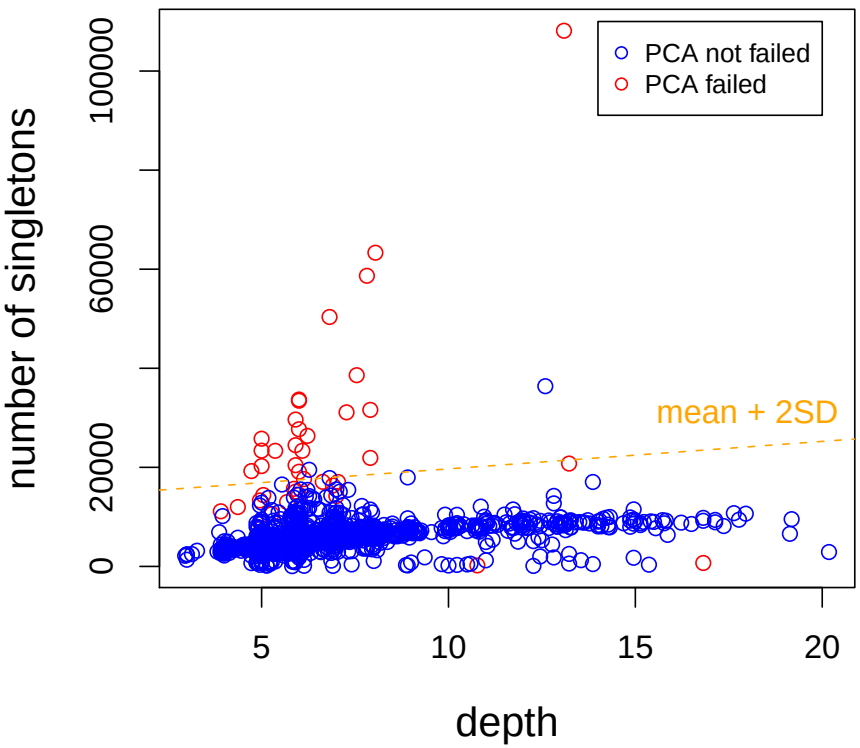
- Het rate $> \mu + 3\,\sigma$
- Discordance rate $> 3\%$

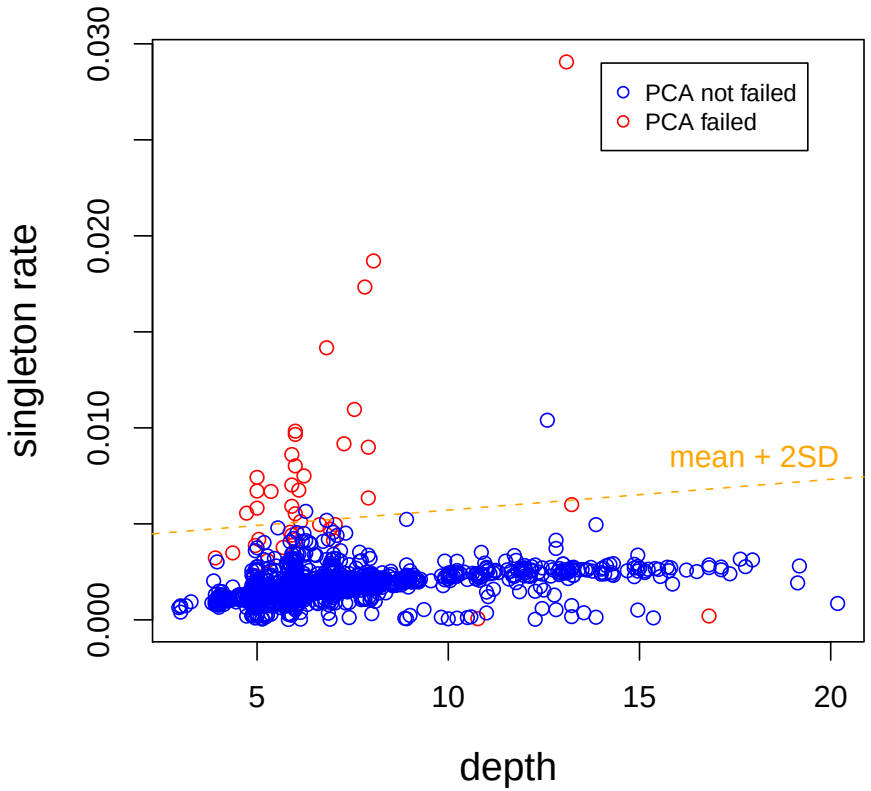
Low het rate and depth of coverage for QTL211899 on chr20



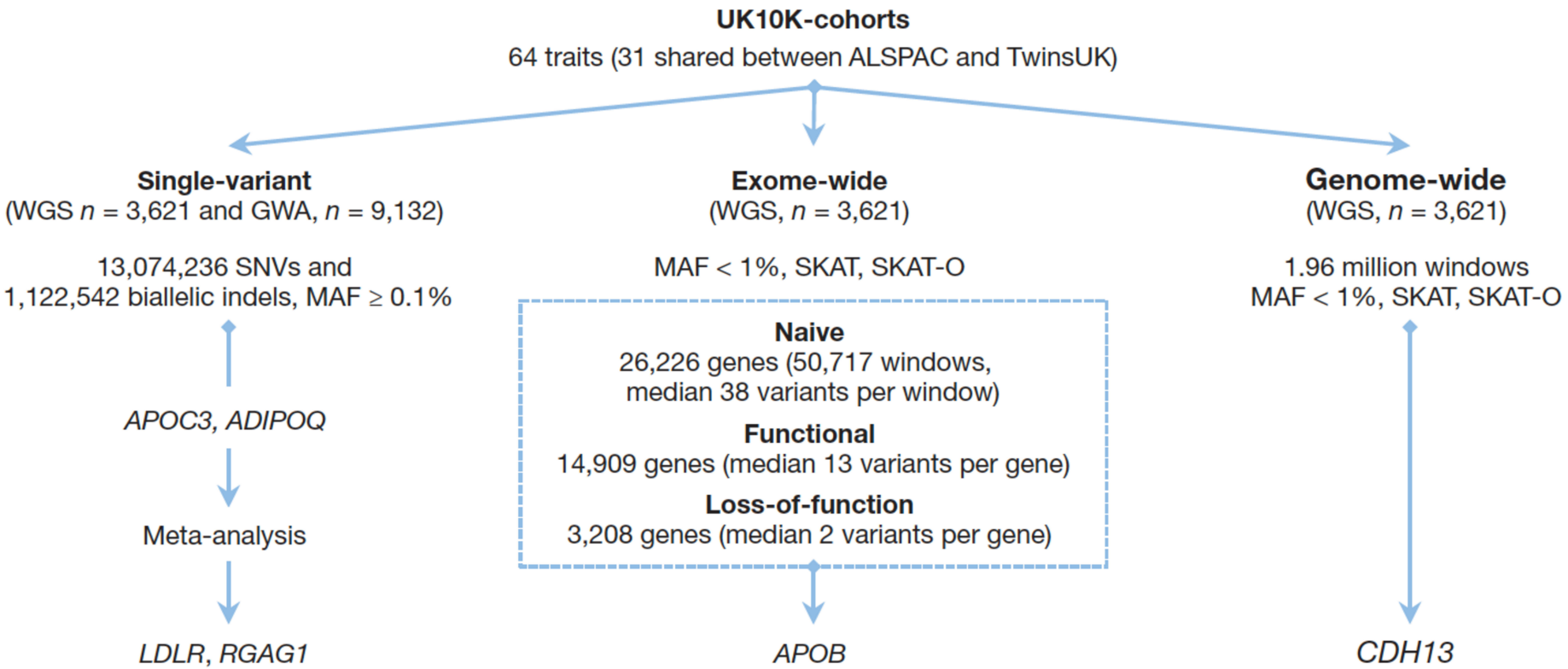
Read depth is not decreased along the $\sim$20 Mb chunk on chr20 for QTL211899, so it is not a deletion. It could be uniparental disomy (UPD), but more likely homozygosity by descent.

