# Our future in big data science

Damjan Vukcevic
http://damjan.vukcevic.net/

13 October 2015
SSA Canberra, Young Statisticians' Workshop

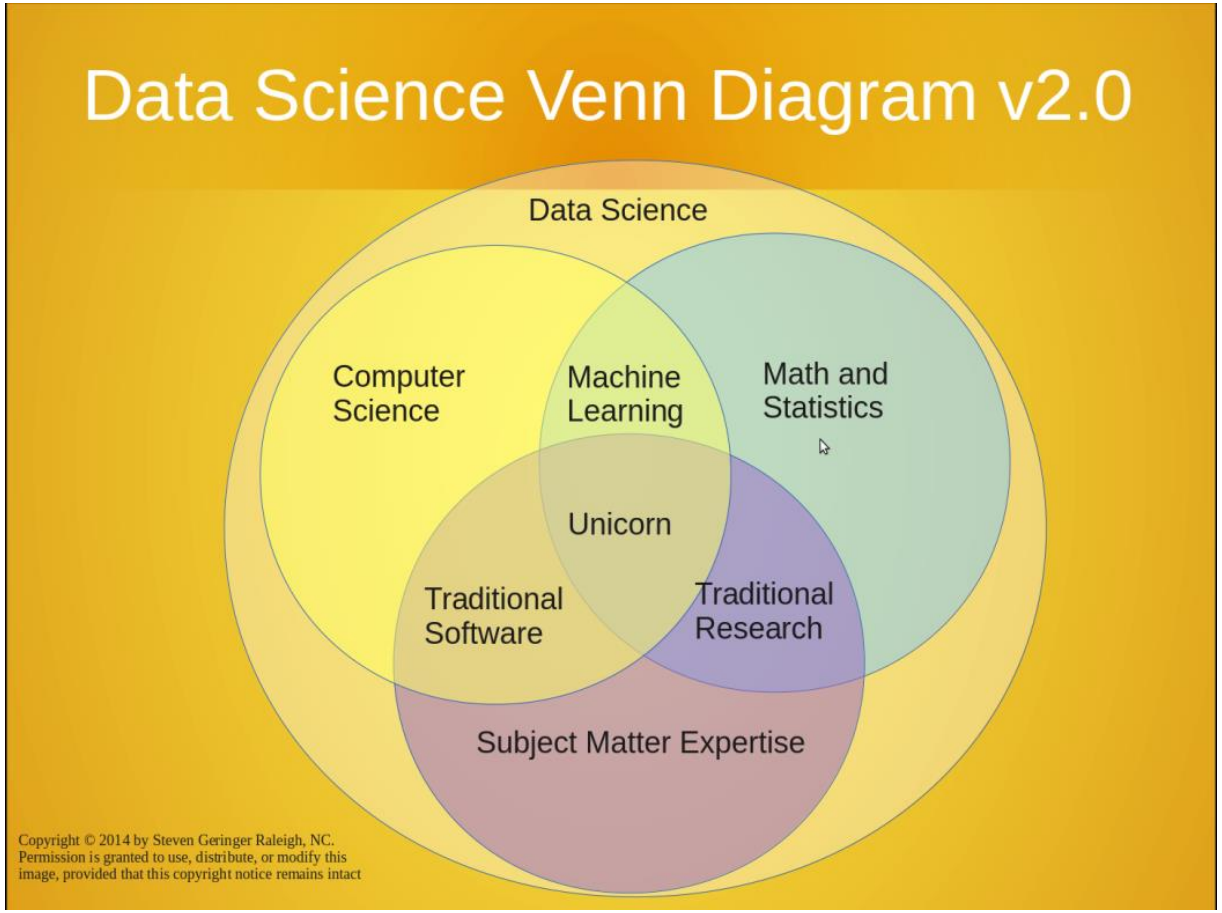
# What is ‘big data’?

You know it when you see it?

