# V10: Bayesian Parameter Estimation

Although the MLE approach seems plausible,
it can be overly simplistic in many cases.

Assume again that we perform the **thumbtack** experiment
and get 3 heads out of 10 $\rightarrow$ assuming $\theta = 0.3$ is then quite reasonable.

But what if we do the same experiment with a **standard coin**,
and also get 3 heads?

Intuitively, we would probably not conclude that the parameter of the coin is 0.3.

Why not?

Because we have a lot more experience with tossing coins,
we have a lot more **prior knowledge** about their behavior.

# Joint probabilistic model

In the Bayesian approach, we encode our prior knowledge about $\theta$ with a probability distribution.

This distribution represents how likely we are **a priori** to believe the different choices of parameters

Then we can create **a joint distribution** over the parameter $\theta$ and the data cases X[1], …, X[M] that we are about to observe.

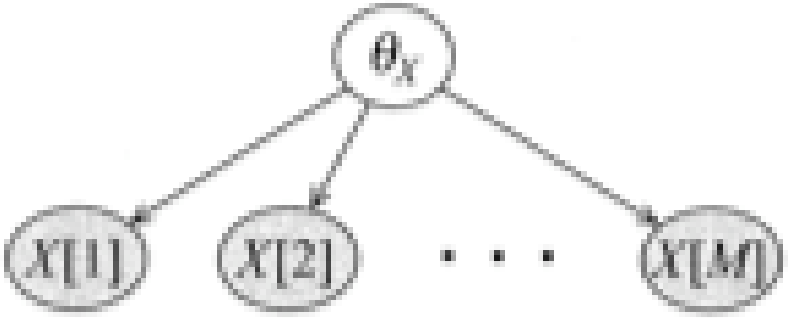
This joint distribution captures our assumptions about the experiment.

As long as we don‘t know $\theta$, the tosses are not marginally independent because each toss tells us something about $\theta$.

One $\theta$ is known, we assume that the tosses are conditionally independent given $\theta$.

# Joint probabilistic model

We can describe these assumptions using the probabilistic model below.



# Joint probabilistic model

Having determined the model structure,
it remains to specify the local probability models in this network.

We begin by considering the probability P(X[m] | $\theta$) :

$$P(x[m] \mid \theta) = \begin{cases} \theta & if\ x[m] = x^1 \\ 1-\theta & if\ x[m] = x^0 \end{cases}$$

We also need to describe the **prior distribution** over $\theta$ , P($\theta$).

This is a continuous density over the interval [0,1].
There are several possible choices for this.

Let us first consider how to use it.

# Joint probabilistic model

The network structure implies that the joint distribution of a particular data set and $\theta$ factorizes as

$$P(x[1], \ldots, x[M], \theta) = P(x[1], \ldots, x[M]|\theta)P(\theta)$$

$$= P(\theta) \prod_{m=1}^{M} P(x[m]|\theta)$$

$$= P(\theta)\theta^{M[1]}(1-\theta)^{M[0]}$$

where M[1] is the number of heads in the data,

M[0] is the number of tails, and

P( x[1], …, x[M] | $\theta$) is simply the likelihood function L($\theta$ : D).

This network specifies a joint probability model over parameters and data.

# Posterior distribution

There are several ways in which we can use this network.

For example, we can take an observed data set D of M outcomes,
and use it to instantiate the values of x[1], …, x[M].

We can then compute the **posterior distribution** over θ:

$$P(\theta|x[1],\ldots,x[M]) = \frac{P(x[1],\ldots,x[M]|\theta)P(\theta)}{P(x[1],\ldots,x[M])}$$

The first term in the numerator is the likelihood,
the second term is the prior over the parameters.

The denominator is a normalizing factor
so that the product is a proper density function [0,1].

# Prediction

Let us consider the value of the next coin toss x[M+1]
given the observations of the first M tosses.
Since θ is unknown, we will consider all its possible values and integrate over them

$$P(x[M+1] \mid x[1], \ldots, x[M])$$

$$= \int P(x[M+1] \mid \theta, x[1], \ldots, x[M]) P(\theta \mid x[1], \ldots, x[M])\ d\theta$$

$$= \int P(x[M+1] \mid \theta)\, P(\theta \mid x[1], \ldots, x[M]) d\theta$$

When going from the second to the third line,
we used the conditional indepencies implied by the meta-network.

→ we are **integrating** the posterior over θ
to predict the probability of heads for the next toss.

# Prediction: revisit thumbtack example

Assume that our prior is uniform (constant) over $\theta$ in the interval [0,1].

Then $P(\theta|x[1],\ldots,x[M])$ is proportional to the likelihood $P(x[1],\ldots,x[M] \mid \theta) = \theta^{M[1]}(1-\theta)^{M[0]}$ .

Plugging this into the integral, we need to compute

$$P(X[M+1] = x^1|x[1],\ldots,x[M]) = \frac{1}{P(x[1],\ldots,x[M])}\int \theta \cdot \theta^{M[1]}(1-\theta)^{M[0]}d\theta$$

$$= \cdots = \frac{M[1]+1}{M[1]+M[0]+1}$$

This so-called **Bayesian estimator** is quite similar to the MLE prediction except that it adds one „imaginary“ sample to each count.

# Priors: Beta distribution

When using nonuniform priors, the challenge is to pick
a continuous distribution that can be written in a compact form
(e.g. using an analytical formula),
and that can be updated efficiently as we get new data.

An appropriate prior is the **Beta distribution**.

Definition: a Beta distribution is parametrized by two real and positive **hyperparameters** $\alpha_1$, $\alpha_0$ and defined as:

$$\theta \sim Beta(\alpha_1, \alpha_0) \text{ if } p(\theta) = \gamma \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

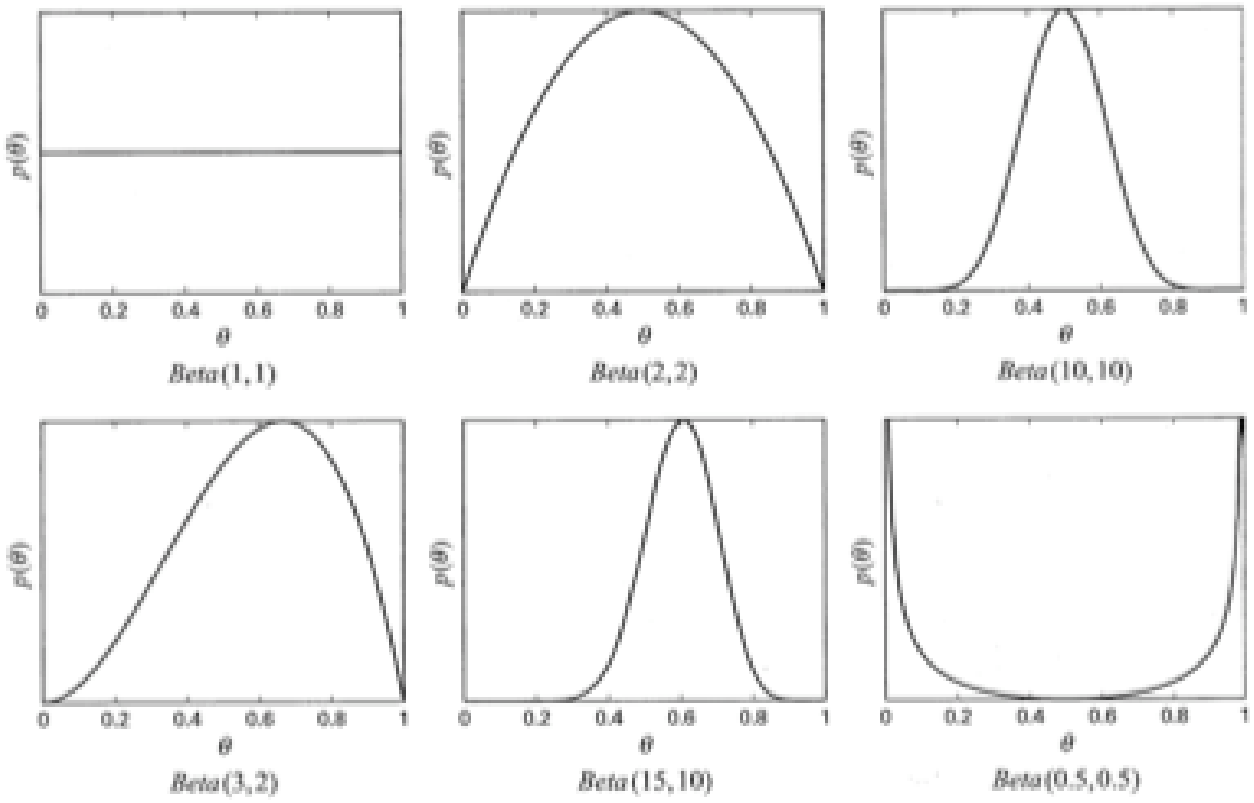
The normalization constant is defined as: $\gamma = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)}$

where $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ is the **Gamma function**.