Depth versus number of singletons and singleton rate





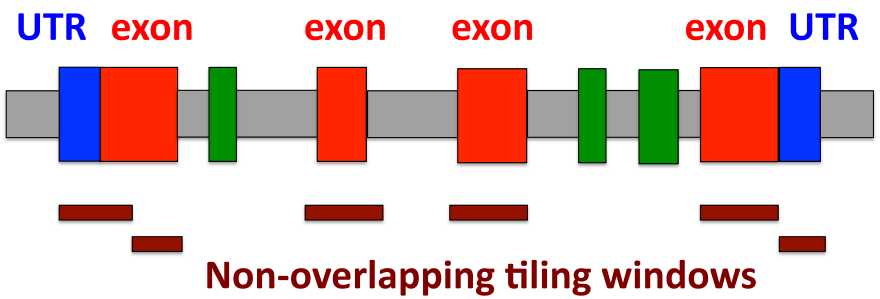
Study design for associations tested



UK10K Consortium, Walter *et al.* (Nature 2015)

Rare variants analysis

Joint effects of multiple variants in a region (SKAT for MAF $< 1\%$)

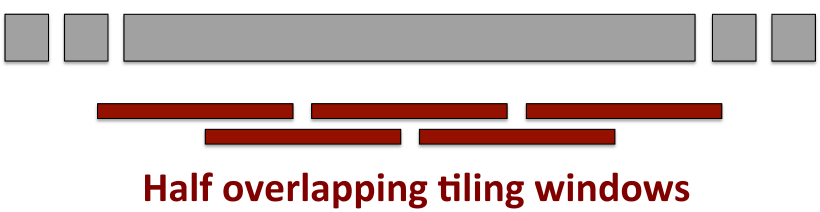


Exome-wide analysis

Variants in CDS+UTR

Non-overlapping windows ≤ 50 SNVs

26,226 genes and 50,717 windows



Genome-wide analysis

3 kb tiled windows overlap by half

Average ~ 38 variants per window

Single-point analysis of common variants

Variants with MAF $> 0.1\%$ were analysed with SNPTEST using an additive model within a frequentist test. For each trait residual y_i and genotype x_i a linear model

$$y_i = \beta_0 + \beta_1 x_i$$

was fitted for $i = 1, 2, ..., n$ where n is the number of samples (WTCCC, Nature 2007).

Single-point meta-analysis of common variants

Meta-analyses were carried out with GWAMA assuming a fixed effects model. GWAMA calculates the combined allelic effect B_j across all studies at the j-th variant as

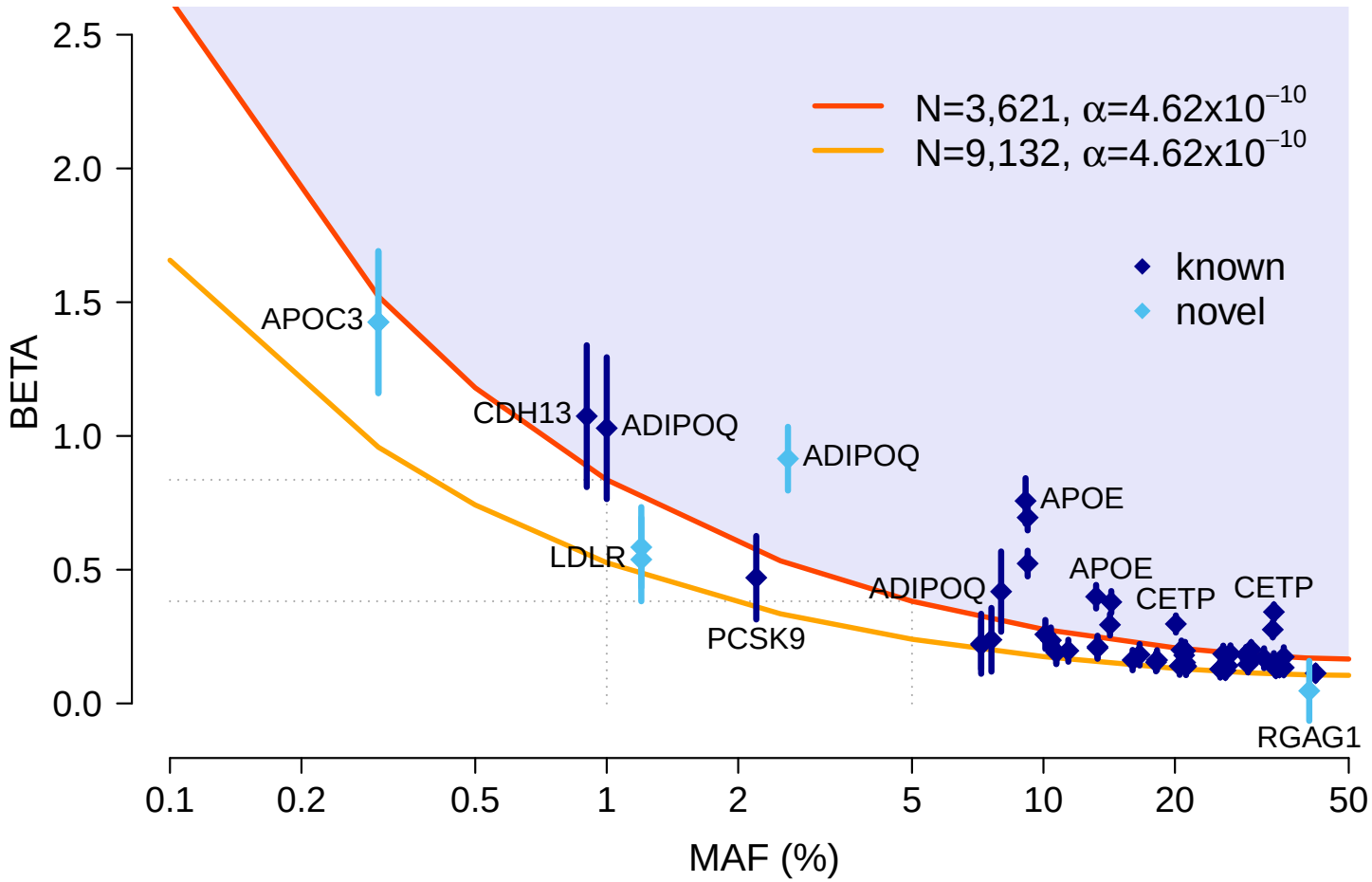
$$B_j = \frac{\sum_{i=1}^{N} \beta_{ij} w_{ij}}{\sum_{i=1}^{N} w_{ij}}$$

β_{ij} represents the effect of the allele at the j-th variant in the i-th study and w_{ij} represents the inverse of the variance of the estimated allelic effect. The combined variance is given by $V_j = (\sum_{i=1}^{N} w_{ij})^{-1}$.
(Magi R & Morris AP, BMC Bioinformatics 2010)

Collapsing and burden tests for rare variants (MAF $< 1\%$)