Tell-tale signs:

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

# What is ‘data science’?



# Why does this matter?

‘The absence of statisticians in Big Data activities is striking’

– Terry Speed (bioinformatician and winner of the Prime Minister's Prize for Science)

‘Let us own data science’

– Bin Yu (IMS Presidential Address 2014)

‘Statistics is foundational to data science’

– ASA policy statement (*The Role of Statistics in Data Science*)

# Overview

1. My previous projects
2. Factors for success
3. Our future

# 1. My previous projects

# About me

## Day job

Murdoch Childrens Research Institute

*Healthier Kids. Healthier Future.*

**Data Science**

**Statistical Genetics**

- Immune system genetics
- Genetic association studies
- Meta-analysis of genetic effects

**CEBU (Biostatistics)**

- Lung function in infants
- Measuring UV exposure from skin samples

## After hours

SSAI

Statistical Society of Australia Inc.

**Victorian Branch**

**Data Science Melbourne**

# My journey

THE UNIVERSITY OF MELBOURNE

2001

UNIVERSITY OF OXFORD

2005 2008

Experian Hitwise

2010

Murdoch Childrens Research Institute
Healthier Kids. Healthier Future.

2012

Mathematics
Statistics
Bioinformatics

Statistical genetics

Web analytics

Statistical genetics
Biostatistics

1

2

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 Randerson

ardian, Mon

THE AGE

News Sport Business Politics Comment Tech Entertainment Life & Style Travel Cars Pr

You are here: Home » Breaking News World » Article

## Vitamin D 'affects more than 200 genes'

August 24, 2010

John Von Radowitz

Tweet

Email article

PAA

Vitamin D influenc
conditions and can

The research highli

Scientists mapped 2
that make up genes

ABC Science

Explore by topic

News in Science | In Depth | Dr Karl | Ask an Expert | Bernie's Basics

Latest News in Science | News Analysis | StarStuff | News Archive | Tag library

News in Science

## Study unravels genetic jigsaw of hormone cancers

Share Print

Stephen Pincock
ABC

Thursday, 28 March 2013

New cancer treatments and better methods for cancer screening could emerge from a huge new international study th
of brea

Scientis
Gene-e
70 new
hormon

Their fi
numbe
ssocia
an 2.

ABC NEWS

News Home | Just In | Local | World | Business | Entertainment | Sport | The Drum | Weather | More

Email

## 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 of developing breast cancer have been uncovered in one of the world's largest cancer studies, helping to identify women who are at greater risk of developing the disease.

The finding, published today in Nature Genetics, nvolved more than 80 research institutions ollaborating with the Breast Cancer Association nsortium (BCAC) and cancer patients from 16 ntries, including Australia.

pears in the journal with another study, led by alian-born cancer expert Professor David , of the Harvard School of Public Health, that entifies new gene variants that predispose to cancer.

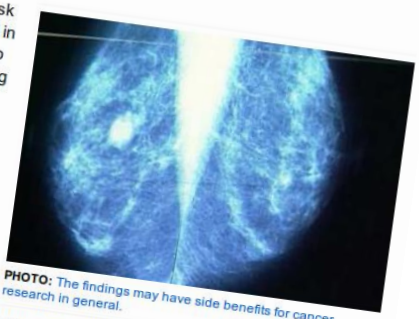


PHOTO: The findings may have side benefits for cancer research in general.

MAP: Sydney 2000

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."