# Beta distribution

The parameters $\alpha_1$ and $\alpha_0$ correspond intuitively to the number of imaginary heads and tails that we have „seen“ before starting the experiment.

These are examples of beta functions



# Gamma function

The Gamma function is simply a continuous generalization of factorials.

It satisfies $\Gamma(1) = 1$ and $\Gamma(x + 1) = x\,\Gamma(x)$.

Hence $\Gamma(n + 1) = n!$

Beta distributions have properties
that make them particularly useful for parameter estimation.

Assume our distribution $P(\theta)$ is *Beta*$(\alpha_1,\alpha_0)$ and consider a single coin toss X.

Let us compute the marginal probability over X, based on $P(\theta)$.

We need to integrate out $\theta$.

# Properties of Beta functions

$$P(X[1] = x^1) = \int_0^1 P(X[1] = x^1|\theta) \cdot P(\theta) d\theta$$

$$= \int_0^1 \theta \cdot P(\theta) d\theta$$

$$= \int_0^1 \theta \cdot \gamma \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_0 - 1} d\theta$$

$$= \int_0^1 \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \theta^{\alpha_1} (1-\theta)^{\alpha_0 - 1} d\theta$$

$$= \cdots = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

This finding matches our inituition that the Beta prior indicates that we have seen $\alpha_1$ (imaginary) heads and $\alpha_0$ (imaginary) tails.

# Properties of Beta distributions

As we get more observations, i.e. M[1] heads and M[0] tails it follows that

$$P(\theta|x[1],\ldots,x[M]) \propto P(x[1],\ldots,x[M]|\theta)P(\theta)$$
$$\propto \theta^{M[1]}(1-\theta)^{M[0]} \cdot \theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}$$
$$= \theta^{\alpha_1+M[1]-1}(1-\theta)^{\alpha_0+M[0]-1}$$

which is precisely *Beta*($\alpha_1$+ M[1], $\alpha_0$+ M[0]).

This result illustrates a key property of the Beta distribution:

If the prior is a Beta distribution, then the posterior distribution,
that is, the prior conditioned on the evidence, is also a Beta distribution.

# Priors

An immediate consequence is that we can compute
the probabilities over the next toss:

$$P(X[M+1] = x^1 | x[1], \ldots, x[M]) = \frac{\alpha_1 + M[1]}{\alpha + M}$$

where $\alpha = \alpha_1 + \alpha_0$ and $M = M_1 + M_0$

In this case, our posterior Beta distribution tells us
that we have seen $\alpha_1$ + M[1] (imaginary) heads and $\alpha_0$ + M[0] tails.

# Effect of Priors

Let us compare the effect of *Beta*(2,2) vs. *Beta*(10,10)
on the probability over the next coin toss.

Both priors predict that the probability of heads in the first toss is $\frac{\alpha_1}{\alpha_1+\alpha_0} = 0.5$ .

How do different priors (*Beta*(10,10) is more narrow) affect further convergence?

Suppose we observe 3 heads in 10 tosses.

Using the first prior, our estimate is $\frac{2+3}{4+10} \approx 0.36$

Using the second prior gives $\frac{10+3}{20+10} \approx 0.43$

But when we obtain much more data, the effect of the prior almost disappears.
If we obtain 1000 tosses of which 300 are heads,

both $\frac{2+300}{4+1000}$ and $\frac{10+300}{20+1000}$ give values close to 0.3

# Priors and Posteriors

Let us assume a general learning problem
where we observe a training set D that contains M IID samples
of a random variable X from an unknown distribution P*(X).

We also assume that we have a parametric model $P(\xi \mid \theta)$
where we can choose parameters from a parameter space $\Theta$.

The MLE approach attempted to find the parameters $\hat{\theta}$ in $\Theta$
that are „best“ given the data.

The Bayesian approach, on the other hand, does not attempt
to find a single best estimate.

Instead, one quantifies the subjective probability for different
values of $\theta$ after seeing the evidence.

# Priors and Posteriors

We need to describe a joint distribution P(D, $\theta$) over the data and the parameters.

We can easily write

$$P(D, \theta) = P(D|\theta)P(\theta)$$

The first term on the right is the **likelihood function**
(see V8 – example on predicting PP complexes).

The second term is the **prior distribution** over the possible values in $\Theta$.

It captures our initial uncertainty about the parameters.

It can also capture our previous experience before we start the experiment.

# Priors and Posteriors

Once we have specified the likelihood function and the prior, we can use the data to derive the **posterior distribution** over the parameters using Bayes rule:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

The term P(D) is the **marginal likelihood** of the data

$$P(D) = \int_{\Theta} P(D|\theta)P(\theta)d\theta$$

what is the integration of the likelihood over all possible parameter assignments.

# Priors and Posteriors

Let us reconsider the example of a multinomial distribution (MD).

We need to describe our uncertainty about the parameters of MD.

The parameter space contains all nonnegative vectors $\Theta = \langle \theta_1, \ldots, \theta_K \rangle$ such that $\sum_k \theta_k = 1$.

As we saw previously, the likelihood function is $L(\theta : D) = \prod_k {\theta_k}^{M[k]}$

Since the posterior is a product of the prior and the likelihood, it is natural to require that the prior also have a form similar to the likelihood.

One such prior is the **Dirichlet distribution** which generalizes the Beta distribution.

# Dirichlet distribution

A Dirichlet distribution is specified by a set of hyperparameters $\alpha_1, \ldots \alpha_K$ so that

$$\theta \sim Dirichlet(\alpha_1, \ldots, \alpha_K) \text{ if } P(\theta) \propto \prod_k \theta_k^{\alpha_k - 1}$$

We use $\alpha$ to denote $\sum_j \alpha_j$.

