# Impact of Machine Learning in Personalised Health: past, present and future

Marta Milo
University of Sheffield
Department of Biomedical Science

The University Of Sheffield.

# Outline

Introduction

- Small history of high throughput data
- how did the data grow and what are we facing?

Complexity of high throughput data analysis: pro and cons

Application of high throughput data analysis

- Clinical study (past –present)
- Stem cell application (present –future)

Future directions

# Once there was the darkness of the microscope room….

The University Of Sheffield.

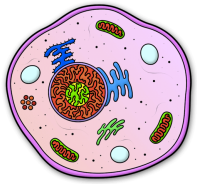
**Past – HGP**
Each gene to be probed one at a time. Biological systems were investigated by examining their parts in isolation. System-level analysis only theoretically possible

**Human Genome Project**

NEW HiSeq 2500

**HGP – present**
Ability to probe and quantify activity of every gene, how it responds to a particular perturbation. high–throughput profiling studies enabling systems–level analyses, with integrative approaches to elucidate the functional associations between differentially expressed genes.

+



+

**Present –Future**
"omics" analysis at single-cell resolution enabling understanding of complex biological phenomena. Characterise cellular cellular composition of complex tissues, find new microbial species and perform genome-wide haplotyping.

The University Of Sheffield.

# Our Motivation

**Personalised Medicine**: a *medical* model that proposes the customisation of healthcare - with medical decisions, practices, and/or products being tailored to the individual patient.

Wikipedia

**My aims …**

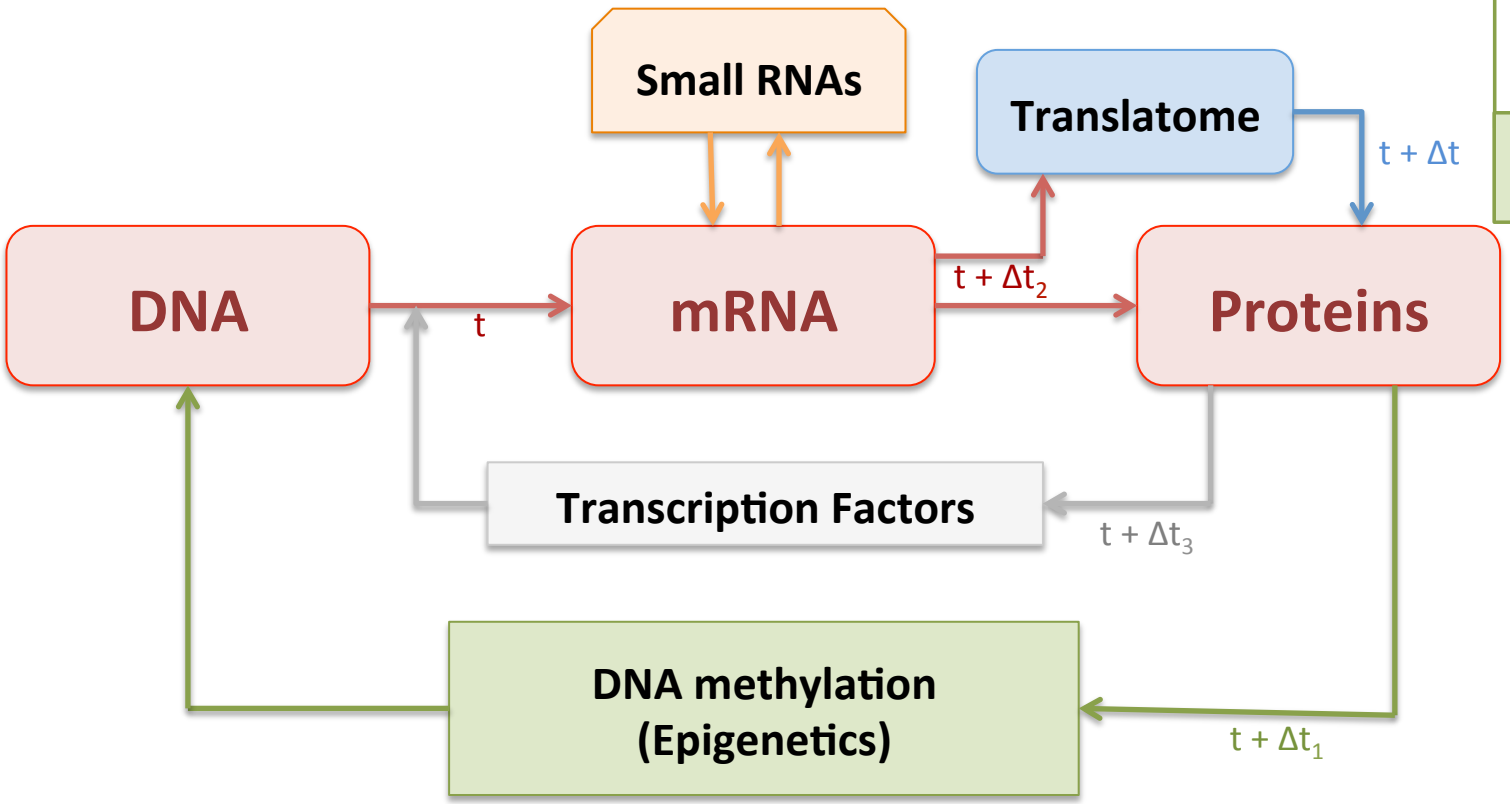
- Combine gene–level analyses with pathway–based methods to generate a comprehensive profile of the functional modules that govern biological processes.

- We want to use high–throughput data to build predictive models at the systems level and define therapeutic intervention and /or genomic predispositions to disease at individual level.

# Biological Data: a network of complex interaction

The University Of Sheffield.

*Regulatory mechanisms in molecular biology*



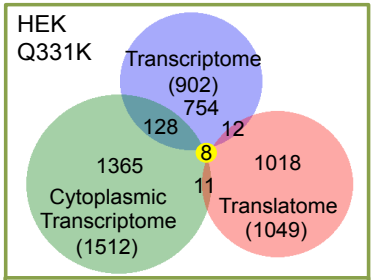




**Fig. 2:** Venn diagrams highlighting differential expression changes between whole-cell or cytoplasmic transcriptomes and translatomes.

G. Hautbergue *et al. (now)*

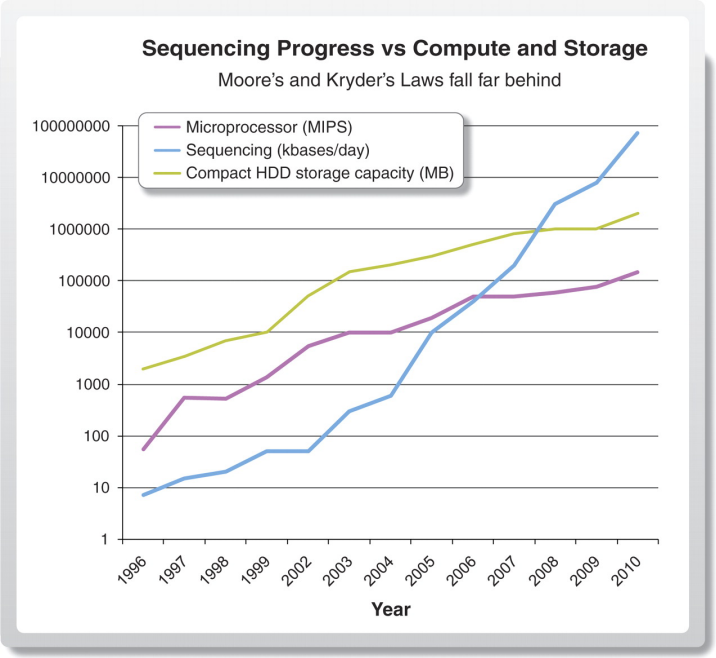
Snapshot at a time *t*?

Adapted from Braun R, *Adv Exp Med Biol*. 2014; 844: 153–187

# Genomic Data: a new challenging era

In February 2001 scientists published the first drafts of the Human Genome, the dawn of the genomic era. The Human Genome Project was completed in 2003.

Challenges in “omics” do not derive only from the *informatics* that is required to **analyse, summarised** the vast amount of raw (sequencing and not) data that is available, but primarly from **intepreting** the findings in complex systems.



