Quantitative and qualitative analysis of the *Bacillus subtilis* whole-transcriptome across lifestyles

Pierre Nicolas

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

Journées NETBIO
November 20, 2012

A “big” joint work

Pierre Nicolas[1]*, Ulrike Mäder[2,3]*, Etienne Dervyn[4]*, Tatiana Rochat[4], Aurélie Leduc[1], Nathalie Pigeonneau[4], Elena Bidnenko[4], Elodie Marchadier[4], Mark Hoebeke[1], Stéphane Aymerich[4], Dörte Becher[2], Paola Bisicchia[5], Eric Botella[5], Olivier Delumeau[4], Geoff Doherty[6], Emma L. Denham[7], Mark J. Fogg[8], Vincent Fromion[1], Anne Goelzer[1], Annette Hansen[5], Elisabeth Härtig[9], Colin R. Harwood[10], Georg Homuth[3], Hanne Jarmer[11], Matthieu Jules[4], Edda Klipp[12], Ludovic Le Chat[4], François Lecointe[4], Peter Lewis[6], Wolfram Liebermeister[12], Anika March[9], Ruben A.T. Mars[7], Priyanka Nannapaneni[3], David Noone[5], Susanne Pohl[10], Bernd Rinn[13], Frank Rügheimer[14], Praveen K. Sappa[3], Franck Samson[1], Marc Schaffer[2], Benno Schwikowski[14], Leif Steil[3], Jörg Stülke[15], Thomas Wiegert[16], Kevin M. Devine[5], Anthony J. Wilkinson[8], Jan Maarten van Dijl[7], Michael Hecker[2], Uwe Völker[3], Philippe Bessières[1], and Philippe Noirot[4]¶

Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in *Bacillus subtilis*. Science. 335. 1099-1103.

SysMO
Systems Biology of Microorganisms

BaSysBio

This talk intends to give you an overview of this study, with particular attention to the methods that we developed for data-analysis.

Systematic exploration of *B. subtilis* transcriptional landscape

- 1 “wild-type” strain, maybe better called “prototype” strain.
- 1 array design (Basysbio tiling array, Nimblegen technology) : 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 (experiments).

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

Outline

Transcriptional landscape estimation from a single tiling array hybridization

Building the *B. subtilis* transcriptional “parts list” from a collection of tiling array hybridizations

Relating transcriptome dynamics to the genome sequence

A focus on antisense transcription

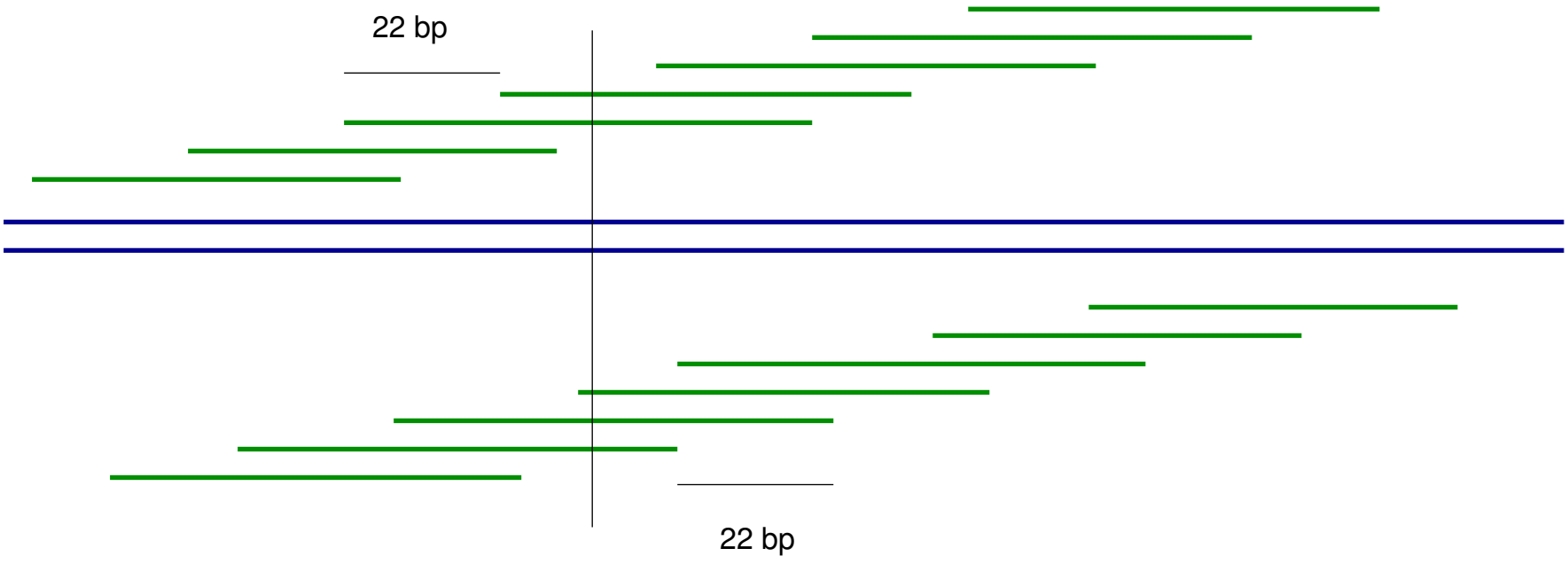
Transcriptional landscape estimation from a single tiling array hybridization

Building the *B. subtilis* transcriptional “parts list” from a collection of tiling array hybridizations

Relating transcriptome dynamics to the genome sequence

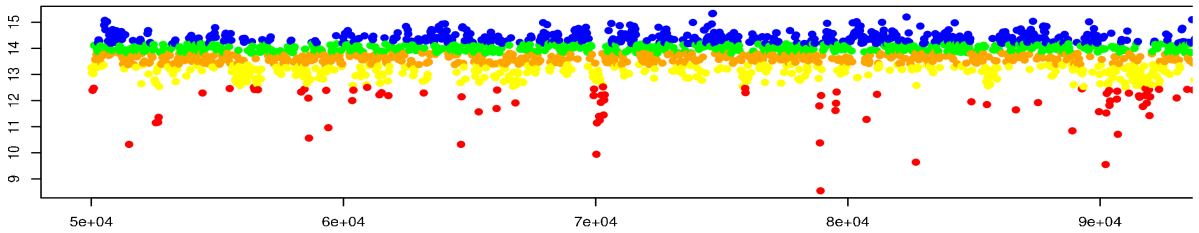
A focus on antisense transcription

BaSysBio tiling array

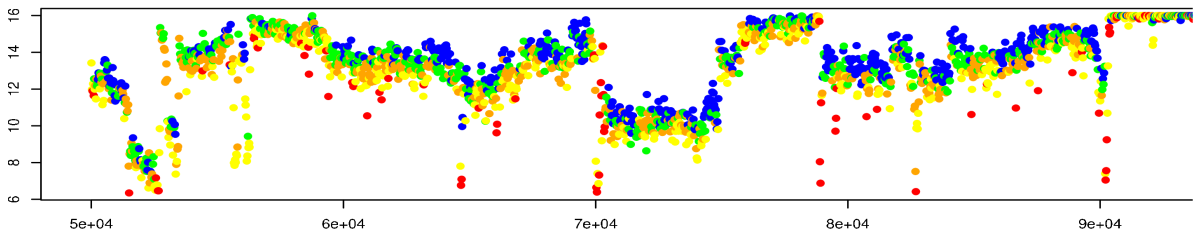


- $\approx 380,000$ probes tiling the 4.2 Mbp *Bacillus subtilis* genome.
- Long probes (45-65 nt), lengths adjusted to achieve relative homogenous affinity (Tm).

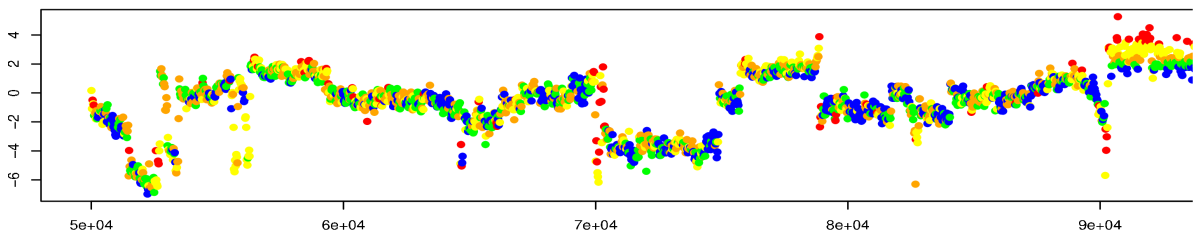
Expression data – log-transformed signal on 50Kbp



log(genomic DNA) from x4 pooled data



log(mRNA)

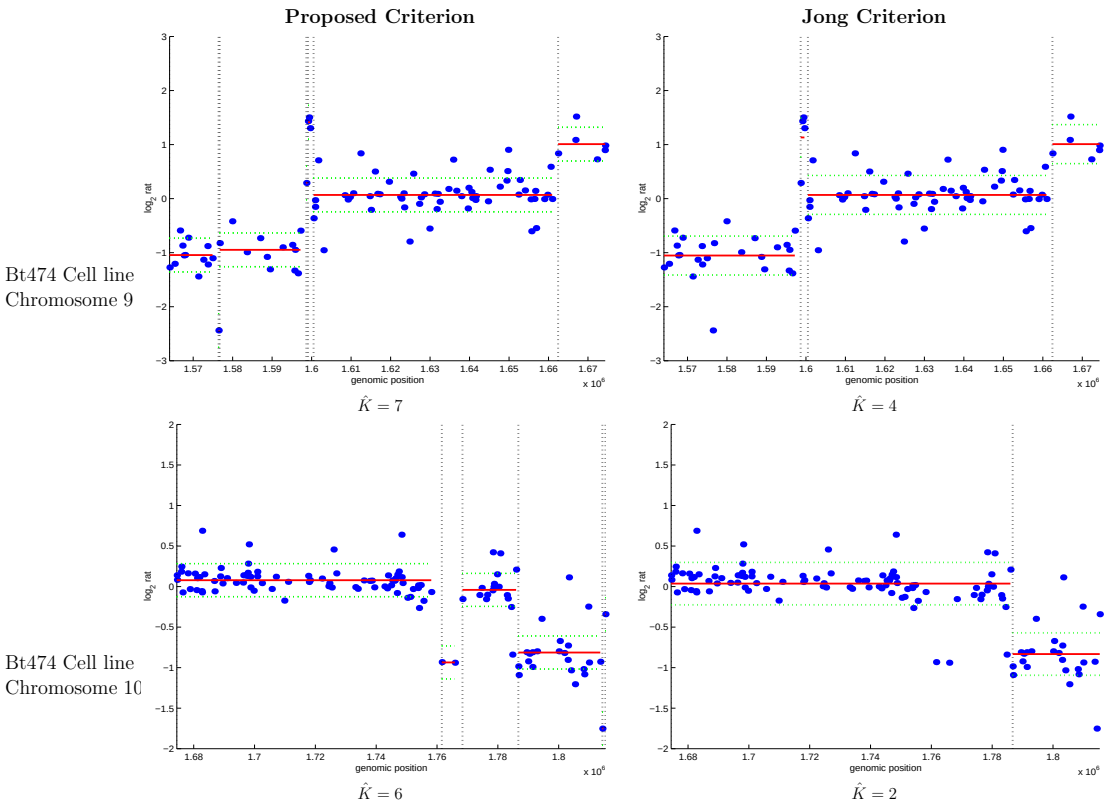


log(mRNA) – log(genomic DNA)

↪ probe affinity is variable, despite the adjustment of probe lengths.

Models for transcriptional landscape reconstruction

Piecewise-constant linear regression (Picard *et al.*, 2005).



Minimizes

$$G(e_1, \ldots, e_{S-1}) \quad = \quad \sum_{s=1}^{S} \sum_{t=e_{s-1}}^{e_s - 1} (x_t - \bar{x}_s)^2,$$

where S is the number of segments (given), e_s is the end of segment s and x_t is the signal at probe t.

Motivations for an alternative method

Confidence interval construction.

- signal level with given breakpoints (easy!)
- breakpoint position (see Huber *et al.*, 2006)
- confidence band for the underlying signal accounting for uncertainty on breakpoints?

Choice of the number of breakpoints.
Model selection problem (non-trivial, see Picard *et al.*, 2005).

Hypothesis of piecewise-constant signal. Shift and drift?

Normalization: how to use the gDNA signal ?
Huber *et al.*, 2006 proposed the following preprocessing step

$$y'_t = \frac{y_t - b(y_t)}{\text{gDNA}_t},$$

with y_t original data (non-log), $b(y_t)$ an estimate of the contribution of background noise.

Our approach

To model not only the noise but also the variations of the underlying signal.

- can solve the problem of confidence band construction (at least conceptually).
- alleviates the problem of choosing the number of breakpoints (at least in principle) a parameter -to be estimated- corresponds to the rate of breakpoints.

Accounting for the correlation between the underlying signal at adjacent probes naturally leads to HMMs.

Other HMMs in related contexts

- CGH data, small number of hidden states (Fridlyand *et al.*, 2004; Marioni *et al.*, 2006, Stjernqvist *et al.*, 2007)
- Classification expressed vs. non-expressed regions (Munch *et al.*, 2006; Du *et al.*, 2006)

Here we aim at “denoising” the data via the modeling of a continuous-valued underlying signal.

