



# ***Genetical genomics applied to hematopoietic development using mouse models***

*Bioinformatics research into Illumina microarray data*

Danny Arends (s1276891)

supervision: Yang Li & Ritsert Jansen

Current work:

Molecular Genetics,

supervision: Sacha v Hijum.



# Presentation Overview

- Experiment
- Illumina
  - Technology overview
- Bioinformatics (Genetical Genomics)
- Results
  - Data quality
  - Differential expression
  - eQTL
- Discussion
- Future perspectives

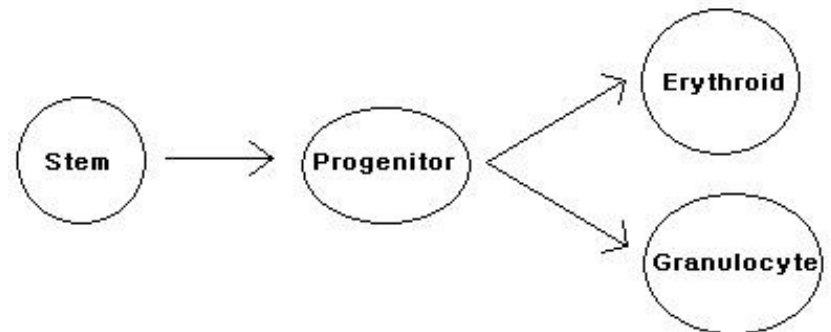
# Experiment

- 2 Recombinant inbred lines
  - C57BL/6 and DBA/2

- Stem cells

- Progenitor cells

- Erythrocytes
- Granulocytes



- RNA-expression measured in each cell type.
- Illumina sentrix beadarrays



# Experiment

- Question:  
Can we find gene(s) responsible for differentiation ?
- Inverse question:  
Can we find gene(s) responsible for proliferation of stemcells ?
- Hypothesis:  
Bioinformatics can help find regulatory elements and interactions governing differentiation/proliferation



# Illumina

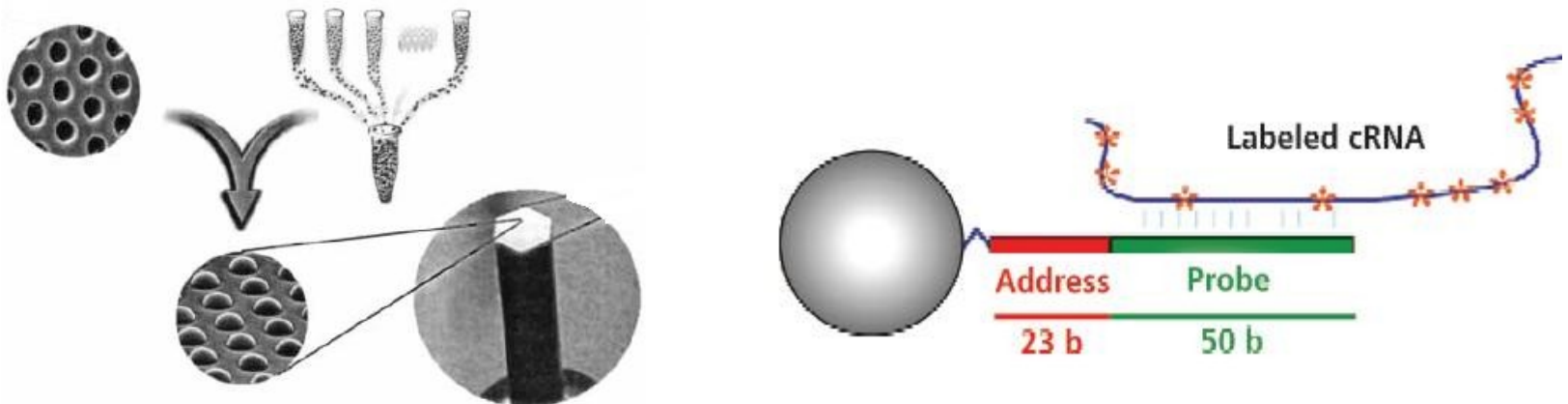
- Founded april 1998
- Innovative technology
- DNA / RNA arrays
- Sentrix Beadarray technology
- 9/1/2005, Mouse whole genome expression
- 9/1/2006, Rat whole genome expression



# Technology overview

## Manufacturing the arrays

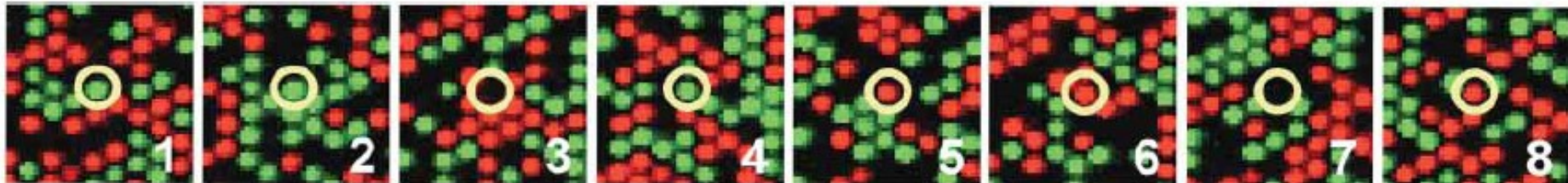
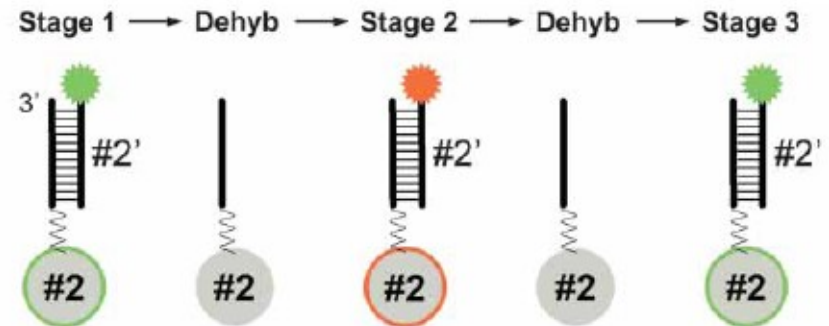
- Long oligonucleotides attached to glass beads
- 50 nucleotide probe
- 23 nucleotide address (bead ID)
- Pooled together in beadpools
- Random assembly in etched wells



# Technology overview

## Decoding the arrays

- Sequential hybridization with labeled decoders
- R-G-0-R-R-G-G-0
- 2 OFF states per code



# Technology overview

- Sentrix technology
  - Mouse (mus musculus)
  - 6 samples per sentrix microarray
  - 5 for this experiment (every 6<sup>th</sup> position failed)
  - Bead intensity measurements
- 62 samples on 13 arrays
  - 20 stem cells
  - 14 progenitor
  - 14 erythrocytes
  - 14 granulocytes





# Problems / Issues

- Random number of replicate beads
  - Between 0 and 60 on average ~30 replicate beads
- Decoding errors
  - Defect beads
  - Non detection
  - Wrong detection
- Unique arrays
- Batch effect / array effect
- Non functional arrays