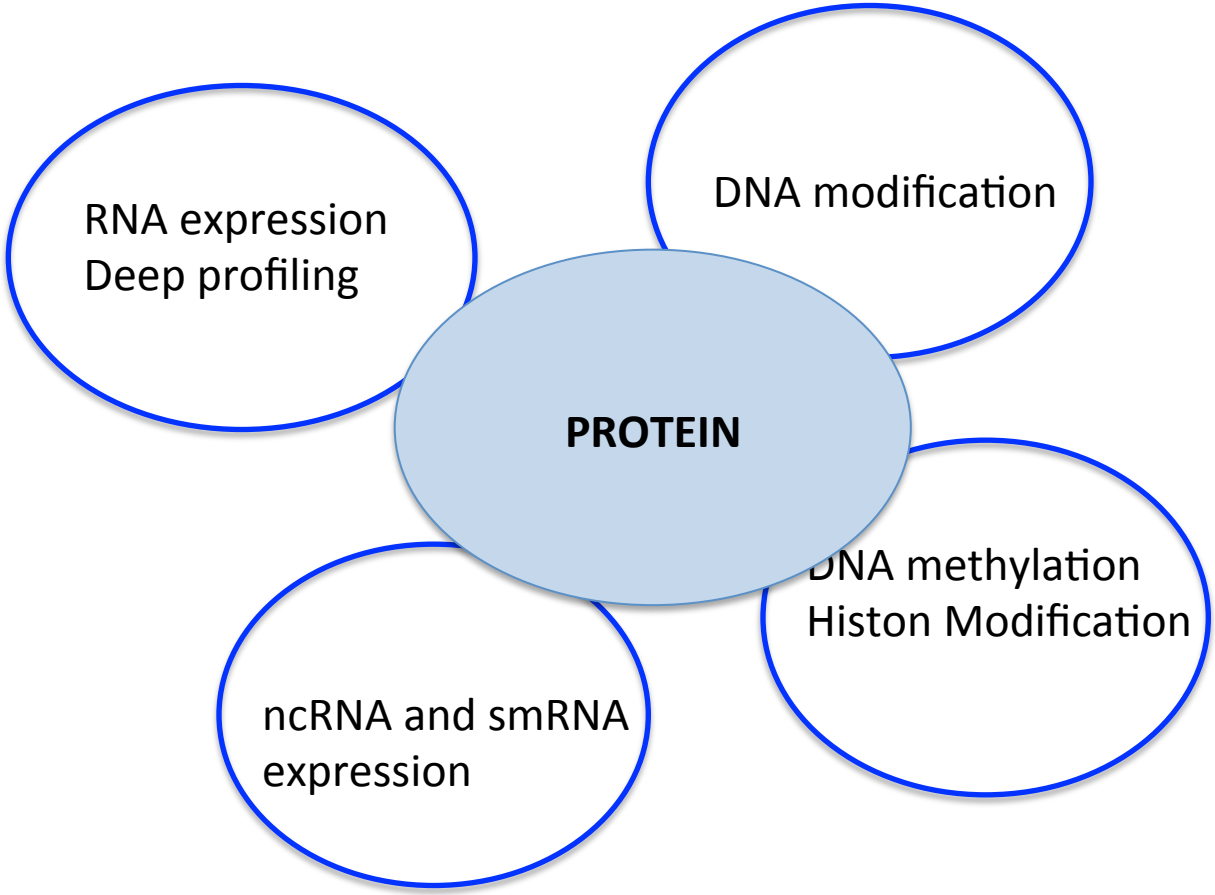

**S D Kahn,** Science 2011;331:728-729

**Moore's law**: The number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years

**Value and definition of raw data?**
**How can we analyse this data effectively?**
**How do we interpret this data?**

The University Of Sheffield.

# Quality of the data is important




The accuracy of systems–level analysis will depend on the quality of the data being analysed. Clear experimental design, appropriate assay technology employed, and the preprocessing of the raw data.

The University Of Sheffield.

# Tools we need

- Optimal Experimental Design – minimise the noise in the measurement
- Mathematical Models – define rules (functions) to describe processes
- Statistical tools – quantify accuracy in prediction and sensitivity in estimation
- Computational Skills – handling large amount of data in automated way
- Visualisation tools – identify patterns in the data

# The study

The University Of Sheffield.

It was funded by the Cardiovascular Biomedical Research Unit (CV BRU) led by Prof D. Crossman under a scheme promoted by the UK Government health research strategy: *Best Research for Best Health.*
CV BRU was an Infrastructure Funding and the aim to:

- Drive innovation in the prevention, diagnosis and treatment of ill-health ( acute coronary syndrome);
- Translate advances in medical research into benefits for patients;

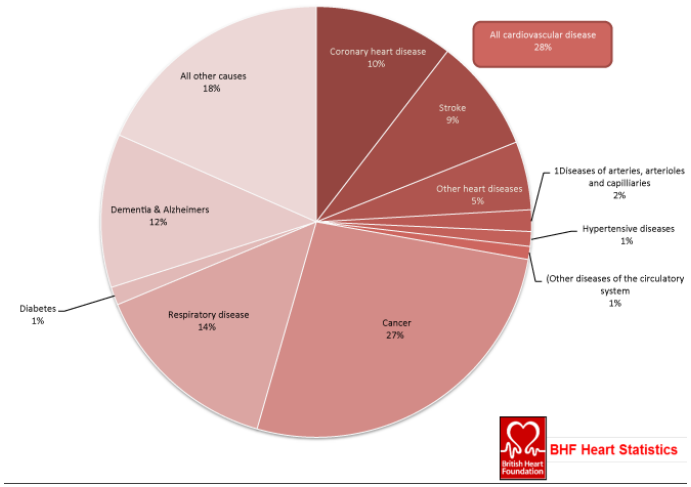
**Genomics of Acute Coronary Syndrome**:

- Explore the transcriptional activity of the biological processes in ACS to predict risk of recurrent events (Gene expression data – transcriptional effects)
- Recategorisation of the acute events (MI) to improve prognostics
- Understanding the role of inflammation in ACS by functional manipulation of affected pathways (**role** of inflammation in ACS particular focus of IL-1)

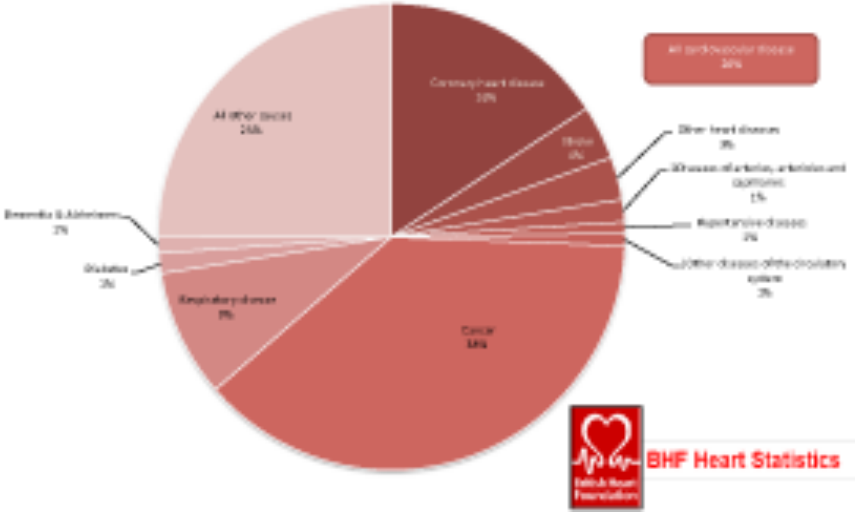
# Why we are interested in Cardiovascular disease (CVD)

The University Of Sheffield.

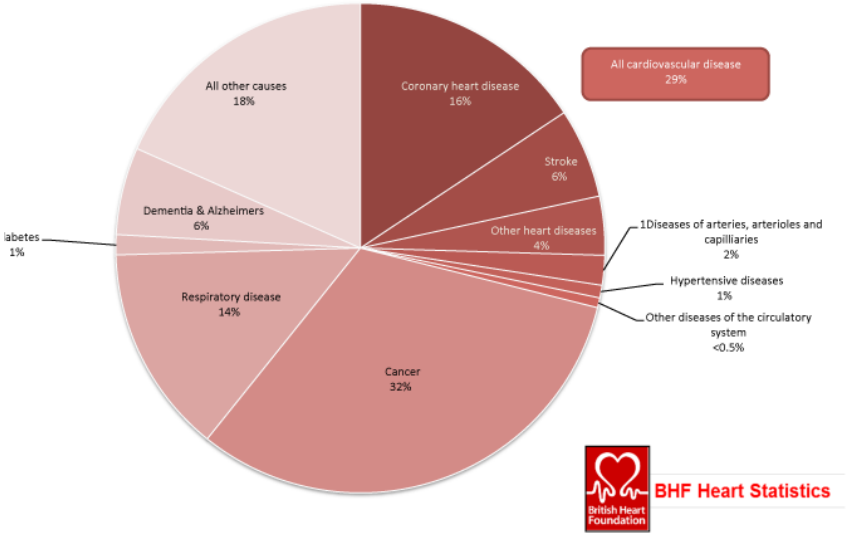
***Deaths by cause in women, United Kingdom 2012***