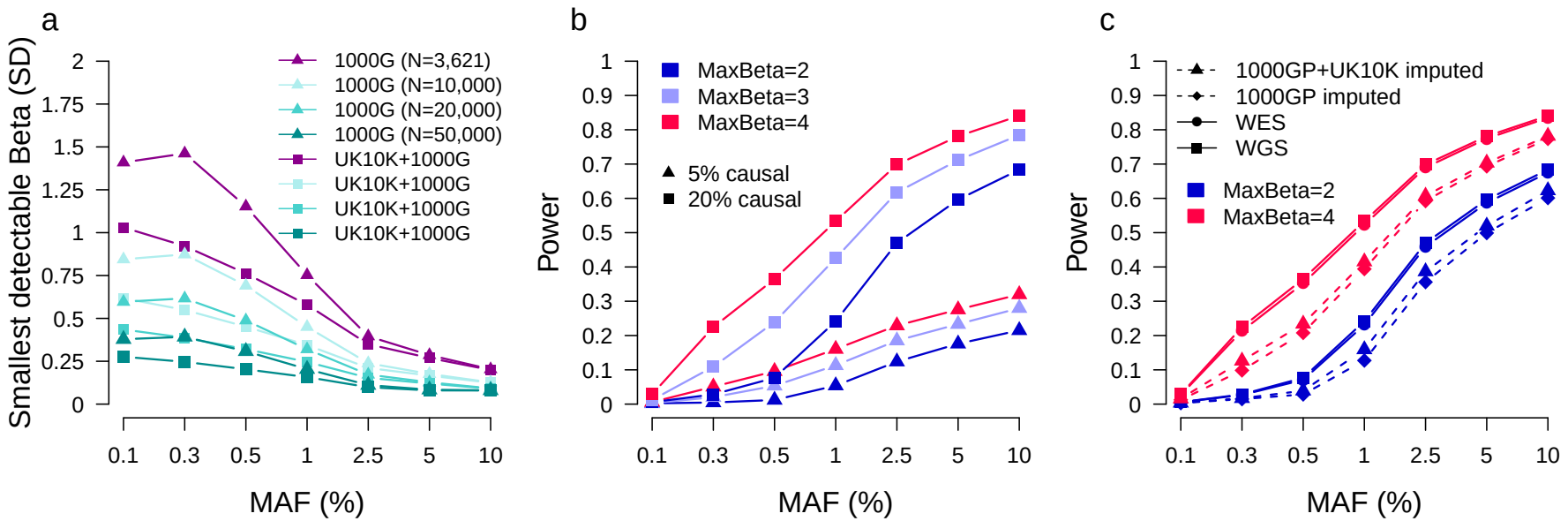
Sequence Kernel Association Tests (SKAT and SKAT-O) were used to test rare variants.
SKAT is a variance-component multiple regression test, it retains power if there are variants with opposite direction of effects.
SKAT-O represents the best linear combination of SKAT and burden tests (Wu *et al.*, AJHG 2011).

Summary of association results



UK10K Consortium, Walter *et al.* (Nature 2015)

Power for single-variant and region-based tests

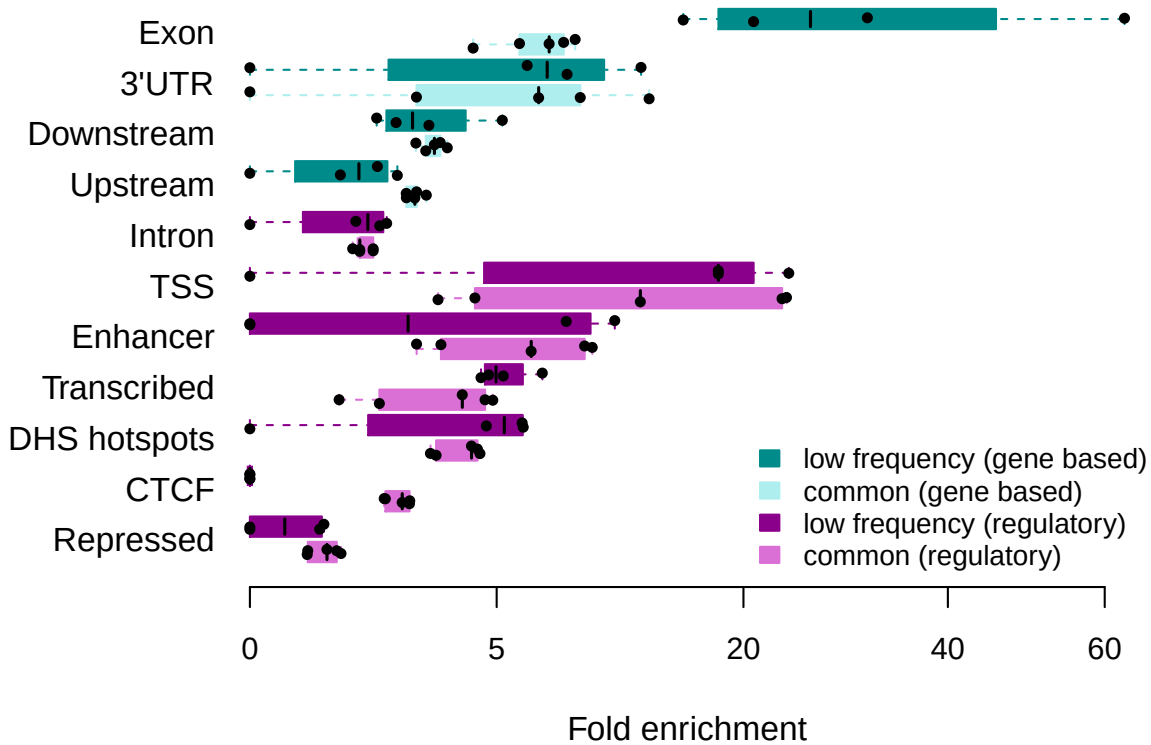


UK10K Consortium, Walter *et al.* (Nature 2015)

Enrichment of single-marker associations by functional annotation

$$\text{Fold enrichment}(t) = \frac{N_a^t}{N^t} / \frac{N_a}{N}$$

N total number of variants
N_a total number of variants that fall in the annotation of interest
N^t total number of variants with p less than threshold
N_a^t number of variants with p less than theshold t that fall in the annotation of interest



UK10K Consortium, Walter *et al.* (Nature 2015)

Introduction to population stratification analysis

- Population structure is a known confounder of association studies
- Are methods to control stratification for common variants equally effective for rare variants?
  (Mathieson & McVean, Nature Genetics 2012)
- Link twins locations to mean longitude and latitude data
- Residuals of 50 phenotypes adjusted for age, sex and other co-variates

Generalized additive models (GAM) by Trevor Hastie and Robert Tibshirani

- Extension of traditional linear statistical model
- Can be applied for standard continuous response regression, categorical or ordered categorical response data, count data, survival data and time series
- Scatterplot smoothing functions
- Overfitting can be a problem

GAM

The model specifies a distribution (such as a normal distribution, or a binomial distribution) and a link function g relating the expected value of the distribution to the m predictor variables, and attempts to fit functions $f_i(x_i)$ to satisfy:

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_m(x_m)$$

The functions $f_i(x_i)$ may be fit using parametric or non-parametric means, thus providing the potential for better fits to data than other methods.

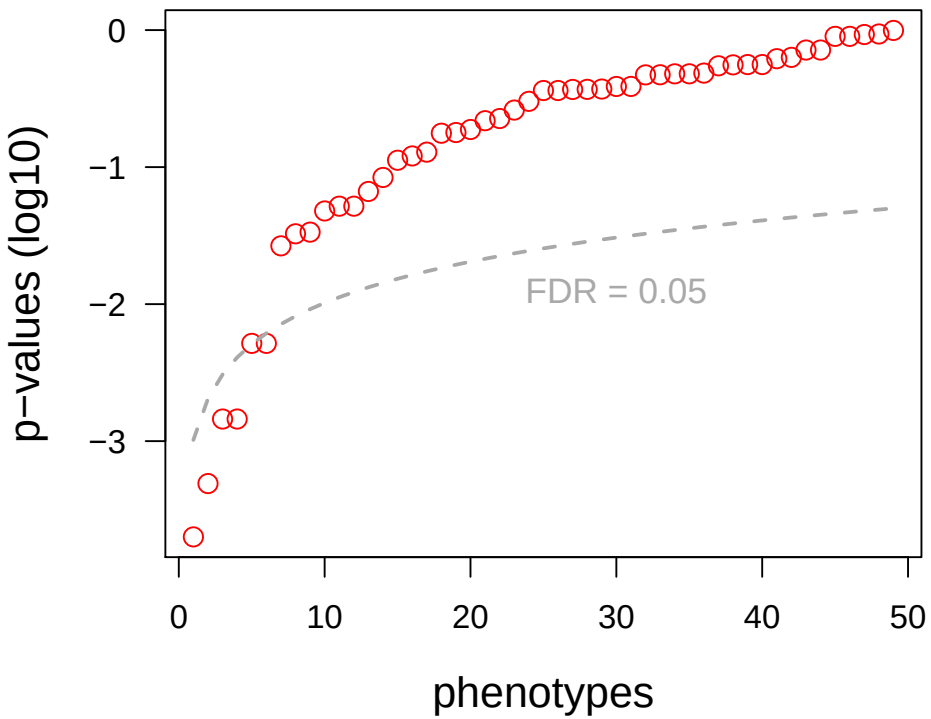
Generalized additive models (GAM)