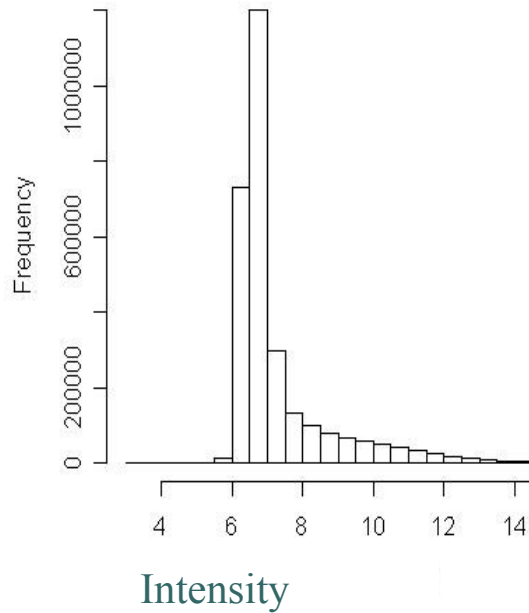
# Bioinformatics

(Genetical Genomics)

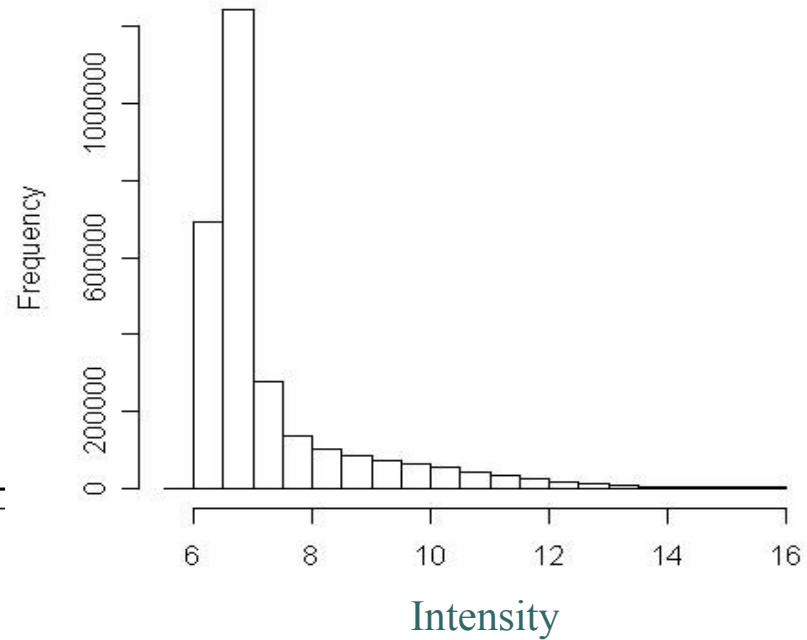
- Data pre-processing
  - Data exploration
  - Quantile Normalization
- Cell type specific effect in expression (ANOVA analysis)
- Genetical Genomics studies
  - Recombinant inbred lines
  - Expression QTL mapping

# Data exploration

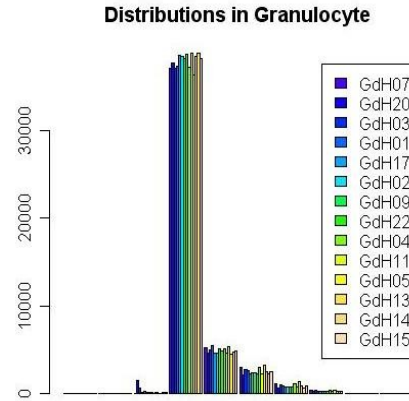
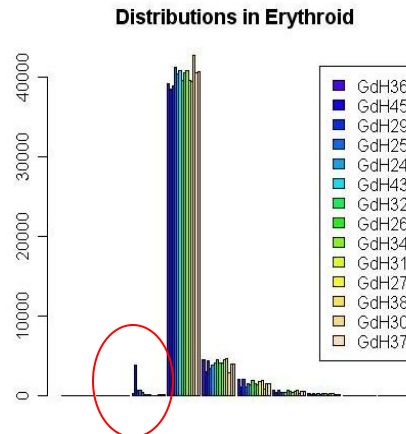
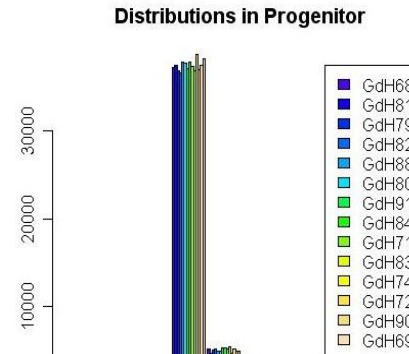
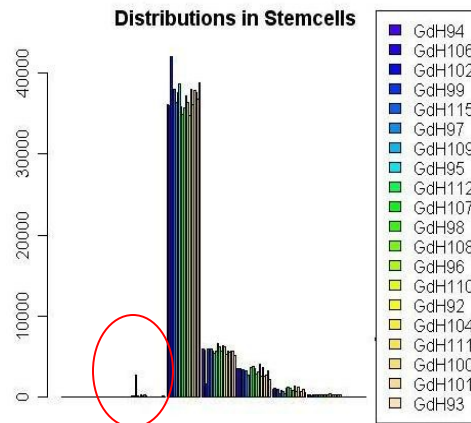
**Before quantile normalization**  
**Histogram Log2**



**After quantile normalization**



# Data exploration



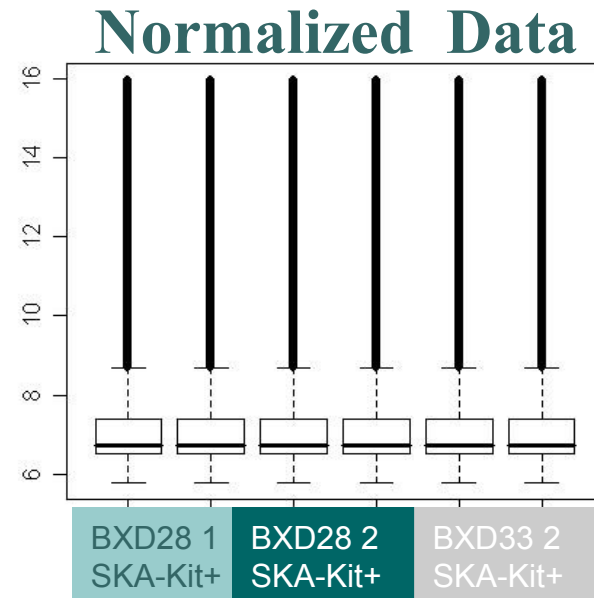
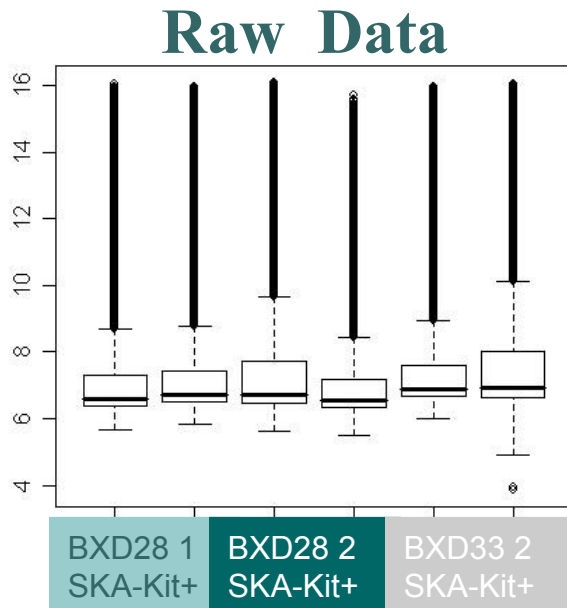