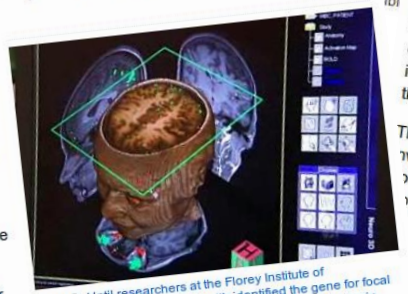
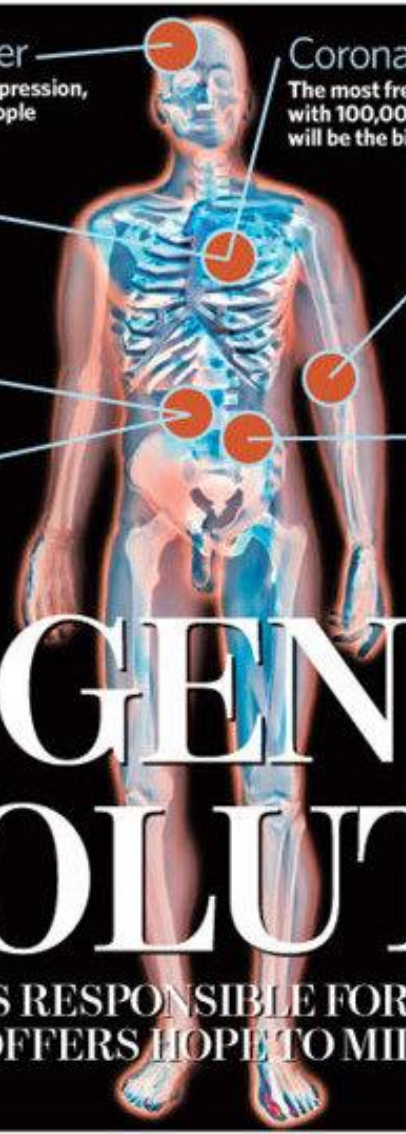


PHOTO: Until researchers at the Florey Institute of Neuroscience and Mental Health identified the gene for focal epilepsy, it was believed the disorder was caused by brain injury or tumours. (ABC News)