Trait	p-value
Urea (BMIadj)	0.00020
Glucose	0.00049
Height	0.00145
Height (std)	0.00145
Leptin	0.00517
Leptin (std)	0.00517
DBP	0.02662
VLDL	0.03260
TG	0.03353
LDL	0.04785
BMI	0.05185
BMI (std)	0.05185
HOMA-ir	0.06643
Uric Acid (BMIadj)	0.08405

GAM models were fitted for each trait against geographical location.

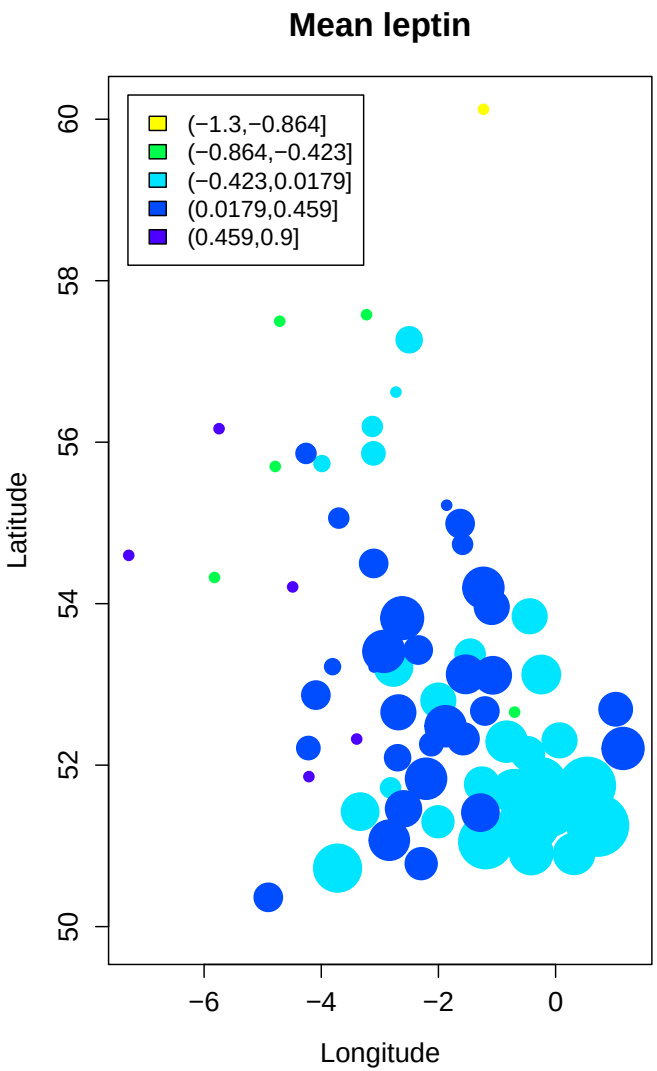
Then the significance of the smoothing functions were tested using ANOVA.

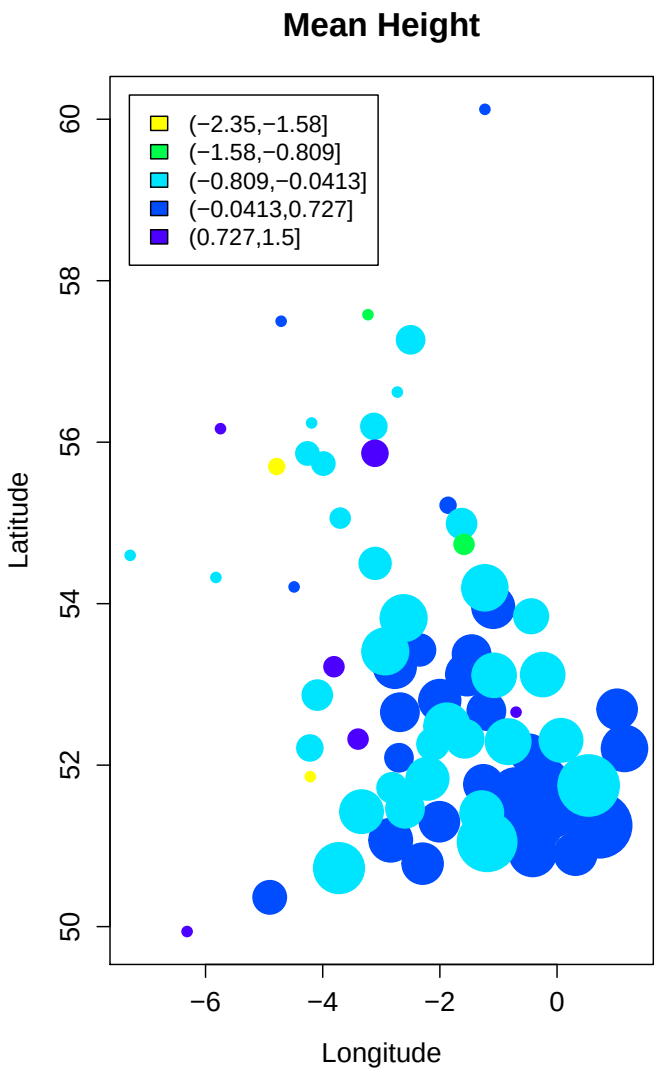
After correcting for multiple testing using FDR with $q = 0.05$ for GAM p-values



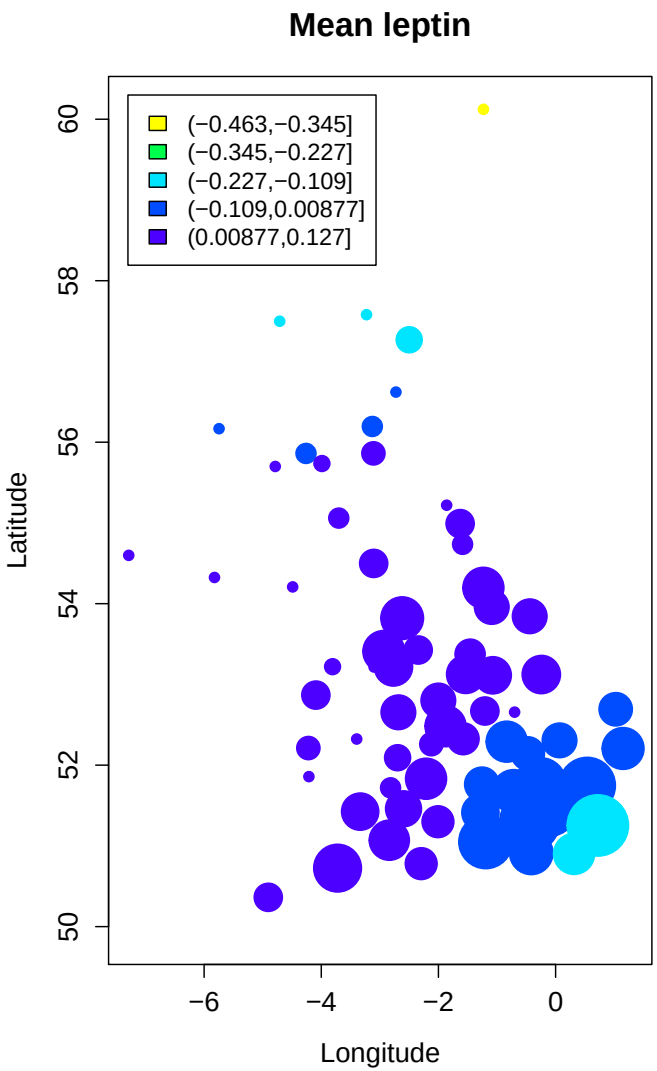
Trait	p-value
Urea (BMIadj)	0.00020
Glucose	0.00049
Height	0.00145
Height (std)	0.00145
Leptin	0.00517
Leptin (std)	0.00517

