# Factors for success in big data science

Damjan Vukcevic

**Data Science**
Murdoch Childrens Research Institute

16 October 2014
Big Data Reading Group
(Department of Mathematics & Statistics, University of Melbourne)

Murdoch Childrens Research Institute
Healthier Kids. Healthier Future.

# About me

POSTERA CRESCAM LAUDE
THE UNIVERSITY OF
MELBOURNE

UNIVERSITY OF
OXFORD

Experian
Hitwise

Murdoch Childrens
Research Institute
Healthier Kids. Healthier Future.

2001

2005

2008

2010

2012

Mathematics
Statistics

Statistical genetics

Web analytics

Statistical genetics
Biostatistics

Bioinformatics

# About this talk

- Examples of ‘Big Data’ science research projects
- Highlight some aspects and common features
- Personal view on the key factors for success
- Suggest some educational reforms
- Recommend adopting a big ‘Data Science’ approach

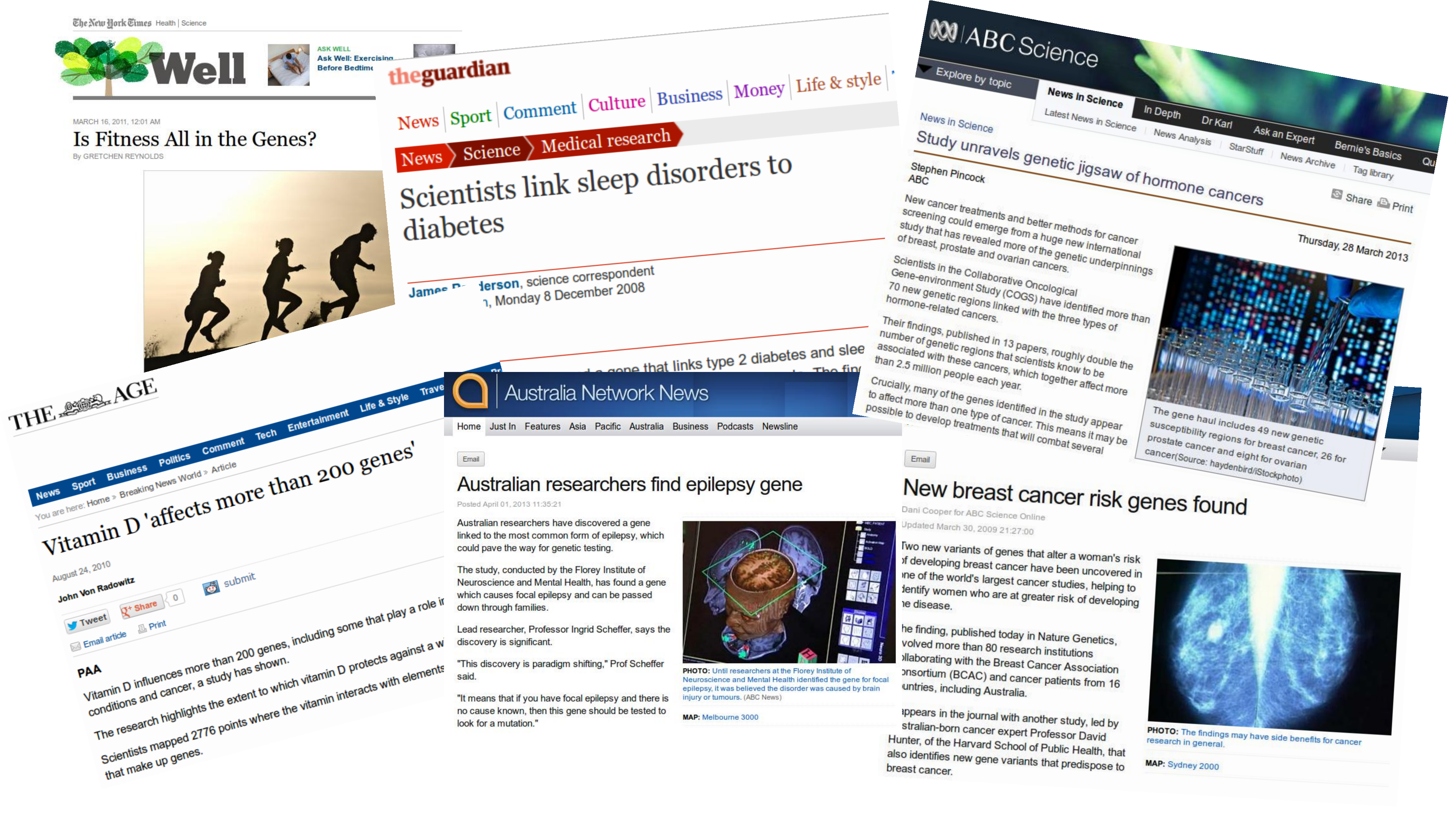
# Overview

1. Genome-wide associations studies (GWAS)
    a) Intro to genetics
    b) Overview of 3 studies

2. Factors for success

3. Statistical education & data science

# 1. Genome-wide association studies (GWAS)



The New York Times Health | Science
Well
ASK WELL
Ask Well: Exercising Before Bedtime

MARCH 16, 2011, 12:01 AM

## Is Fitness All in the Genes?

By GRETCHEN REYNOLDS

theguardian
News | Sport | Comment | Culture | Business | Money | Life & style
News › Science › Medical research

## Scientists link sleep disorders to diabetes

James R… lerson, science correspondent
…, Monday 8 December 2008

…a gene that links type 2 diabetes and slee… The fin…

ABC Science
Explore by topic
News in Science | In Depth | Dr Karl | Ask an Expert | Bernie's Basics | Qu…
Latest News in Science | News Analysis | StarStuff | News Archive | Tag library
News in Science

## Study unravels genetic jigsaw of hormone cancers

Stephen Pincock
ABC

Share Print

Thursday, 28 March 2013

New cancer treatments and better methods for cancer screening could emerge from a huge new international study that has revealed more of the genetic underpinnings of breast, prostate and ovarian cancers.

Scientists in the Collaborative Oncological Gene-environment Study (COGS) have identified more than 70 new genetic regions linked with the three types of hormone-related cancers.

Their findings, published in 13 papers, roughly double the number of genetic regions that scientists know to be associated with these cancers, which together affect more than 2.5 million people each year.

Crucially, many of the genes identified in the study appear to affect more than one type of cancer. This means it may be possible to develop treatments that will combat several

The gene haul includes 49 new genetic susceptibility regions for breast cancer, 26 for prostate cancer and eight for ovarian cancer(Source: haydenbird/iStockphoto)

THE AGE
News | Sport | Business | Politics | Comment | Tech | Entertainment | Life & Style | Trave…

You are here: Home » Breaking News World » Article

## Vitamin D 'affects more than 200 genes'

August 24, 2010

John Von Radowitz

PAA

Vitamin D influences more than 200 genes, including some that play a role i… conditions and cancer, a study has shown.

The research highlights the extent to which vitamin D protects against a w…

Scientists mapped 2776 points where the vitamin interacts with elements… that make up genes.

Australia Network News
Home | Just In | Features | Asia | Pacific | Australia | Business | Podcasts | Newsline
Email

## Australian researchers find epilepsy gene

Posted April 01, 2013 11:35:21

Australian researchers have discovered a gene linked to the most common form of epilepsy, which could pave the way for genetic testing.

The study, conducted by the Florey Institute of Neuroscience and Mental Health, has found a gene which causes focal epilepsy and can be passed down through families.

Lead researcher, Professor Ingrid Scheffer, says the discovery is significant.

"This discovery is paradigm shifting," Prof Scheffer said.

"It means that if you have focal epilepsy and there is no cause known, then this gene should be tested to look for a mutation."

PHOTO: Until researchers at the Florey Institute of Neuroscience and Mental Health identified the gene for focal epilepsy, it was believed the disorder was caused by brain injury or tumours. (ABC News)

MAP: Melbourne 3000

## New breast cancer risk genes found

Dani Cooper for ABC Science Online

Updated March 30, 2009 21:27:00

Two new variants of genes that alter a woman's risk …f developing breast cancer have been uncovered in …ne of the world's largest cancer studies, helping to …dentify women who are at greater risk of developing …e disease.

…he finding, published today in Nature Genetics, …volved more than 80 research institutions …llaborating with the Breast Cancer Association …onsortium (BCAC) and cancer patients from 16 …untries, including Australia.

…ppears in the journal with another study, led by …stralian-born cancer expert Professor David Hunter, of the Harvard School of Public Health, that also identifies new gene variants that predispose to breast cancer.

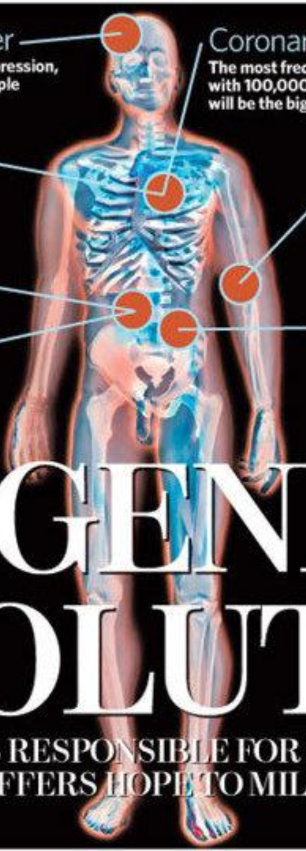
PHOTO: The findings may have side benefits for cancer research in general.

MAP: Sydney 2000

(Ireland, €1) 70p

Thursday 7 June 2007
www.independent.co.uk
• NUMBER 6,44

THE INDEPENDENT



Bipolar disorder
Also known as manic depression, it affects 100 million people around the world

Coronary heart disease
The most frequent cause of death in Britain, with 100,000 victims every year. By 2020, it will be the biggest killer in the world

Hypertension
High blood pressure affects 16 million people in Britain. Can lead to stroke, heart disease and kidney failure

Rheumatoid arthritis
Nearly 400,000 people in Britain are afflicted with this auto-immune disease of the joints

Type 1 diabetes
Diabetic condition in which sufferers have to inject insulin. Affects 350,000 people in UK

Crohn's disease
Up to 60,000 people are affected by this debilitating bowel condition which can cause distress and pain for a lifetime

