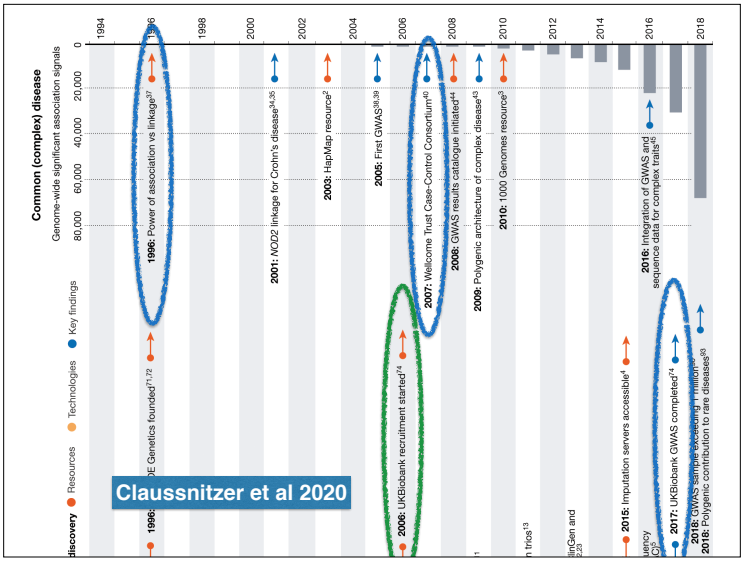
QC GWAS

Hae Kyung Im, PhD

THE UNIVERSITY OF CHICAGO

January 25, 2021



In 1996, Risch and Merikangas published a highly influential perspective paper that show the benefits of GWAS.
We have come a long way since the landmark publication of the WTCCC GWAS in 17,000 cases of 7 common diseases and controls.
Sample sizes continued to grow over the years, identifying increasing number of genomic loci associated with complex traits.
In 2006, recruitment for the UK Biobank project started.
Today we will look look at the UK Biobank GWAS performed half a million participants.

GWAS in 2020

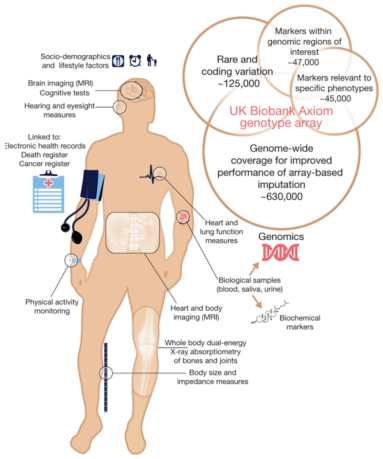
UK Biobank

Biological Samples in a Storage Freezer at the UK Biobank Nancy Cox, UK Biobank shares the promise of big data, 2018, Nature

UK Biobank started recruitment of participants in 2006, even before the publication of the WTCCC GWAS study publication, a testament of the forwarding looking vision of the proponents of the study. This is a picture of the huge freezer with automated handing of the biological samples. This gives us a sense of the scale of the biobank.

Their liberal data sharing policies made it possible for thousands of investigators to examine this data yielding more than 1000 publications to date (as of beginning 2020).

UK Biobank Genotype and Phenotype Data



- Large prospective population-based cohort study.
- Over 500,000 participants enrolled.
- Participants aged 40-69, between 2006 and 2010
- Deeply phenotyped
- questionnaires
- physical & biological measurements
- electronic health records (EHR)
- images, accelerometer measurements (subset)
- Genotype data (488K)

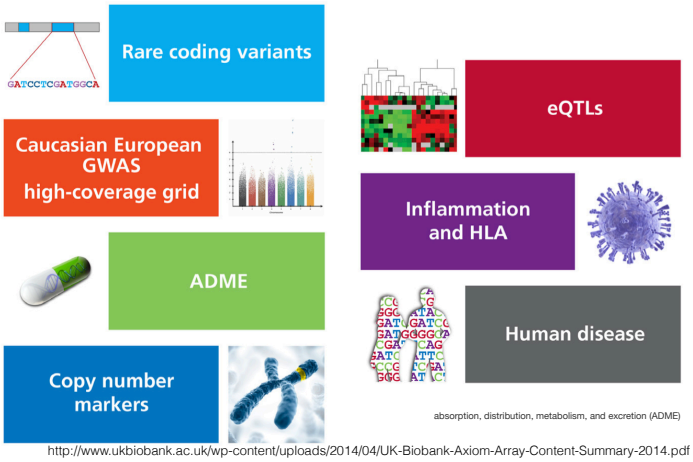
The depth of the phenotypes is astonishing. Electronic health records have been linked with the participants. The single payer health care system in the UK is a huge component that made possible to get this amount of information in a relatively uniform fashion.

