Latent variable models for bacterial transcriptome analysis : from data processing to biology

Pierre Nicolas

pierre.nicolas@jouy.inra.fr

Mathématique Informatique et Génome (MIG)
INRA Jouy-en-Josas

Journées MAS (SMAI)
Toulouse – August 28, 2014

Outline

0- A brief introduction to post-genome high-throughput biology

1- Discovering transcription units

Expression (read counts) profiles along genomes

Segmentation with State Space Models
Bogdan Mirauta (PhD thesis), co-advisor Hugues Richard (UPMC)

2- Partitioning the promoter space wrt RNAPol binding sites

DNA sequence

Mixture Models on trees and sequences

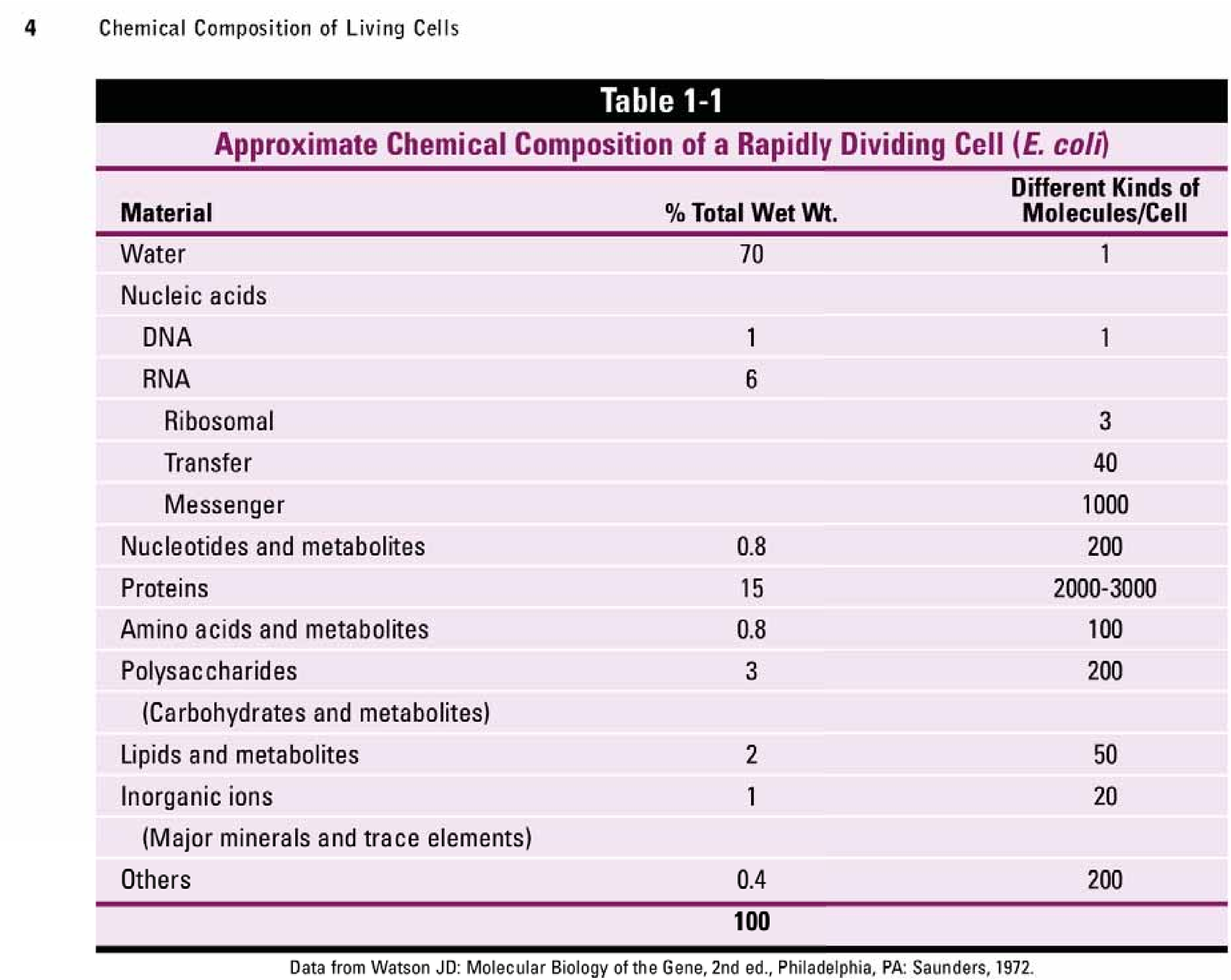
3- Assessing a model of transcriptional regulation

Expression profiles across conditions

Inference for a mechanistic model
joint work with Vincent Fromion (INRA)

Key wet-biologists partners (EU Basysbio and Basynthec projects):
Etienne Dervyn, Philippe Noirot (INRA), Ulrike Mäder (U. Greifswald).

0- A brief introduction to post-genome high-throughput biology



4 Chemical Composition of Living Cells

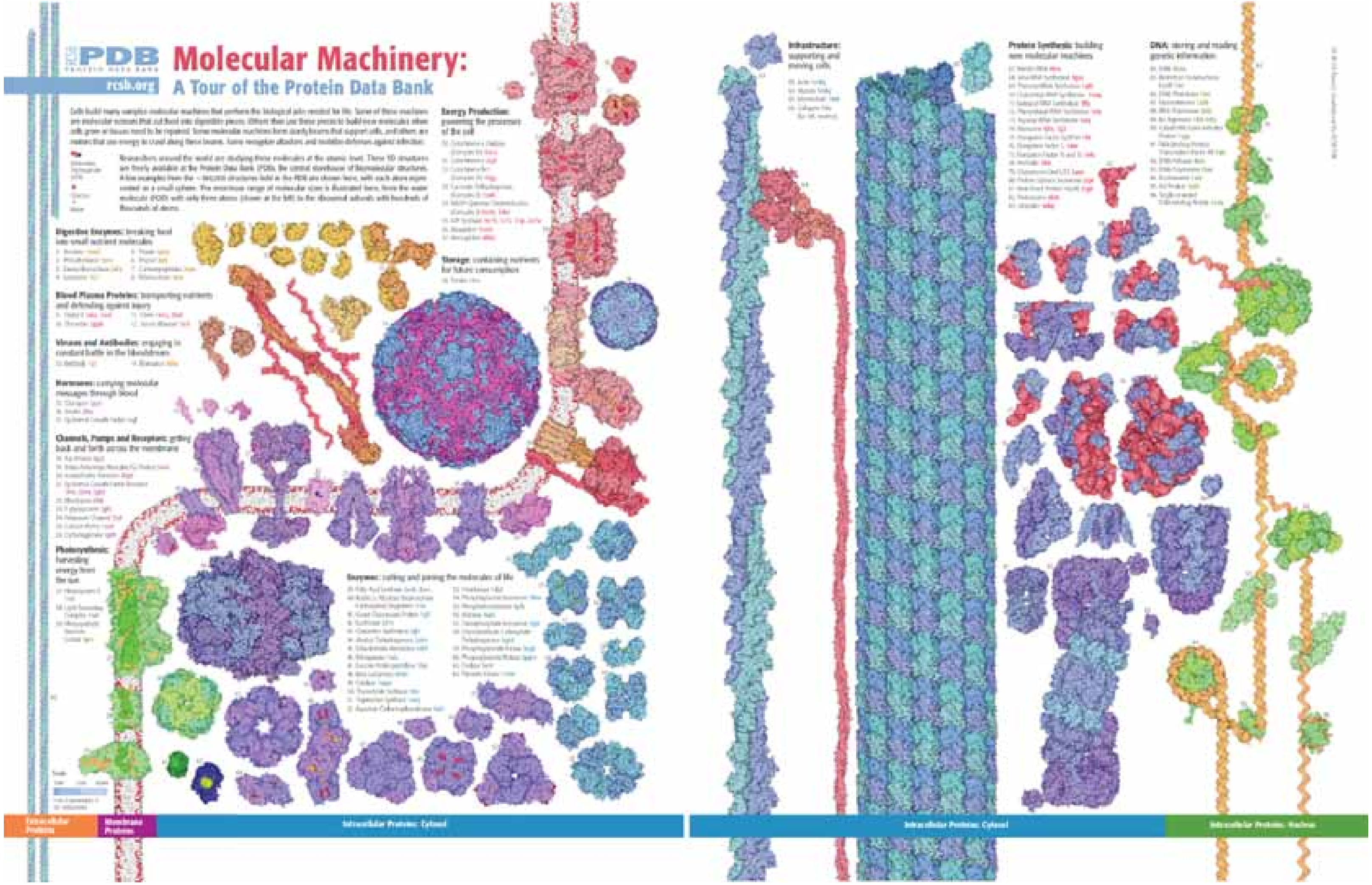
Table 1-1

Approximate Chemical Composition of a Rapidly Dividing Cell (*E. coli*)

Material	% Total Wet Wt.	Different Kinds of Molecules/Cell
Water	70	1
Nucleic acids		
DNA	1	1
RNA	6	
Ribosomal		3
Transfer		40
Messenger		1000
Nucleotides and metabolites	0.8	200
Proteins	15	2000-3000
Amino acids and metabolites	0.8	100
Polysaccharides	3	200
(Carbohydrates and metabolites)		
Lipids and metabolites	2	50
Inorganic ions	1	20
(Major minerals and trace elements)		
Others	0.4	200
	100	

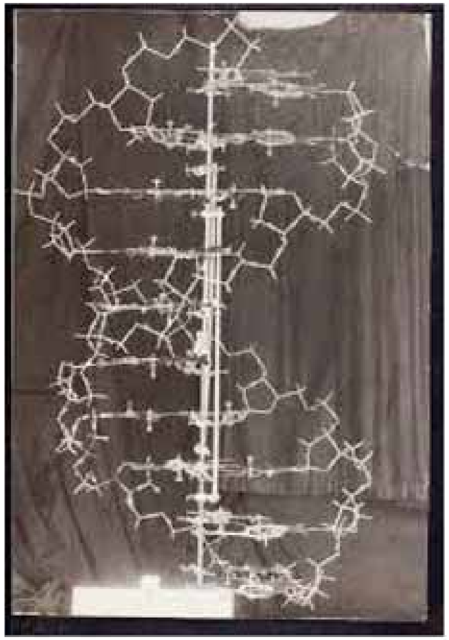
Data from Watson JD: Molecular Biology of the Gene, 2nd ed., Philadelphia, PA: Saunders, 1972.

Proteins: enzymes, cell signaling, structural proteins.



Protein function depends on 3D/4D structure that remains very difficult to assess. July 2014: 101,948 structures in the Protein Data Bank (vs. 173,353,076 nucleotide sequences in NCBI-GenBank).

Flow of genetic information within a biological system



Crick and Watson (1953)

NATURE VOL. 227 AUGUST 8 1970

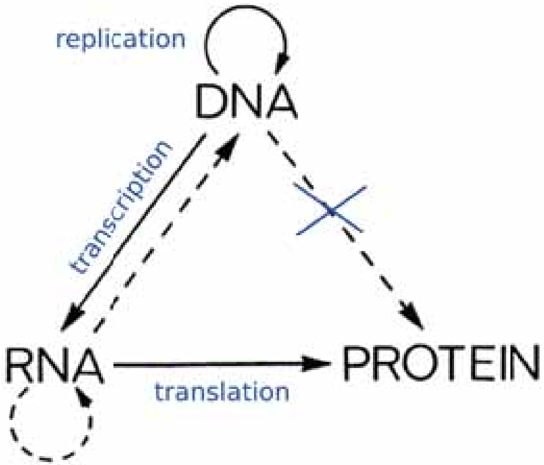
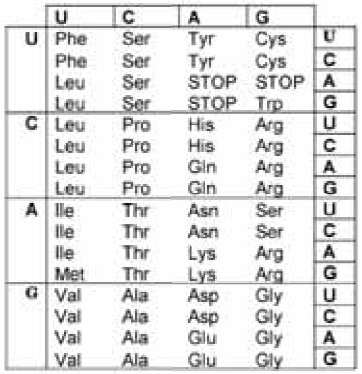


Fig. 3. A tentative classification for the present day. Solid arrows show general transfers; dotted arrows show special transfers. Again, the absent arrows are the undetected transfers specified by the central dogma.



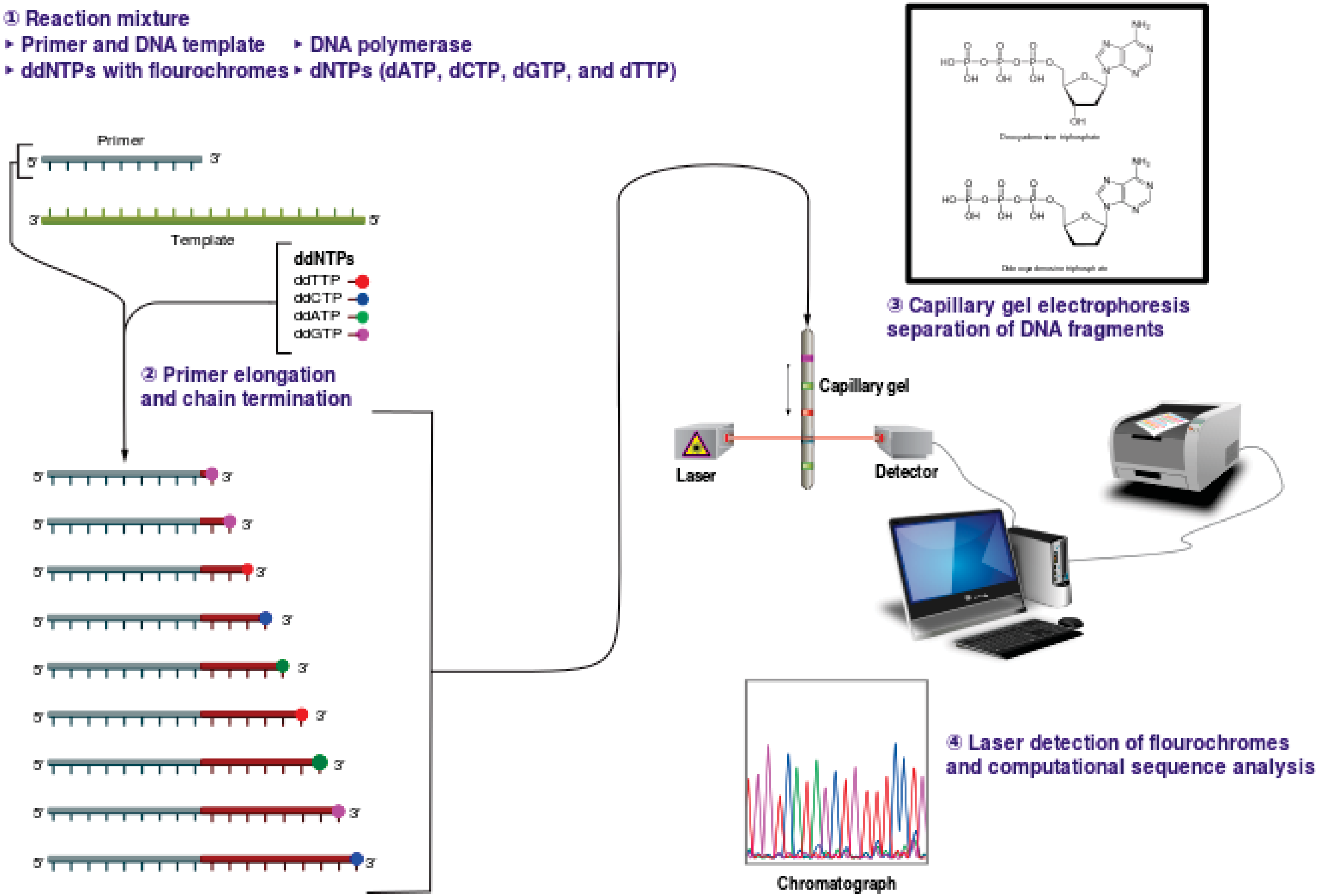
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Khorana, Holley and Nirenberg Nobel (1968)