***Deaths under 75 by cause in men, United Kingdom 2012***

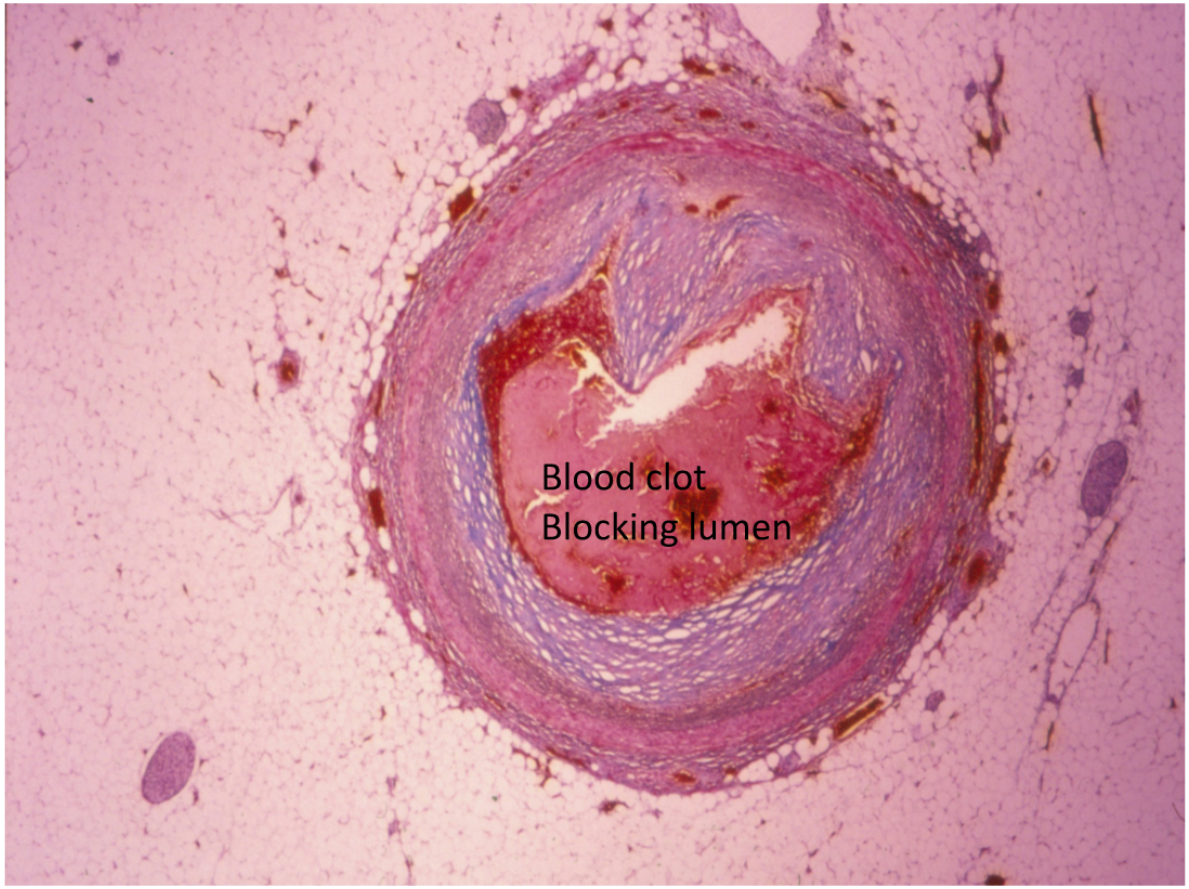



***Deaths by cause in men, United Kingdom 2012***



# Why we are interested in Cardiovascular disease (CVD)

The University Of Sheffield.




Multifactorial disease: ideal for a system-level approach

Cardiovascular disease (CVD) is the United Kingdom's biggest killer.

Each year 198,000 people die from CVD(BHF, 2009) and although the rate of death in the UK has fallen markedly over the past two decades it remains among the highest in Europe.

It costs the EU more than €192bn (£145bn) annually, equivalent to nearly €400 a head.

# Patients cohort and experimental design

The University Of Sheffield.

Pilot Study: Recruited ~170 patients

Patients recruited presenting to A&E with acute cardiac chest pain with first occurrence in the last 48 hrs.
Both STEMI and NSTEMI we recruited and control cohort is represented by patients presenting with Unstable Angina (tropinin negative).

ACS patients are deeply screened over time after their first acute event
Time course: **Day1** **Day2** **Day3** **Day7** **Day 30** **Day 90** Day 365

Tests performed: Bloods, Urine, Platelet Tests, GTT, ETT, Echo
Sample stored: Serum, Plasma, DNA, RNA(PAXgene), RNA(Tempus)

mRNA quantification with Affymetrix arrays from Whole blood:
microRNA quantification with TaqMan (TILDA cards) from whole blood
**Full time course at Day 90: 33 Patients**

The University Of Sheffield.

# Example: genomic analysis of diseases

Dealing with a very heterogeneous genetic background:

- splice variants and *de novo* splice junctions
- functional SNPs
- different predisposition to disease, due to a different factors

Difference font of variations, not always linear and not always possible to model globally.

Missing data points is a HUGE problem.

Adding complexity to the model: increase of computational time and compromising tractability of the models

Often not big enough sample size, in common diseases need large collections. No replicated experiments.

Collection of samples over long period: high technical variation.
Difficult to evaluate variation, complex data handling and storage

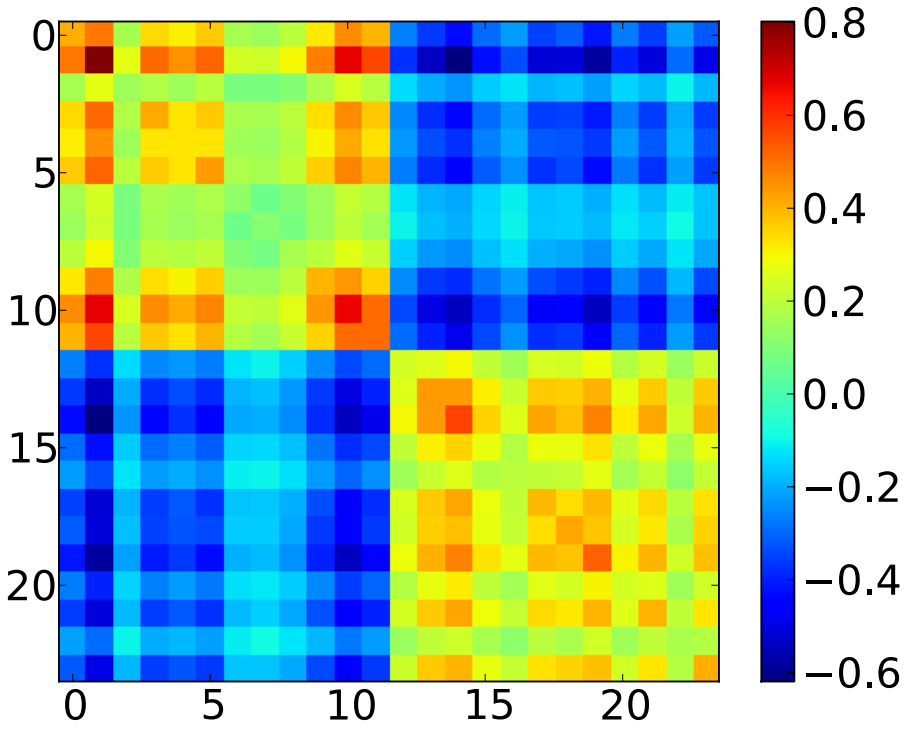
# Challenges in the data: diagnostics of confounding factors

Whole blood for Transcriptomics

Down:
- Many cell types: erythrocytes, leukocytes, thrombocytes, Circulating microvesicles, apoptotic bodies, circulating cells
- Hemoglobin and its mRNA
- Sensitive to donor conditions
- Specificity

Up:
- Easy accessible tissue
- Dynamic interaction with the whole body
- RNA, microRNA and DNA from the same sample
- Stored in biorepositories
- Complex but richer information
- Possible diagnostic tool

Environmental effects and confounding factors influencing gene expression




Probabilistic ANAlysis of genoMic datA (PANAMA) – *Fusi et al.*

The University Of Sheffield.