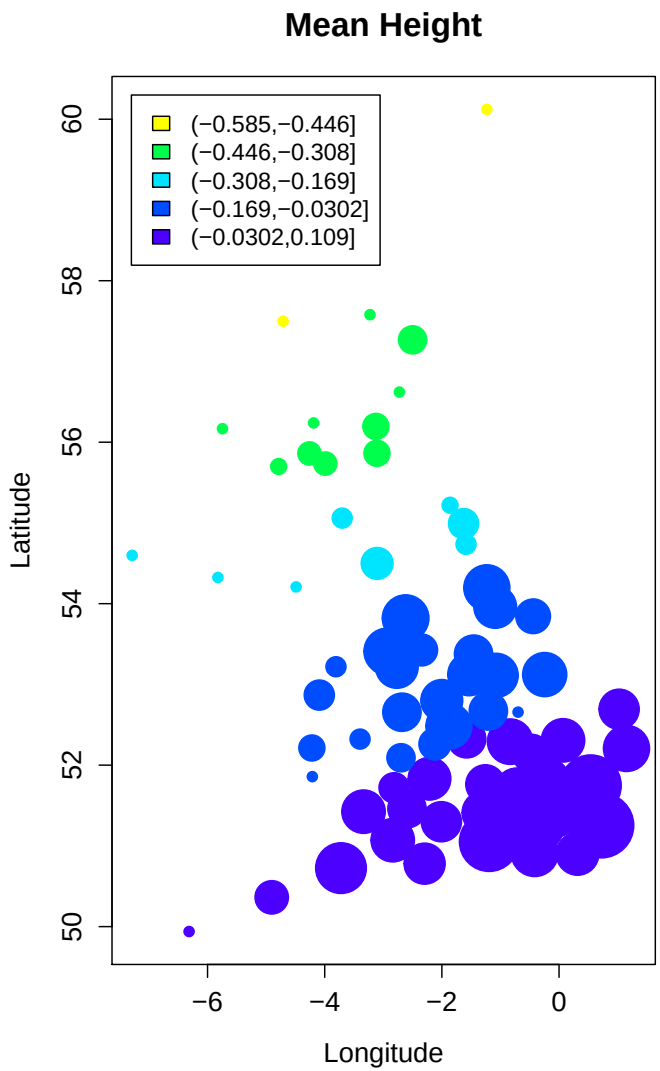
Leptin and Height



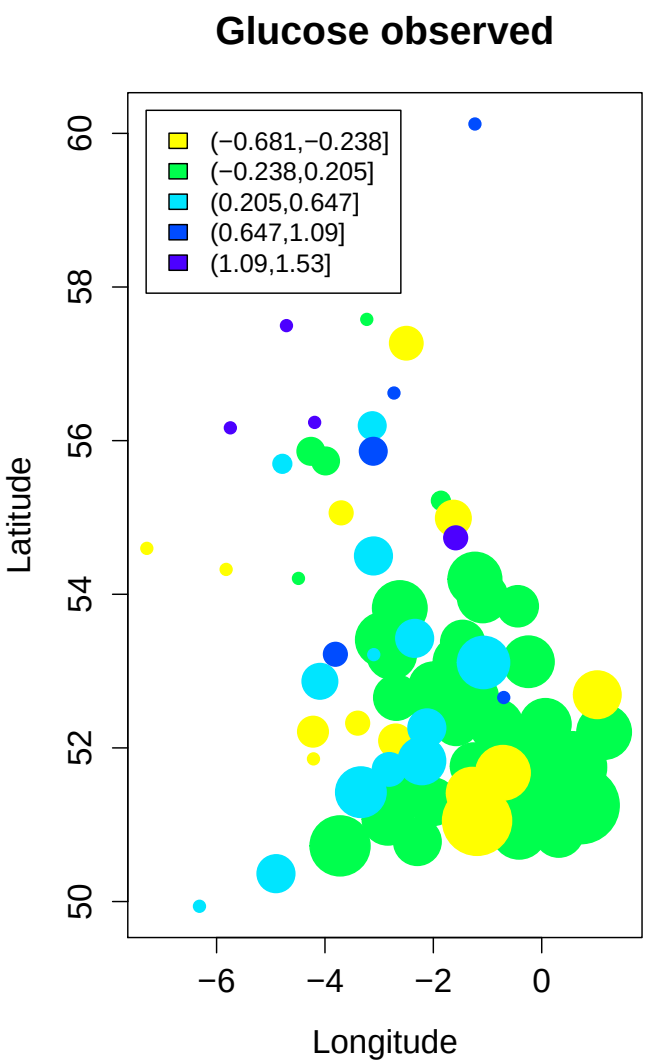


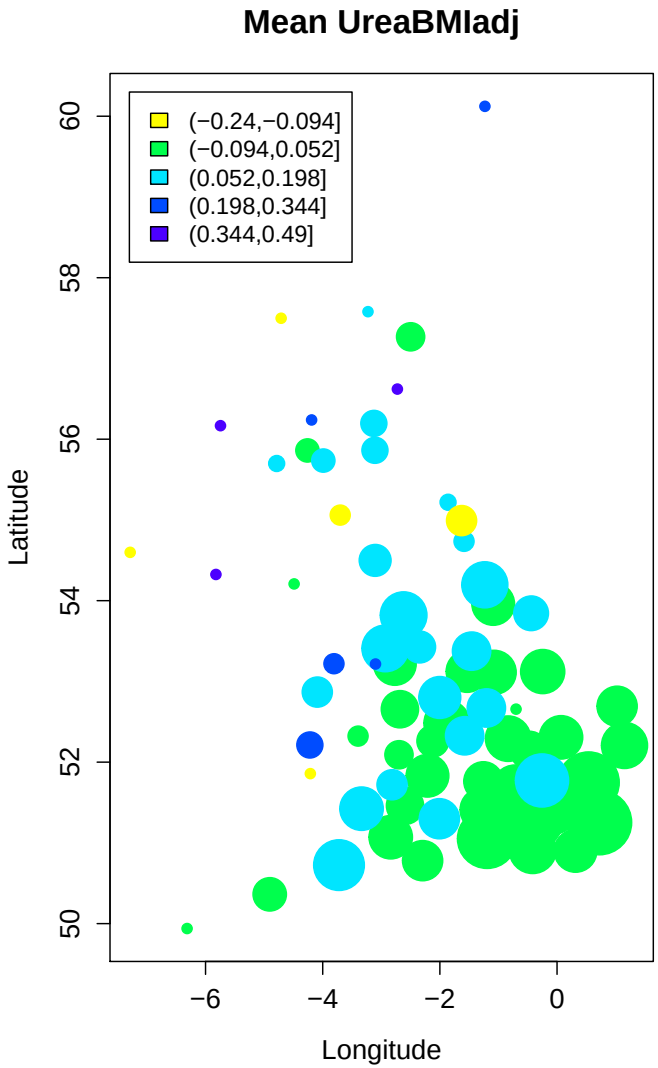
Predicted values for Leptin and Height



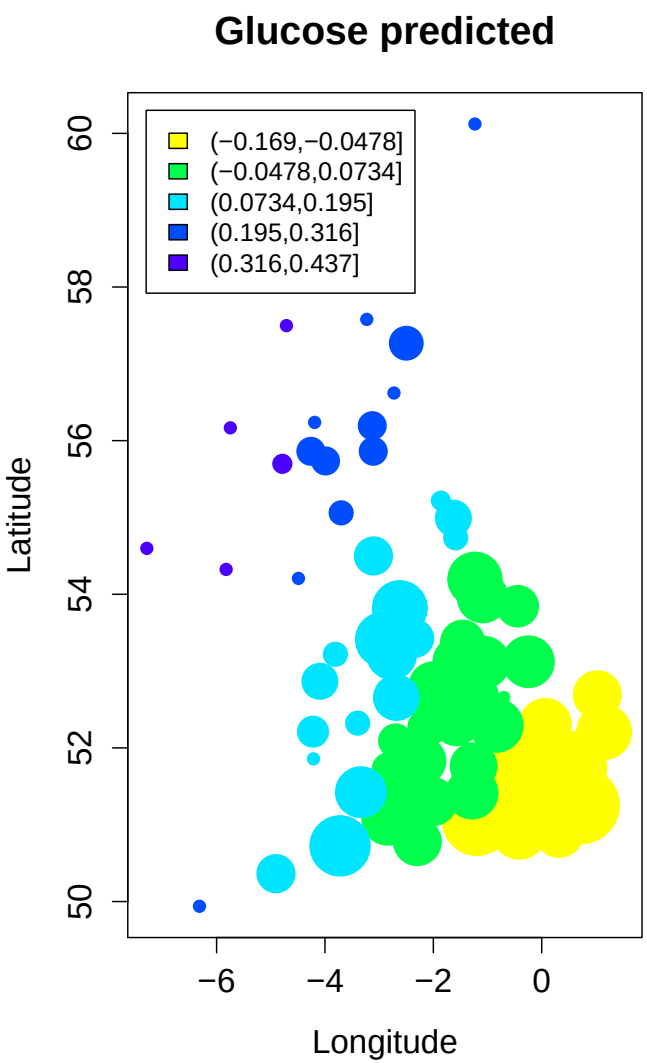


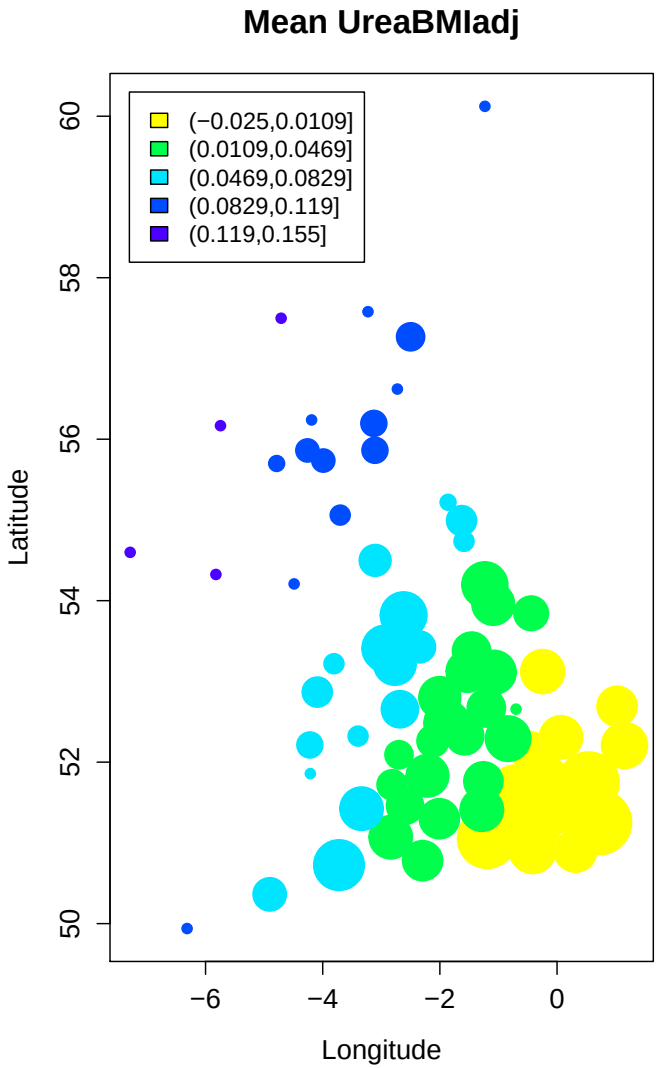
Glucose and Urea adjusted for BMI