The central dogma of molecular biology :
My mind was, that a dogma was an idea for which there was no reasonable evidence. You see?! I just didn't know what dogma meant. And I could just as well have called it the 'Central Hypothesis,' or — you know. Which is what I meant to say. Dogma was just a catch phrase. (Crick as reported by H.F. Judson in The Eighth Day of Creation)

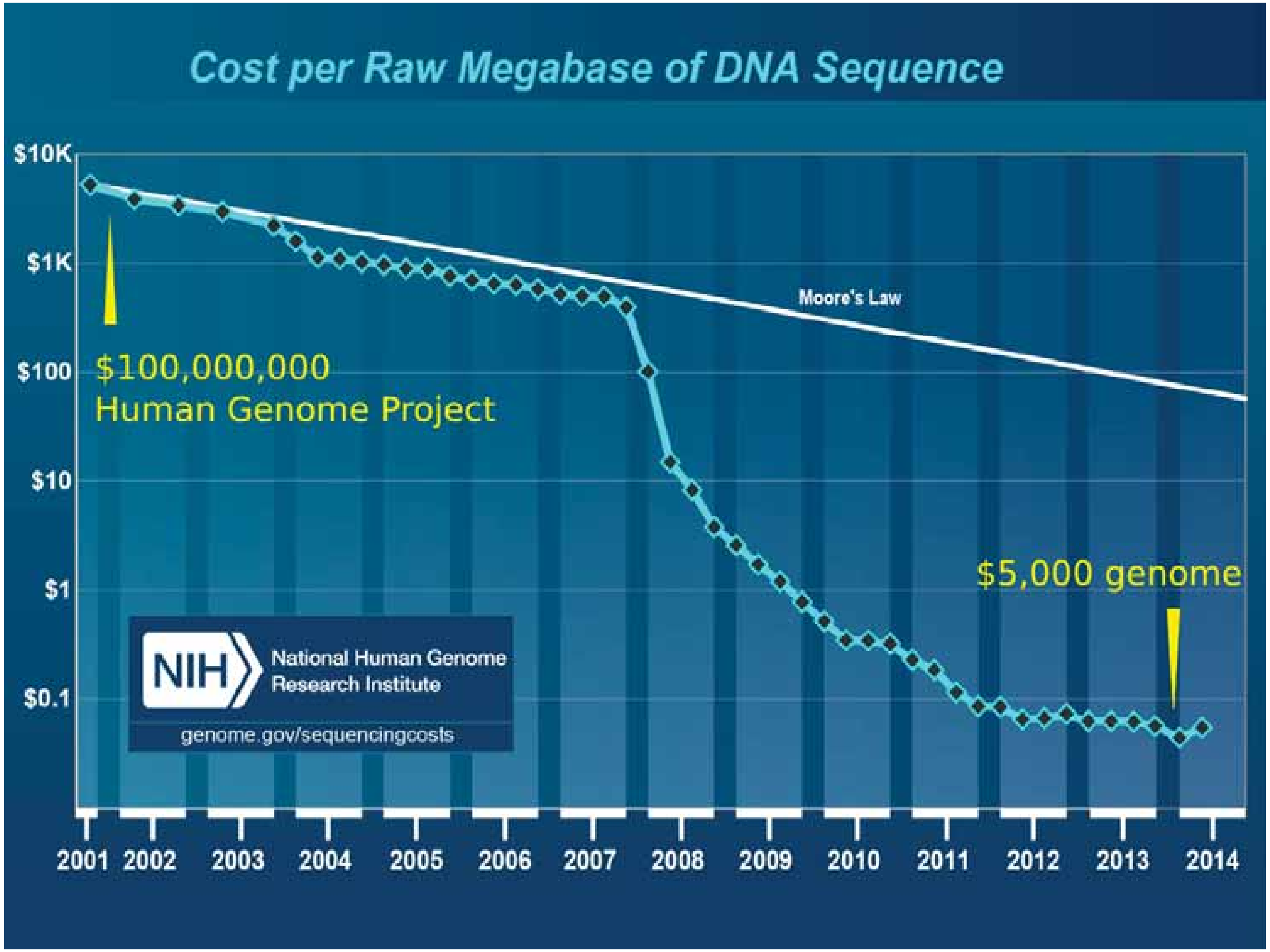
DNA sequencing - Sanger method (1977-)

The chain termination method initially proposed in 1977 was refined/automated up to the beginning of the 2000s.


① Reaction mixture
‣ Primer and DNA template
‣ DNA polymerase
‣ ddNTPs with flourochromes
‣ dNTPs (dATP, dCTP, dGTP, and dTTP)
Primer
5'
3'
3'
5'
Template
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
② Primer elongation and chain termination
NH₂
HO
OH
③ Capillary gel electrophoresis separation of DNA fragments
Capillary gel
Laser
Detector
Chromatograph
④ Laser detection of flourochromes and computational sequence analysis

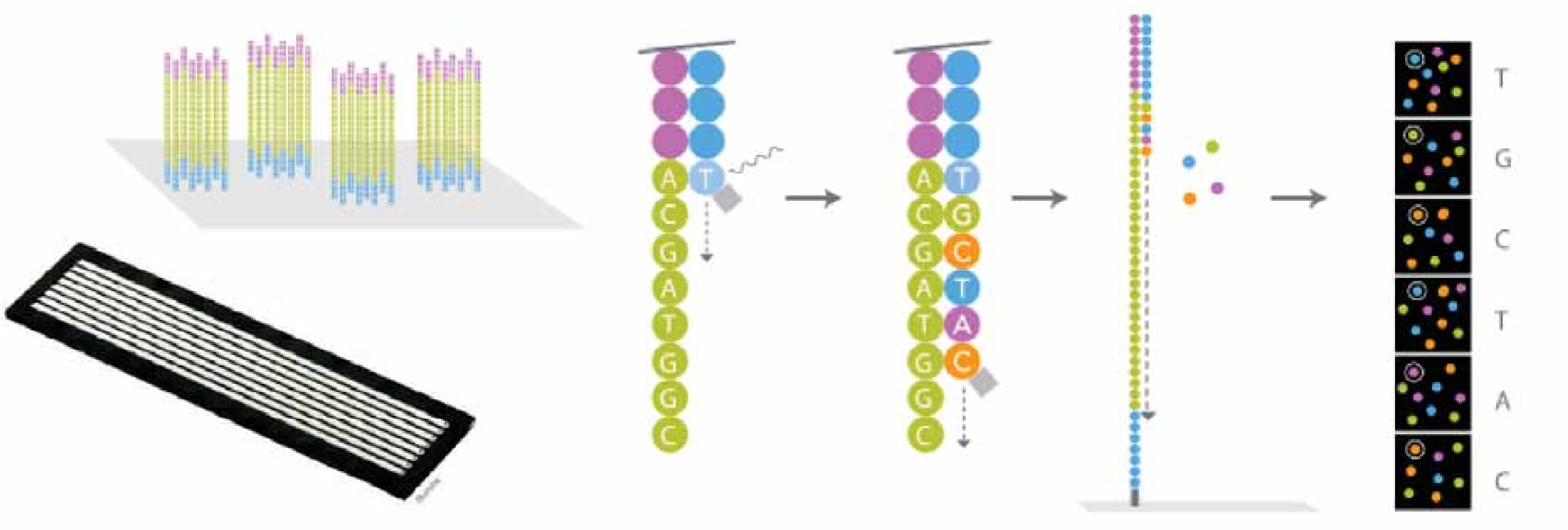
Recent technological breakthroughs in sequencing

best summarized by sequencing costs...



DNA sequencing - High Throughput Sequencing (2007-)

Many different methods exist (available since 2007) and are currently developed. One of the most popular is the sequencing by synthesis on Illumina platforms. It involves three steps: library preparation (not shown), cluster generation and sequencing.



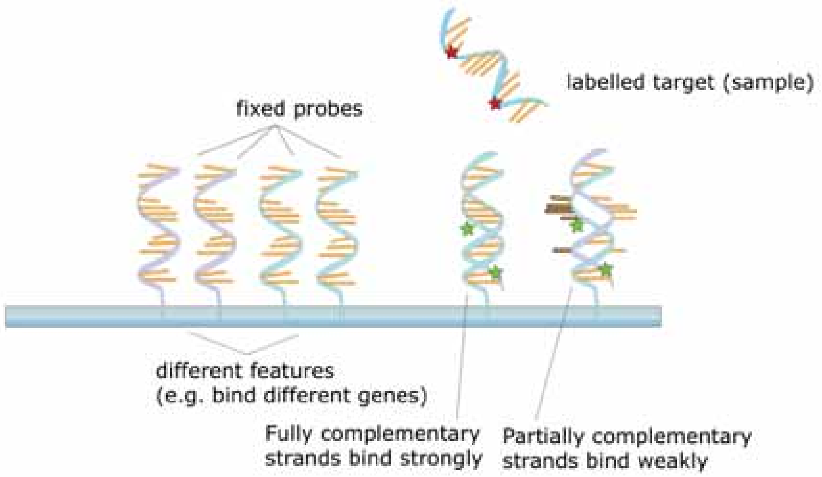
>1,000,000 sequences reads are produced simultaneously: the approach is massively compared to classical (Sanger) sequencing.

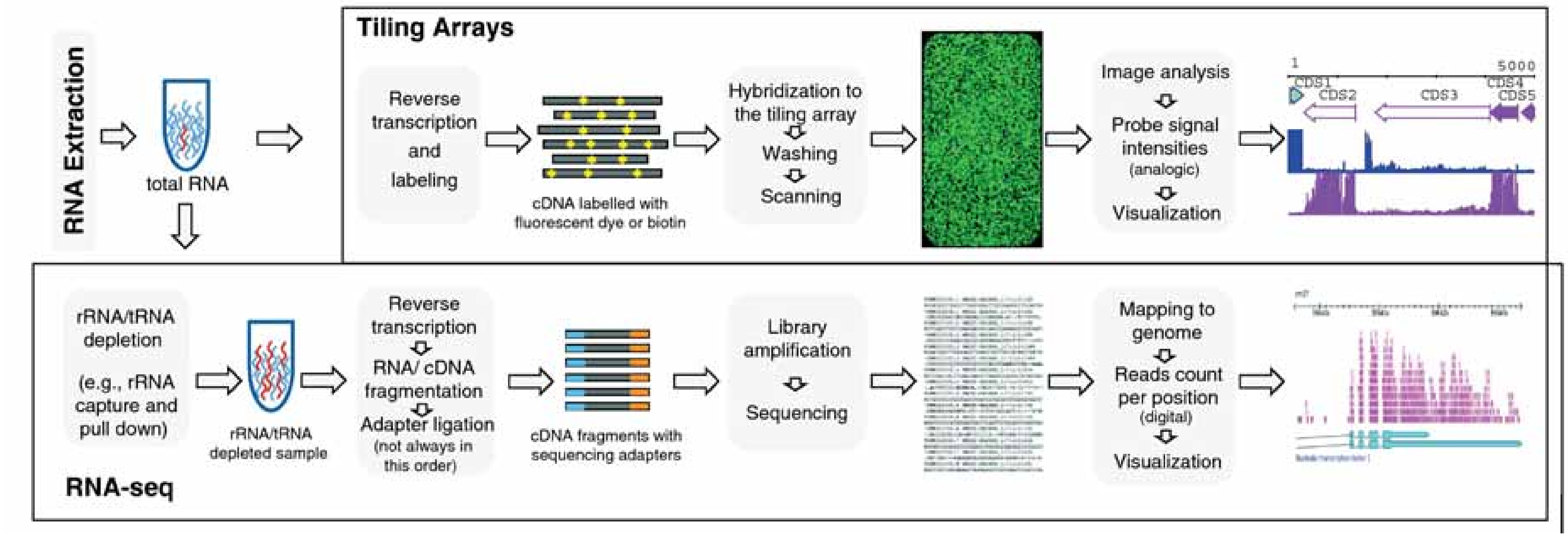
HTS-based approaches have a broad range of applications: de novo sequencing (genomes, metagenomes); resequencing (genetic diversity surveys within species); transcriptome sequencing; genomic location analyses (ChIP-Seq, chromosome conformation capture).

Genome-wide transcriptomics

Some major landmarks

- 1987: DNAs in arrays for expression profiling (Kulesh *et al.*).
- 1994: Sequencing (Sanger) of cDNA/EST libraries (Boguski *et al.*)
- 1995: Miniaturized arrays (microarrays) (Schena *et al.*).
- 2004: Genome tiling arrays (Bertone *et al.*)
- 2008: HTS for expression profiling (Nagalakshmi *et al.*)





The post-genome era

Major research fields in the post-genome era include

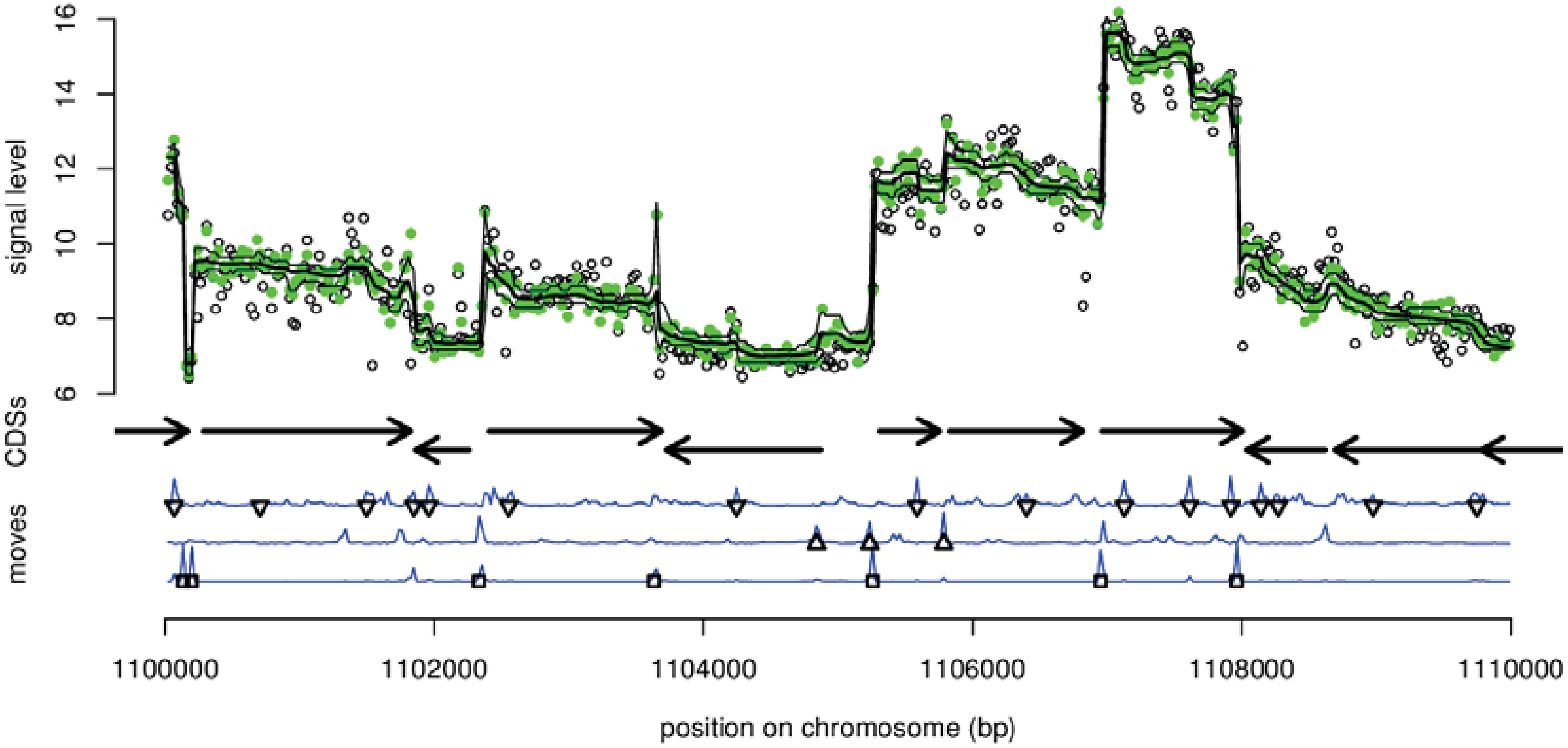
- Use DNA data-sets to identify genetic determinants of interesting phenotypes (e.g. disease gene, personalized medicine).
- Use DNA data-sets for evolutionary/population genetics analyzes.
- Use DNA/RNA/other data sets to investigate microbial ecosystem samples (meta-genomics, meta-transcriptomics).
- Use DNA/RNA/other data sets in integrative approaches understand/model living organisms (systems biology).