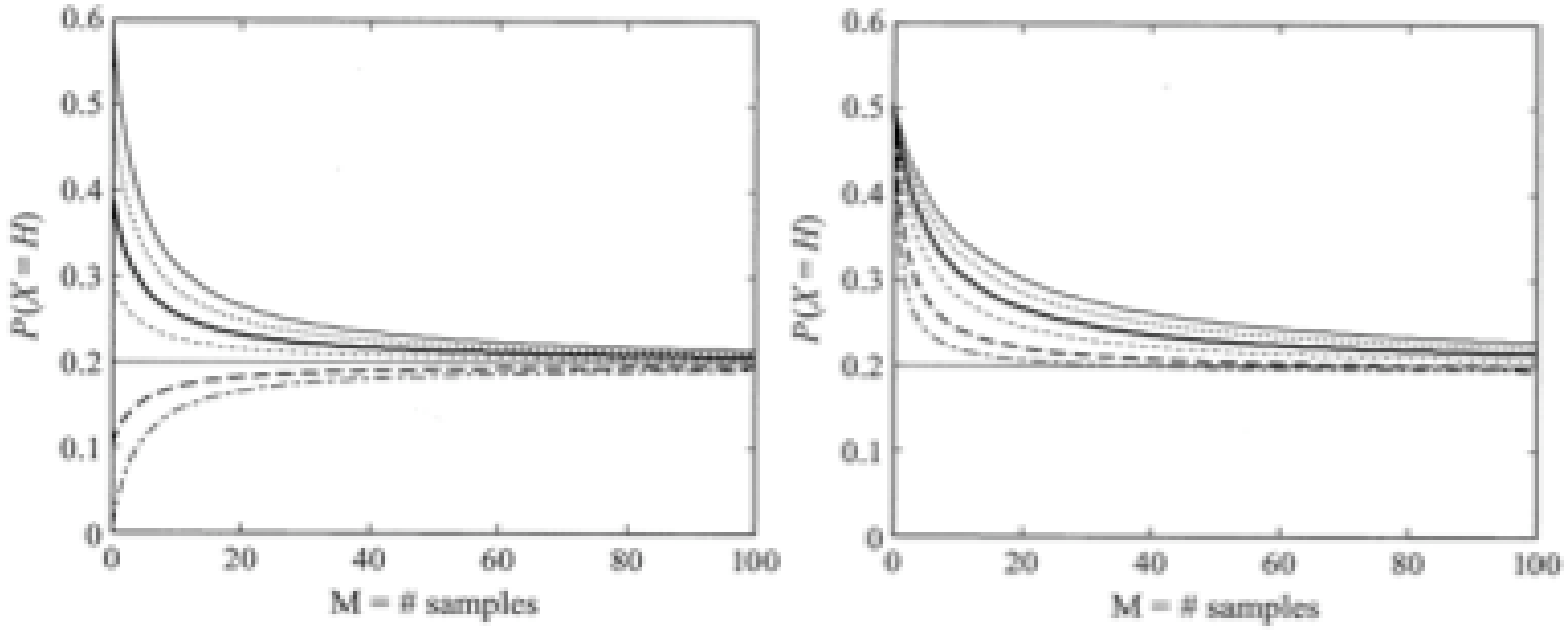
If we use a Dirichlet prior, then the posterior is also Dirichlet:

Proposition: If P($\theta$) is $Dirichlet(\alpha_1, \ldots, \alpha_K)$ then P($\theta$ | D) is $Dirichlet(\alpha_1 + M[1], \ldots, \alpha_K + M[K])$,
where M[K] is the number of occurrences of $x^k$.

Priors such as the Dirichlet are useful since they ensure that the posterior has a nice compact description and uses the same representation as the prior.

We will see on 2 examples the effects of priors on posterior estimates.

# Effect of Beta prior on convergence of posterior estimates

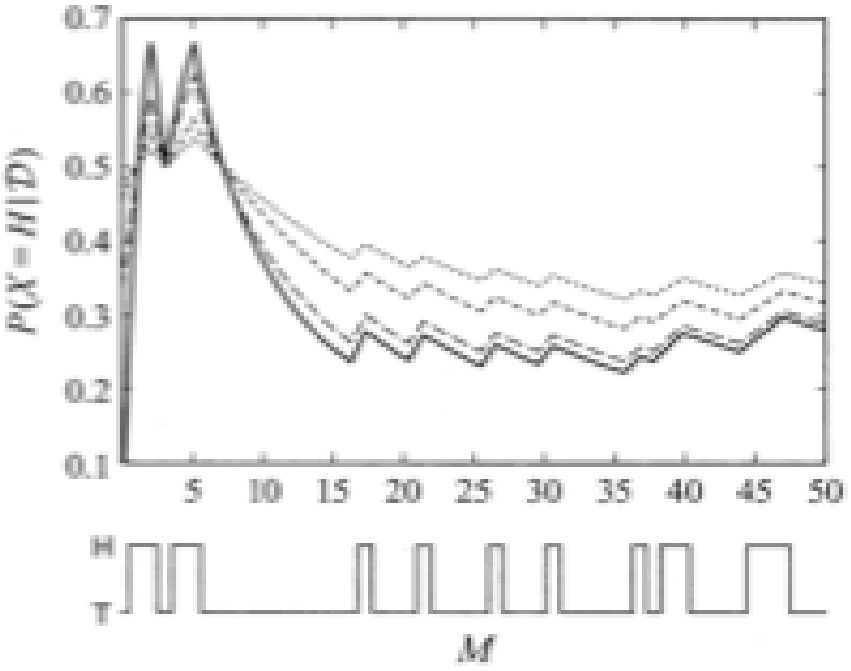



For a given data set size M, we assume that D contains 0.2 M heads and 0.8 M tails. As the amount of real data grows, our estimate converges to the true underlying distribution, regardless of the starting point.

(Left): effect of varying prior means $\theta_1'$, $\theta_0'$ for a fixed prior strength $\alpha$.

(Right): effect of varying prior strength for a fixed prior mean $\theta_1' = \theta_0' = 0.5$

# Convergence of parameter estimate




Dotted line: Beta(10,10)
Small-dash line: Beta(5,5)
Large-dash line: Beta (1,1)

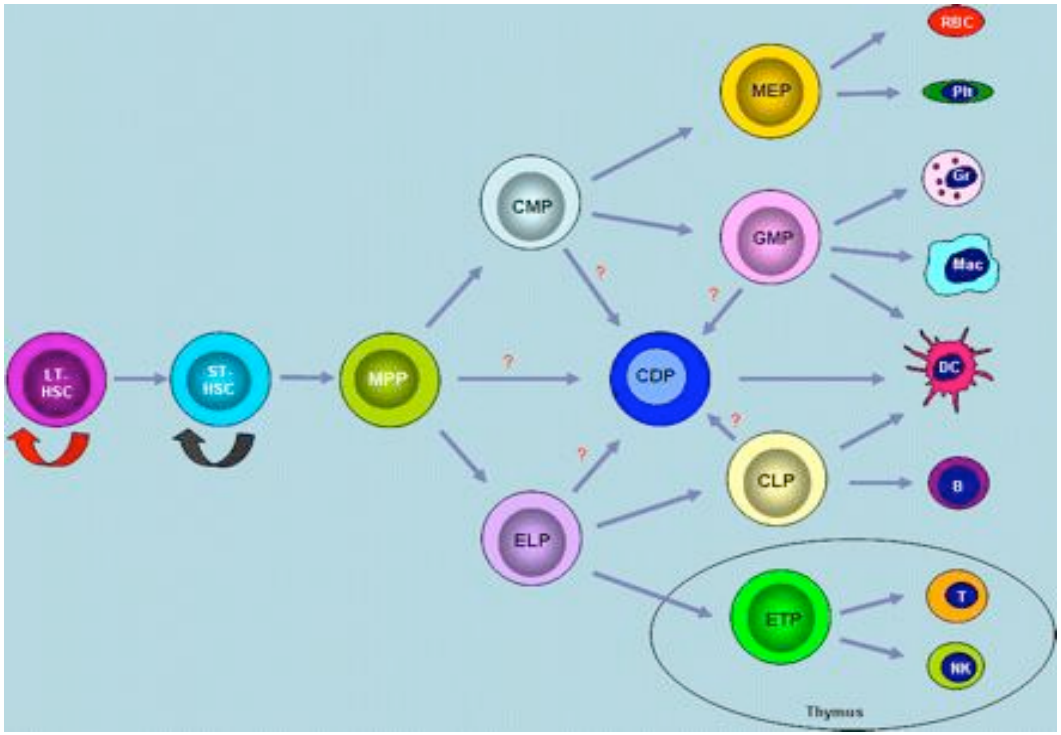
→ Beta(10,10) has longer „memory“ about initial conditions

Effect of different priors on smoothing the parameter estimates.

Below the graph is shown the particular sequence of tosses.
Solid line: MLE estimate
Dashed lines: Bayesian estimates with different strengths and uniform prior means.

