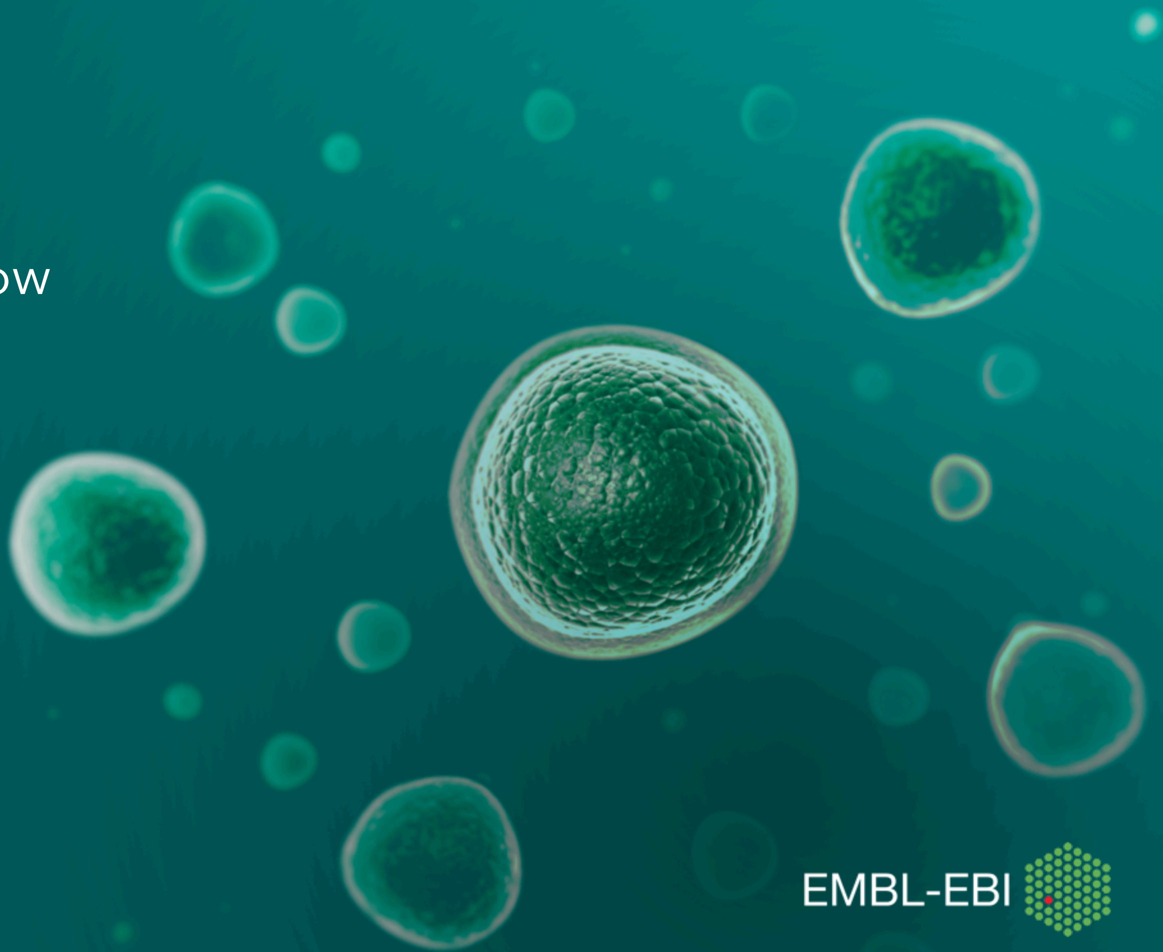
Using R and Bioconductor to explore genetic effects on single-cell gene expression

Davis McCarthy
NHMRC Early Career Fellow
Stegle Group, EMBL-EBI

@davisjmcc

www.ebi.ac.uk

www.hipsci.org



EMBL-EBI

1. (How) Can we carry out single-cell QTL studies?

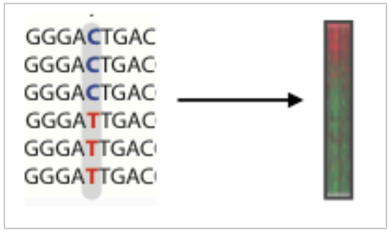
2. How will we scale Bioconductor single-cell tools to datasets of millions of cells?

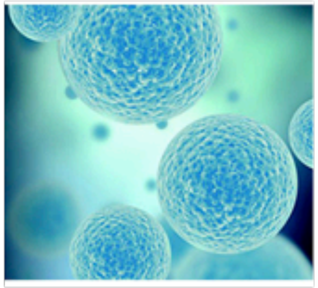
Single-cell QTL studies

Combining individual-to-individual and cell-to-cell heterogeneity

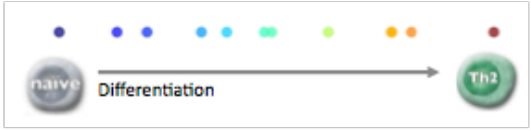
variation of interest

population variation



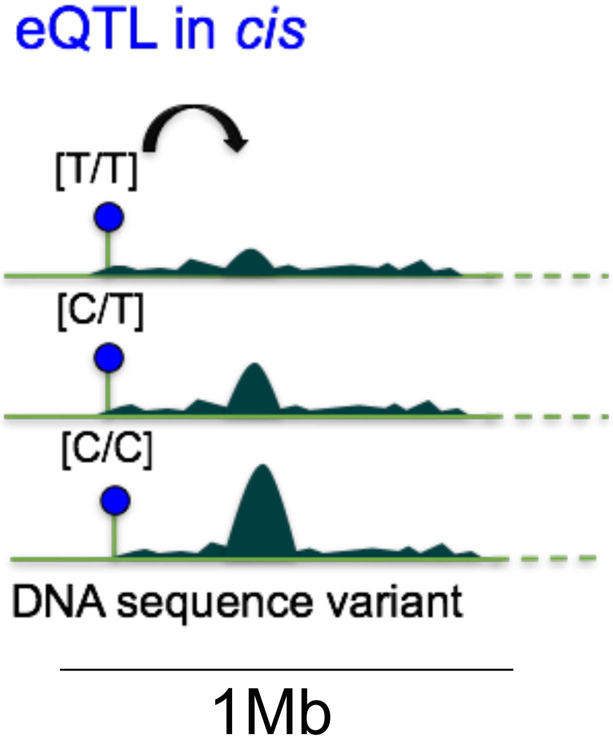


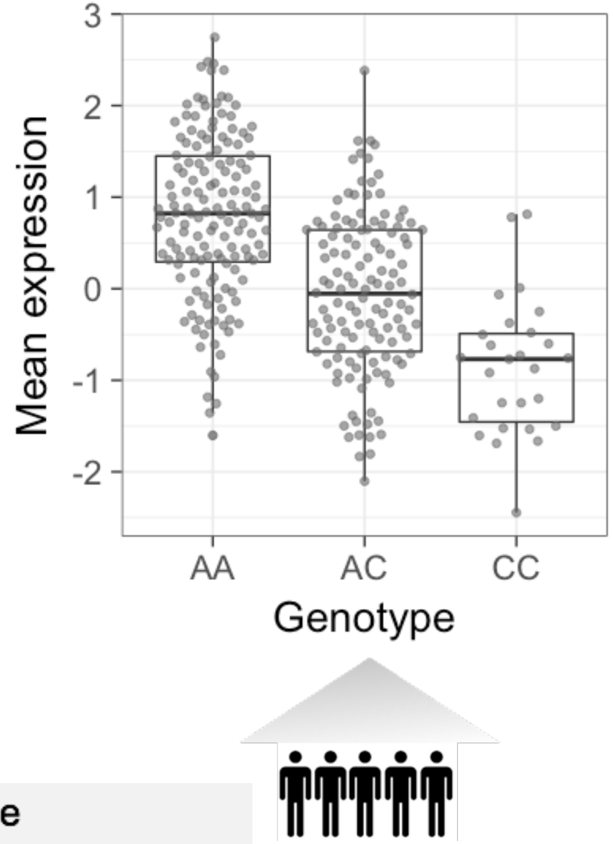
single-cell variation



Single-cell QTL mapping

Recap: QTL in population variation datasets

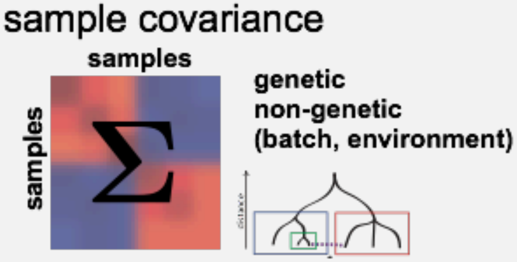




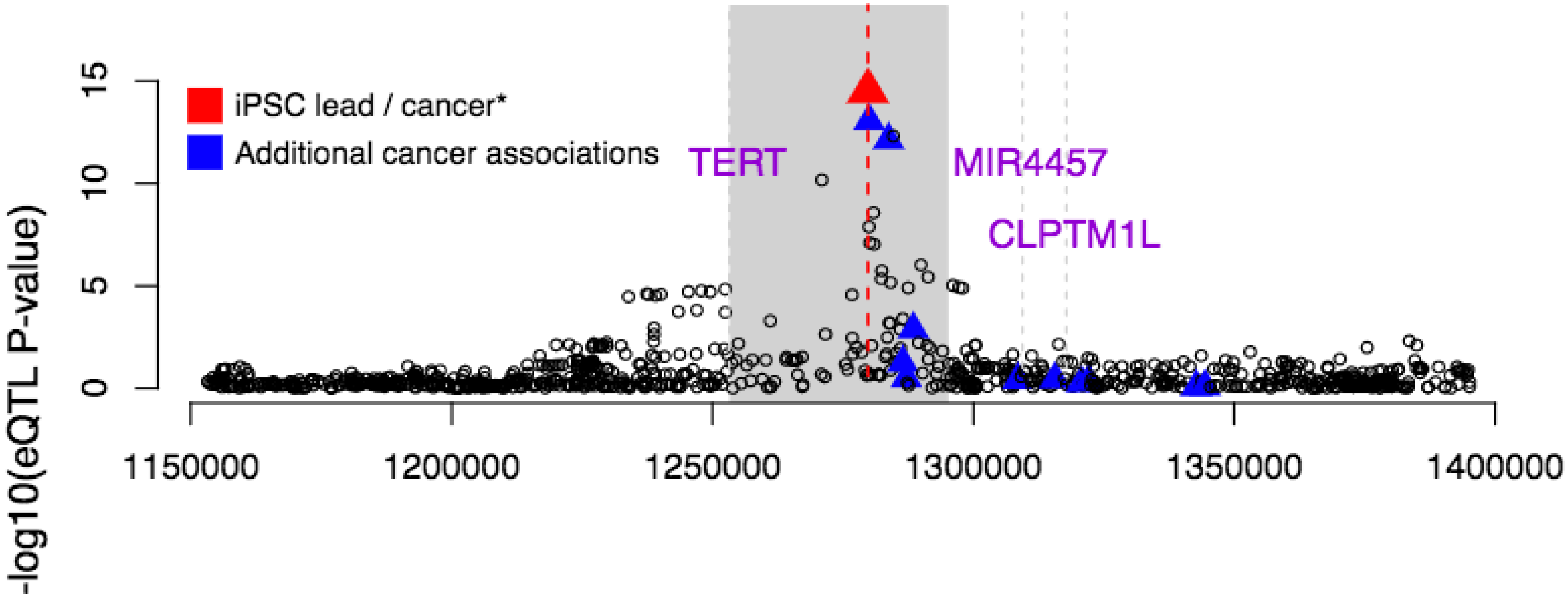
Linear mixed model:

Y = covars + SNP + g + e

$g \sim N(0, K\sigma_g)$; $e \sim N(0, I\, \sigma_e)$



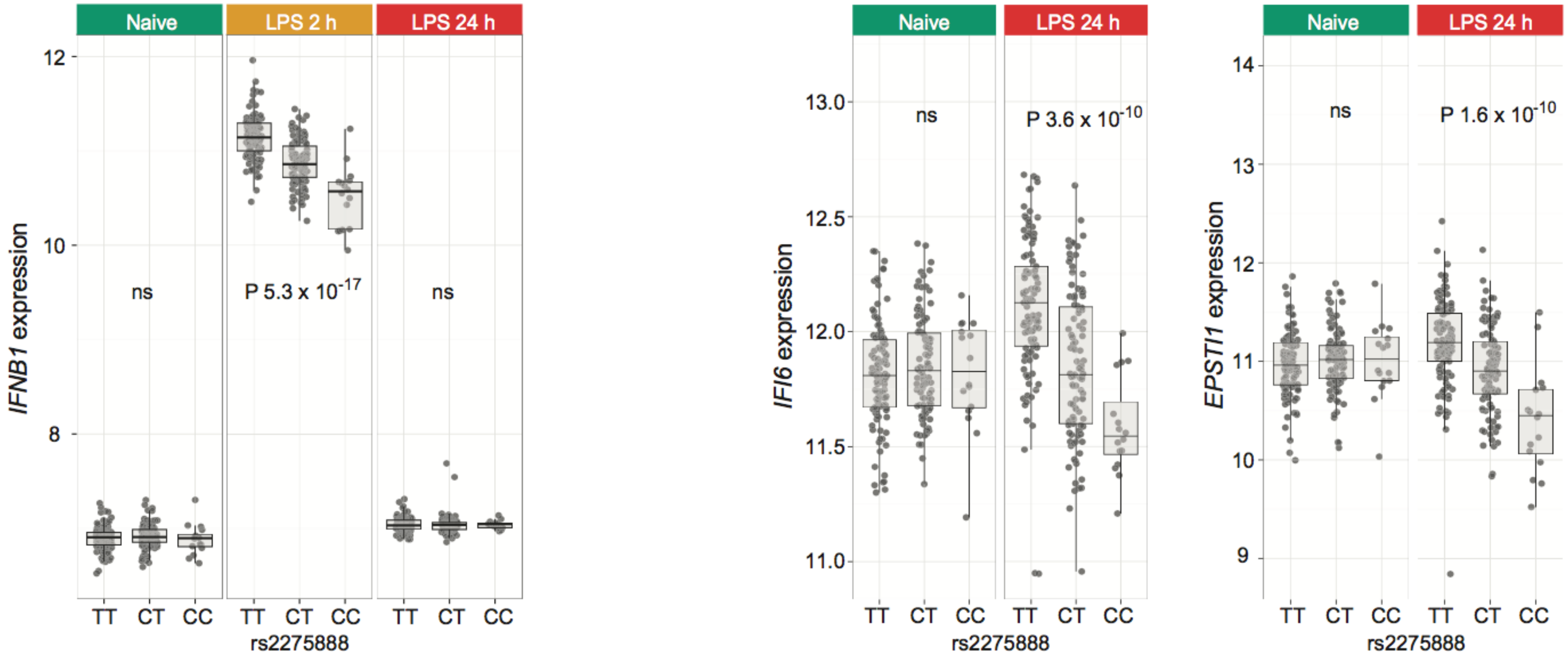
Motivating example (I): in induced pluripotent stem cells we can link disease risk variants to gene expression



TERT has an iPS eQTL that overlaps a cancer risk variant.

Kilpinen, Goncalves et al, *Nature*, 2017

Motivating example (II): genetic effects on gene expression (can) depend on context

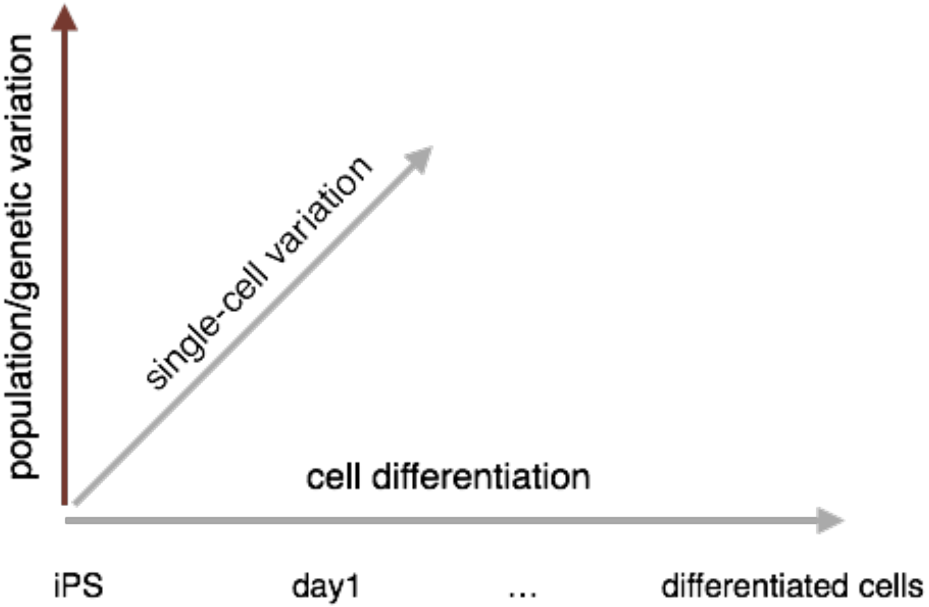


Fairfax et al, *Science*, 2014: Fig. 3

EMBL-EBI

scRNA-seq as a readout for QTL analyses offers new phenotypes to study with unprecedented characterisation of cell types and states

HipSci



population/genetic variation

single-cell variation

cell differentiation

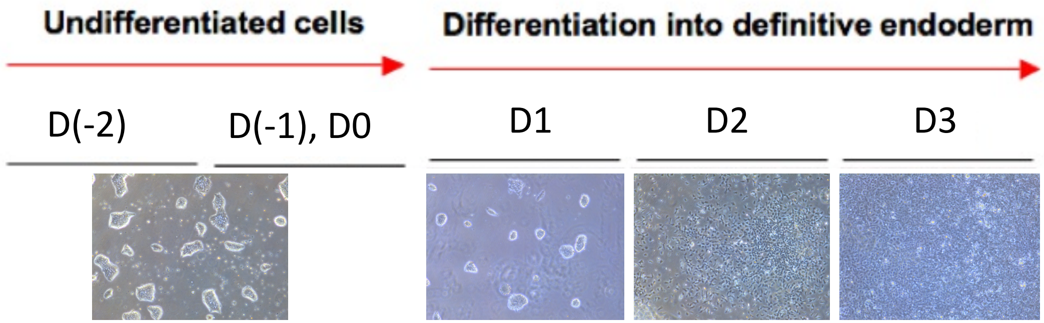
iPS

day1

...

differentiated cells

Definitive endoderm differentiation from iPSCs

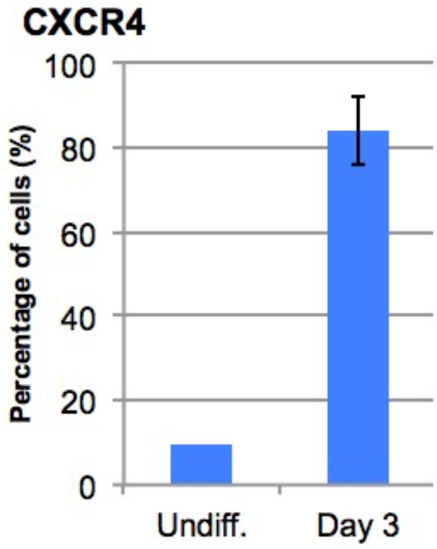


Oct4$^+$
Nanog $^+$
Sox2 $^+$

Oct4$^+$
Nanog $^+$
Sox2 $^-$
Bra$^+$
Eomes$^+$

Oct4$^+$
Nanog $^-$
Sox2 $^-$
Bra$^+$
Eomes$^+$
Mixl1$^+$

Oct4$^-$
Nanog $^-$
Sox2 $^-$
Bra$^-$
Eomes$^+$
Sox17$^+$
Mixl1$^+$
GSC$^+$
CXCR4$^+$
……



Adapted from Touboul et al, 2010 Hepatology

- How do we characterise the heterogeneity of transcriptome states in iPSCs during differentiation?

- How do genetic variants influence single-cell states?

- How do genetic effects differ in differentiated cells?

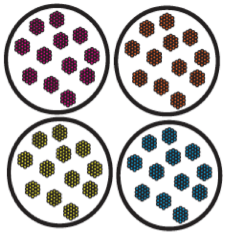
- **(How) Can we map QTLs for single-cell phenotypes?**

- How can we design a single-cell QTL study that:
  1. Can feasibly assay cells from a large enough number of individuals?
  2. Is robust to batch effects?

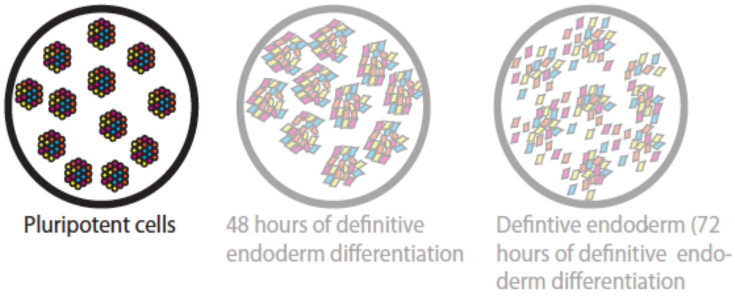
Donor pooling can increase throughput and ameliorate batch effects

HipSci

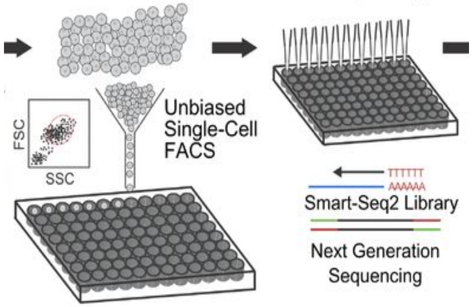
4-6 lines



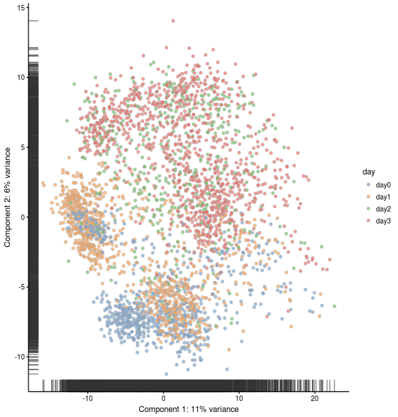
Grown together in mixed population


Pluripotent cells
48 hours of definitive endoderm differentiation
Defintive endoderm (72 hours of definitive endo-derm differentiation

Dissociate into single cells for RNA-Seq analysis



Li et al, *EMBO Rep.*, 2016



scRNA-seq data for 100s cells per donor

Shradha Amatya, Mariya Chhatriwala, Jose Garcia-Bernardo, Ludovic Vallier

EMBL-EBI

Computational challenge: Donor ID

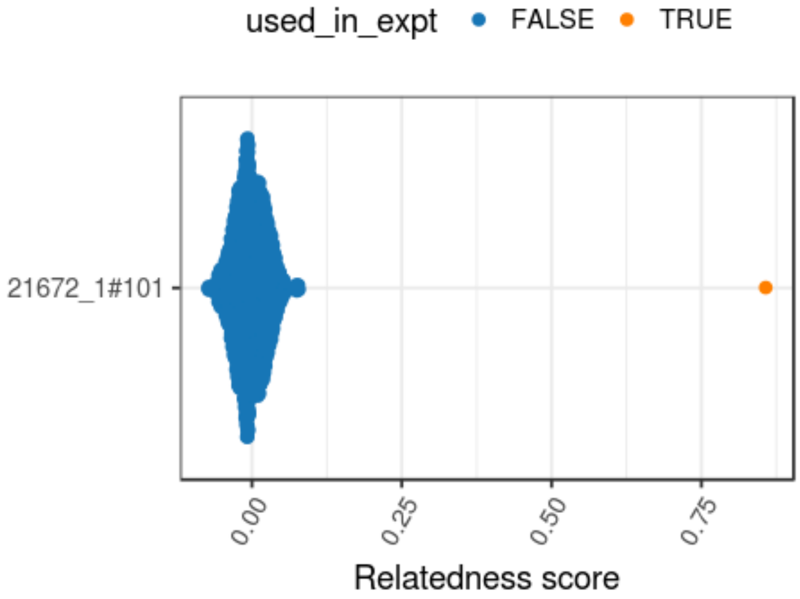
At the point of sequencing, we do not know which individual a cell came from.

So can we:

- Identify the donor for each cell?
  - When the donor genotypes are known?
  - When the donor genotypes are unknown?