Statistical methods are essential, illustrated in this session

- Regression with many variables and few independent measurements (i.e. establishing genotype-phenotype links, regulatory networks) → Jean-Michel Bécu
- Probabilistic models for sequences and trees (i.e. evolutionary and population biology) ↪ Alexis Huet
- Data processing (i.e. normalization in transcriptomics, metagenomics) ↪ Marine Jeanmougin
- Latent variable models for uncovering underlying structures in data-sets ↪ Marine Jeanmougin, here

1- Discovering transcription units

Initial approach for tiling array data[1]. Here Basysbio microarray with overlapping probes starting every 25 bp on each strand of the *Bacillus subtilis* chromosome.



Upshifts and downshifts indicate the position of transcription start and termination sites; the smoothed signal comes with a credibility interval and aims at capturing the underlying local transcription level.

[1] P. Nicolas, A. Leduc, S. Robin, S. Rasmussen, H. Jarmer and P. Bessières. (2009) Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. Bioinformatics.

Discrete-state HMM approximating a continuous state-space

Let x_t denote the log-transformed data and u_t the underlying signal.

"Emission" model

$$x_t \mid u_t \quad \sim \quad \mathcal{N}(u_t, \sigma^2)\,.$$

(simplified: the actual emission model σ^2 depends on u_t and the mean also depends on covariates)

Transition kernel

$$u_{t+1} \mid u_t \quad \sim \quad \pi(u_{t+1}, u_t)$$

Difficulty: (u_t) is continuous-valued whereas the HMM machinery works well for discrete and typically small number of hidden states (Forward-Backward, Viterbi, ... have complexity $O(TK^2)$ in their general form).

$\hookrightarrow$ Use a transition matrix structure that allows algorithms in $O(TK)$ and choose a discretization-step h small enough.

A transition kernel accounting for shift and drift

Hidden state space: grid with K points

$$K = \frac{u_{\max} - u_{\min}}{h} + 1\,.$$

Mixture of 4 types of moves

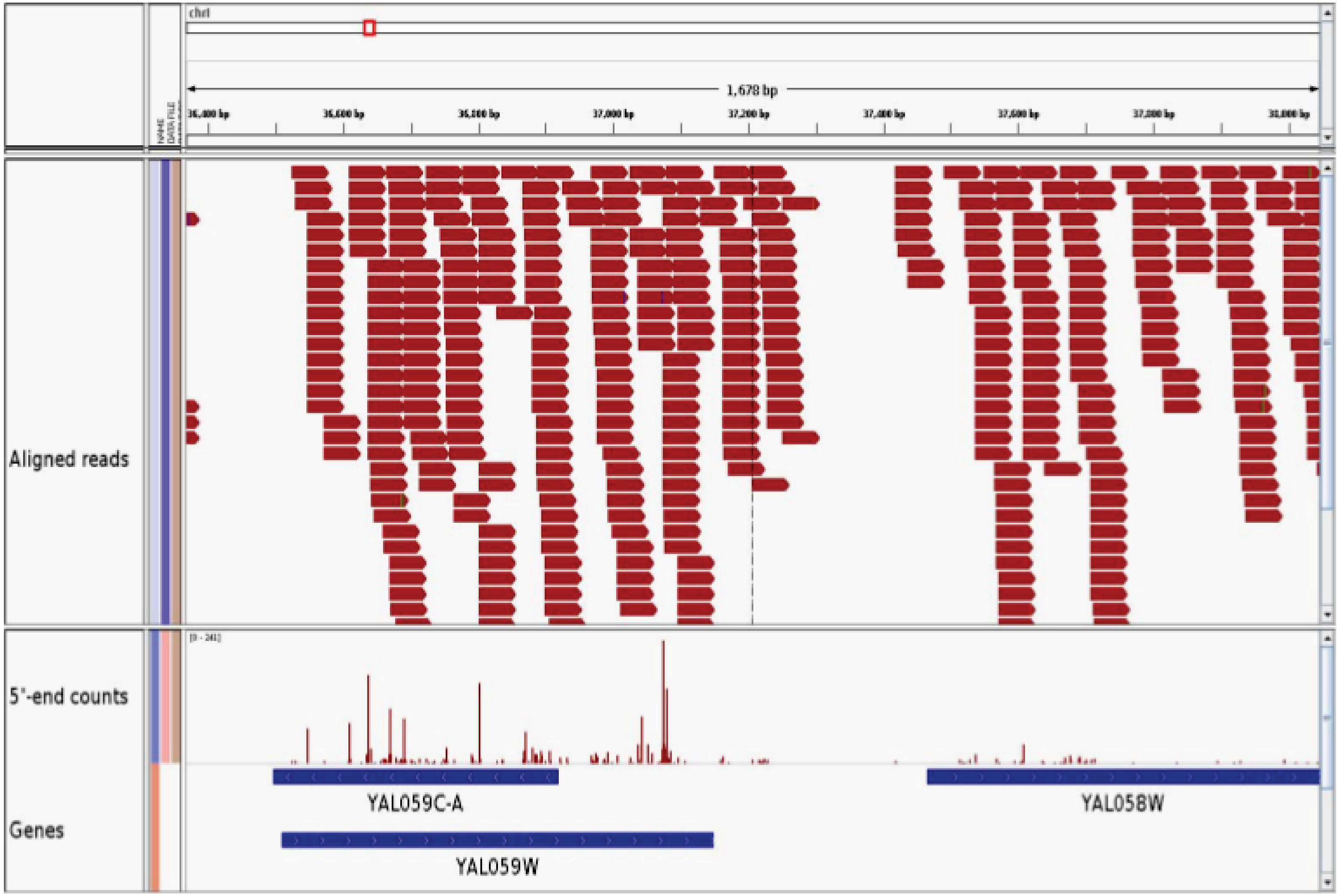
$$\begin{aligned}\pi(u_t, u_{t+1}) &= \alpha_n \mathbb{I}_{\{u_{t+1}=u_t\}} + \alpha_s \eta(u_{t+1}) \\ &\quad + \alpha_u \mathbb{I}_{\{u_{t+1}>u_t\}} \lambda_u^{\frac{u_{t+1}-u_t}{h}-1}(1-\lambda_u) \\ &\quad + \alpha_d \mathbb{I}_{\{u_{t+1}<u_t\}} \lambda_d^{\frac{u_t-u_{t+1}}{h}-1}(1-\lambda_d)\,,\end{aligned}$$

with $0 \le \alpha_n, \alpha_s, \alpha_u, \alpha_d \le 1$, $\alpha_n + \alpha_s + \alpha_u + \alpha_d = 1$ et $0 \le \lambda_u, \lambda_d < 1$.

- α_n, probability of not moving,
- α_s, probability of shift,
- α_u and α_d, probabilities of upward and downward drifts.

$\hookrightarrow$ When $h \to 0$ and $h/(1-\lambda) \to \gamma$ the discrete kernel converges towards a continuous kernel (State Space Model, HMM with continuous-valued underlying process).

RNA-Seq count data



S.cerevisiae, chromosome I, Watson strand

RNA-Seq count data vs. tiling array data

Count reads starting (i.e. 5'-ends) at each position. Ideally we would expect independence given the underlying transcription level.

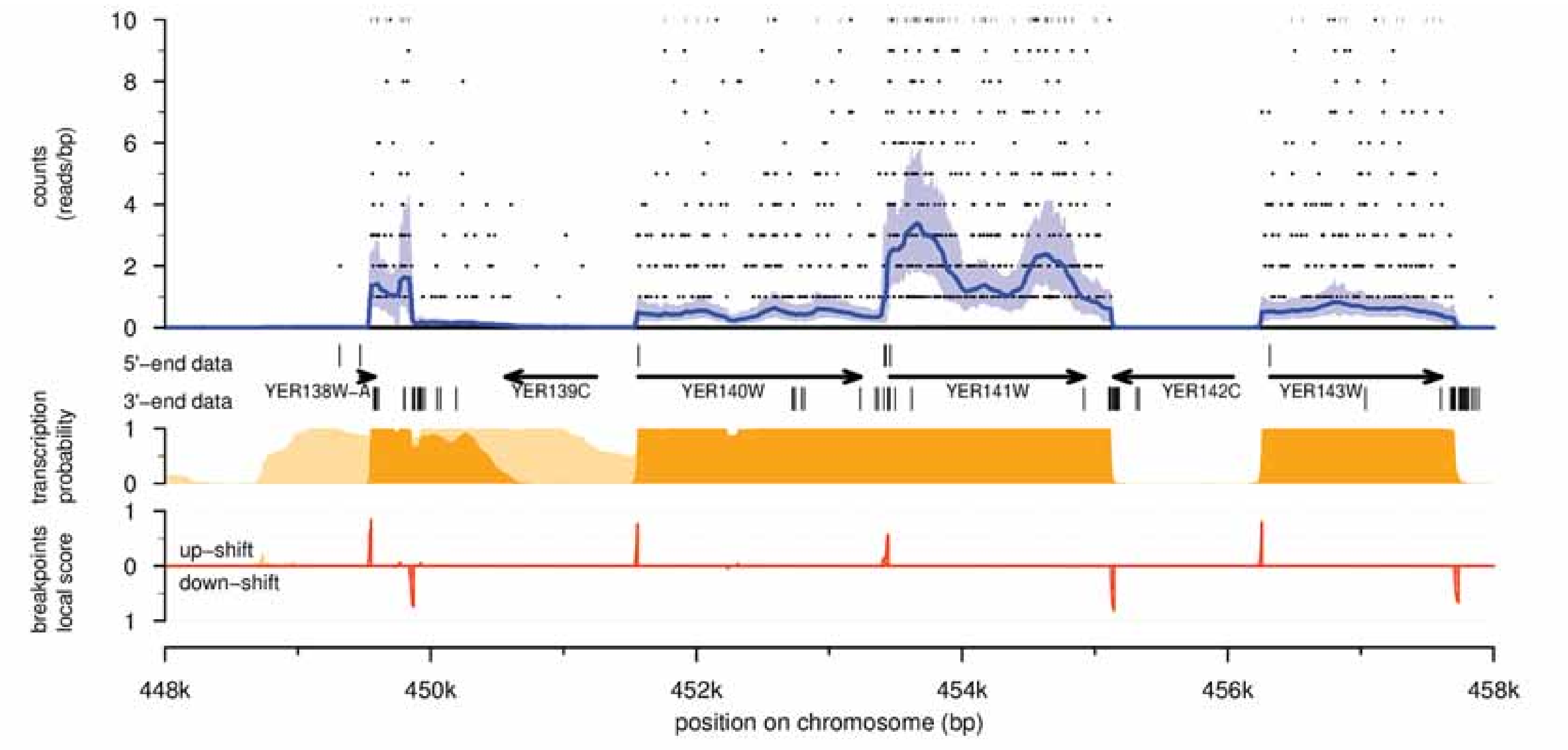
RNA-Seq count data vs. tiling array data

- 1 bp resolution instead of tiling step (≈25 bp)
- Count data instead of logarithm of fluorescence intensity
- Observations are discrete $\mathbb{N}^\star = \{0, 1, 2, \ldots\}$ instead of continuous $\mathbb{R}$
- Underlying signal (expression level) lives in $\mathbb{R}^+$ with discrete mass at zero instead of $\mathbb{R}$

We also made different modeling and inference choices

- State Space Model instead of the discrete space HMM approximation
- Particle MCMC algorithm proposed by Andrieu, Doucet and Holenstein (2010) ($\hookrightarrow$ Bayesian estimation) instead of EM ($\hookrightarrow$ ML estimation)

Smoothing and break-point detection from RNA-Seq count data



Main unanticipated issues encountered in this work[2]

- finding a reasonably good emission model
- finding a solution for smoothing out local heterogeneity inducing short range autocorrelation of read counts (local scaling term)

[1] B. Mirauta, P. Nicolas, and H. Richard (2014) Parseq: reconstruction of microbial transcription landscape from RNA-Seq read counts using state-space models. Bioinformatics.