UK Biobank: Genotyping Chips

- UK BiLEVE (UK Biobank Lung Exam Variant Evaluation)
  - N ~ 50K
  - UK BiLEVE Axiom Array by Affymetrix (807K markers)
- UK Biobank (remaining)
  - N ~ 438K
  - UK Biobank Axiom Array by Applied Biosystems

Two genotyping chips were used. Initially the UK BiLEVE Axiom Array with 807K markers that was used for about 50K participants. The remaining 90% of the participants were genotyped with the newer UK Biobank Axiom Array, specifically designed for this study.

UK Biobank Axiom Array Content (825K markers)



http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf

The UK Biobank Axiom Array contained 805K markers chosen with several criteria. One was to obtain a high coverage of the common variation to be used as a scaffold for genotype imputation. Rare coding variants were included, with the thinking that rare variation must be relevant for health and disease. eQTLs, variants that are known to regulate expression of genes, an important mechanism underlying the genotype-phenotype associations. Higher coverage of the complex HLA region and other immune implicated variants. Markers implicated in human diseases. Copy number variants (deletions, insertions, beyond SNPs).

UK Biobank: Ancestries

- White 94.23% (88.26% British)
- Asian 1.92%
- Black 1.57%
- Chinese 0.31%
- Mixed 0.58%
- Other 1.38%

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature, 1–25.

The biobank is mostly composed on White British individuals, with a small portion of Asian (1.92%), Black (1.57%), Chinese (0.31%), and others.

GWAS QC

Why Is QC Important?

Science. 2010 Jul 1;2010. doi: 10.1126/science.1190532. Epub 2010 Jul 1.

Genetic signatures of exceptional longevity in humans

Sebastiani P[1], Solovieff N, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wil

Author information

Retraction in
Retraction. [Science. 2011]

10

QC is the least glamorous part of research and analysis. So why should we care about QC? Well, to avoid huge pitfalls and draw spurious conclusions. General rule: if something sounds too good to be true? Well, it is highly likely it is not true. So before making big claims and causing media splash make sure your QC is super solid. Think of every possible confounders that could lead to the "interesting" results.

In this paper, the authors had found many SNPs associated with being centenarian, i.e. they thought they had found the "longevity genes"

Why Is QC Important?

RETRACTED ARTICLE

See: Retraction Notice

Science. 2010 Jul 1;2010. doi: 10.1126/science.1190532. Epub 2010 Jul 1.

Genetic signatures of exceptional longevity in humans

Sebastiani P[1], Solovieff N, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wil

Author information

Retraction in
Retraction. [Science. 2011]

11

But then they found out that there was a problem with some of the chips that affected more of the centenarians than the controls. Faulty chips were confounded with being a case leading to false positive results. Their original claim that they had 77% accuracy to predict longevity could not be supported with the QC'd data.

Title Text

nytimes.com

The New York Times

Scientists Retract Report on Predicting Longevity

By Nicholas Wade

July 22, 2011

Scientists who asserted last year that they could predict with 77 percent accuracy who would live past 100 have retracted their report in the journal Science, yet say they are right anyway.