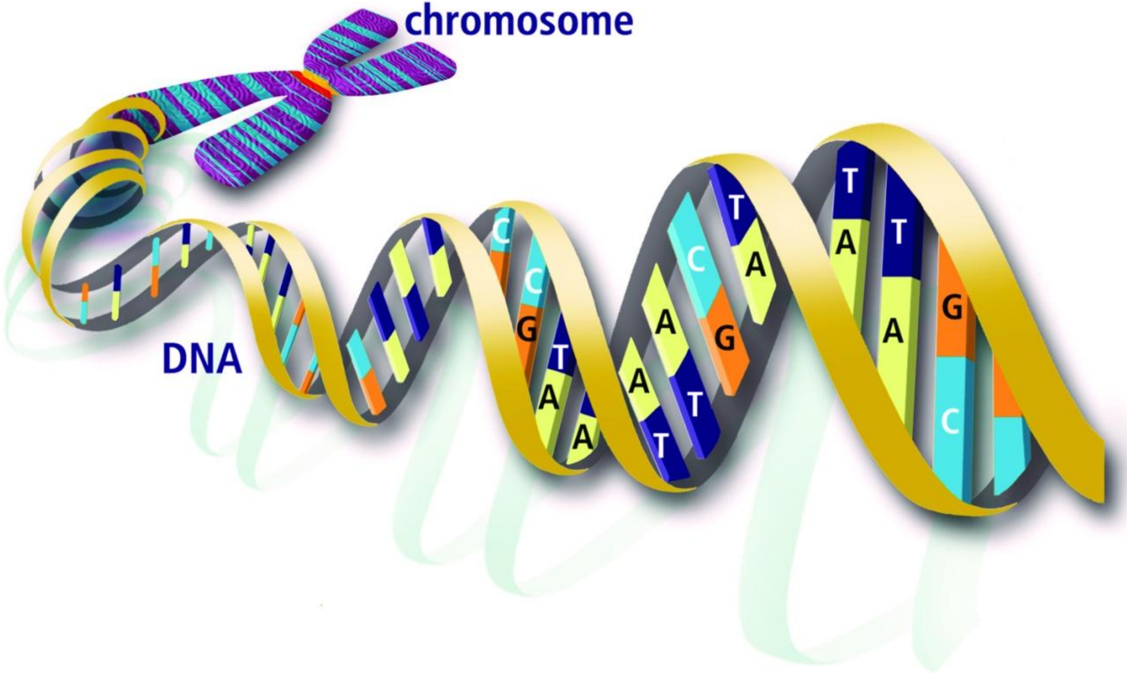
Type 2 diabetes
Almost 2 million Britons are affected by this late-onset disease, which is linked with the growing obesity epidemic

# THE GENETIC REVOLUTION

## DISCOVERY OF GENES RESPONSIBLE FOR SEVEN OF THE MOST COMMON ILLNESSES OFFERS HOPE TO MILLIONS OF SUFFERERS

FULL STORY, PAGE 2

# Human genome



Biological and Environmental Research Information System, Oak Ridge National Laboratory, genomicscience.energy.gov and genomics.energy.gov

# Human genome

- Total length = 3 billion bases/nucleotides
- Each person inherits 2 complete copies (one each from mother & father)

# DNA fragment

…TAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTAAACCTGATCAA…

…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTAAACCTGATCAA…

# Single nucleotide polymorphisms (SNPs)

…TAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGAGATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGAAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTAAACCTGATCAA…
…TAACGCGATAAGATATTAGCCCAAAAACACAGACATGGTAATAGCGTAAACCTGATCAA…

# Single nucleotide polymorphisms (SNPs)

```
--------------G------------------------A---------------------
--------------G------------------------A---------------------
--------------G------------------------A---------------------
--------------G------------------------A---------------------
--------------T------------------------A---------------------
--------------T------------------------A---------------------
--------------T------------------------A---------------------
--------------T------------------------T---------------------
--------------T------------------------T---------------------
--------------T------------------------T---------------------
--------------T------------------------T---------------------
```

# Single nucleotide polymorphisms (SNPs)

```
--------------0------------------------0---------------------
--------------0------------------------0---------------------
--------------0------------------------0---------------------
--------------0------------------------0---------------------
--------------1------------------------0---------------------
--------------1------------------------0---------------------
--------------1------------------------0---------------------
--------------1------------------------1---------------------
--------------1------------------------1---------------------
--------------1------------------------1---------------------
--------------1------------------------1---------------------
```

# Single nucleotide polymorphisms (SNPs)

*Individual 1*

```
-------------0------------------------0---------------------
-------------0------------------------0---------------------
```

*Individual 2*

```
-------------1------------------------0---------------------
-------------1------------------------1---------------------
```

*Individual 3*

```
-------------1------------------------1---------------------
-------------1------------------------1---------------------
```

# Single nucleotide polymorphisms (SNPs)

*Individual 1*

```
-------------0------------------------0---------------------
```

*Individual 2*

```
-------------2------------------------1---------------------
```

*Individual 3*

```
-------------2------------------------2---------------------
```

Count the 1 types at each SNP to create **genotypes**

# SNP facts

Best current knowledge:

- 10 million SNPs in the human genome
- One in every ~300 bases, on average
- (Total human genome = 3 billion bases)

Other facts:

- Nearby SNPs are correlated due to shared inheritance

# Genotyping arrays

https://commons.wikimedia.org/wiki/File:Affymetrix-microarray.jpg

# Overview of studies

1. **WTCCC (2007)**
   Case-control study of 7 diseases
2. **WTCCC (2010)**
   Case-control study of 8 diseases
3. **IMSGC & WTCCC2 (2011)**
   Meta-analysis of case-control studies for 1 disease

# Is this ‘Big Data’?

Four V’s:

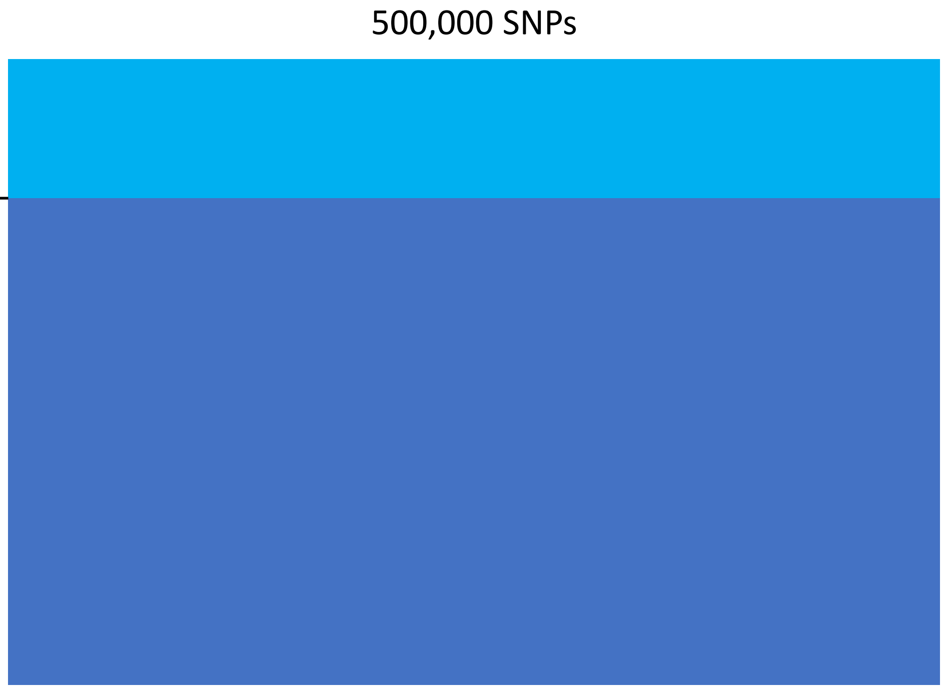
- **Volume** – scale of data
- **Velocity** – streaming data
- **Variety** – different forms of data
- **Veracity** – bias, noise, artefacts

Tell-tale signs:

- Need >1 computer
- Need >1 piece of software
- Need >1 analyst

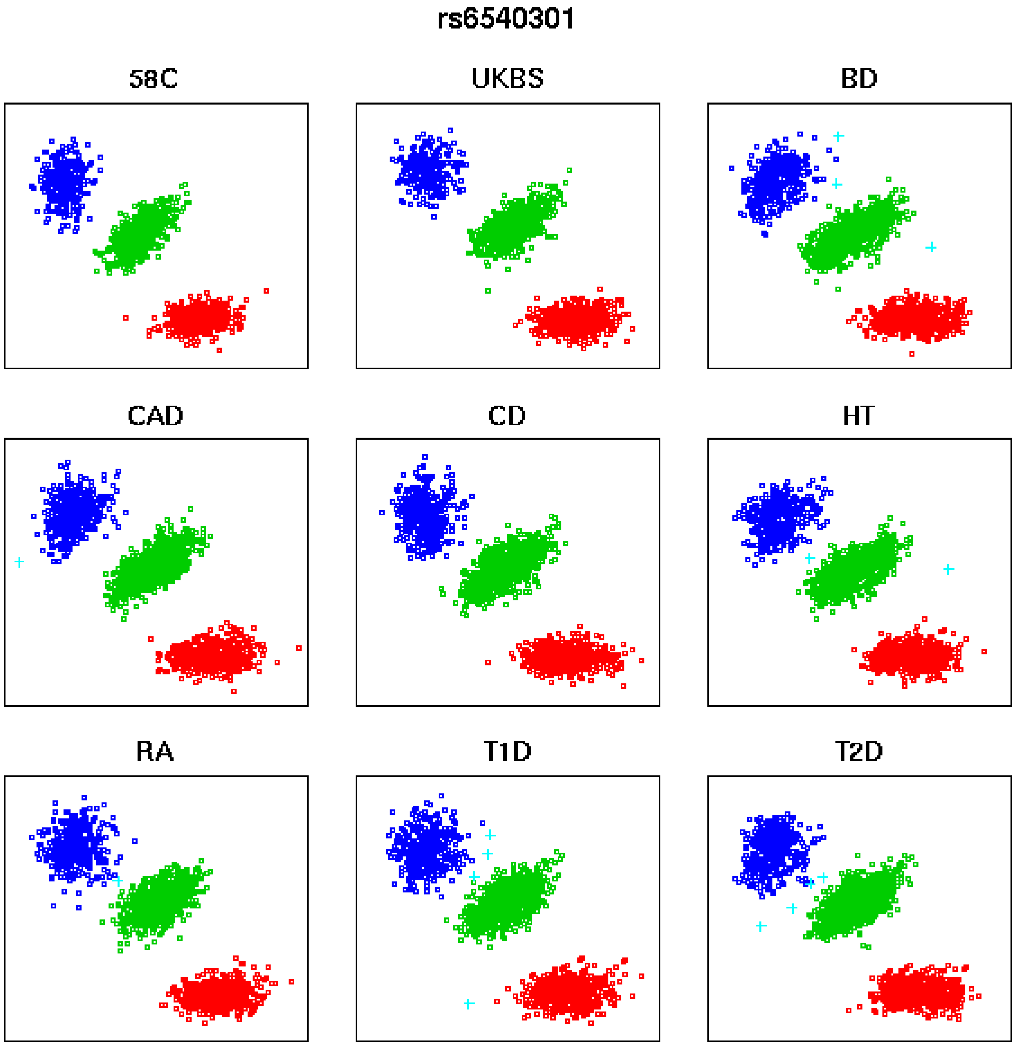
# WTCCC (2007) study design

| | |
|---|---|
| 3,000 controls | **1958 Birth Cohort** |
| | **UK Blood Service** |
| 2,000 cases | **Bipolar disorder** |
| 2,000 cases | **Coronary artery disease** |
| 2,000 cases | **Crohn’s disease** |
| 2,000 cases | **Hypertension** |
| 2,000 cases | **Rheumatoid arthritis** |
| 2,000 cases | **Type 1 diabetes** |
| 2,000 cases | **Type 2 diabetes** |