# Imprinting effects during hematopoietic differentiation?

- One of the most well studied developmental systems
- Mature cell line models




Rathinam and Flavell 2008

Mohamed Hamed (unpublished)

# Blood lineages

| Lineage | Developmental cell line |
|---|---|
| B cell | *LTHSC → STHSC → LMPP→GMLP→CLP→PREPROB→BLP→MZB* |
| NK cell | *LTHSC → STHSC → LMPP→GMLP→CLP→MNK* |
| T cell | *LTHSC → STHSC → LMPP→CLP→ETP→THY-DP→THY-SP→TCELL* |
| Erythrocytes | *LTHSC → STHSC → MPP→CMP →PMEP →MEP →PROE→ ERY* |
| Monocytes | *LTHSC → STHSC → MPP→CMP→PGMP→GMP→MONO→MACRO or DENDRITIC* |
| Megakaryocytes | *LTHSC → STHSC → MPP→CMP→PMEP→MEP→MKP→MKE* |

Mohamed Hamed
(unpublished)

# Motivation I

- Identify cellular events that drive cell differentiation and reprogramming

- Construct gene-regulatory network (GRN) that governs
  - transitions between the different states along the developmental cell lines and
  - pausing at specific states.

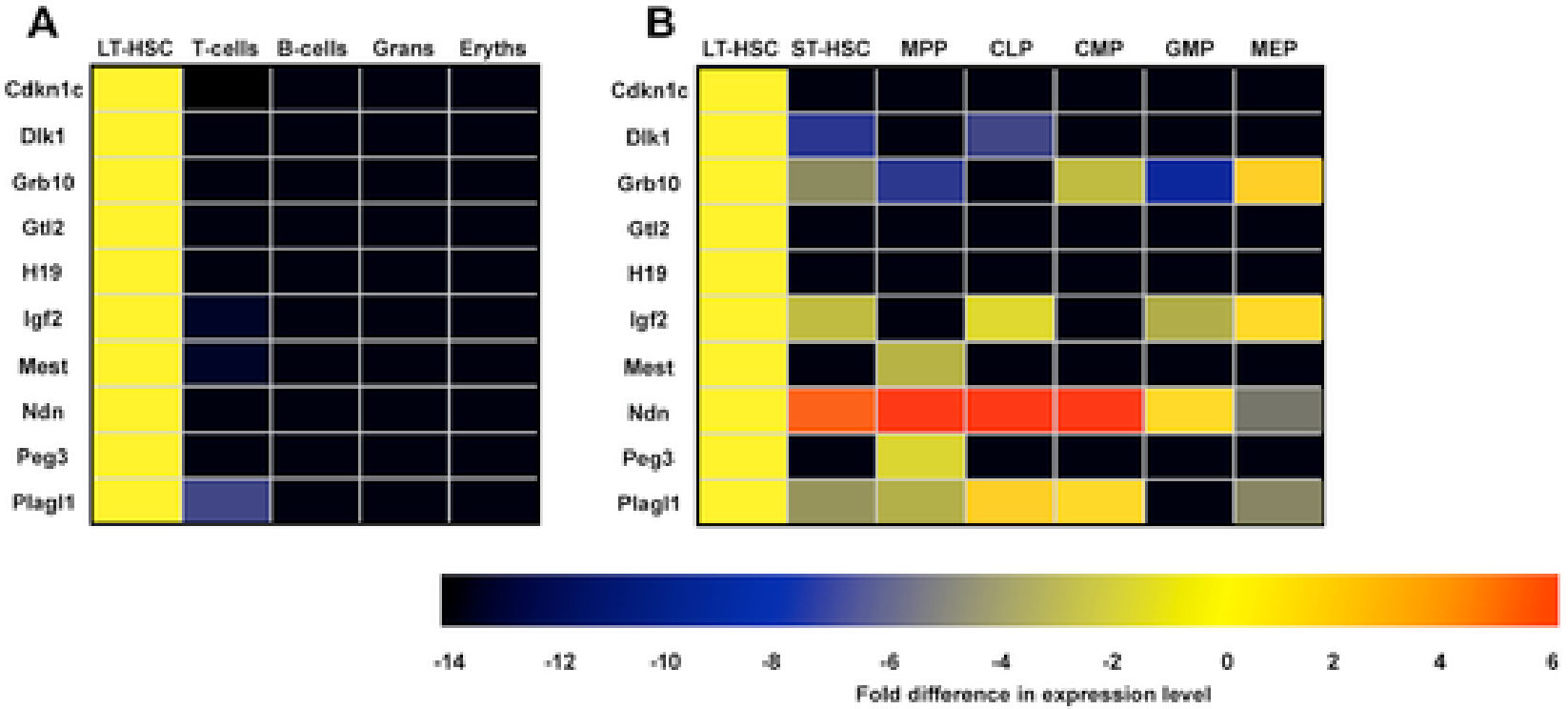
- Do **imprinted genes** play a role in regulating differentiation?.

Mohamed Hamed
(unpublished)

# Motivation II

Berg, Lin et al. (2011)

**Real-time PCR analysis of imprinted gene expression in hematopoietic cells**




Imprinted genes drastically down-regulated in differentiated cells.

during the earliest phases of hematopoietic development, imprinted genes may have distinct roles

Mohamed Hamed (unpublished)

# Imprinted genes

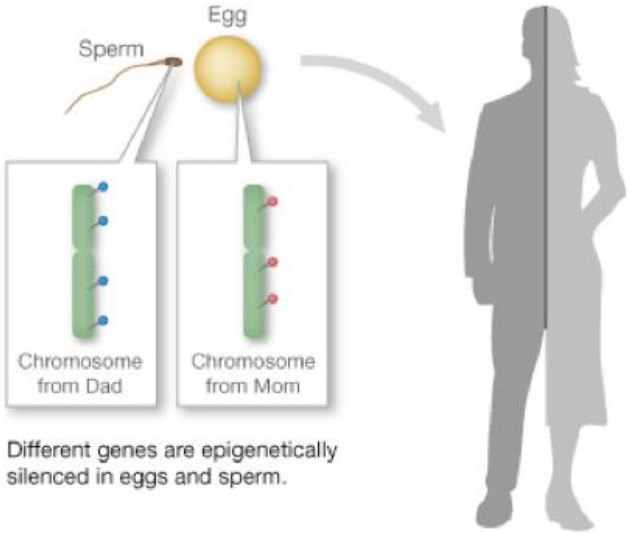
- violate the usual rule of inheritance
- **bi-allelic** genes :

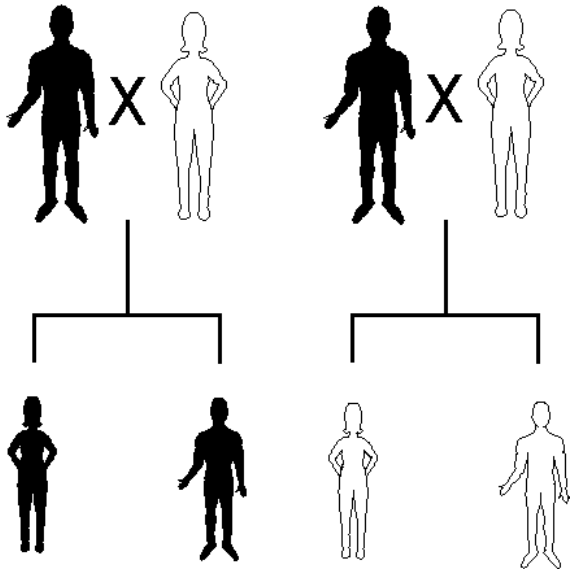
  gene copy (allele) encoding hemoglobin from dad