Break-point detection from RNA-Seq count data is not easy

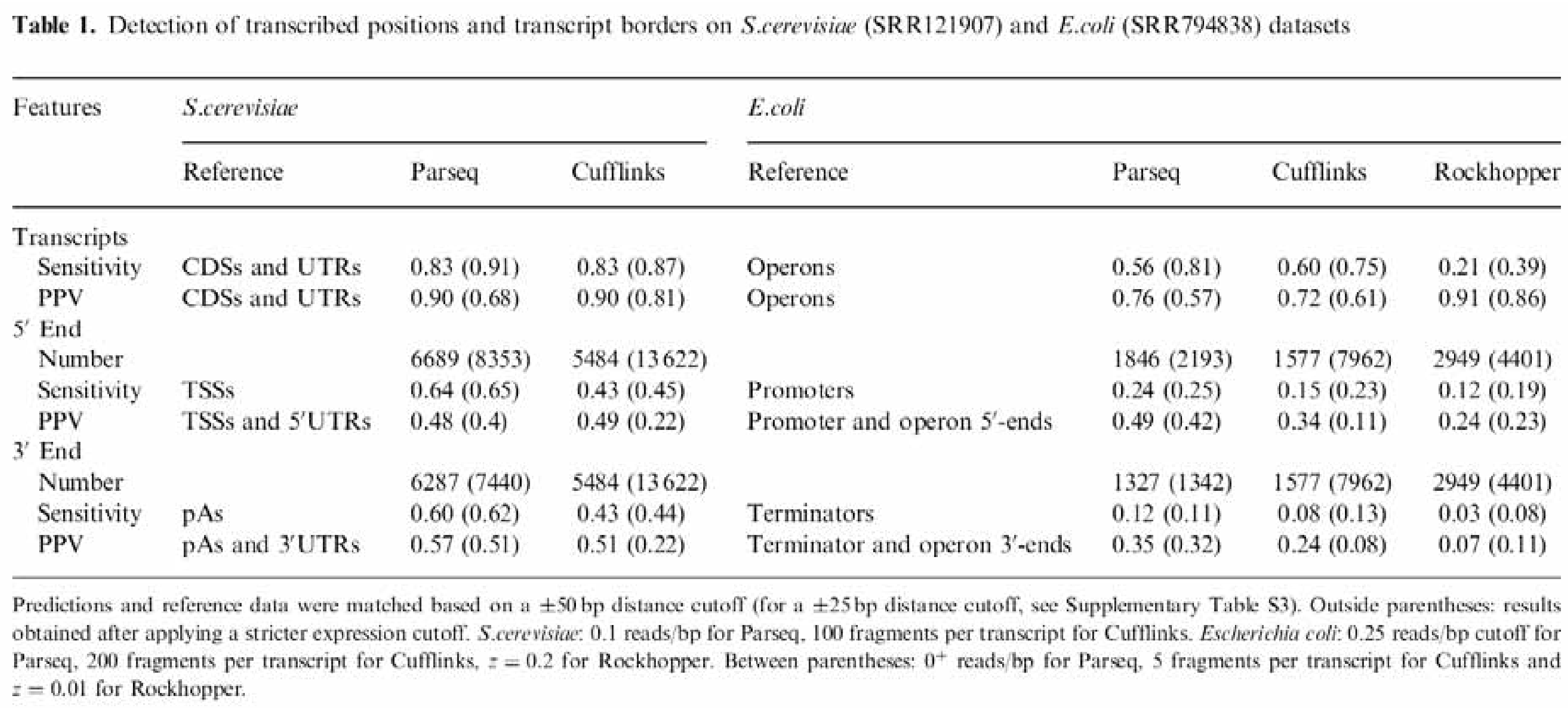


Table 1. Detection of transcribed positions and transcript borders on *S.cerevisiae* (SRR121907) and *E.coli* (SRR794838) datasets

Features	*S.cerevisiae*			*E.coli*			
	Reference	Parseq	Cufflinks	Reference	Parseq	Cufflinks	Rockhopper
Transcripts							
Sensitivity	CDSs and UTRs	0.83 (0.91)	0.83 (0.87)	Operons	0.56 (0.81)	0.60 (0.75)	0.21 (0.39)
PPV	CDSs and UTRs	0.90 (0.68)	0.90 (0.81)	Operons	0.76 (0.57)	0.72 (0.61)	0.91 (0.86)
5′ End							
Number		6689 (8353)	5484 (13 622)		1846 (2193)	1577 (7962)	2949 (4401)
Sensitivity	TSSs	0.64 (0.65)	0.43 (0.45)	Promoters	0.24 (0.25)	0.15 (0.23)	0.12 (0.19)
PPV	TSSs and 5′UTRs	0.48 (0.4)	0.49 (0.22)	Promoter and operon 5′-ends	0.49 (0.42)	0.34 (0.11)	0.24 (0.23)
3′ End							
Number		6287 (7440)	5484 (13 622)		1327 (1342)	1577 (7962)	2949 (4401)
Sensitivity	pAs	0.60 (0.62)	0.43 (0.44)	Terminators	0.12 (0.11)	0.08 (0.13)	0.03 (0.08)
PPV	pAs and 3′UTRs	0.57 (0.51)	0.51 (0.22)	Terminator and operon 3′-ends	0.35 (0.32)	0.24 (0.08)	0.07 (0.11)

Predictions and reference data were matched based on a ±50 bp distance cutoff (for a ±25 bp distance cutoff, see Supplementary Table S3). Outside parentheses: results obtained after applying a stricter expression cutoff. *S.cerevisiae*: 0.1 reads/bp for Parseq, 100 fragments per transcript for Cufflinks. *Escherichia coli*: 0.25 reads/bp cutoff for Parseq, 200 fragments per transcript for Cufflinks, $z = 0.2$ for Rockhopper. Between parentheses: 0^+ reads/bp for Parseq, 5 fragments per transcript for Cufflinks and $z = 0.01$ for Rockhopper.

Our program (Parseq) does better than some others . . .

Drift and local scaling contribute to improve the results . . .

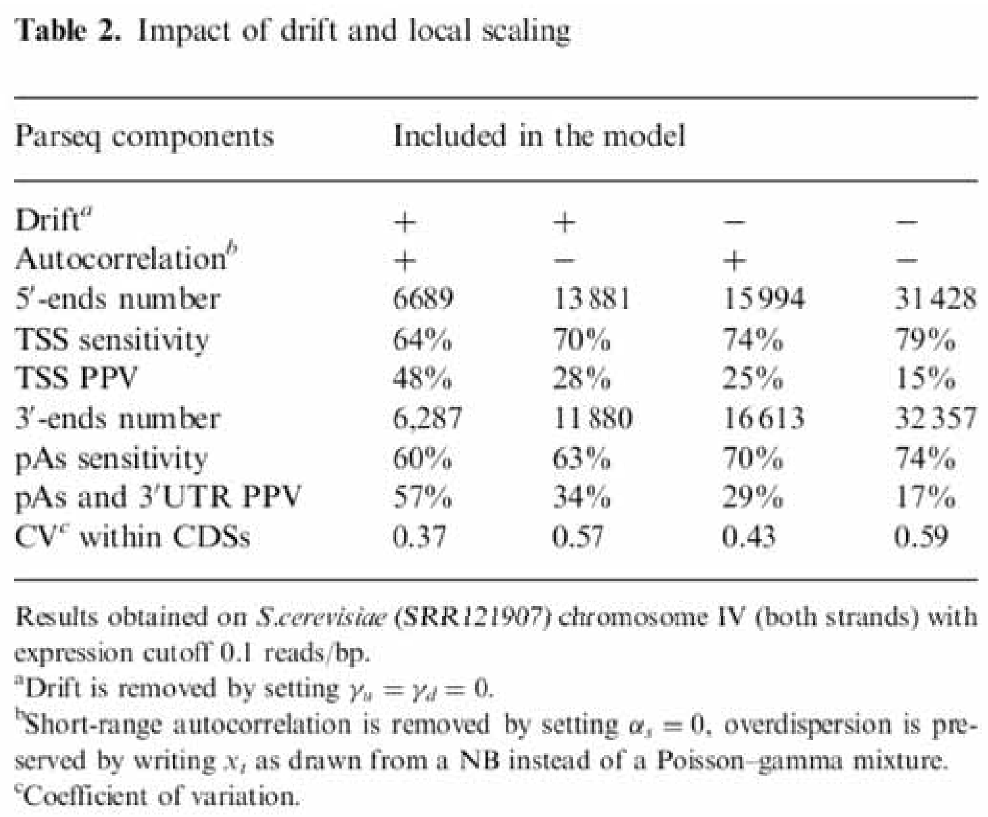


Table 2. Impact of drift and local scaling

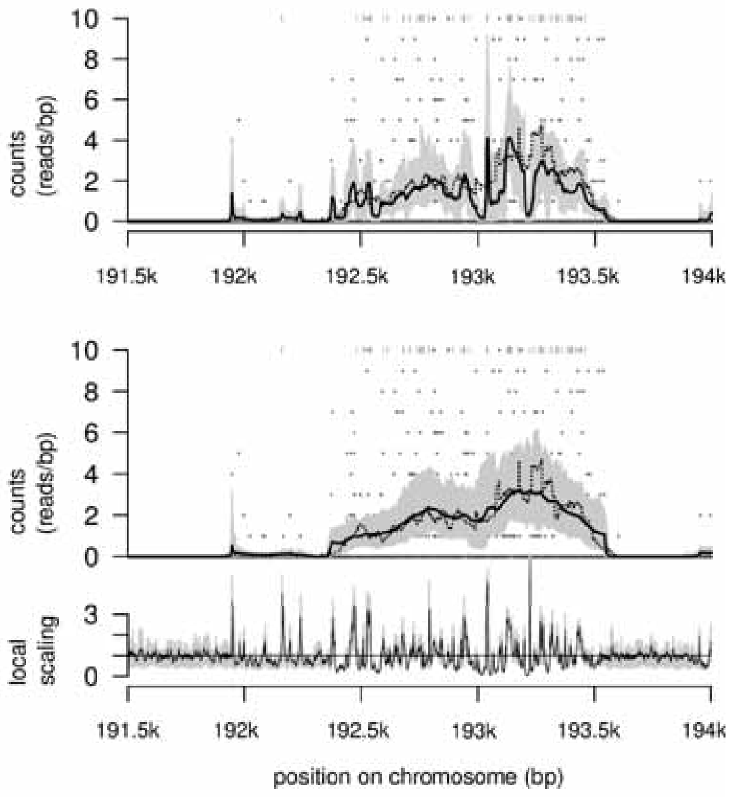
Parseq components	Included in the model			
Drift[a]	+	+	−	−
Autocorrelation[b]	+	−	+	−
5′-ends number	6689	13 881	15 994	31 428
TSS sensitivity	64%	70%	74%	79%
TSS PPV	48%	28%	25%	15%
3′-ends number	6,287	11 880	16 613	32 357
pAs sensitivity	60%	63%	70%	74%
pAs and 3′UTR PPV	57%	34%	29%	17%
CV[c] within CDSs	0.37	0.57	0.43	0.59

Results obtained on *S.cerevisiae* (*SRR121907*) chromosome IV (both strands) with expression cutoff 0.1 reads/bp.

[a] Drift is removed by setting $\gamma_u = \gamma_d = 0$.

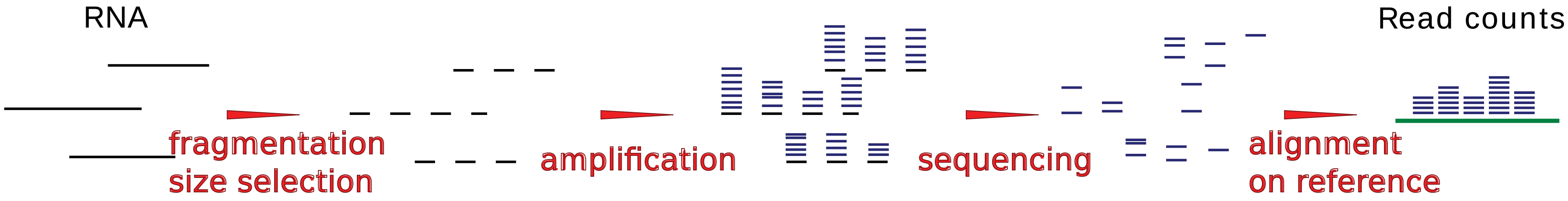
[b] Short-range autocorrelation is removed by setting $\alpha_s = 0$, overdispersion is preserved by writing x_t as drawn from a NB instead of a Poisson–gamma mixture.

[c] Coefficient of variation.



. . . by decreasing the rate of “false positives” in break-point calling (Positive Predictive Value increases).

The Negative Binomial distribution, a natural choice?



- the molecule abundance before sequencing writes as a product $u_t a_t$ whose distribution given the amount u_t of the RNA species is

$$u_t a_t \mid u_t \quad \sim \quad \Gamma(shape = \kappa, scale = u_t/\kappa)$$

  where a_t ($a_t \sim \Gamma(shape = \kappa, scale = 1/\kappa)$, $\mathbb{E}(a_t) = 1$ and $\mathbb{V}(a_t) = 1/\kappa$) captures bias and randomness in fragmentation, size selection, amplification – and sequencing if needed.

- sequencing plays the role of a sampling procedure (Poisson distribution)

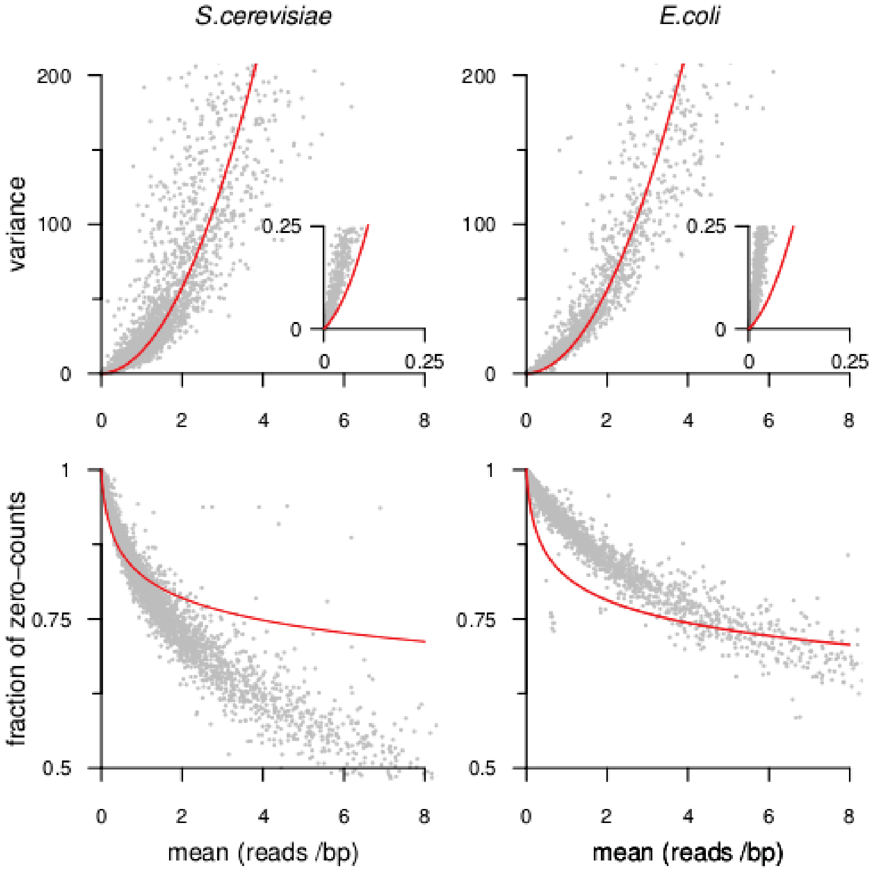
$$\begin{aligned} y_t \mid u_t, a_t &\quad \sim \quad \mathcal{P}(u_t a_t) \\ y_t \mid u_t &\quad \sim \quad \mathcal{NB}(mean = u_t, size = \kappa) \end{aligned}$$

RNA-Seq counts vs. Negative Binomial distribution

$$y_t \mid u_t \sim \mathcal{NB}(mean = u_t, size = \kappa)$$

$$\mathbb{V}(y_t \mid u_t) = u_t + (1/\kappa)u_t^2$$

$$\mathbb{P}(y_t = 0 \mid u_t) = 1/(1 + (u_t/\kappa))^\kappa$$



The variance vs. mean and fraction of zero-counts vs. mean relationships are not properly captured . . . which impedes correct reconstruction of the underlying expression level.