# Results histogram analysis

- No overhybridization seen
- Normalization doesn't seem to significantly change distribution
- Samples look equally distributed
- Stem cells and erythrocytes each seem to have 1 array with lower intensities

# Data exploration

- Biological replicates present
- Equal variation after normalization
- Boxplots of the biological replicates





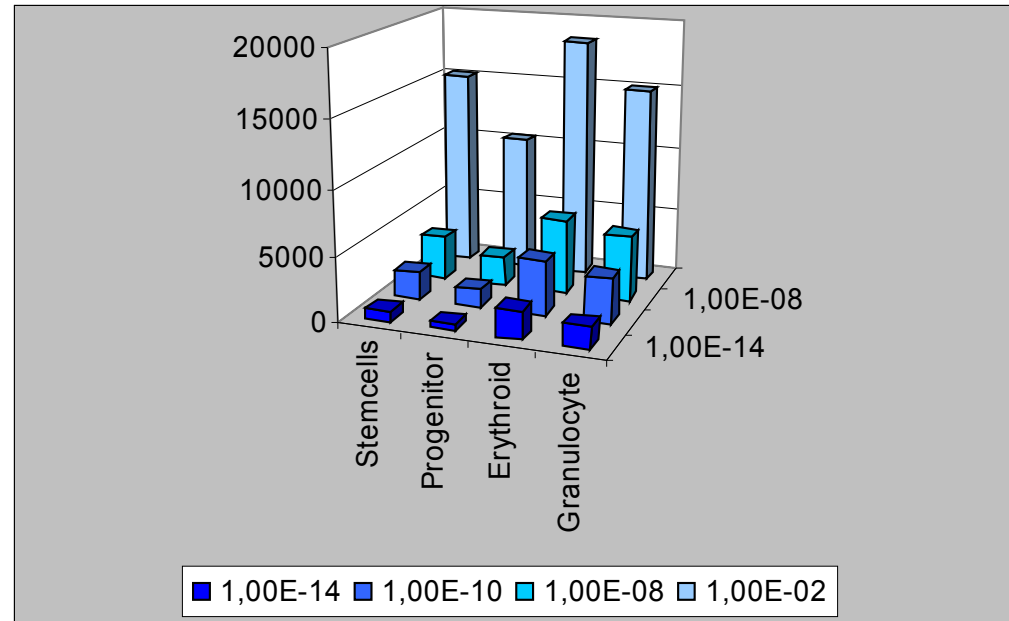
# Differential expression

- Differential expression using ANOVA
- ANOVA Model:  $y_i = \mu + C_i + \varepsilon$
- T-test for celltype specificity
- Test compared one cell type with the other 3 cell types

$$T = \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

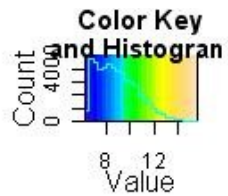
# Number of probes with cell specific effect

## ○ Threshold vs. Hits

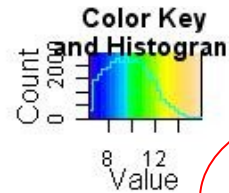
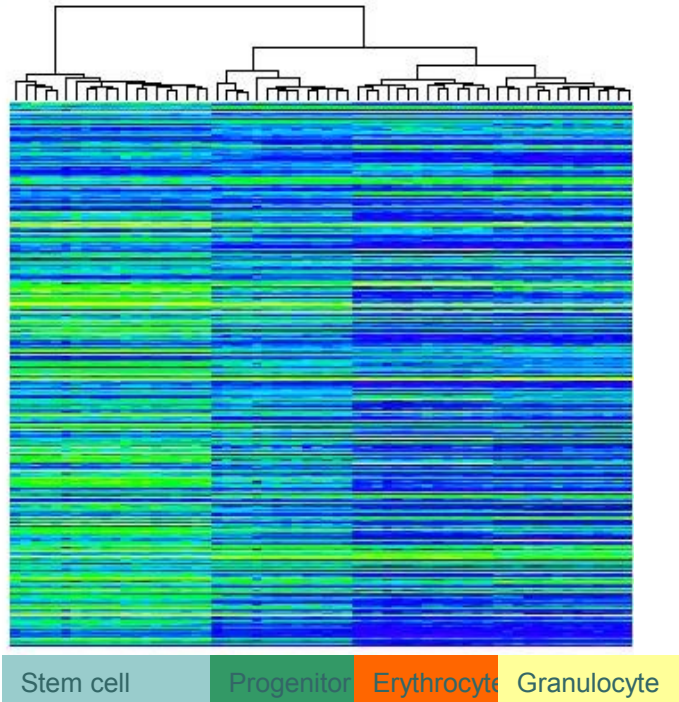


	1,0E-02	1,0E-08	1,0E-10	1,0E-14
Stemcells	15176	3340	2088	815
Progenitor	10416	2234	1383	501
Erythrocyte	18523	5769	4200	2128
Granulocyte	15028	5030	3460	1655

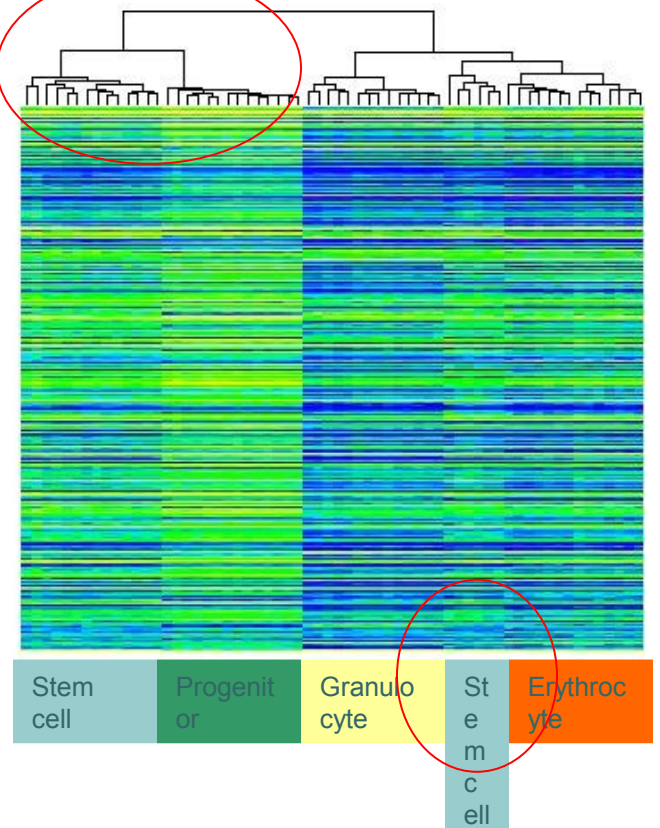
# Heatmap of gene expression for cell specific probes



