Analysis of RNA-Seq experiments

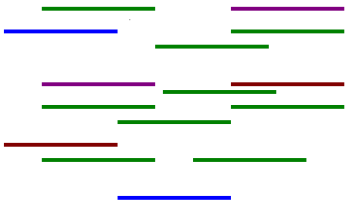
Matthias Lienhard

Seminar RNA bioinformatics

Dec 10th 2014

RNA-Seq Pipeline

Short Reads



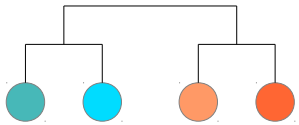
Mapping

Alignment

Quantification

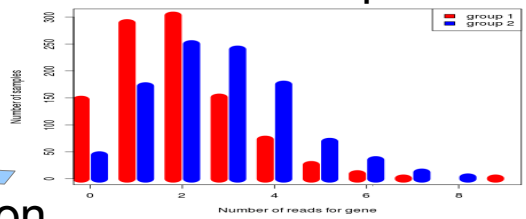
Count Table

	BaP 12h	Ct 12h	...
GENE 1	321	422	12
GENE 2	32	50	20
GENE 3	132466	72921	43223



Exploratory Analysis

Differential Expression



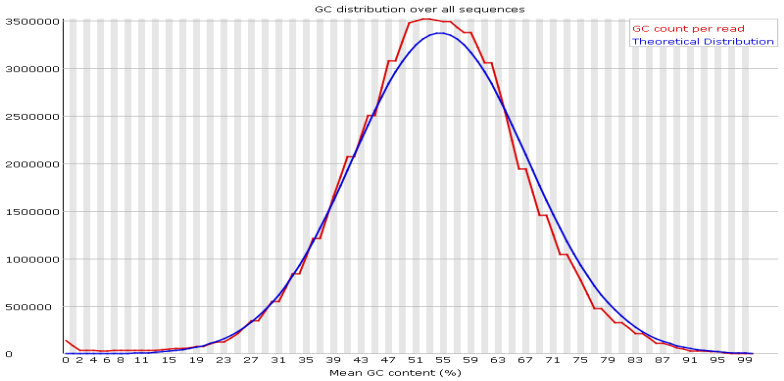
Statistical Test

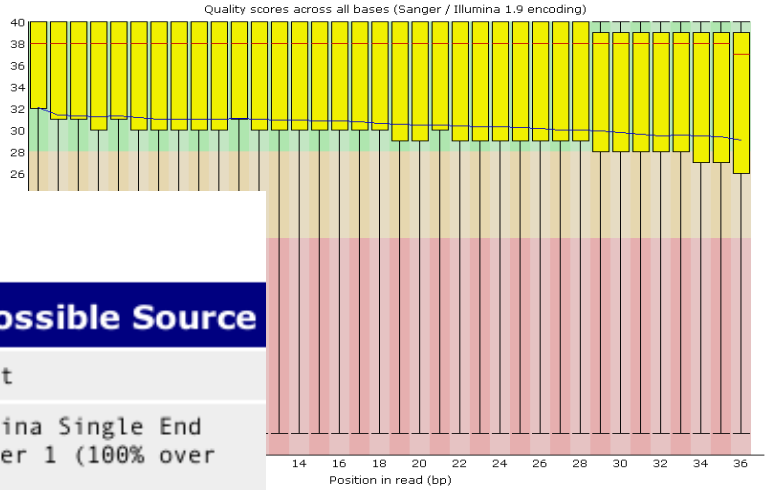
Interpretation

M. Lienhard (2011), “Analysis of RNA-seq Experiments” Master thesis, Freie Universität Berlin.

Quality Control

- Base Call Quality
- PCA artifacts
- Base composition
- Positional base frequency
- Adapter contamination





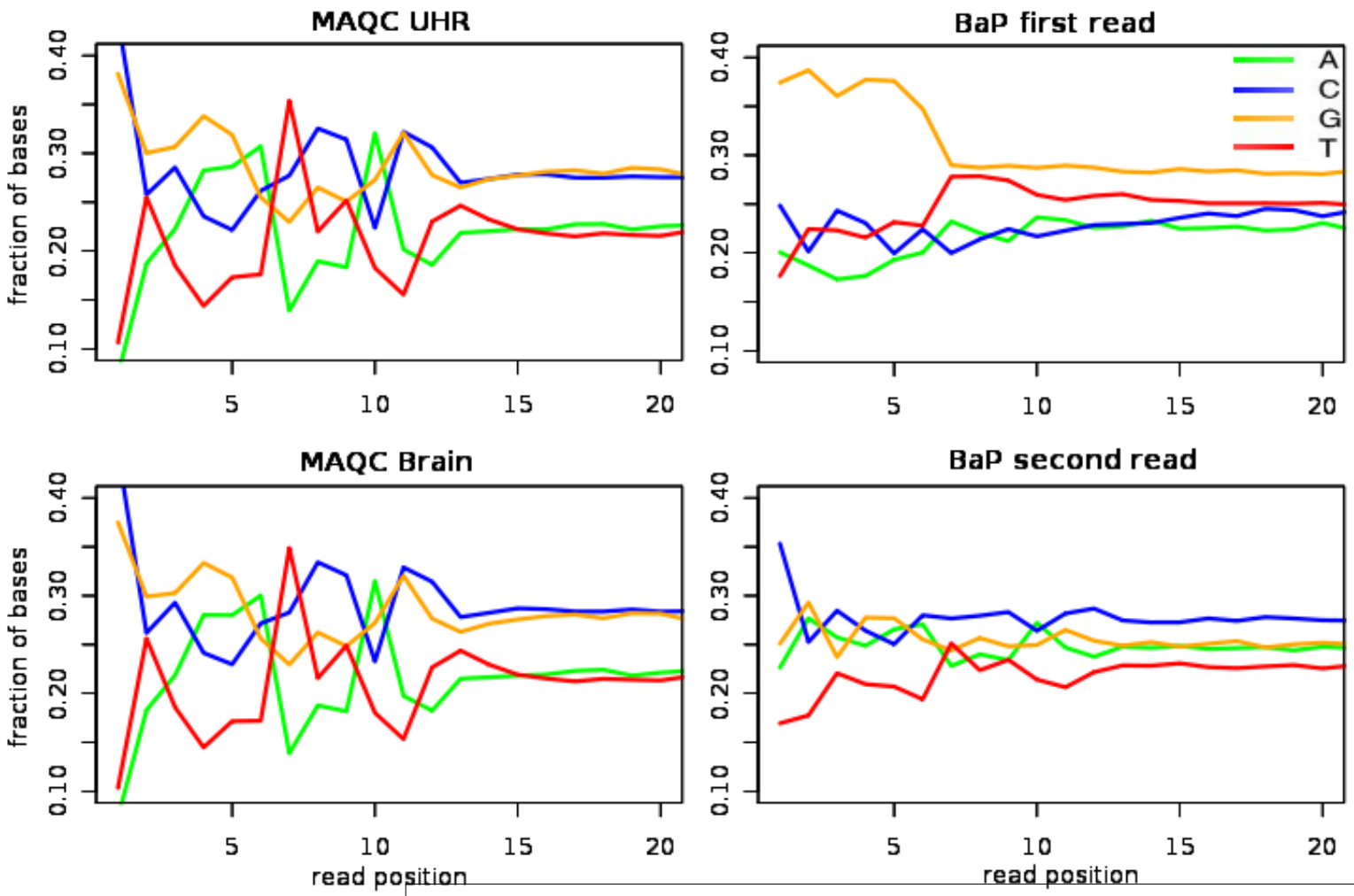
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	119709	0.30700917731983646	No Hit
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAA	113700	0.29159832144003717	Illumina Single End Adapter 1 (100% over 33bp)

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

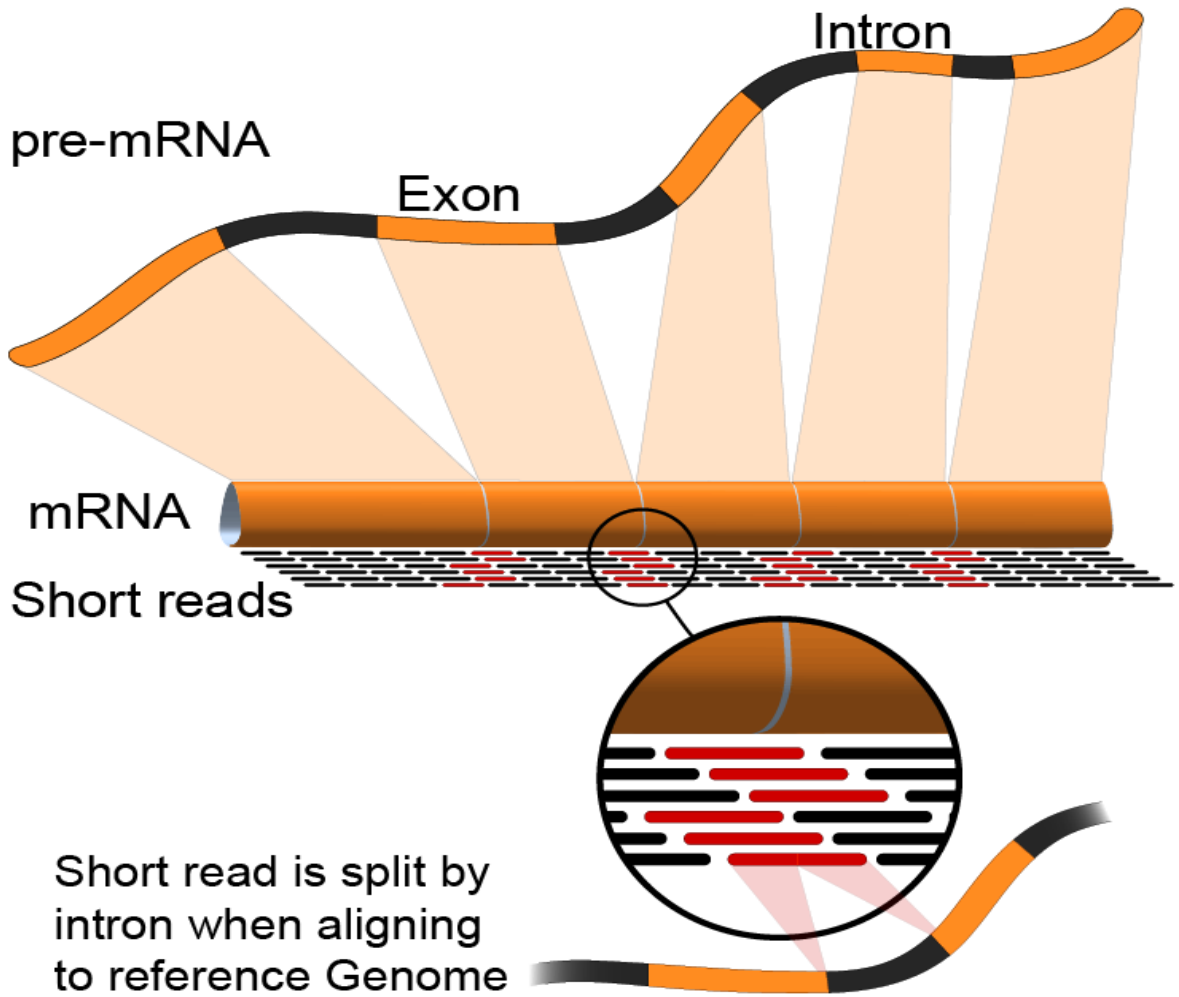
Quality Control

RNA seq has strange base-position patterns



KD Hansen et al. (2010): Biases in Illumina transcriptome sequencing caused by random hexamer priming, NAR

Read Mapping



Tophat

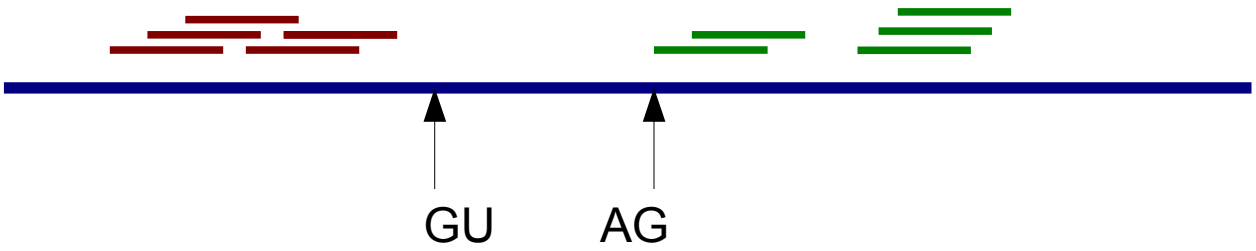
1.Genome alignment

Trapnell, C et al. (2009), TopHat: discovering splice junctions with RNA-Seq; Bioinformatics

Tophat

1. Genome alignment

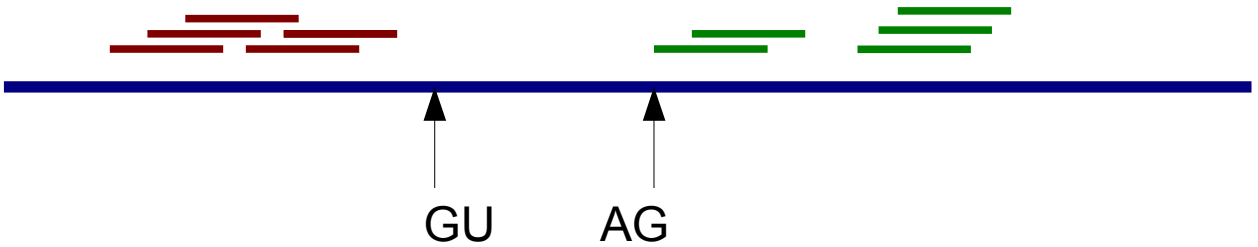
2. look for canonical splice sites



Trapnell, C et al. (2009), TopHat: discovering splice junctions with RNA-Seq; Bioinformatics

Tophat

1. Genome alignment



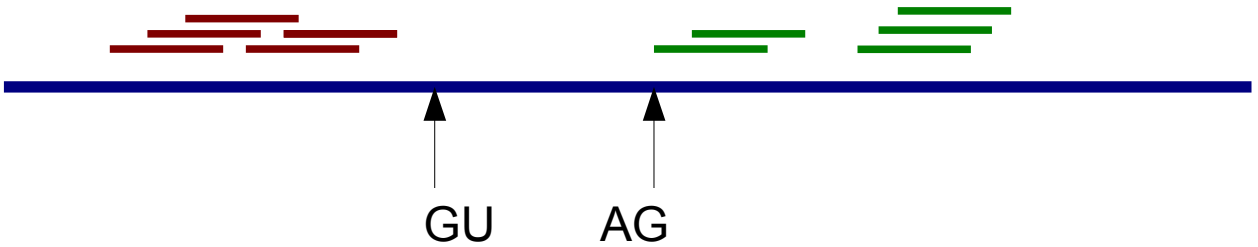
2. look for canonical splice sites

3. add known splice sites from database (eg refseq)

Trapnell, C et al. (2009), TopHat: discovering splice junctions with RNA-Seq; Bioinformatics

Tophat

1. Genome alignment



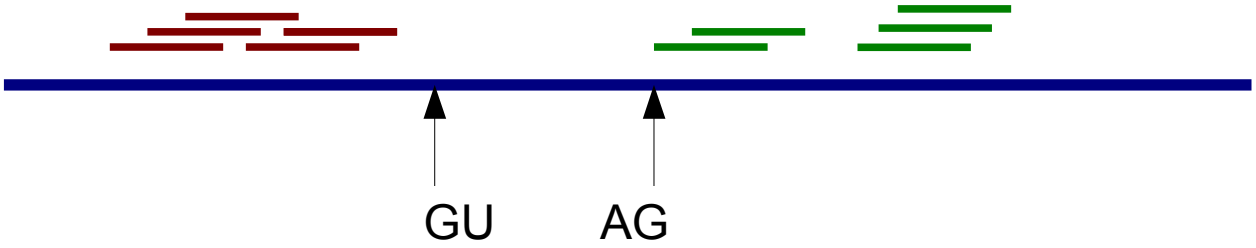
2. look for canonical splice sites

3. add known splice sites from database (eg refseq)

Trapnell, C et al. (2009), TopHat: discovering splice junctions with RNA-Seq; Bioinformatics