MAP: Melbourne 3000

(Ireland, €1) 70p
Thursday 7 June 2007
www.independent.co.uk

# 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

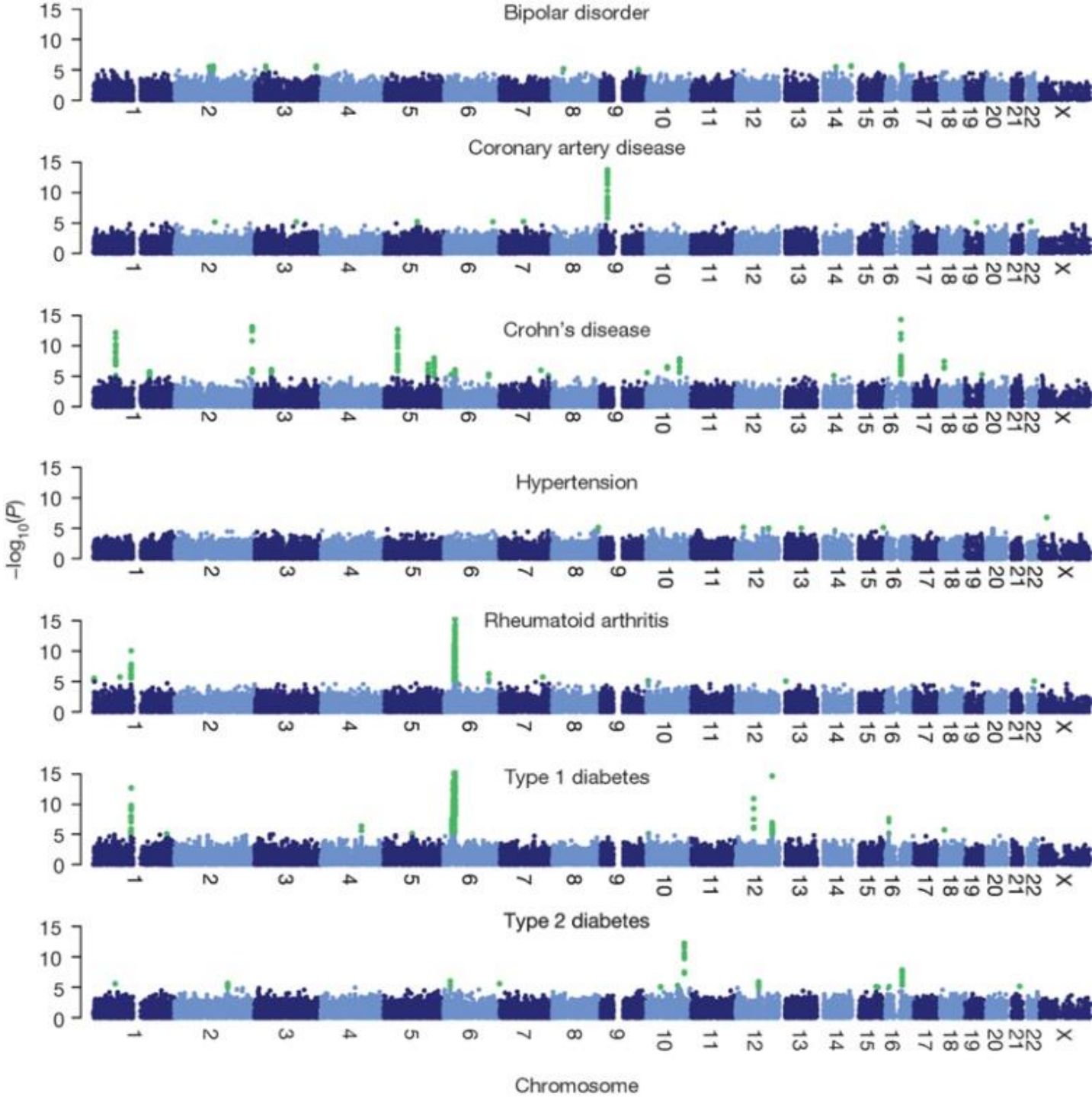
FULL STORY, PAGE 2

# Study design (WTCCC, 2007)

500,000 genetic markers

| | |
|---|---|
| 3,000 controls | **1958 Birth Cohort**<br>**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** |

# Results


Bipolar disorder
Coronary artery disease
Crohn's disease
Hypertension
Rheumatoid arthritis
Type 1 diabetes
Type 2 diabetes
$-\log_{10}(P)$