# Choosing the correct methods

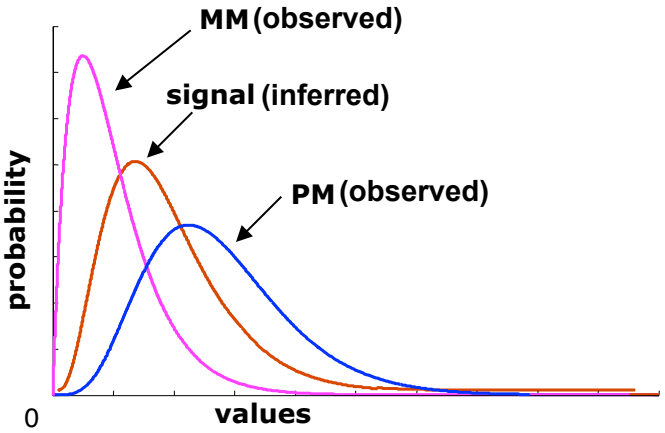
## Transcriptome analysis

Robust gene expression estimates depend on how well we are able to quantify the uncertainty.

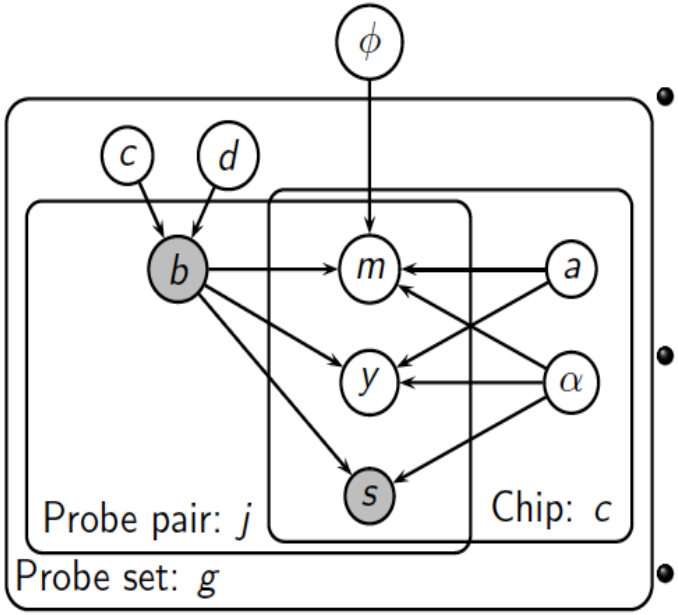
Down stream analysis will be more effective and number of false positives will be reduced.

**Good measure of uncertainty helps to denoise the data and consequently gives confidence in the results**

# *puma:* Propagating Uncertainty in Microarray Data




**Milo M *et al*, Biochem transcation 2003**
**Liu X *et al*, Bioinformatics 2005**
***Pearson R et al*, BMC Bioinformatics, 2009**



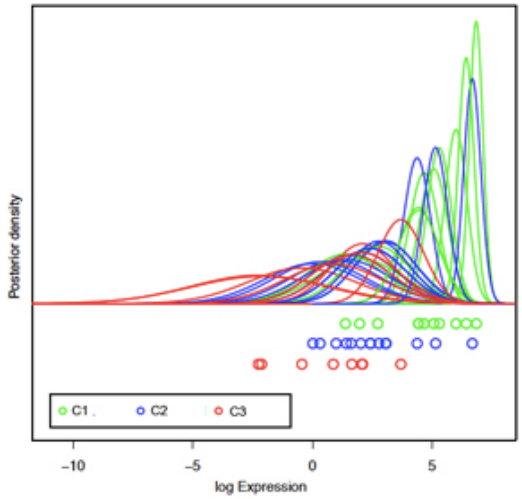

- latent true signal :
  $s_{gjc} \sim \mathrm{Ga}(\alpha_{gc}, b_{gj})$
  - specific hybridization $\alpha$.
  - latent rate ("precision") :
    $b_{gj} \sim \mathrm{Ga}(c_g, d_g)$.
- PM signal :
  $y_{gjc} \sim \mathrm{Ga}(a_{gc} + \alpha_{gc}, b_{gj})$
  - $a$ models background hybridization.
- MM signal :
  $m_{gjc} \sim \mathrm{Ga}(a_{gc} + \phi\alpha_{gc}, b_{gj})$

**specific MM binding and multiple information across chips**

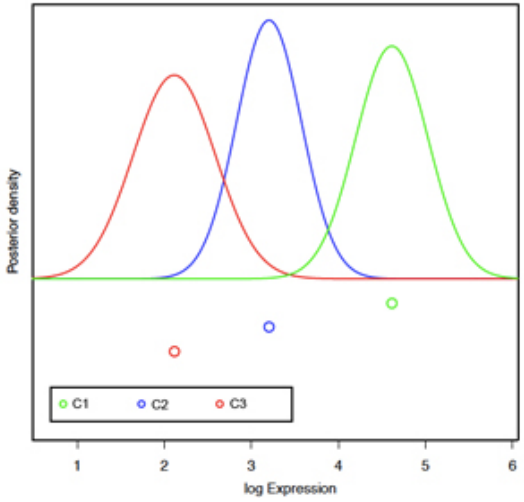
# Combining replicates

The University Of Sheffield.

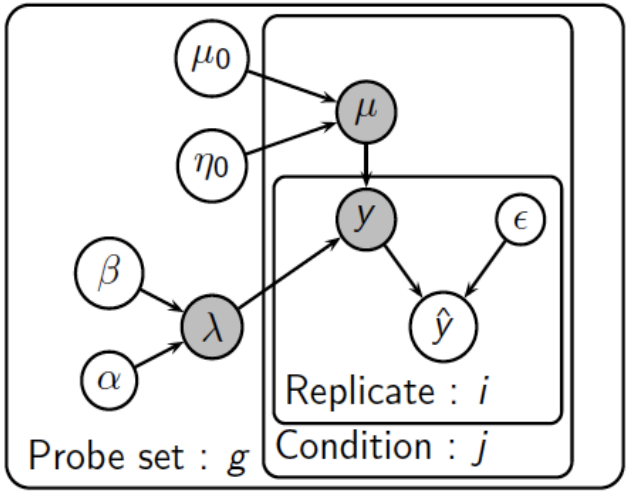
## ACS profiles




## Combined profiles




## The model




- true signal : $y_{gij} \sim \mathrm{N}(\mu_{gj}, \lambda_{gj}^{-1})$
- observed signal : $\hat{y}_{gij} \sim \mathrm{N}(\mu_{gj}, \lambda_{gj}^{-1} + \nu_{gij}^{-1})$
- noise : $\epsilon_{gij} \sim \mathrm{N}(0, \nu_{gij}^{-1})$
- $\mu_{jg} \sim \mathrm{N}(\nu_{0,g}, \eta_{0,g}^{-1})$
- $\lambda_g^{-1} \sim \mathrm{Ga}(\alpha_g, \beta_g)$

Observed variance:

$$\sigma_g = \sqrt{\frac{1}{n-1}\sum_c (E[\log(s_{gjc})] - \langle E[\log(s_{gjc})]\rangle)^2}$$

Predicted variance:

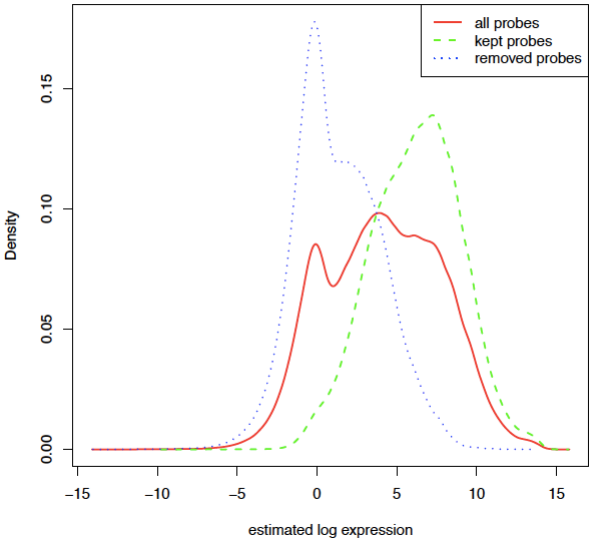
$$\langle Var[\log(s_{gjc})]\rangle = \frac{1}{n}\sum_c Var[\log(s_{gjc})]$$