Hidden Markov Models

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

Simplest "Emission" model (a more sophisticated model is implemented)

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

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 $O(nK^2)$ in their general form).

$\hookrightarrow$ Use a transition matrix structure that allows algorithms in $O(nK)$ 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

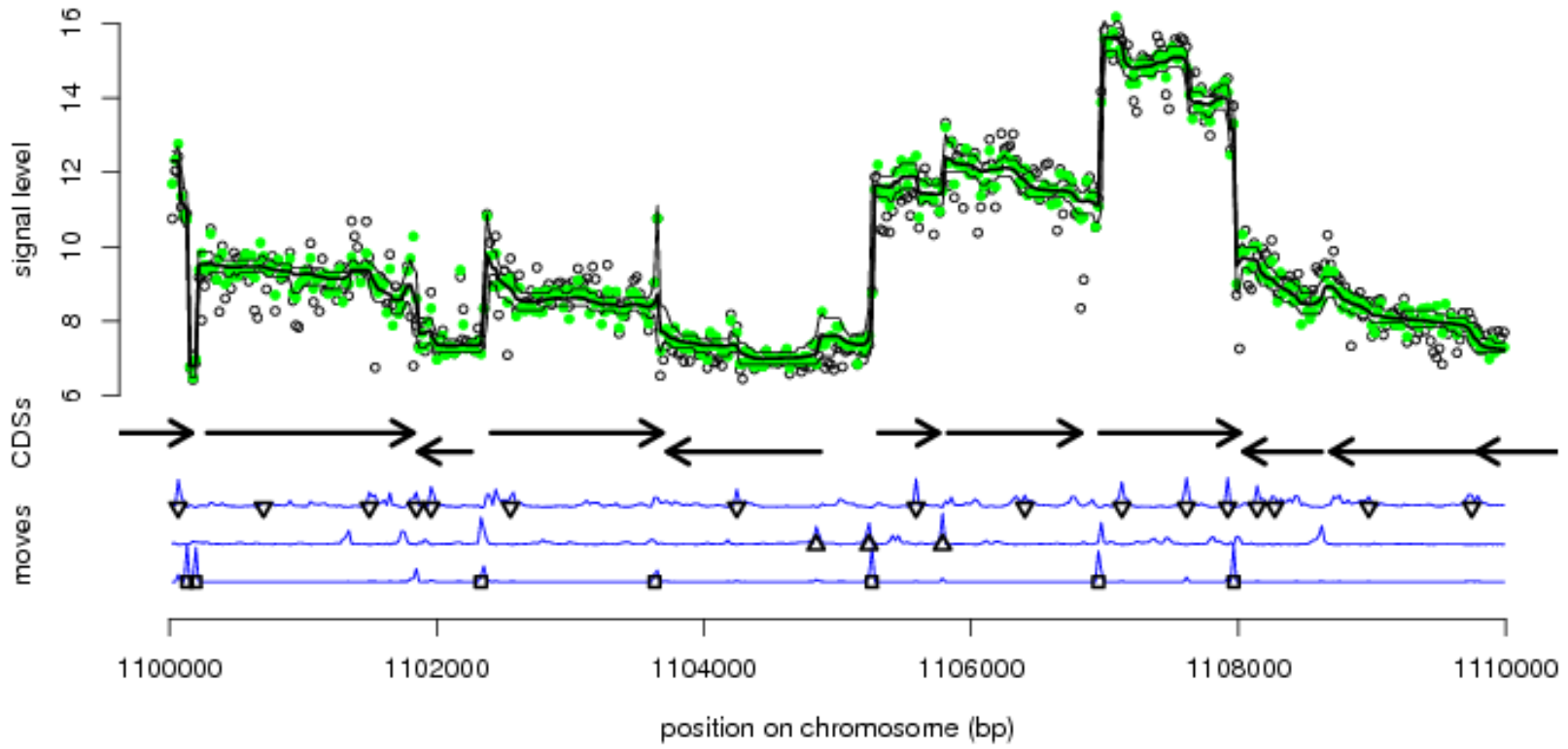
$$\begin{aligned}\pi(u_t, u_{t+1}) = {} & \alpha_n \mathbb{I}_{\{u_{t+1}=u_t\}} + \alpha_s \eta(u_{t+1}) \\ & + \alpha_u \mathbb{I}_{\{u_{t+1}>u_t\}} \lambda_u^{\frac{u_{t+1}-u_t}{h}-1}(1-\lambda_u) \\ & + \alpha_d \mathbb{I}_{\{u_{t+1}<u_t\}} \lambda_d^{\frac{u_t-u_{t+1}}{h}-1}(1-\lambda_d)\,,\end{aligned}$$

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

- 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 (HMM with continuous-valued underlying process).

Transcriptional landscape reconstruction

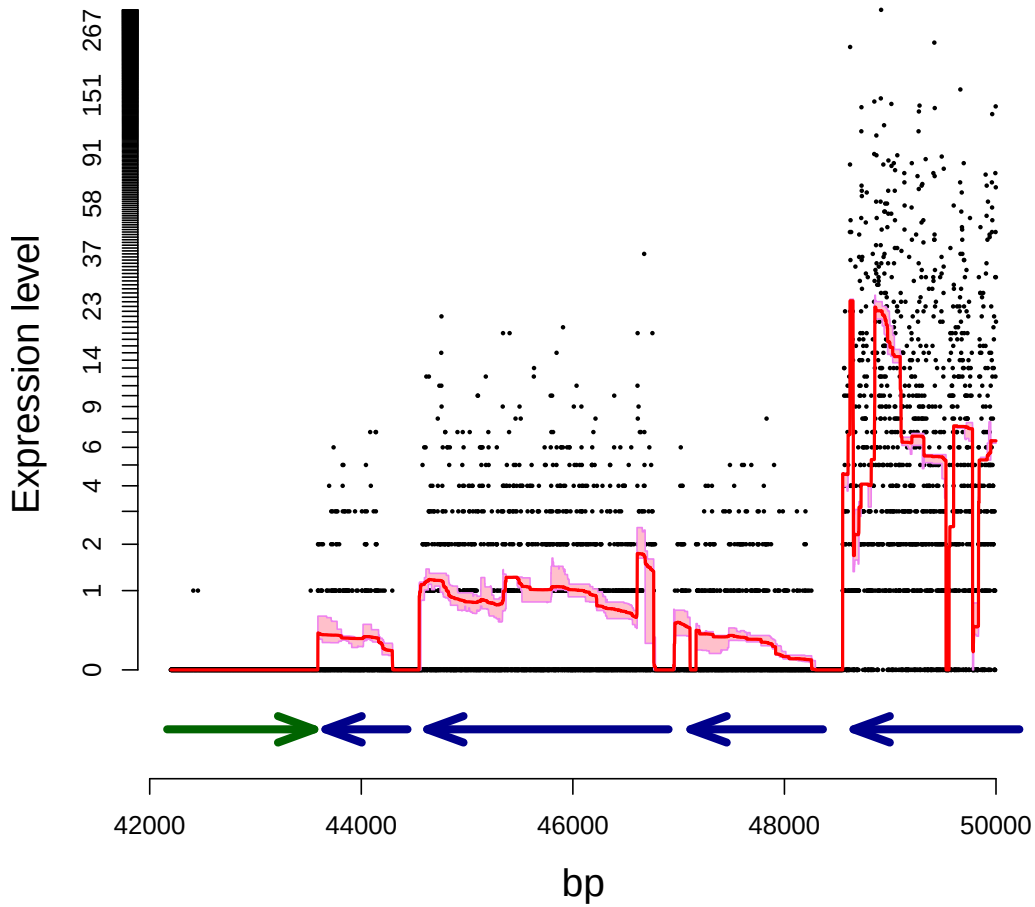


Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. P. Nicolas, A. Leduc, S. Robin, S. Rasmussen, H. Jarmer and P. Bessières. Bioinformatics. 2009. 25. 2341-2347

Ongoing work on RNA-Seq

Bogdan Mirauta (PhD student) and Hugues Richard (Laboratoire de Génomique des Microorganismes — UP6).

- A framework alleviates the need for discretization of the hidden state space.
- A generative State Space Model fitted by Sequential Monte Carlo algorithms.



Part II

Transcriptional landscape estimation from a single tiling array hybridization

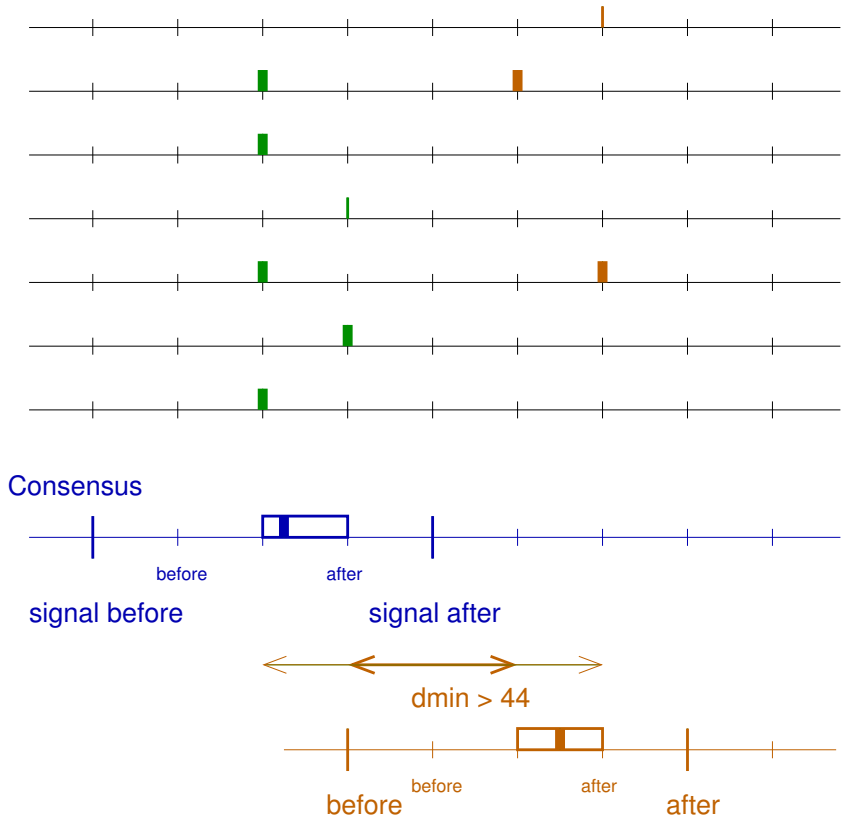
Building the *B. subtilis* transcriptional “parts list” from a collection of tiling array hybridizations

Relating transcriptome dynamics to the genome sequence

A focus on antisense transcription

Building catalogs of breakpoints - Method

- The confidence of a breakpoint is computed as the sum of the probability of upward (or downward) shift over two adjacent probes.
- A cutoff is applied on the confidence value
- Adjacent breakpoints across the hybs are merged



Building catalogs of breakpoints - Results

Upward shifts

cutoff p	#	DBTBS prom.	> 2 probes	single hyb.	within CDS	E(FP)
0.9975	2983	613/733	323	4	431	0.2
0.9950	3086	620	345	6	471	0.6
0.9900	3240	626	381	17	534	1.7
0.9800	3432	631	438	34	619	4.6
0.9600	3711	638	518	61	744	12.8
0.9200	4102	644	635	125	934	36.8

Downward shifts

cutoff p	#	Petrin term.	> 2 probes	single hyb.	within CDS	E(FP)
0.9975	1850	1411/3510	216	5	129	0.2
0.9950	1958	1462	256	8	152	0.6
0.9900	2123	1517	292	18	192	1.9
0.9800	2327	1564	353	28	250	4.9
0.9600	2614	1613	422	56	352	13.6
0.9200	3003	1666	522	106	484	37.5

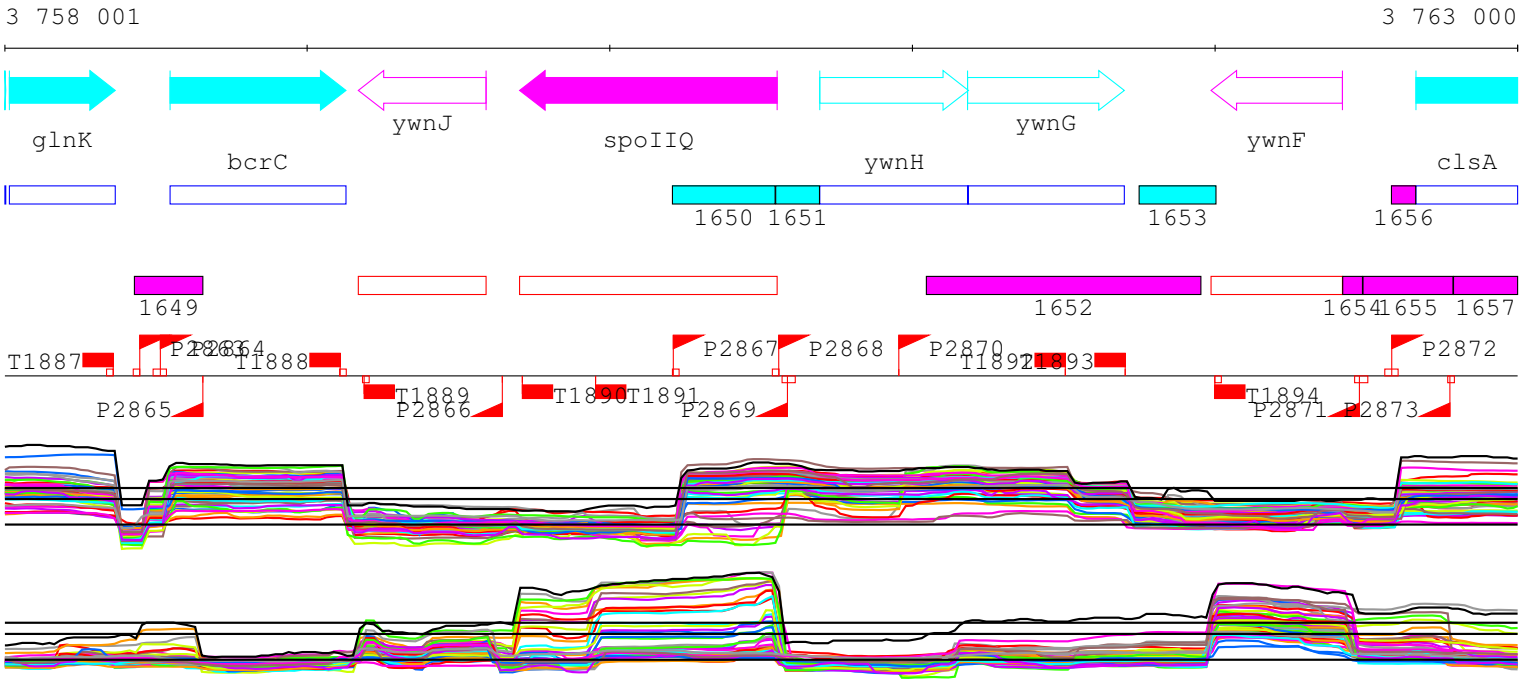
Catalog: 3240 putative promoters (upward shifts), 2123 putative terminators (down shifts).