Towards an alternative to the Negative Binomial

At least two obvious “empirical” modifications could be envisioned

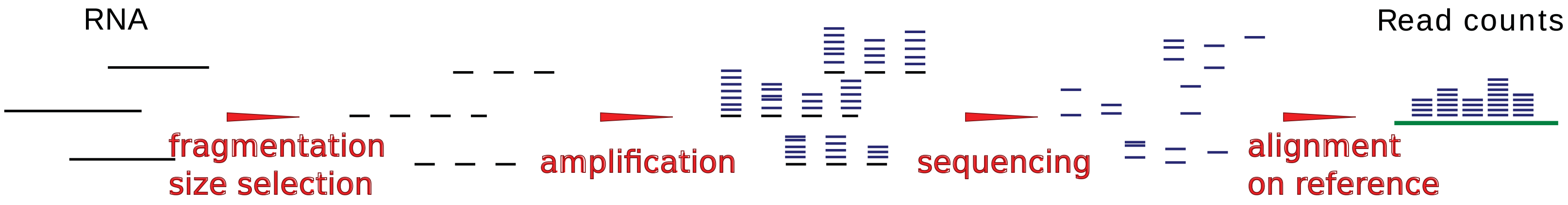
- making size parameter a function of u_t to capture the correct relationship between variance and mean,
  but such a “plug-in” of an arbitrary link between variance and mean would not fix the lack of fit of the zero-counts.
- adding mass at zero (zero-inflation),
  but the zeros are not always in excess.

Can we think of more ingenious approaches to find an alternative to the Negative Binomial?

Of note, the choice of the Gamma distribution for a_t was mostly done for convenience but the relationships $\mathbb{V}(y_t \mid u_t) = u_t + \mathbb{V}(a_t)u_t^2$ holds as long as we write $y_t \sim \mathcal{P}(u_t a_t)$ with $u_t \perp a_t$ and $\mathbb{E}(a_t) = 1$ (law of total variance).

$$\begin{aligned}
\mathbb{V}(y_t \mid u_t) &= \mathbb{E}(\mathbb{V}(y_t \mid a_t, u_t)) + \mathbb{V}(\mathbb{E}(y_t \mid a_t, u_t)) \\
&= \mathbb{E}(a_t u_t) + \mathbb{V}(a_t u_t) \\
&= u_t + \mathbb{V}(a_t) u_t^2
\end{aligned}$$

Proposed alternative : a more complex/realistic compound distribution



- read sampling

$$y_t \quad \sim \quad \mathcal{P}(x_t a_t),$$

  where x_t is the number of molecules before "amplification" and a_t is the amplification coefficient.

- "amplification" and secondary variance inflation

$$a_t \quad \sim \quad \Gamma(shape = \kappa, scale = \theta),$$

  mean $\kappa\theta$ corresponds to the average number of reads (y_t) per initial molecule (x_t)

- molecule sampling and initial variance inflation

$$\begin{aligned} x_t \mid u_t, s_t \quad &\sim \quad \mathcal{P}((u_t/\kappa\theta)s_t), \\ s_t \quad &\sim \quad \Gamma(shape = \kappa_S, scale = 1/\kappa_S) \end{aligned}$$

In our SSM context s_t is made piecewise constant (to capture short-range autocorrelation) and the emission density writes

$$\begin{aligned} s_t \mid s_{t-1} \quad &\sim \quad (1-\alpha_S)\delta(s_{t-1}) + \alpha_S \Gamma(\kappa_S, 1/\kappa_S) \\ \mathbb{P}(y_t = k \mid u_t, s_t) \quad &= \quad \sum_{x_t \geq 0} \mathcal{P}(x_t; (u_t/\kappa\theta)s_t)\mathcal{NB}(y_t; \kappa, x_t\theta/(x_t\theta + 1)) \end{aligned}$$

Fitting emission parameters to empirical observations

- Variance vs. mean

$$\mathbb{V}(y_t \mid u_t) \quad = \quad \Big(\frac{1}{\kappa} + \frac{1}{\kappa_s} + \frac{1}{\kappa_s \kappa}\Big) \cdot u_t^2 + \cdot (1 + \theta + \kappa\theta) \cdot u_t$$

- Zero-counts vs. mean

$$\mathbb{P}(y_t = 0 \mid u_t) \quad = \quad \sum_{x_t \geq 0} \mathcal{NB}(x_t; \kappa_s, \frac{u_t}{\kappa_s \kappa \theta + u_t}) \cdot \mathcal{NB}(0; \kappa, \frac{x_t \theta}{x_t \theta + 1})$$

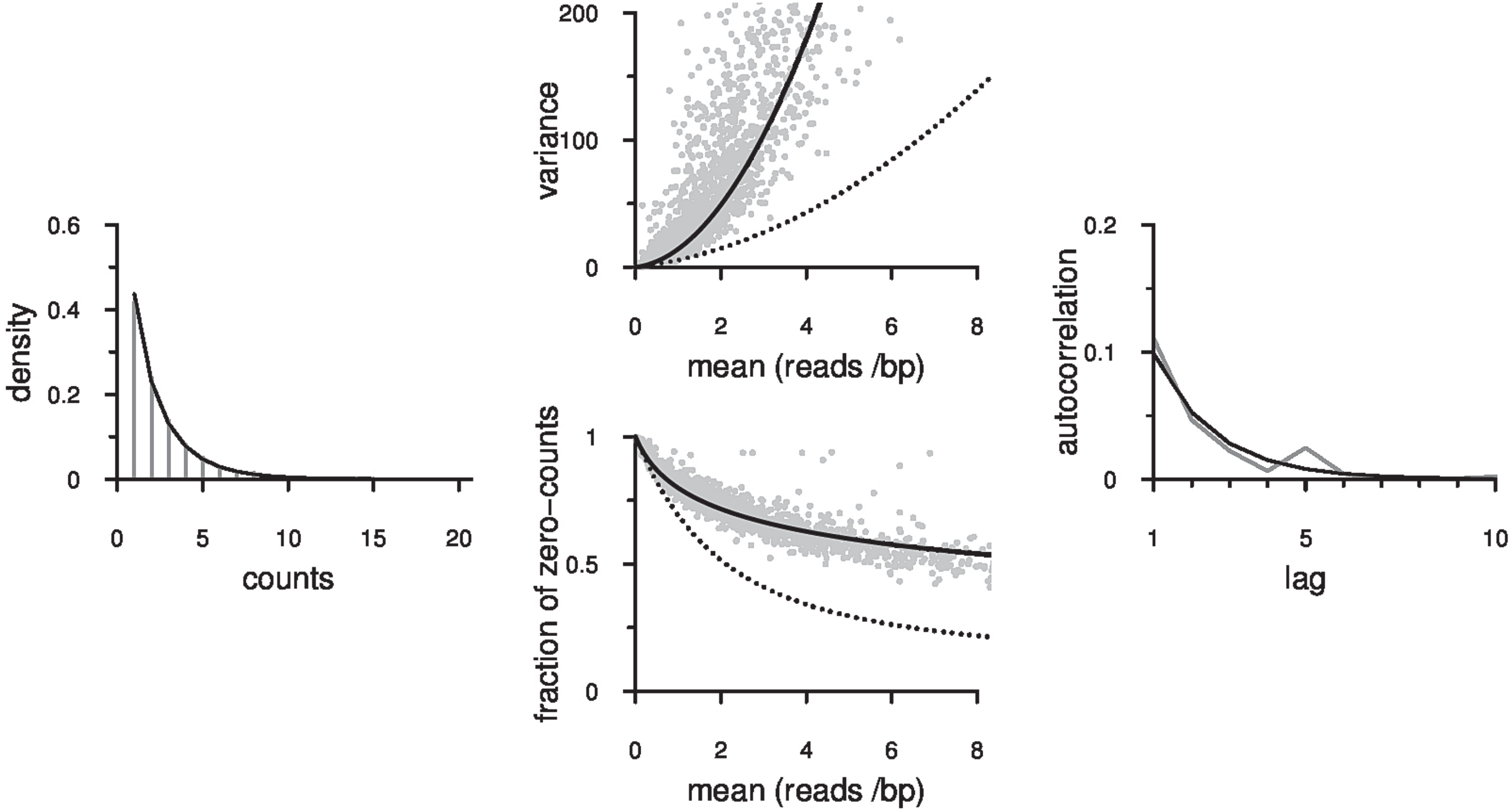
- Distribution of counts in regions of low expression level ($u_t \to 0^+$) we have $\lim_{u_t \to 0^+} \mathbb{P}(x_t = 1 \mid y_t \geq 1) = 1$, and thus

$$y_t \mid u_t, y_t \geq 1 \quad \to_{u_t \to 0^+} \quad \mathcal{NB}_{-\{0\}}(mean = \kappa, size = \frac{\theta}{\theta + 1})$$

- Short range auto-correlation in a region where the underlying expression level is constant ($u_t = u_{t+1} = u_g$).

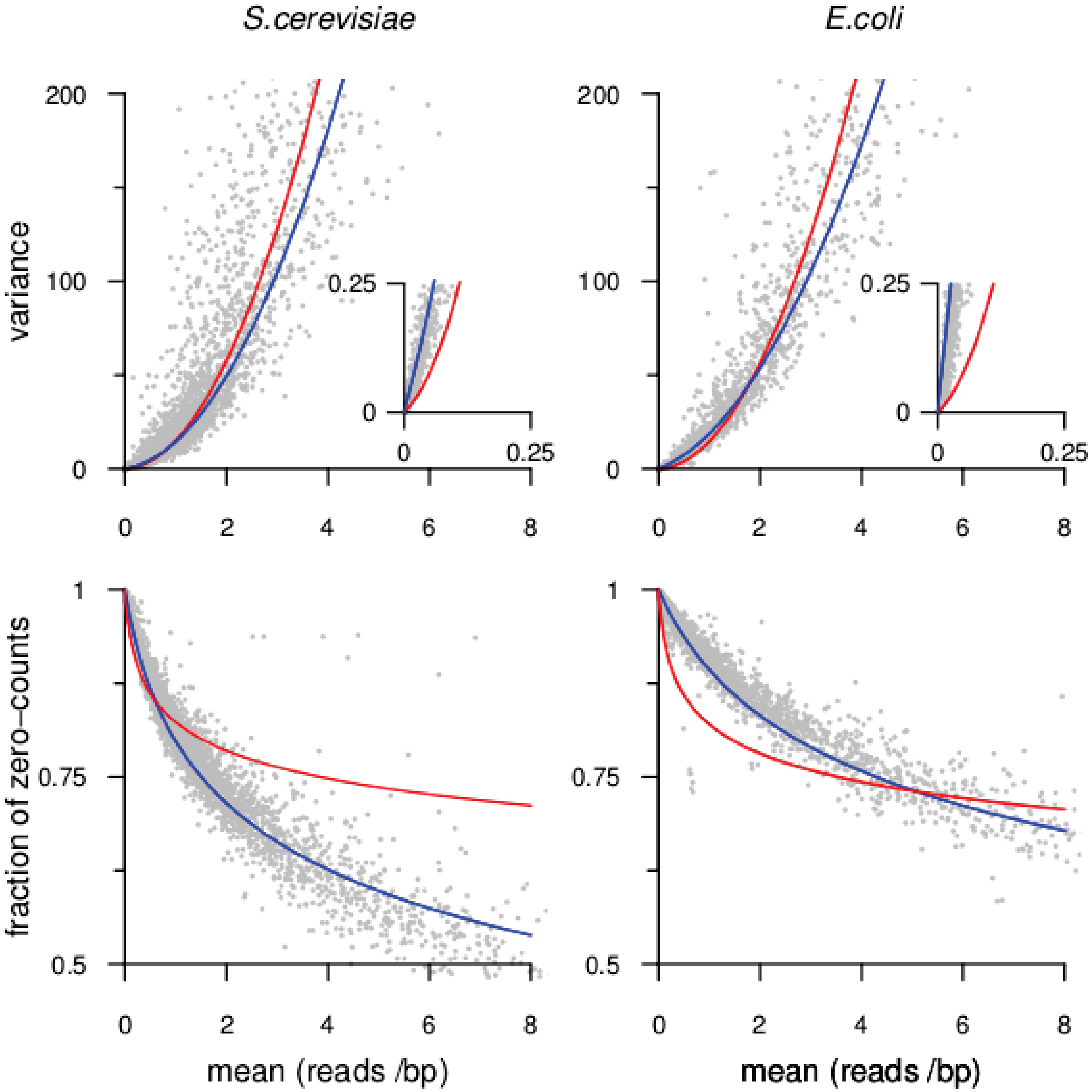
$$cor_g(y_t, y_{t+l}) \quad = \quad \frac{u_g^2}{\kappa_s} \cdot \frac{1}{\mathbb{V}(y_t \mid u_g)} \cdot \alpha_s^l, \qquad l \geq 1$$

Parameter choice for a real RNA-Seq data-set



a crude procedure involving three steps : (i) select κ_s and θ based on read-counts in regions of low expression level (ii) select κ based on variance vs. mean and zero-counts vs. mean (iii) select α_s based on autocorrelation.

A much better model than NB for RNA-Seq count data



2- Partitioning the promoter space wrt RNAPol binding sites

Systematic exploration of *B. subtilis* transcriptional landscape based on large data-sets (Basybio and Basynthec projects)^3:

- 1 prototype strain (“wild type”).
- 1 tiling array design providing a strand-specific expression signal with a 22 bp step.
- 269 hybridizations sampling a maximum variety of lifestyles,
- 104 different biological conditions, most with 2-3 biological replicates.

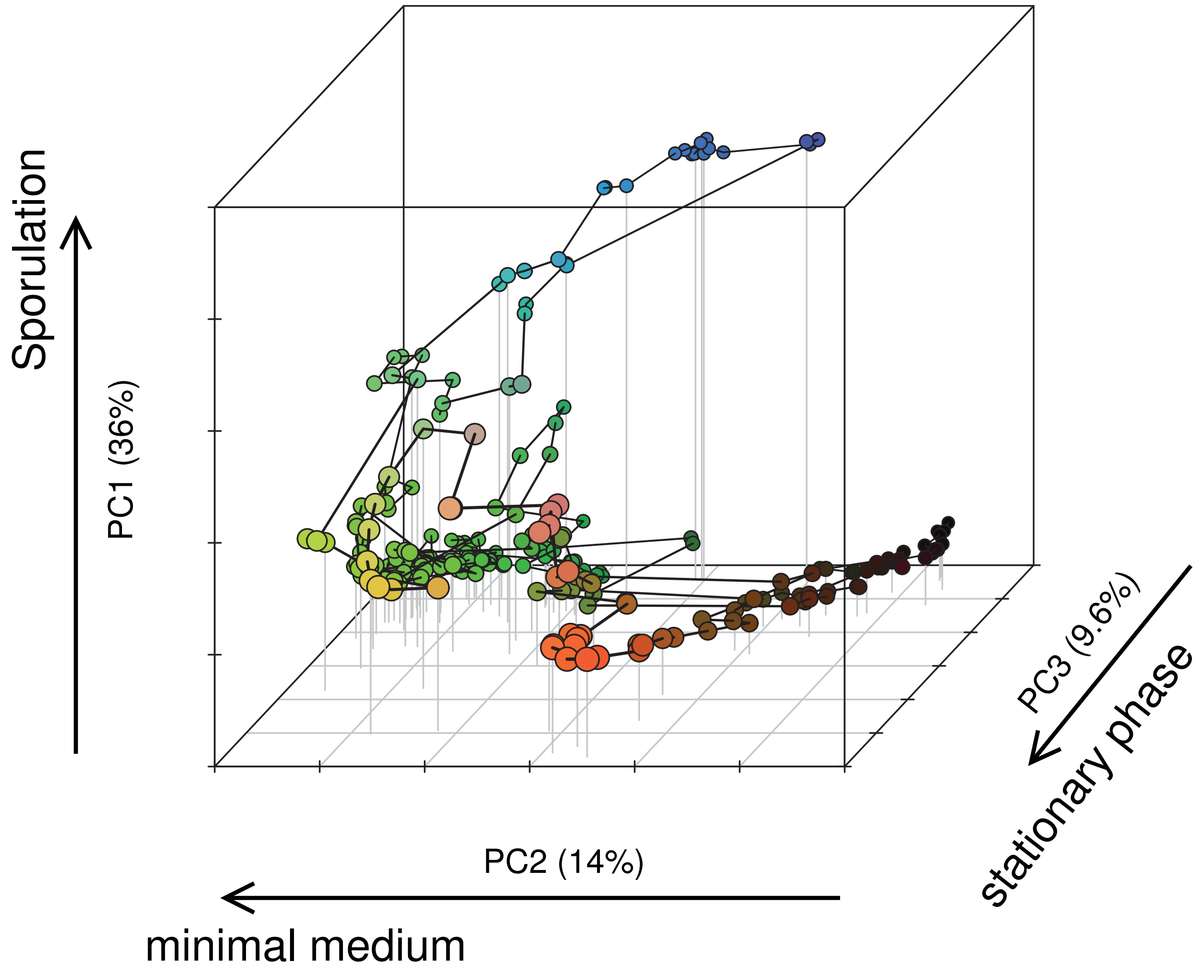
Growth on various media (rich/poor, solid/liquid, aerobic/anaerobic), variety of stresses (ethanol, salt, temperature, oxidative), landmark adaptations (sporulation, germination, competence) . . .