Predicted values for Glucose and Urea adjusted for BMI





Random traits with spikes and clines (1)

Random traits were generated using a normal distribution $N(0, 1)$, adding a regional spike and a north-south cline

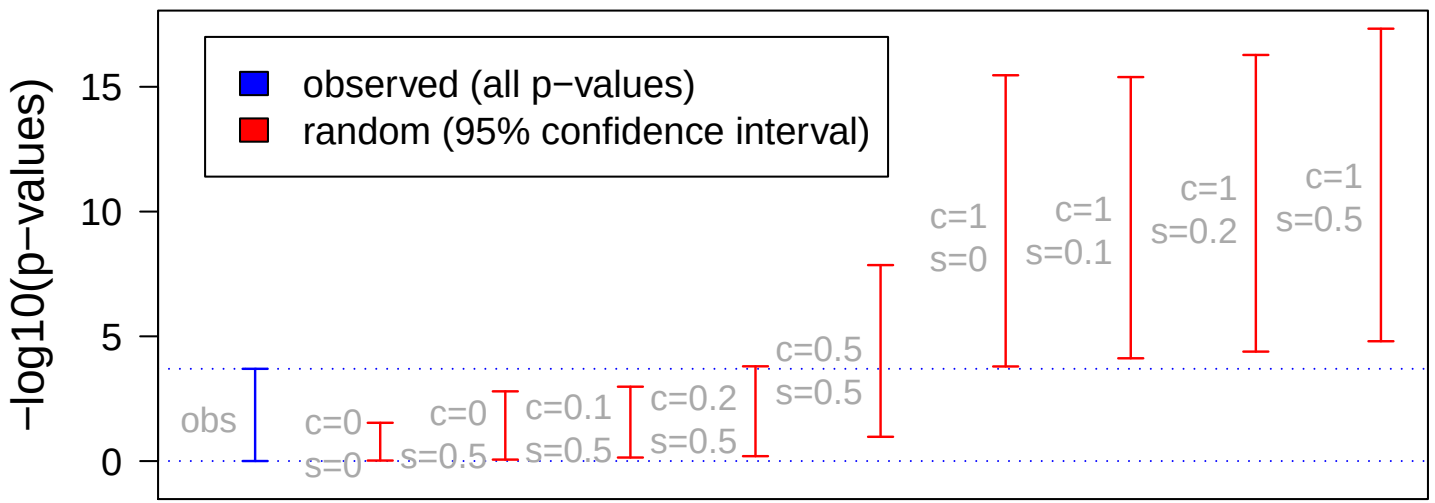
Selected spikes: 0, 0.1, 0.2, 0.5 (in SD)
Selected clines: 0, 0.1, 0.2, 0.5 (in SD)