Building a catalog of new transcripts - Method

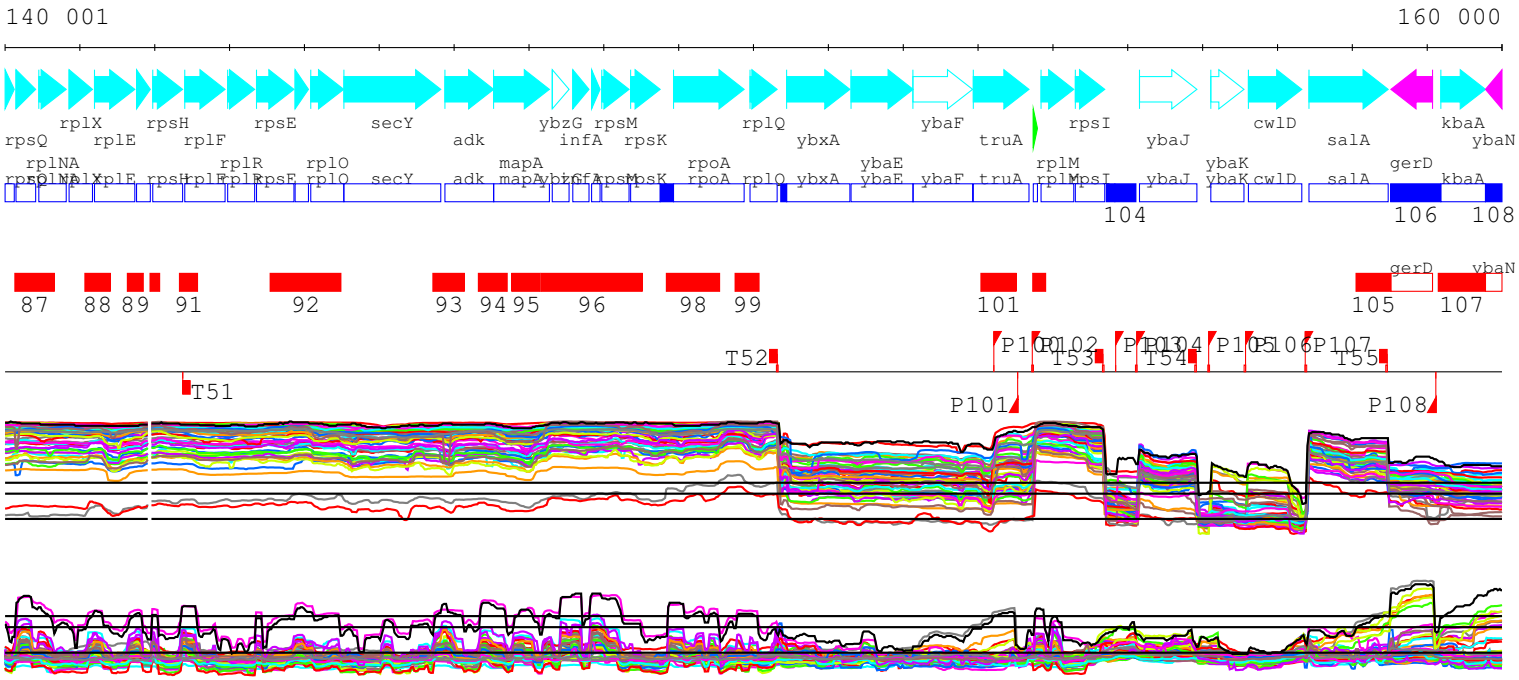
The local 95% confidence interval is compared to the overall median of the signal on the array (we expect less than 50% of a single strand to be expressed).

- We search for regions where the lower bound of this CI is 10$\times$ above the median in at least one hyb. 90.8% of the annotated CDSs are called with this cutoff.
- Regions are extended on the left and on the right as far as the CI is 5$\times$ above the median.
- Probes outside annotated CDS or RNA genes define unannotated transcribed regions.
- Regions without a single probe satisfying the 10$\times$ criterion are discarded.

Building a catalog of new transcripts - Results

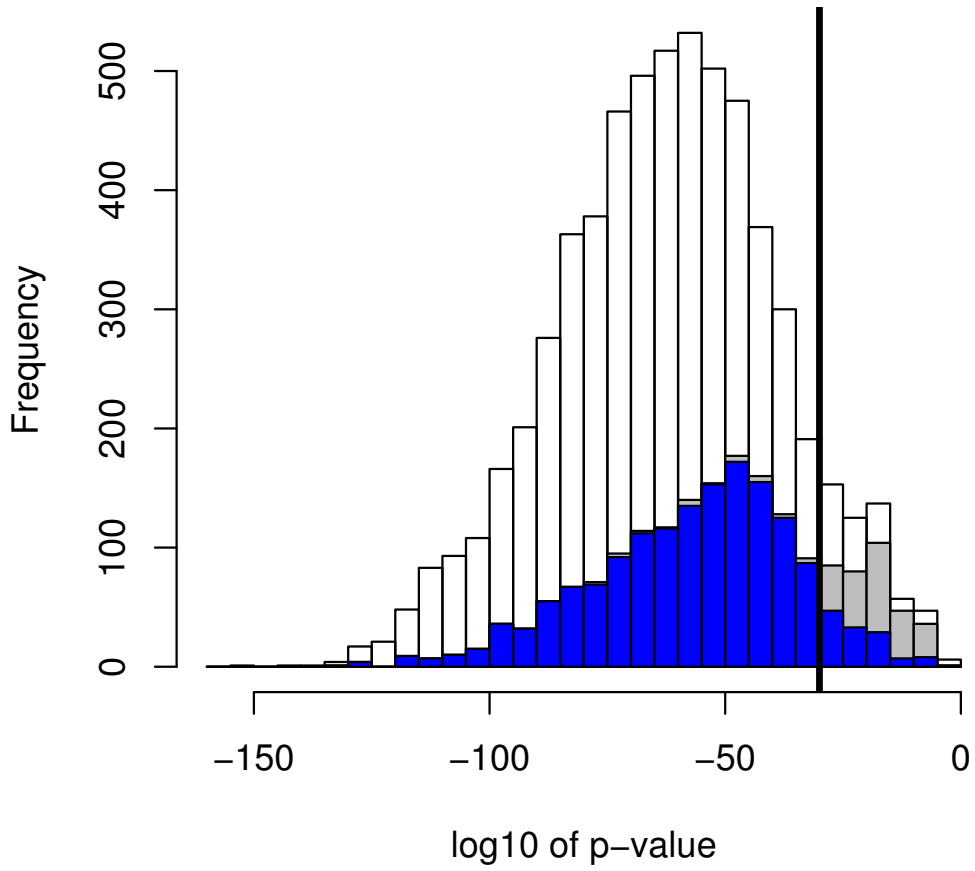


Antisense artifacts

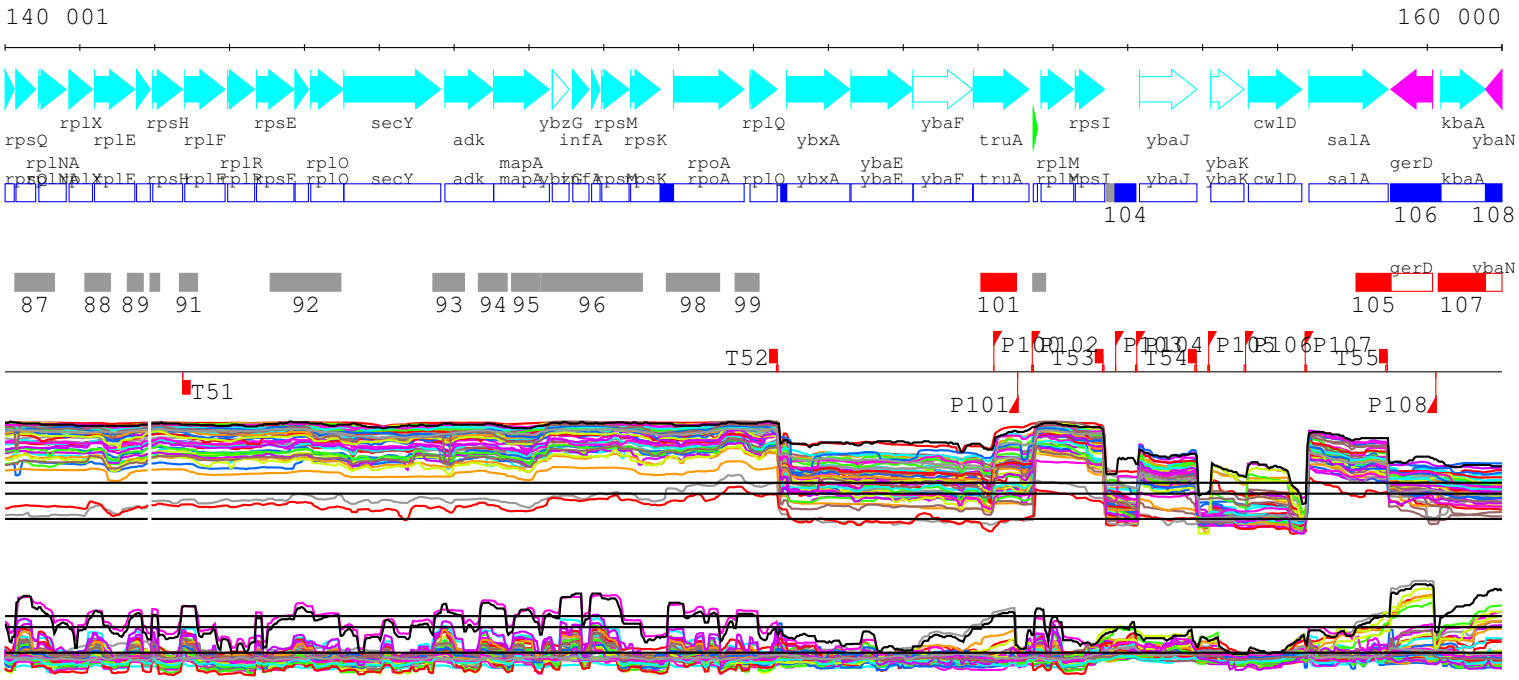


Trimming the artifacts

Based on p-value assessing the statistical significance of variations between biological conditions, 268 regions were proposed to be discarded (cutoff set to 10^{-30}). Final manual validation (+18 -42 and 15 merged into 6) lead to 1583 regions.