Approach when donor genotypes are known

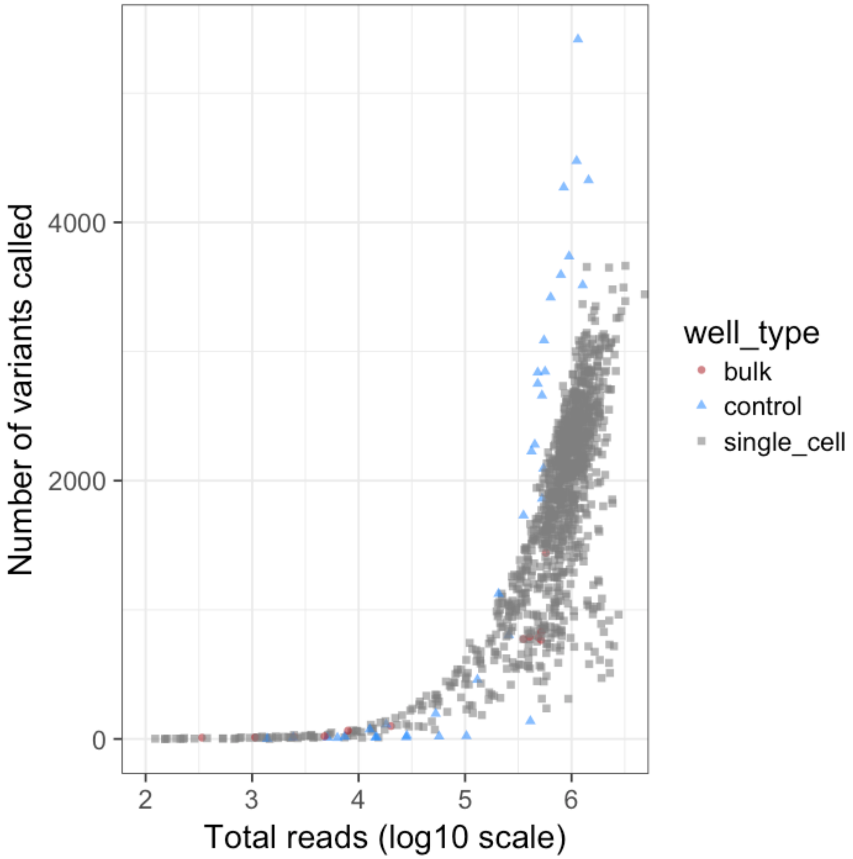
- Variants called with GATK HaplotypeCaller from scRNA-seq reads
- Matched against genotypes for 400 HipSci donors by estimating “genomic relatedness” (average allelic correlation) between cell and line
- Use highest relatedness score to identify line from which cell came

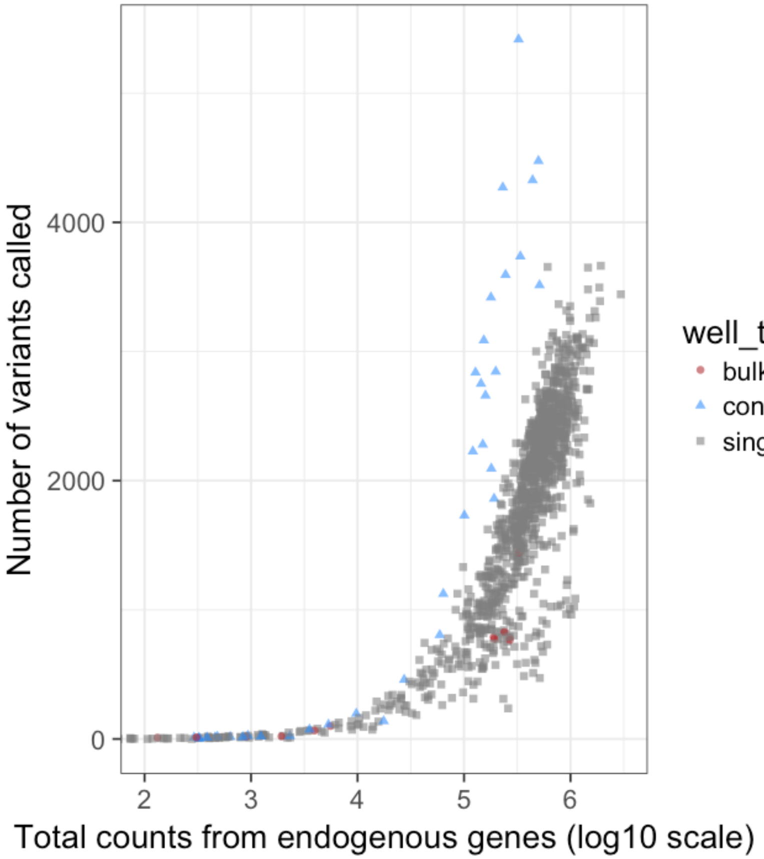


Approach when donor genotypes are known

- De novo variant calling from RNA-seq reads?
  - Too variable; not enough overlap with genotyped sites; bias to variant allele
- Call variants at known sites (e.g. dbSNP variants)?
  - Too slow; too many uninformative sites
- Call variants at known sites in the 1000 highest expressed genes in bulk iPSC samples?
  - Right balance between informative sites, speed and accuracy 😸

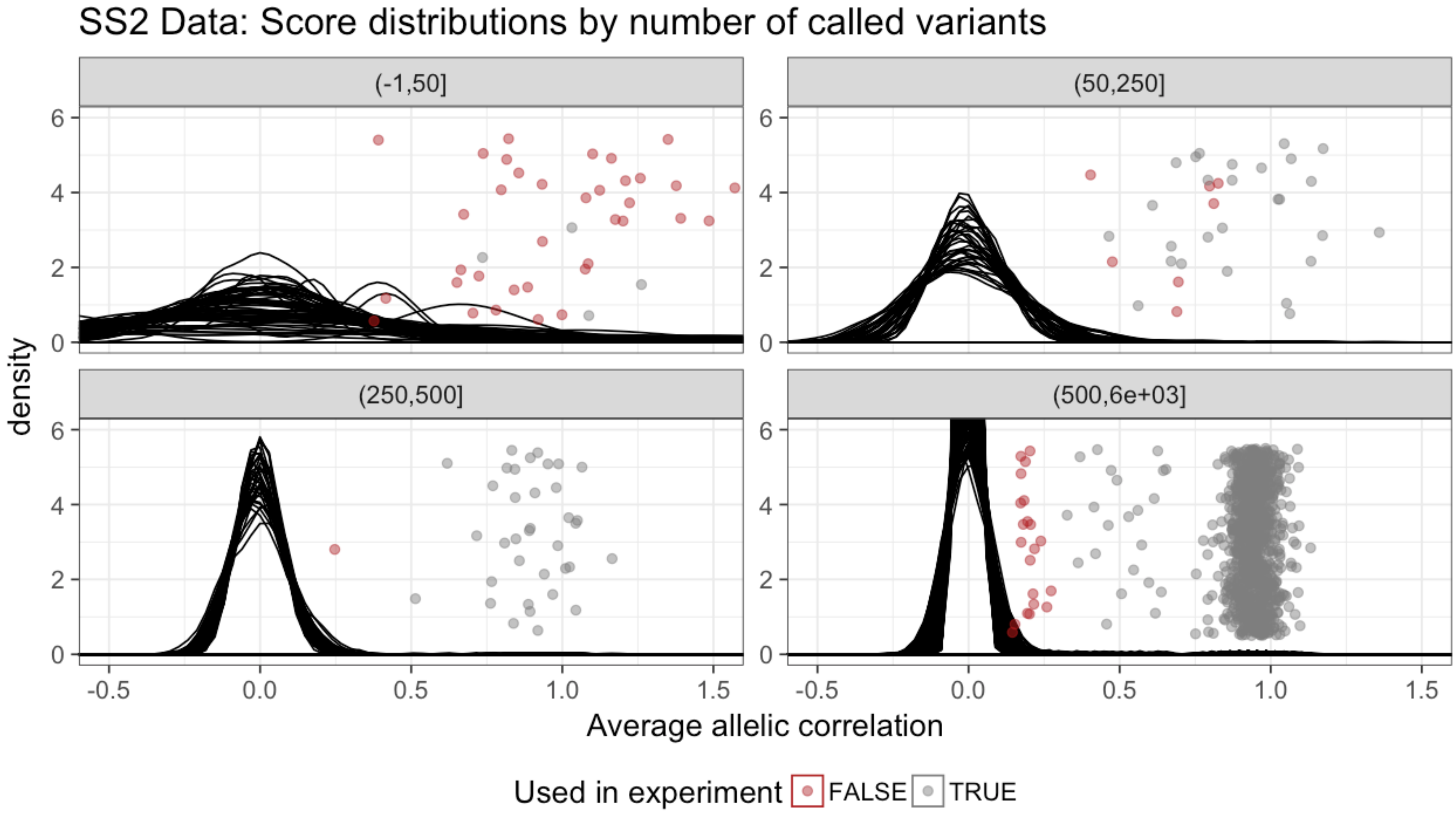
Variants called from Smartseq2 fibroblast data