Combinations: cline $= 1$ and all spikes,
spike $= 0.5$ and all clines

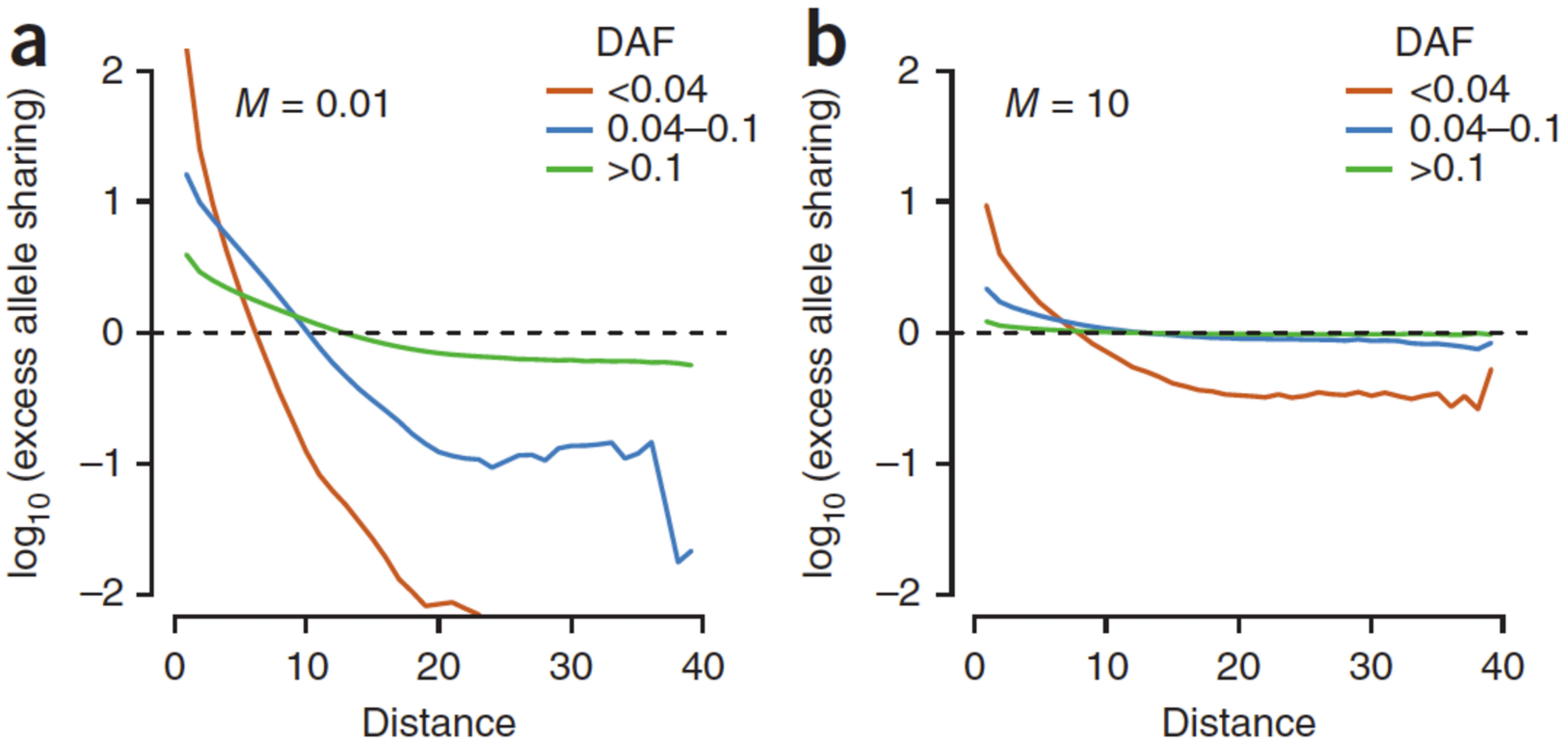
Control: cline $= 0$ and spike $= 0$

Random traits with spikes and clines (2)

- Random traits for $\sim$ 1500 samples
- Scenarios are combinations of clines (= c) and spikes (= s)
- For each scenario 1000 traits were generated
- GAM models were fitted for each trait versus location
- 95% confidence intervals were generated for GAM p-values



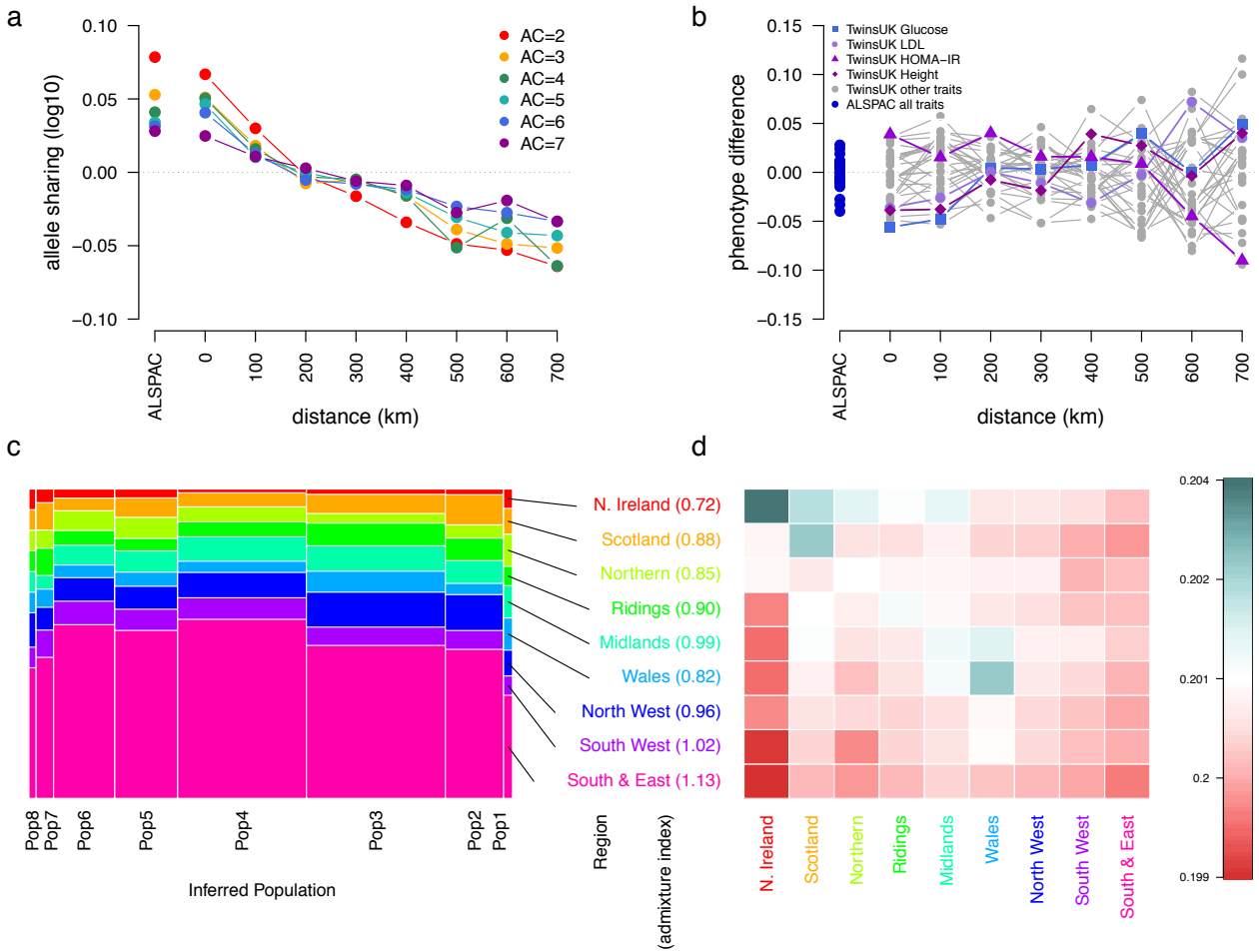
Excess allele sharing by distance (Mathieson *et al.* 2012)