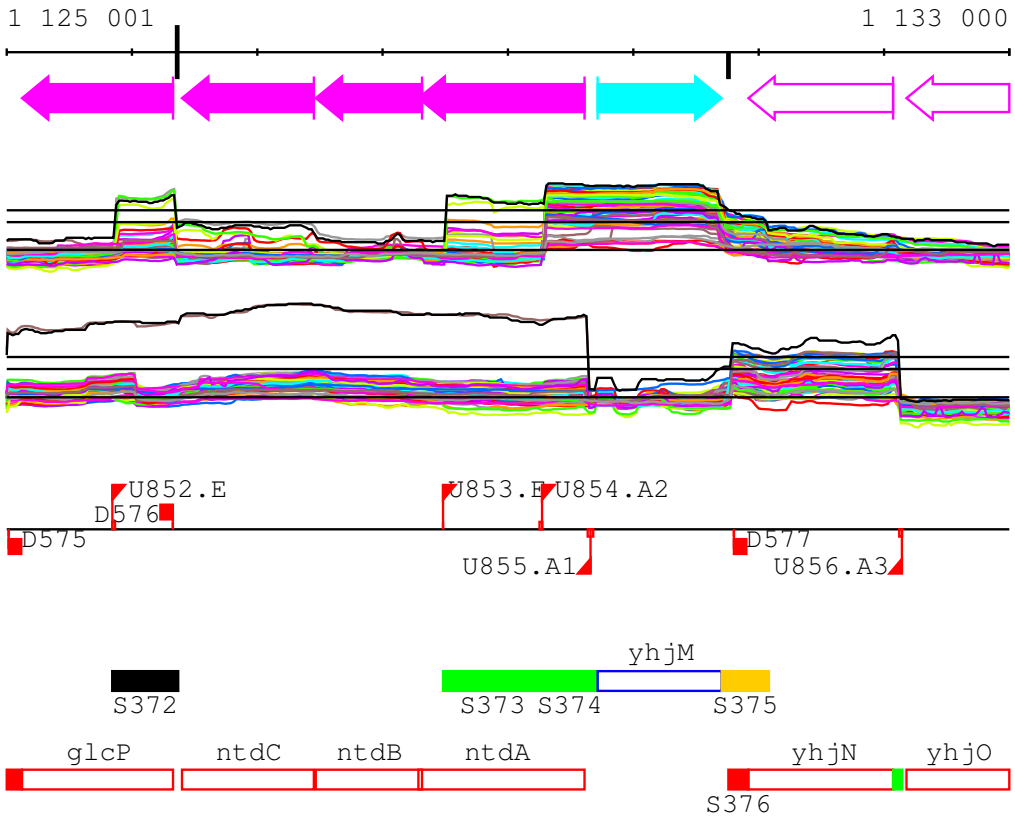


After trimming



Structural classification of transcriptional contexts

- 5’
- 3’UTR, 3’NT (no termination), 3’PT (partial termination)
- indep, indep-NT
- inter, intra

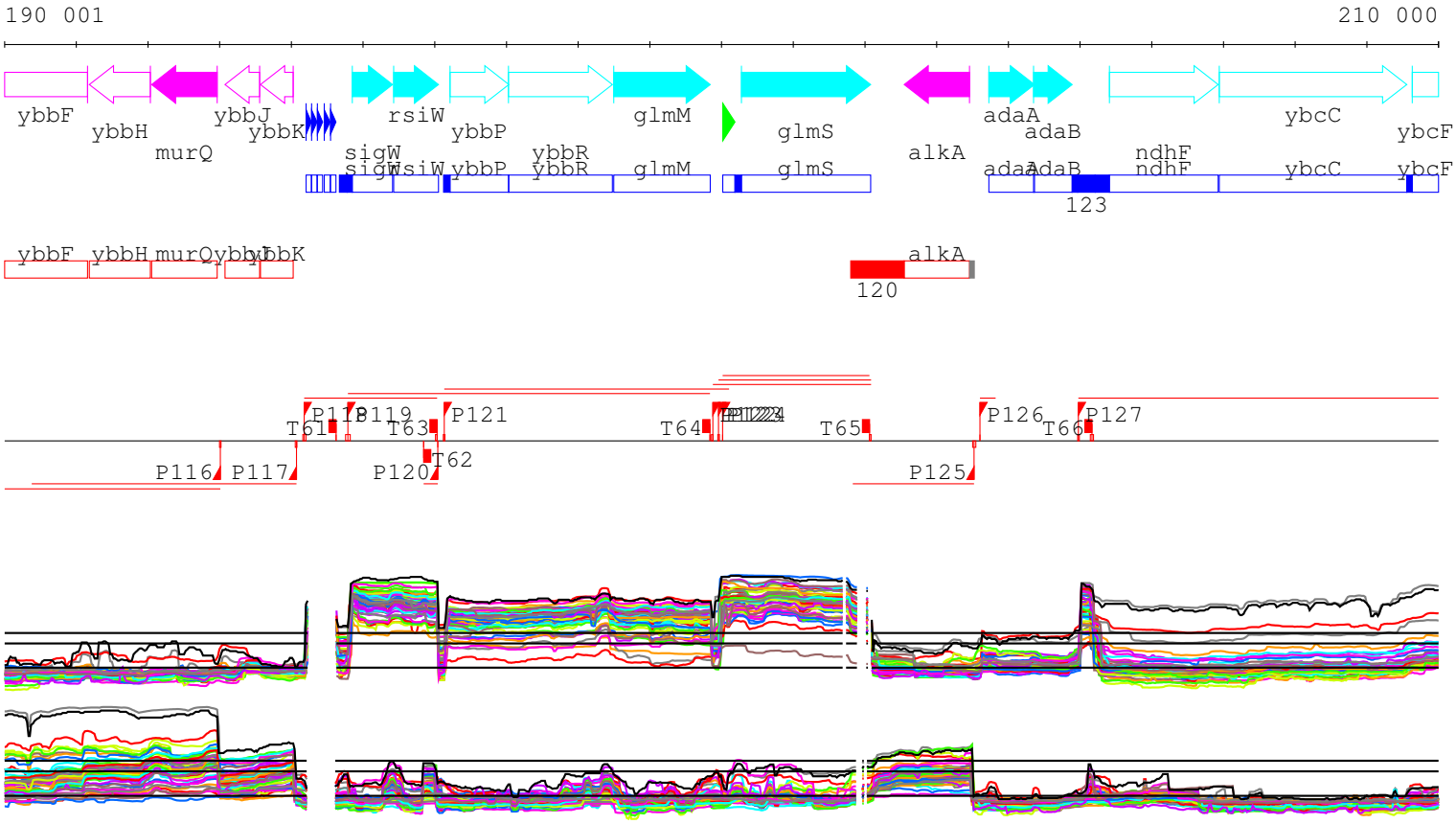


Summary of new regions

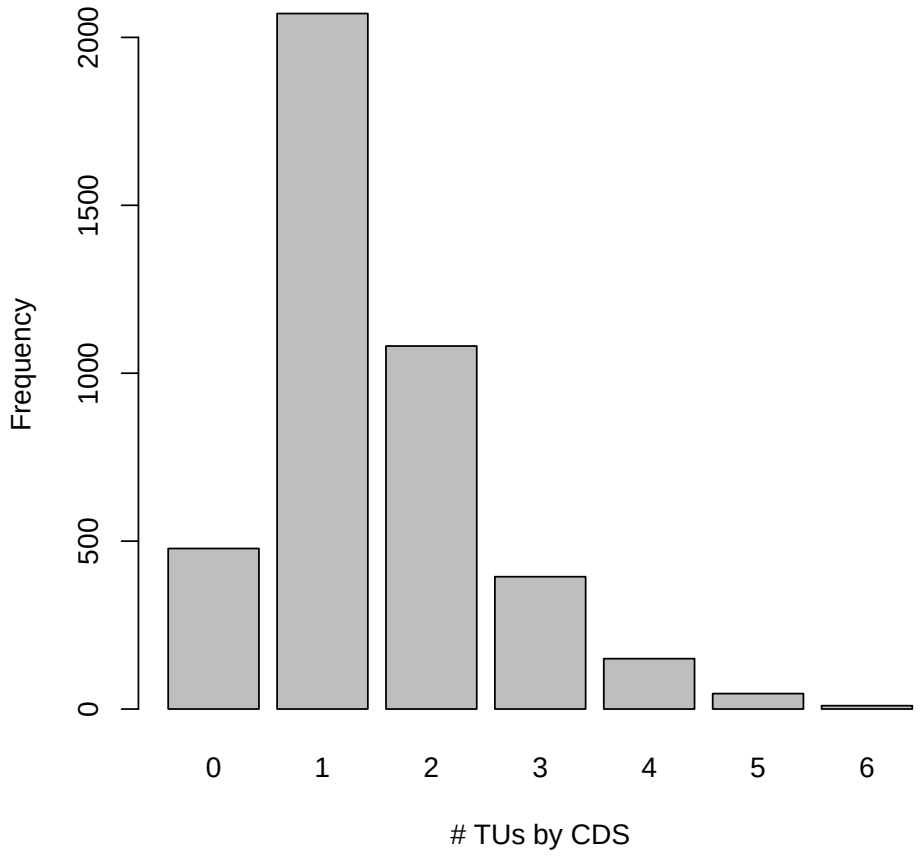
In addition to transcriptional context, we flagged antisenses and short unannotated CDSs.

	50 bp $\leq$ length $<$ 150 bp				length $\geq$ 150 bp			
Type	#	bp	u-CDS	AS	#	bp	u-CDS	AS
3'UTR	64	5,955	0	5	61	25,993	5	26
3'NT	2	180	0	0	44	48,060	1	40
3'PT	4	316	0	1	74	60,294	0	69
5'	462	44,310	2	5	214	90,165	13	85
indep	17	1,985	2	3	62	28,573	14	21
indep-NT	2	256	0	0	72	68,187	1	64
inter	182	17,770	0	3	137	84,103	3	86
intra	132	12,555	1	0	54	18,101	2	15
total	865	83,327	5	17	718	423,476	39	406

Identifying Transcription Units



Number of Transcription Units per CDS



Part III

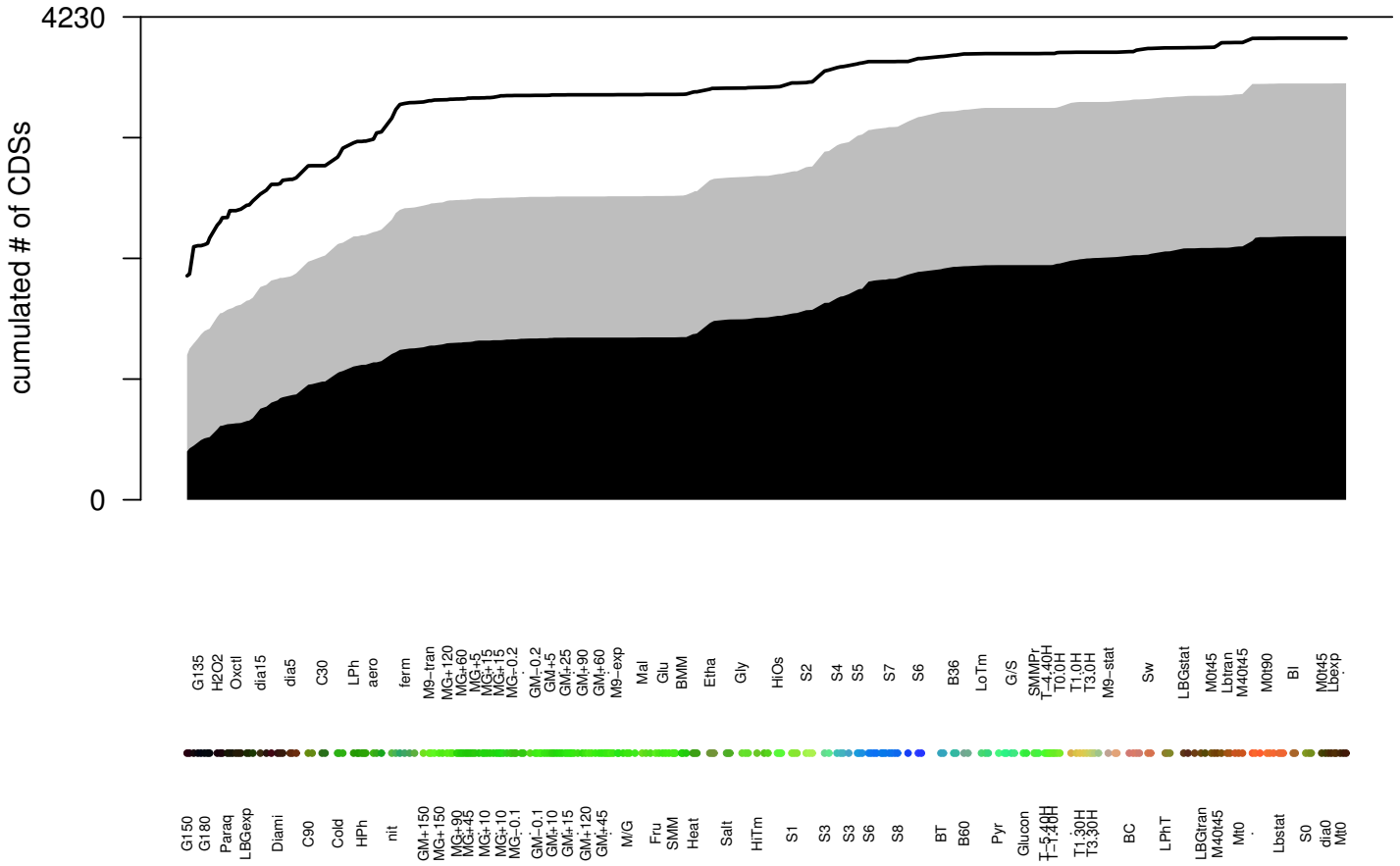
Transcriptional landscape estimation from a single tiling array hybridization

Building the *B. subtilis* transcriptional “parts list” from a collection of tiling array hybridizations