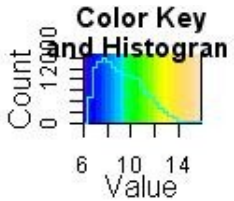
## Stemcell specific



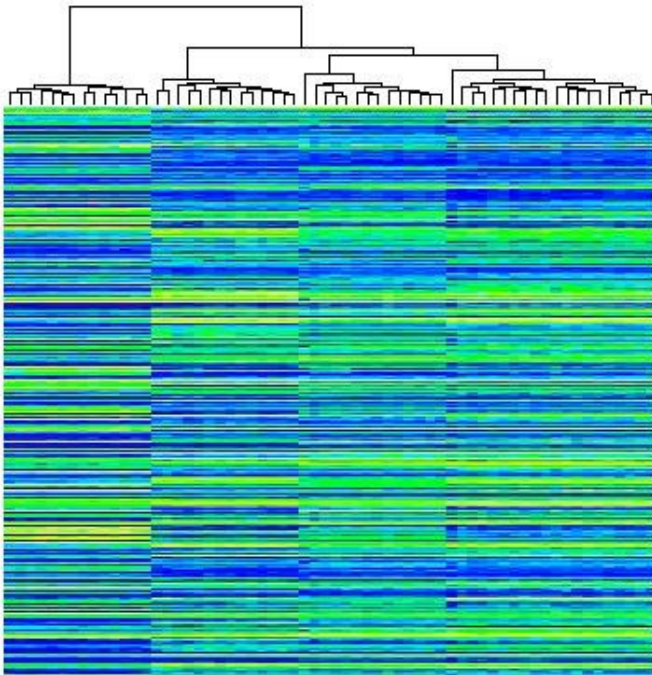
## Progenitor specific



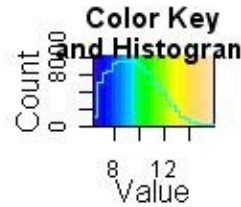
# Heatmap of gene expression for cell specific probes



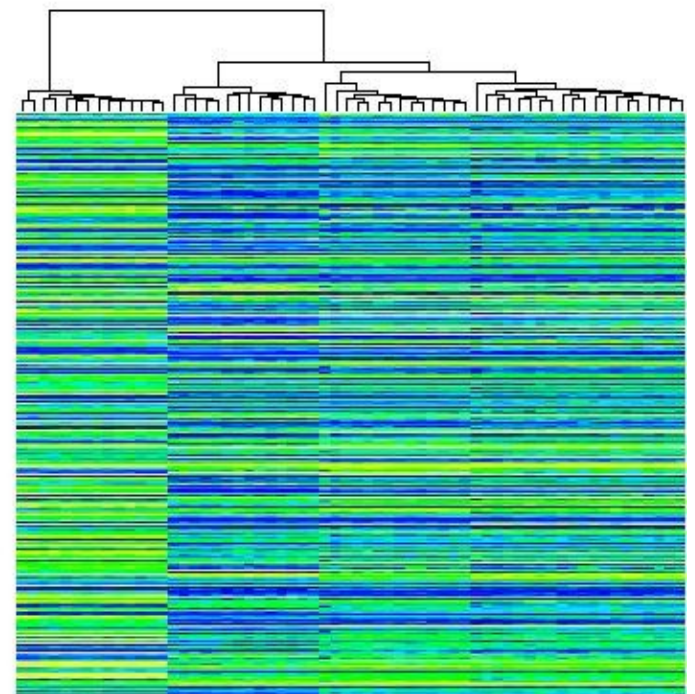
## Erythroid specific



Erythrocyte Granulocyte Progenitor Stem cell



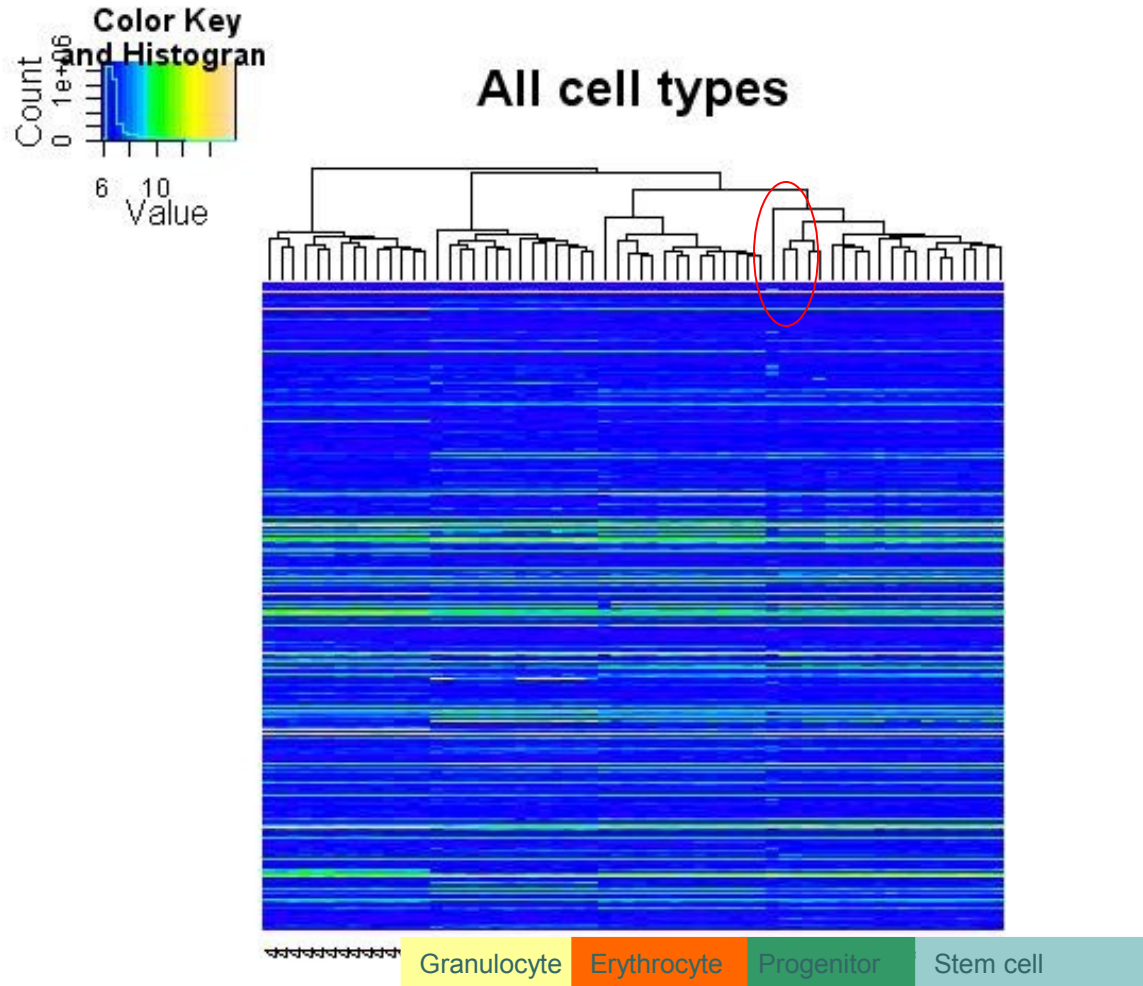
## Granulocyte specific



Granulocyte Erythrocyte Progenitor Stem cell

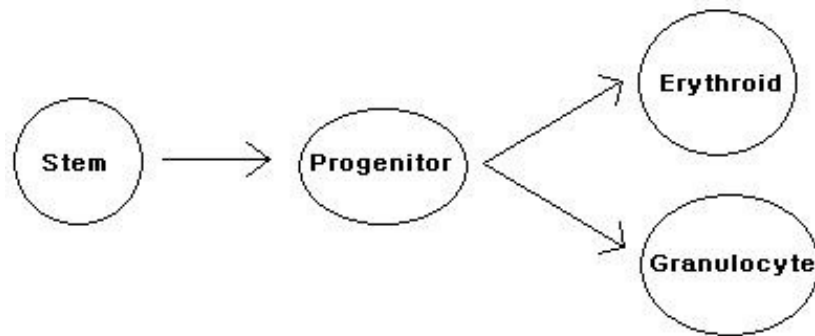
# Heatmap of gene expression for all probes