Filtering:

$$\frac{\sigma_g}{Var[\log(s_{gjc})]} < 1$$

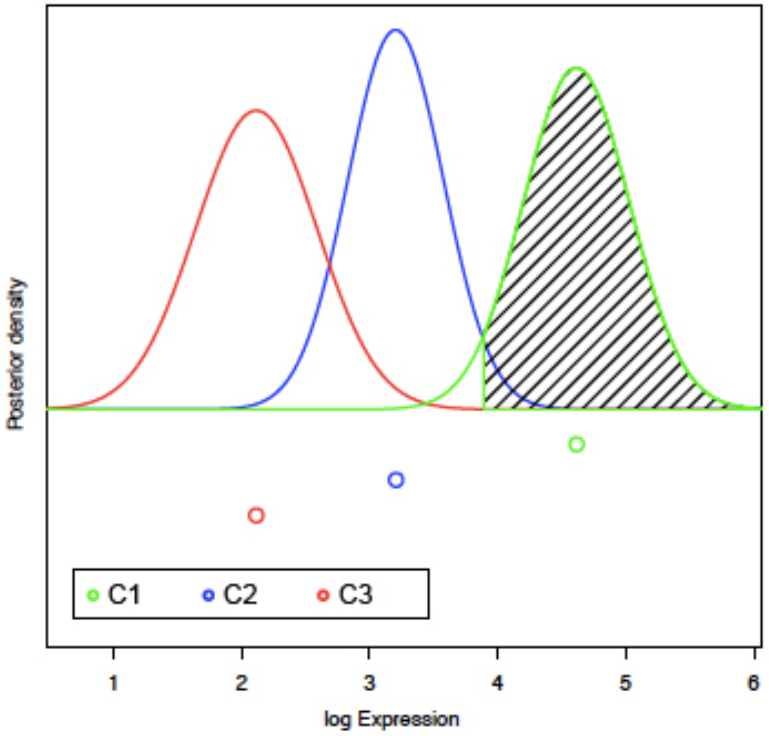
J. Nielsen *et al*, in preparation

Densityplot of log expression



The University Of Sheffield.

# Differential Expression: pumaDE and PPLR




**Probability of Positive Log ratio: PPLR**

$$P(\mu_1 > \mu_2 | D, \phi) = \int_0^\infty P(\mu_1 - \mu_2 | D, \phi) d(\mu_1 - \mu_2)$$

**PPLR and False Discovery Rate**

$$FDR(1..n) = \frac{1}{n} \sum_{i=1}^{n} 1 - PPLR(i)$$

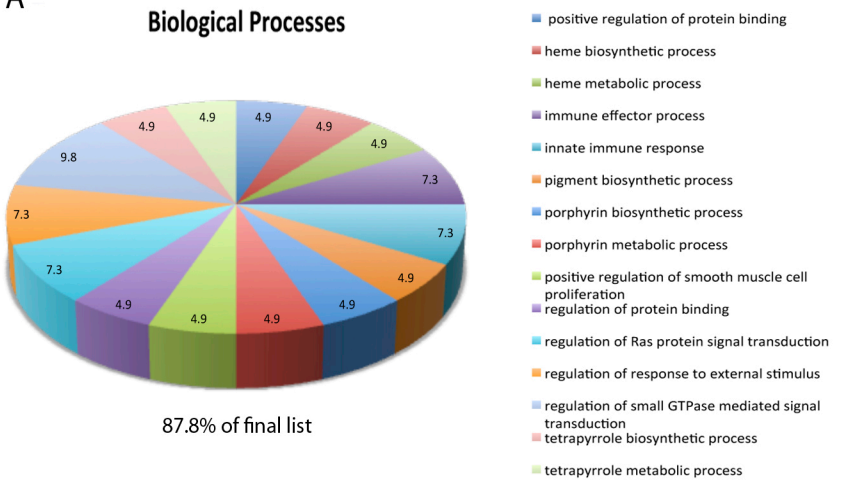
J. Nielsen *et al*, in preparation

# The results

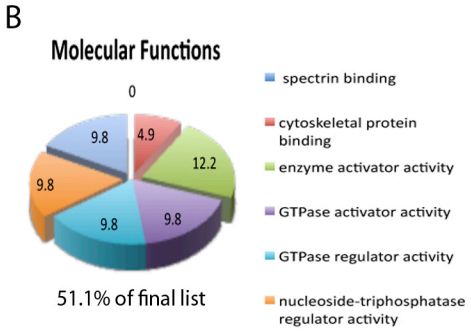
1. Robust quantification of expression levels - *puma*
2. Correction for environmental effects and confounding factors
3. Effective filtering of the data
4. Down stream analysis of robust targets
5. Target Validation and Specificity

We selected 93 genes across the 5 time points that were significantly differentially expressed.
The top 44 (plus 3 CTRLs) were technically validated using TLDA cards ( 70.5% were validated)

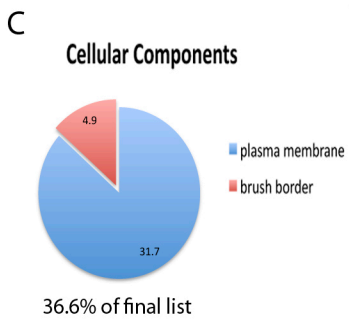
A



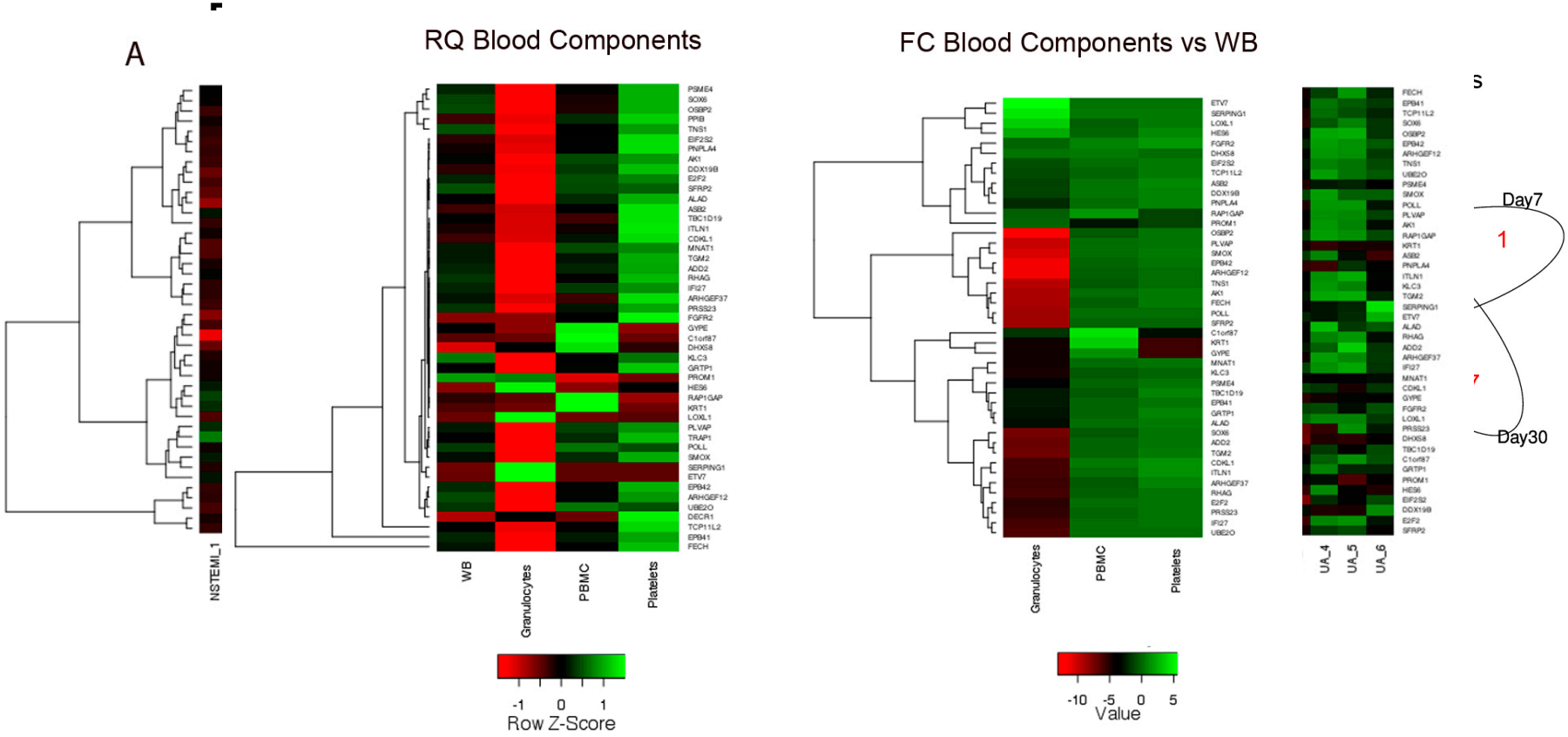

B




C



# The Results (cont.)




Healthy volunteers a pool of 5

ACS: is it all about plateles?
We identify pathways connected to inflammatory responses that can be regulated or dysregulated even after full recovery. Are platelets involved in modulating inflammation?