# Measuring SNPs

(X,Y) for each SNP for each individual



# Testing association

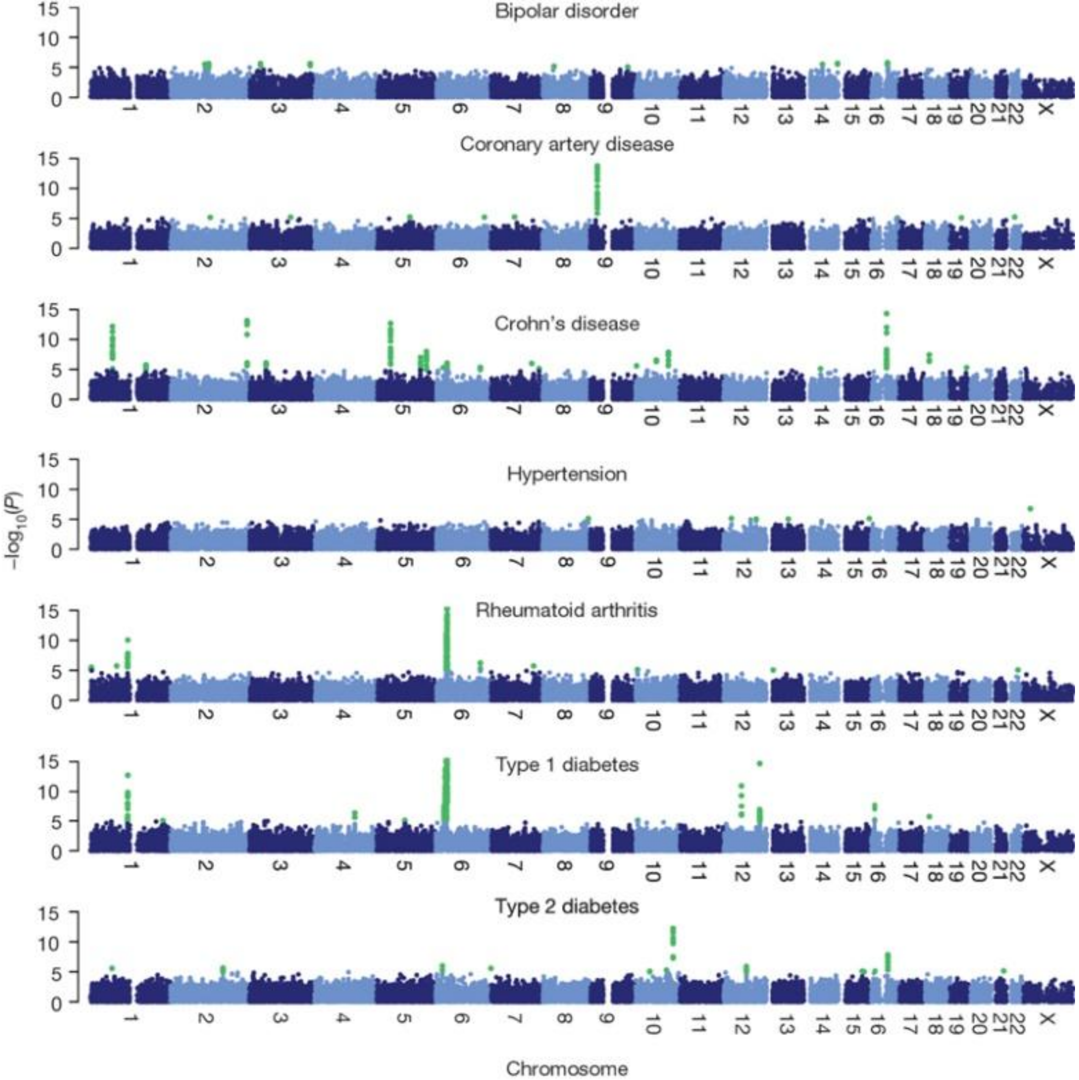
- Data: $3 \times 2$ contingency table at each SNP
- Test for association ($\chi^2$ with 1 degree of freedom)

| | **Genotype** | | |
|---|---|---|---|
| | **0** | **1** | **2** |
| **Cases** | 109 | 546 | 1659 |
| **Controls** | 89 | 478 | 1503 |

$\Rightarrow$ p-value

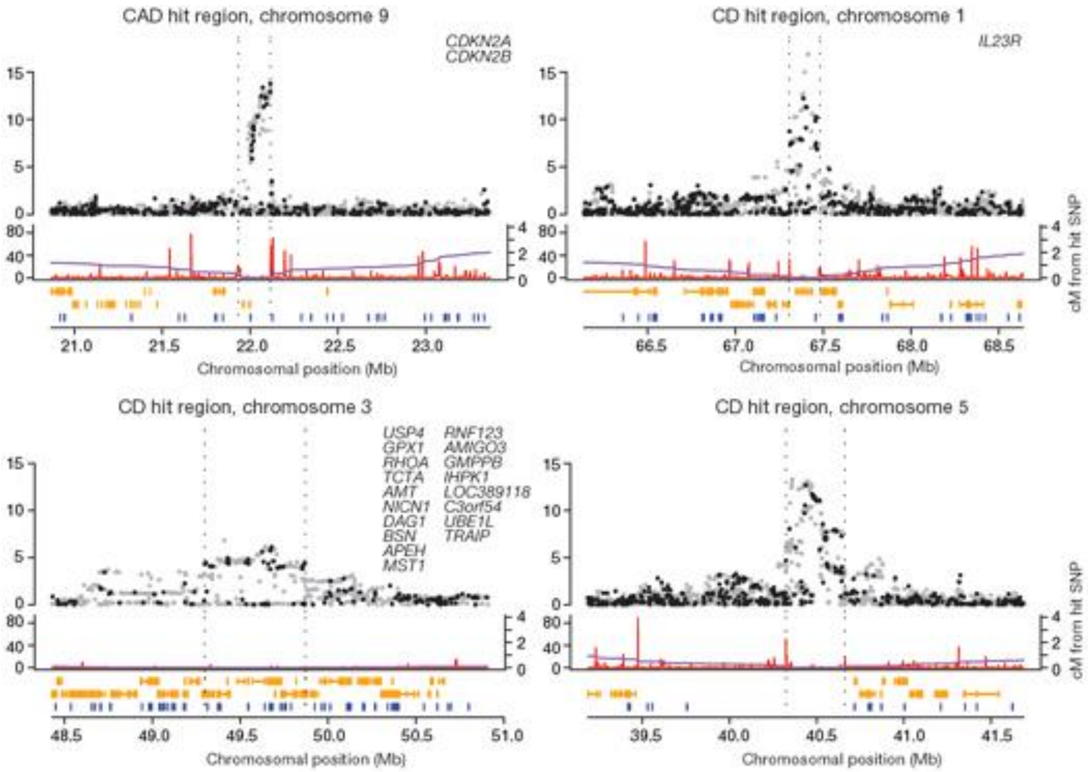
# Results

'Manhattan' plot



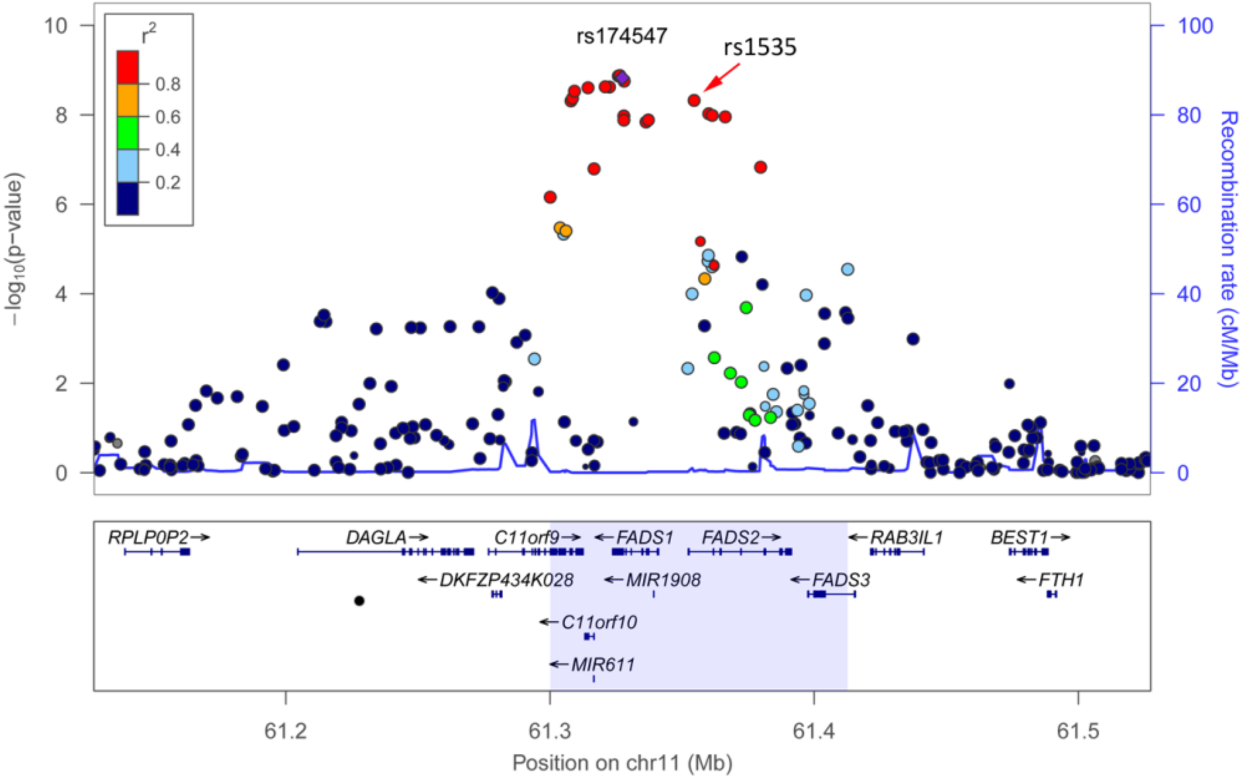
# Results

'Signal' plots



# Results

Signal plot from another study



Lettre *et al.* 2011

# Findings

- **Doubled** the number of known genetic associations (12 → 24)
- Found genetic effects present in **more than one** disease
- Hints of different genetic architectures for different disease classes: **autoimmune** vs **metabolic** vs **other**

*Definitely a success!*

# Inferring genotypes

‘Genotype calling’