  gene copy (allele) encoding hemoglobin from mom

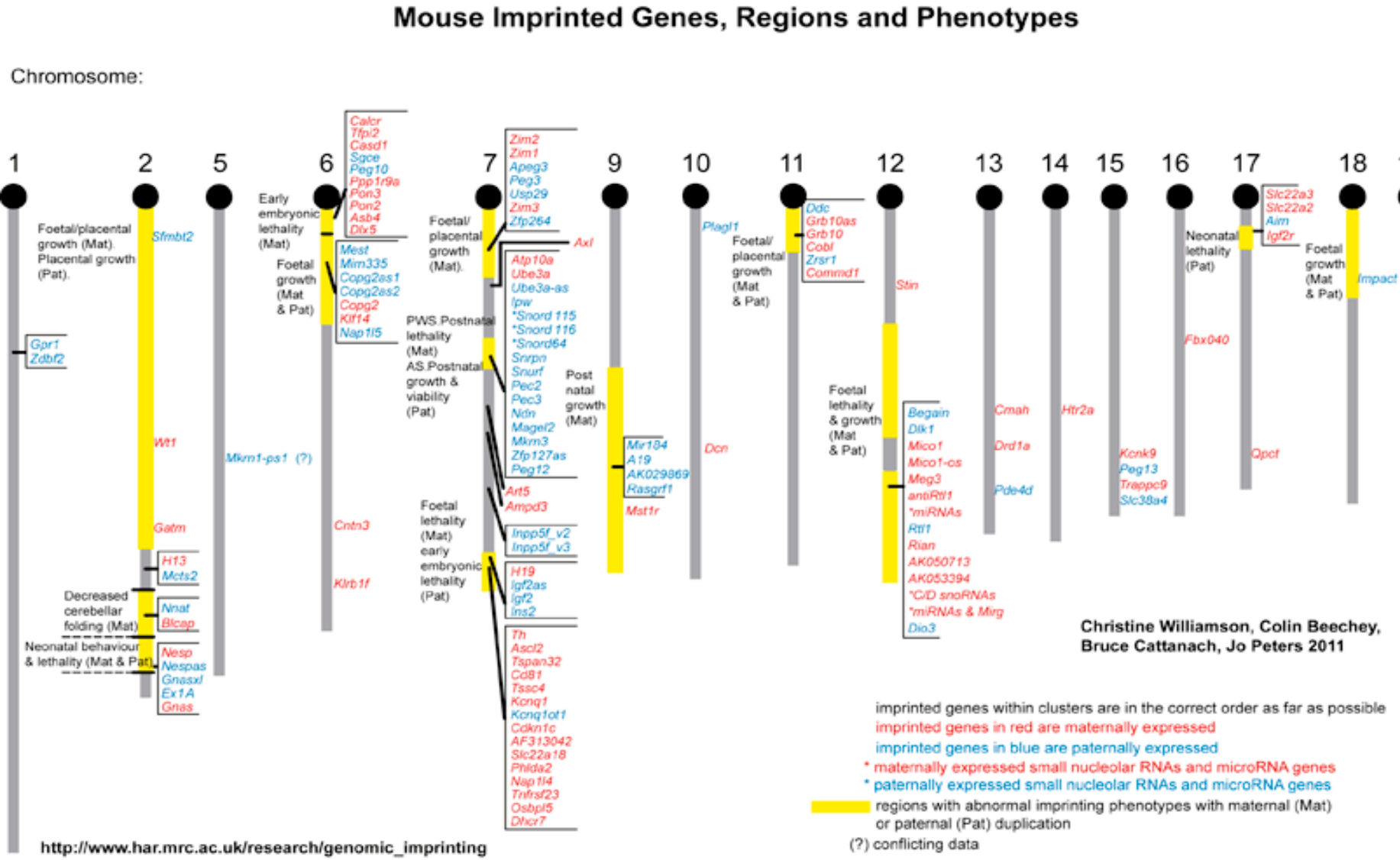
  Child: expresses equal amounts of the 2 types of hemoglobin

- **mono-allelic** (imprinted) genes : one allele silenced by **DNA methylation**



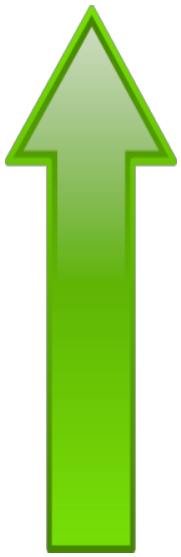



# Imprinted genes cluster in the genome



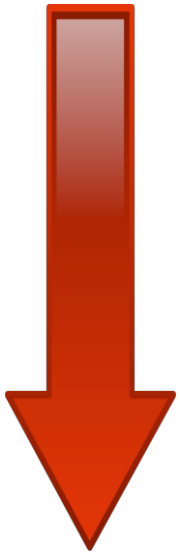
# Parental conflict hypothesis = “battle of the sexes”