All probes (no specificity)



# Results heatmap analysis

- Differential expression is seen
- Distance clustering shows outlier in stemcell
- Clustering into 4 distinct cell types
- Progenitor specific genes show overlap with type 1 (stemcells)
- Model seems to fit distance clustering for all probes





# goClustering

- Gene ontology groups
  - controlled vocabulary to describe **gene** and **gene** product attributes in any organism.
- Detect overrepresentation of goGroups
- Statistical process



# Summary of ontology clustering

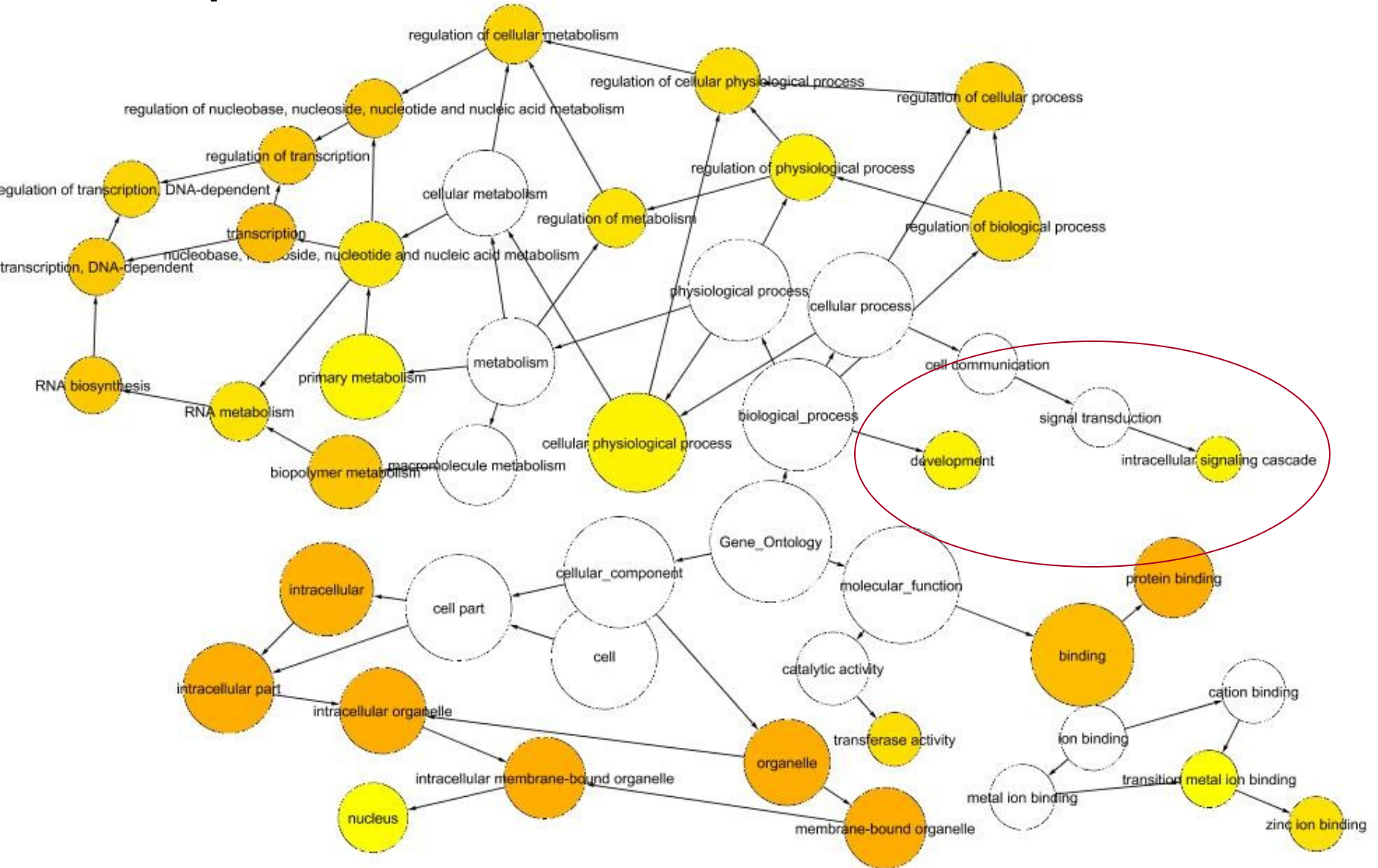
	Stem cells	Progenitor	Granulocyte	Erythrocyte
Biological process	Ubiquitin Cycle	cellular macromolecule metabolic process	nucleobase, nucleoside, nucleotide	cellular carbohydrate metabolic process
		immune system process	nucleic acid metabolic process	methylation-dependent chromatin silencing
			macromolecule metabolic process	M phase
			cell communication	cell cycle
				biogenic amine catabolic process
Molecular Function	None Found	unfolded protein binding	nucleic acid binding	cobalt ion transporter activity
		monovalent inorganic cation transporter activity	RNA binding	
		phosphoric monoester hydrolase activity	cytoskeletal protein binding	
			signal transducer activity	
			urea transporter activity	
			phospholipid-translocating ATPase activity	
		microtubule motor activity		