Migration rate $= 0.01$
$F_{ST} = 0.1$

Migration rate $= 10$
$F_{ST} < 0.01$

Allele sharing by distance and FineSTRUCTURE

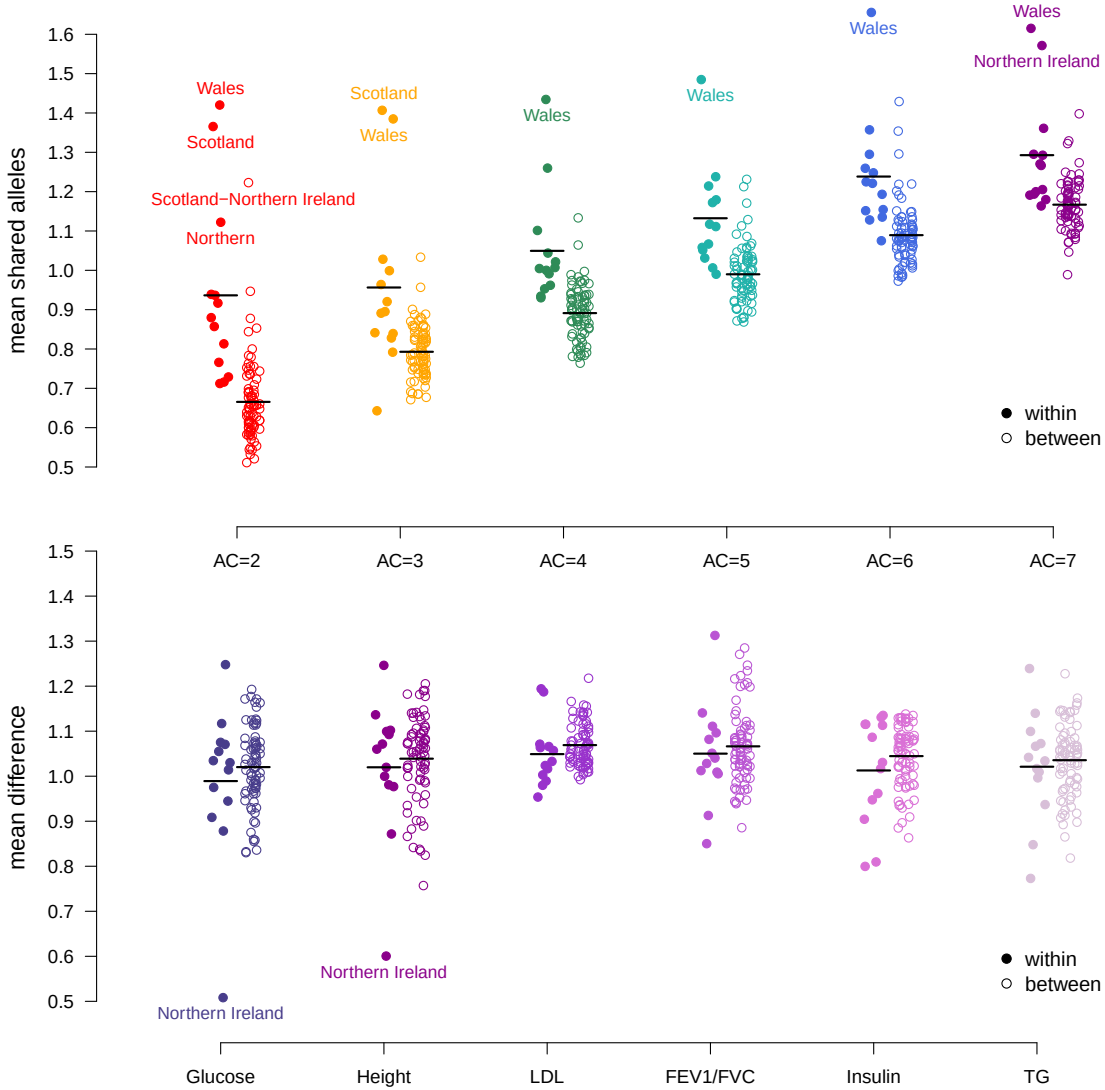


UK10K Consortium, Walter *et al.* (Nature 2015)

Allele sharing within and between regions

- Sequence data for 1139 twins from TwinsUK
- Phenotype and place of birth available
- Count shared doubletons (AC=2) and shared alleles for allele counts AC 3 to 7 between each pair
- Summarise counts within and between 12 regions
- Correct for number of pairs within region $(n \times (n-1))/2)$ and between regions $(n \times m)$ for n samples in one region and m samples in the other region

Genotype and phenotype similarities by regions



Mantel tests

AC=2		AC=3		AC=4	
Height	0.048	Height	0.052	Adiponectin	0.190
LDL	0.063	Weight	0.055	TRFM	0.231
Adiponectin	0.071	Adiponectin	0.075	Insulin	0.297
Weight	0.177	FEV1/FVC	0.117	Weight	0.317
Waist	0.192	Waist	0.183	Gripstrength	0.359

AC=5		AC=6		AC=7	
Gripstrength	0.137	FEV1/FVC	0.142	Insulin	0.100
Adiponectin	0.206	Adiponectin	0.144	ApoA1	0.108
ApoB	0.298	Height	0.144	Gripstrength	0.119
Insulin	0.310	Glucose	0.144	TFM	0.175
ApoA1	0.318	LDL	0.150	FEV1	0.206

A resource for the community

- Data access conditions
  - Data deposited to European Genome-Phenome Archive (EGA)
  - Application to Data Access Committee (DAC) (www.uk10k.org/data_access)
- Genotype
  - All primary sequence data submitted to EGA
  - Final variant calls passing QC submitted to the EGA
- Phenotype
  - Exomes: disease status
  - Cohorts: Core phenotypes released with genetic data (raw data, data dictionaries, trait protocols and standardized residuals)
  - Other phenotypes accessible through cohort DACs: longitudinal phenotypes and non-core phenotypes
- Reference panel for imputation

Summary

- The UK10K project has generated an enormous amount of genotype data
- There are already studies with many more sequenced individuals (e.g. INTERVAL, 100,000 Genomes Project)
- Quality control is important

Acknowledgments

PI and Co-chairs
Richard Durbin
Nicole Soranzo
Nicholas Timpson

Production
Shane McCarthy
Petr Danecek
Jim Stalker
Yasin Memari

UK10K Manager
Dawn Muddyman

UK10K Exome Groups
Lucy Crooks
Jamie Floyd
Audrey Hendricks

UK10K Cohorts Group
Josine Min
and many others

Team 151
Lu Chen
Jie Huang
Max Cocca
Valentina Iotchkova

UK10K Pop-Strat Working Group
Eleftheria Zeggini
Nicole Soranzo
Sarah Metrustry
Nicholas Timpson
Jennifer Asimit
Audrey Hendricks

UK10K cohorts team

Chairs
Nicole Soranzo, WTSI
Nic Timpson, Bristol University

WTSI
Aaron Day-Williams
Andrew Brown
Audrey Hendricks
Chris Franklin
Eleftheria Zeggini
Ines Barroso
Ioanna Tachmazidou
Jie Huang
Jim Stalker*
Julian Hughes
Kalliope Panoutsopoulou
Kim Wong*
Klaudia Walter
Lorraine Southam
Lu Chen
Margarida Lopes
Petr Danecek*
Richard Durbin
Shane McCarthy*
So-Youn Shin
Yasin Memari

Bristol University
Beate St Pourcain
Chris Boustred
Dave Evans
George Davey-Smith
Ghazaleh Fatemifar
Ian Day
John Kemp
Josine Min
Lavinia Paternoster
Tom Gaunt

Kings College London
Alireza Moayyeri
Feng Zhang
Genevieve Lachance
John Perry
Kerrin Small
Kirsten Ward
Lydia Quaye
Massimo Mangino
Pirro Hysi
Sarah Metrustry
Scott Wilson
Tim Spector
Yalda Jamshidi

Leicester University
Louise Wain
Martin Tobin

BGI Shenzen
Jing Tian
Jun Wang
Sifei He
Yingrui Li

EBI
Graham Ritchie
Paul Flicek

Oxford University
Jonathan Marchini

McGill University
Brent Richards
Celia Greenwood
Houfeng Zheng
Rui Li

* production