Main contributors of experimental data:

- Etienne Dervyn, Philippe Noirot (Biologie des systèmes, MICALIS, INRA),
- Ulrike Mäder (Univ. Greifswald).

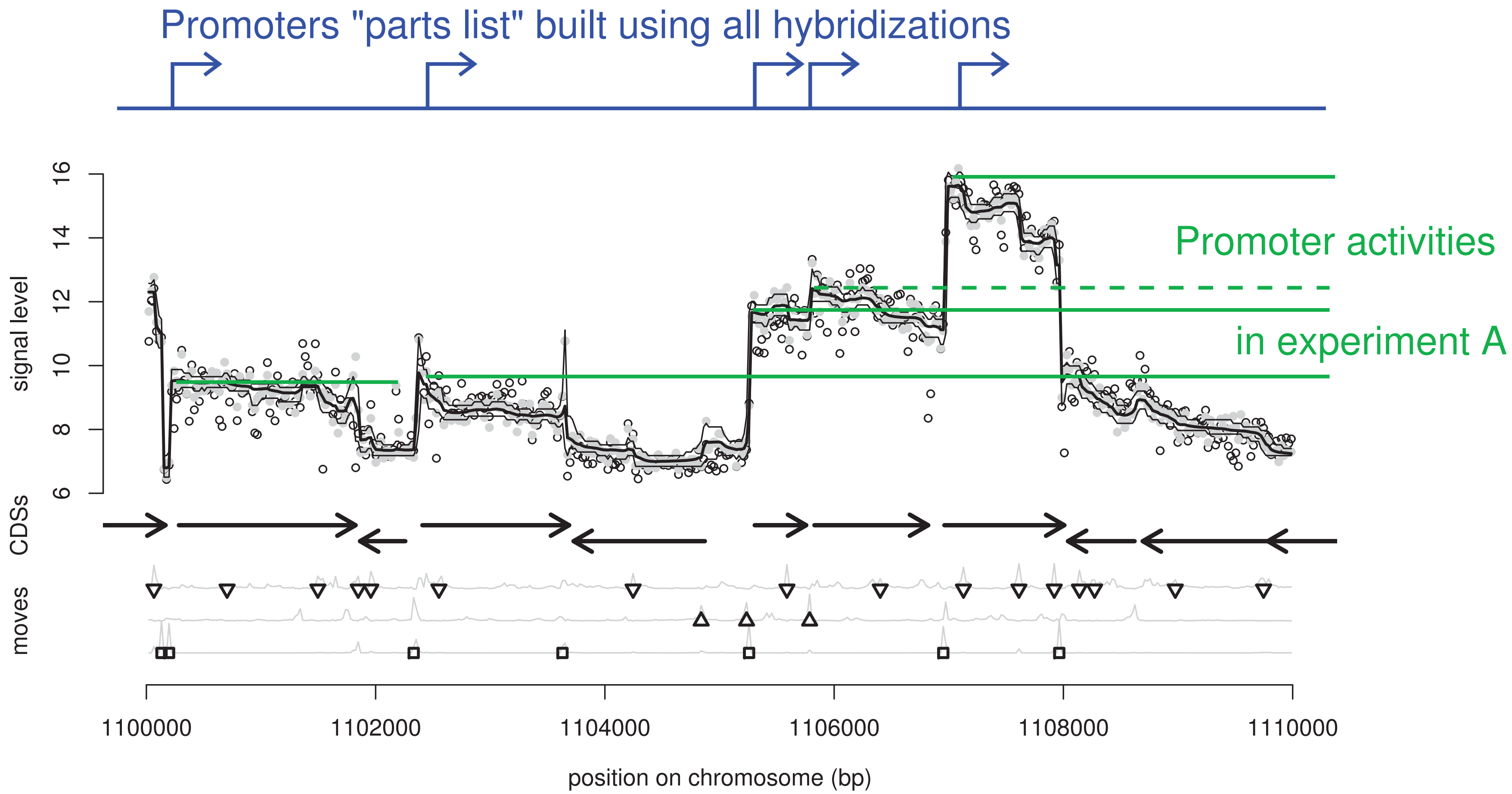
[1] P. Nicolas, U. Mäder, E. Dervyn, (47 authors), and P. Noirot. Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis. Science. 2012.

Highly coordinated changes of gene expression levels

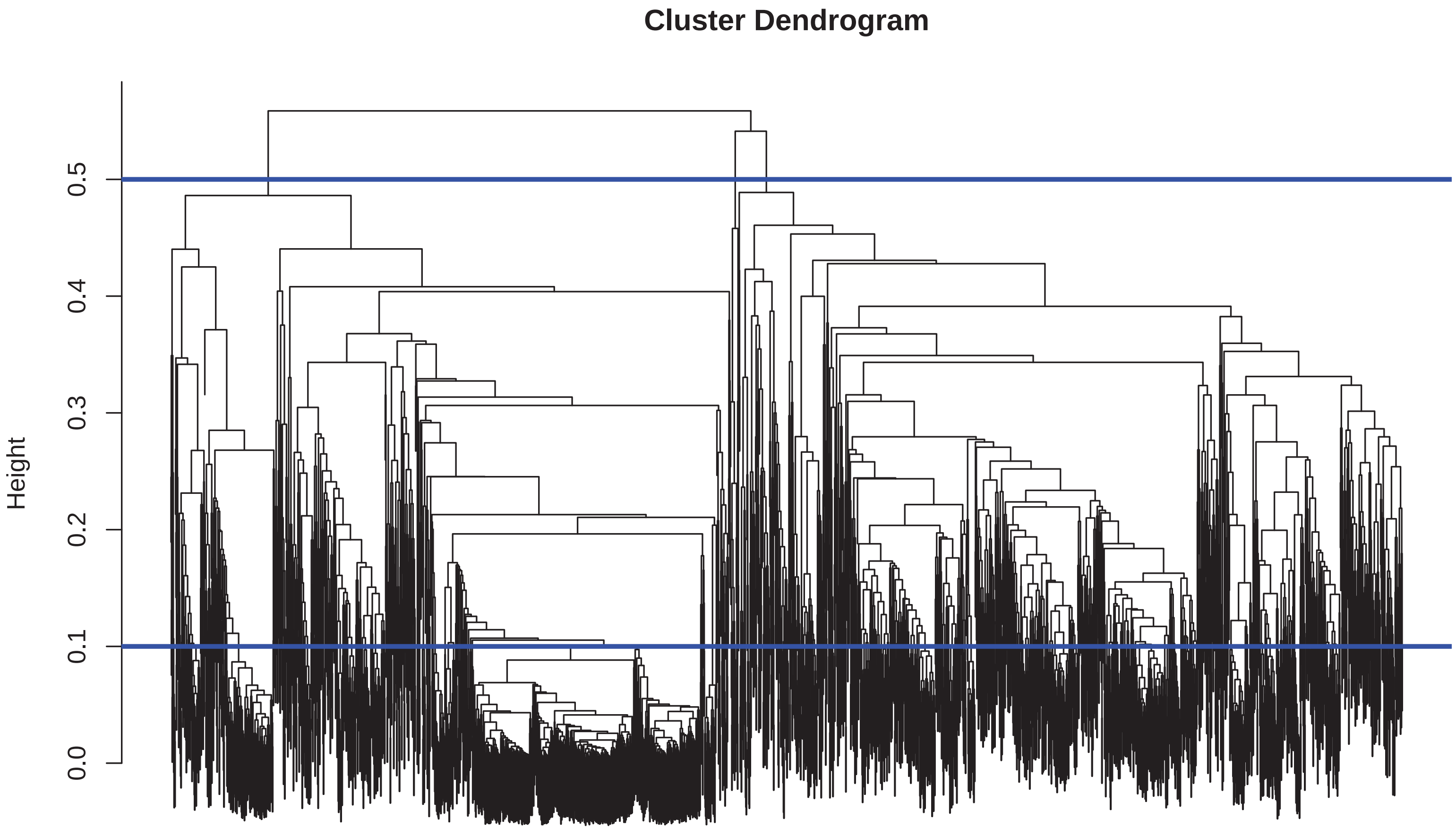


Each point represents an experiment (hybridization) in which the transcriptional activity of 5875 genome segments are recorded.

Monitoring promoter activity



Summarizing correlations between activities of 3243 promoters

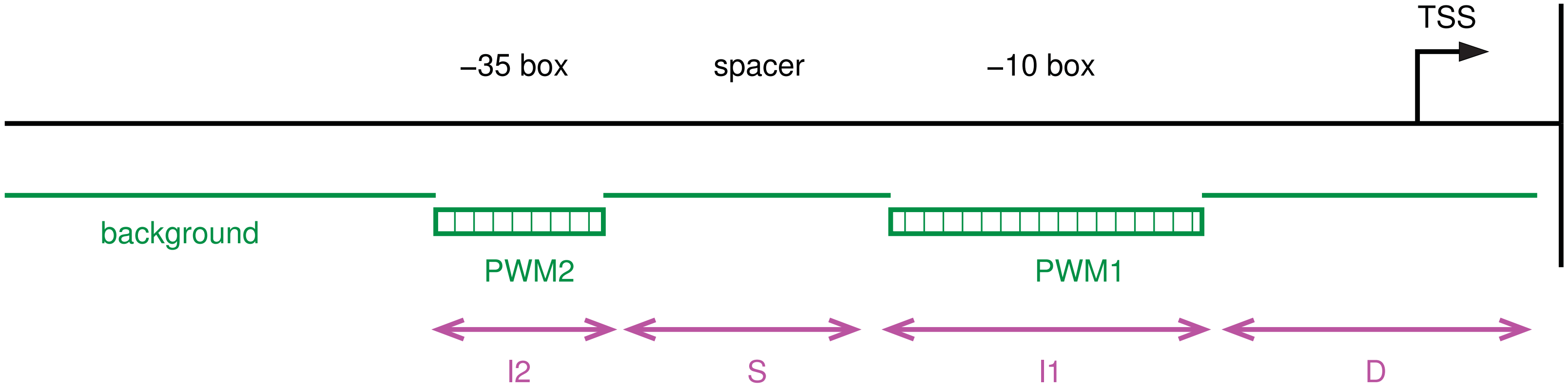


A 'promoter tree' is built by hierarchical clustering using average linkage on the dissimilarity matrix $d_{i,j} = (1 - r_{i,j})/2 \in [0, 1]$ where $r_{i,j}$ is the correlation between activities of promoters i and j.

Identifying Sigma-binding site sequence motifs

Sequence modeling[4]

- the model expresses $\mathbb{P}(x_i \mid U_i = k)$, the probability of sequence x_i given the presence of a Sigma-binding site motif of type $U_i = k$.



- a probability is associated to each motif $\mathbb{P}(U_i = k) = \alpha_k$, $\sum_k \alpha_{k=1}^K = 1$.

Searching for binding sites in a set of n sequences

- motif finding based on parameter estimation
- binding site predictions based on computation of $\mathbb{P}(U_i = k \mid x_i) \propto \mathbb{P}(x_i \mid U_i = k)\alpha_k$ for each sequence $i \in \{1, \ldots, n\}$.

[1] Sequence model and transdimensional MCMC algorithm adapted from P. Nicolas, A.-S. Tocquet, V. Miele, F. Muri (2006) A reversible jump Markov chain Monte Carlo algorithm for bacterial promoter motifs discovery. J Comput Biol. 13. 651-67.

Identifying sequence motifs: taking into account the correlation tree

We introduce a joint model where the motif allocations $U_1^n = (U_1, U_2, \ldots, U_n)$ result from an "evolution" along the tree.

- Change-points follow a Poisson process with rate λ along the branches of the tree.
- At each change-point the new value of the allocation variable is drawn according to the proportions $\alpha = (\alpha_1, \ldots, \alpha_K)$.
- Allocation is allowed to change at the leaf level with probability ϵ.

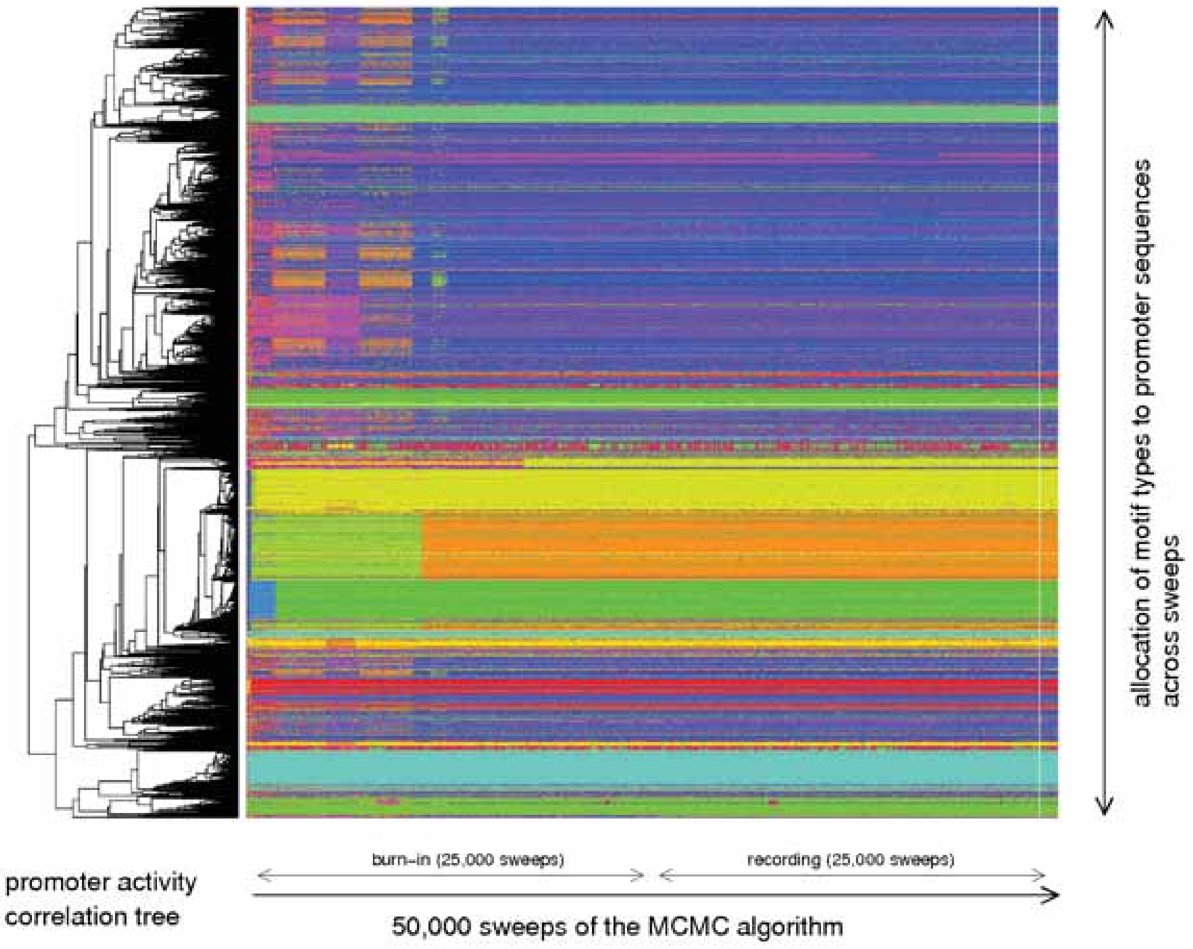
$$\mathbb{P}(U_1^n = u_1^n) \quad = \quad \sum_{(v)} \left[\pi_\alpha(v_{\text{root}}) \prod_{j \in \text{nodes}} \pi_{\lambda,\alpha}(v_{a_j} \to v_j) \prod_{i \in \text{leaves}} \pi_{\epsilon,\alpha}(v_{a_i} \to u_i) \right]$$

where v_j is the motif allocation variable associated with internal node j of the tree, a_j is the ancestor of node j.