Relating transcriptome dynamics to the genome sequence

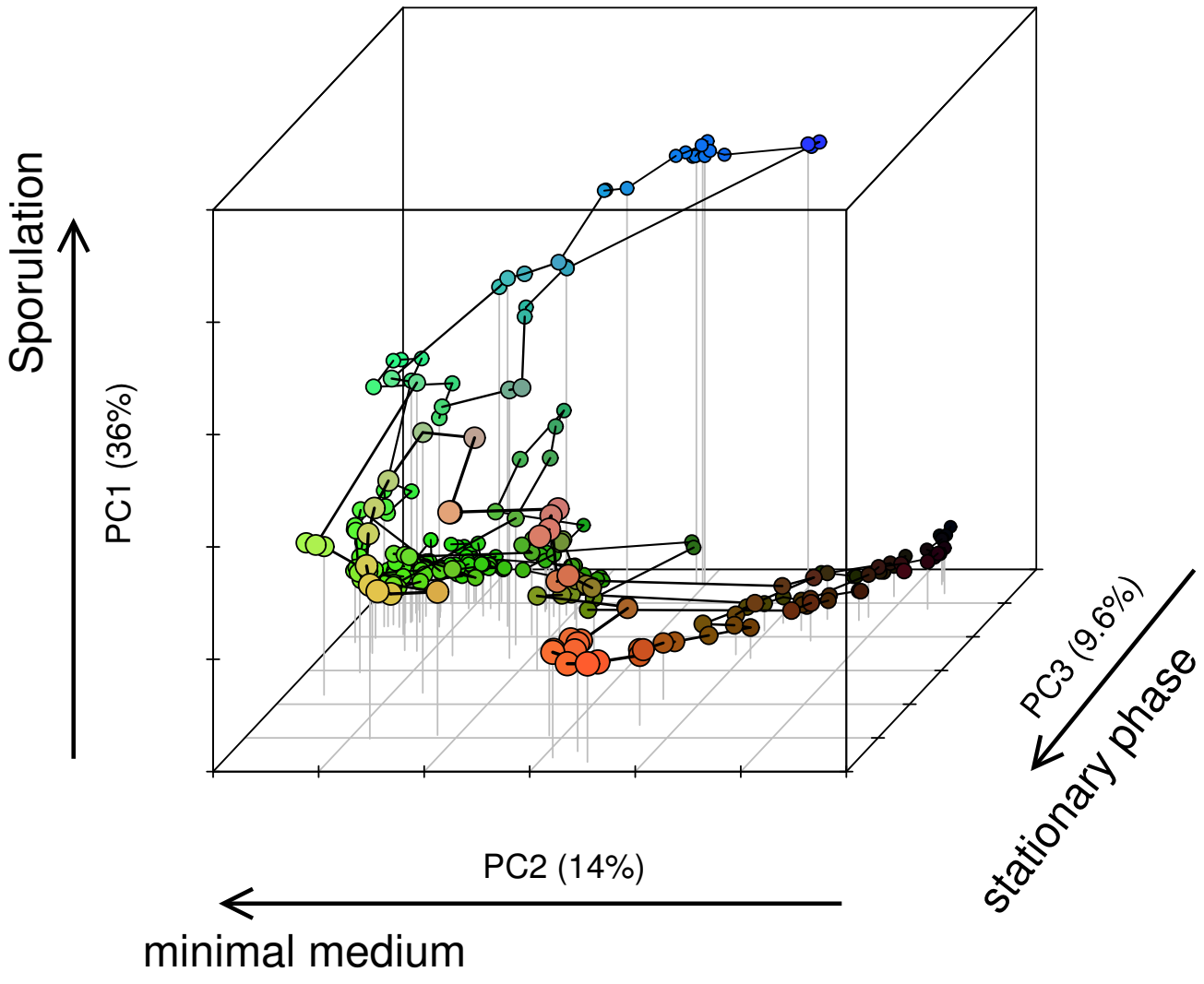
A focus on antisense transcription

A transcriptional context for almost every gene

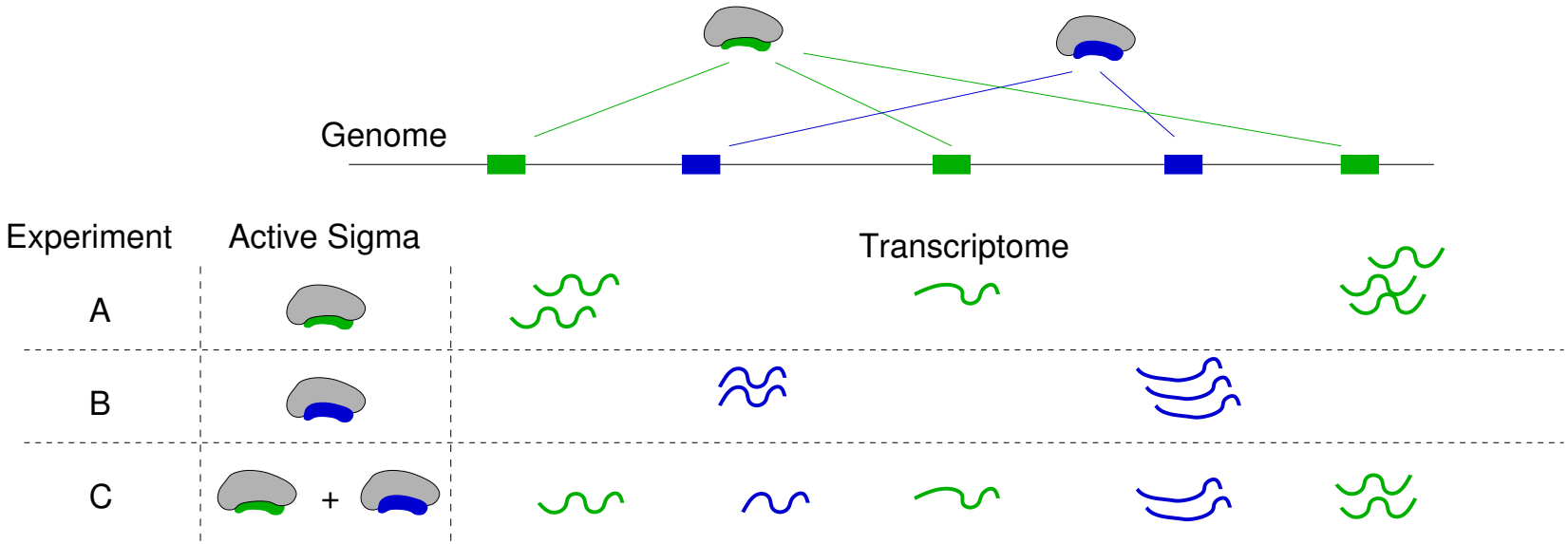


- Only 186 (<5%) annotated CDSs never seen expressed (below 5x background).
- 85% of the CDSs highly-expressed in at least one experiment.
- Only 144 genes highly expressed in all the conditions.

Highly coordinated changes of gene expression levels



Regulation of gene expression (simplified!)



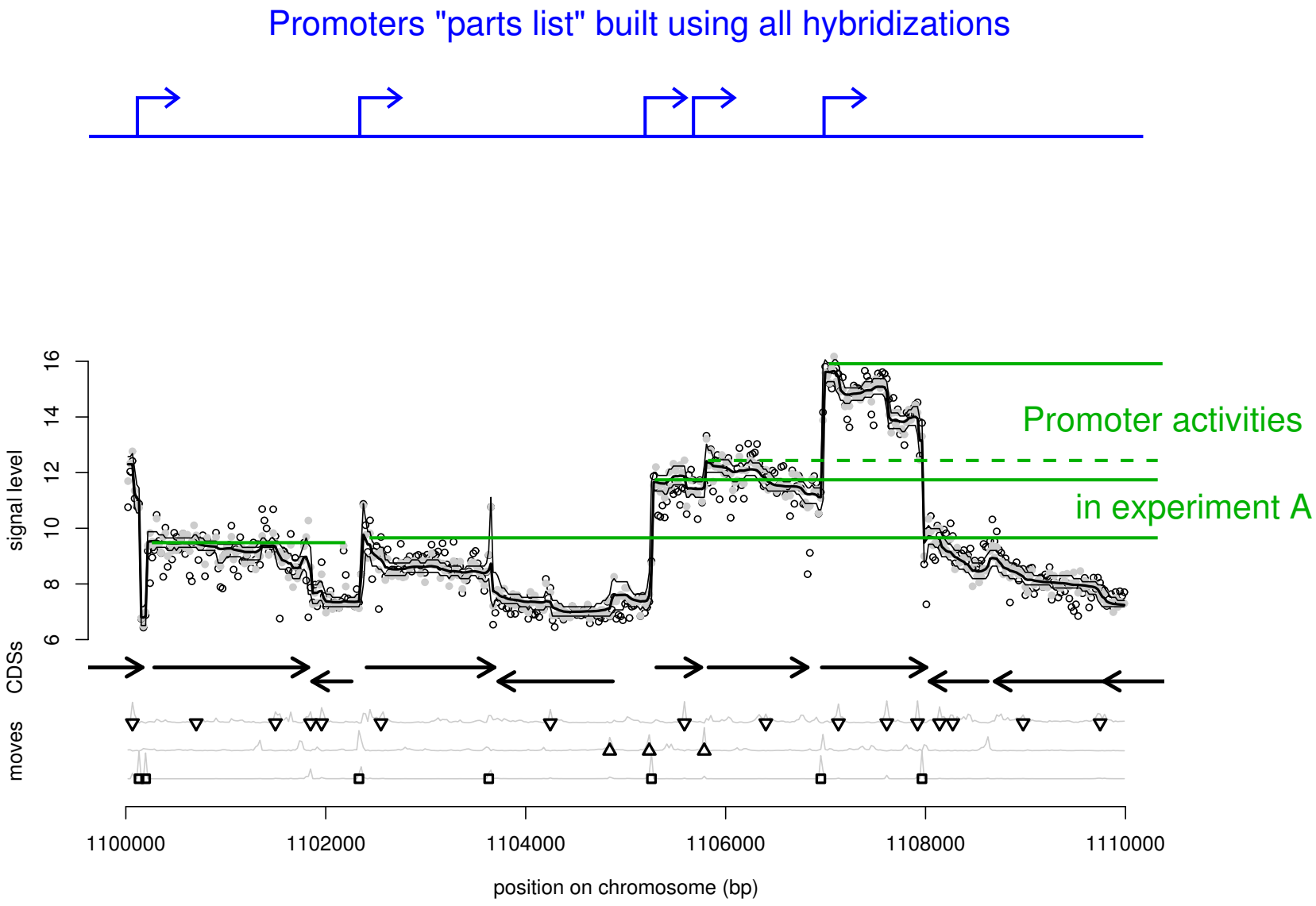
Questions

- Which promoter is dependent of which Sigma factor ?
- Which Sigma factor is active in which experiment ?
- How much of the promoter expression variance is explained by this basic model ?

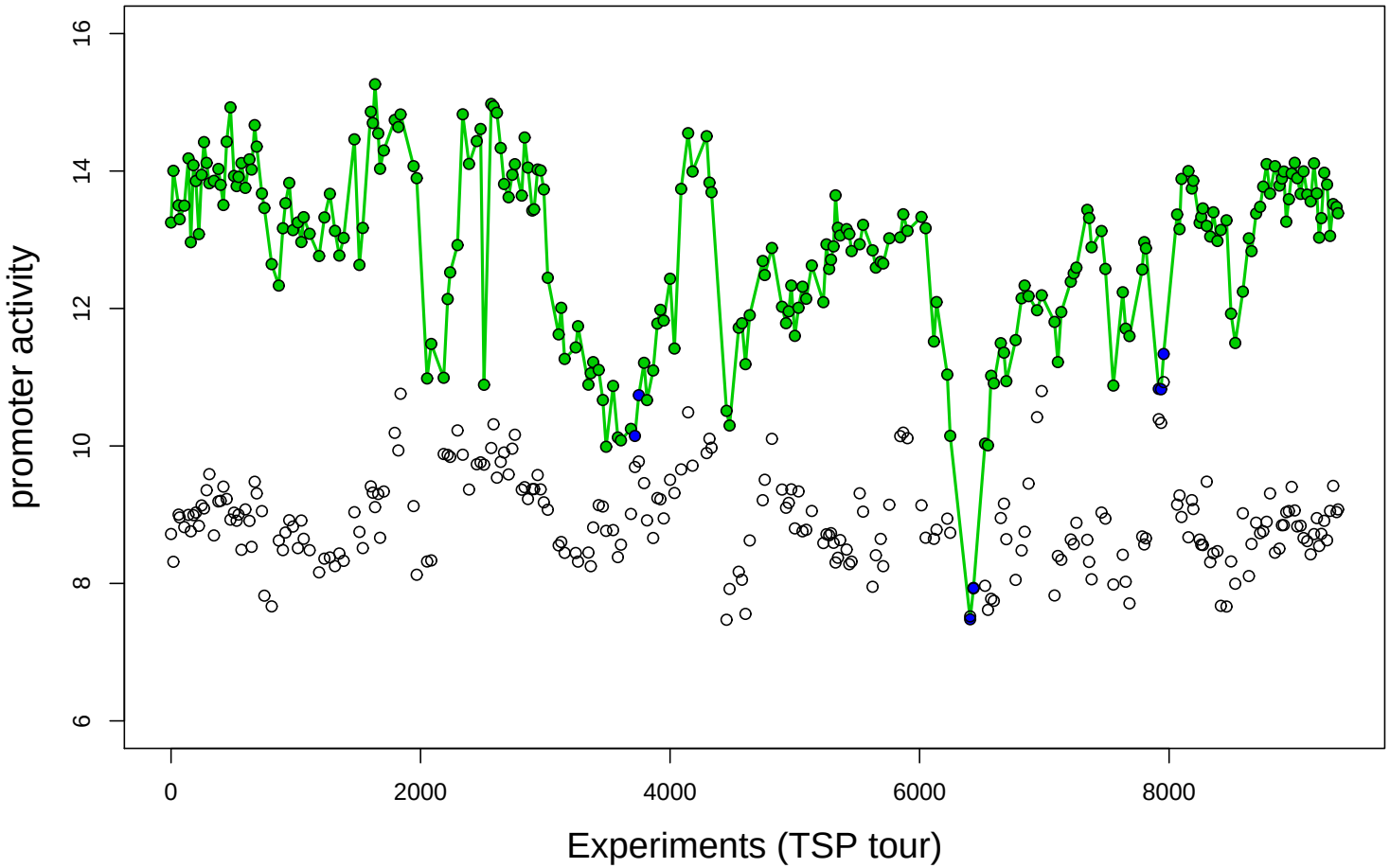
Steps

- To measure promoter activities
- To identify motifs in promoter sequences, taking into account the promoter activity
- To quantify the explained variance