Chromosome

# Findings

- **Doubled** the number of known genetic associations (12 → 24)
- Found common genetic effects **common to more than one** disease
- Evidence of different genetic architectures:
  **autoimmune disease** vs **other diseases**

# 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 languages environments (C++, R, bash,…)
- Developed a suite of software in tandem with analysis

# Inferring genotypes

'Genotype calling'

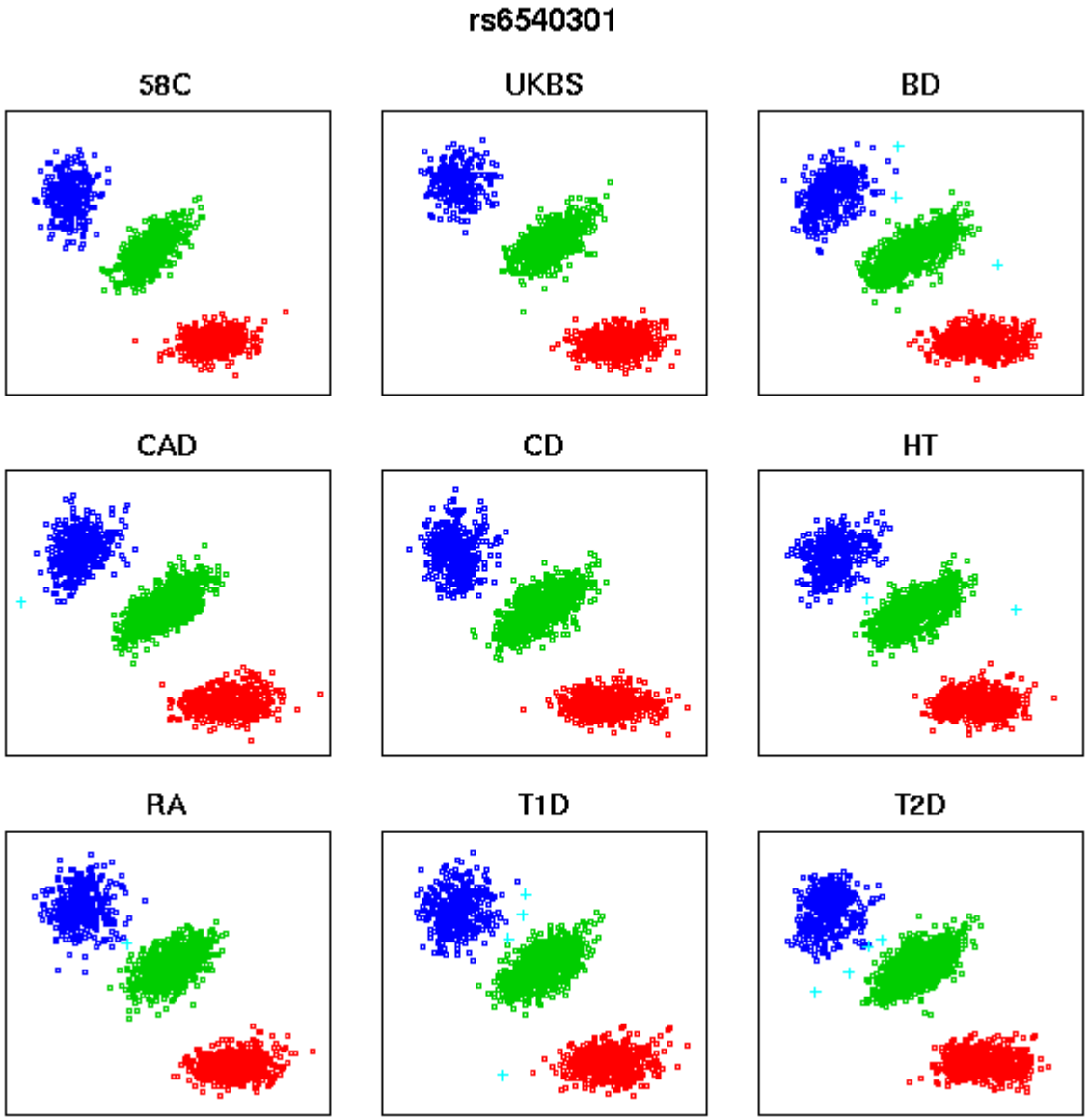
Designed new method (CHIAMO)