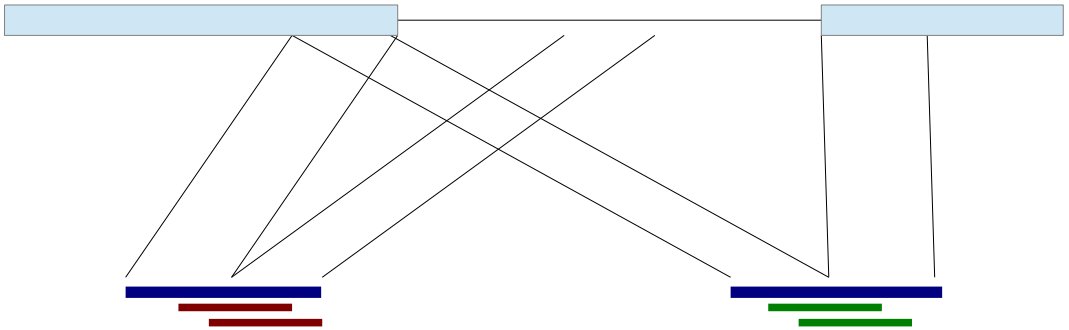
Tophat

1. Genome alignment



2. look for canonical splice sites

3. add known splice sites from database (eg refseq)

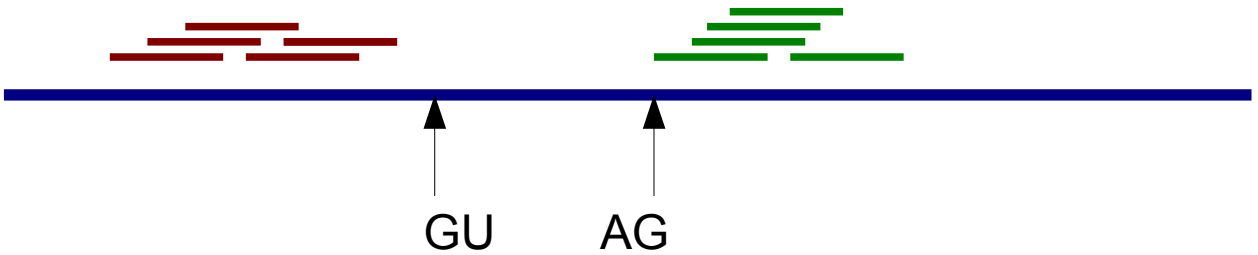


4. assemble sequences at splice sites and map reads

Trapnell, C et al. (2009), TopHat: discovering splice junctions with RNA-Seq; Bioinformatics

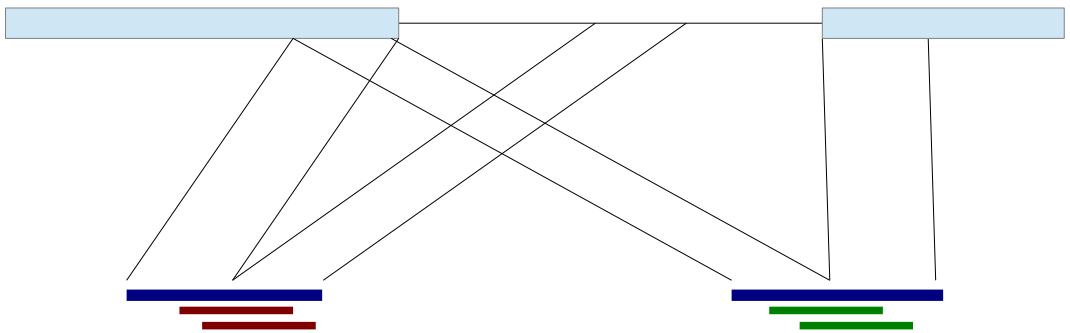
Tophat

1. Genome alignment



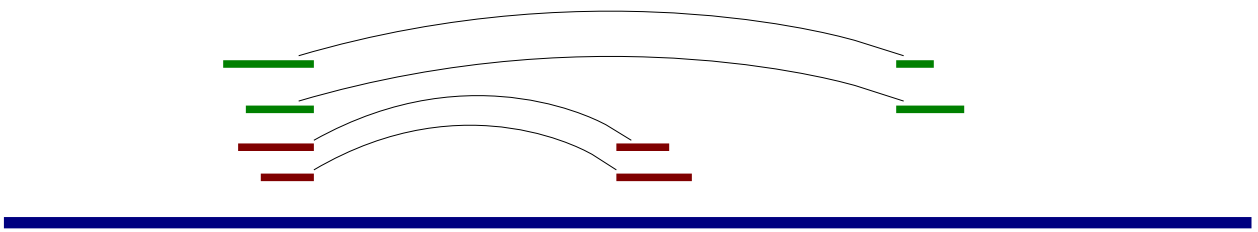
2. look for canonical splice sites

3. add known splice sites from database (eg refseq)



4. assemble sequences at splice sites and map reads

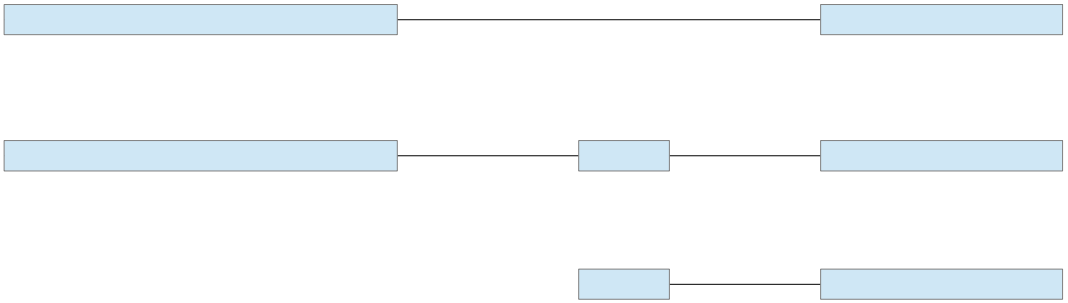
5. Map reads to genome



Trapnell, C et al. (2009), TopHat: discovering splice junctions with RNA-Seq; Bioinformatics

Quantification

- Gene / Isoform / Exon Level
- Reads on boundaries / intronic reads
  - htseq-count strategies
- ambiguous reads



S Anders *et al.*(2014): HTSeq — A Python framework to work with high-throughput sequencing data. Bioinformatics

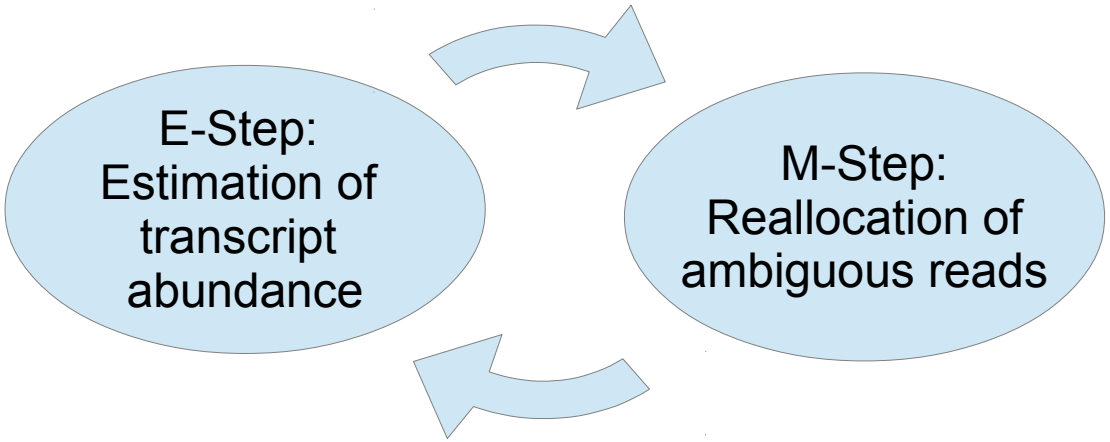
Quantification

Ambiguous reads:

~ 20% gene-level

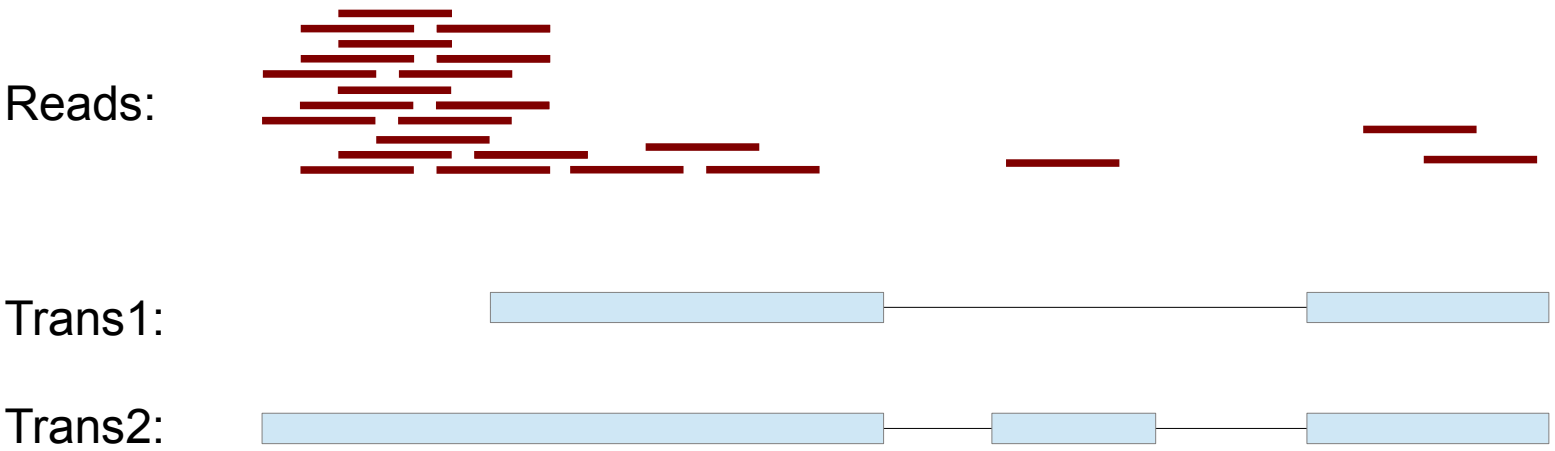
~ 30-90% on isoform level

RSEM multiread assignment EM algorithm:



B Li et al, (2011): RSEM: accurate transcript quantification from RNA-Seq data. BMC Bioinformatics

Isoform Quantification

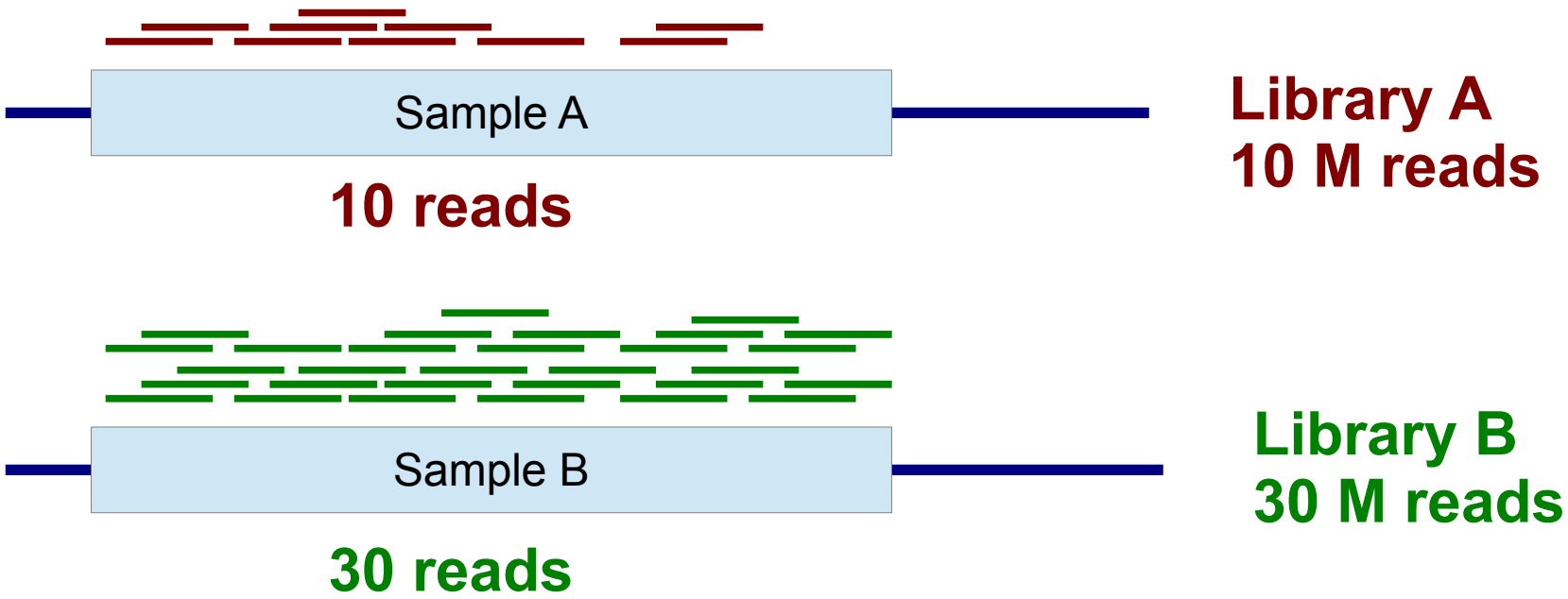


- Problems:
  - Uneven read distribution
  - Imperfect transcript annotation

  → be careful


Sample A
10 reads
Sample B
30 reads

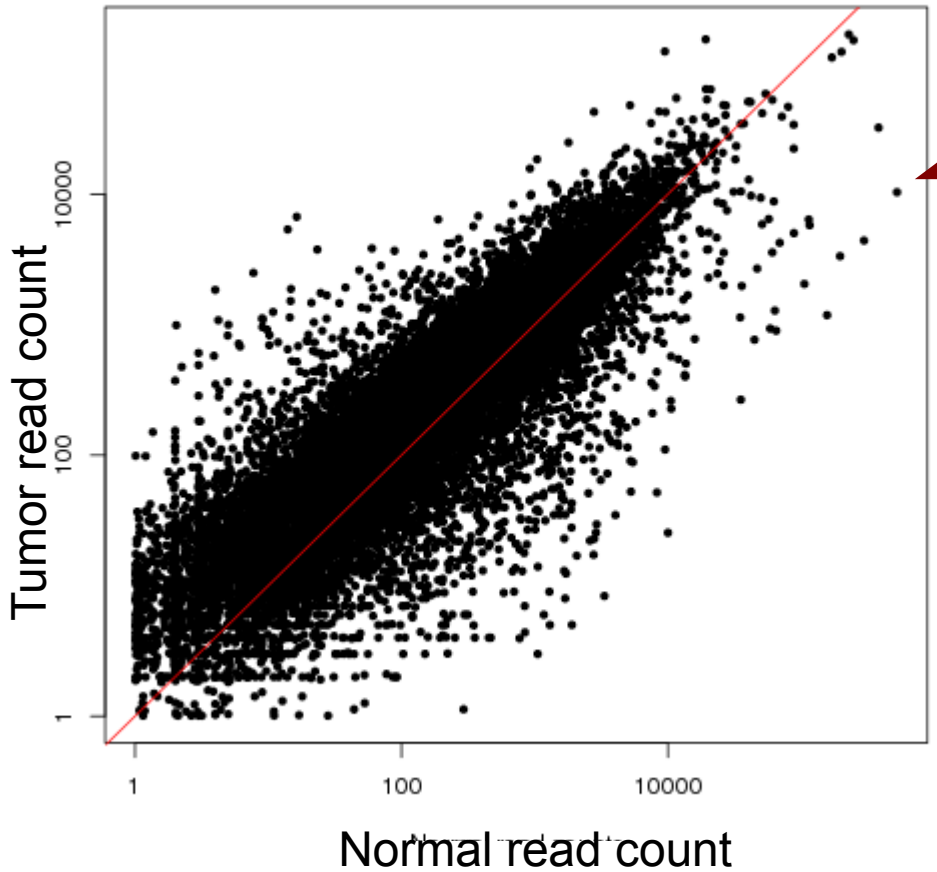
Library Size Normalization



→ Scale Samples to common size

Scaling Library Size Normalization

- Divide by total number
  - Highly expressed genes predominate factor



MYH11
Tumor: 10.371 reads
Normal: 523.926 reads

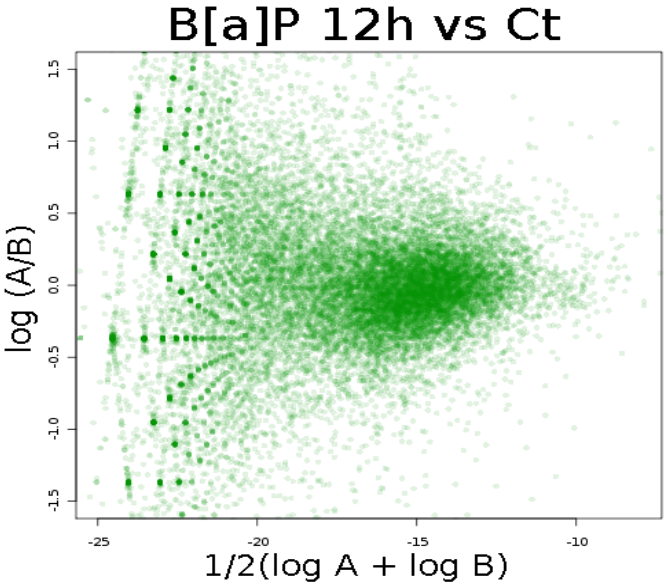
Scaling Library Size Normalization

- Divide by total number
  - Highly expressed genes predominate factor
- Upper Quantile Normalization (DESeq)
  - Median is perturbed by 1 and 2 read genes (noise)
  - 75% quartile usually works