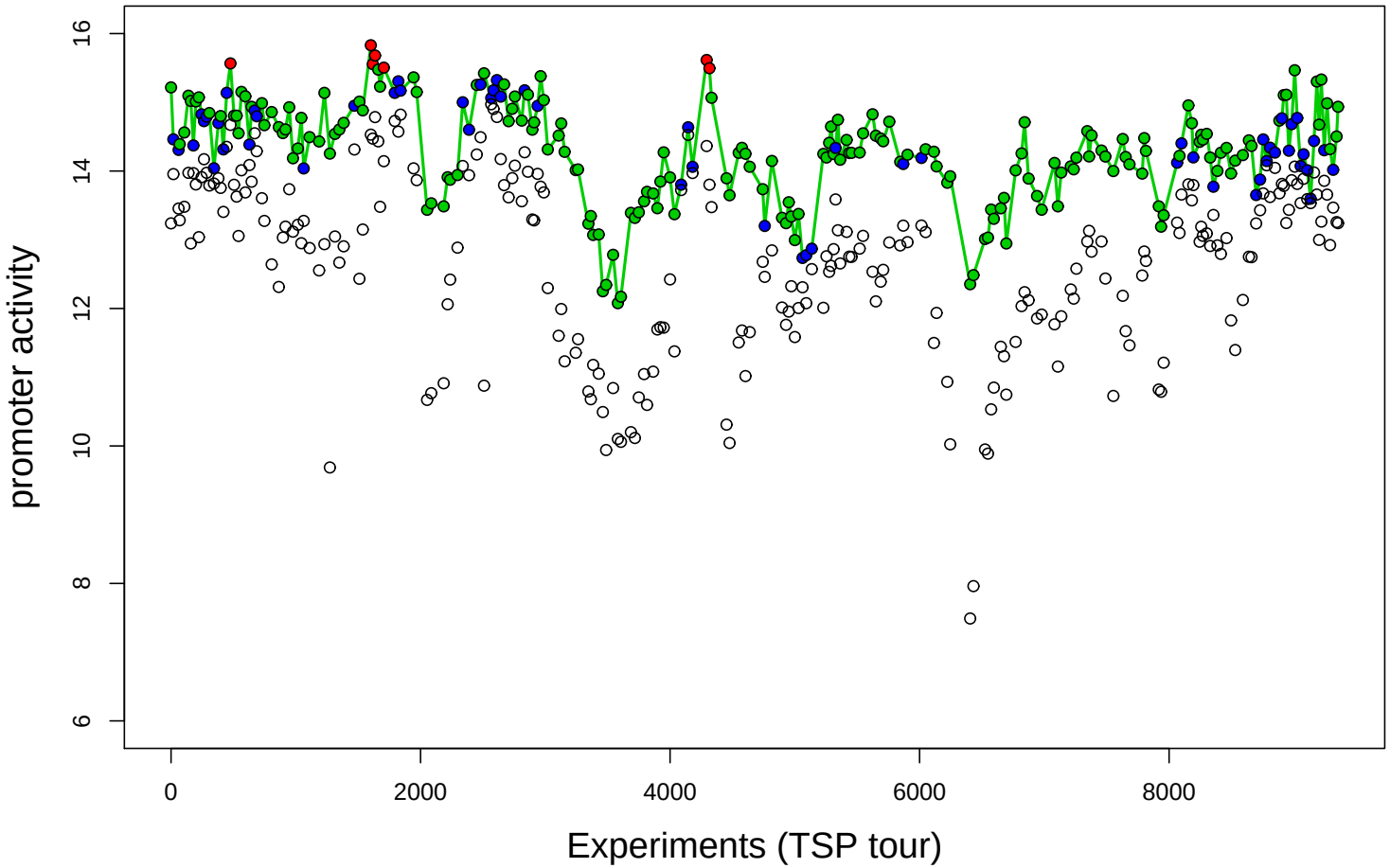
Measuring promoter activities



Example of activity of one promoter across experiments



Data on promoter activity are censored



Computing correlation between promoters

The Pearson correlation coefficient between x and y writes

$$\begin{aligned} r &= \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}} \\ &= \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} \end{aligned}$$

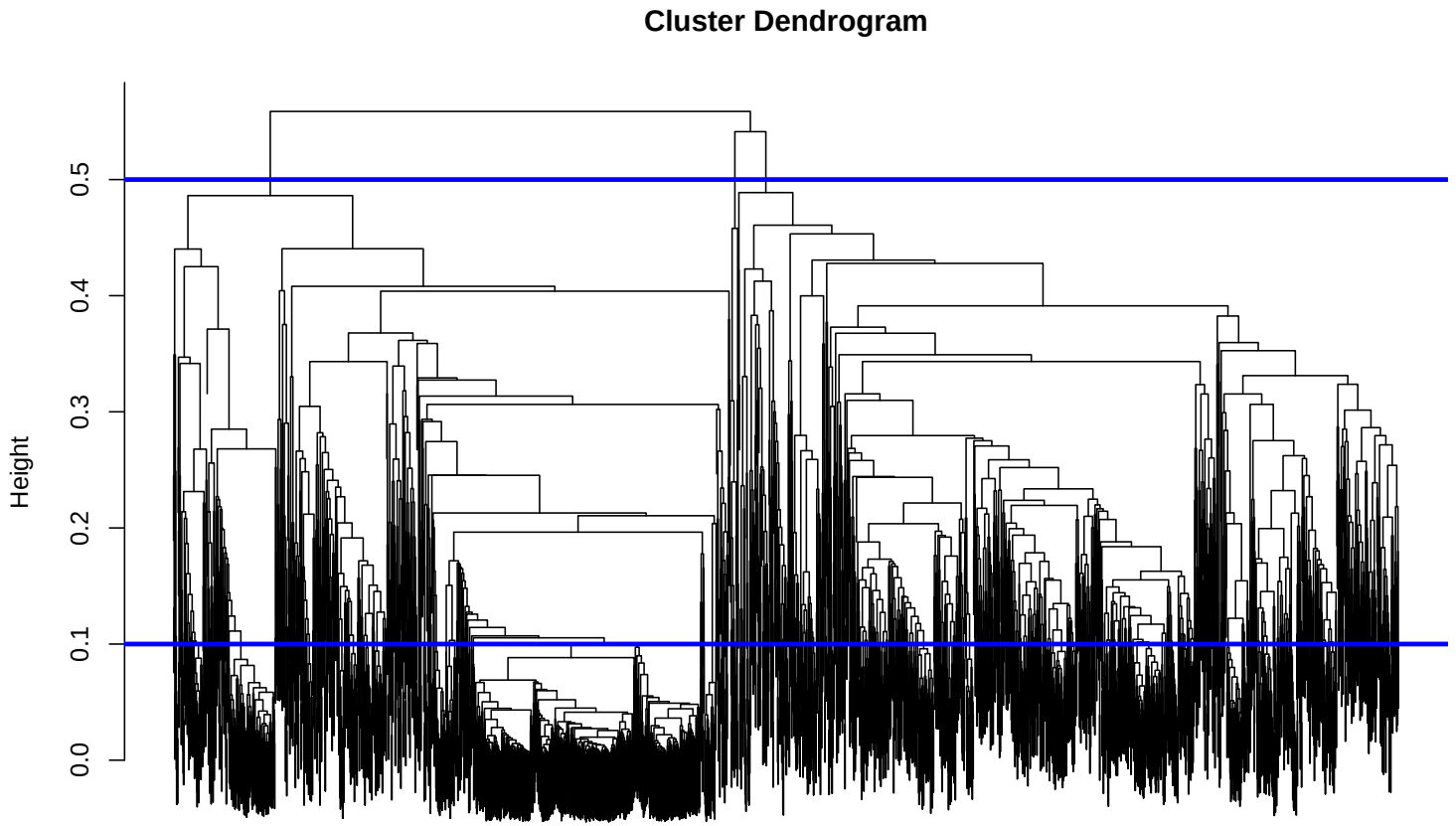
To account for censored data we fit the bivariate Gaussian distribution with covariance matrix Σ using a likelihood approach

$$(X,Y) \sim \mathcal{N}(\mu, \Sigma)$$

and we compute

$$r = \frac{\Sigma_{1,2}}{\sqrt{\Sigma_{1,1}\Sigma_{2,2}}}$$

Summarizing correlations between promoter activities

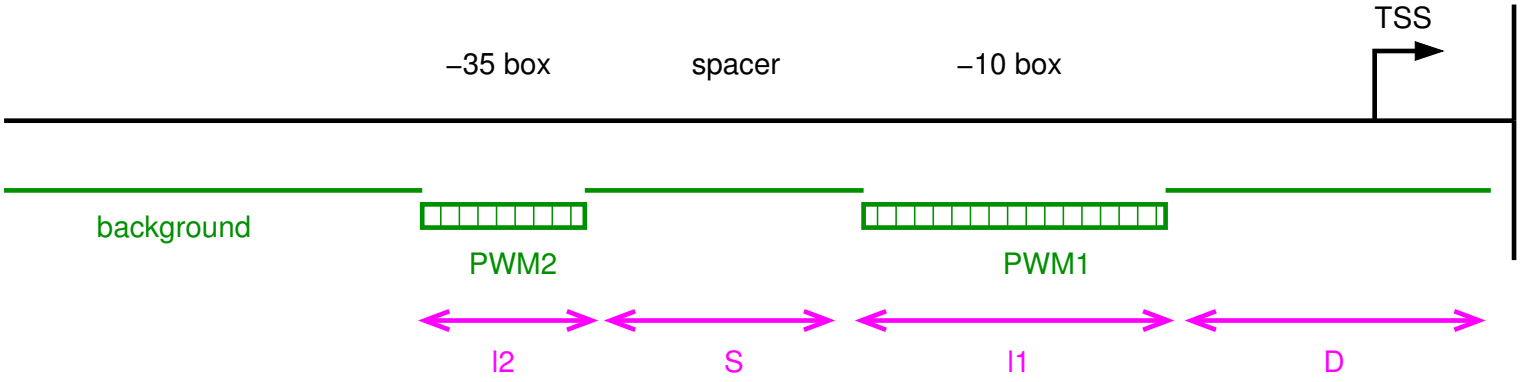


A 'promoter tree' is built by hierarchical clustering using average linkage on the dissimilarity matrix $d_{i,j} = (1 - r_{i,j})/2 \in [0, 1]$ where $r_{i,j}$ is the correlation between activities of promoters i and j.

Identifying sequence motifs

Sequence modeling

- the model expresses $\mathbb{P}(x_i \mid U_i = k)$, the probability of sequence x_i given the presence of a 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 $i \in \{1, \ldots, n\}$.

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.

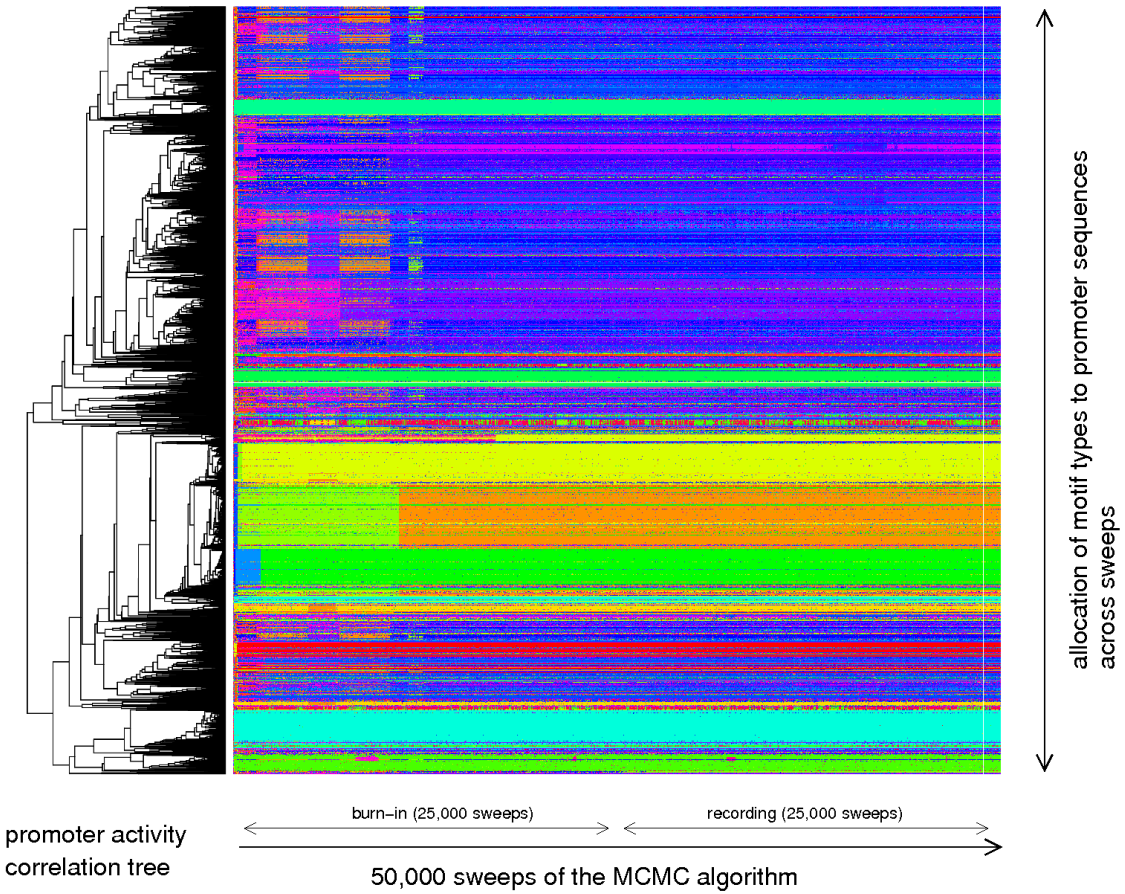
$$\mathbb{P}(U_1^n = u_1^n) \quad = \quad \sum_{(v)} \left[\pi_\alpha(v_{\text{root}}) \prod_{j \in \text{nodes}} \pi_{\lambda,\alpha}(v_{a_j} \to v_j) \prod_{i \in \text{leaves}} \pi_{\epsilon,\alpha}(v_{a_i} \to u_i) \right]$$