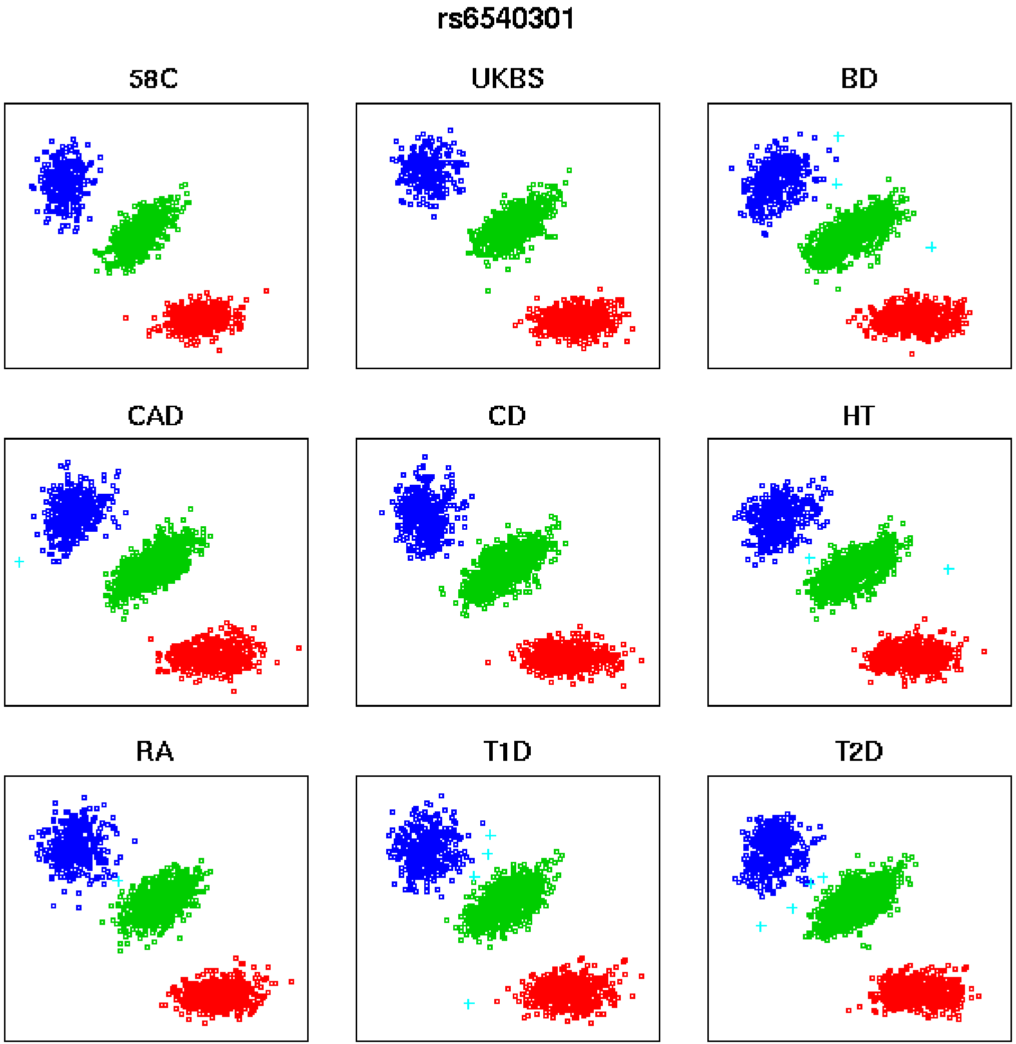
Designed new method (CHIAMO)

Hierarchical Bayesian clustering with informative priors

Used data from **all** individuals

Allowed for variation between cohorts

Showed Affymetrix data is actually reasonably good

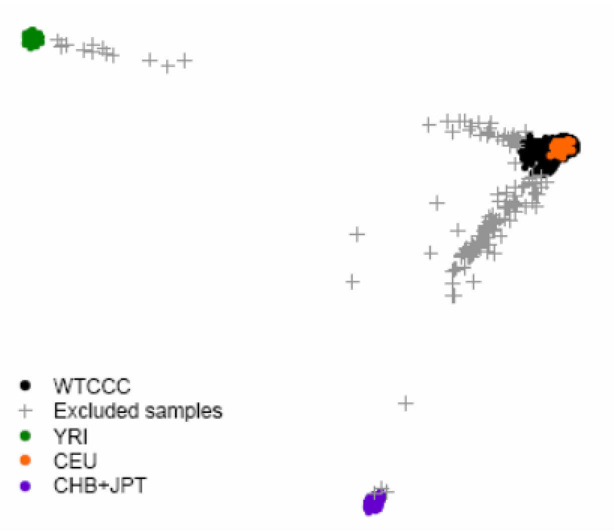


# Population structure

Principal components analysis (PCA)

Reference panel with known ancestry

Uses data across the whole genome



# Combination analyses

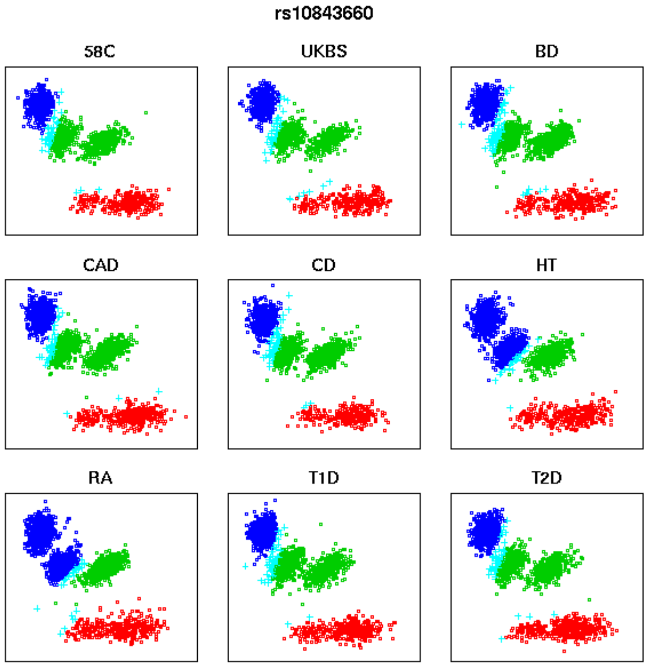
- **Combined cases**
  e.g. autoimmune diseases

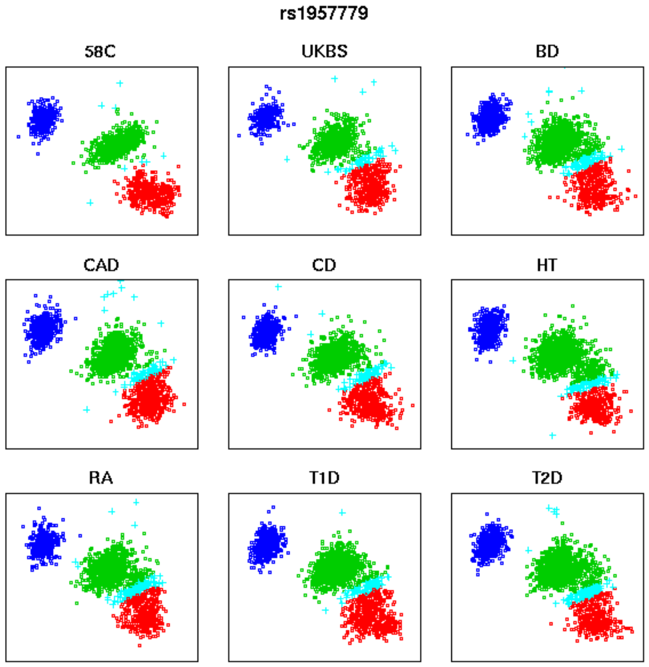
- **Combined controls**
  ('expanded reference set')

# Quality control (QC) & filtering

- Big data ⇒ 'rare' errors become numerous
- Artefacts and random noise unavoidable
- Systematic QC is mandatory
    - Samples
    - SNPs
    - Putative associations
- Automated & manual procedures

# ‘Cluster plot’ inspection





# QC ‘epic fail’

- The letter to Nature…

# Team

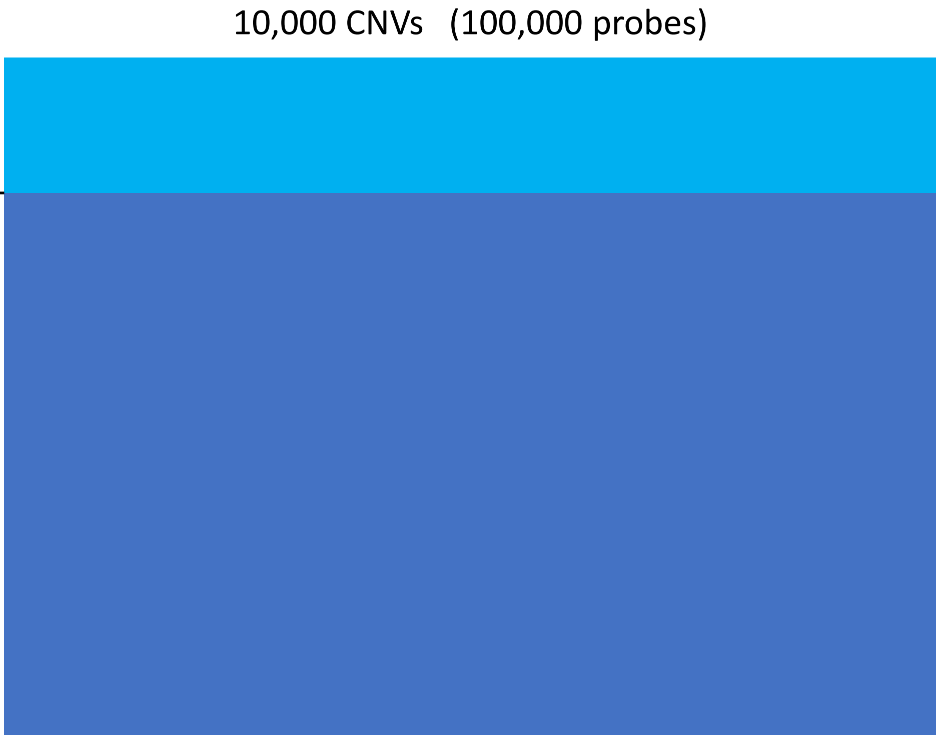
- 20 statisticians/analysts, across 4 institutions
- Full-time scientific programmer
- Diversity, parallelisation, and sometimes duplication of work
- Regular meetings
- Frequent collaboration and communication

# Computation

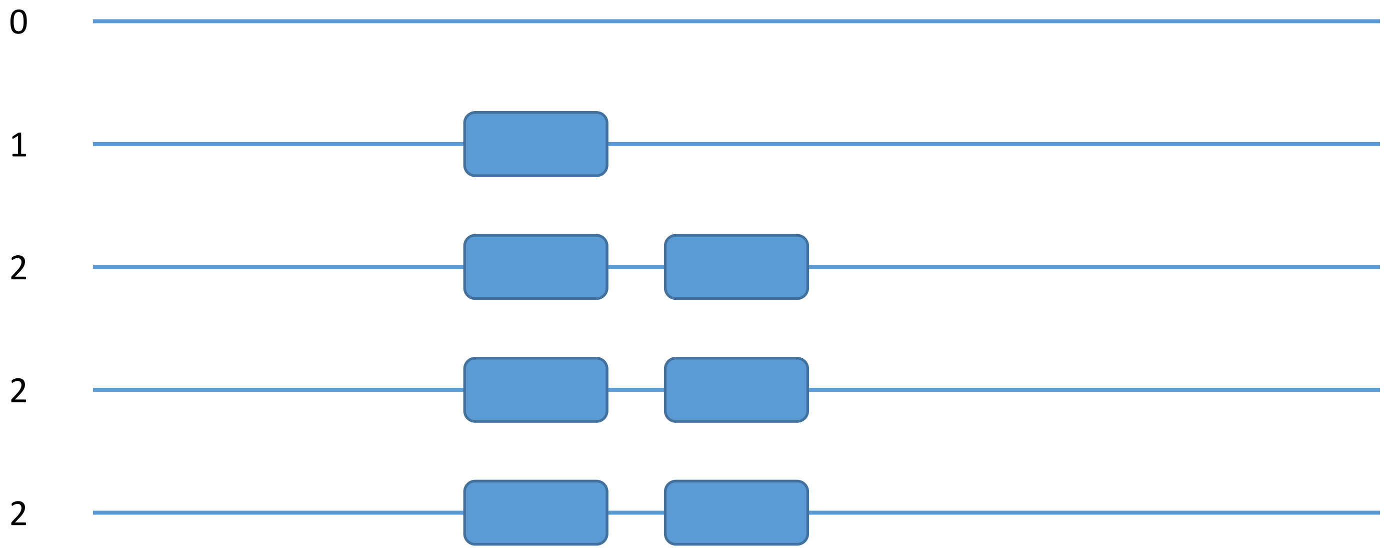
- Every statistician was also a programmer
- Computing cluster
- Multiple programming environments: C++, R, bash,…
- Developed a suite of software in tandem with analysis