Fibroblast cells from 3 individual donors

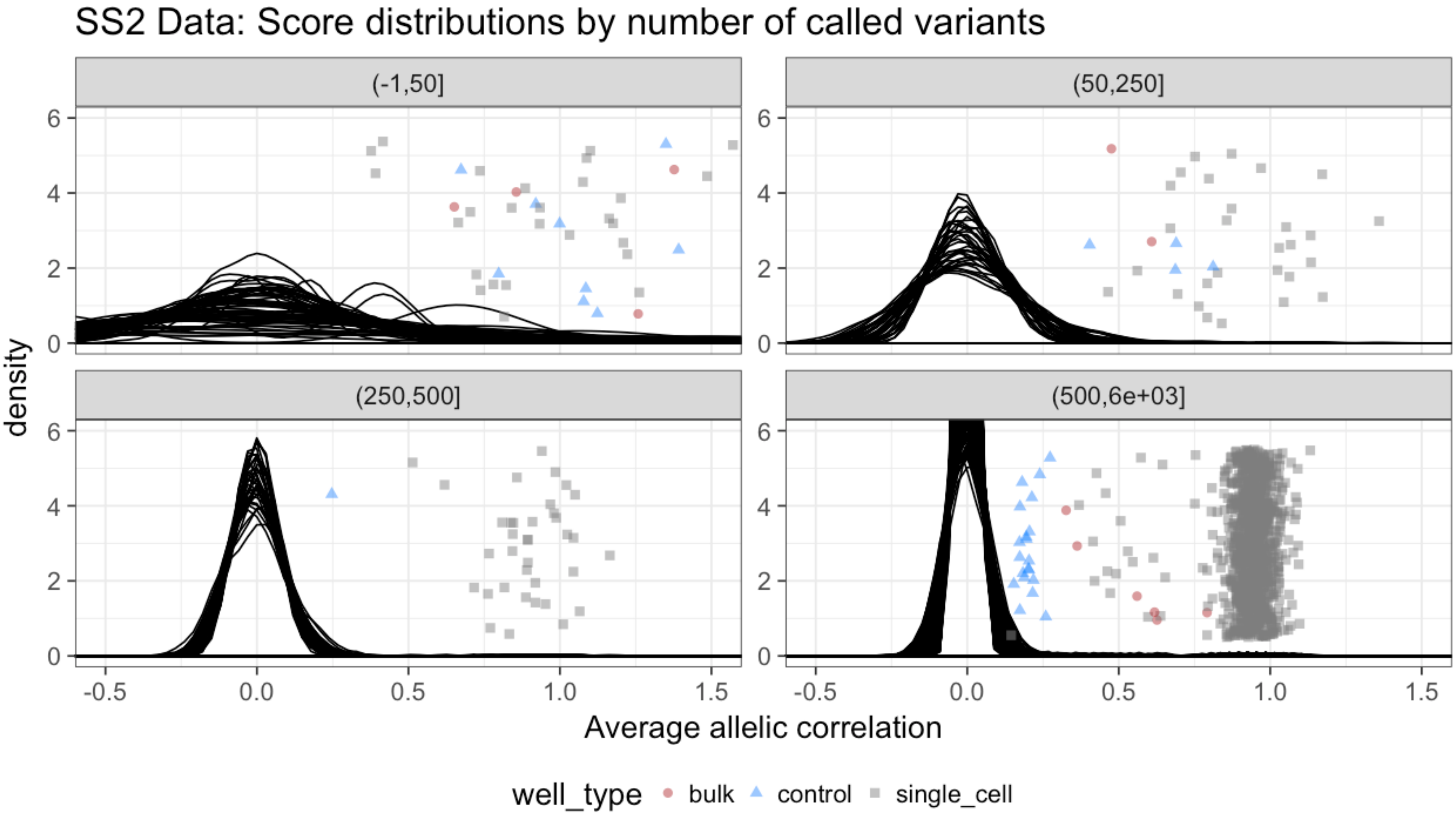
Score distributions for Smartseq2 data



Fibroblast cells from 3 individual donors

EMBL-EBI

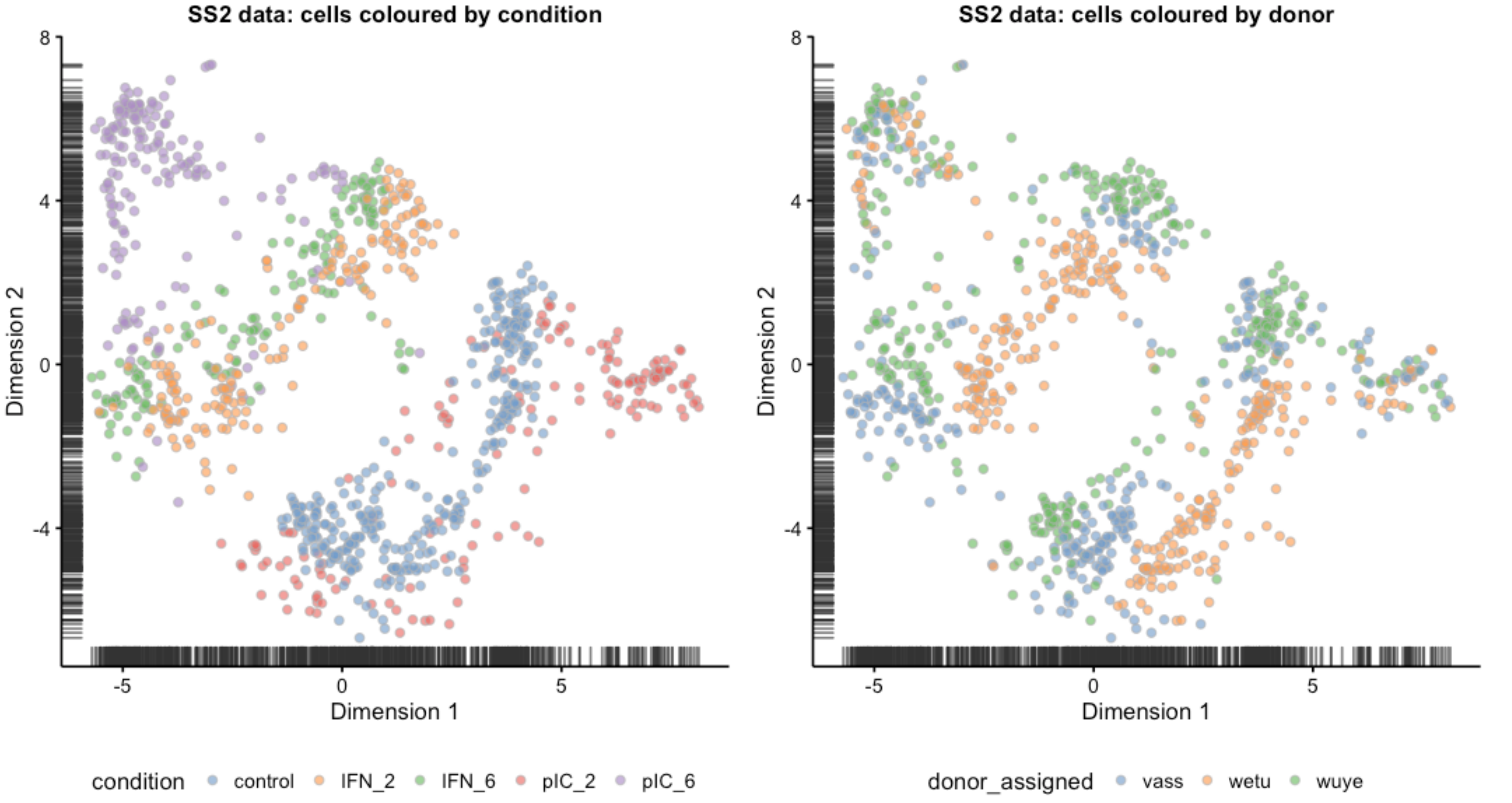
Score distributions for Smartseq2 data



Fibroblast cells from 3 individual donors

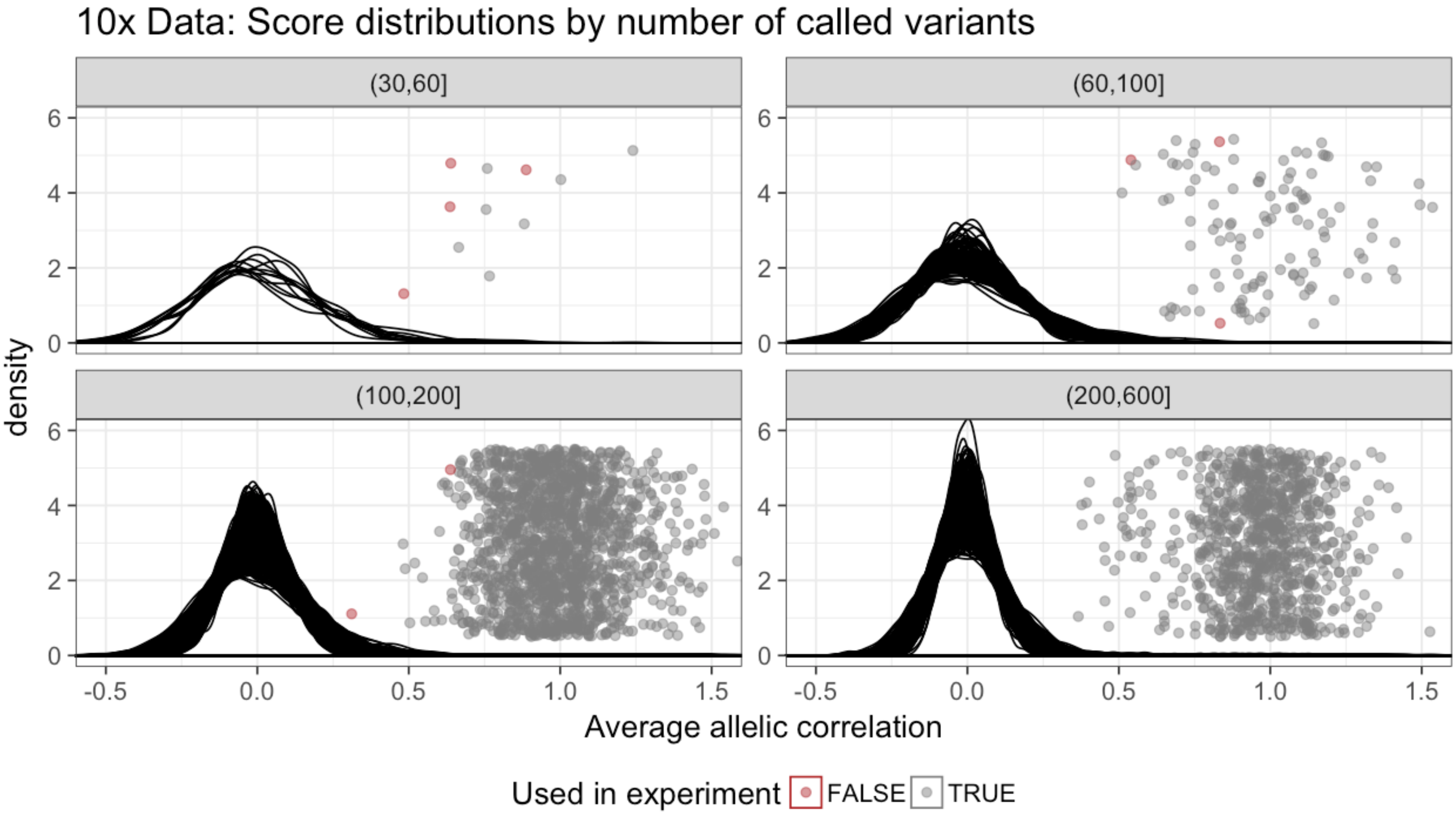
EMBL-EBI

There are large-scale differences in gene expression between donors



Fibroblast cells from 3 individual donors

Donor ID also works for sparser 10x data

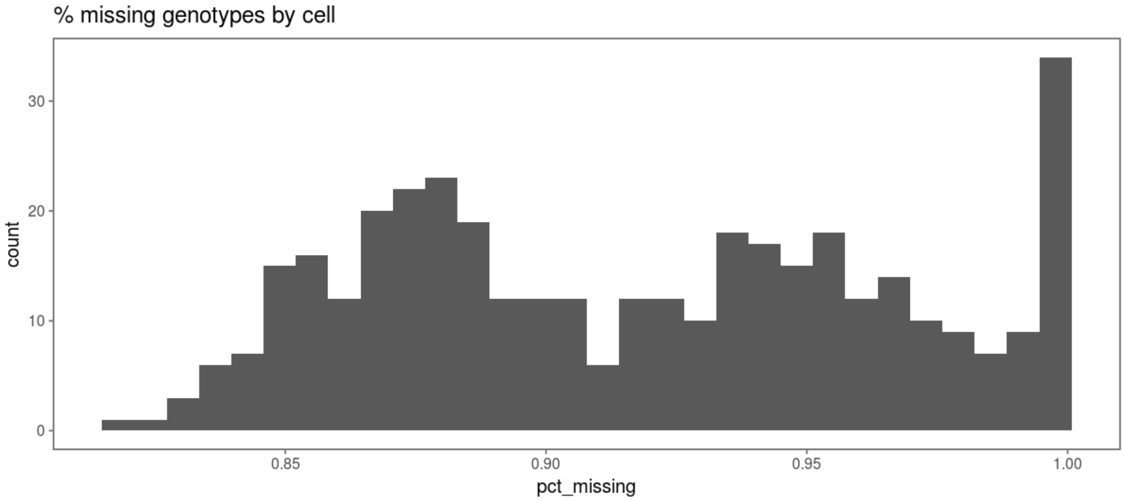


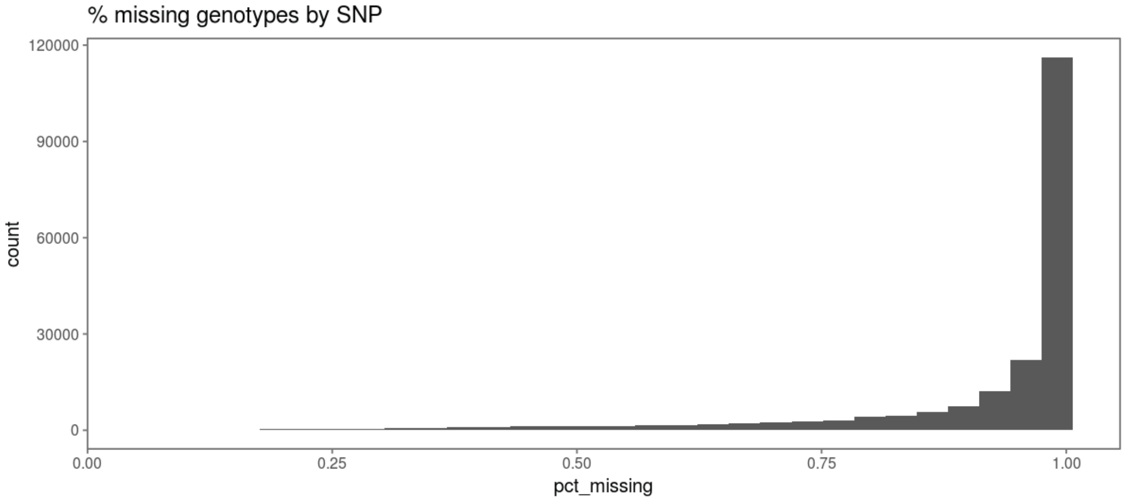
Fibroblast cells from 3 individual donors