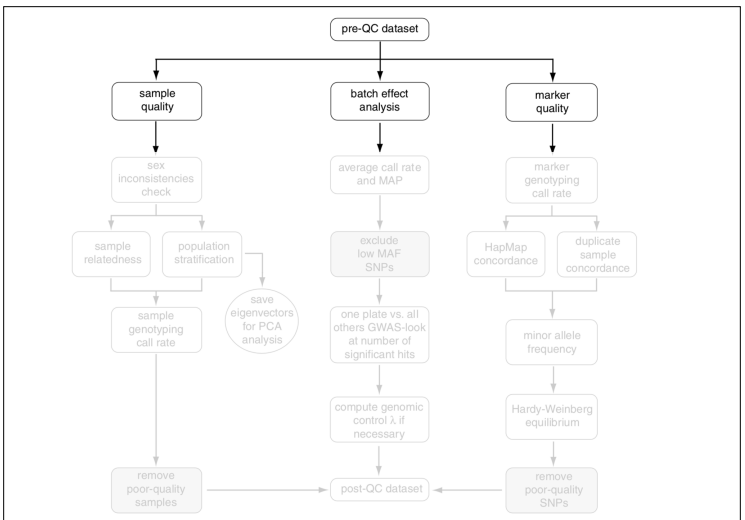
The researchers, Paola Sebastiani and Dr. Thomas T. Perls of Boston University, wrote in Science last July that they had found 150 genetic variants that correlated with extreme longevity.

https://www.nytimes.com/2011/07/23/science/23retract.html

12

You really don't want to appear in the NY Times as the scientist who had to retract a paper because of a faulty QC. After the publication, they realized that a 10% of the centenarians had been genotyped in faulty chips.

How would they have detected the confounding between the chip and the longevity status?



S. Turner et al, 2011, "Quality control procedures for genome-wide association studies", Current Protocols in Human Genetics

Here is a summary of the workflow for QC in GWAS. Three con

Marker-based QC

Test	Average number of SNPs failed per batch (sd)	Fraction of all genotype calls affected
Affymetrix cluster QC	1109 (699)	0.00140
1. Batch effect	197 (86)	0.000249
2. Plate effect	284 (266)	0.000358
3. Departure from Hardy-Weinberg equilibrium	572 (77)	0.000723
4. Sex effect	45 (5)	0.0000569
5. Array effect*	5417	0.00683
6. Discordance across controls**	622 and 632	0.000796
Total	7704 (721)	0.00971

UKB QC pipeline was designed specifically to accommodate the large-scale dataset of ethnically diverse participants, genotyped in many batches (106), using two slightly different novel arrays, and which will be used by many researchers to tackle a wide variety of research questions.

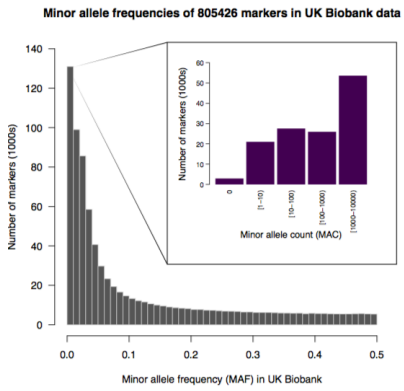
Clare Bycroft, et al. Nature 2018

14

Markers were filtered out according to several criteria.

Manufacture's criterion: failure to clustering used for calling the genotypes. On average, 1109 SNPs per bach failed Affymetrix cluster QC.
Batch effects: there were 106 batches of about 5000 individuals.
Plate effects: participants DNA were placed on 96 well plates.
Departure from HW equilibrium
Sex effects
Array effects. In the next slides, we will see examples of these QC measures.
In total, less than 1% of markers were excluded due to low quality.

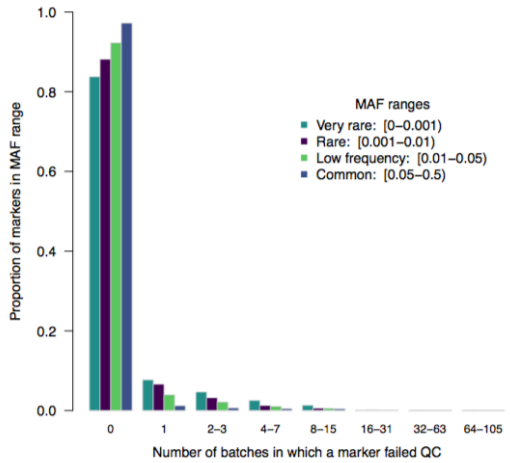
Minor Allele Frequencies of 805K Markers



15

Here is the distribution of the minor allele frequencies in the UKB.
About 130K markers had allele frequencies below 1%. Half of rare variants were found in at least 1000 individuals. 20K were present in less than 10 participants. (Given rarity, there probably wasn't two copies of these rare alleles in one individual)