Paternally expressed genes genes



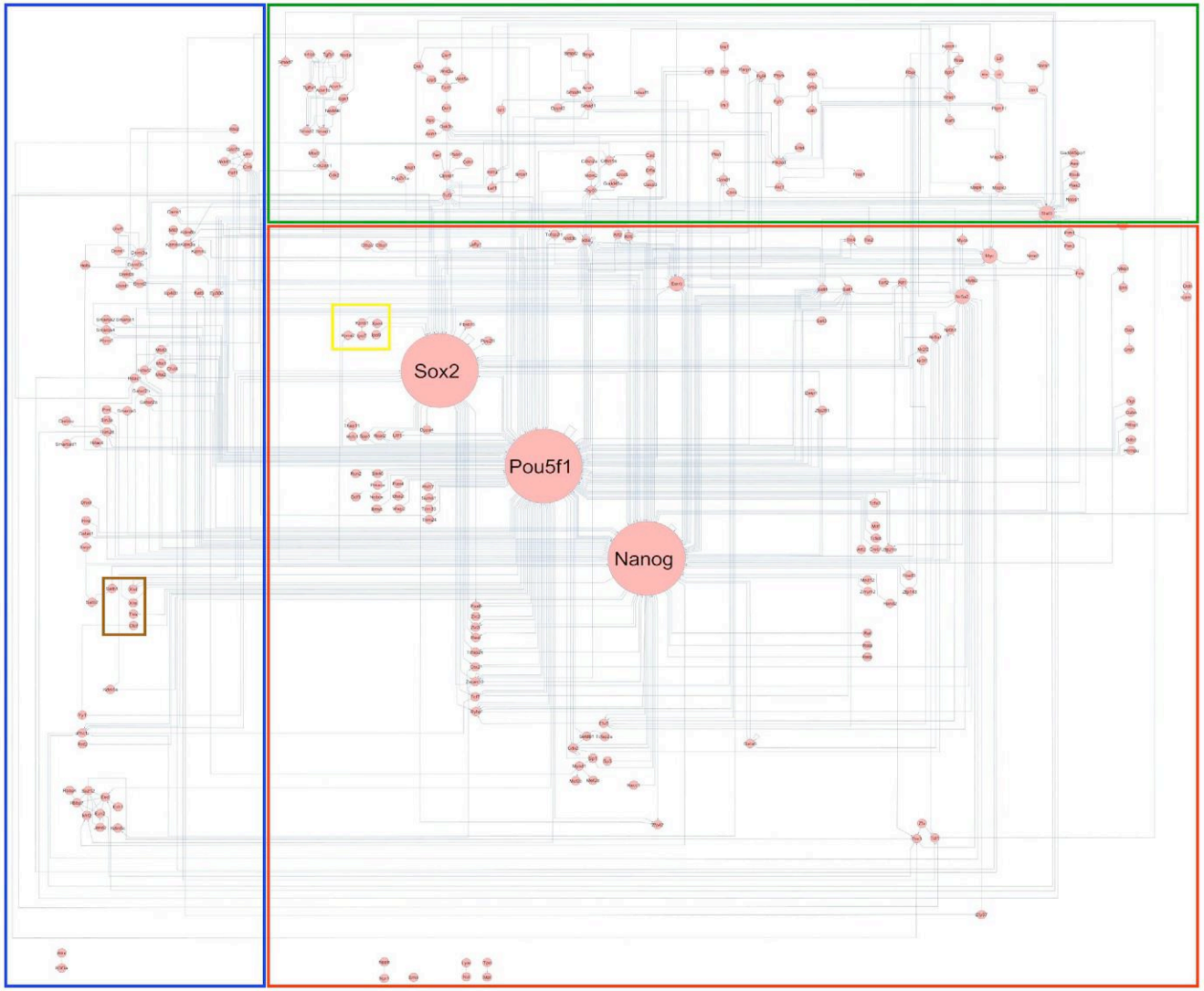
embryonic growth in placenta

Maternally expressed



embryonicg growth in placenta

# Mouse Pluripotency network (Plurinet)




Pluripotency network in mouse

G. Fuellen et al. (2010)

based on 177 publications

274 genes
574 stimulations / inhibitions/ and interactions

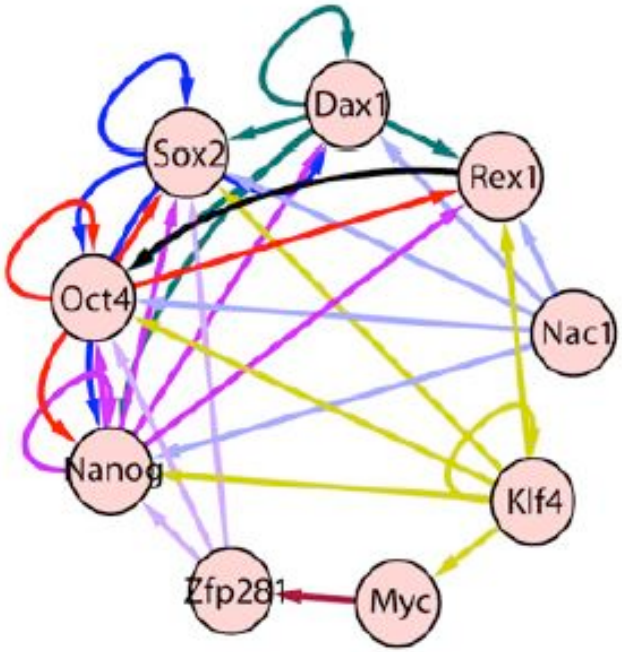
# Gene regulatory network around Oc4 controls pluripotency

Tightly interwoven network of 9 transcription factors keeps ES cells in pluripotent state.

6632 human genes have binding site in their promoter region for at least one of these 9 TFs.

Many genes have multiple motifs.

800 genes bind ≥ 4 TFs.



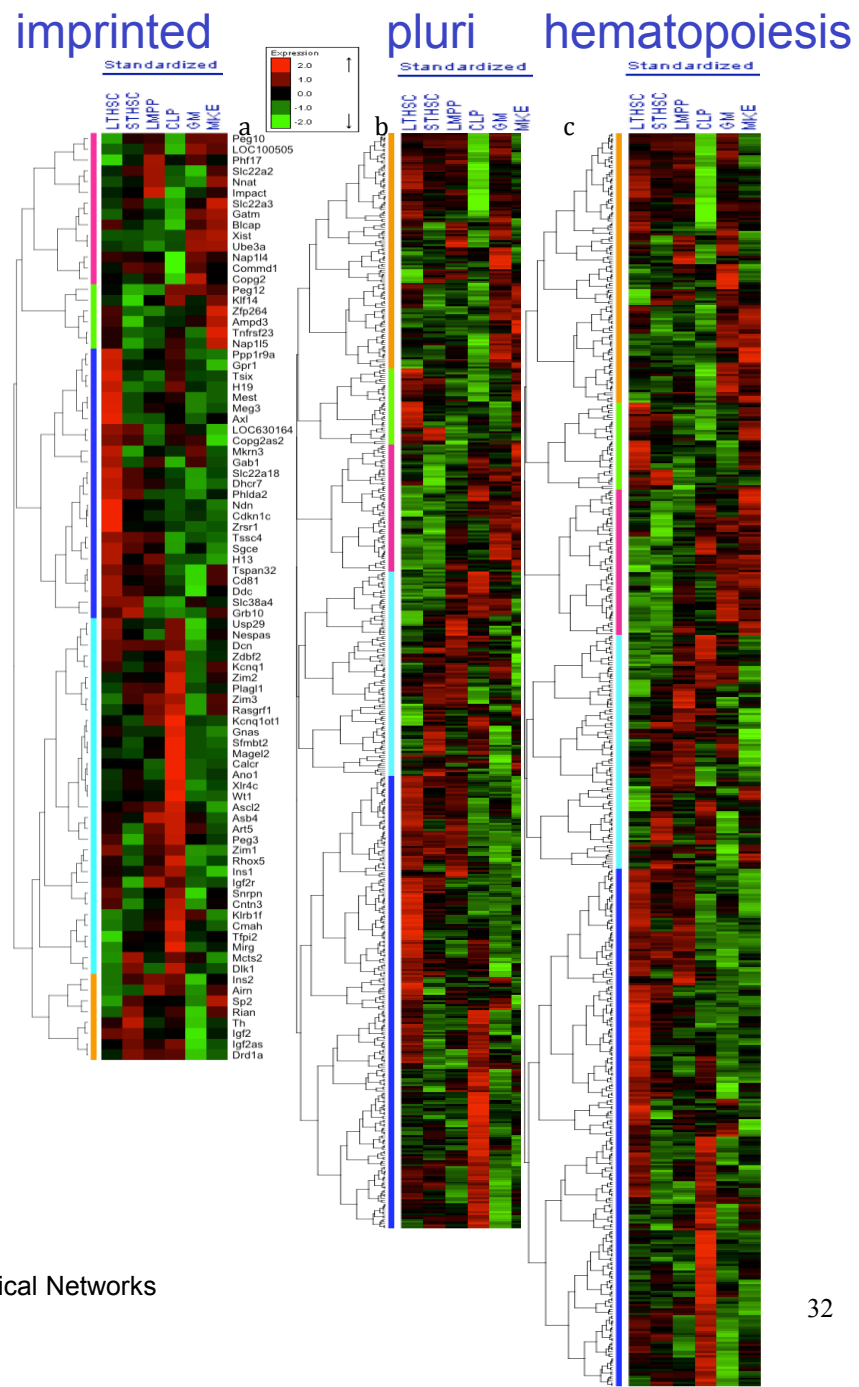
# Gene expression profiles

(1) long and short-term hematopoietic stem cells

(2) Intermediate progenitor populations such as Lymphoid primed multipotent progenitor (*LMPP*), common lymphoid progenitor (CLP), and granulocyte–monocyte progenitor (*GMP*), and