Approach when donor genotypes are unknown

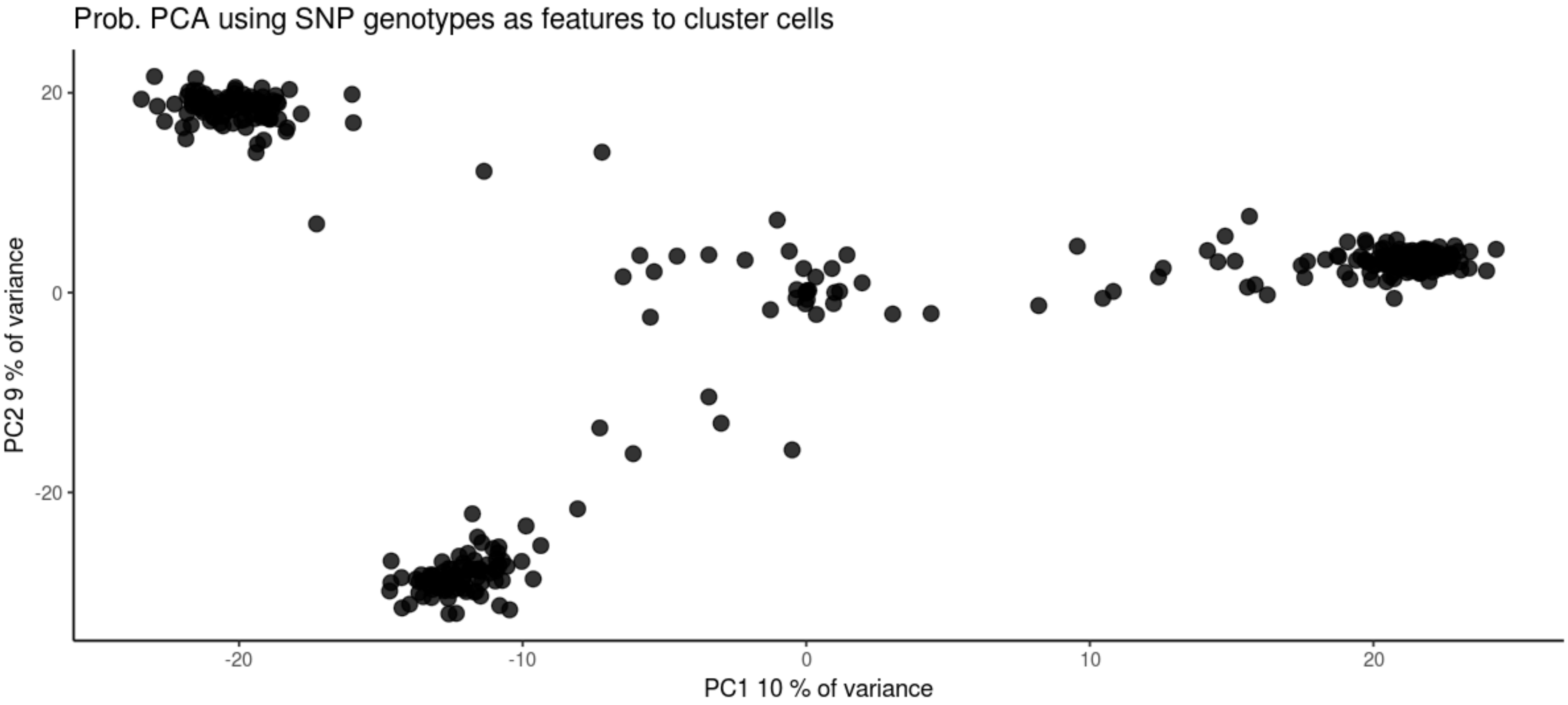
- Genotype cells at a list of HipSci variant sites
  - This need not be HipSci-specific. 1000G sites or similar would work just as well
- Merge cell VCFs to one big VCF (high % missing genotypes)
- Filter to SNPs on % missing genotypes threshold
  - <75% missing genotypes for SS2 data
  - <90% missing genotypes for 10x data
- Probabilistic PCA (*pcaMethods*)
- model-based clustering on PCs (*mclust*)

For Smartseq2 data, 250k SNPs are called, but most genotypes are missing





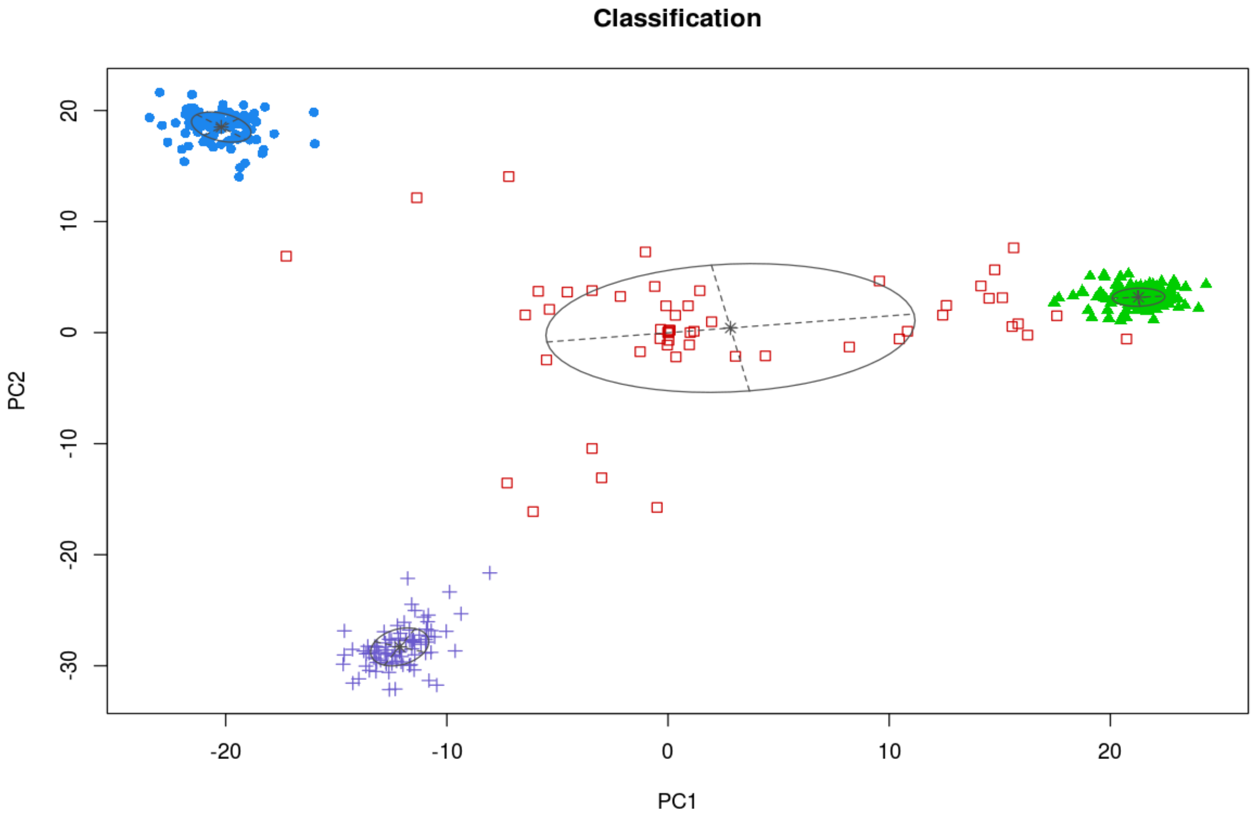
Prob. PCA on 22k filtered SNP genotypes works well



Fibroblast cells from 3 individual donors

EMBL-EBI

Specifying 4 clusters for mclust VEV model yields clean results

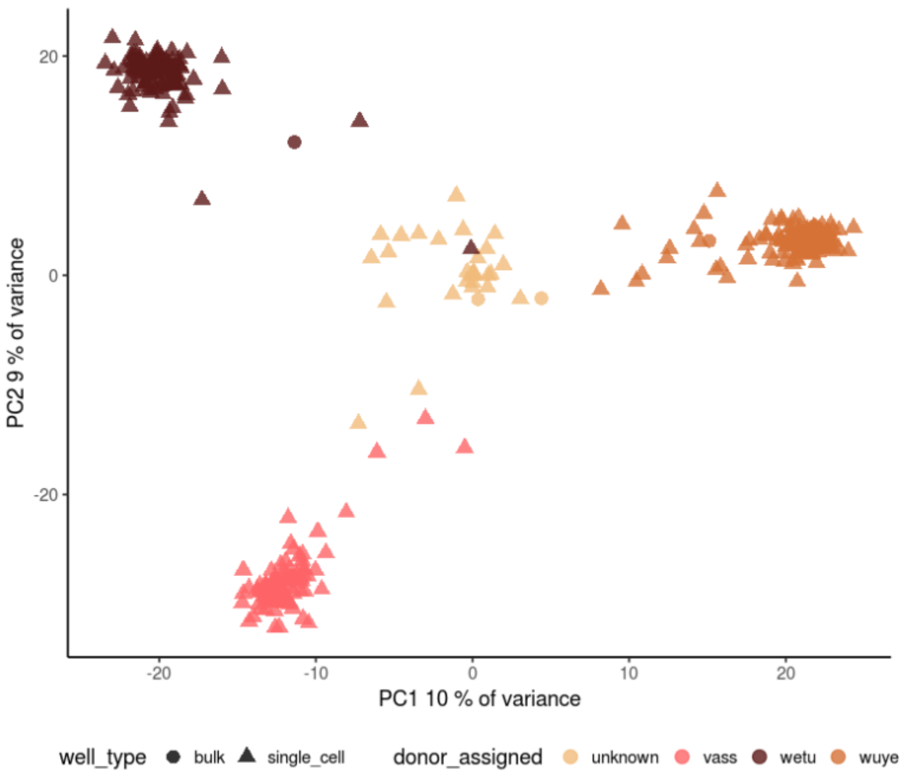


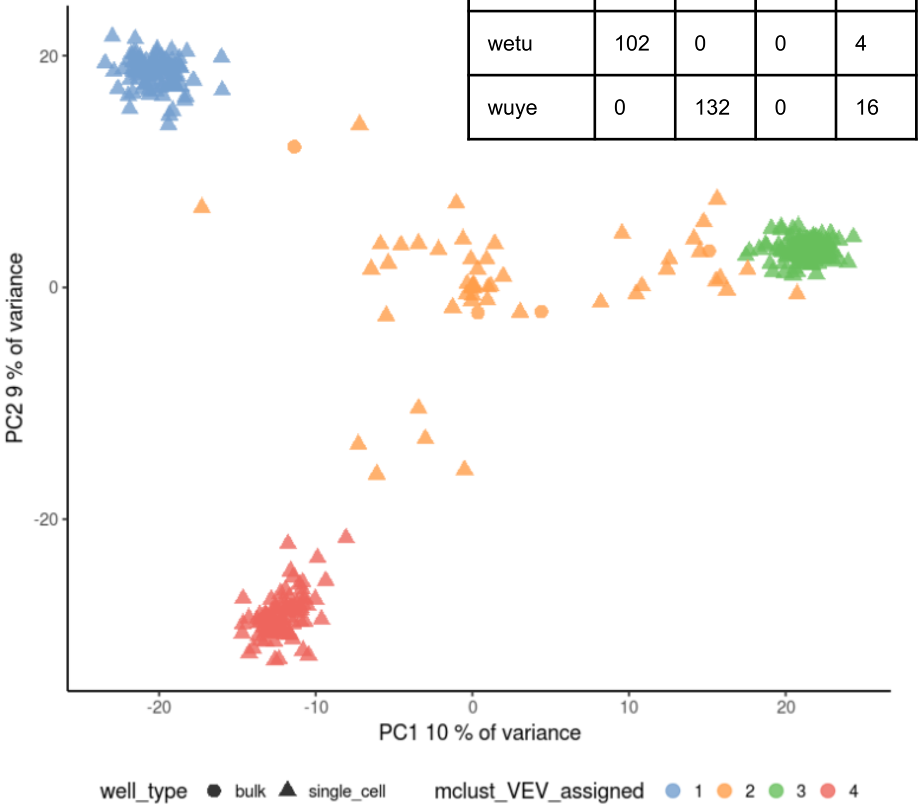
Interpret this as 3 “donor” clusters and an “unassigned” cluster

Favourable comparison of these results with donor ID using genotypes

	1	2	3	4
unknown	0	0	0	31
vass	0	0	84	3
wetu	102	0	0	4
wuye	0	132	0	16

Prob. PCA using SNP genotypes as features to cluster cells



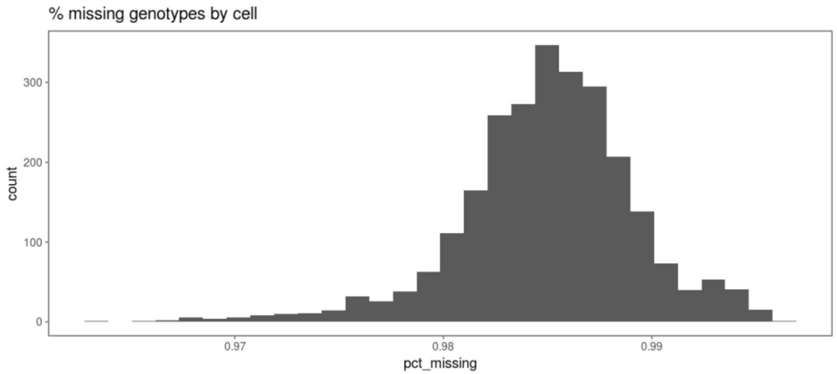


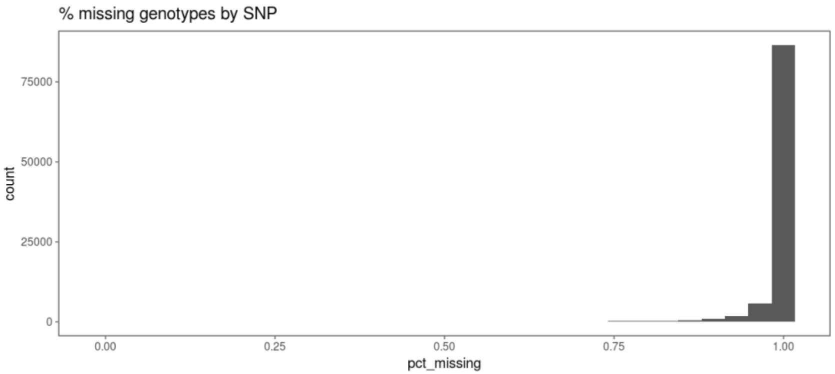
Adjusted Rand Index: 0.87 (1 is perfect agreement between donor assignments)

Fibroblast cells from 3 individual donors

- Donor ID without known genotypes works well for Smartseq2 protocol, which yields full-length transcript data.
- What about for 3' tag methods like 10x Chromium?