$$\begin{aligned} \pi_{\lambda,\alpha}(v_{a_j} \to v_j) &= (1 - e^{-\lambda d_j})\mathbb{I}\{v_j = v_{a_j}\} + e^{-\lambda d_j} \alpha_{v_j} \\ \pi_{\epsilon,\alpha}(v_{a_i} \to u_i) &= (1 - \epsilon)\mathbb{I}\{u_i = v_{a_i}\} + \epsilon \alpha_{u_i} \end{aligned}$$

All parameters are estimated jointly with the MCMC algorithm. Only two additional parameters compared to the classical mixture model λ and ϵ.
The approach is very different from the "regression" perspective adopted by others to identify motifs that explain the expression patterns (REDUCE, FIRE, . . .).

Behavior of the MCMC algorithm, with $K = 20$ motifs

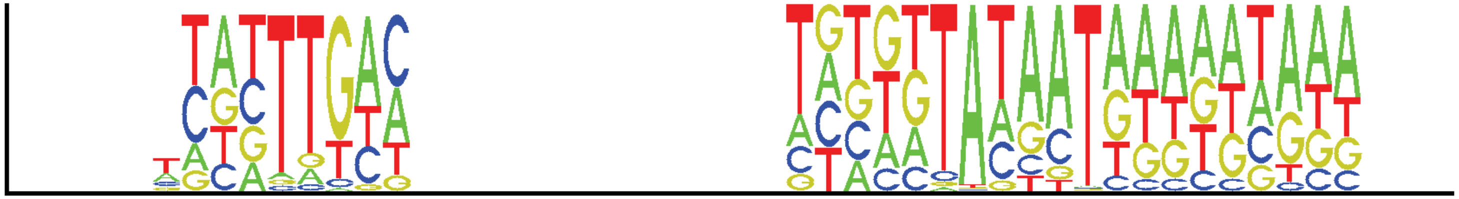


Comparison with known Sigma factor binding sites

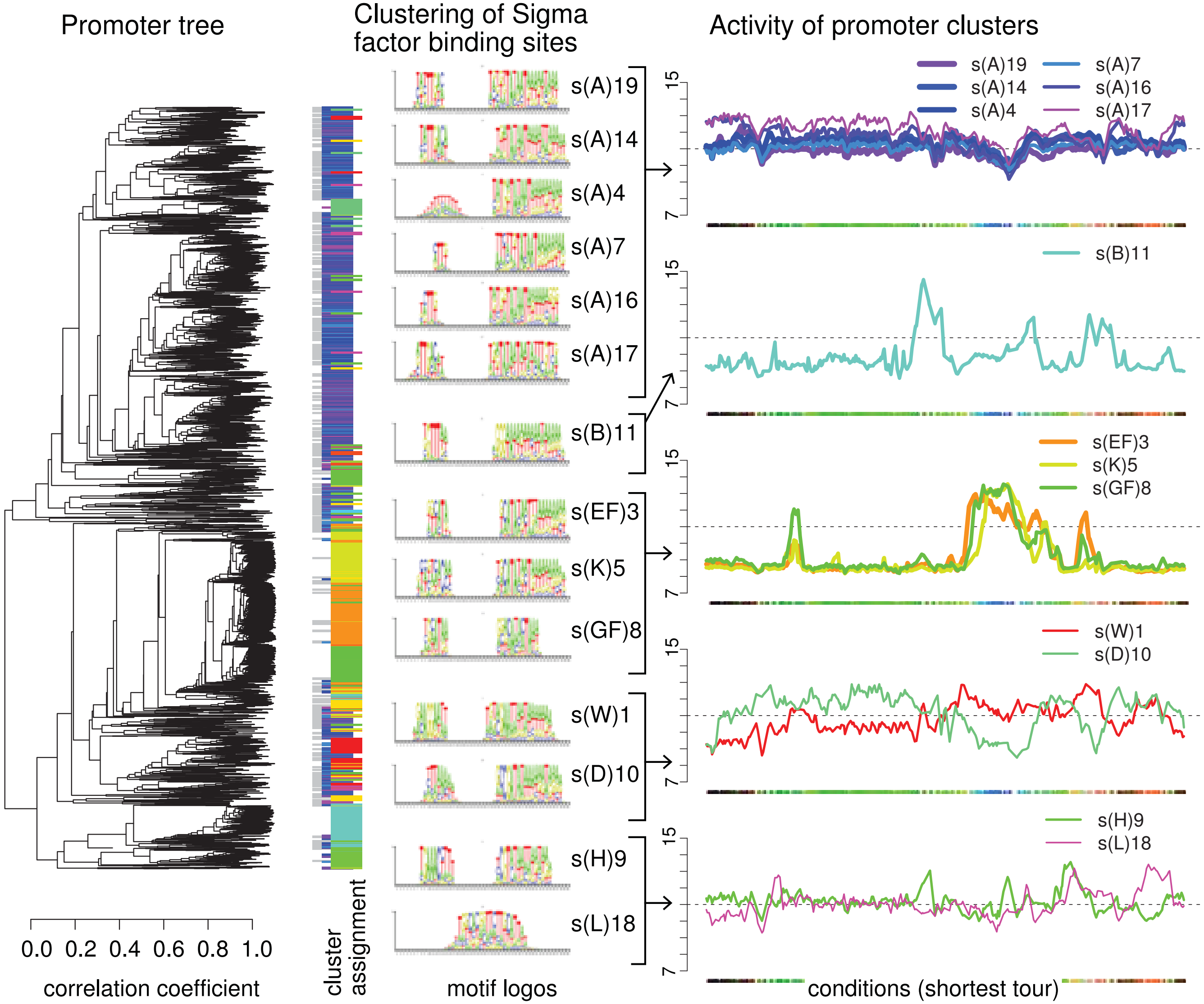
DBTBS: a database of transcriptional regulation in *Bacillus subtilis*

DBTBS	M19	M14	M4	M3	M7	M5	M16	M8	M11	M13	M17	M9	M1	M15	M10	-	M2	M18	M20	M6	M12
-	401	369	349	213	218	170	170	134	127	113	80	43	63	72	48	44	16	11	12	4	5
SigA	59	90	49	1	33	1	22	0	1	0	19	0	1	0	1	1	0	0	0	7	0
SigB	0	0	0	0	0	0	0	0	44	0	0	0	0	0	0	0	0	0	0	0	0
SigD	0	0	0	0	1	0	0	0	0	0	1	0	0	0	23	0	0	0	0	0	0
SigE	0	0	1	54	0	4	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
SigF	0	0	0	8	0	0	0	10	1	0	0	0	0	1	0	0	0	0	0	0	0
SigG	0	0	0	0	0	0	0	42	0	0	0	0	0	0	0	0	0	0	0	0	0
SigH	0	0	0	1	0	0	1	1	0	0	0	1	12	0	0	0	0	0	0	0	0
SigI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
SigK	1	0	0	1	0	38	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
SigL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0
SigM	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
SigW	0	0	1	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0
SigX	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
SigY	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0

Sequence logos to represent motifs



Average activity of the promoters for each Sigma binding site motif



Relating promoter activity to Sigma factor activity

The activity $m_{i,t}$ of promoter i in experiment t is modeled as a linear function of the mean activity $a_{k(i),t}$ of all the promoters with the same motif $k(i)$

$$m_{i,t} \quad = \quad \alpha_i + \beta_i a_{k(i),t} + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, \sigma_{1,i}^2).$$

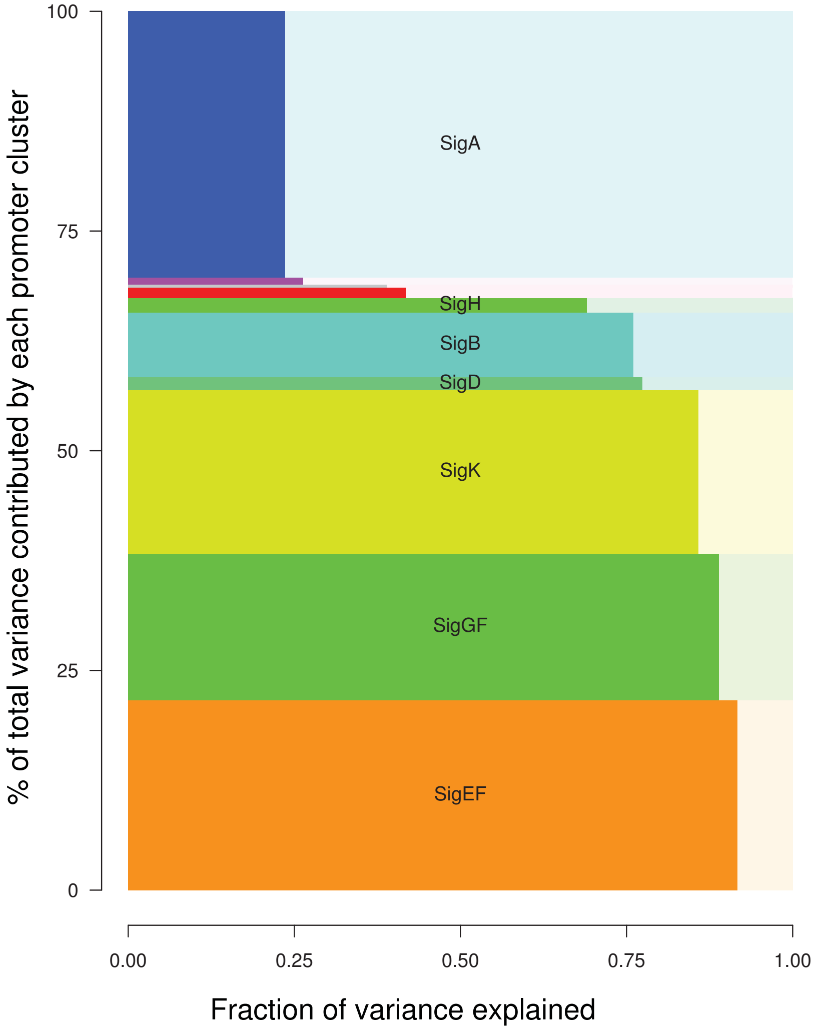
To be compared with

$$m_{i,t} \quad = \quad \alpha'_i + \epsilon'_{i,t}, \quad \epsilon'_{i,t} \sim \mathcal{N}(0, \sigma_{0,i}^2).$$

The activity of each promoter *i* can be summarized with three numbers

- α_i and β_i quantify the “strength” of the promoter and its “sensitivity” to the activity of the Sigma factor.
- $1 - \sigma_{1,i}^2 / \sigma_{0,i}^2$ the fraction of variance that is explained by the activity of the Sigma factor.

Fraction of variance linked to Sigma factors

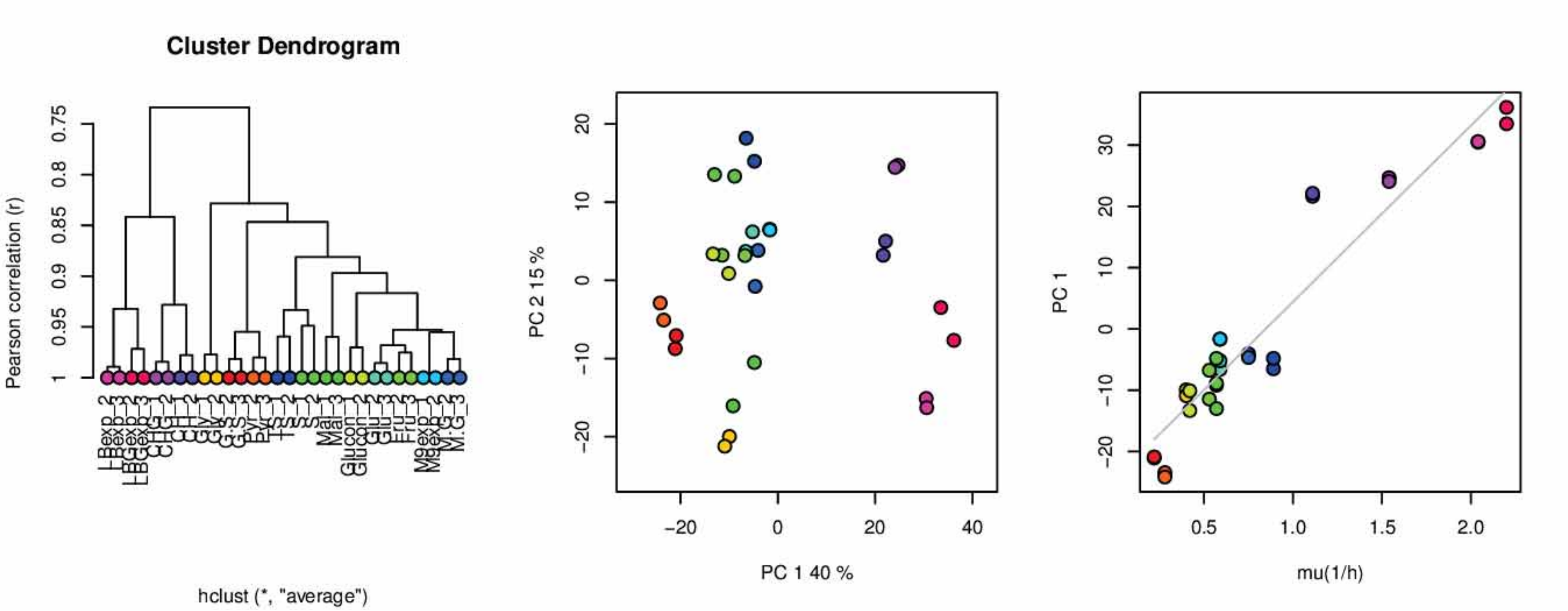


- 66% of the total variance can be linked to direct regulation by Sigma factors.
- Variance of SigA-dependent promoters is poorly explained; due to the activity of other transcription factors (>150 in *B. subtilis*), or not.