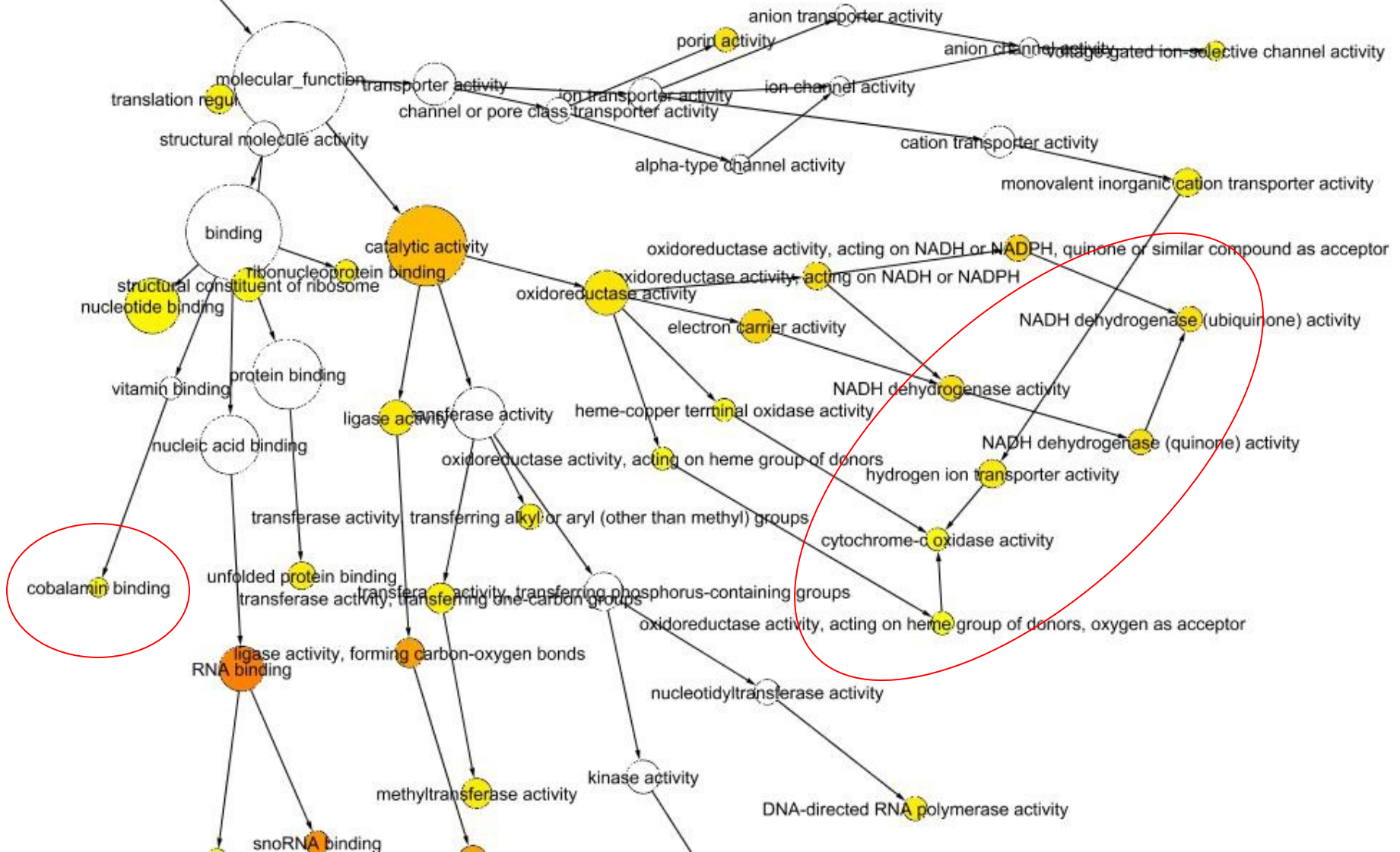
# Visualizing gene ontology

- Cytoscape for visualizing networks
  - Much used bioinformatics tool
  - Loads of plug-ins available
  - Still in development
- Using BINGO
- Database ontology search
- Hierarchical groups
  - Gene ontology