Fewer SNPs called from 10x data and most genotypes for a cell and a SNP are missing

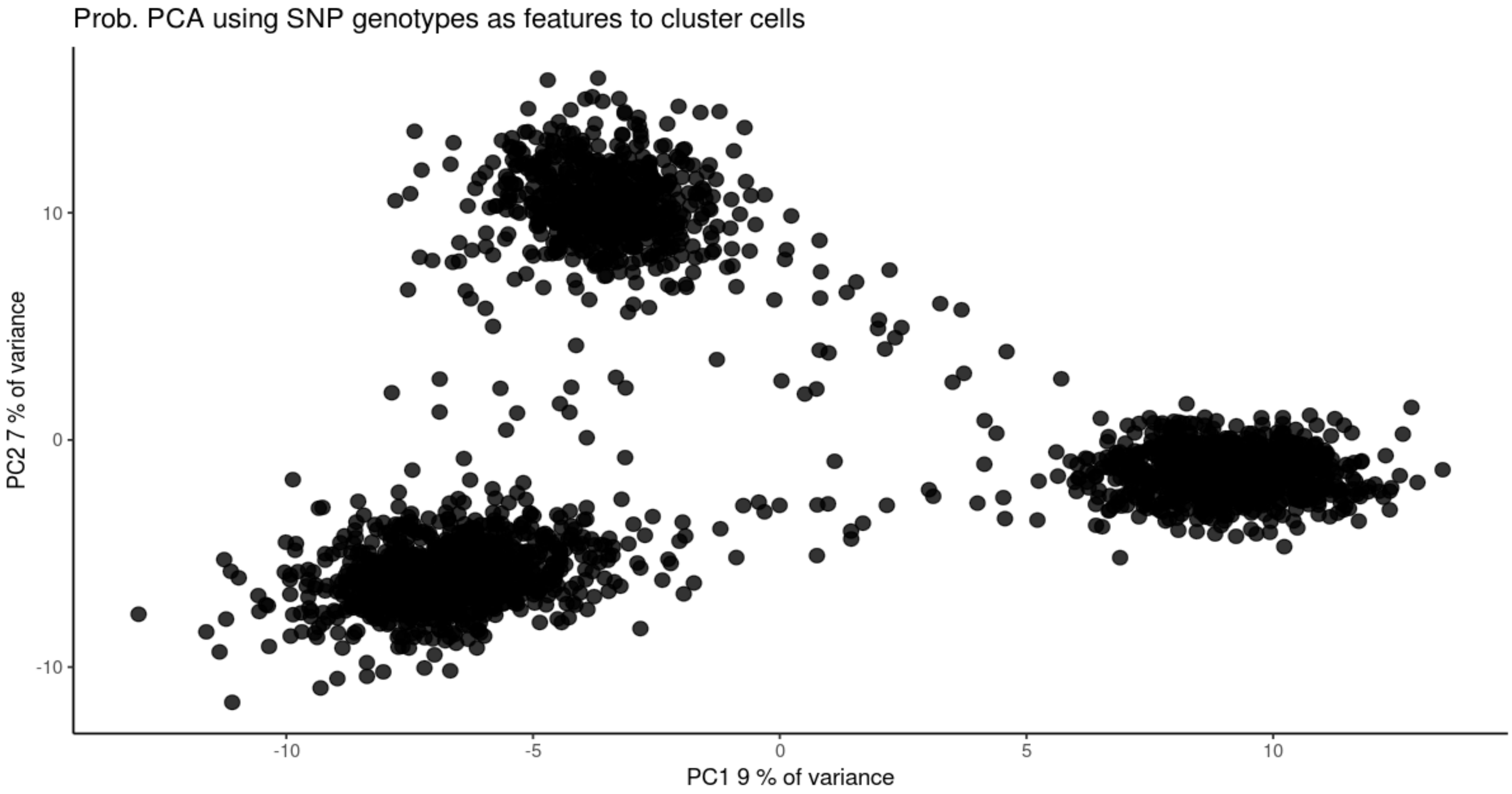




Total of 100k SNPs called across all 2553 cells. Few shared across cells.

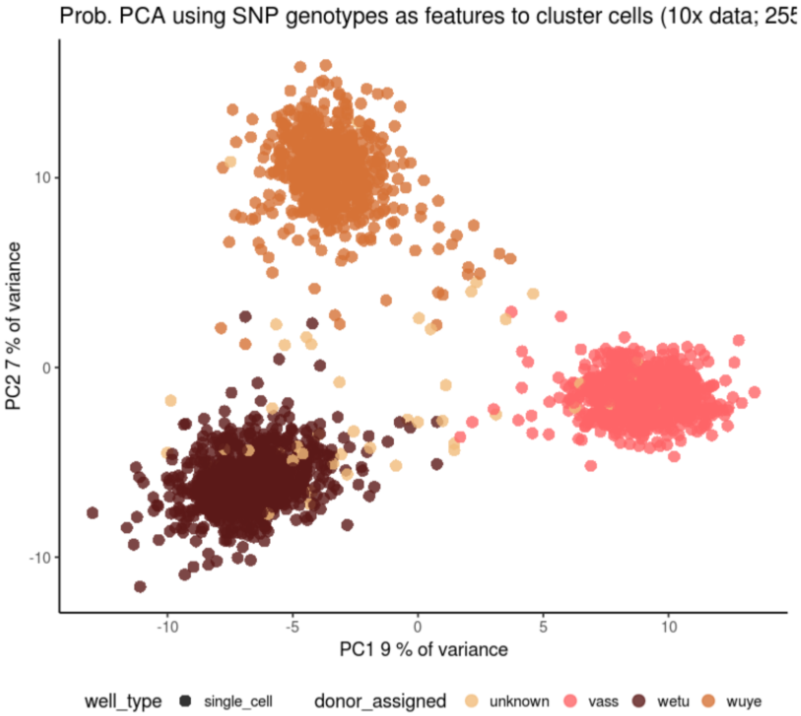
3110 SNPs with <90% missing genotypes across cells. Use these.

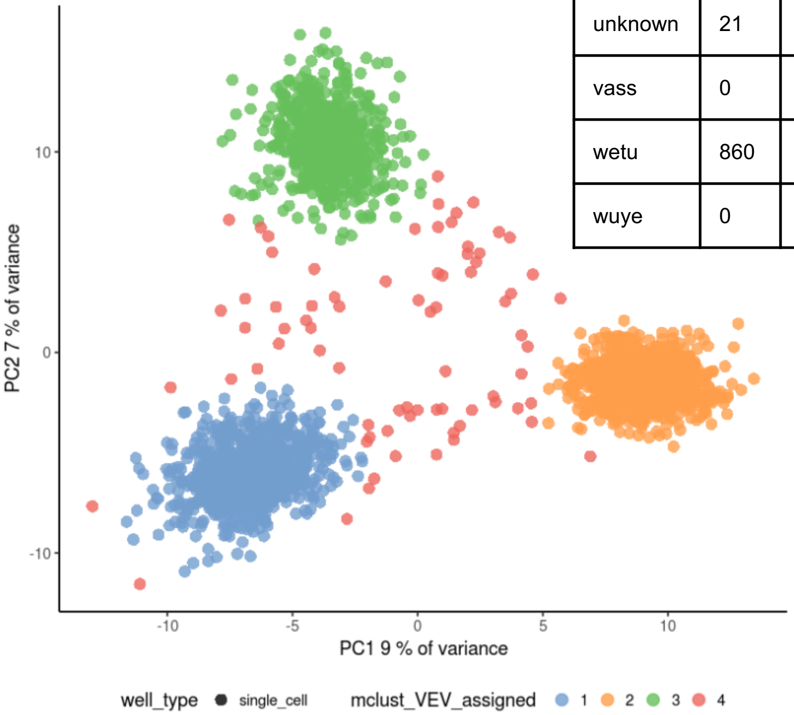
Prob. PCA on 3110 SNPs from 10x yields distinct clusters



Fibroblast cells from 3 individual donors

Excellent agreement with donor ID using donor genotypes for 10x data





	1	2	3	4
unknown	21	6	4	21
vass	0	944	0	12
wetu	860	0	0	18
wuye	0	0	642	25

Adjusted Rand Index: 0.95 (1 is perfect agreement between donor assignments)

Even better agreement than for SS2 data. Some cells with “unknown” donor assignment from approach with donor genotypes look “confidently” assigned to cells without using donor genotypes

Fibroblast cells from 3 individual donors

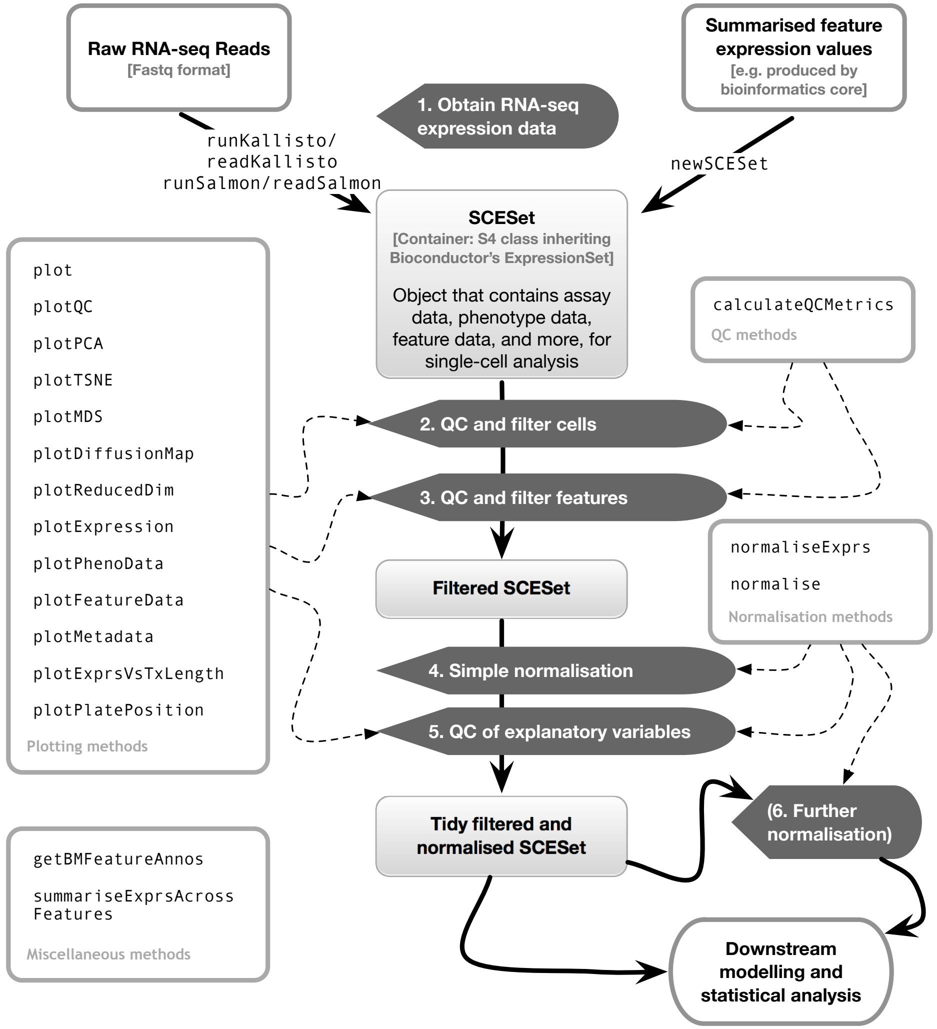
Donor ID summary and conclusions

- Genetic donor can be identified from SNP genotypes called from scRNA-seq reads.
- Donor ID works both from full-length transcript data (Smartseq2) and 3′ tag data (10x).
- Successful donor ID enables pooling of cells from multiple donors per experiment/run:
  - Scale up donor numbers necessary for QTL studies in minimal runs
  - Efficient use of expensive protocols
  - Enable experimental designs that are robust to batch effects
- Single-cell RNA-seq expands the phenotypes we can study with QTL mapping