Hierarchical Bayesian clustering with informative priors

Used data from **all** individuals

Allowed for variation between cohorts

Showed Affymetrix data is actually reasonably good

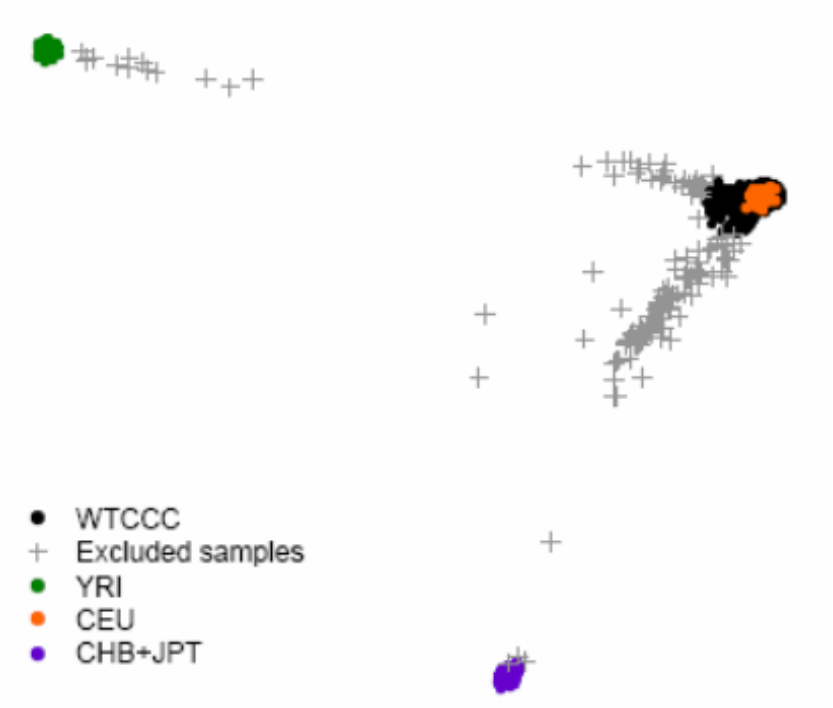


# Population structure

Principal components analysis (PCA)