# Looking at the future….

# Stem Cell therapy using hPluripotent Stem Cells

The University Of Sheffield.

Stem cell therapy is the use in living individuals of specifically derived stem cells to treat or prevent a disease .

- hPSCs in culture acquire mutation in their genome. Using Inhibitors in culture medium and specific culture protocols, can improve the proliferation of cell carrying CNVs.

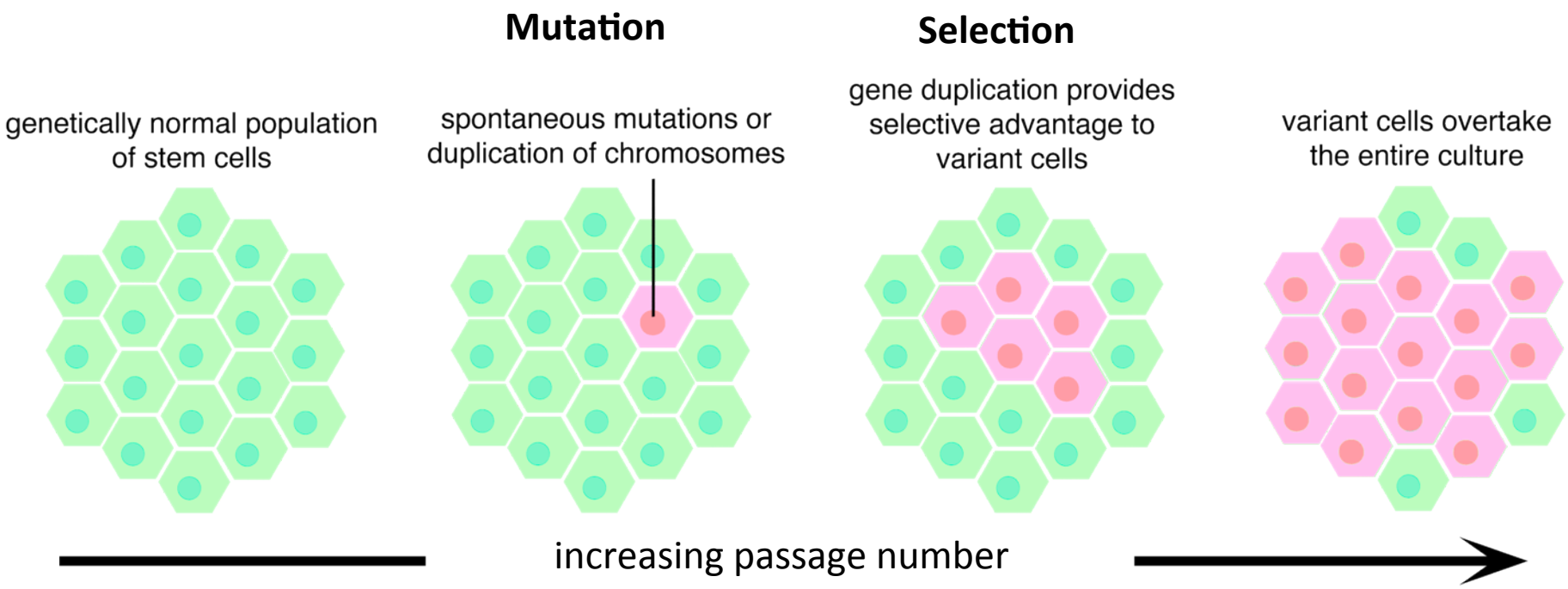
- Useful model to investigate methods of suppressing the selective advantage of variants. Mechanisms of selective advantage are not yet known.

- In clinical grade line it is required ABSENCE of CNVs for cell therapy. This is very difficult to obtain. Limitations in advances in the field.

    1. **How do we predict occurrence of variants so to minimise them?**

    2. **How can we identify functional significant genetic and epigenetic variants during hPSCs production for efficient translational use at individual bases?**

# Selective advantage of variants drives *culture adaptation*




**Biological / manufacturing consequences:**

- abnormal growth characteristics
- compromised differentiation potential
- no longer representative for screening

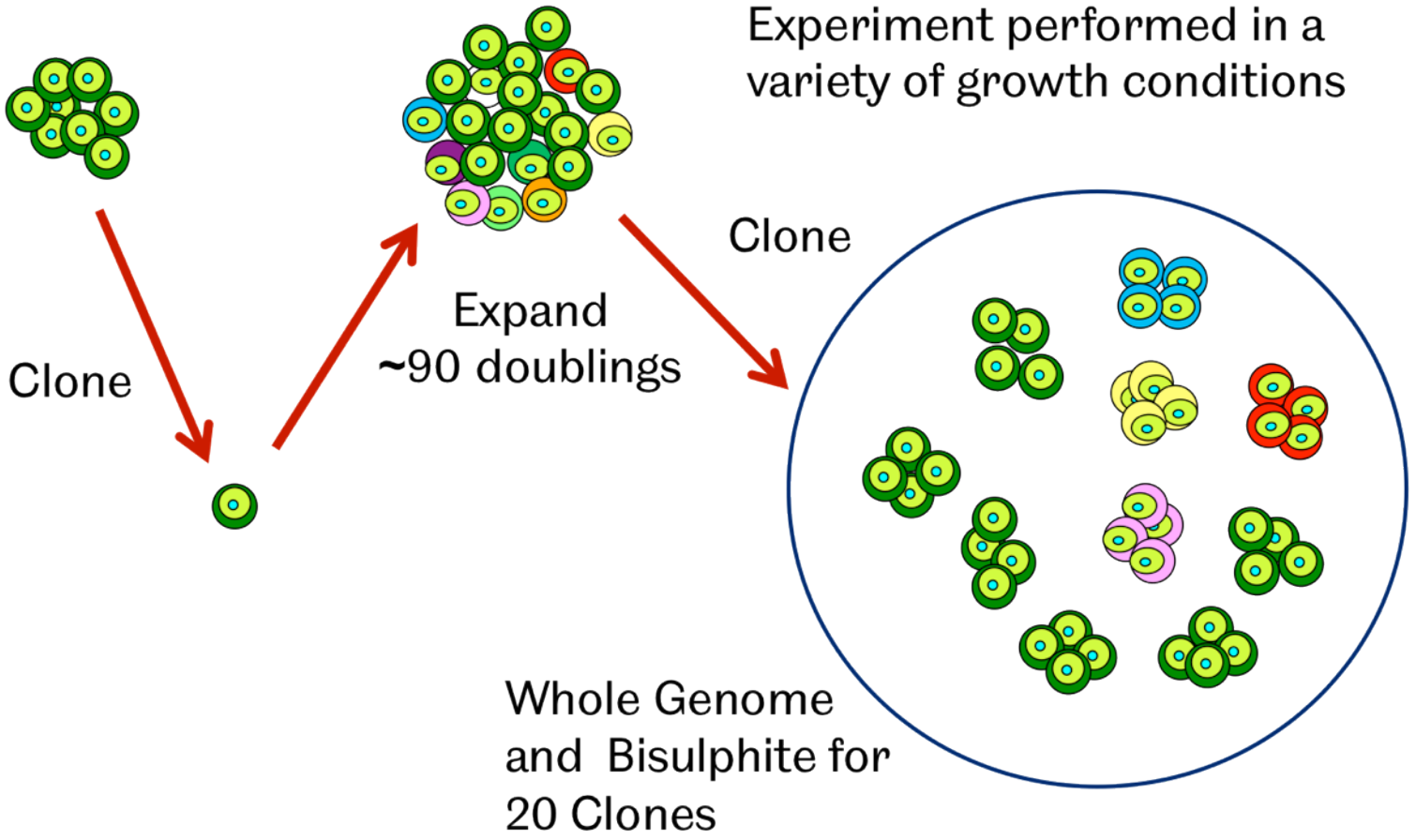
**Clinical consequences:**

- poor therapeutic products
- genetic disease
- danger of cancer following transplantation

## 1. How do we predict occurrence of variants so to minimise them?

The University Of Sheffield.