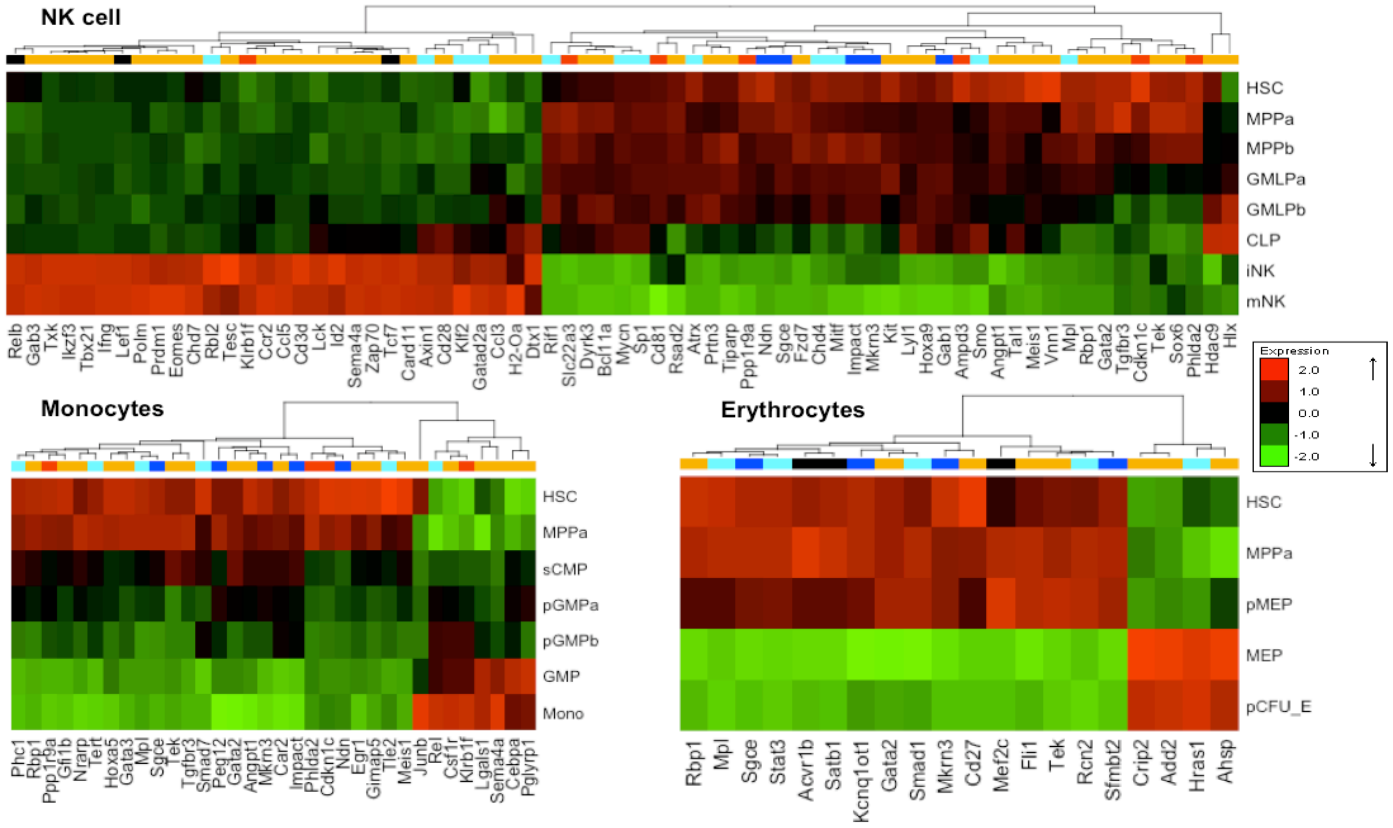
(3) Terminally differentiated blood progeny such as *NK cells* and granulocyte- monocyte (*GM*).

All 3 gene sets contain genes that are upregulated either in (1), (2) or (3) stages




Mohamed Hamed
(unpublished)

# Lineage-specific marker genes from all 3 gene sets cluster together




**red** : maternally expressed imprinted genes

**blue** : paternally expressed imprinted genes

**cyan** : pluripotency genes

**orange**: hematopoietic genes

Mohamed Hamed (unpublished)

# Imprinted gene network (IGN)

## Aim: explain surprisingly similar expression profiles of 3 gene sets

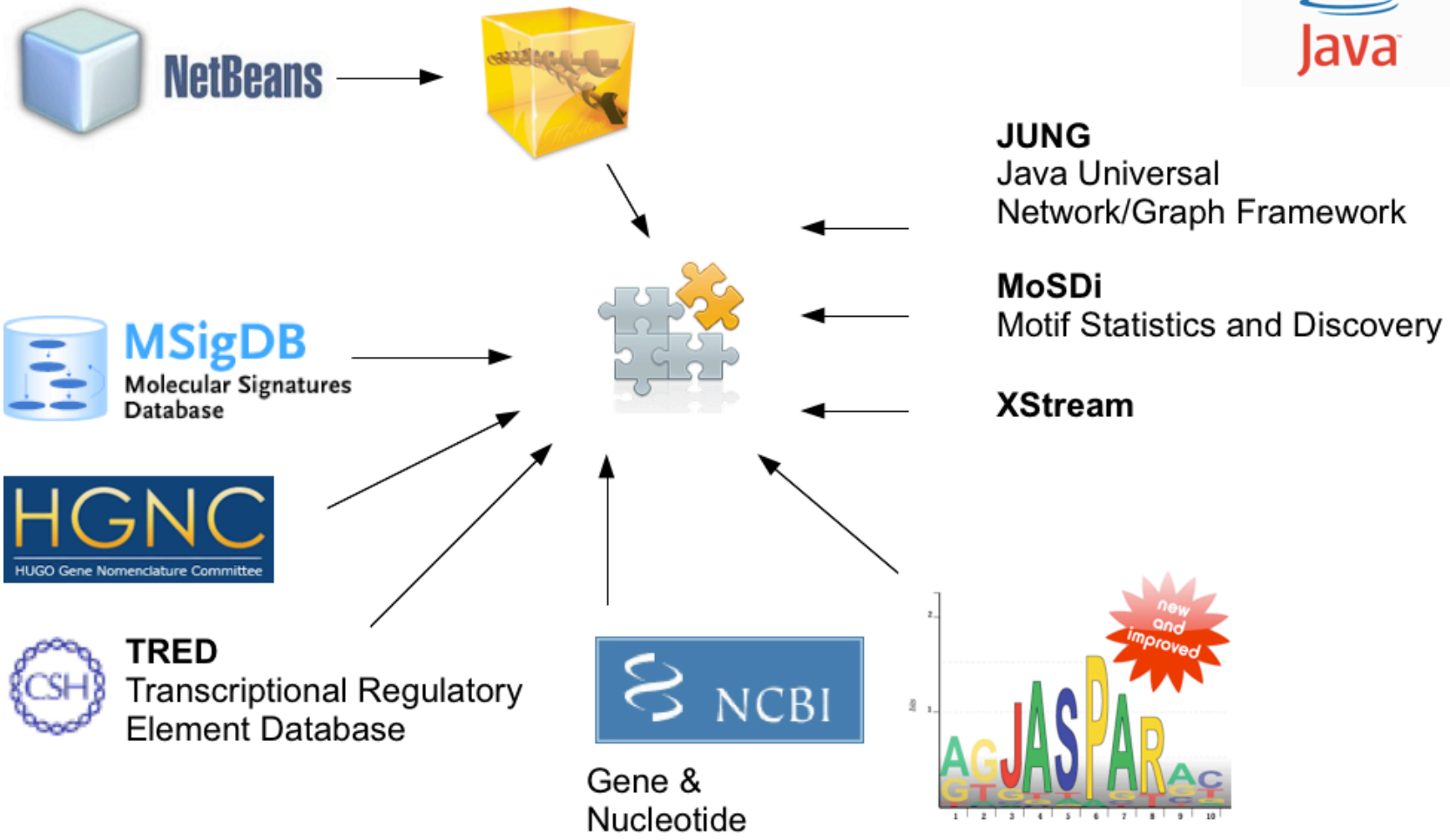
- only 5 imprinted genes (Gab1, Ins1, Phf17, Tsix, and Xist) are present in the pluripotency list and
- only 3 imprinted genes (Axl, Calcr, and Gnas) belong to the hematopoietic list.