# WTCCC (2010) study design

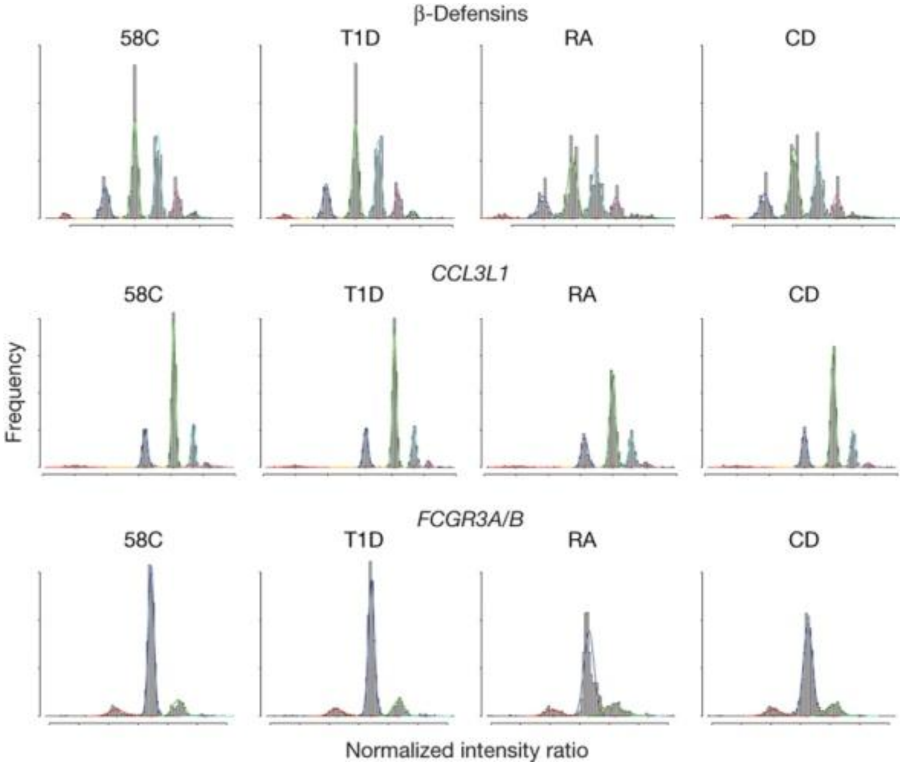
| | |
|---|---|
| 3,000 controls | **1958 Birth Cohort** |
| | **UK Blood Service** |
| 2,000 cases | **Bipolar disorder** |
| 2,000 cases | **Breast cancer** |
| 2,000 cases | **Coronary artery disease** |
| 2,000 cases | **Crohn's disease** |
| 2,000 cases | **Hypertension** |
| 2,000 cases | **Rheumatoid arthritis** |
| 2,000 cases | **Type 1 diabetes** |
| 2,000 cases | **Type 2 diabetes** |



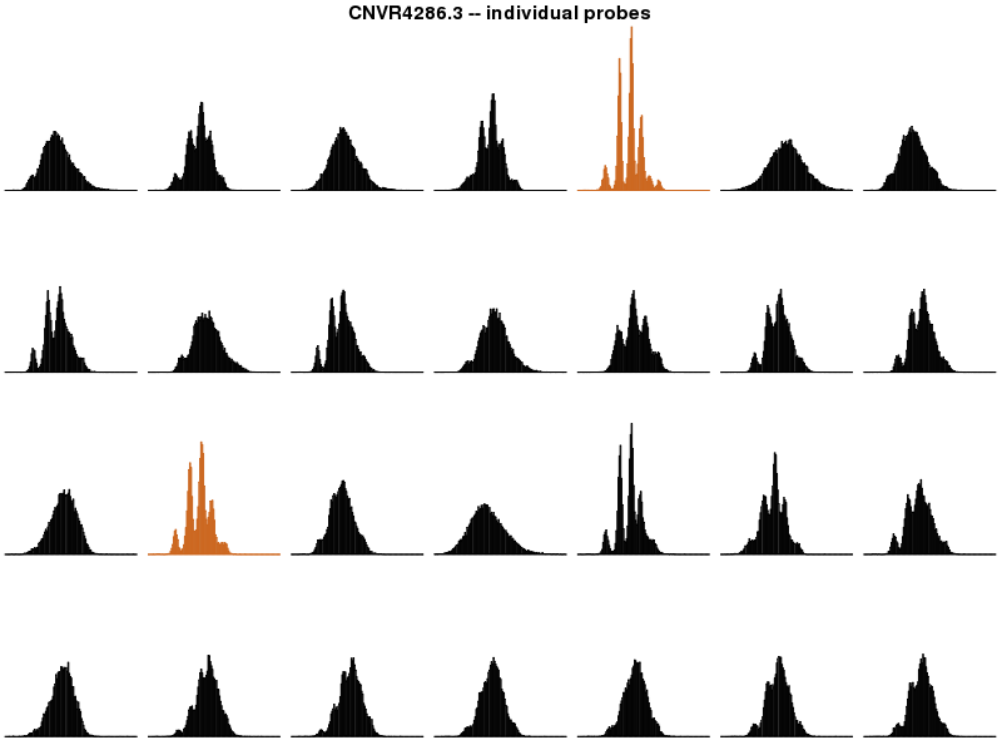
# Copy number variants (CNVs)



# Measuring CNVs



# Measuring CNVs

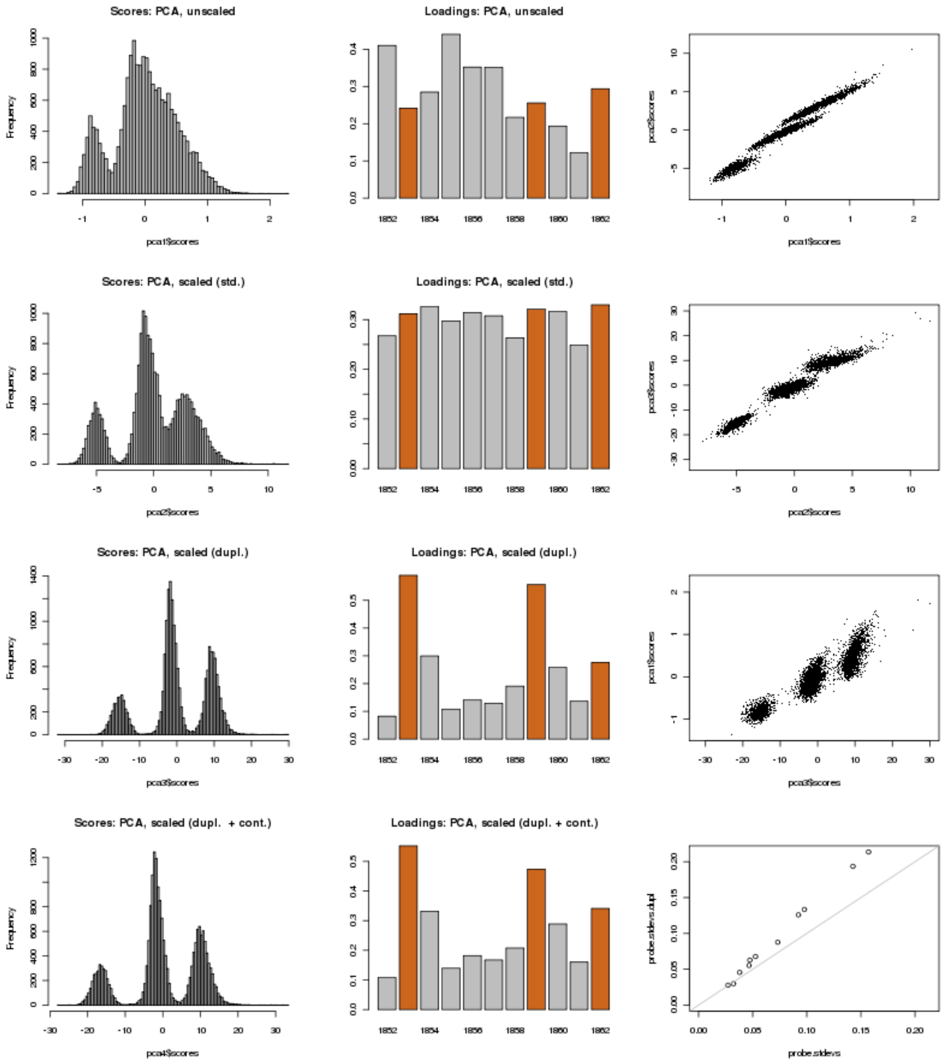


# Probe variance scaling

Replicate measurements
(duplicates & controls)

Use replicates to calculate per-probe variance

Rescale each probe



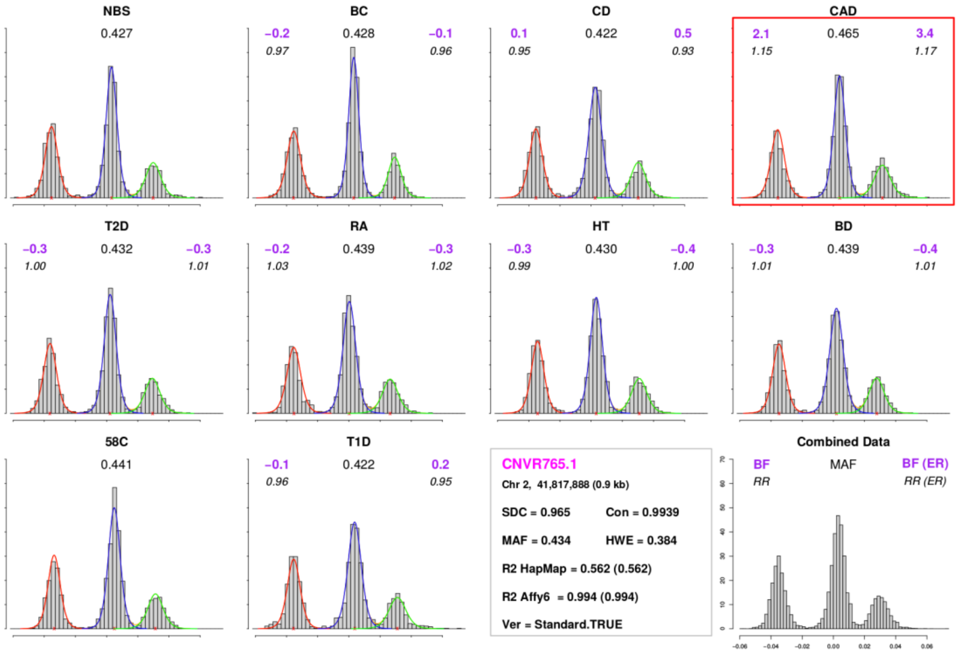
# Inferring (‘calling’) CNVs

Developed **two** different methods (Oxford vs Cambridge)