JH Bullard (2010): Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. BMC Bioinformatics

Scaling Library Size Normalization

- Divide by total number
  - Highly expressed genes predominate factor
- Upper Quantile Normalization (DESeq)
  - Median is perturbed by 1 and 2 read genes
  - 75% quartile usually works
- TMM: trimmed mean of M Values (edgeR)
  - Idea: center the “main dot cloud” in M vs A plot

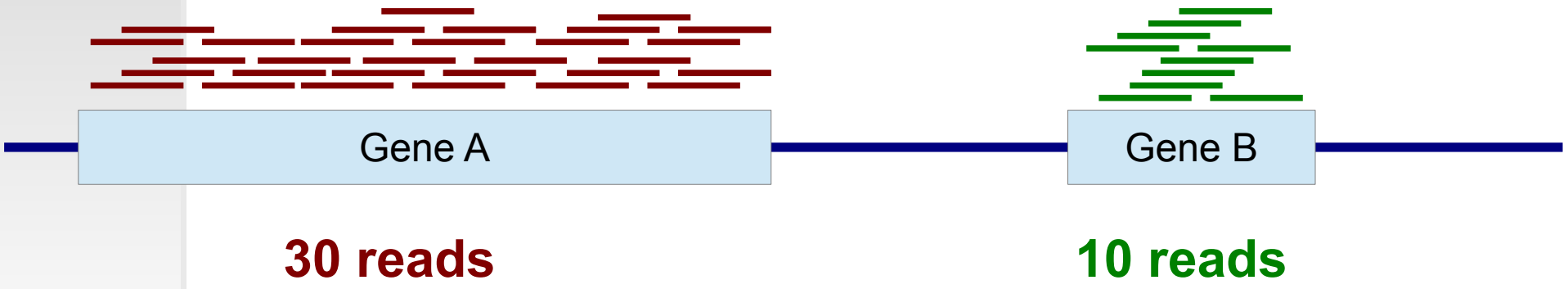


MD Robinson et al.(2010):A scaling normalization method for differential expression analysis of RNA-seq data; Genome Biology

Read length normalization

- Read length normalization
  - Compare different genes

    → Reads per kilobase of RNA per million (rpkm)



Exploratory Analysis

First Check of Quality and Hypotheses:

- How related are the samples?
- Are there distinct groups?
- Are the samples assigned correctly?
- Do we have contamination?
- Do technical differences have effects?
- Do other factors (sex, age, ...) have major influence?

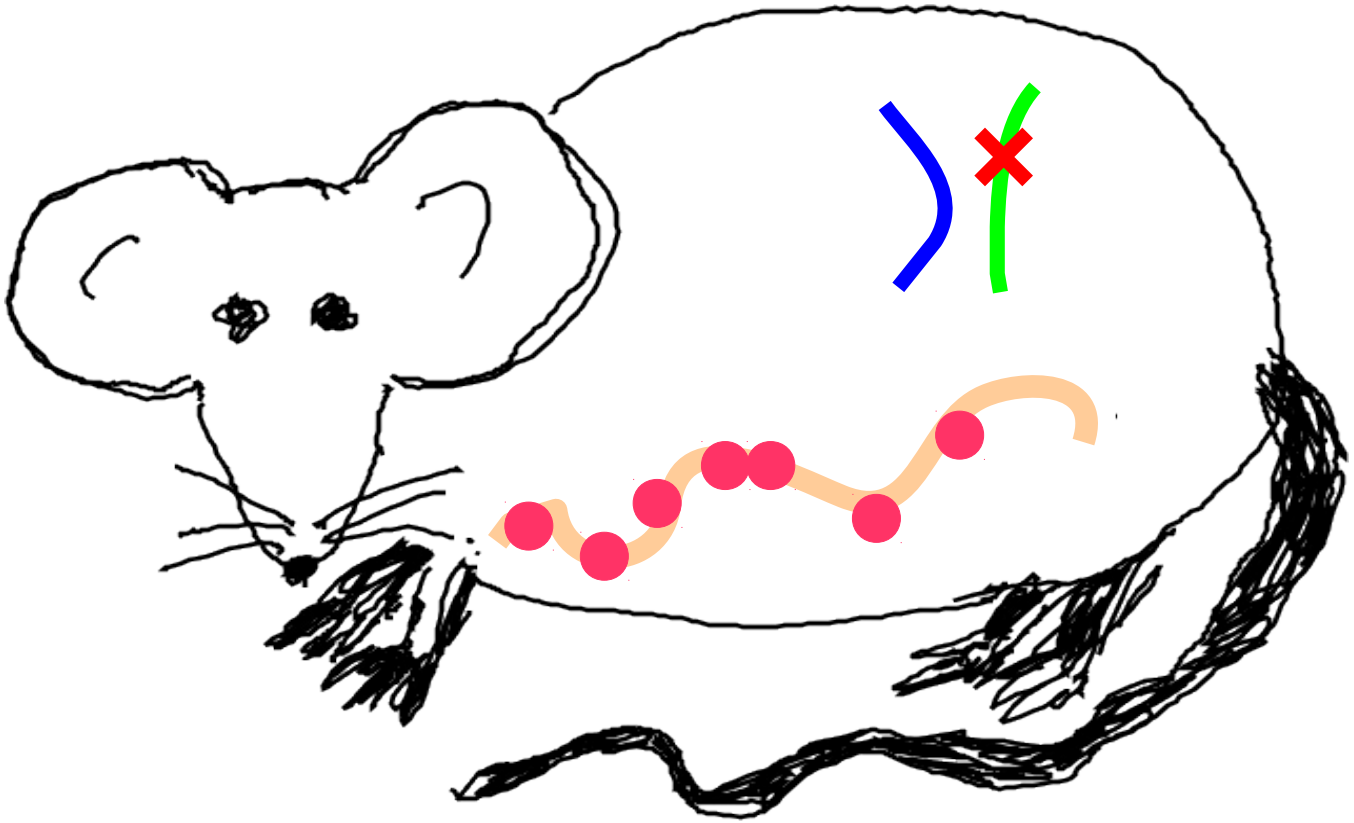
Hierarchical Clustering

Two choices:

- Clustering function
  - Single, complete, avg (UPGMA), Ward, centroid ...
- Distance function
  - Euclidean
  - Correlation based: $d(x,y)=1-cor(x,y)$

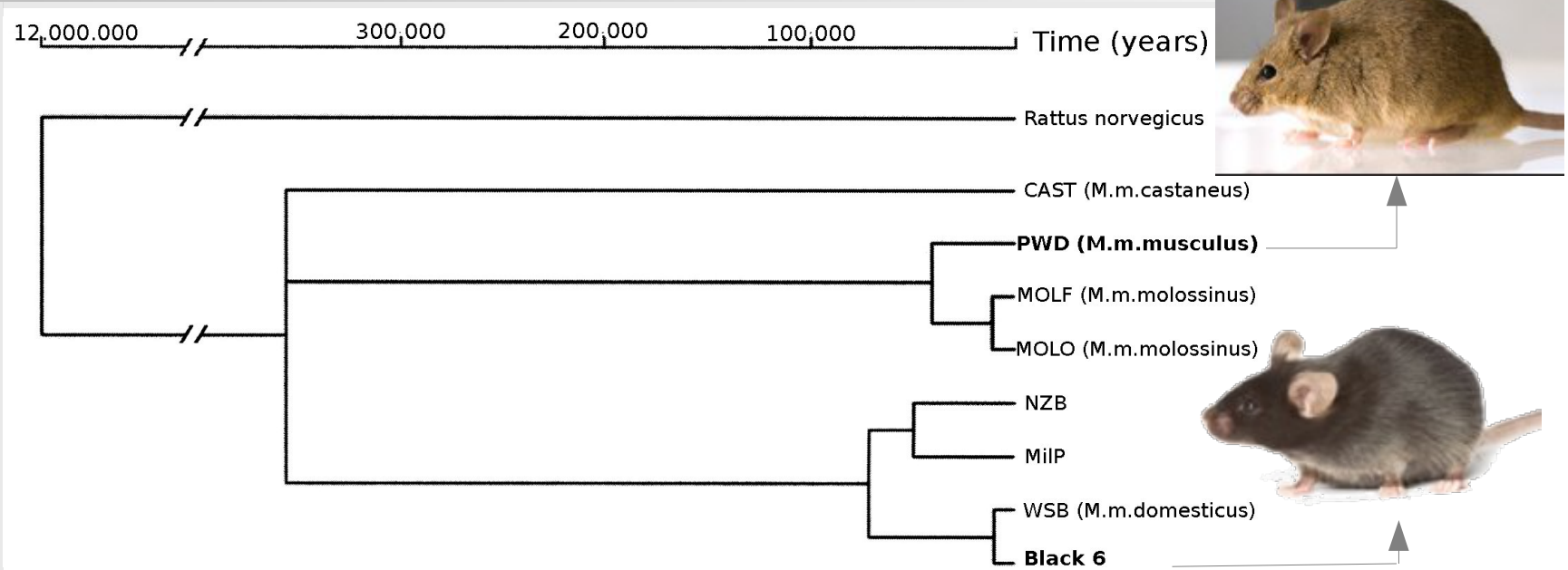
Example: Colon Cancer Mouse Model

APC $^{min/+}$ Mouse



Multiple **I**ntestinal **N**eoplasia

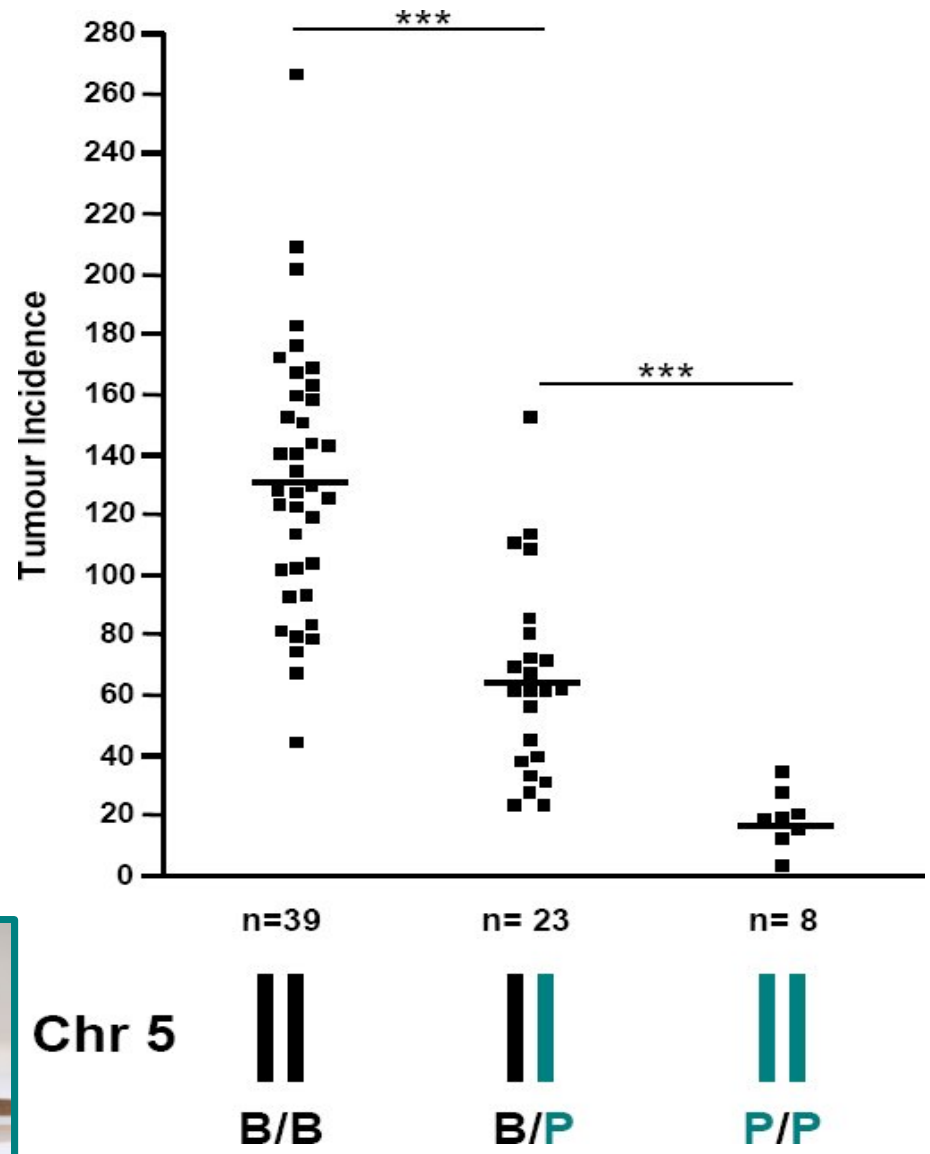
Genetic variation



- Genetically divergent
- High degree of sequence polymorphisms

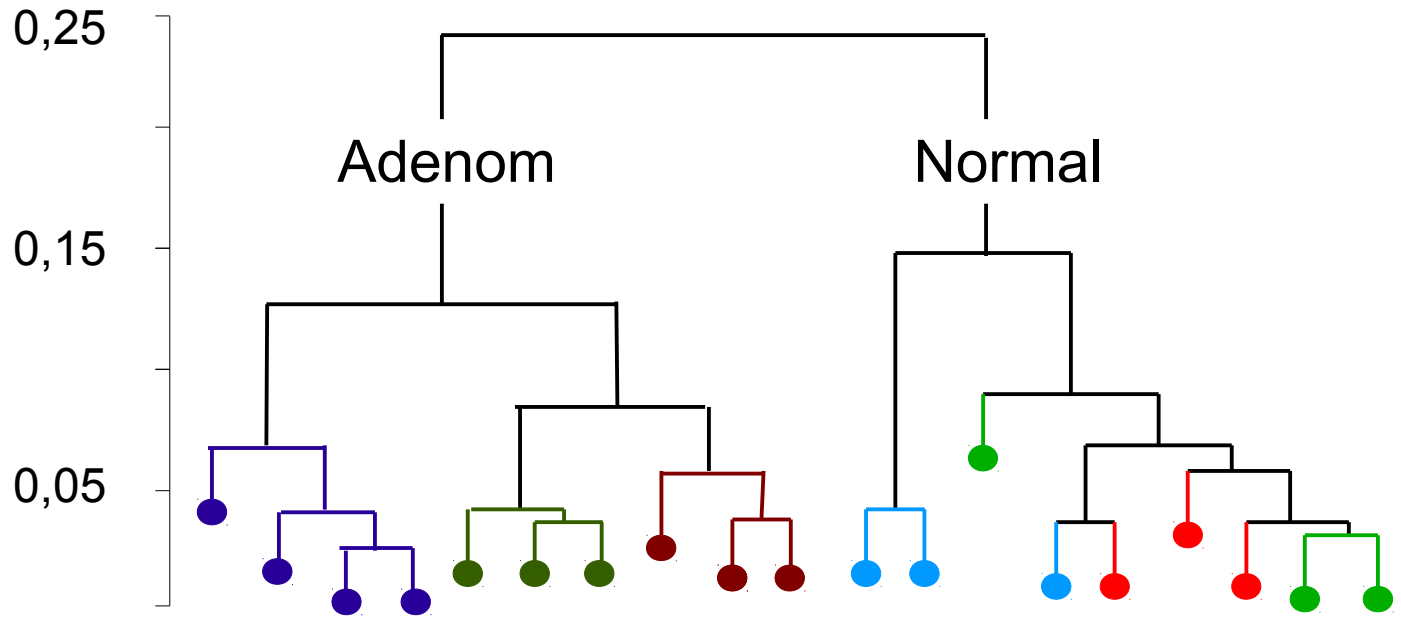
Goios *et al* (2007): mtDNA phylogeny and evolution of laboratory mouse strains, Genome Res.