Estimating mutation rate in different culture condition using whole genome sequencing



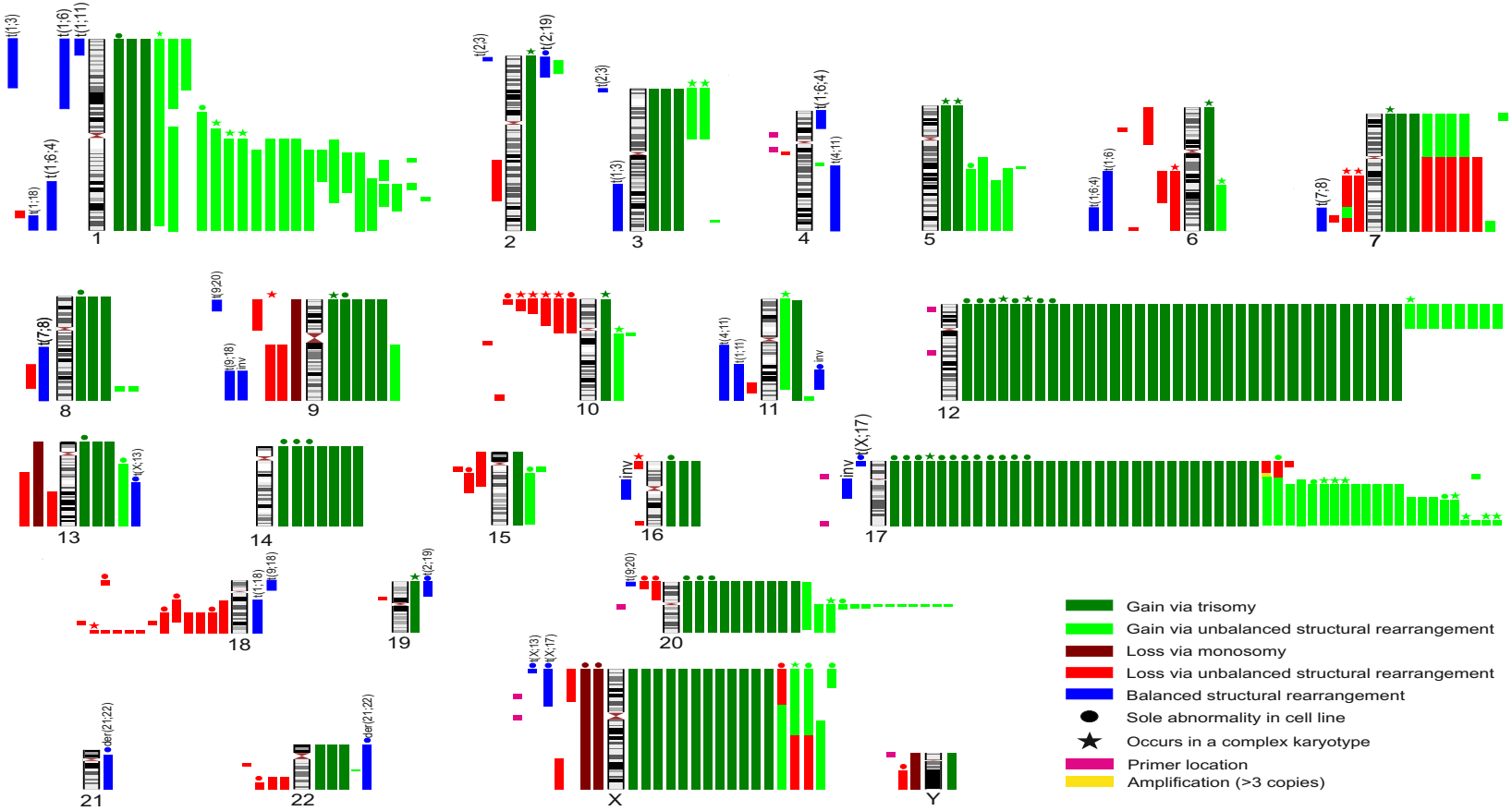

MShef4 (relatively unstable)

MShef11 (stable)

In collaboration with the Sanger

1. **How can we identify functional significant genetic and epigenetic variants during hPSCs production for efficient translational use at individual bases?**

Define areas of “hot spot” in the genome by studying stem cell lines from different population – ISCI project. Use the information as *a prior* to rank the impact of CNVs



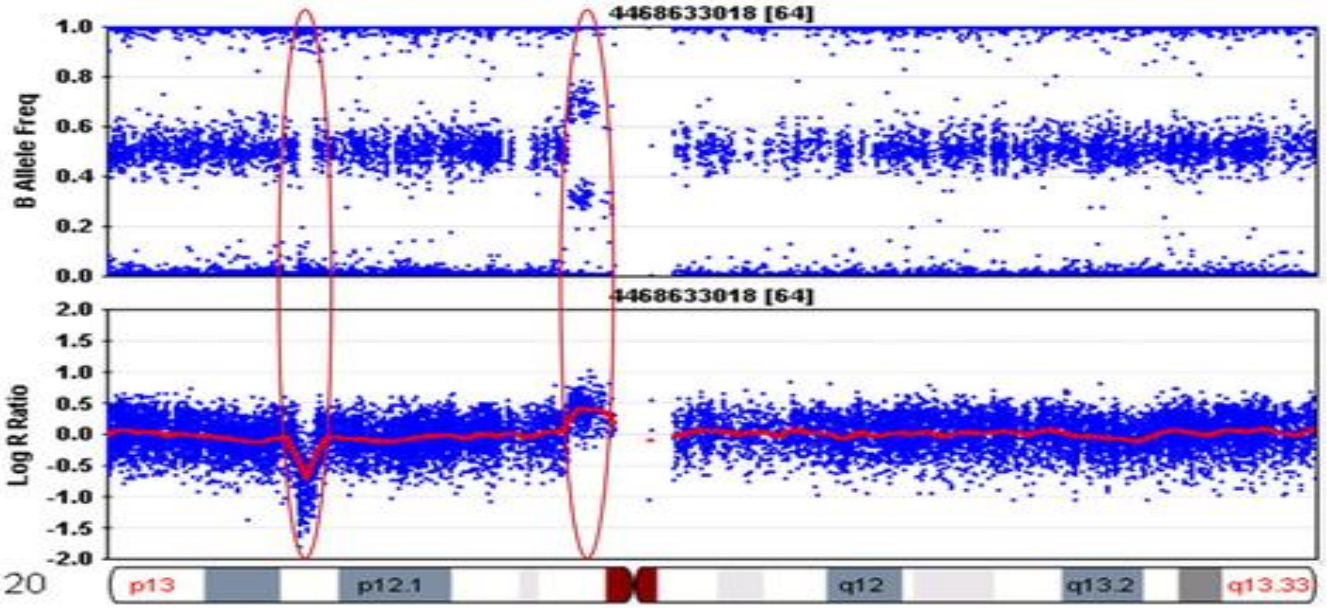
# Illumina SNP Bead Chip Arrays: the analysis

The University Of Sheffield.

Across the whole genome we aim to identify genomic alterations:

- Identify loss-of-heterozygosity (LOH)
- Identify Copy Number Variation (CNV)
- Identify hot spot with statistical significance over noise
- Identify new loci where CNVs can occur in small amplicons

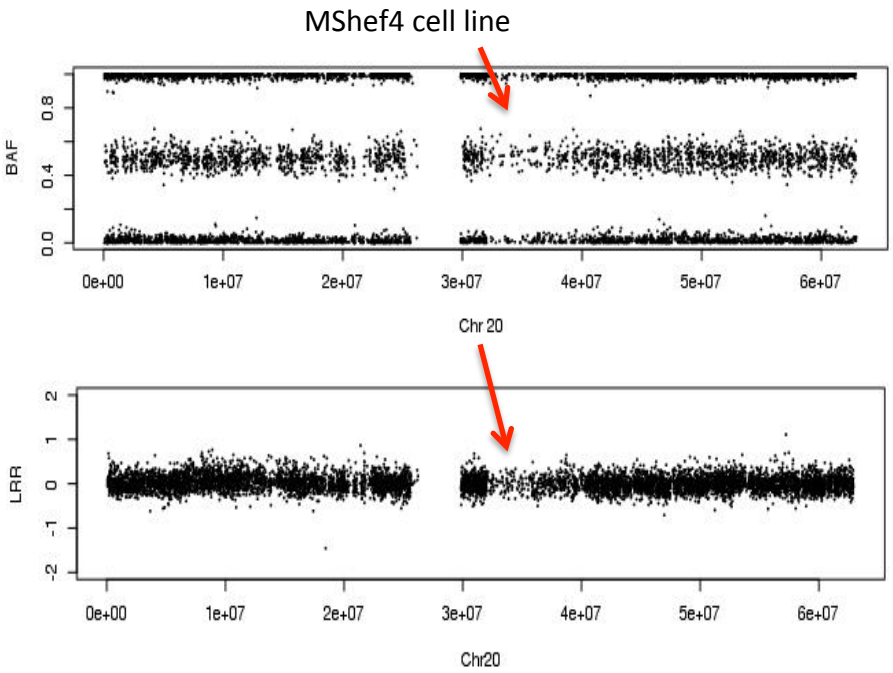
SNP arrays contains over 2.3 millions markers of most common and rare SNPs from the 1kGP (MAF>2.5%) for diverse world populations.