Uses data across the whole genome

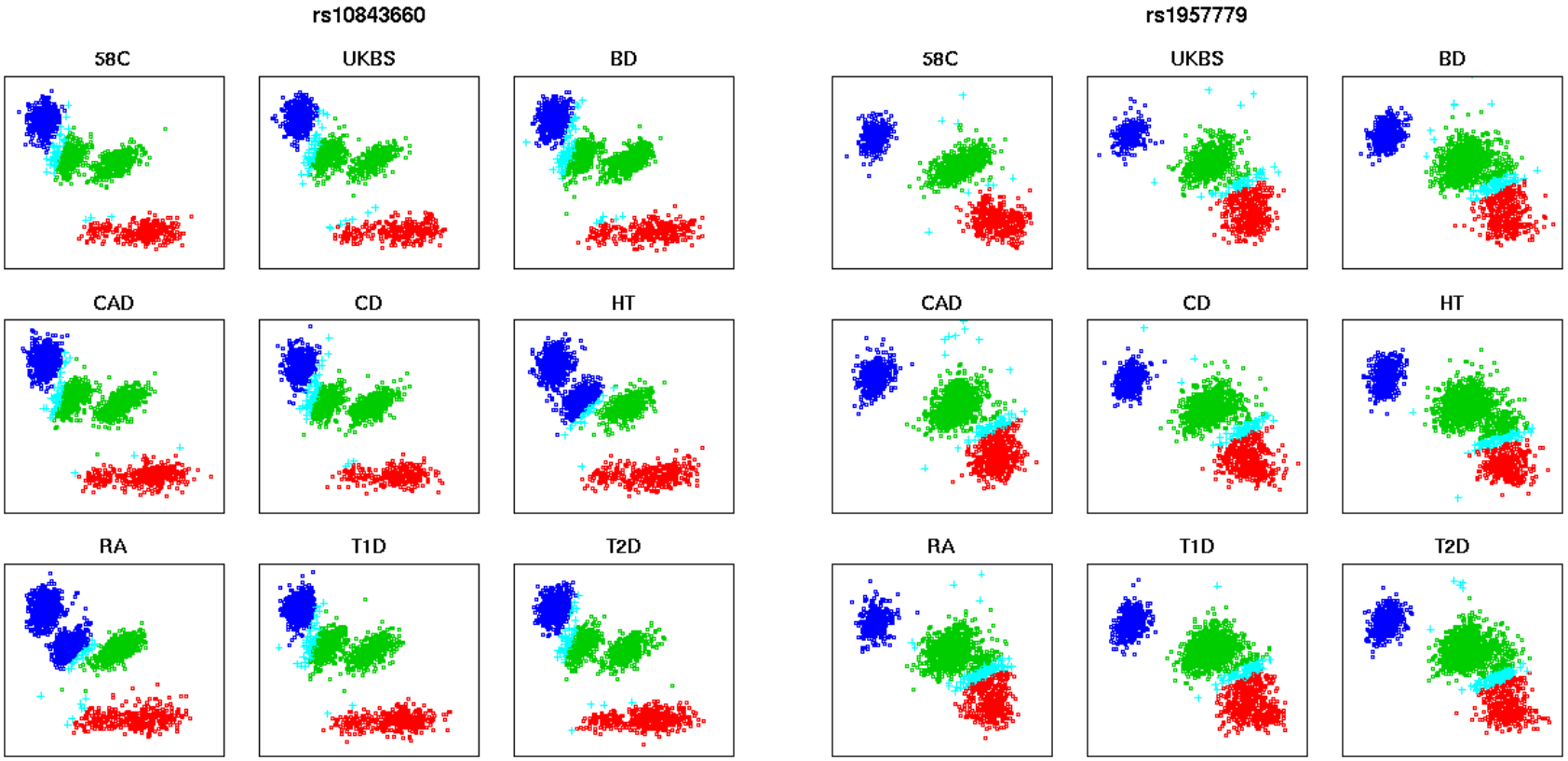
Reference panel with known ancestry



# Quality control (QC) & filtering

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

# ‘Cluster plot’ inspection



# Quality control ‘epic fail’

- The letter to Nature…

# 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

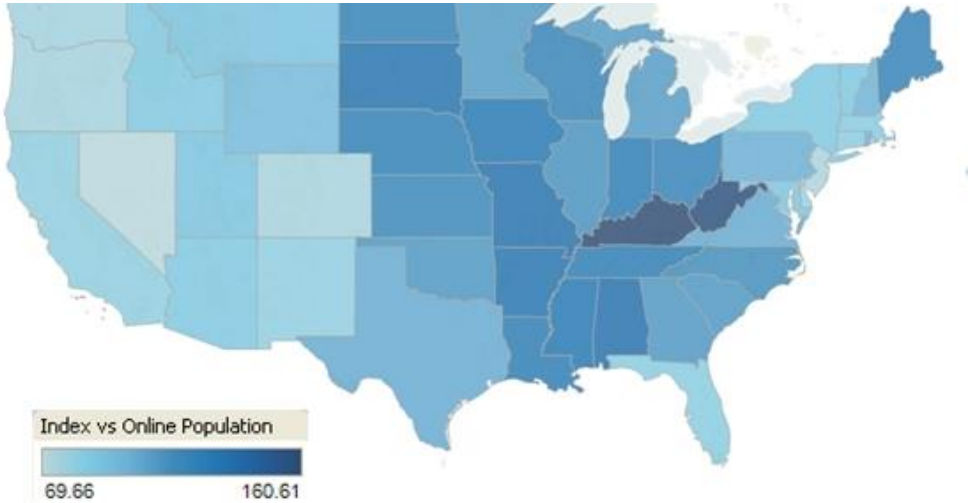
We used 16 different analysis pipelines where different aspects of the data pre-processing were varied. Sup-

# Web analytics at Experian Hitwise

- **Hitwise** – a Melbourne internet start-up from 1997
- Acquired by **Experian** in 2007
- I worked there 2010–12

- Large team of programmers, web developers, database administrators, data managers, data analysts, project managers, sales teams,…
- No statistical expertise

# Web analytics




**2011 Black Friday Shoppers by State**

Week ending November 26, 2011, compared with the Online Population

| State | Visits Share Black Friday 2011 Shoppers | Visits Share Online Population | Index |
|---|---|---|---|
| 1 California | 10.05% | 12.45% | 81 |
| 2 Texas | 7.93% | 7.72% | 103 |
| 3 New York | 5.23% | 6.08% | 86 |
| 4 Illinois | 4.88% | 4.34% | 112 |
| 5 Florida | 4.88% | 5.92% | 82 |
| 6 Ohio | 4.49% | 3.66% | 123 |
| 7 Pennsylvania | 4.03% | 4.01% | 100 |
| 8 Georgia | 3.57% | 3.23% | 111 |
| 9 Michigan | 3.51% | 3.23% | 109 |

# Questions I tackled

Population projections (‘total visits’)

Multilevel models to reduce variance for rare event estimates

Estimates of proportion of unobserved search terms

Detection of marketing campaigns from web traffic

# Major challenges

Lack of statistical strategic planning

No statistical team
(no mentors, no peers, lack of manpower)

Poor sample design (opportunistic sampling)

# 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

# 3. Our future

# Our future in (big) data science

**Engage** with data analysts from other disciplines

**Embrace** projects beyond our traditional domains

**Educate** the next generation, reform statistical curricula

# **Your** future in (big) data science

## Learn **software engineering** skills

- Learn to program (R)
- Version control (Git)
- Modularisation (R packages)
- Learn another language (Python)

## Get experience with **real data**

- Hackathons
  (HealthHack, GovHack)
- Kaggle
- Open Knowledge Australia

## Seek out good **mentors**

- Supervisor / line manager
- Group head / senior manager
- Informal/formal mentors
- Knowledgeable peers

## Cultivate a wide **network**

- Attend Meetup events
  (Canberra R Users Group,
  Canberra Data Science)
- Enter competitions
- Organise events
  (e.g. through SSA Canberra)

# Your future in (big) data science

- **Software engineering**
- **Real data**
- **Mentors**
- **Network**

# Contact me

**Web** http://damjan.vukcevic.net/

**Email** damjan@vukcevic.net

**Twitter** @VukcevicD