Genetic Background Matters Chromosome Substitution Stains:



RNA-Seq Clustering

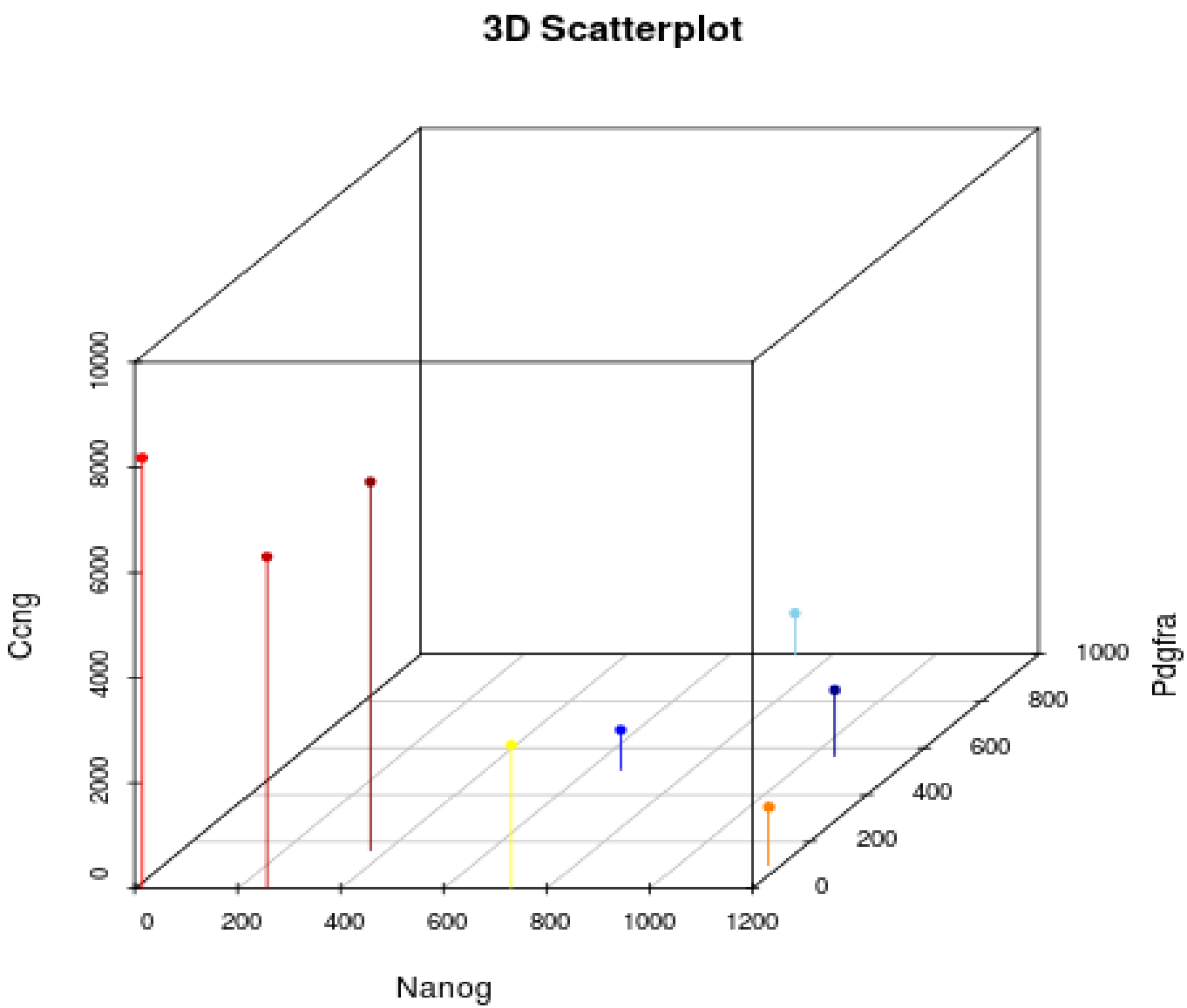
Cluster Dendrogramm w/o chr 5



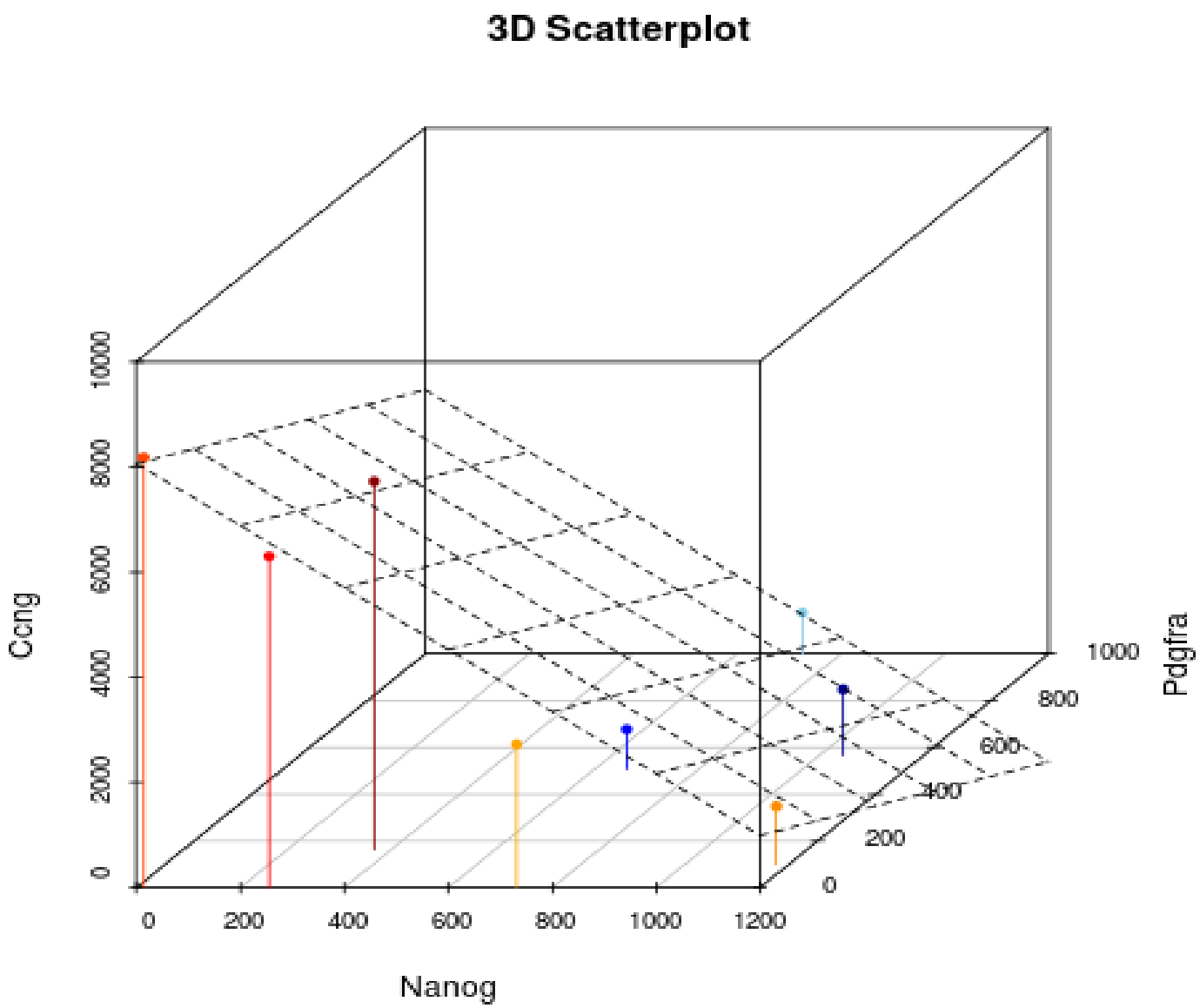
- Black 6
- C5 PWD/B6
- C5 PWD

Data: Tumor Modifiers (unpublished)

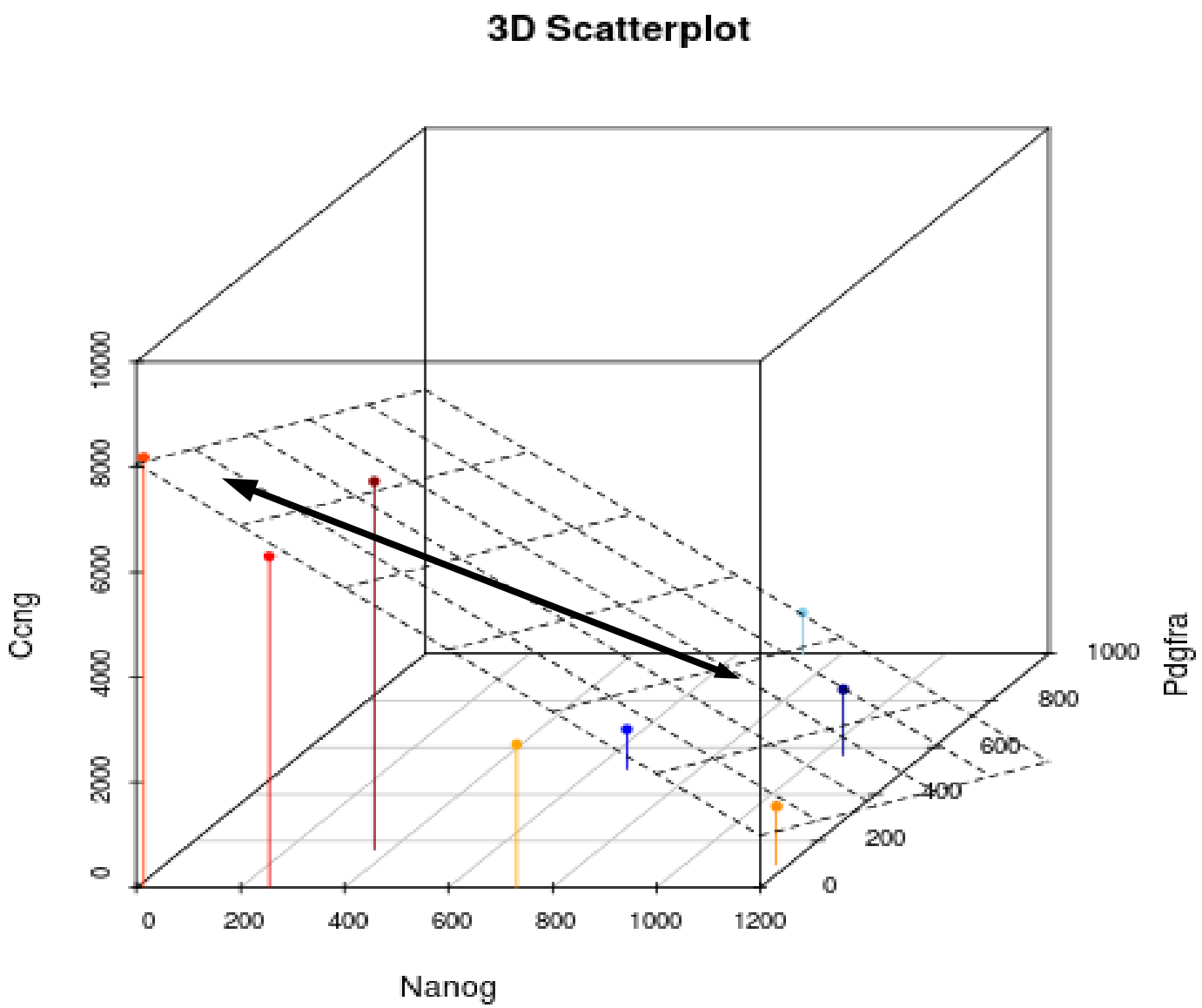
PCA


3D Scatterplot
Ccng
0
2000
4000
6000
8000
10000
Nanog
0
200
400
600
800
1000
1200
Pdgfra
0
200
400
600
800
1000

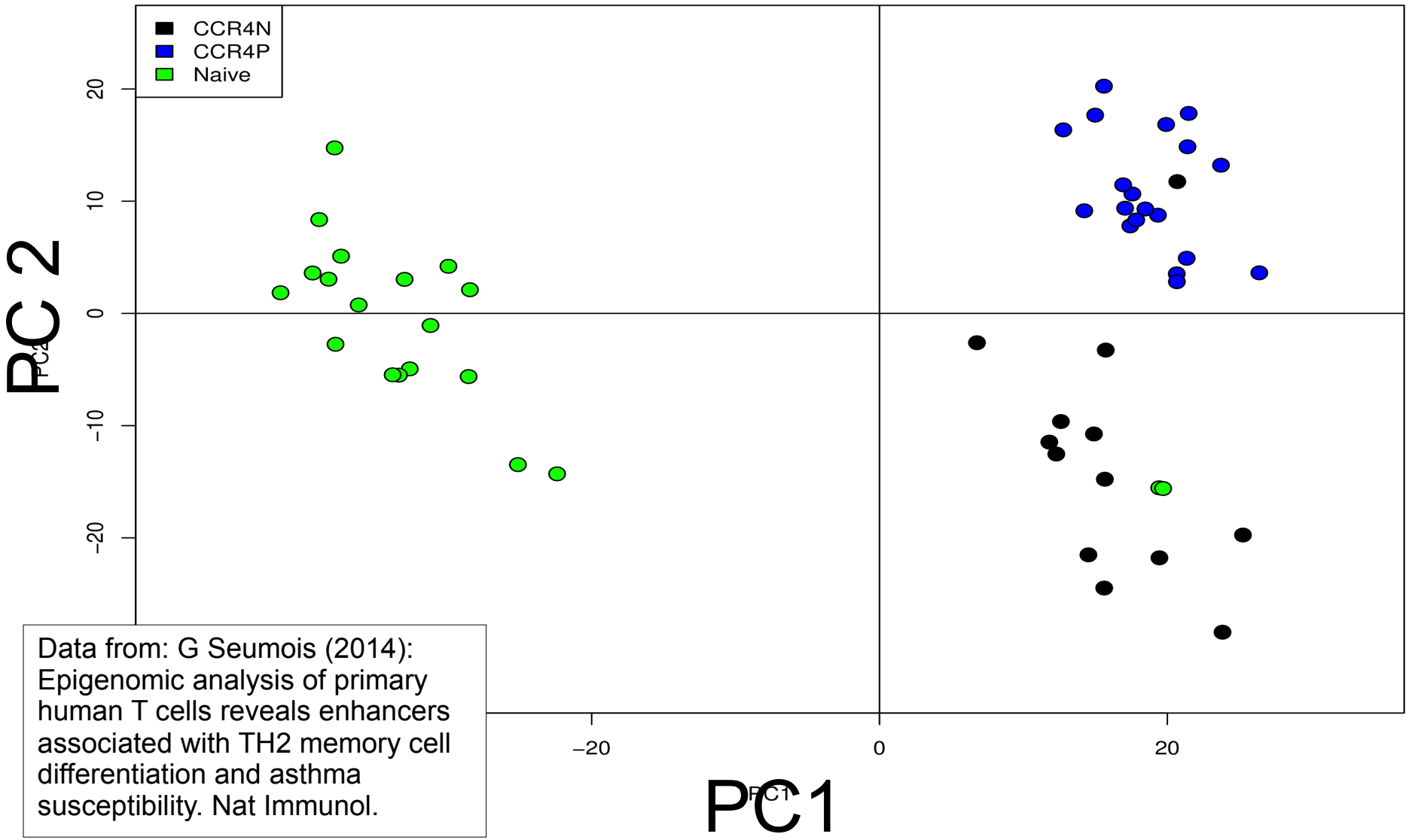
PCA


3D Scatterplot
Ccng
0
2000
4000
6000
8000
10000
Nanog
0
200
400
600
800
1000
1200
Pdgfra
0
200
400
600
800
1000