What is the actual share of the transcription factors in the regulation of the SigA-dependent promoters ? Let’s attempt to explain regulations without transcription factors !

Ongoing joint work with Vincent Fromion. Motivation: explain tuning of transcription levels wrt growth rate.

Data-set 30 hybridizations assessing transcriptome activity during exponential growth ($N_t = N_0 e^{\mu t}$, $T_g = \log 2/\mu$) in 15 growth media; we focus our analysis on 1514 SigA-promoter with TSS known at the 1 bp resolution.



The hardware hypothesis

Our guess on this level of transcription regulation

- unlikely to rely on specific regulators.
- might be hard-coded in gene-specific kinetic parameters governing the rates of synthesis and degradation.

$$
\begin{array}{rcccccc}
G_i + P & \overset{k_i^+,k_i^-}{\leftrightarrows} & (G_i, P) & & & & \\
 & & (G_i, P) & \overset{k_i^s}{\rightarrow} & M_i & & \\
 & & & & M_i & \overset{k_i^d}{\rightarrow} & \emptyset
\end{array}
$$

Predictions of the amount of mRNA *i* is obtained by solving the “steady-state” relationships

$$
\begin{aligned}
k_i^+[G_i][P] - k_i^-[G_i.P] - k_i^s[G_i.P] &= 0 \quad \left(\frac{d[G_i.P]}{dt} = 0\right) \\
k_i^s[G_i.P] - k_i^d[M_i] &= 0 \quad \left(\frac{d[M_i]}{dt} = 0\right)
\end{aligned}
$$

The core equation of the hardware model

This yields to the amount of messenger M_i for gene i written as a function of the amount of ’free’ polymerase p

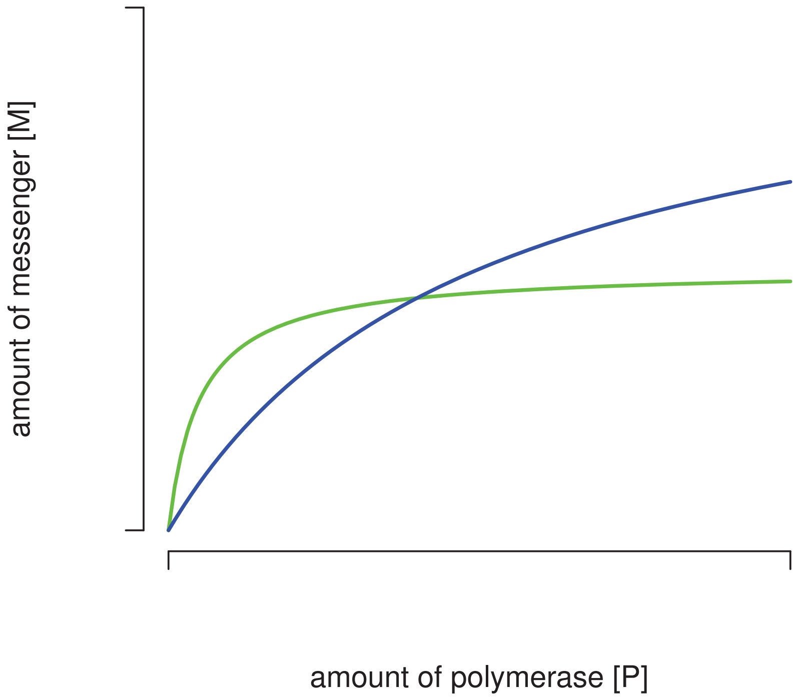
$$[M_i] \quad = \quad \frac{k_i^s.g_i.p}{\frac{k_i^- + k_i^s}{k_i^+} + p} \cdot \frac{1}{k_i^d}$$

where

- $g_i = [G_{i,tot}] = [G_i] + [G_i.P]$, $p = [P]$
- $k_i^s g_i$ is the maximal rate of synthesis,
- $\frac{k_i^- + k_i^s}{k^i_+}$ gives the concentration of the polymerase that allows half of the maximal rate (i.e. Michaelis constant),
- $\frac{1}{k_i^d}$ is proportional to mRNA half-life.

Link between *p* and expression data

This equation allows a variety of expression profiles and can thus model changes in relative mRNA concentrations.



Our goal here is to estimate the amount of free polymerase *p* from (relative) mRNA concentrations.

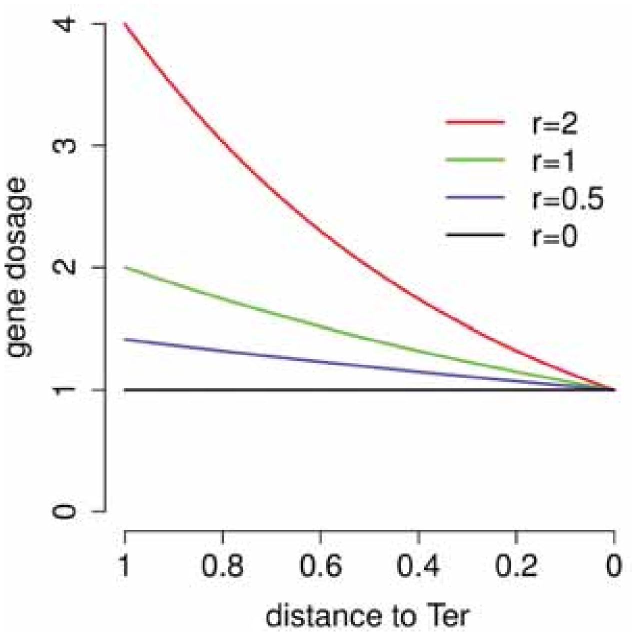
Probabilistic model for inference

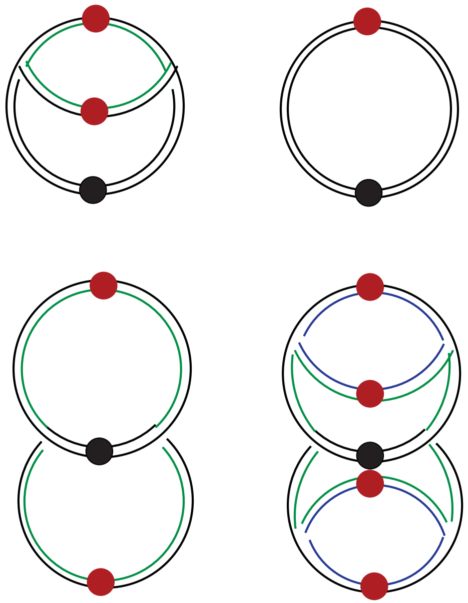
We make a log transformation (as usual) to stabilize the variance wrt expression level

$$\begin{aligned} m_{i,t} \quad &= \quad \log_2\left(\frac{k^s_{i,t}k^+_i}{k^-_i k^d_i}\right) + \log_2(g_{i,t}) + \log_2(p_t) + z_t \\ &\quad - \log_2\left(1 + \frac{k^s_{i,t}}{k^-_i} + \frac{k^+_i}{k^-_i}p_t\right) + \epsilon_{i,t} \end{aligned}$$

- where z_t is a ’normalizing’ constant (we measure relative abundances).
- $\epsilon_{i,t} \sim \mathcal{N}(0, \sigma_i^2)$ is the “error term” whose variance is gene-specific (accounting for other levels of regulation and noise). This gives less weight to the genes whose expression is poorly explained by the model.

A simple model for gene dosage ($g_{i,t}$)



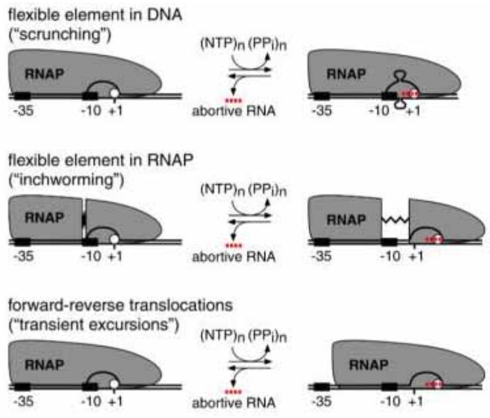


$$g_{i,t} \quad = \quad 2^{c_i r_t}$$

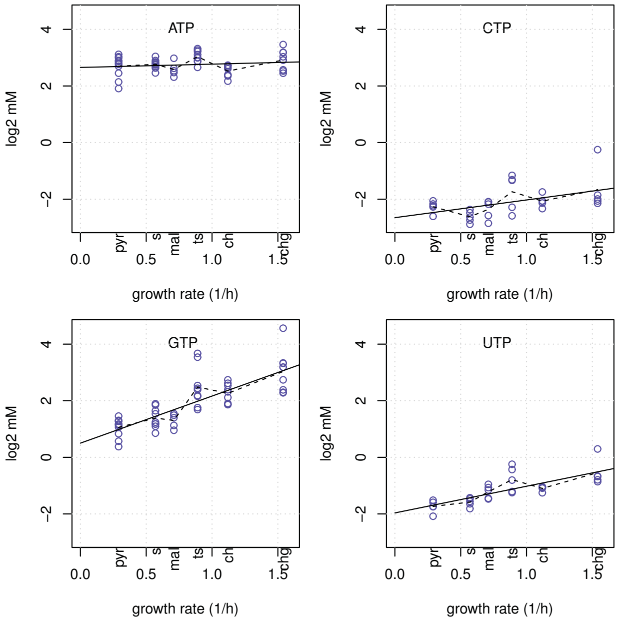
where

- c_i is the distance of the locus relative to the replication terminus (1 for Ori, 0 for Ter).
- $r_t \approx T_c / T_g$ reflects the chromosome replication rate, corresponds to number of replication forks on 1/2 chromosome.

A simple model for rNTP-dependent abortive initiation



From Revyakin et al., 2006



We set

$$k_{i,t}^{s} \quad = \quad k_{i}^{s} \prod_{k=1}^{k_{\max}} (\tau_{x_{i,k},t})^{a_k} \ ,$$

with $c_k \in (0, 1)$ and where $x_{i,k}$ is the rNTP needed for elongation at position k of transcript i and τ_x refers to the concentration of rNTP x in condition t.

This formula arises as a crude approximation of the rate of success in a multi-step competition between elongation (whose rate is proportional to τ_x) and abortion.

From biological to statistical parameters

the initial probabilistic model

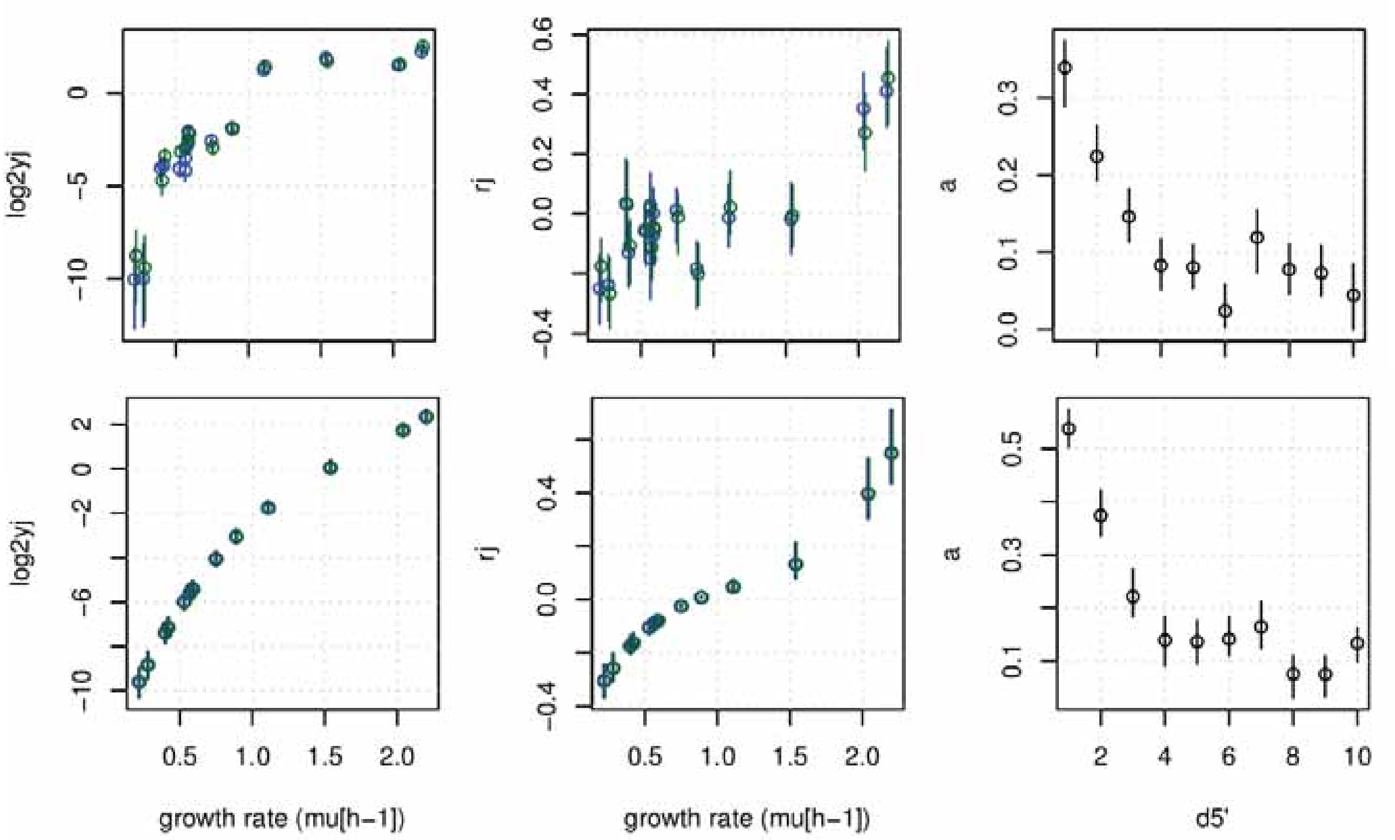
$$\begin{aligned} m_{i,t} \quad &= \quad \log_2\left(\frac{k_{i,t}^s k_i^+}{k_i^- k_i^d}\right) + \log_2(g_{i,t}) + \log_2(p_t) + z_t \\ &\qquad - \log_2\left(1 + \frac{k_{i,t}^s}{k_i^-} + \frac{k_i^+}{k_i^-} p_t\right) + \epsilon_{i,t} \end{aligned}$$

is rewritten

$$m_{i,t} \quad = \quad \alpha_i + c_i \rho_t + \sum_k a_k \log \tau_{x_{i,k},t} + \zeta_t - \log_2(1 + \beta_i \psi_t + \gamma_i \nu \prod_k (\tau_{x_{i,k},t})^{a_k}) + \epsilon_{i,t}$$

There are a number of identifiability issues that makes the mapping between the parameters that we can estimate and their biological counterparts not trivial.

Preliminary inference results on the “hardware model”



The model captures $\approx 50\%$ of the total variance, the fit of the more constrained model where r_t and ψ_t are written as polynomial functions of the growth-rate is almost as good.

Results to be extended on larger data sets and analyzed from a biological perspective . . .

Concluding remarks and acknowledgments

Illustration of

- examples of current research questions in computational biology
- the diversity of uses of latent variable models in this field (from data processing to biological/mechanistic modeling).

Acknowledgments

- Bogdan Mirauta and Hugues Richard for part 1.
- Partners of the Basysbio and Basynthec projects for parts 2 and 3 (in particular Vincent Fromion for part 3 and the wet-biologists Philippe Noirot, Etienne Dervyn and Ulrike Mäder).