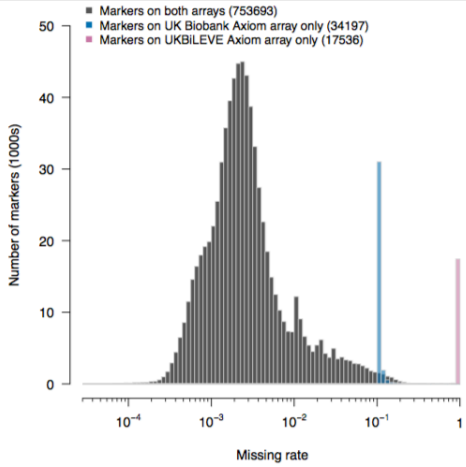
QC Sucesss/Failure Rates by MAF Ranges



16

Most common variants, more than 95%, (in blue to the right) passed QC in all batches. Quality was overall pretty good even for low frequency: over 80% of the very rare variants passed QC in all batches (left most bar above 0).

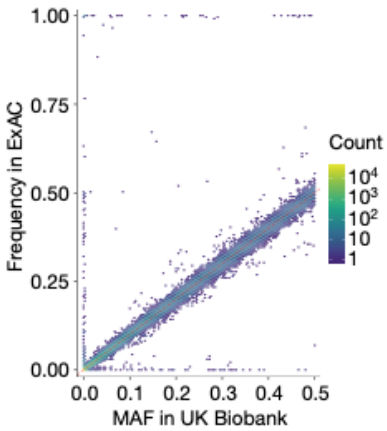
Missing Rates for Markers

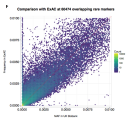


17

Missing rates were low, with most of the mass under 0.01. Pink bar correspond to markers that were exclusively present in the old array, i.e. 90% of the people did not have a value for those. Blue corresponds to markers only available in the new array, so. about 10% of participants did not have those genotypes measured.

Comparison of Allele Frequency with ExAC

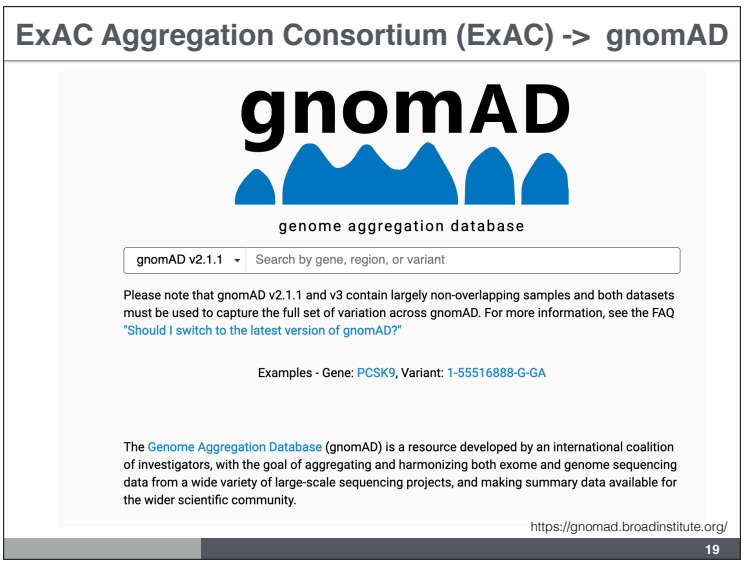




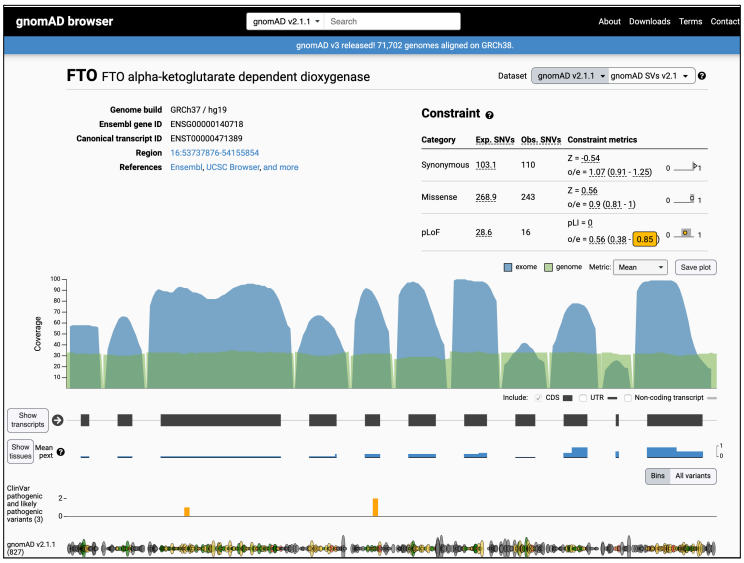
18

Allele frequencies of all variants were compared to "population" frequencies available from the ExAC consortium. The markers lie nearby the identity line, providing reassurance that the genotyping was reliable.