PCA

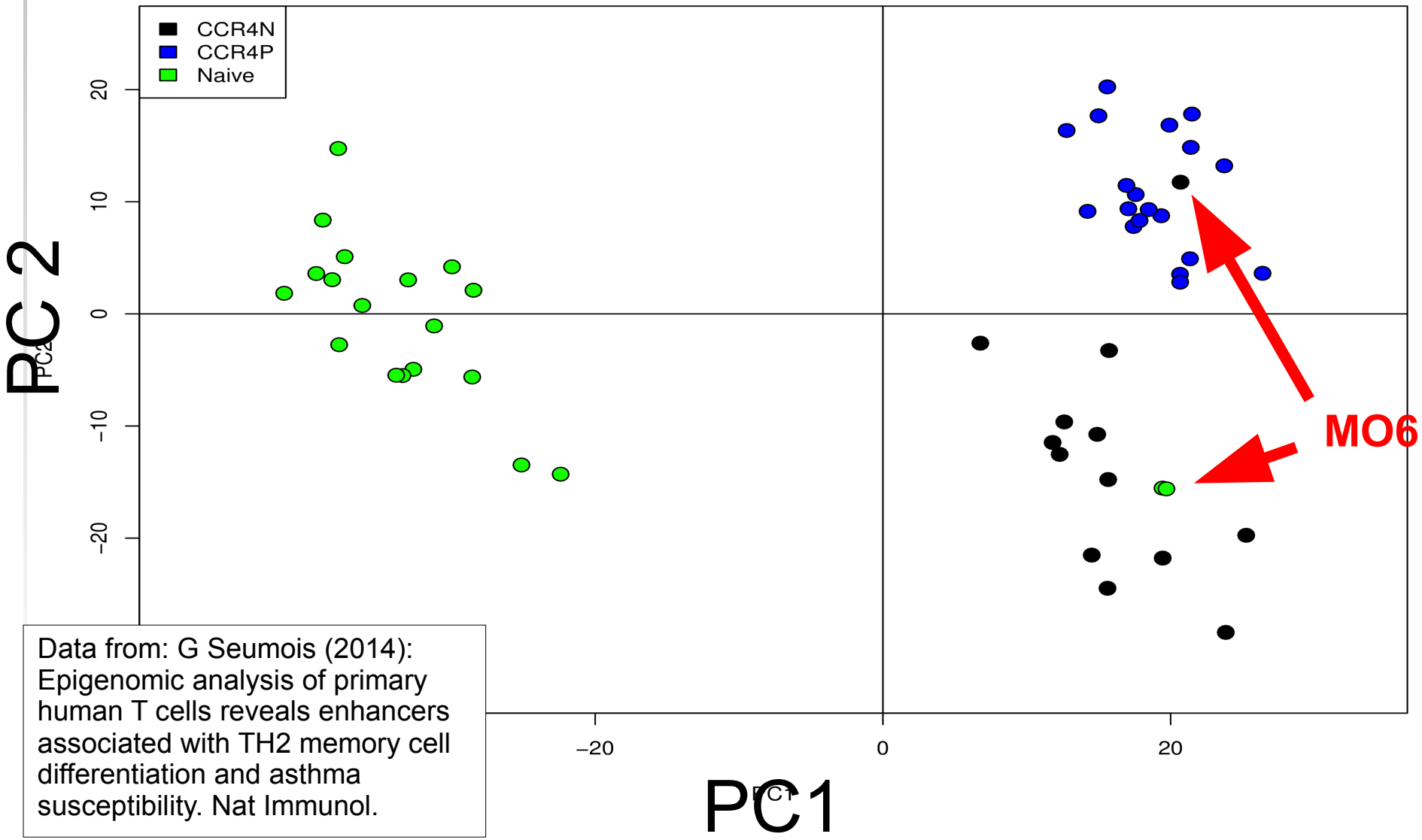

3D Scatterplot
Ccng
0
2000
4000
6000
8000
10000
Nanog
0
200
400
600
800
1000
1200
Pdgfra
0
200
400
600
800
1000

PCA



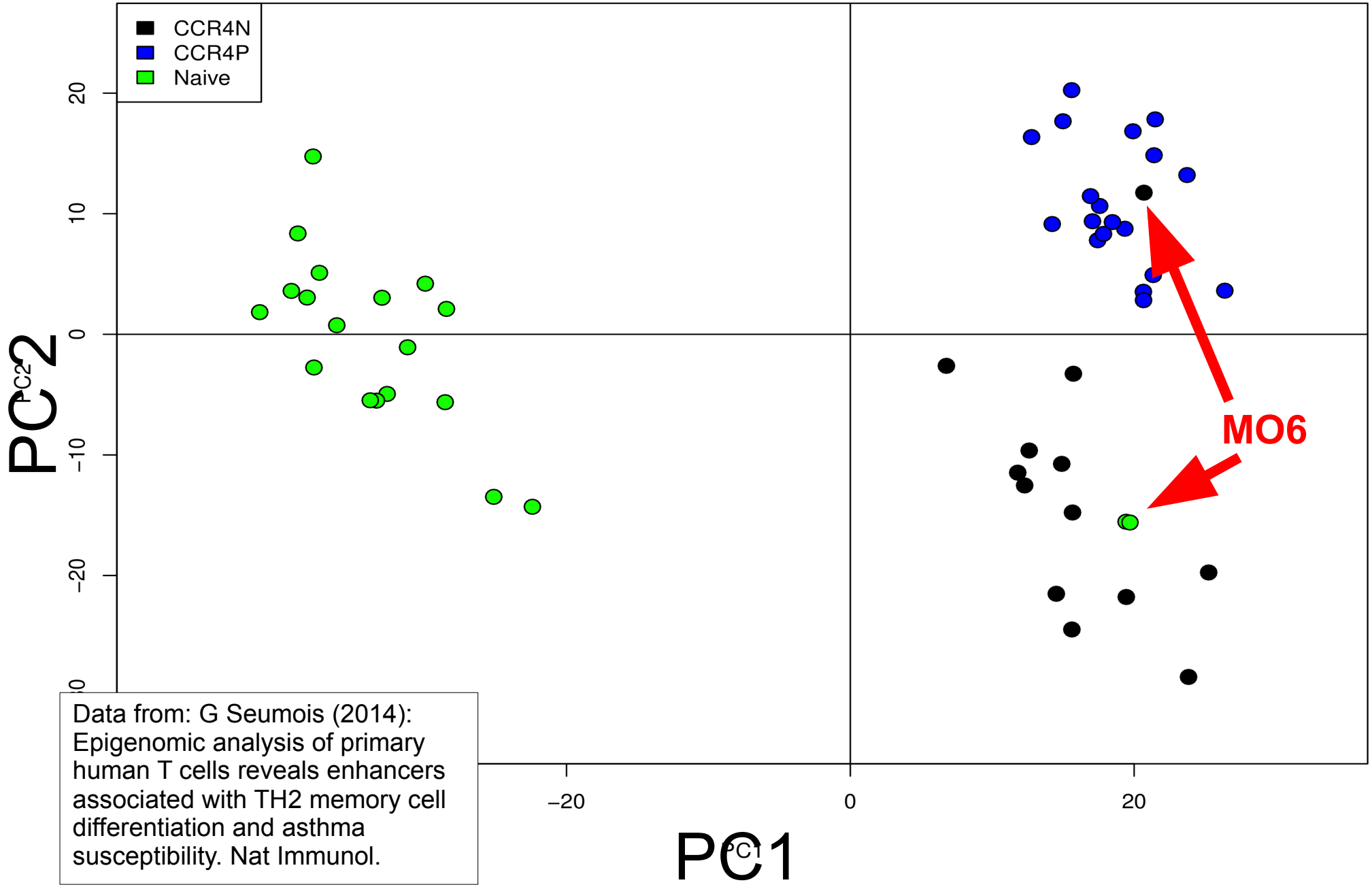
Data from: G Seumois (2014): Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. Nat Immunol.

PCA



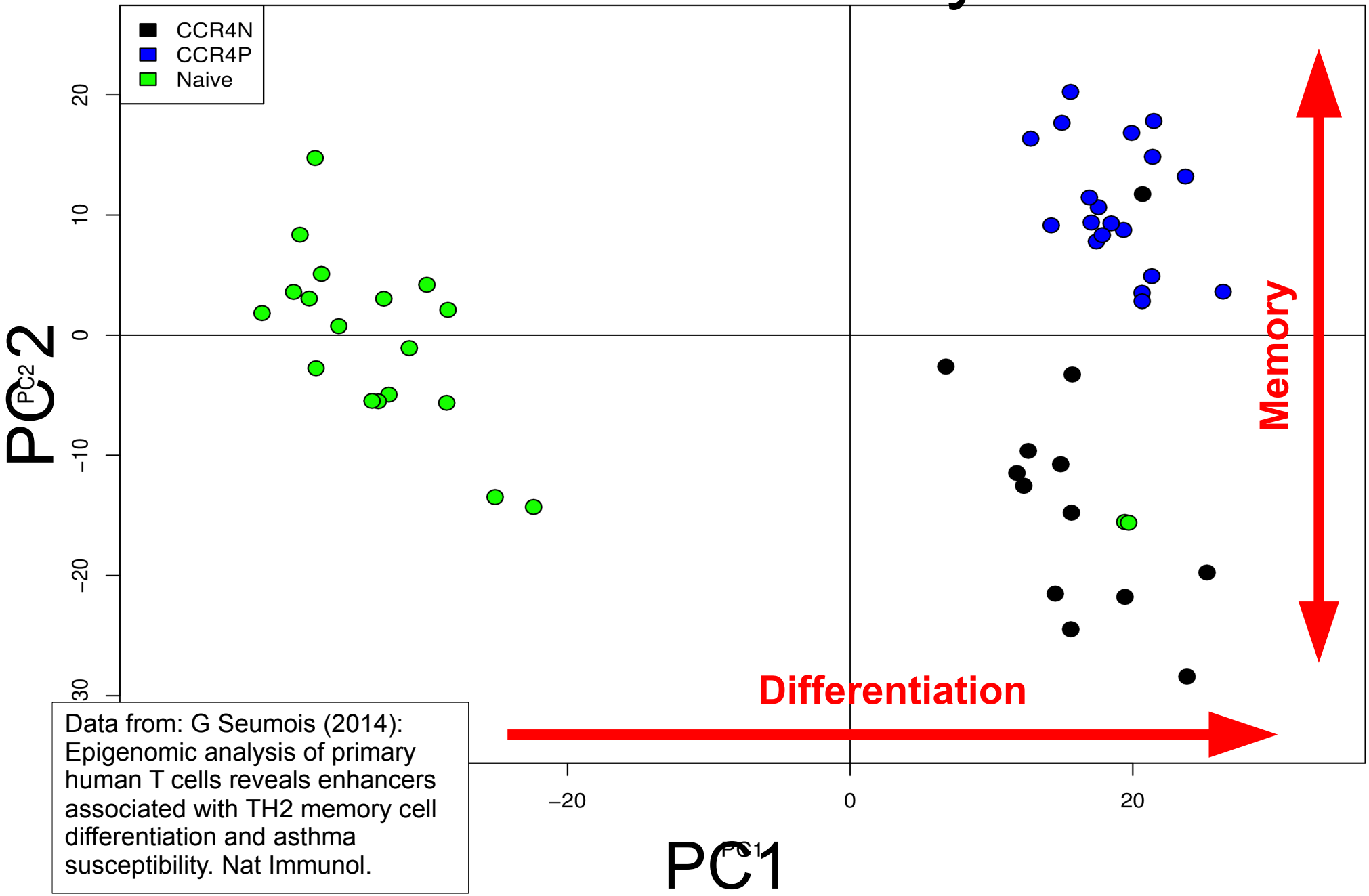
Data from: G Seumois (2014): Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. Nat Immunol.

PCA – sample verification



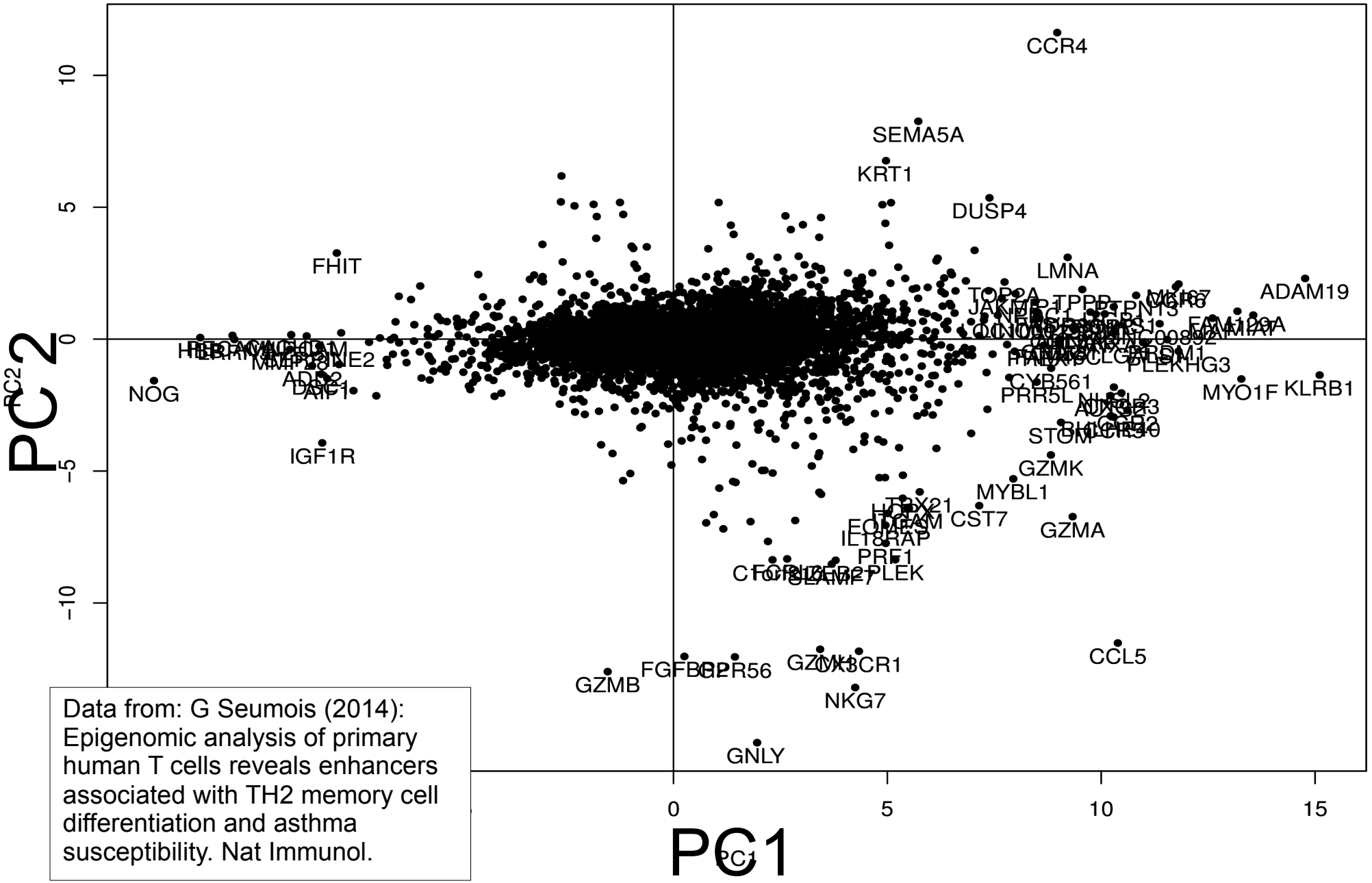
Data from: G Seumois (2014): Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. Nat Immunol.

PCA – factor analysis



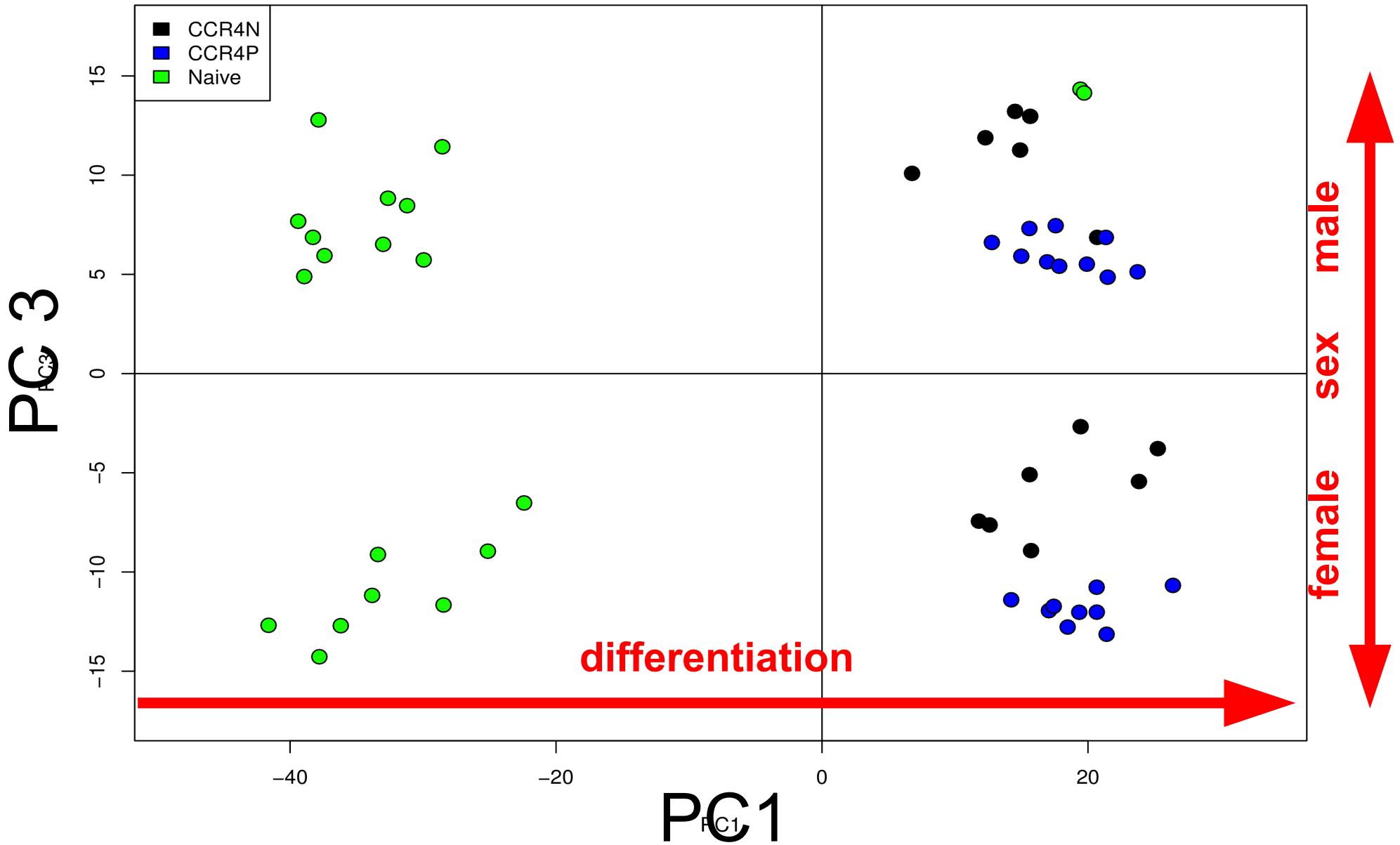
Data from: G Seumois (2014): Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. Nat Immunol.