Scaling Bioconductor single-cell tools to millions of cells

scater pre-processing and quality control workflow

From raw RNA-seq reads to a clean, tidy dataset ready for downstream analysis



Raw RNA-seq Reads
[Fastq format]

Summarised feature expression values
[e.g. produced by bioinformatics core]

1. Obtain RNA-seq expression data

runKallisto/
readKallisto
runSalmon/readSalmon

newSCESet

SCESet
[Container: S4 class inheriting Bioconductor's ExpressionSet]

Object that contains assay data, phenotype data, feature data, and more, for single-cell analysis

```
plot
plotQC
plotPCA
plotTSNE
plotMDS
plotDiffusionMap
plotReducedDim
plotExpression
plotPhenoData
plotFeatureData
plotMetadata
plotExprsVsTxLength
plotPlatePosition
```

Plotting methods

```
calculateQCMetrics
```

QC methods

2. QC and filter cells

3. QC and filter features

Filtered SCESet

```
normaliseExprs
normalise
```

Normalisation methods

4. Simple normalisation

5. QC of explanatory variables

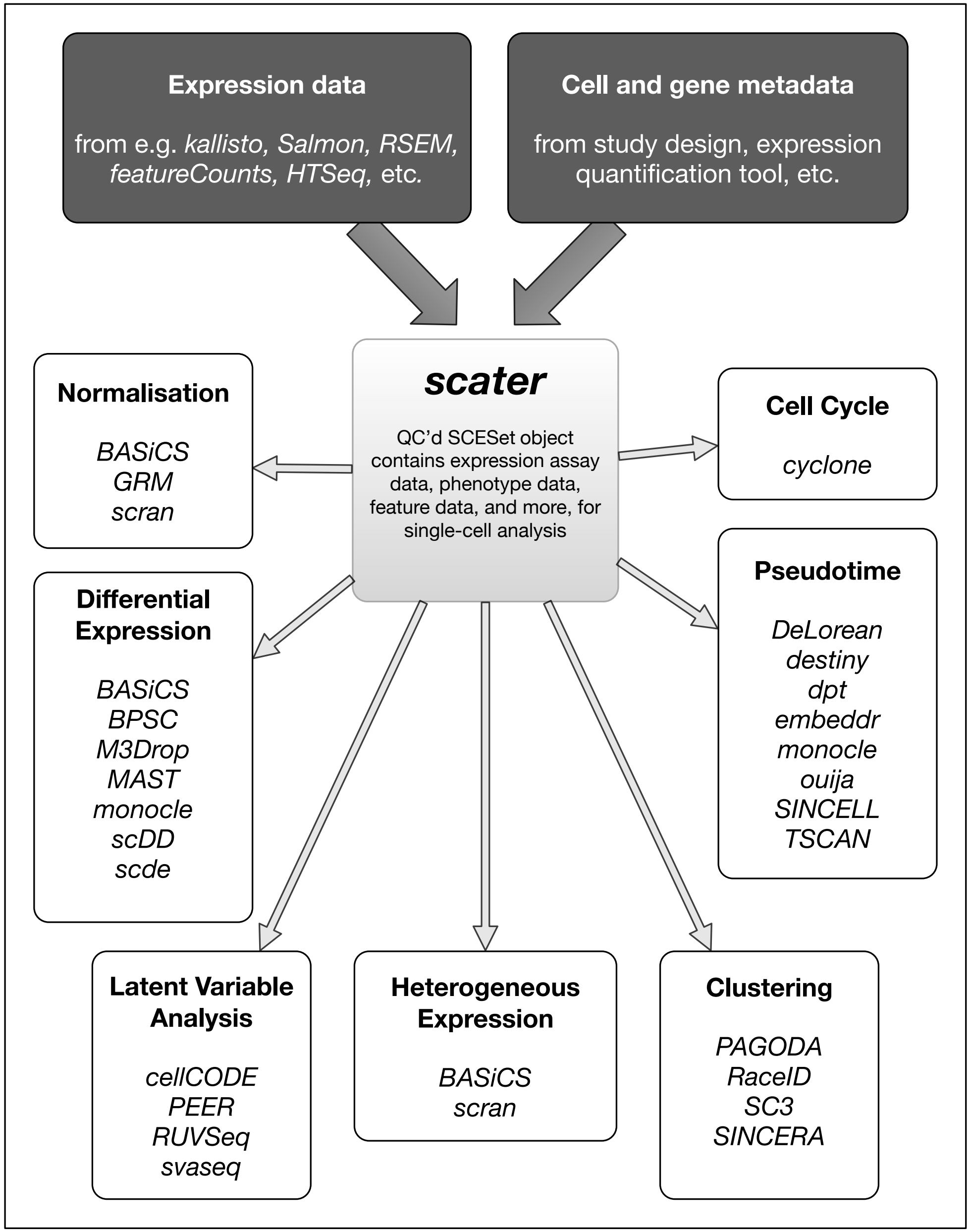
Tidy filtered and normalised SCESet

(6. Further normalisation)

Downstream modelling and statistical analysis

```
getBMFeatureAnnos
summariseExprsAcross
Features
```

Miscellaneous methods



Expression data

from e.g. *kallisto, Salmon, RSEM, featureCounts, HTSeq,* etc.

Cell and gene metadata

from study design, expression quantification tool, etc.

scater

QC'd SCESet object contains expression assay data, phenotype data, feature data, and more, for single-cell analysis

Normalisation

BASiCS
GRM
scran

Cell Cycle

cyclone

Differential Expression

BASiCS
BPSC
M3Drop
MAST
monocle
scDD
scde

Pseudotime

DeLorean
destiny
dpt
embeddr
monocle
ouija
SINCELL
TSCAN

Latent Variable Analysis

cellCODE
PEER
RUVSeq
svaseq

Heterogeneous Expression

BASiCS
scran

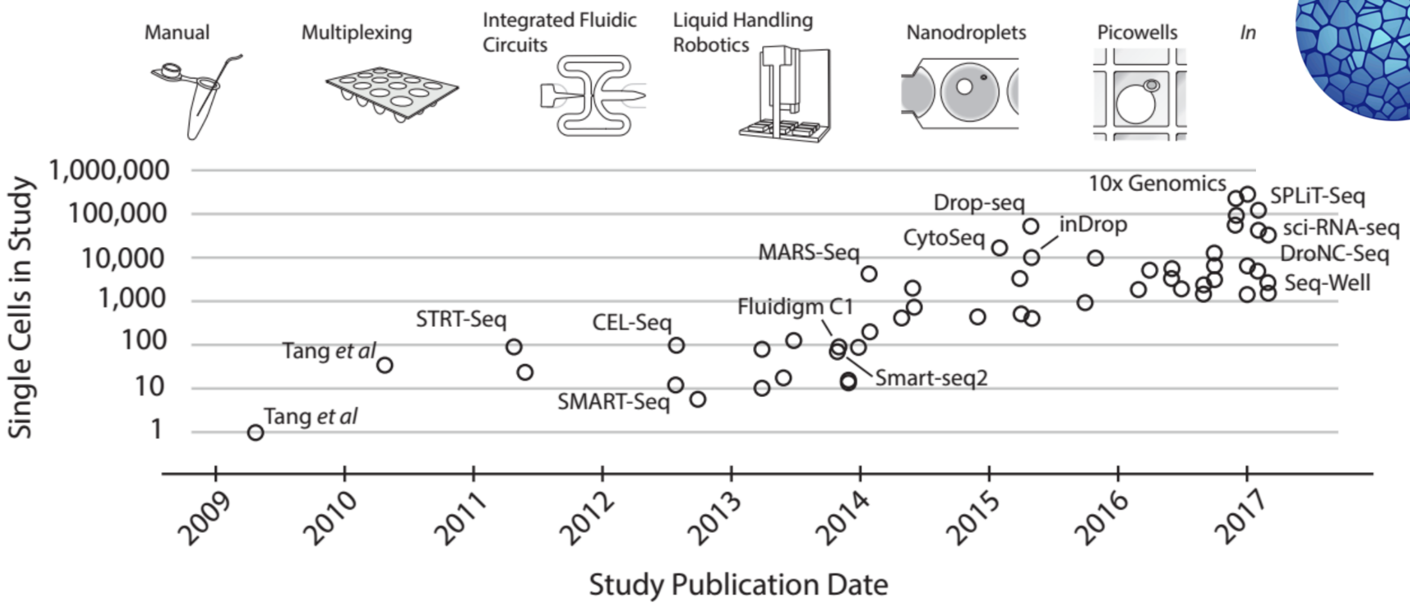
Clustering

PAGODA
RaceID
SC3
SINCERA

scater ecosystem: take advantage of many other R/Bioconductor packages

cf. ExpressionSet, data classes in *Seurat*, *monocle*

Technological developments drive Moore's Law in single-cell transcriptomics



HUMAN CELL ATLAS

Svensson V, Vento-Tormo R, Teichmann SA. Moore's Law in Single Cell Transcriptomics, *arXiv,* 2017. Available: http://arxiv.org/abs/1704.01379

Two key developments…

- *SingleCellExperiment* (Davide Risso)
  - Base class for single-cell data with out-of-memory representations of assay data.
  - Advantages for pkg developers; interoperability

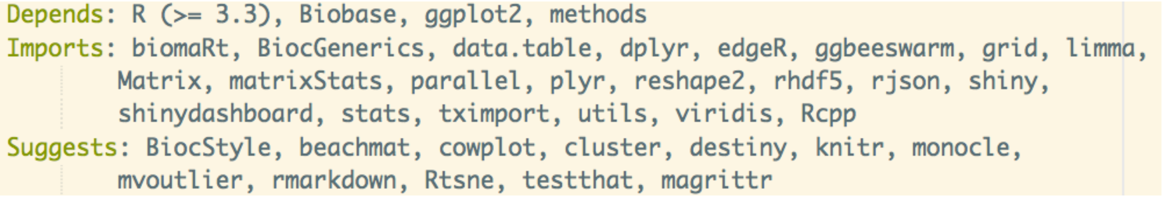
- *Beachmat* (Aaron Lun, Hervé Pages, Mike Smith)
  - C++ API that allows developers to implement computationally intensive algorithms in C++ that can be immediately applied to a wide range of R matrix classes, including simple matrices, sparse matrices from the Matrix package, and HDF5-backed matrices from the HDF5Array package [Lun et al, *bioRxiv*, 2017]

Adoption of SingleCellExperiment and beachmat will be better for users and devels

- *scater* and *scran* will move to SingleCellExperiment and beachmat under the hood for the next release.
- Other developers: you should too!

Acknowledgements: R/Bioconductor pkgs

- **Bioconductor:**
  scater
  scran
  VariantAnnotation
  snpStats
  pcaMethods



```
Depends: R (>= 3.3), Biobase, ggplot2, methods
Imports: biomaRt, BiocGenerics, data.table, dplyr, edgeR, ggbeeswarm, grid, limma,
        Matrix, matrixStats, parallel, plyr, reshape2, rhdf5, rjson, shiny,
        shinydashboard, stats, tximport, utils, viridis, Rcpp
Suggests: BiocStyle, beachmat, cowplot, cluster, destiny, knitr, monocle,
        mvoutlier, rmarkdown, Rtsne, testthat, magrittr
```

- **CRAN:**
  tidyverse
  vcfR
  adegenet
  mclust

Many, many thanks to:

- Bioconductor core team
- Bioconductor developers
- scater users
- All open-source software developers

Acknowledgements

- ***Stegle Lab (EMBL-EBI):***
  Oliver Stegle
  Raghd Rostom (Stegle/Teichmann)
  Anna Cuomo (Stegle/Marioni)
  Marc Jan Bonder

- ***Vallier Lab (Sanger):***
  Shradha Amatya
  Mariya Chhatriwala
  Jose Garcia-Bernardo
  Ludovic Vallier

- ***Scater developers:***
  Aaron Lun, Kieran Campbell, Quin Wills

Sarah Teichmann (Sanger)
John Marioni (EMBL-EBI/CRI)
Helena Kilpinen (UCL/Sanger)
Ian Streeter (EMBL-EBI)

Sanger single cell core facility (SCGCF)
Sanger FACS facility
Sanger sequencing facility

Everyone in HipSci!