**Who regulates the imprinted genes?**

- Identify regulators (TFs) of imprinted genes and target genes regulated by imprinted genes

Mohamed Hamed
(unpublished)

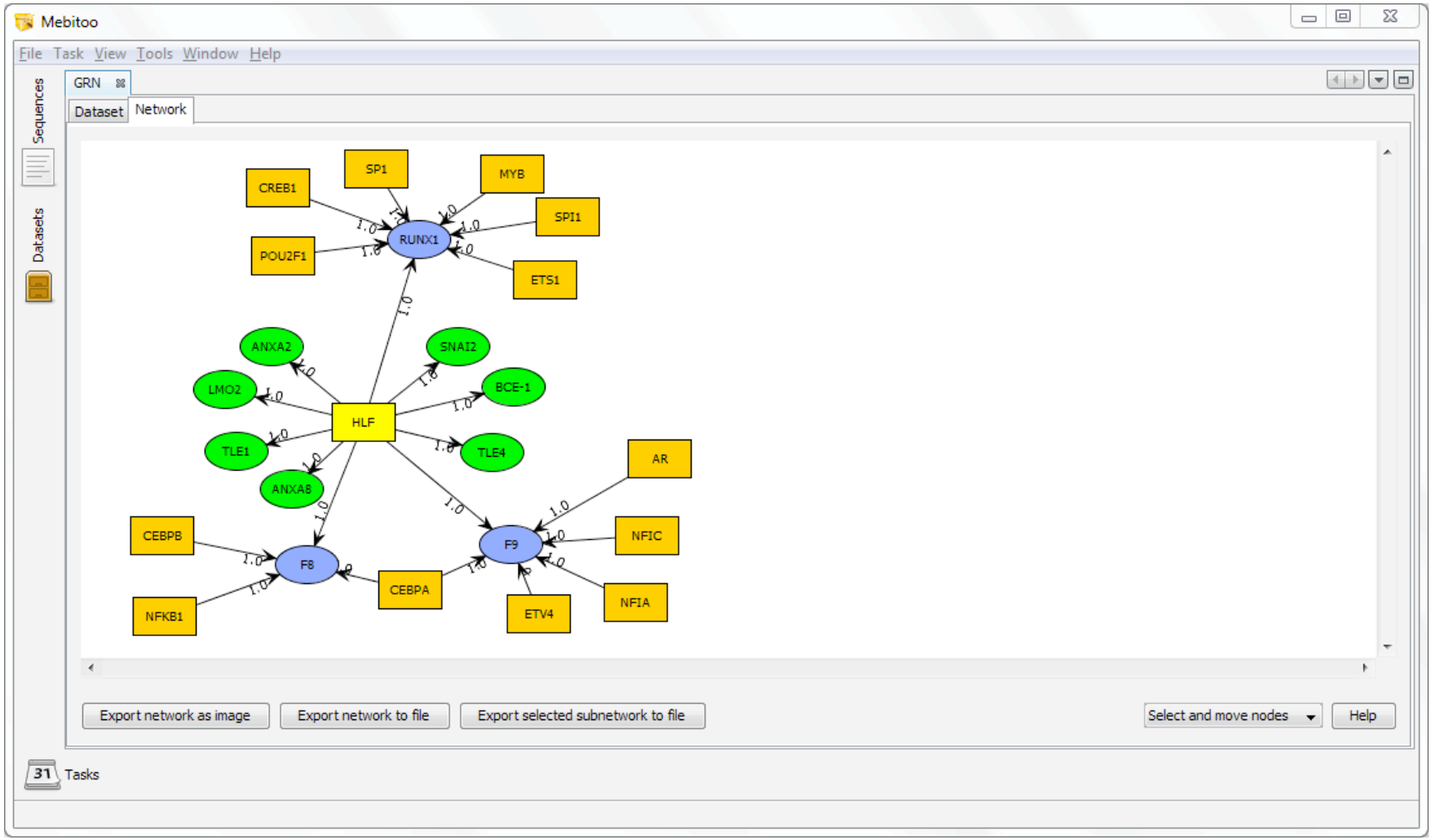
# Mebitoo GRN Plugin




Johannes Trumm, MSc thesis ,CBI, 2011.

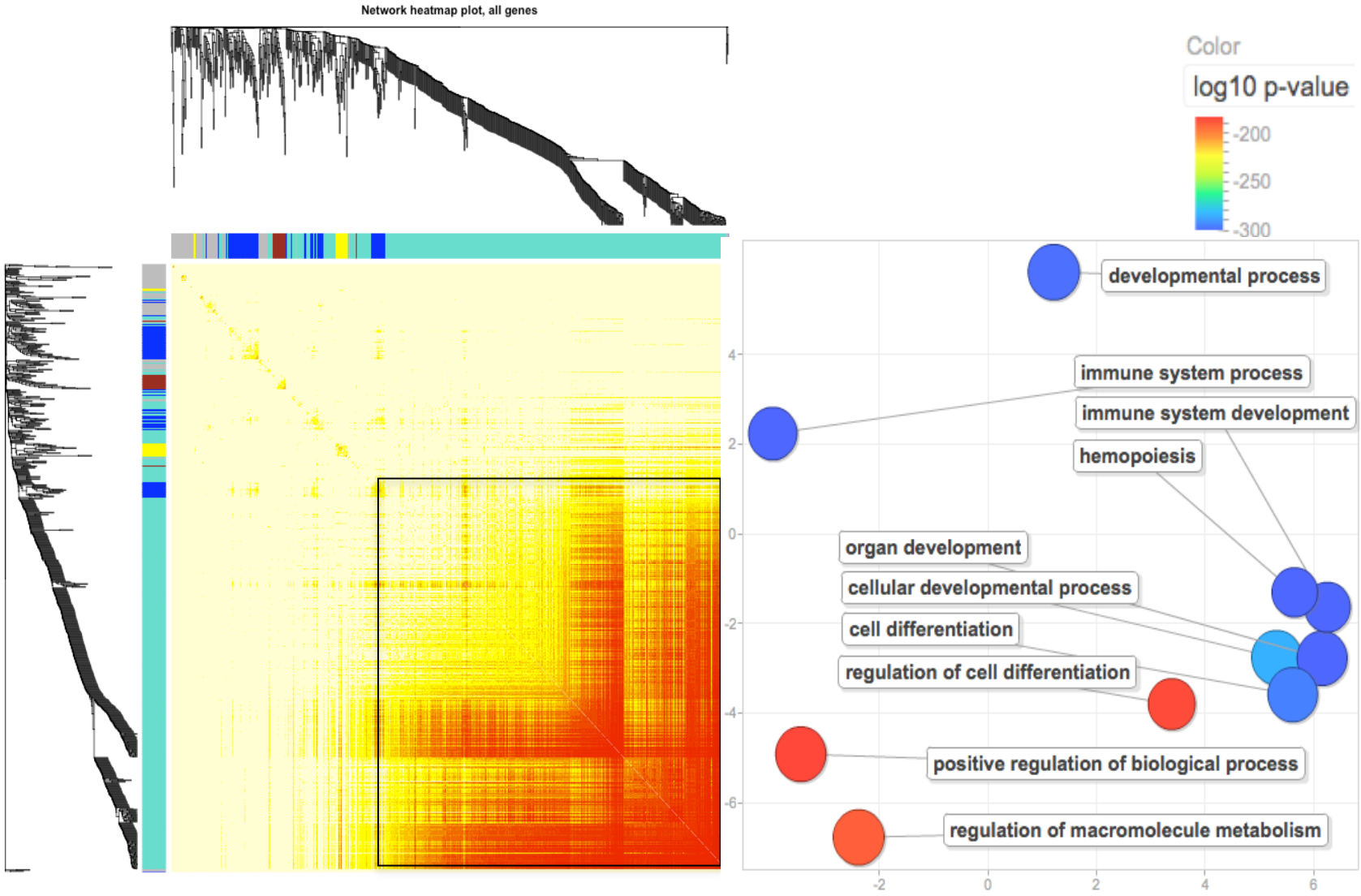
Mohamed Hamed
(unpublished)

# Mebitoo GRN Plugin




Johannes Trumm, MSc thesis ,CBI, 2011.

# gene sets are (largely) co-expressed and enriched with developmental GO terms




Mohamed Hamed (unpublished)

# Summary

Parameter learning from data is an important research field.
We entered into some basics about MLE and Bayesian parameter estimation.

Powerful and efficient priors need to be estimated, see Beta function.

V11: enter into structure learning.

Application example: construct GRN to derive genes that drive hematopoiesis.

Intersection with pluripotency and imprinted genes reveals interesting module of co-expressed genes with homogenous involvement in development.