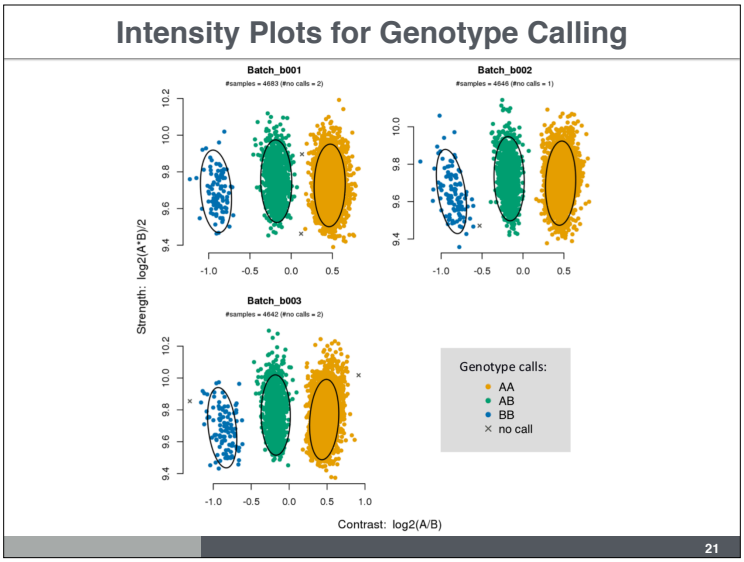
- Some high frequency in ExAC not found in UKB,
- very few the other way around, high frequency in UKB and not observed in ExAC.



The ExAC database, now renamed gnomAD, is a huge publicly available resource with summaries of a very large number of whole exome and whole genome sequenced data. This resource is critical to evaluate the pathogenicity of rare variants. For example, if a variant appears in relatively high numbers in this database, we can safely assume that reasonably healthy life is possible with the mutation. When first appeared, many variants that had been catalogued as highly pathogenic ended up being reclassified as variants of uncertain significance VUS.

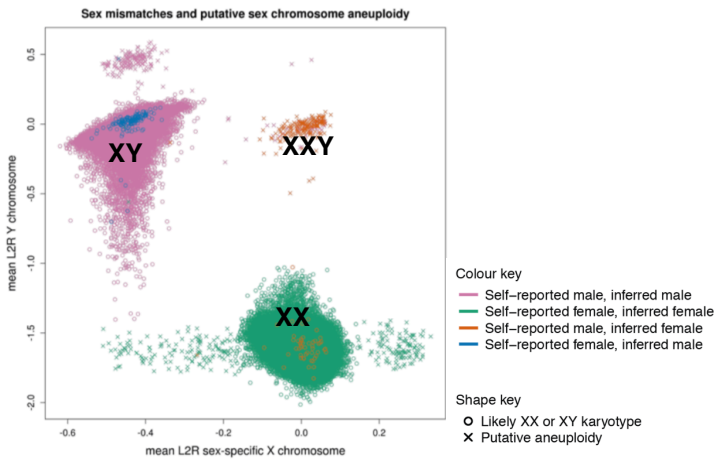


A snapshot of the gnomAD webpage search for the FTO gene.