Richard Durbin

Dan Gaffney

wellcome trust
Strategic Award

MRC Medical Research Council

Australian Government
National Health and Medical Research Council

HipSci

Get in touch

@davisjmcc

davis@ebi.ac.uk

Workflow with Aaron Lun and John Marioni:

http://bioconductor.org/help/workflows/simpleSingleCell/

http://bioconductor.org/packages/scater/

Single-cell course with Martin Hemberg, Vlad Kiselev, Tallulah Andrews:

https://hemberg-lab.github.io/scRNA.seq.course/

#bioc2017
#RCatLadies
#dataparasites

HipSci

Acknowledgement

WTSI
Richard Durbin
Anja Kolb-Kokocinksi
Andreas Leha
Yasin Memari
Phil Carter
Petr Danecek
Shane McCarthy
Sendu Balasubramaniam
Danielle Walker
Thomas Keane

Daniel Gaffney
Andrew Knights
Natsuhiko Kumasaka
Angela Goncalves

Ludovic Vallier
Filipa Soares
Katarzyna Tilgner
Mariya Chhatriwala
Jose Garcia-Bernardo

CGaP
Chris Kirton
Minal Patel
Rachel Nelson
Alistair White
Sharad Patel
Heather James
Anthi Tsingene
Maria Imaz
Clair Stribling
Chloe Allen
Rizwan Ansari
Leighton Sneade
Lucinda Weston-stiff
Alex Alderton
Jose Garcia-Bernardo
Sarah Harper
Chukwuma Agu

Carol Smee
Ros Cook

EBI
Ewan Birney
Laura Clarke
Ian Streeter
David Richardson
Helen Parkinson

Oliver Stegle
Helena Kilpinen
Marc Jan Bonder
Bogdan Mirauta
Anna Cuomo
Daniel Seaton

Dundee
Angus Lamond
Dalila Bensaddek
Yasmeen Ahmad

KCL
Fiona Watt
Davide Danovi
Annie Kathuria
Nathalie Moens
Oliver Cullley
Darrick Hansen
Natalia Palasz
Andreas Reimer
Ruta Meleckyte

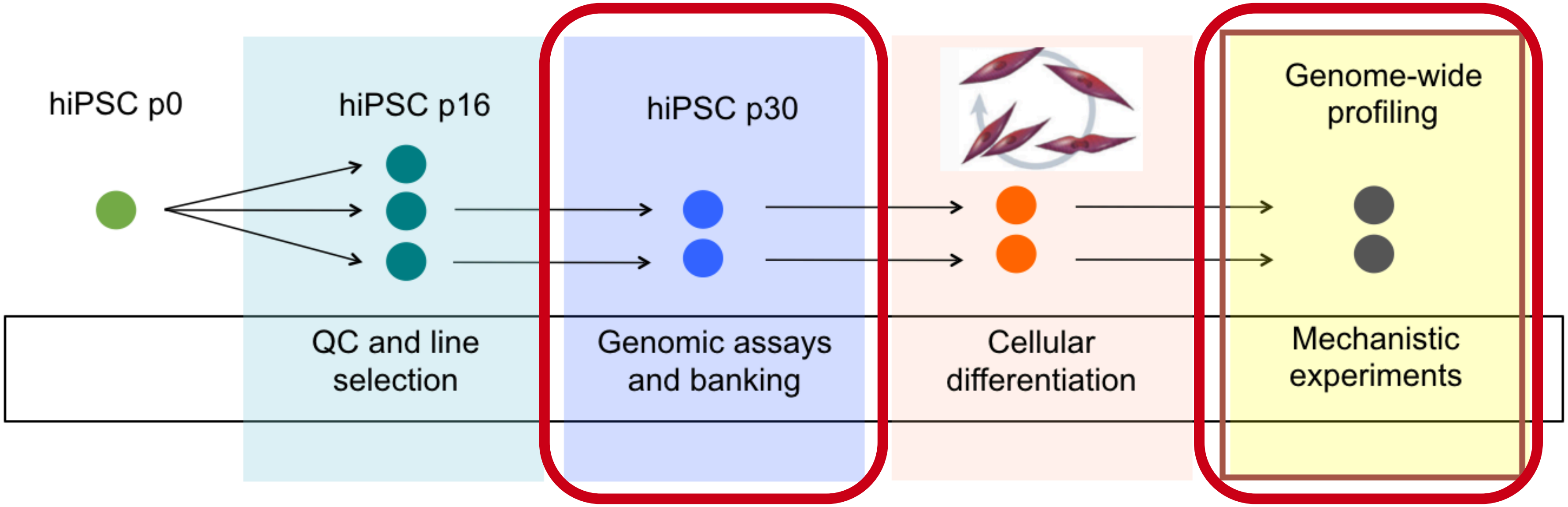
CBR
Willem Ouwehand
Sofie Ashford
Karola Rehnstrom
BRC hIPSCs core facility
Monika Madej
Juned Kadiwala

DNA pipeline teams
Illumina High Throughput pipeline - Emma Gray
Sample Management - Emily Wilkinson
Illumina Bespoke - Richard Rance

wellcome trust
Strategic Award

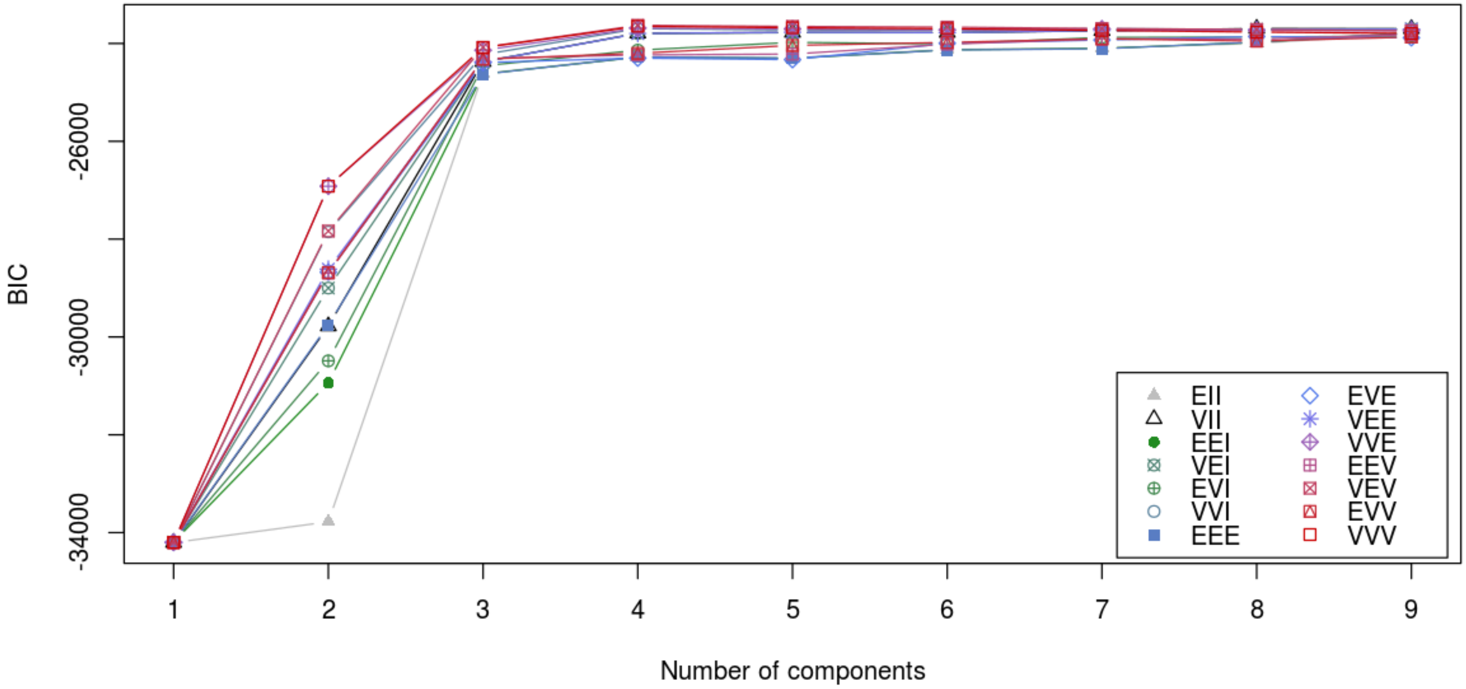
MRC Medical Research Council

Cell differentiation experiments leverage iPSCs to look at downstream effects

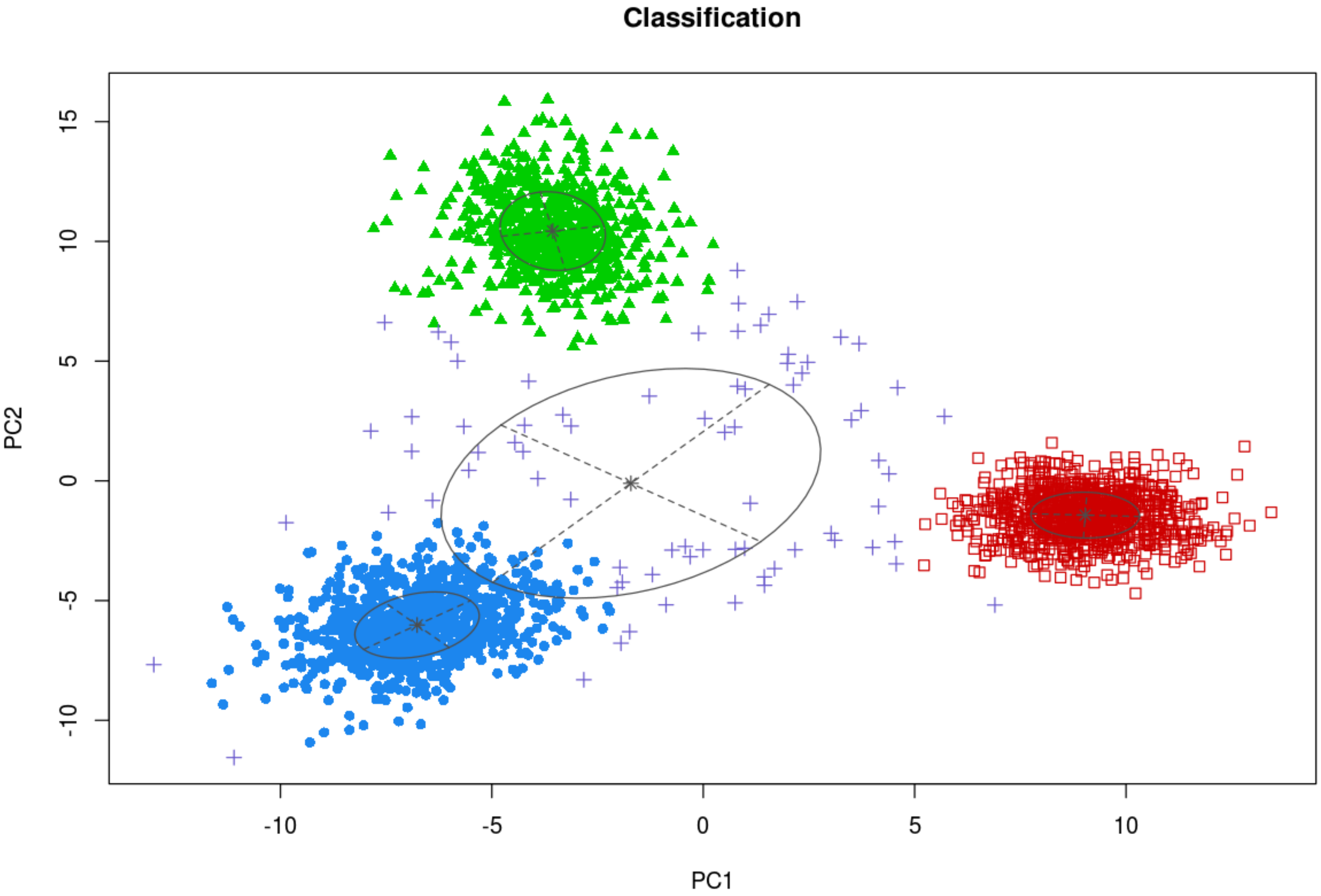


iPSCs provide models for genetic diseases in which we can assay regulatory effects of disease variants in differentiated cells.

mclust BIC selects VEV model with 4 groups



Automated mclust approach yields optimal(?) clustering - no further tweaking looks required


Classification
PC2
PC1