Genes Contributing to PCs

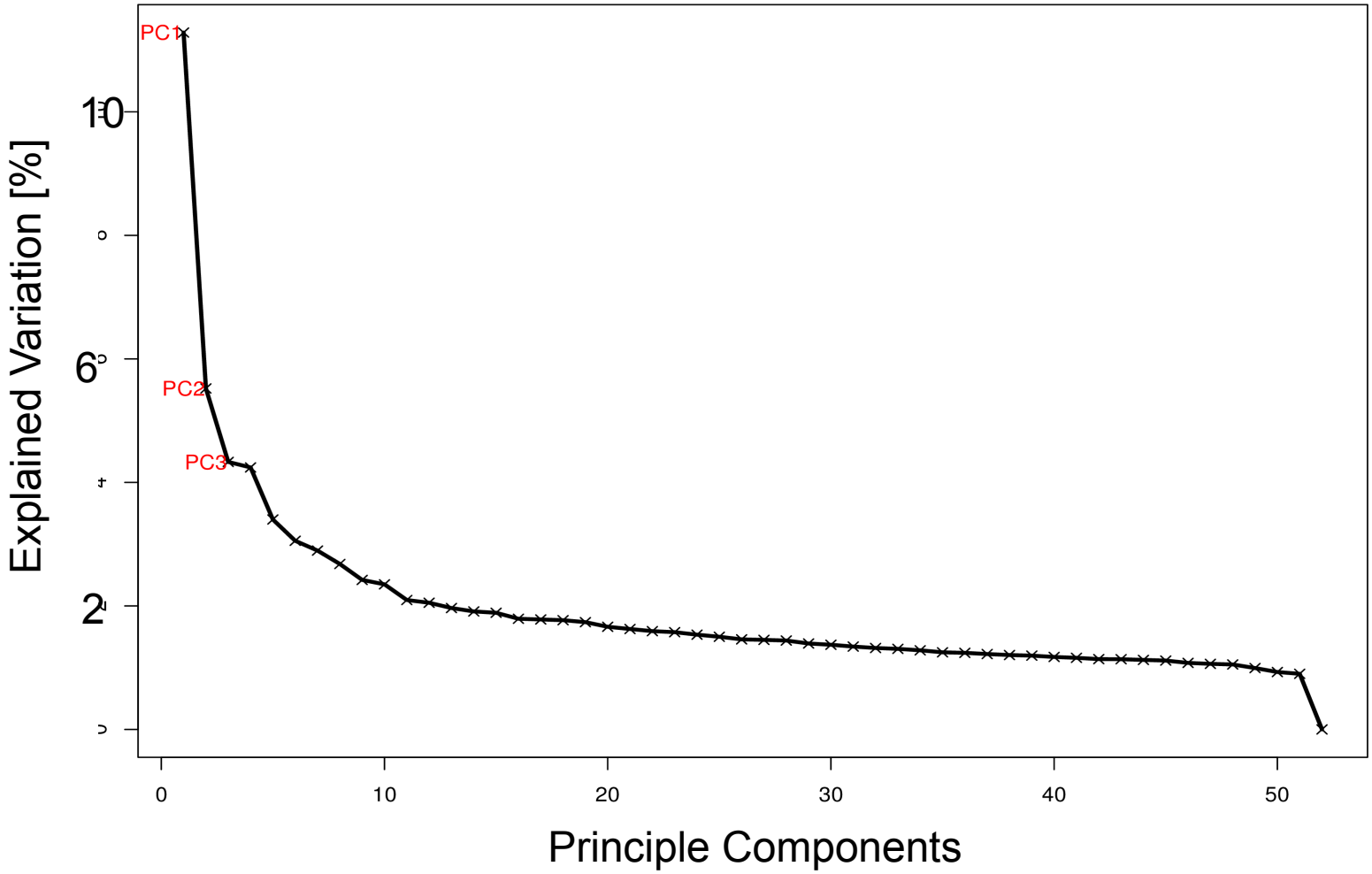


Data from: G Seumois (2014): Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. Nat Immunol.

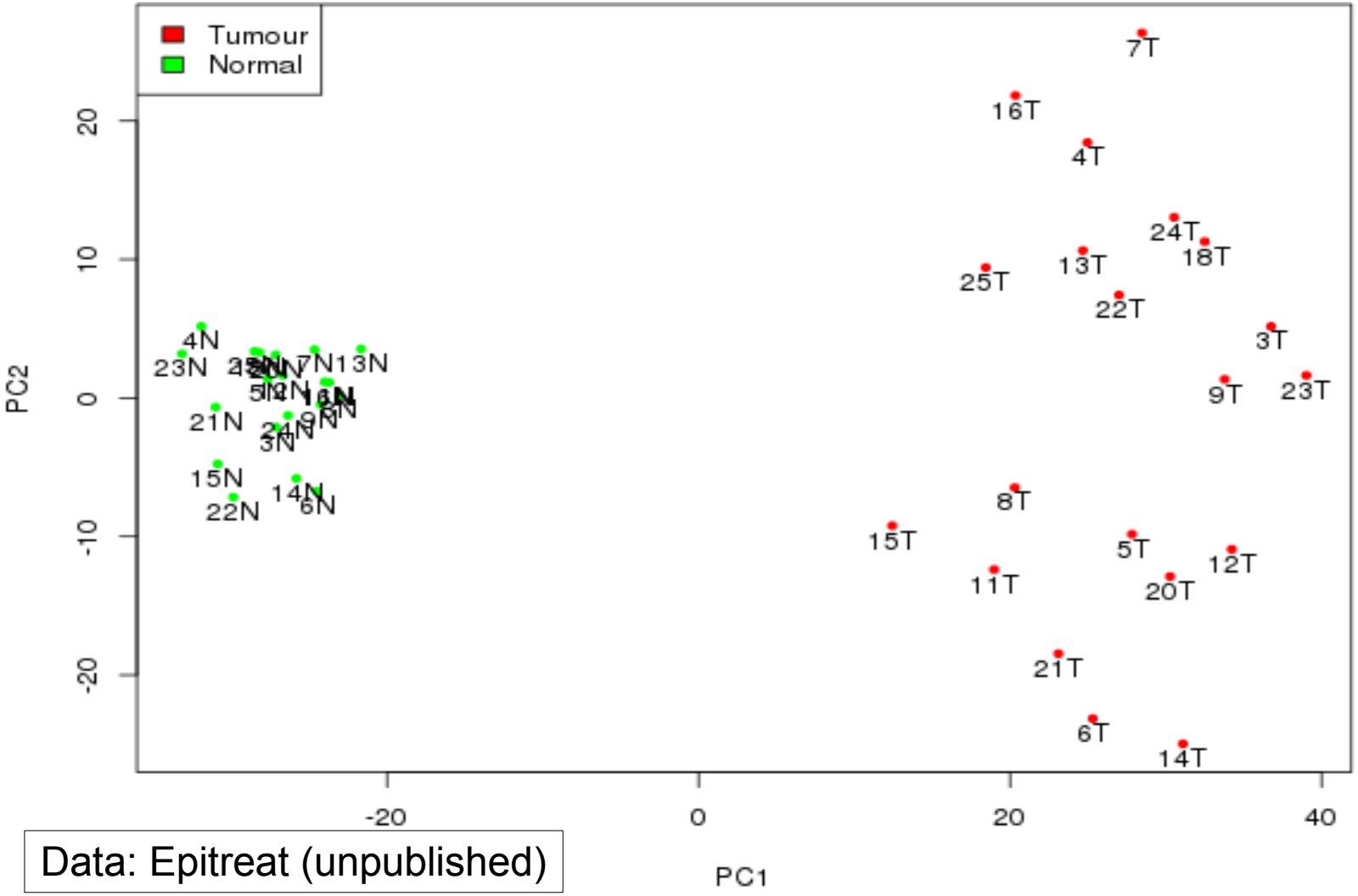
Further Components



Explained Variation of PC



Principle Component Analysis Tumour Example



Data: Epitreat (unpublished)

Summary: PCA / Clustering

Clustering:

- ➕ flexible (distance and clustering functions)
- ➕ Can display nested properties
- ➖ “Binary” decisions
- ➖ “One dimenisonal”

PCA:

- ➕ Separates independent factors
- ➕ PC interpretable as factors
- ➖ Needs sufficient samples
- ➖ Problems with nonlinear dependencies

Differential Gene Expression

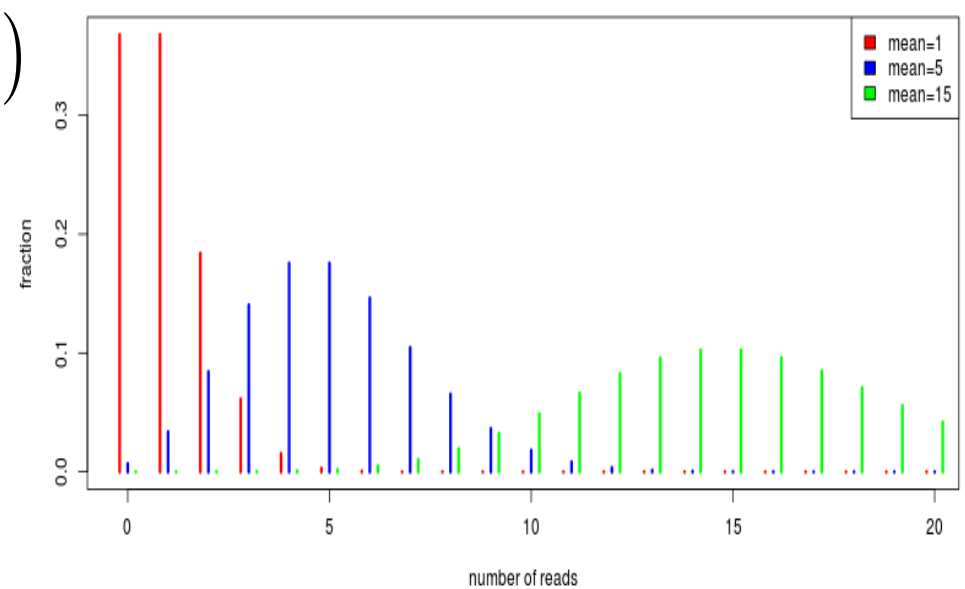
- p-value: probability that H0 is true: gene is expressed in A and B at the same level
  - assuming negative binomial distribution and estimated dispersion
  - If p is “very low” → H0 is “very unlikely”
- Test each gene
  - Multiple testing problem
  - FDR

Poisson Model

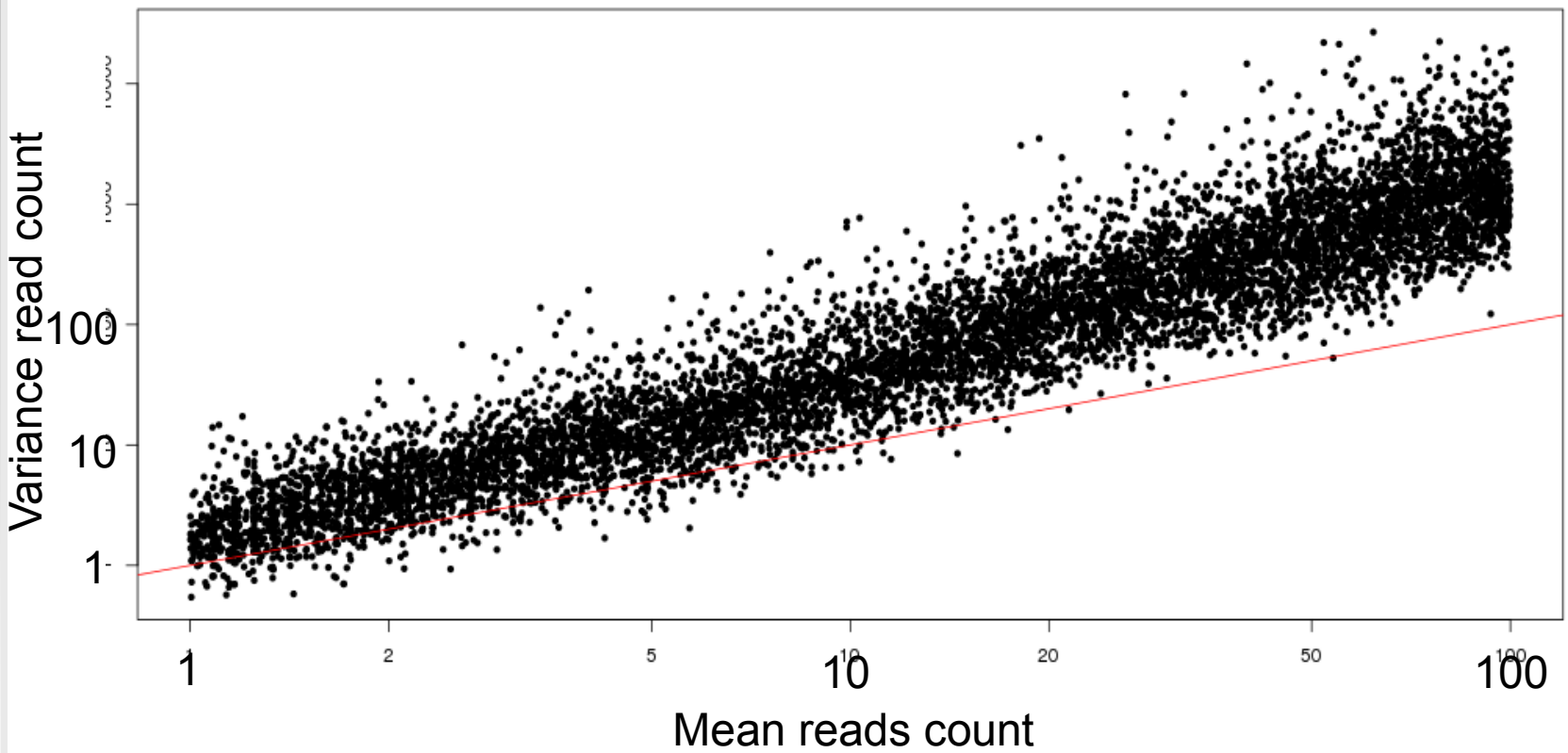
- N_i: total number of reads from sample *i*
- $\pi_{i,g}$: fraction of fragments from gene *g* in *i*
- y_{ig}: number of reads from *g* in *i*
- $\mu_{i,g} = E(Y_{i,g}) = N_i \pi_{i,g}$