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

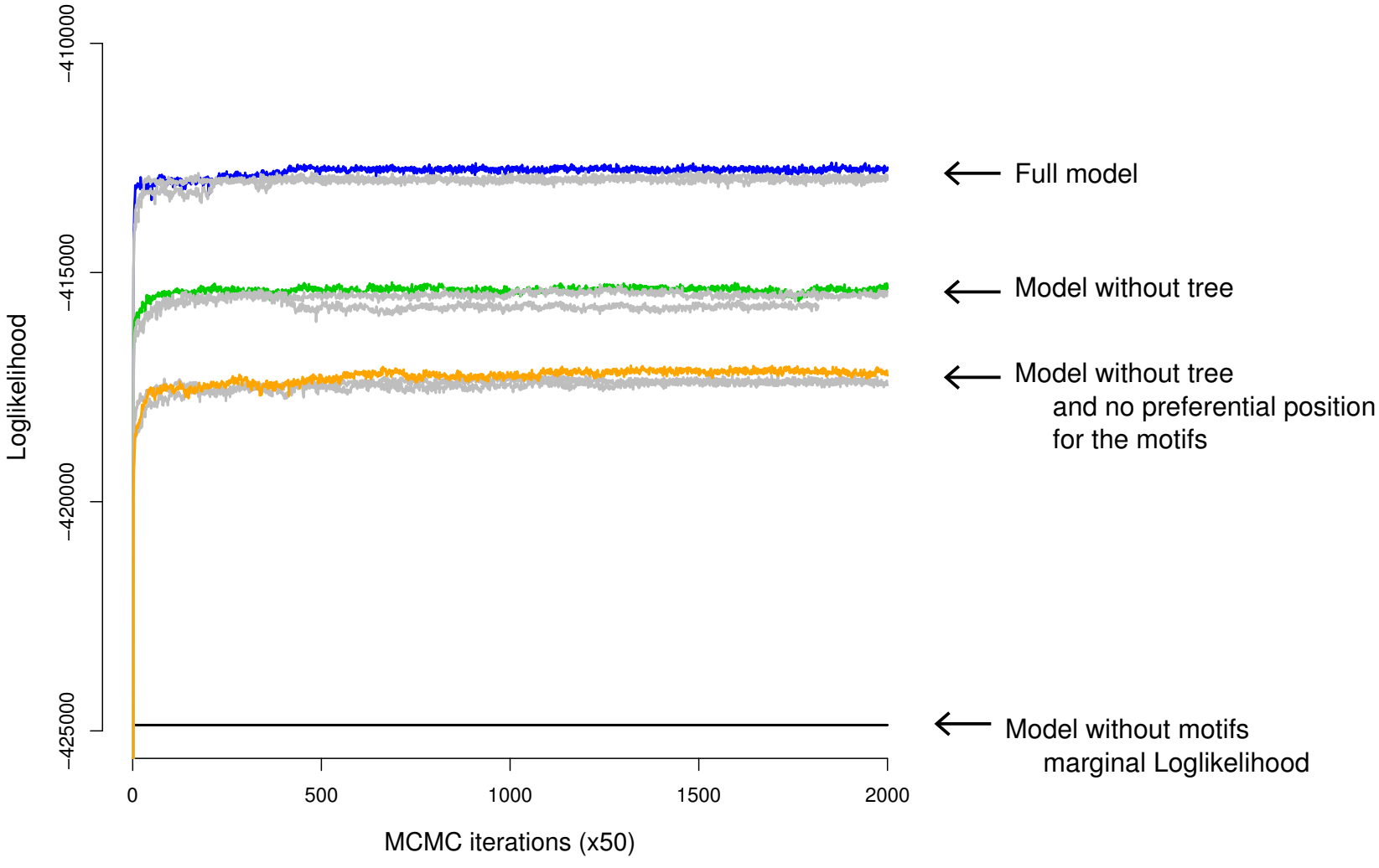
$$\begin{aligned} \pi_{\lambda,\alpha}(v_{a_j} \to v_j) &= (1 - e^{-\lambda d_j}) \mathbb{I}\{v_j = v_{a_j}\} + e^{-\lambda d_j} \alpha_{v_j} \\ \pi_{\epsilon,\alpha}(v_{a_i} \to u_i) &= (1 - \epsilon) \mathbb{I}\{u_i = v_{a_i}\} + \epsilon \alpha_{u_i} \end{aligned}$$

All parameters are estimated jointly with the MCMC alogrithm. 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



Model comparison: the tree improves the model

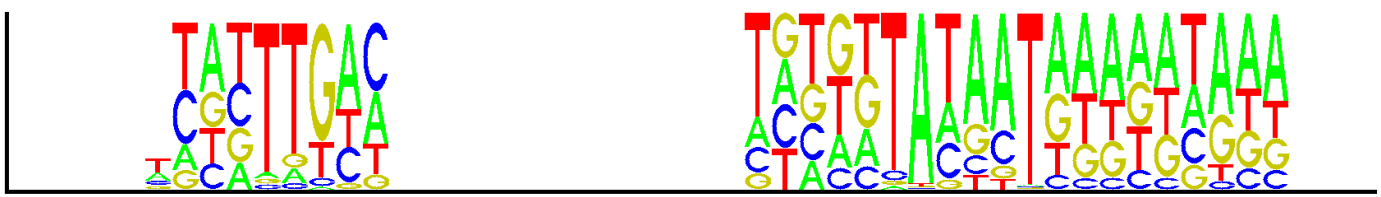


Comparison with known Sigma factor binding sites

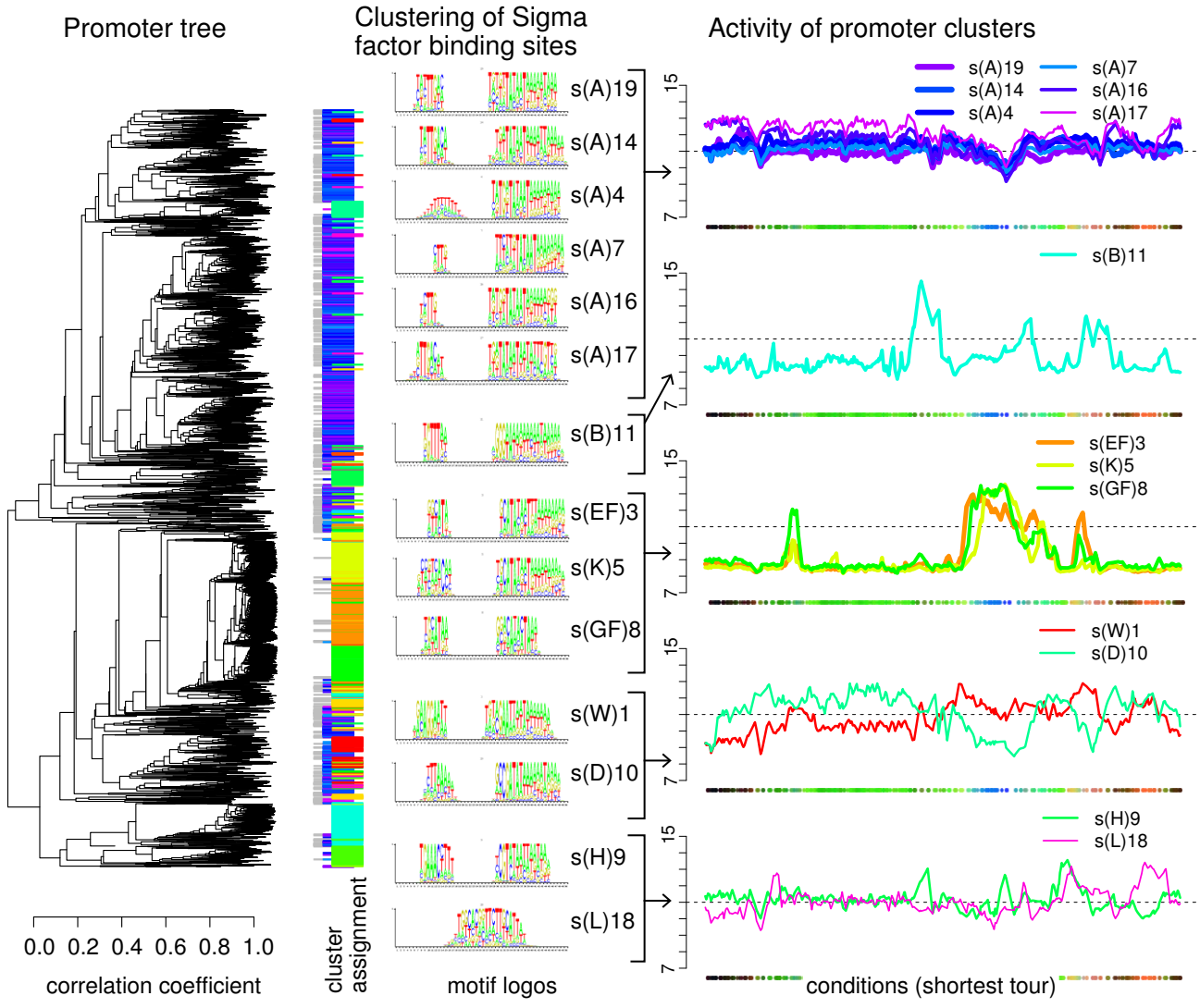
DBTBS: a database of transcriptional regulation in *Bacillus subtilis*

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 motif



Summarizing the activity of each individual promoter

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

$$y_{i,t} \quad = \quad \alpha_i + \beta_i a_{k,t} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_i^2).$$

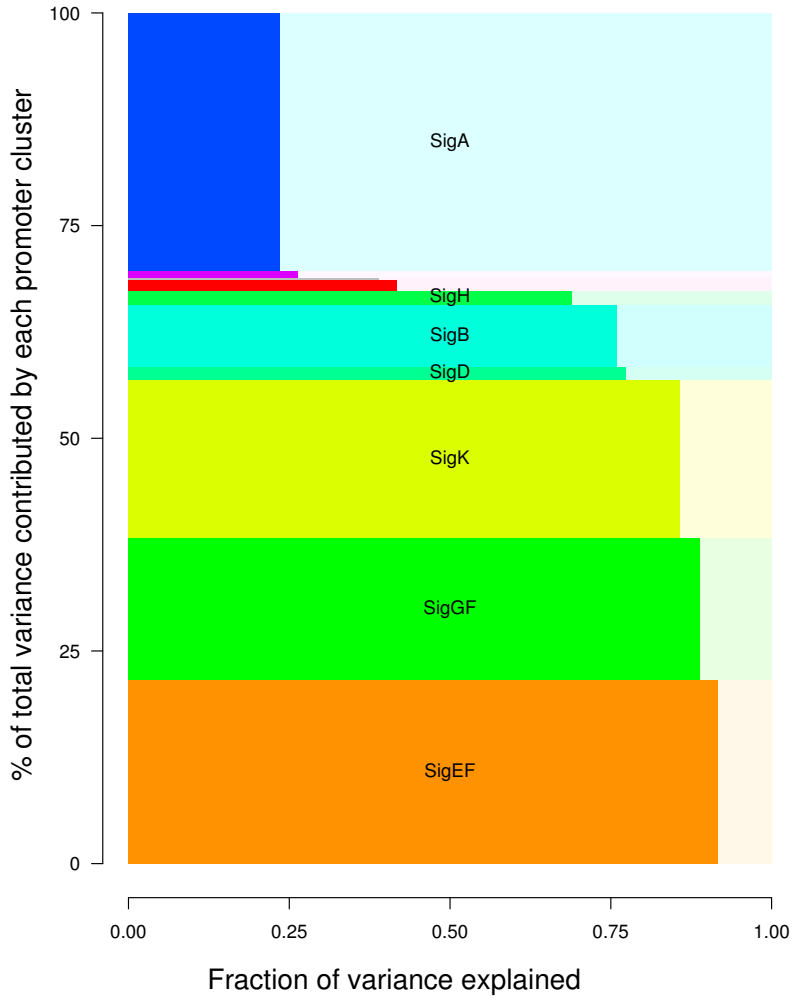
To be compared with

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

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.
- $1 - \sigma_i^2/\tau_i^2$ the fraction of variance that is explained by the activity of the Sigma factor.