Methods were complementary

Served as sanity check

Boosted our confidence in our results

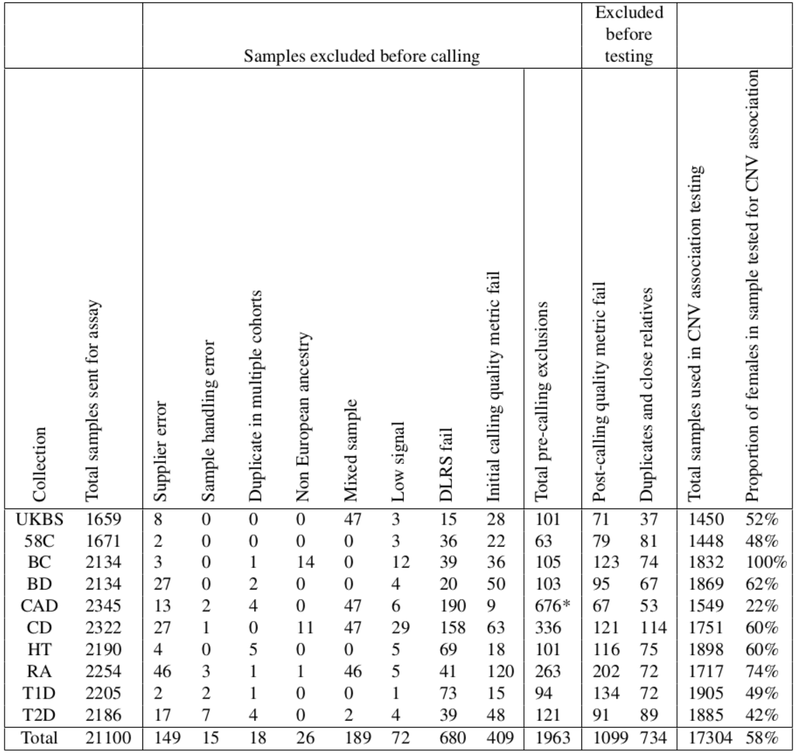


# Extensive QC

Multiple QC stages

Multiple QC criteria

Consumed by far the bulk of our time!



| | | Samples excluded before calling | | | | | | | | | Excluded before testing | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Collection | Total samples sent for assay | Supplier error | Sample handling error | Duplicate in multiple cohorts | Non European ancestry | Mixed sample | Low signal | DLRS fail | Initial calling quality metric fail | Total pre-calling exclusions | Post-calling quality metric fail | Duplicates and close relatives | Total samples used in CNV association testing | Proportion of females in sample tested for CNV association |
| UKBS | 1659 | 8 | 0 | 0 | 0 | 47 | 3 | 15 | 28 | 101 | 71 | 37 | 1450 | 52% |
| 58C | 1671 | 2 | 0 | 0 | 0 | 0 | 3 | 36 | 22 | 63 | 79 | 81 | 1448 | 48% |
| BC | 2134 | 3 | 0 | 1 | 14 | 0 | 12 | 39 | 36 | 105 | 123 | 74 | 1832 | 100% |
| BD | 2134 | 27 | 0 | 2 | 0 | 0 | 4 | 20 | 50 | 103 | 95 | 67 | 1869 | 62% |
| CAD | 2345 | 13 | 2 | 4 | 0 | 47 | 6 | 190 | 9 | 676* | 67 | 53 | 1549 | 22% |
| CD | 2322 | 27 | 1 | 0 | 11 | 47 | 29 | 158 | 63 | 336 | 121 | 114 | 1751 | 60% |
| HT | 2190 | 4 | 0 | 5 | 0 | 0 | 5 | 69 | 18 | 101 | 116 | 75 | 1898 | 60% |
| RA | 2254 | 46 | 3 | 1 | 1 | 46 | 5 | 41 | 120 | 263 | 202 | 72 | 1717 | 74% |
| T1D | 2205 | 2 | 2 | 1 | 0 | 0 | 1 | 73 | 15 | 94 | 134 | 72 | 1905 | 49% |
| T2D | 2186 | 17 | 7 | 4 | 0 | 2 | 4 | 39 | 48 | 121 | 91 | 89 | 1885 | 42% |
| Total | 21100 | 149 | 15 | 18 | 26 | 189 | 72 | 680 | 409 | 1963 | 1099 | 734 | 17304 | 58% |

## 5 Quality control procedures

### 5.1 Sample quality control filters

Two sample exclusion lists were constructed and used in the analysis of the data. The first list (pre-calling exclusion list) was used to exclude samples from the final calling of the CNVs using the processed intensity data. The second list (pre-testing exclusion list) was used to exclude samples from the testing for CNV association based on the final set of CNV calls. A full break down of excluded samples is given in Supplementary Table 8.

#### Pre-calling exclusions

1963 samples were excluded from the final CNV calling based on several different criteria described below. Some of the filters were applied to the raw intensity data while others were based on CNV calls obtained from an initial calling run on the data.

**Supplier error** 149 samples were excluded due to evidence that the samples were not the same as those indicated by the supplier manifest. Sequenom QC and calling gender on the CNV array were used to confirm these discrepancies.

**Sample handling error** 15 samples were excluded due to evidence of an error during arraying the samples for CNV screening.

**Multi-cohort duplicates** 18 samples (9 pairs) were detected that showed high correlation with another sample from a different cohort, indicating a sample that has genuinely been collected twice as the patient has at least two of diseases. No sample handling issue could be detected, and the data matched for both samples with the Sequenom and WTCCC1 SNP data. Both samples in the pair were excluded. The samples were identified by taking the summarised probe-level signal (first principal component) over 1,500 good quality polymorphic CNVs and running an all-vs-all correlation analysis (Pearson) to identify highly correlated samples.

**Non-European samples** 26 samples were excluded due to evidence of non-European ancestry. A PCA analysis was carried out on CNV calls from an initial calling run, that included HapMap individuals from the CEU, YRI and JPT+CHB panels. Examination of the loadings and scores of this analysis indicated that only the first principal component was discriminating European samples from the YRI and JPT+CHB samples. Supplementary Figure 12 shows the scores for each sample from the first principal component and highlights 14 outlying BC samples that were excluded. A further 11 CD samples and 1 RA samples were also excluded based on self-reported ancestry information.

**Mixed sample** 189 samples were excluded due to the samples having a high correlation with another sample on the same well of the screening plate pair or an adjacent well in the same plate suggesting that these samples consist of a mixture of DNA from two or more non-identical individuals.

**Low signal** 72 samples were excluded due to having a low signal intensity for either the green or the red channel ( $< 100$). The precise quantities used are the metrics named "SignalIntensityRed" and "SignalIntensityGreen" from the Agilent Feature Extraction software[109]. These give a measure of the median background-subtracted red and green channel signals respectively (not logged) across all non-control probes on the array.

**High derivative log ratio spread** Samples were excluded based on a measure of the variability in log-ratio ($\log_2(R/G)$) across all probes for each sample. The Agilent DLRS metric was used which is measures the spread of the differences between the log ratio values of consecutive probes[109]. High values of this metric indicate a poor sample. We excluded samples if DLRS was either $> 0.35$, or $> 0.3$ if it is a repeat and the original sample had a DLRS $> 0.35$.

**Outlying CAD samples** 405 CAD samples were identified that noticeably reduced the ability to distinguish different CNV classes when the samples were included. Removing these samples lead to a clear improvement in the ability to cluster some CNVs in the CAD cohort. This problem was observed for multiple probes in this study and is illustrated in Supplementary Figure 13 (see first and second panels) where we extracted from CNV ILMN_1M_4 a subset of probes (A_16_P30155705, chr1_047654910_047654955, A_16_P30155706, chr1_047654921_047654966, chr1_047654923_047654968, A_16_P30155708) that showed no sign of CNV polymorphism in the non CAD cohorts. However, a set of CAD samples was clearly separated from the main distribution at these probes.

To identify the subset of problematic CAD samples we used two probe sets (average signal for ILMN_1M_4 probes described above and probes A_18_P20232231, A_16_P40333900, A_16_P02994736 in CNV CNVR6314.1) outside of CNV regions for which the separation of outlying CAD samples was particularly obvious. For both probe sets, we manually set cutoffs for the mean normalized signal value and we excluded samples that exceeded both cutoffs (see the third panel of Supplementary Figure 13 with excluded samples marked in red).

Further analysis of the processing pipeline indicated that the likely source of the problem was mis-calibrated DNA concentration. Variable DNA concentrations differentially affected each probe, thus altering the within sample probe intensity rankings. In quantile normalisation, probe intensities were first ranked within the sample, and each intensity data point was then replaced by the appropriate quantile of the marginal distribution of probe intensities over all samples. Therefore, altered probe rankings eventually affected the normalized signal distribution.

**Initial-calling quality metric** 409 samples were identified based on 3 metrics designed to measure the quality of samples from an initial set of calls. The three metrics were (a) average CNV call rate measured as the proportion of CNV calls made on each sample using a calling threshold of 0.95, (b) average posterior probability of the most likely CNV class across all CNVs for a sample, and (c) average log-density (from the final model fit after merging) across all CNVs for a sample. Samples were ranked according to the minimum of the ranks on these three metrics and sample excluded so that the total number of exclusions was 2% of the total sample size.

#### Pre-testing exclusions

A further 1832 samples were excluded before testing for association of CNVs with the disease phenotypes. This resulted in a total of 17304 samples used in testing.

**Post-calling quality metric** 1099 samples were excluded based on thresholding three metrics applied to a final set of calls from the CNVCALL and CNVtools standard calling pipelines.

**Dispersion metric** A set of hard calls were made using CNVtools. A hard call is the genotype with the maximum likelihood given the estimates of the model parameters. For each CNV these hard calls were used to generate empirical means and standard deviations of the components that individuals were assigned to (the sample means conditional on the calls). Then for each individual at each CNV the absolute distance from the mean of the distribution that individual was assigned to was calculated. These were then averaged across CNVs to get the dispersion statistic for each individual. A threshold of 1.3 was chosen after visual inspection, all individuals that exceeded this threshold were excluded from testing (see Supplementary Figure 14).

**Posterior** Probablistic calls were made at each CNV using CNVCALL. For each individual the probability of assignment to the most-likely (non-null) class was averaged across all the CNVs polymorphic after merging. A threshold of 0.967 was chosen after visual inspection, all individuals that failed to exceed this threshold were excluded from testing (see Supplementary Figure 15).

**Heterozygosity** Using hard-calls from the CNVCALL (thresholded at a value of 0.95) the proportion of heterozygote calls in each individual was calculated on the CNVs polymorphic after merging. As this is a sum of independent binomials the Central Limit Theorem Applies. Modelling this as a normal distribution using the median as a robust estimator of the mean of the distribution, individuals were excluded if they lay in either tail with the probability of exclusion set at 1/2000 under the null (see Supplementary Figure 16).