$\rightarrow Y_{i,g} \sim Pois(\mu_{i,g})$

$\rightarrow Var(Y_{i,g}) = \mu_{i,g}$



Mean Variance Relation: Overdispersion



Samples are from different cells or organisms

Poisson Mixture Model

$$Y_{i,g} \sim Pois(\mu_{i,g} * \theta)$$

θ : Random variable with

- $E(\theta) = 1$
- $Var(\theta) = \Phi$

$Var(Y_{i,g}) = ?$ (backbord)

$$\rightarrow CV^2(Y_{i,g}) = CV^2_{technical} + CV^2_{biological}$$

Poisson Gamma Mixture

$$Y_{i,g} \sim Pois\left(\mu_{i,g} * \theta\right)$$

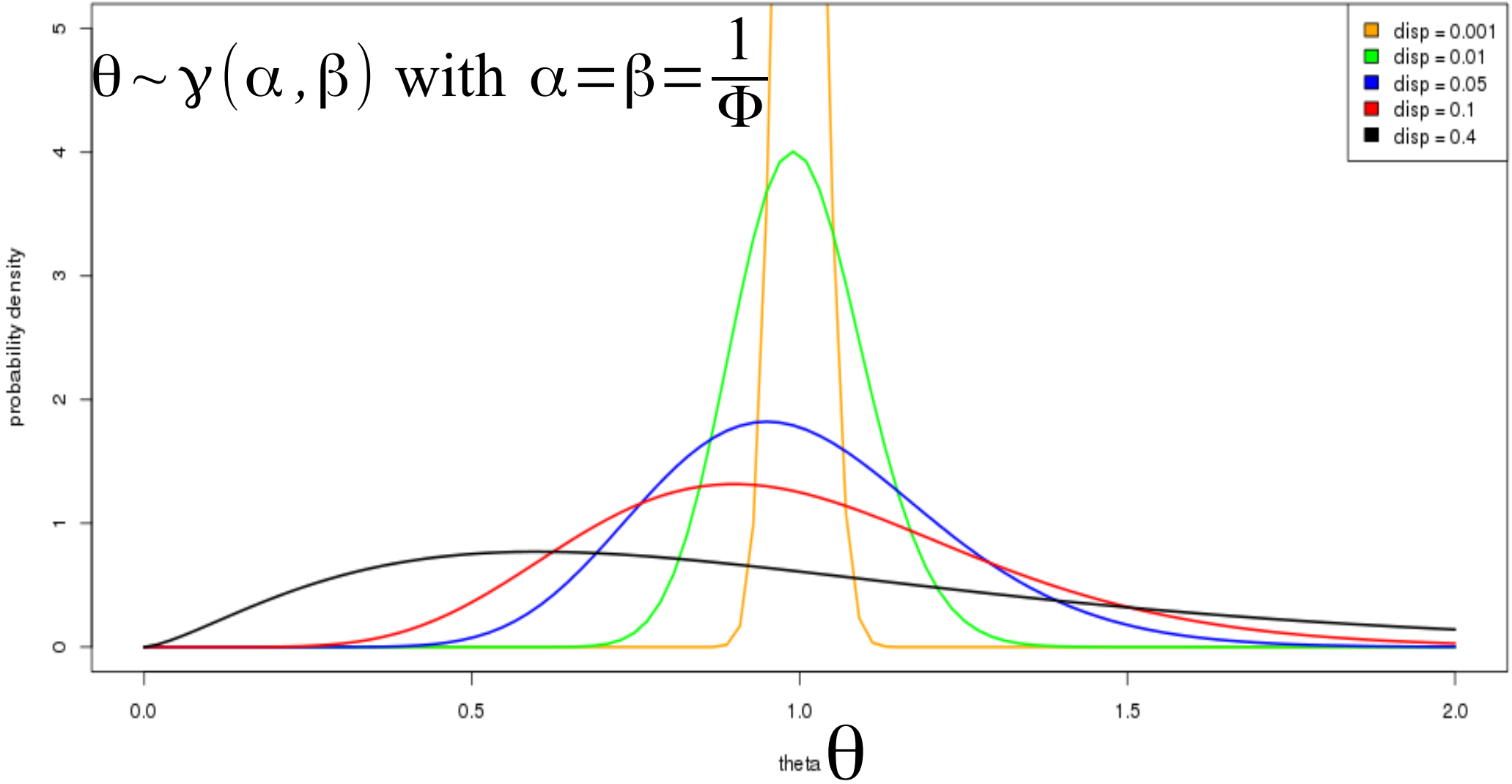
$$\theta \sim \gamma\left(\alpha, \beta\right) \text{ with } \alpha = \beta = \frac{1}{\Phi}$$

$$E\left(\theta\right) = \frac{\alpha}{\beta} = 1$$

$$Var\left(\theta\right) = \frac{\alpha}{\beta^2} = \Phi$$

Poisson Gamma Mixture

Gamma distribution



Poisson Gamma Mixture

$$Y_{i,g} \sim Pois(\mu_{i,g} * \theta)$$

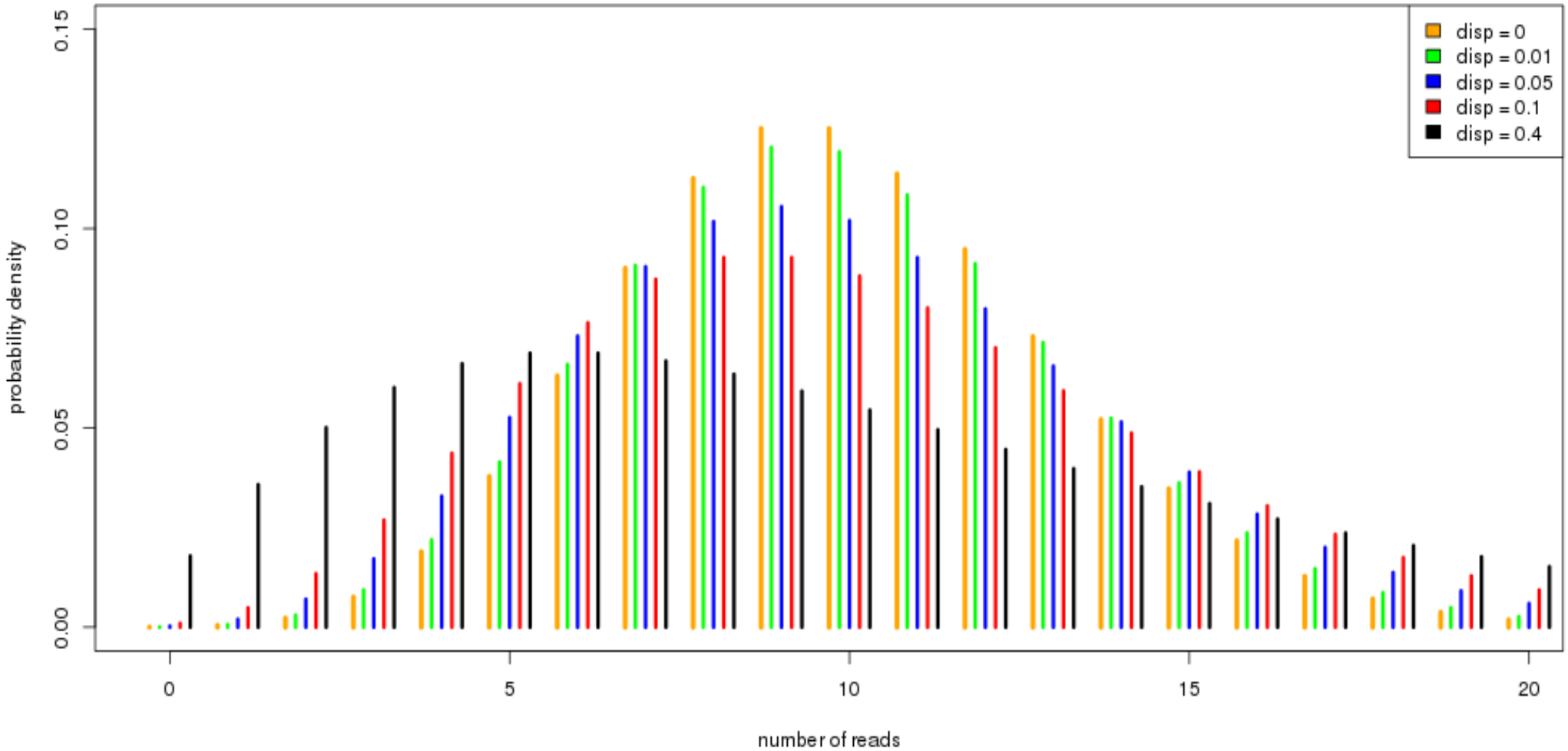
$$\theta \sim \gamma(\alpha, \beta) \text{ with } \alpha = \beta = \frac{1}{\Phi}$$

$$E(\theta) = \frac{\alpha}{\beta} = 1$$

$$Var(\theta) = \frac{\alpha}{\beta^2} = \Phi$$

$$\rightarrow Y_{i,g} \sim NB(k, r) \text{ with } k = \frac{1}{\Phi} \text{ and } r = \frac{1}{\mu * \Phi + 1}$$

Poisson Gamma Mixture



Estimating Dispersion

- edgeR: quantile adjusted conditional maximum likelihood estimate
- Problem: Few samples
  - Share information over genes

    → Common dispersion for all genes

    → Trended dispersion (expression level)

    → dispersion squeezed towards trend

    → dispersion cut by trend

MD Robinson (2008): Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics

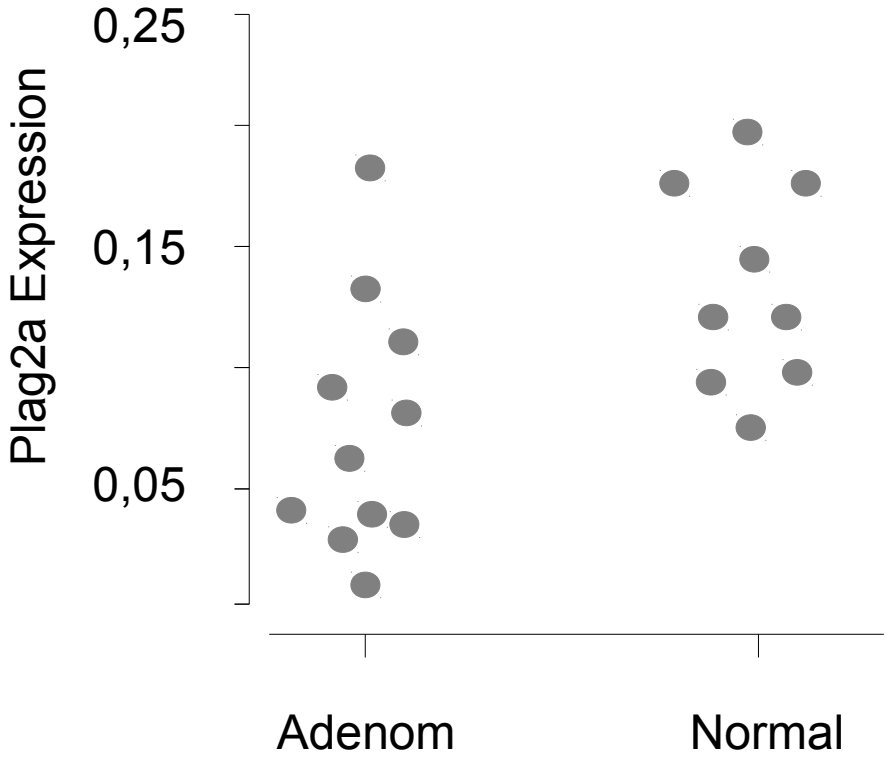
Testing Differential Expression

- H0: Reads for gene g are drawn from the same distribution for groups a and b
- y_T, y_a, y_b: # reads form gene in total, a and b
- N_T, N_a, N_b: total # reads in all, a and b
- $\mu_{0,a}, \mu_{0,b}: E(Y \mid H_{0,} N_x) = \frac{y_T}{N_T} * N_x$

$$pvalue = \sum_{i=0}^{y_T} pr(i \mid \mu_{0,a}, \Phi) * pr(y_T - i \mid \mu_{0,b}, \Phi) * I$$

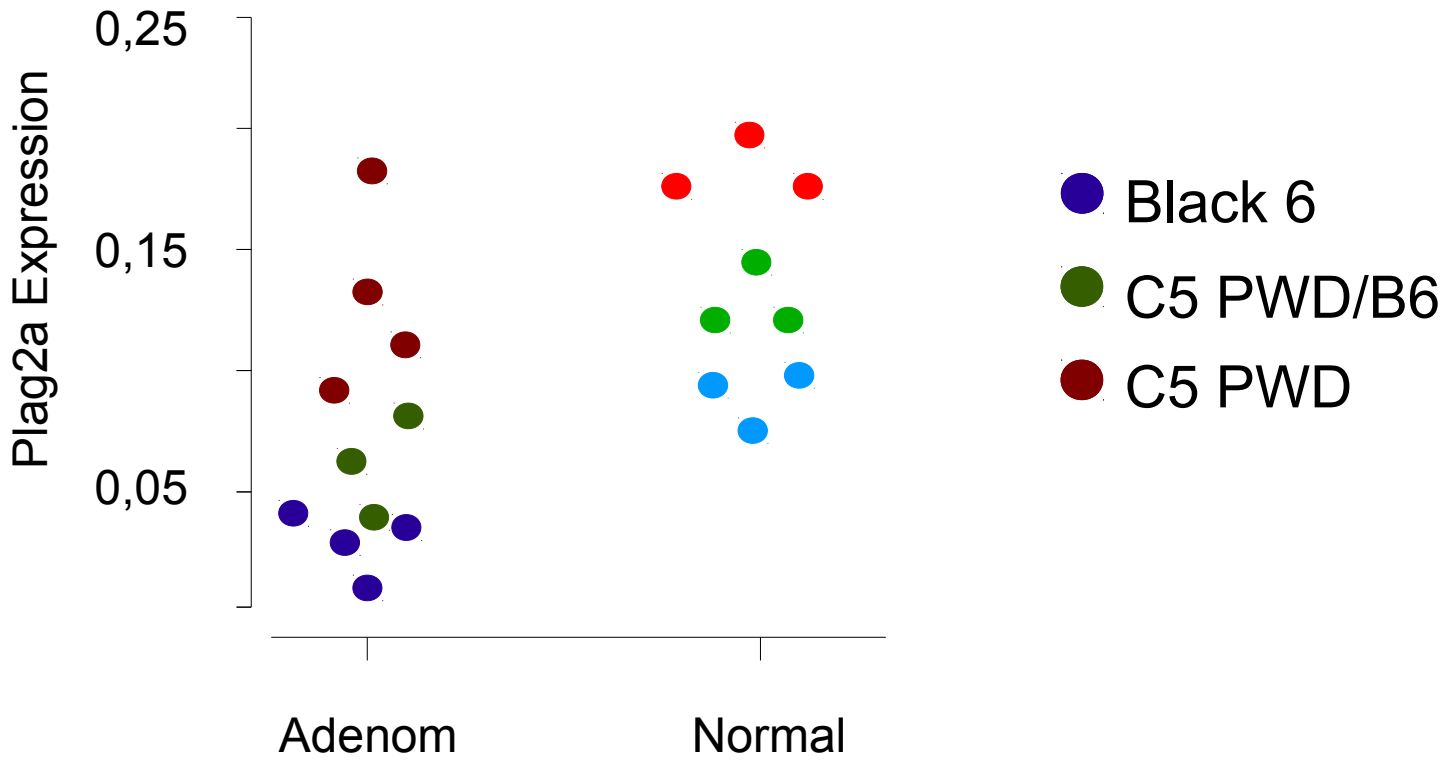
$$I \begin{cases} 1 \text{ if } pr_0(y_a) * pr_0(y_b) \geq pr_0(i) * pr_0(y_T - i) \\ 0 \text{ else} \end{cases}$$

Differentially Expressed?



More Complex Experimental Design

Example: APC-min mice



GLM

$$\log\left(\frac{\mu_{g,i}}{N_i}\right)=\beta_{0,g}+\beta_{1,g}*x_{1,i}+\ldots+\beta_{n,g}*x_{n,i}$$

$$\log(\mu_{g,i})=x_i^T\beta_g+\log(N_i) \longleftrightarrow$$

x_i: Vector of covariats from model matrix

β_g: Vector of regression coefficients

model matrix:

Sample	B6_ad1	B6_ad2	B6_no1	C5F1_ad1	C5F1_no1	C5_ad1	C5_no1
Intercept	1	1	1	1	1	1	1
Adenom	1	1	0	1	0	1	0
Chr5 PWD	0	0	0	1	1	2	2

GLM

$$\log(\mu_{g,i}) = x_i^T \beta_g + \log(N_i)$$

- Find estimates for beta for reduced (null) and full model
- Estimate dispersion under GLM
- Test for DE: likelihood ratio test

$$\frac{L(M_0)}{L(M_1)} \sim \chi^2$$

JD McCarthy et al.(2012): “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.” Nucleic Acids Research

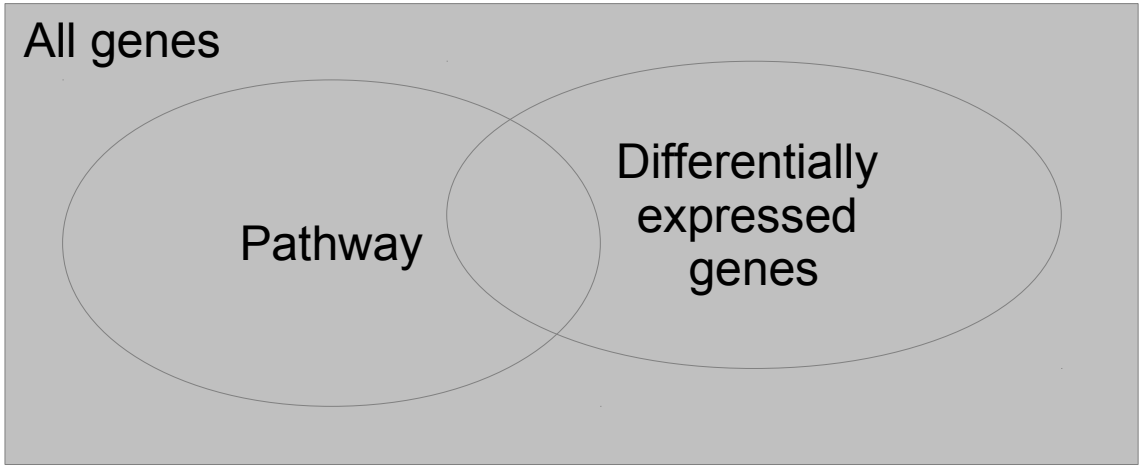
Summary: Tests for Differentially Expressed Genes

- Group A vs group B: exact NB test
- Multi factor test: GLM

→ List of differentially expressed genes

Interpretation of DE Genes

- Overrepresentation analysis
  - Web tools: CPDB, DAVID, ...
  - Integrate GO, Pathway databases, interaction databases, ...
  - Hypergeometric test: is overlap significant?