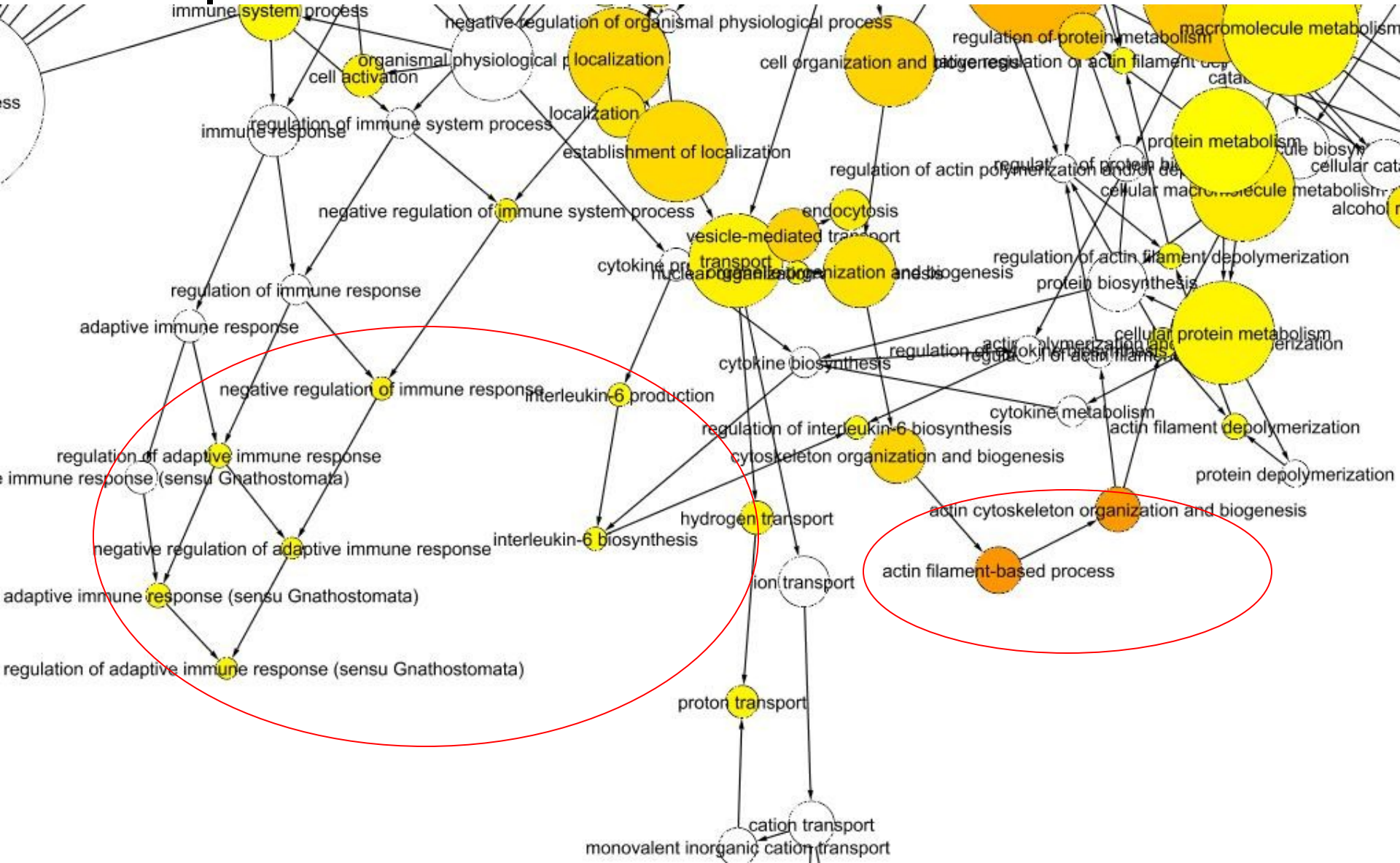
# Gene ontology of stem cells



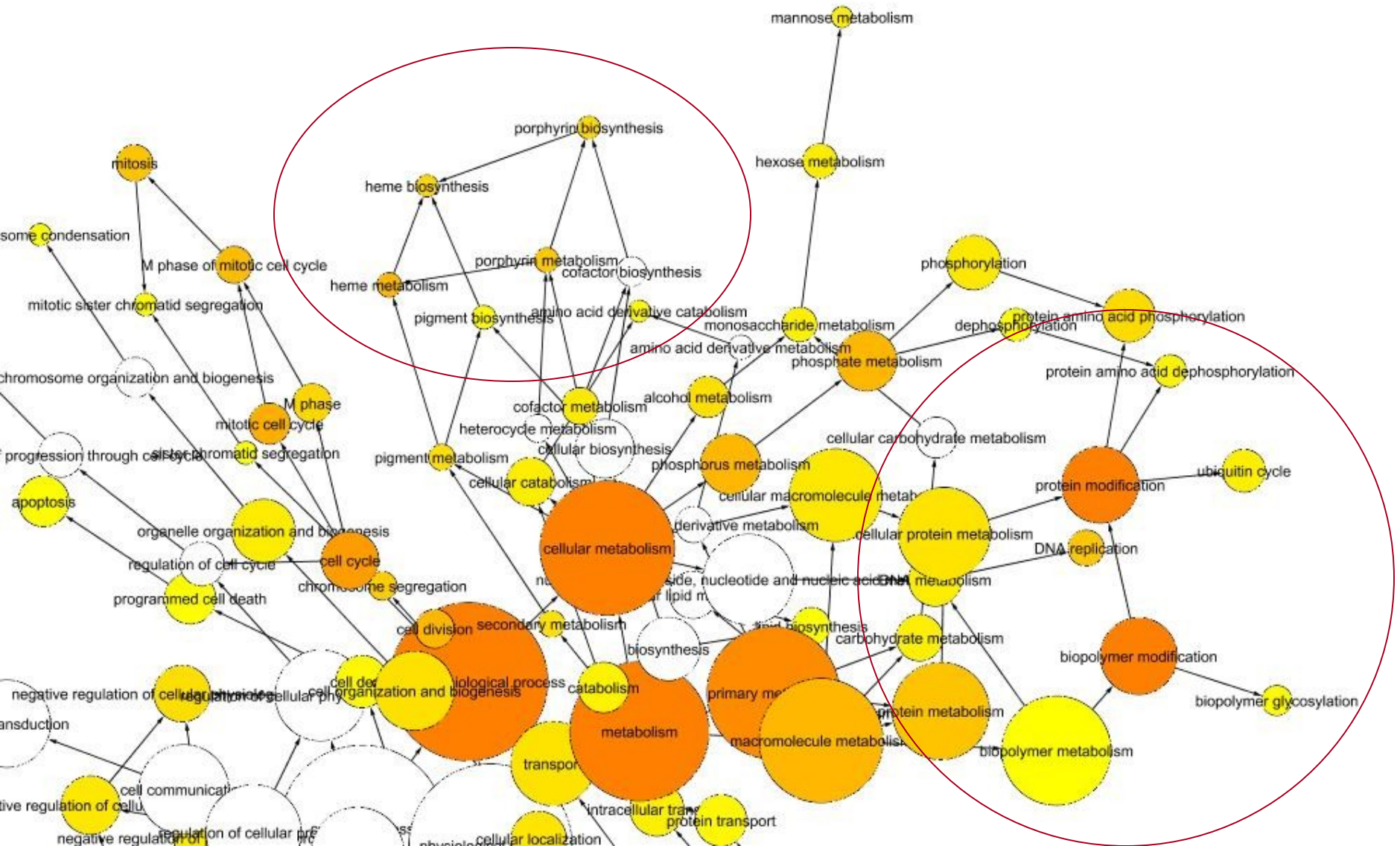
# Gene ontology of progenitor



# Gene ontology of granulocyte



# Gene ontology of erythrocyte cells





# Interaction networks

- Made by BioNetBuilder
  - plugin for cytoscape
- Public databases searched:
  - KEGG/ProLinks/BIND/DIP/BioGrid
- Results for differentially expressed genes:
  - Stem: 110 interaction
  - Prog: 48 interactions
  - Granulocyte: 469 interactions
  - Erythrocyte: 552 interactions
- Results are not fully shown





# Genetical Genomics

- Recombinant Inbred Lines
  - B6 and DBA mouse strains
- Genotyping
  - Every genetic location can either be inherited from B6 or DBA
- Expression data
- Mapping together
  - Predicting regulators (influential) regions using genotype data and expression data
- Statistical process

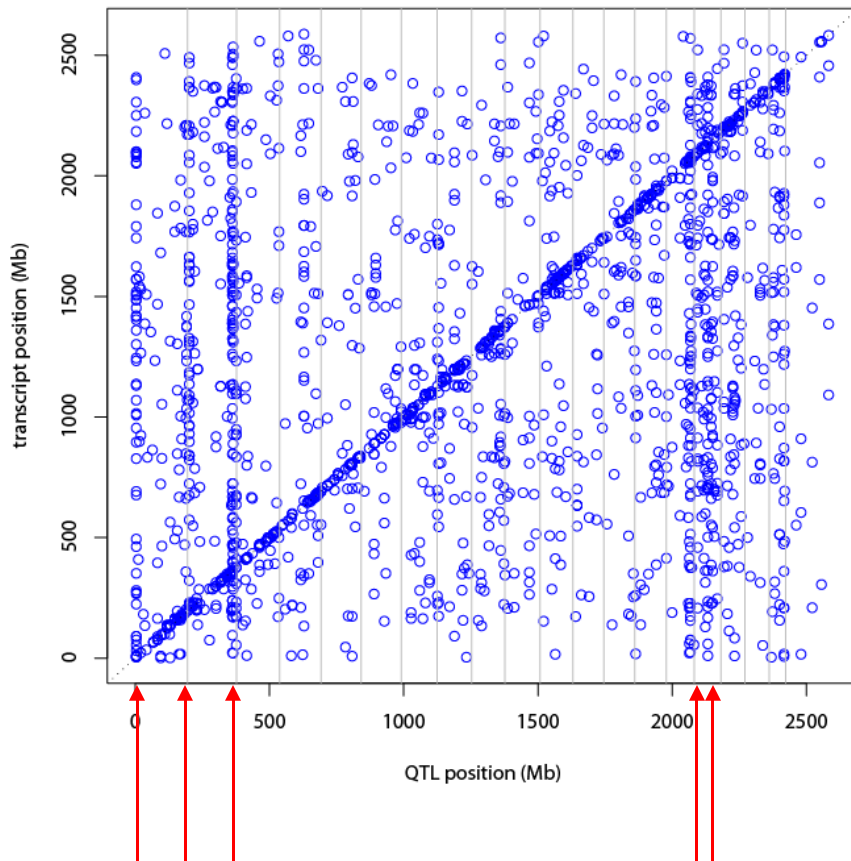


# eQTL mapping

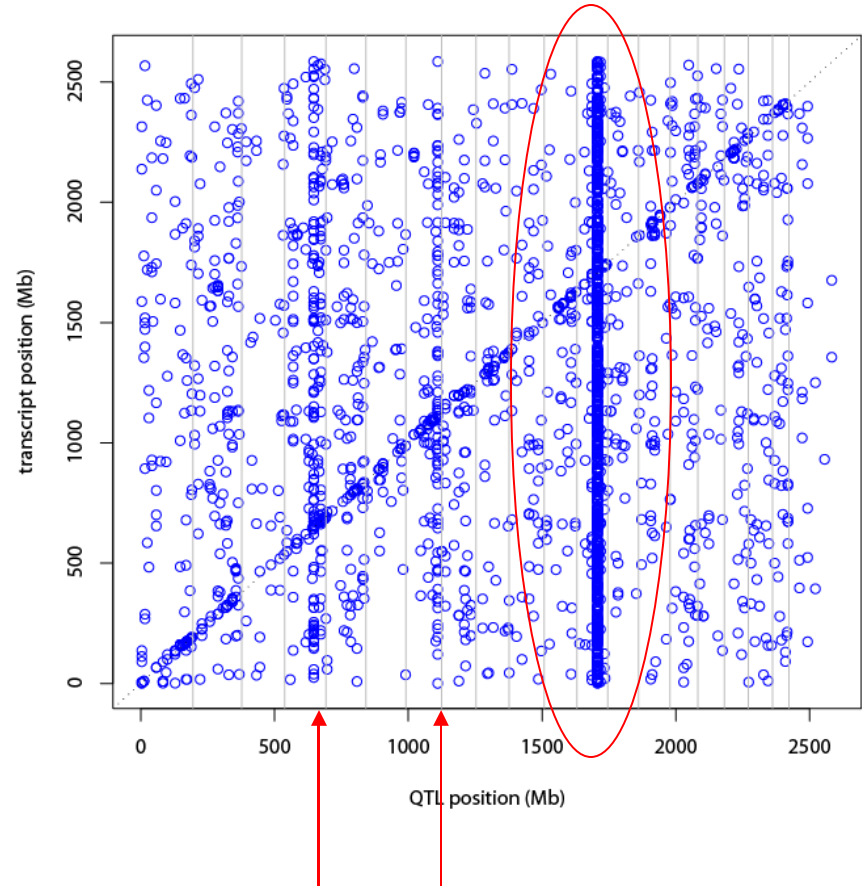
- ANOVA Model:  $y_i = \mu + Q_i + \varepsilon$
- Integration of genetic information and gene expression profiling
- *Cis* and *Trans* effects seen in all cell types
- Progenitor cell type has very strong transband