# ES cell lines screening with Illumina Arrays

The University Of Sheffield.

| Cell Line | Platform | Cell Bank | Abnormalities | Notes |
|---|---|---|---|---|
| MShef2 | Illumina HumanCytoSNP-12 | mShef2 PreMCB | None significant | Noisy BAF in: Chr9, Chr20, Chr17 |
| MShef3 | Illumina HumanCytoSNP-12 | mShef3 PreMCB | None significant | Noisy BAF in: Chr9, Chr17, Chr20, ChrX |
| Mshef4 | Illumina HumanCytoSNP-12 | mShef4 PreMCB | None significant | Noisy BAF in: Chr9, Chr17, Chr20, ChrX |
| MShef5 | Illumina HumanCytoSNP-12 | mShef5 PreMCB | None significant | Noisy BAF in: Chr9, Chr20, ChrX |
| Shef6 | Illumina HumanCytoSNP-12 | UKSCB material | None significant | Noisy BAF in: Ch1, Ch9, Chr11, Ch17, Chr20, ChrX |
| MShef7 | Illumina HumanCytoSNP-12 | mShef7 PreMCB | None significant | Noisy BAF in: Chr9, Chr17,Chr20,ChrX |
| MShef8 | Illumina HumanCytoSNP-12 | mShef8 PreMCB | None significant | Noisy BAF in: Chr6, Chr9,Chr16, Chr17 ,Chr20, ChrX |
| MShef10 | Illumina HumanCytoSNP-12 | mShef10 PreMCB | None significant | Noisy BAF in: Chr7, Chr9, Chr17, Chr20, ChrX |
| MShef11 | Illumina HumanCytoSNP-12 | mShef11 PreMCB | None significant | Noisy BAF in: Chr6, chr9, Chr17, Chr20,ChrX |
| MShef12 | Illumina HumanCytoSNP-12 | mShef12 PreMCB | None significant | Noisy BAF in: Chr9, Chr19, Chr20, ChrX |
| MShef13 | Illumina HumanCytoSNP-12 | mShef13 PreMCB | None significant | Noisy BAF in: Chr3, Chr6, Chr7,Chr9,Chr17,Chr20,ChrX |
| MShef14 | Illumina HumanCytoSNP-12 | mShef14 PreMCB | None significant | Noisy BAF in: Chr1, Chr9, Chr17, Chr20, ChrX |
| MShef4 | Illumina HumanOmin2.5-8 | mShef4 QC1 clone B1 | None significant | Clone B1 |
| MShef4 | Illumina HumanOmin2.5-8 | mShef4 QC1 clone B4 | None significant | Clone B4 |
| MShef4 | Illumina HumanOmin2.5-8 | mShef4 QC1 clone B5 | None significant | Clone B5 |
| MShef4 | Illumina HumanOmin2.5-8 | mShef4 QC1 clone B8 | None significant | Clone B8 |
| MShef4 | Illumina HumanOmin2.5-8 | mShef4 QC1 clone B9 | None significant | Clone B9 |




**SNPs analysis**:

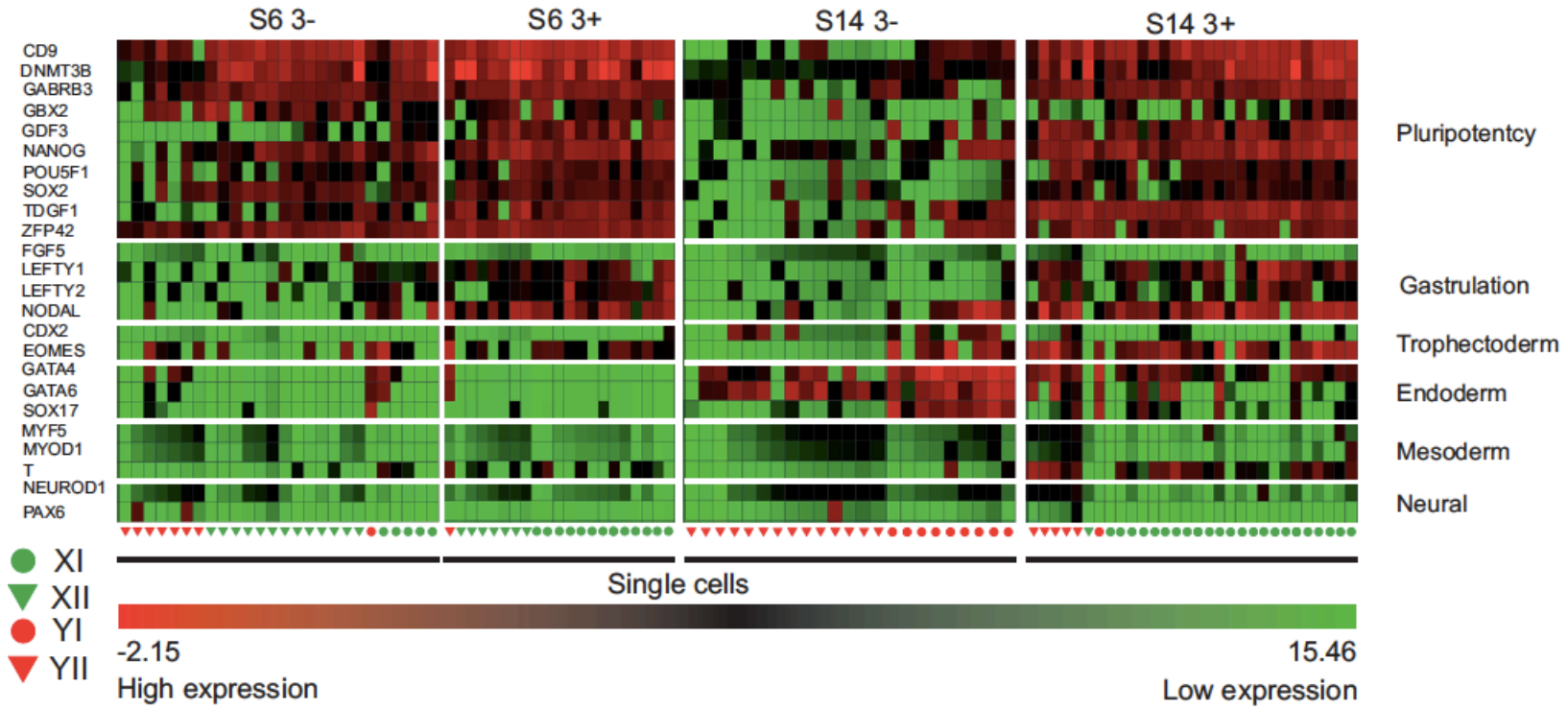
- Correlation of BAF and LRR for each line and each chromosome
- No CNVs and LOH were identified
- "Noisy" loci were identified across all lines

Illumina **HumanOmin2.5-8 BeadChips**:
>2.3 millions markers of most common and rare SNPs from the 1kGP (MAF>2.5%) for diverse world populations

The University Of Sheffield.

Studying the impact of biophysical force in culture and in 3D culture at system-level for each cell.

- integrated data approach
- predictive models of selection in specific condition
- single cell approach




Improve the efficacy of stem cell therapy, studying the impact of silent mutation in transplanted line.

# Summary

Modern biology studies fine details on large scale and has generated challenges that we can only approach by integrating disciplines in a system-level approach

With large data and assays in high throughput fashion we need to quantify uncertainty to be able to estimate the proportion of false discovery we have in our selections. Impact of Machine Learning in analysing and using this data is huge

Import to model the uncertainty when signal is very close to noise, i.e. single cell data. Successful use of Machine Learning methods need to be supported by appropriate experimental designs.

The road to Personalise Medicine is now clear:

Computational Models -->Bench -->Computational Model --> bedside

# Acknowledgements


***puma* group:**
Magnus Rattray
Xuejun Liu
Neil Lawrence
Giudo Sanguinetti
Antii Honkela
Peter Glaus

Marcelo Rivolta
Oliver Thompson
Peter Andrews
Harry Moore

**Pluripotent Stem Cell Platform – PSCP**

Jens Neilsen
Adam Kneebone
Sarah Craig
Lisa Flaherty
Elizabeth Boggis
Alex Rothman

**The Patients**
**The volunteers**
**at CV department**

Tim Chico
Sarah Langridge
David Crossman
Jane Arnold
Allison Morton
Rob Storey

Guillaume Hautbergue
J. Cooper-Knox
Winston Hide