**Duplicates and close relatives** 734 samples were excluded because they were identified to be duplicates or closely related samples. Samples from the same individual (duplicated samples) were identified as those having a calls correlation (using hard calls at a 0.95 threshold) of $> 0.9$. Closely related samples were identified as those having a calls correlation of between 0.6 and 0.9. Supplementary Figure 17 shows a plot of maximum calls correlation for each sample with any other sample. For each set of samples from the same individual, only the sample with the highest average posterior was retained. Likewise, for closely related samples from the same collection, only the sample with the highest average posterior was retained.

### 5.2 CNV quality control filters

We used 16 different analysis pipelines where different aspects of the data pre-processing were varied. Sup-

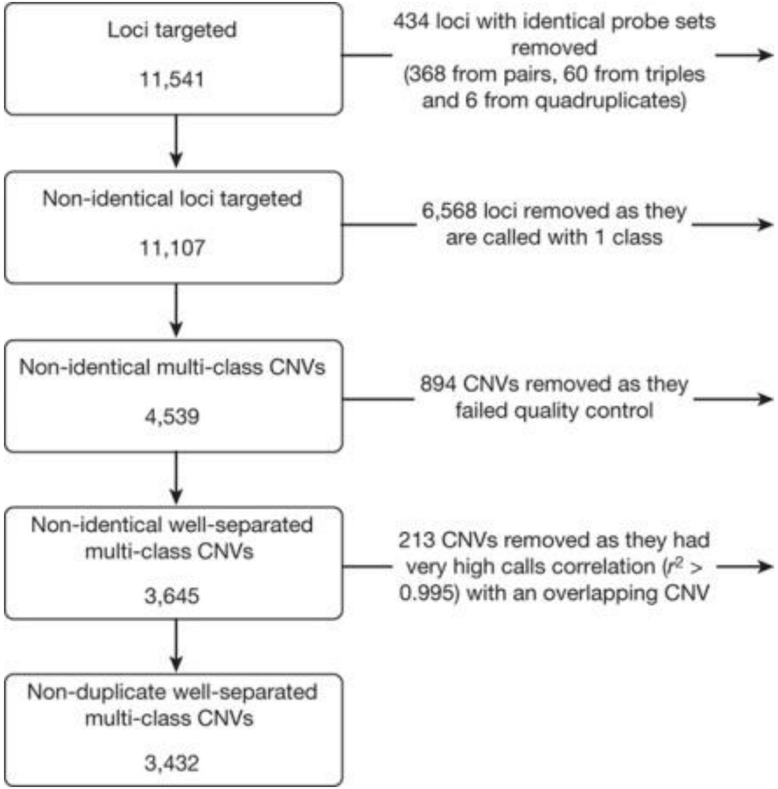
# Pipelines

16 normalisation schemes

2 calling algorithms

No single method always the best

Run them all, pick the best for each CNV



Loci targeted

11,541

434 loci with identical probe sets removed (368 from pairs, 60 from triples and 6 from quadruplicates)

Non-identical loci targeted

11,107

6,568 loci removed as they are called with 1 class

Non-identical multi-class CNVs

4,539

894 CNVs removed as they failed quality control

Non-identical well-separated multi-class CNVs

3,645

213 CNVs removed as they had very high calls correlation ($r^2 >$ 0.995) with an overlapping CNV

Non-duplicate well-separated multi-class CNVs

3,432

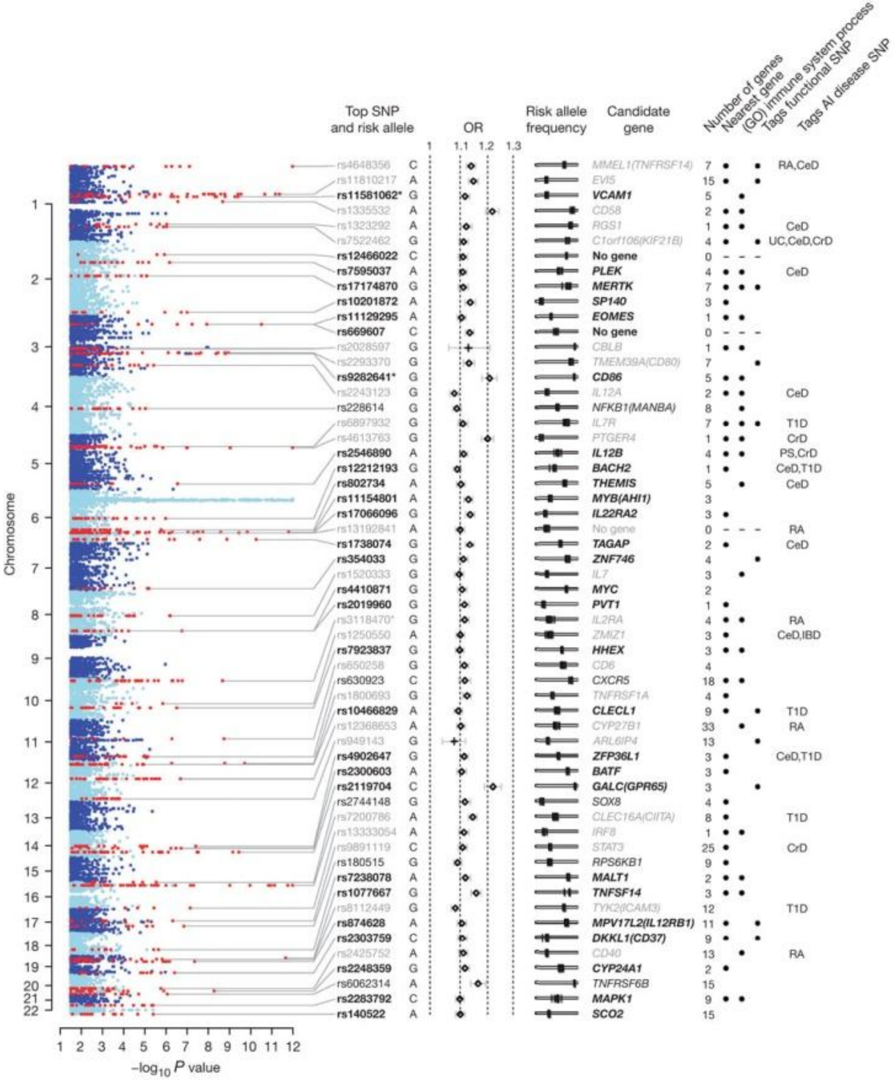
# IMSGC & WTCCC2 (2011) study design

Large GWAS meta-analysis:

- **23 research groups, from 15 countries**
- 10,000 cases (multiple sclerosis)
- 17,000 controls
- 460,000 SNPs

# Large meta-analysis

Big Data ⇒ many findings!



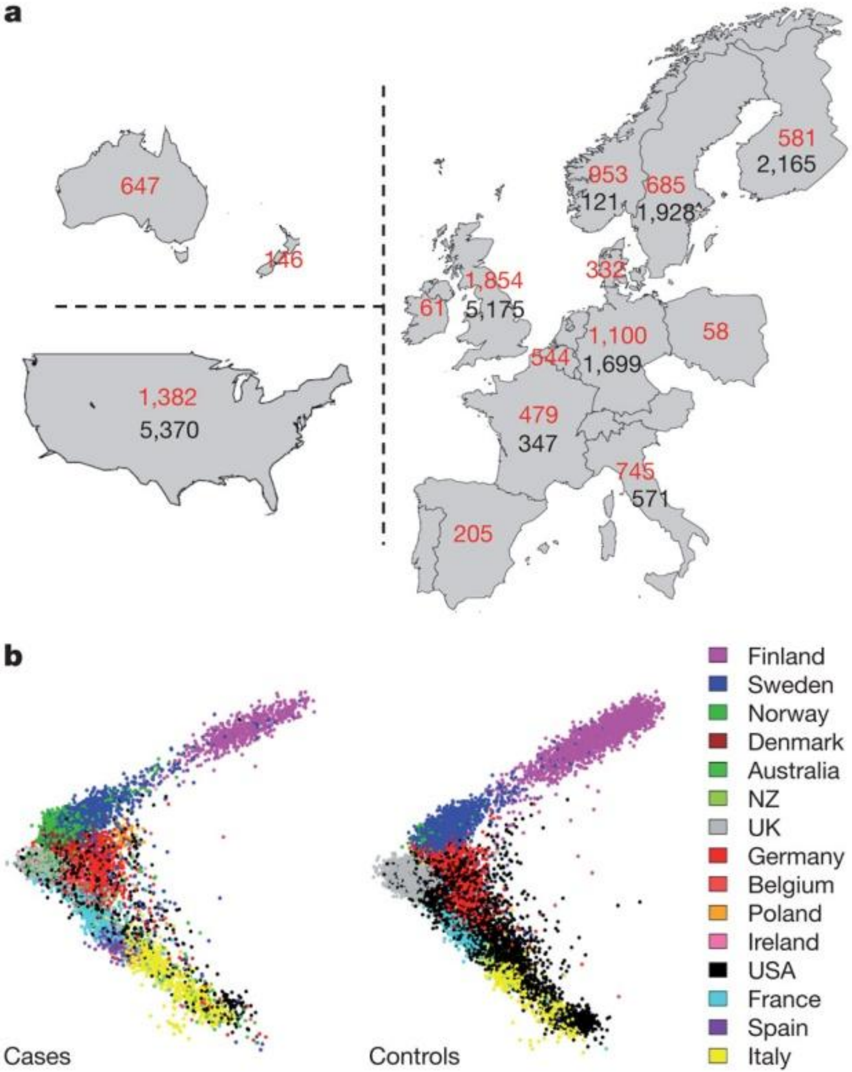
| Top SNP and risk allele | | Candidate gene | Number of genes | Nearest gene | (GO) immune system process | Tags functional SNP | Tags AI disease SNP |
|---|---|---|---|---|---|---|---|
| rs4648356 | C | *MMEL1(TNFRSF14)* | 7 | • | | • | RA,CeD |
| rs11810217 | A | *EVI5* | 15 | • | | • | |
| **rs11581062*** | G | ***VCAM1*** | 5 | | • | | |
| rs1335532 | A | *CD58* | 2 | • | • | | |
| rs1323292 | A | *RGS1* | 1 | • | • | | CeD |
| rs7522462 | G | *C1orf106(KIF21B)* | 4 | • | | • | UC,CeD,CrD |
| **rs12466022** | C | **No gene** | 0 | – | – | – | |
| **rs7595037** | A | ***PLEK*** | 4 | • | • | | CeD |
| **rs17174870** | G | ***MERTK*** | 7 | • | • | • | |
| **rs10201872** | A | ***SP140*** | 3 | • | | | |
| **rs11129295** | A | ***EOMES*** | 1 | • | • | | |
| **rs669607** | C | **No gene** | 0 | – | – | – | |
| rs2028597 | G | *CBLB* | 1 | • | • | | |
| rs2293370 | G | *TMEM39A(CD80)* | 7 | | | • | |
| **rs9282641*** | G | ***CD86*** | 5 | • | • | | |
| rs2243123 | G | *IL12A* | 2 | • | • | | CeD |
| rs228614 | G | *NFKB1(MANBA)* | 8 | | • | | |
| rs6897932 | G | *IL7R* | 7 | • | • | • | T1D |
| rs4613763 | G | *PTGER4* | 1 | • | • | | CrD |
| **rs2546890** | A | ***IL12B*** | 4 | • | • | | PS,CrD |
| **rs12212193** | G | ***BACH2*** | 1 | • | | | CeD,T1D |
| **rs802734** | A | ***THEMIS*** | 5 | | • | | CeD |
| **rs11154801** | A | ***MYB(AHI1)*** | 3 | | | | |
| **rs17066096** | G | ***IL22RA2*** | 3 | • | | | |
| rs13192841 | A | No gene | 0 | – | – | – | RA |
| **rs1738074** | G | ***TAGAP*** | 2 | • | | | CeD |
| **rs354033** | G | ***ZNF746*** | 4 | | | • | |
| rs1520333 | G | *IL7* | 3 | | • | | |
| **rs4410871** | G | ***MYC*** | 2 | | | | |
| **rs2019960** | G | ***PVT1*** | 1 | • | | | |
| rs3118470* | G | *IL2RA* | 4 | • | • | | RA |
| rs1250550 | A | *ZMIZ1* | 3 | • | | | CeD,IBD |
| **rs7923837** | G | ***HHEX*** | 3 | • | • | | |
| rs650258 | G | *CD6* | 4 | | | | |
| rs630923 | C | *CXCR5* | 18 | • | • | | |
| rs1800693 | G | *TNFRSF1A* | 4 | • | | | |
| **rs10466829** | A | ***CLECL1*** | 9 | • | | • | T1D |
| rs12368653 | A | *CYP27B1* | 33 | | • | | RA |
| rs949143 | G | *ARL6IP4* | 13 | | | • | |
| **rs4902647** | G | ***ZFP36L1*** | 3 | • | | | CeD,T1D |
| **rs2300603** | A | ***BATF*** | 3 | • | | | |
| **rs2119704** | C | ***GALC(GPR65)*** | 3 | | | • | |
| rs2744148 | G | *SOX8* | 4 | • | | | |
| rs7200786 | A | *CLEC16A(CIITA)* | 8 | • | | | T1D |
| rs13333054 | A | *IRF8* | 1 | • | • | | |
| rs9891119 | C | *STAT3* | 25 | • | | | CrD |
| rs180515 | G | *RPS6KB1* | 9 | • | | | |
| **rs7238078** | A | ***MALT1*** | 2 | • | • | | |
| **rs1077667** | G | ***TNFSF14*** | 3 | • | • | | |
| rs8112449 | G | *TYK2(ICAM3)* | 12 | | | | T1D |
| **rs874628** | A | ***MPV17L2(IL12RB1)*** | 11 | • | | • | |
| **rs2303759** | C | ***DKKL1(CD37)*** | 9 | • | | • | |
| rs2425752 | A | *CD40* | 13 | | • | | RA |
| **rs2248359** | G | ***CYP24A1*** | 2 | • | | | |
| rs6062314 | A | *TNFRSF6B* | 15 | | | | |
| **rs2283792** | C | ***MAPK1*** | 9 | • | • | | |
| **rs140522** | A | ***SCO2*** | 15 | | | | |