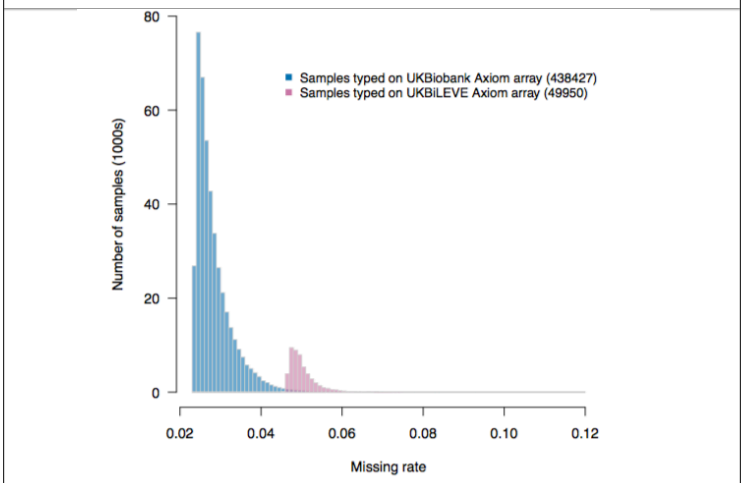
Here are examples of the intensity plots used for genotype calling. By plotting the strength vs the contrast of the intensities, we can visualize distinct clusters which are used for genotype calling.

Sex Mismatches and Sex Chromosome Aneuploidy



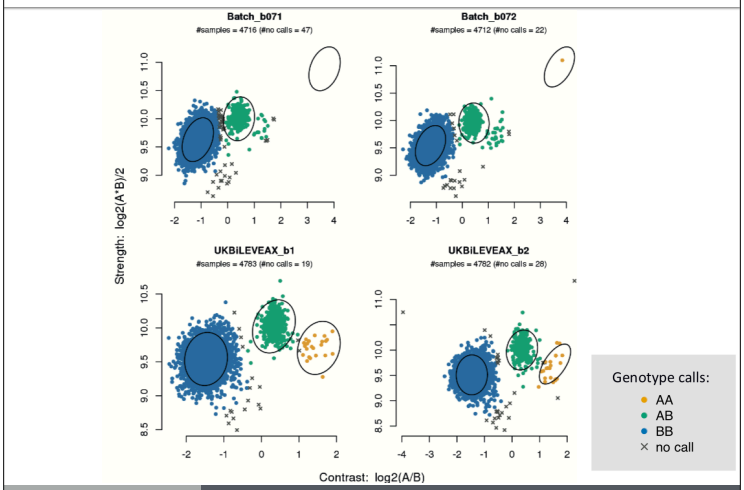
Average log intensities (normalized) of Y chromosome and X chromosome markers can help infer the sex of participants. Green cluster has "deficient" Y chromosome markers whereas the pink cluster shows X chromosome marker deficiency. XXY is centered at 0 due to the choice of normalization.

Missing Rates for Samples



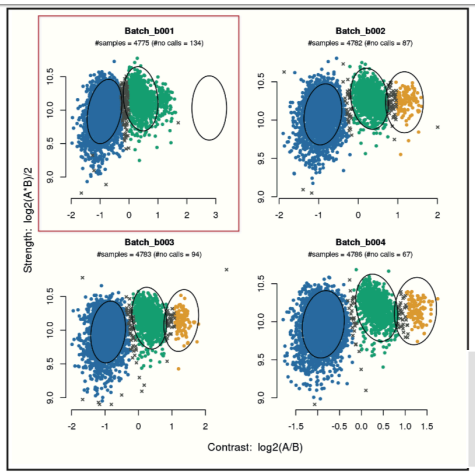
Missing rates are different for the first 50K individuals and the remaining driven by the difference in array. The first 50K were genotyped with the UkBiLEVE Axion array whereas the remaining 438K idnividuals were genotyped with the UK Bionbak Axiom array.

QC: Array Effect



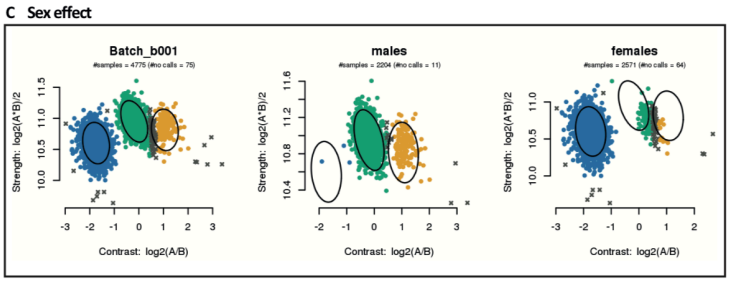
Two arrays (UKBiLEVE and UKBiobank Axiom Arrays). This marker has an outlier for UKBiLEVE batch that is not present in the newer array.