# Comparison of eQTL position with transcript position

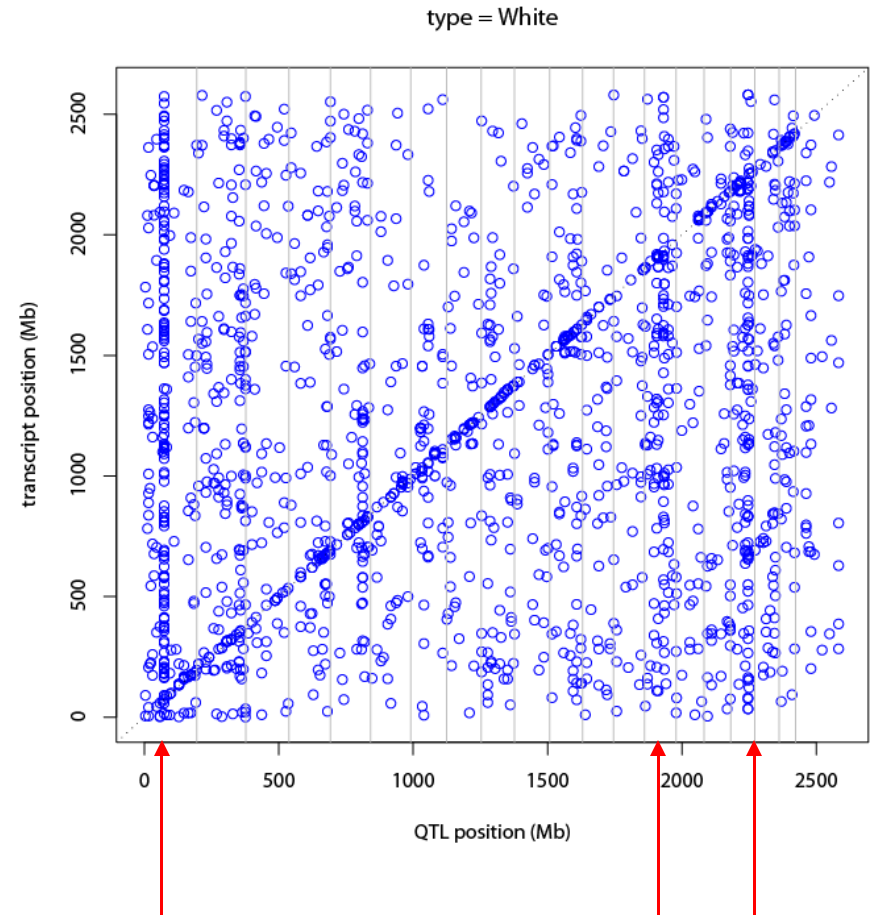
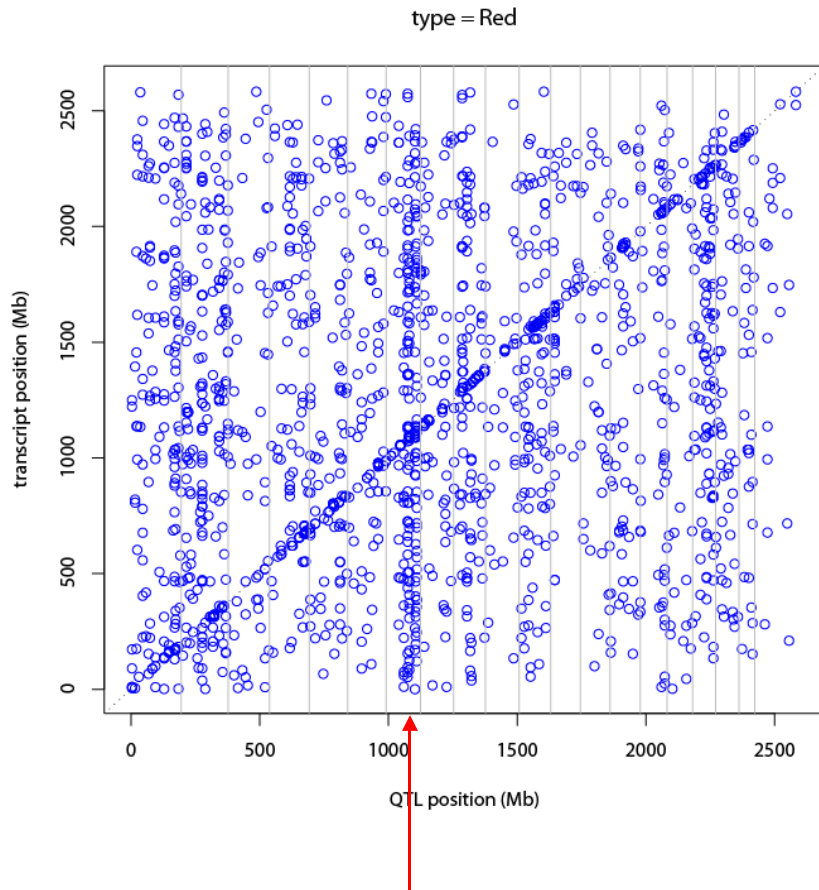
type = Stem



type = Pro



# Comparison of eQTL position with transcript position





# Conclusion / Discussion

- Illumina beadarray is a new technology
- Data obtained from illumina seem to be good
- Bioinformatics tools can be used to obtain knowledge about biological systems.
- Gene ontology can give researchers a clue in which pathway to look for regulation
- Interaction networks show known literature
- eQTLs can give hints into regulatory mechanisms
- Changes in transbands seen between cell-types
  - Regulation locations



# Future perspectives

- Bead level analysis of Illumina data
- Comparison of Illumina with Affy
- Scaling up experiment
- Further eQTLs studies
- Combining eQTLs with gene ontology
- Combining eQTLs with known interactions
- More research into stemcell differentiation is needed (Biology).

# Questions/Suggestions ?

I'd like to thank the following people:

- From GBIC (Haren)
  - Ritsert Jansen
  - Yang Li
  - Bruno Tesson
  - Morris Swertz
  - Gonzalo vera Rodriguez
- From Stemcell biology (UMCG)
  - Leonid v Bystrykh
  - Gerald de Haan