# Population structure

Multiple methods evaluated
(PCA covariates, genomic control, matching by clustering…)

Linear mixed model approach developed

Accounts for correlations due to multiple levels of relatedness

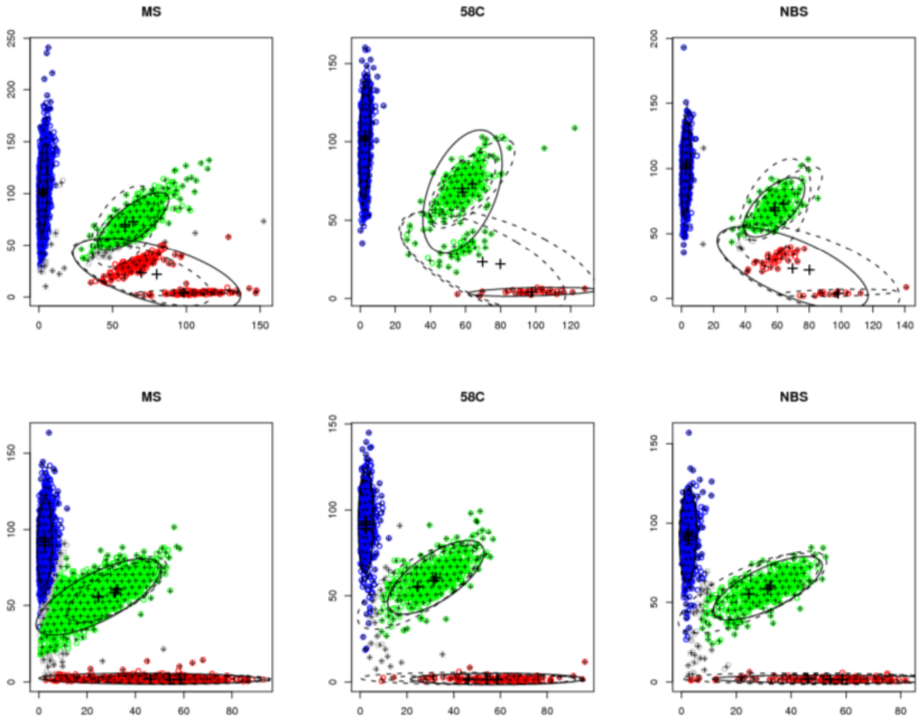


# Maturing QC

Increasing automation of QC procedures

Reducing human intervention

‘Automated cluster checking’,
using genotype calls from multiple cohorts



# 2. Factors for success

Informed by these studies and my general experience

# Factors in 3 parts

- Projects
- Methods
- People

# Projects

**The basics**

- Ask the right questions
- Collect relevant data
- Collect *quality* data

**Good experimental design**

- Replicates & controls
- Representative samples
- Use reference datasets

**Pragmatic analysis**

- Sanity checks and visualisation
- Systematic quality control
- Try multiple methods

**Capture the ‘Big’ value**

- Use all of the data
- Combine datasets
- Use reference datasets

# Methods

**Keep it real, make it easy**

- Solve a 'real' problem (i.e. one that people want solved)
- Provide a software implementation
- Write documentation
- Show examples

*Without an implementation, your method won't be used by practitioners, will be excluded in comparisons, and possibly ignored in reviews*

**Make it robust**

- Follow standards
- Implementation should work most of the time
- Cope with unexpected/unusual data
- Fail gracefully as a last resort

*Robustness beats optimality*

# People

**Statistical knowledge**

- Statistical insight, ‘data savvy’
- Knowledge of variety of methods

**Data analysis skills**

- Data management & manipulation
- Visualisation & exploratory analysis
- Can run a variety of methods

**Computational skills**

- Programming
- Unix & cluster computing
- Software engineering tools & principles (version control, code reusability)

**Collaboration & communication skills**

- Can work in teams
- Can talk to non-experts

# Factors with little impact

- Methods with no implementation
- Methods with no relevant real data examples
- Theoretical optimality

# 3. Statistical education & data science

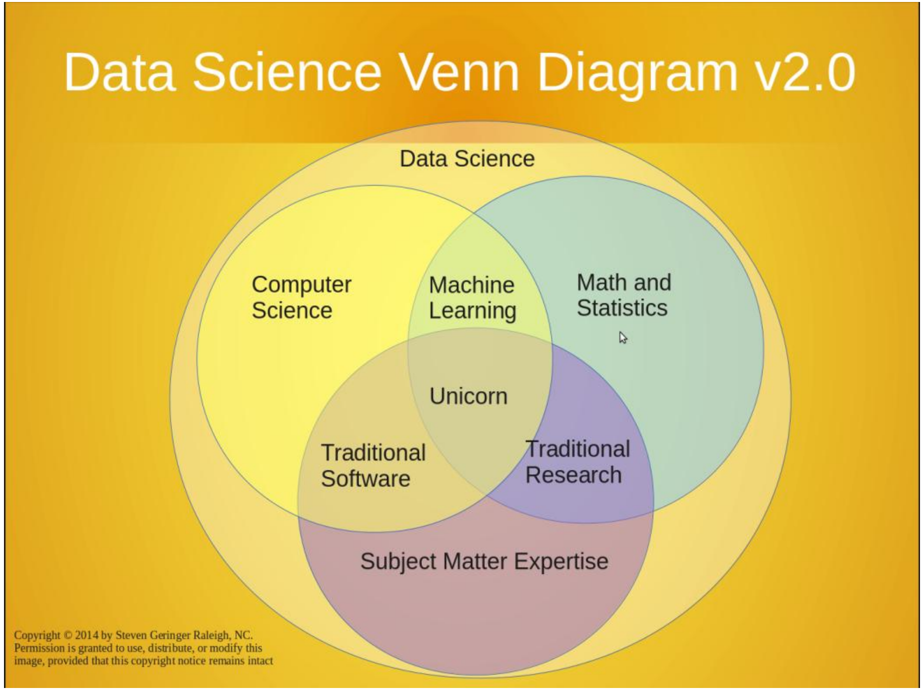
# The gap between education and practice

- Strong focus on theory
- Less focus on practice
- Fresh graduates are not equipped for real data analysis
- ‘Big Data’ problems are even further out of reach

# What I wish I was taught earlier

- Real data is messy, how do I deal with it?
- There is no single best method: how do I embrace plurality?
- *Ad hoc* procedures: when and how to use them safely?
- Data management
- Software engineering
- Working as part of a team

# What is ‘data science’?



# What is ‘data science’?

Bin Yu’s portrait of a data scientist:

- Statistics
- Domain/science knowledge
- Computing
- Collaboration/teamwork
- Communication to outsiders

# Bin Yu, on embracing data science

We need to...reform statistical curricula

We need to fortify our position in data science by focusing on training skills of:

- Critical thinking
- Computing
- Leadership, interpersonal and public communication

# Rafael Irizarry, on teaching applied statistics

Challenges:

- Applied statisticians don't teach what we actually **do**
- Applied statistics work is published outside of the ‘flagship’ statistics journals
- Resistance from students to open-ended assignments(...?)

# Mathematical vs applied statistics

- Undergraduate education is foundational
- Relevant for **all** statisticians
- Need to understand real data analysis in order to develop relevant theory

# Suggestions

1. Foundational skills subjects:

- Principles of data management
- Programming for statisticians
- Software engineering for statisticians (perhaps as a service course?)

2. Final year major project:

- Real, messy data
- Teamwork
- Deliverables to include an R package (or similar)

3. **Every** subject to have **one** main project using **real** data

4. Collaborative projects with computer science students

5. External ‘industry’ guest lecturers

6. Develop assessment schemes that focus on the solution process rather than on getting the ‘right’ answer

# Discussion questions

Are these proposals relevant to the Department of Mathematics & Statistics?

What changes can/should be made?

What are the main barriers to reform?

What is our role in these changes?

# More discussion questions

Is the Department of Mathematics & Statistics able to teach programming & software engineering skills?

How much flexibility/creativity is possible with assessment schemes?

Should we try to emulate how engineers are taught?