QC: Batch Effect (109 batches)



Genotype calls:
- AA
- AB
- BB
- × no call

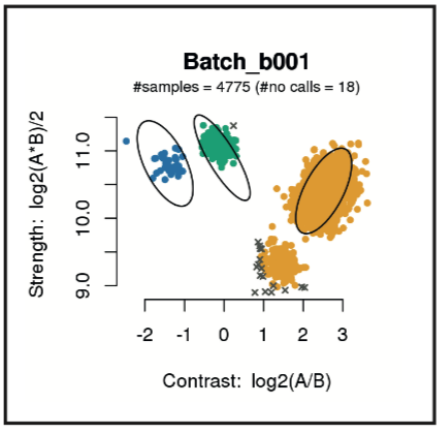
QC: Sex Effect



Genotype calls:
- AA
- AB
- BB
- × no call

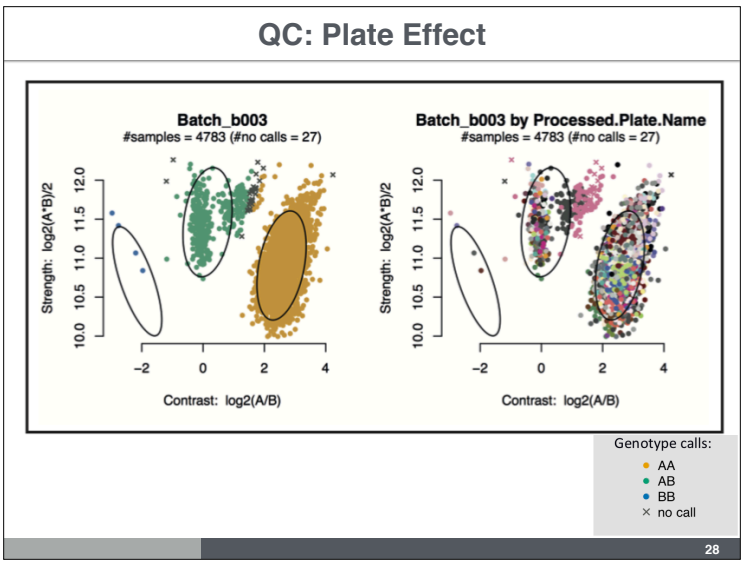
Data points cluster by sex rather than genotype. Unreliable.

Hardy Weinberg Disequilibrium

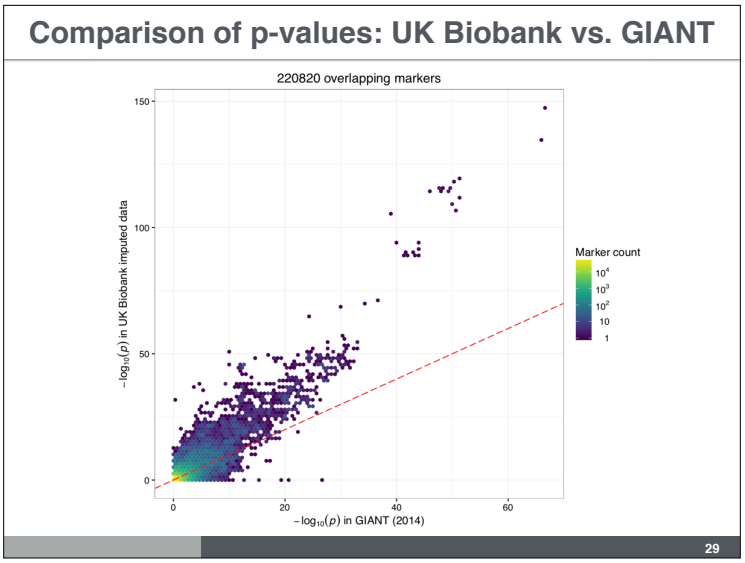


Genotype calls:
- AA
- AB
- BB
- × no call

Example of a marker that does not pass HWE test



Right figure shows plates with different colors. Pink plate data clusters (right figure) on its own cluster messing up the calling.



GWAS results of height phenotype in UKB are compared to an independent GWAS of height from the GIANT consortium. UKB p values are more significant than GIANT's p-values due to larger sample size in UKB as well as less heterogeneity.