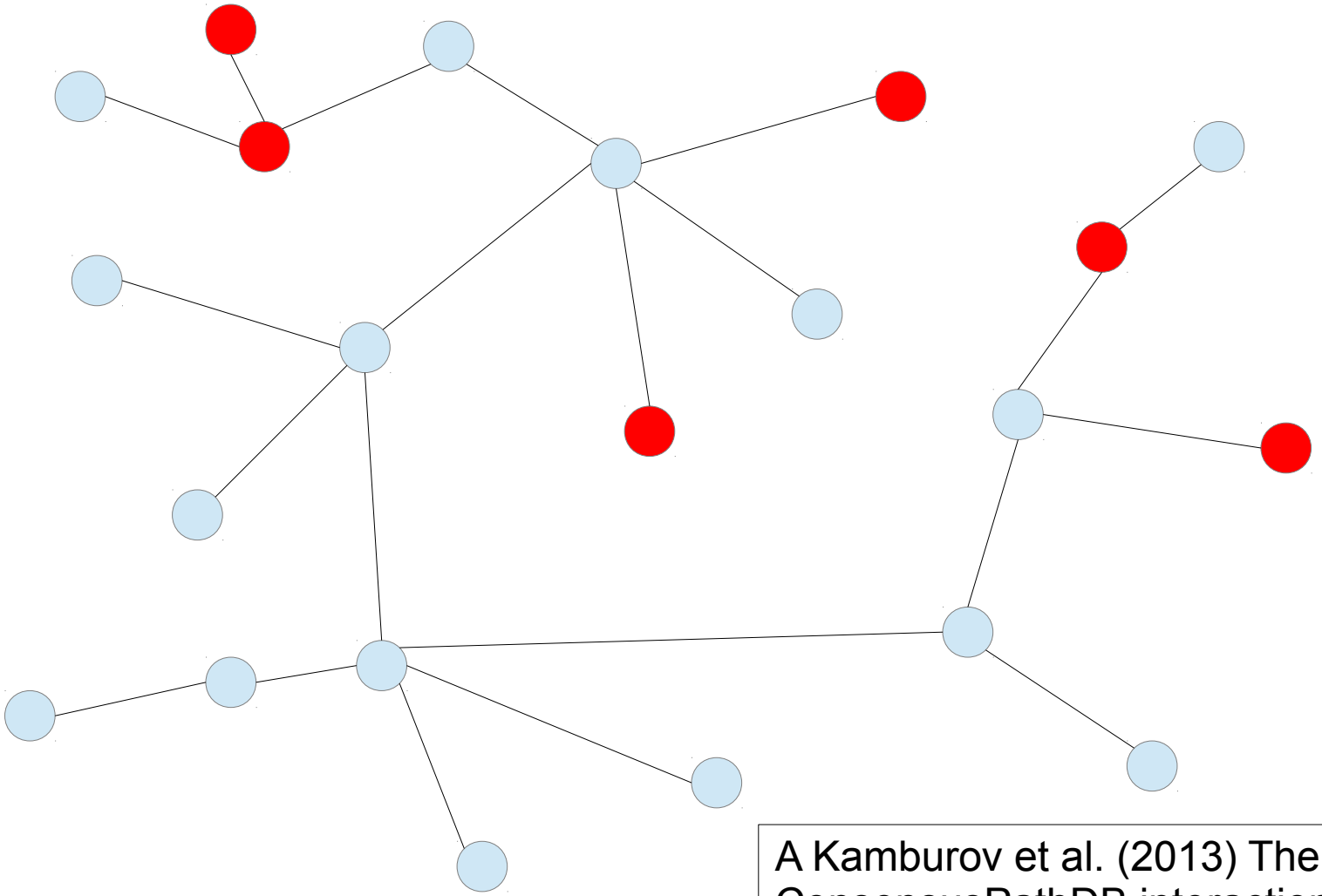
Pathway Analysis

CPDB Over-representation Analysis

pathway name	set size	candidates contained	p-value	q-value	pathway source
Direct p53 effectors	242 (135)	33 (24.4%)	6.52e-06	0.00729	PID
fibrinolysis pathway	20 (14)	8 (57.1%)	3.4e-05	0.019	BioCarta
p53 signaling pathway - Homo sapiens (human)	69 (67)	19 (28.4%)	7.09e-05	0.0244	KEGG
Facilitative Na+-independent glucose transporters	12 (12)	7 (58.3%)	9.15e-05	0.0244	Reactome
Benzo(a)pyrene metabolism	9 (9)	6 (66.7%)	0.000109	0.0244	Wikipathways
Oxidative Stress	29 (26)	10 (38.5%)	0.000251	0.0467	Wikipathways

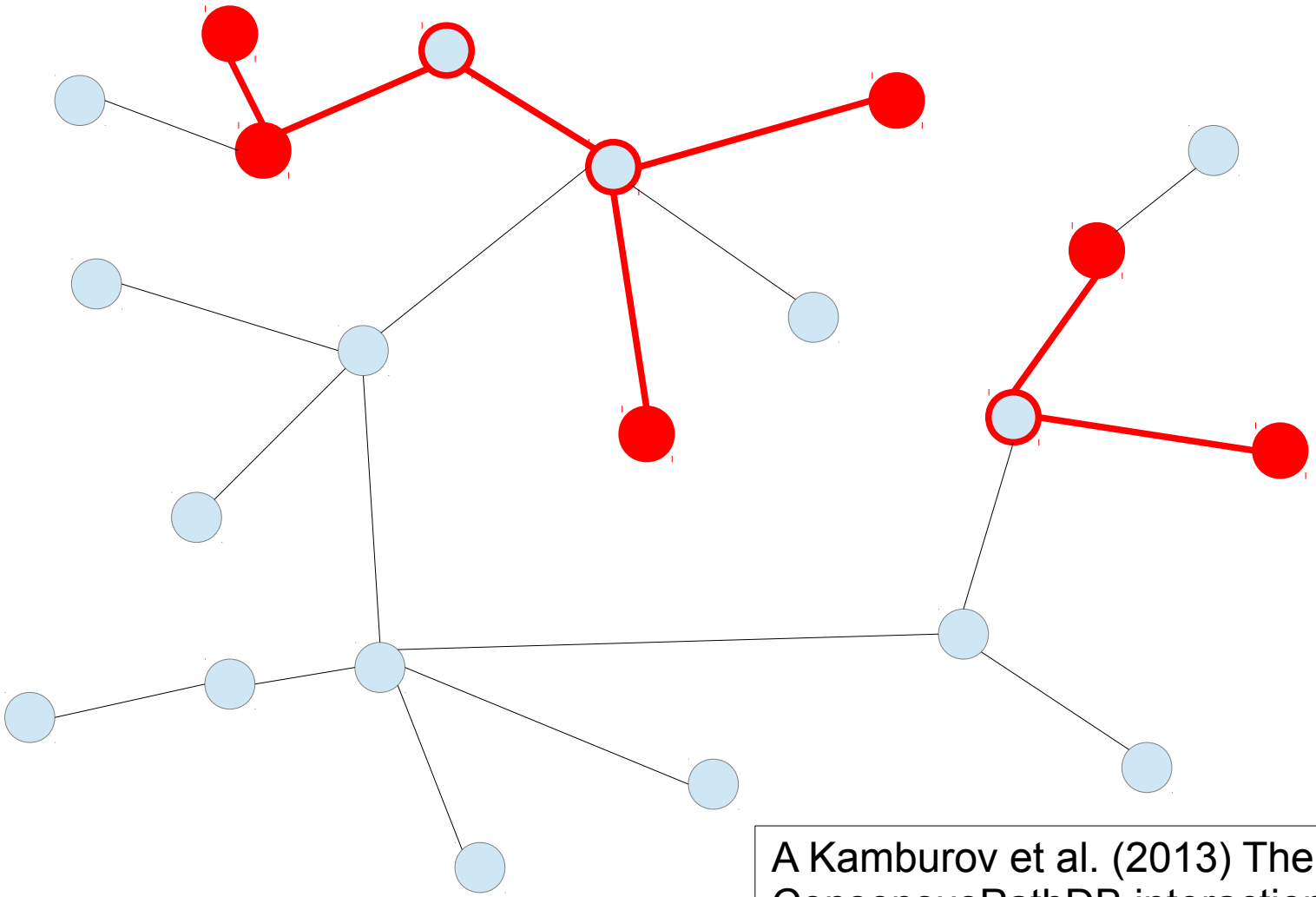
Kamburov, A. et al. (2009) ConsensusPathDB--a database for integrating human interaction networks. Nucleic Acids Res.37:D623-628.

CPDB: Induced networks



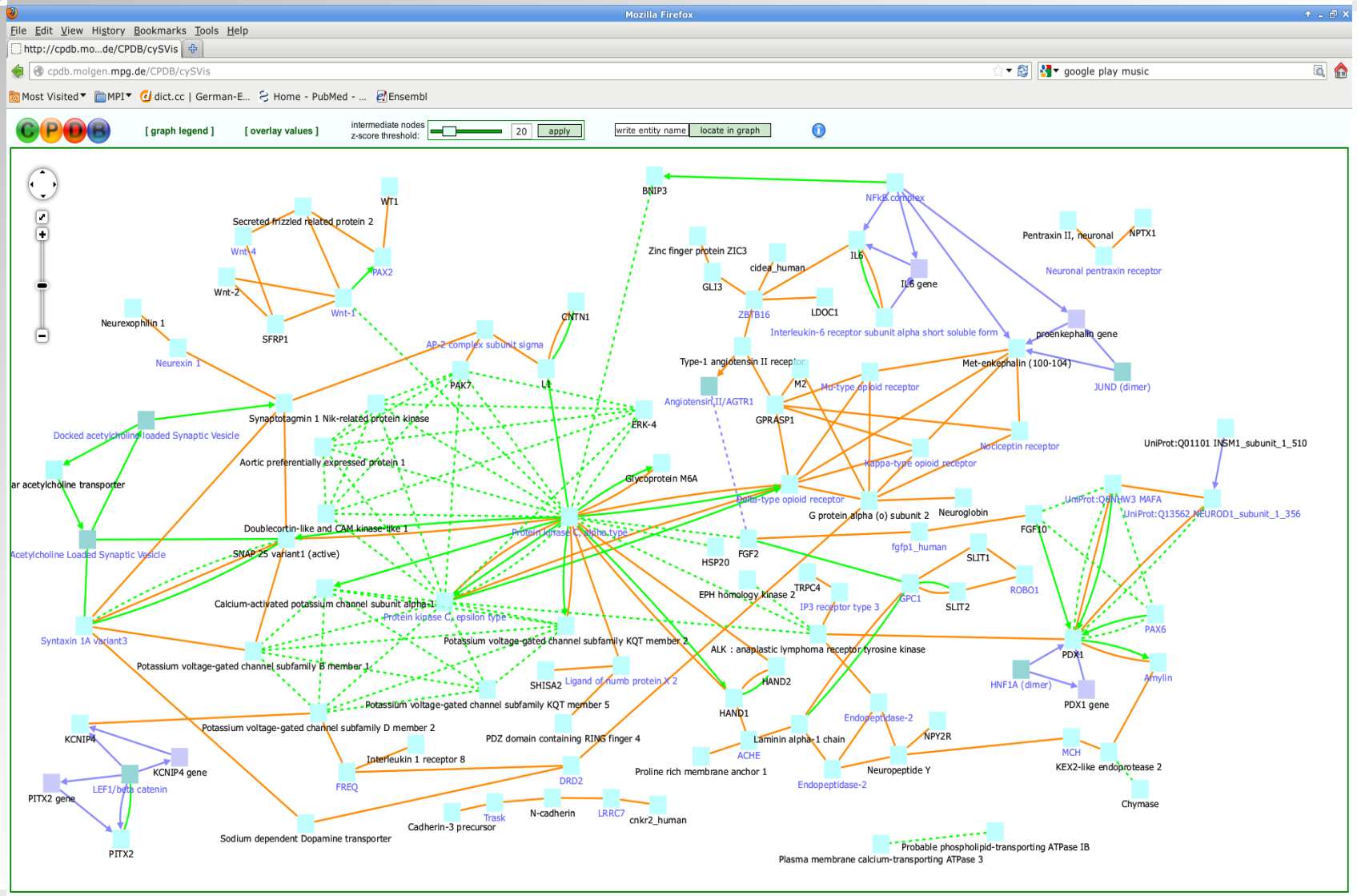
A Kamburov et al. (2013) The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res.

CPDB: Induced networks

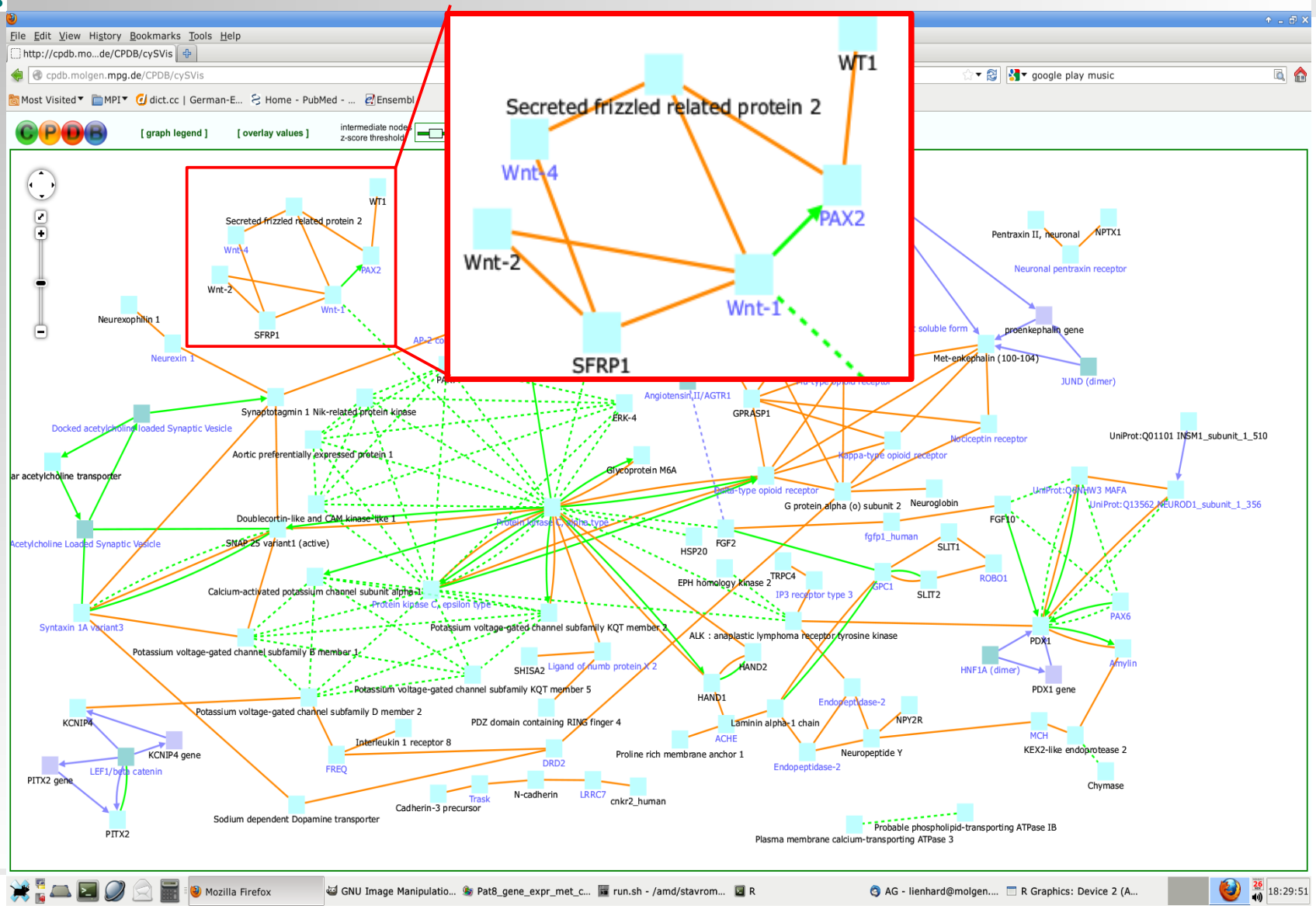


A Kamburov et al. (2013) The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res.

Induced Network



Induced Network



Summary

- Quality Control: Do data look OK?
- Mapping: Handle reads across exon boundaries
- Quantification: Gene/isoform/exon level
- Exploratory analysis: Relation of Samples?
- Differential Expression: A vs B or GLM?
- Over representation and network analysis: Make sense out of gene lists.