Fraction of explained variance



66% of the total variance can be linked to direct regulation by Sigma factors.

Part IV

Transcriptional landscape estimation from a single tiling array hybridization

Building the *B. subtilis* transcriptional “parts list” from a collection of tiling array hybridizations

Relating transcriptome dynamics to the genome sequence

A focus on antisense transcription

AS transcription before the whole-transcriptome area

Only a very limited number of cases of regulation by AS transcription were known before genome-wide transcriptome studies.

JOURNAL OF BACTERIOLOGY, Oct. 2005, p. 6641–6650 Vol. 187, No. 19
0021-9193/05/$08.00+0 doi:10.1128/JB.187.19.6641–6650.2005

Small Untranslated RNA Antitoxin in *Bacillus subtilis*†

Jessica M. Silvaggi,[1] John B. Perkins,[2] and Richard Losick[1]*

Department of Molecular and Cellular Biology, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138,[1] and DSM Nutritional Products, Ltd., Biotechnology R&D, P.O. Box 3255, Building 203/20A, CH-4002 Basel, Switzerland[2]

Received 27 May 2005/Accepted 7 July 2005

JOURNAL OF BACTERIOLOGY, Feb. 2009, p. 1101–1105 Vol. 191, No. 3
0021-9193/09/$08.00+0 doi:10.1128/JB.01530-08

Extracytoplasmic Function σ Factors Regulate Expression of the *Bacillus subtilis yabE* Gene via a *cis*-Acting Antisense RNA▽

Warawan Eiamphungporn and John D. Helmann*

Department of Microbiology, Cornell University, Ithaca, New York 14853-8101

Received 29 October 2008/Accepted 23 November 2008

Bacillus subtilis yabE **encodes a predicted resuscitation-promoting factor/stationary-phase survival (Rpf/Sps) family autolysin. Here, we demonstrate that *yabE* is negatively regulated by a *cis*-acting antisense RNA which, in turn, is regulated by two extracytoplasmic function σ factors: σ^X and σ^M.**

- AS regulation has been demonstrated for the toxin *txpA* and for the autolysin *yabE*. In the second case, the biological role of this regulation is unknown.
- Other AS transcripts have been described but regulation still needs to be demonstrated (*surA*).

AS regulation in the whole-transcriptome area: a new paradigm?

This and previous genome-wide transcriptome profiling studies have revealed widespread AS transcription.

The recent review by Thomason and Storz (2010) lists mechanisms by which AS RNAs act.
AS-mediated regulation can virtually affect all the aspects of mRNA life:

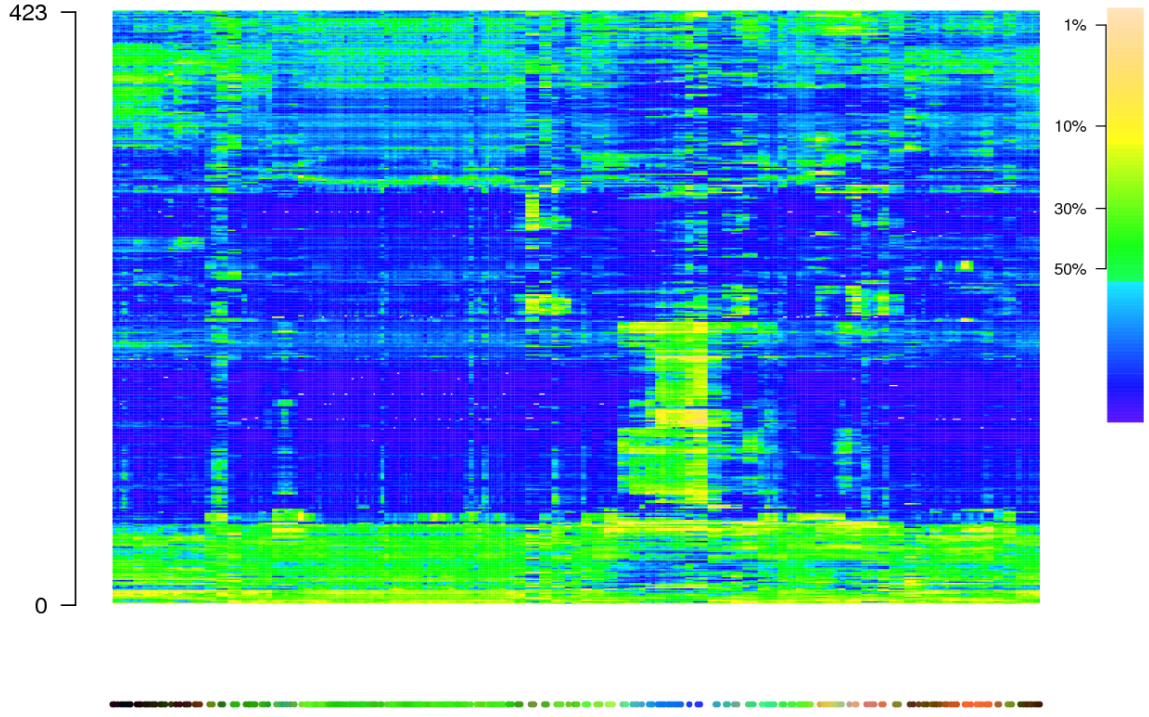
- transcription interference (transcription initiation and elongation)
- transcription attenuation (transcription termination)
- endonucleases and exonucleases (RNA degradation)
- ribosome binding (RNA activity)

Any AS RNA found is presumed to be cis-encoded regulatory RNAs.

Lead to the well-shared idea that AS RNAs constitute an important but overlooked class of regulatory molecules.

What have we learned on ASs in our study?

Using stringent cut-offs for calling a region “transcribed”, we mapped 423 unannotated transcription segments with a significant overlap with an annotated gene on the opposite direction (>100bp or 50% of the transcript length).



Of note, *ratA* (the antitoxin) and *surA* are detected but do not fulfill our overlap criterion. This number may thus underestimate the full repertoire of transcripts involved in AS regulation.

The patterns of AS transcription

Numerous AS with different coordinated expression profiles that may even suggest functional niches for AS regulations (such as during the sporulation or in some stress responses).

As many as 597 pairs of sense-antisense transcripts with documented expression contexts have been listed. For a number of these pairs nice biological stories could be imagined. Testing them individually would require a formidable amount of experimental work . . .

The fact is that no global story emerged from the analysis of the sense-antisense pairs.

We will thus now describe a few facts on the global pattern of AS transcription:

- Where and when AS transcription arises on the chromosome ?
- Are the amounts of the sense and AS transcripts correlated ?
- What are the expression levels of the AS transcripts ?

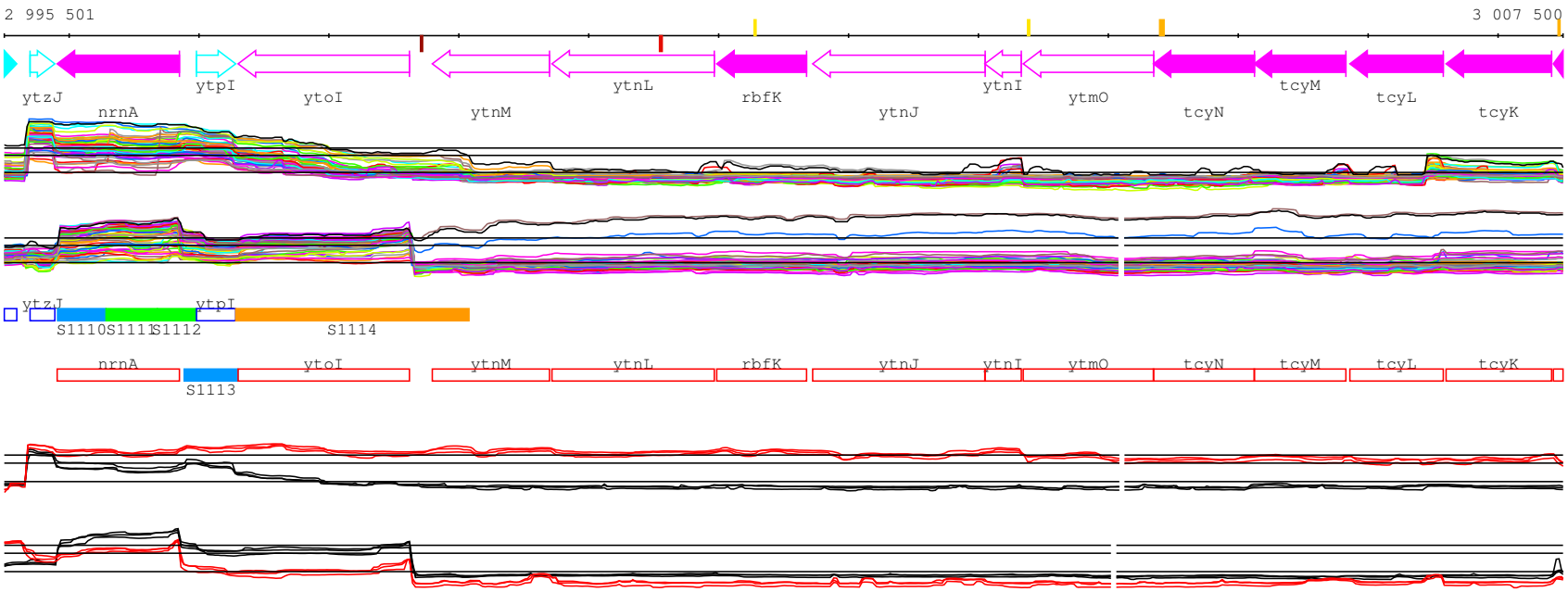
The transcriptional contexts of AS transcription

	50 bp $\leq$ length $<$ 150 bp				length $\geq$ 150 bp			
Type	#	bp	u-CDS	AS	#	bp	u-CDS	AS
3'UTR	64	5,955	0	5	61	25,993	5	26
3'NT	2	180	0	0	44	48,060	1	40
3'PT	4	316	0	1	74	60,294	0	69
5'	462	44,310	2	5	214	90,165	13	85
indep	17	1,985	2	3	62	28,573	14	21
indep-NT	2	256	0	0	72	68,187	1	64
inter	182	17,770	0	3	137	84,103	3	86
intra	132	12,555	1	0	54	18,101	2	15
total	865	83327	5	17	718	423476	39	406

Many AS transcripts (62%) arise in transcriptional contexts corresponding to incomplete termination of the transcription (categories 3'PT, 3'NT, Indep-NT and Inter).

AS transcription: The role of Rho

We also found that the protein Rho plays a key role in limiting ASRNAs by preventing transcription beyond the 3’ boundaries of a subset of TUs.



Contexts of AS transcription (continued): the 5'-end

[data not shown]

The prevalence of SigA-dependent transcription is much lower for AS RNAs than for protein coding genes: only 52% of the AS RNAs are predicted to be transcribed from a SigA promoter whereas this fraction is 74% for protein coding genes.

This trend is most pronounced for the classes of AS RNAs that have their own promoters (Indep and Indep-NT) as only a small minority (23%) is predicted to be SigA-dependent.

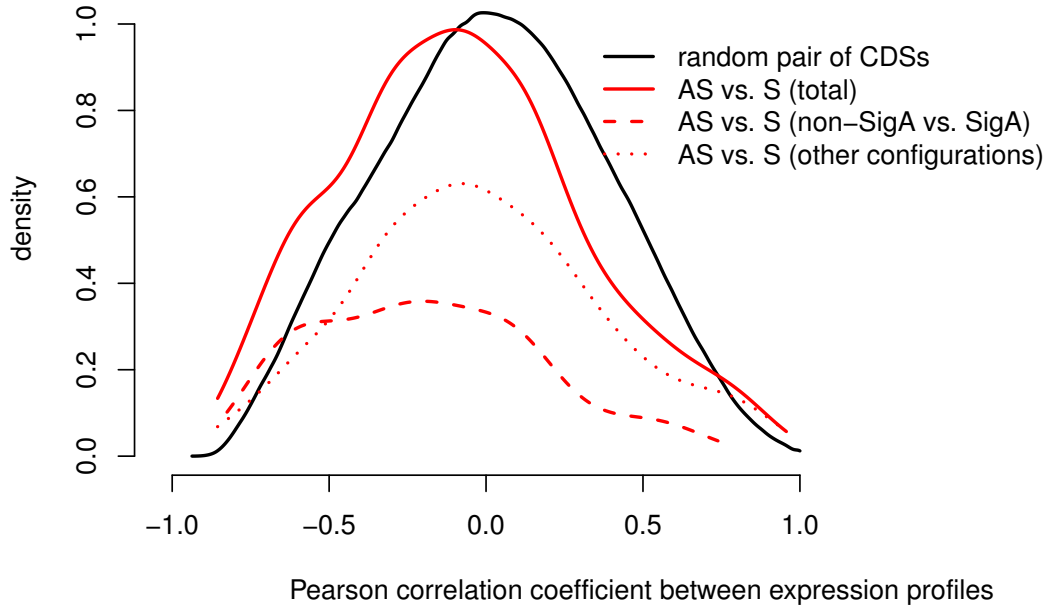
Overall, 82% (347/423) of the AS transcripts are accounted for by incomplete termination of transcription or by initiation of transcription from promoter controlled by alternative Sigma factors

Pairwise correlation patterns between sense and AS expression

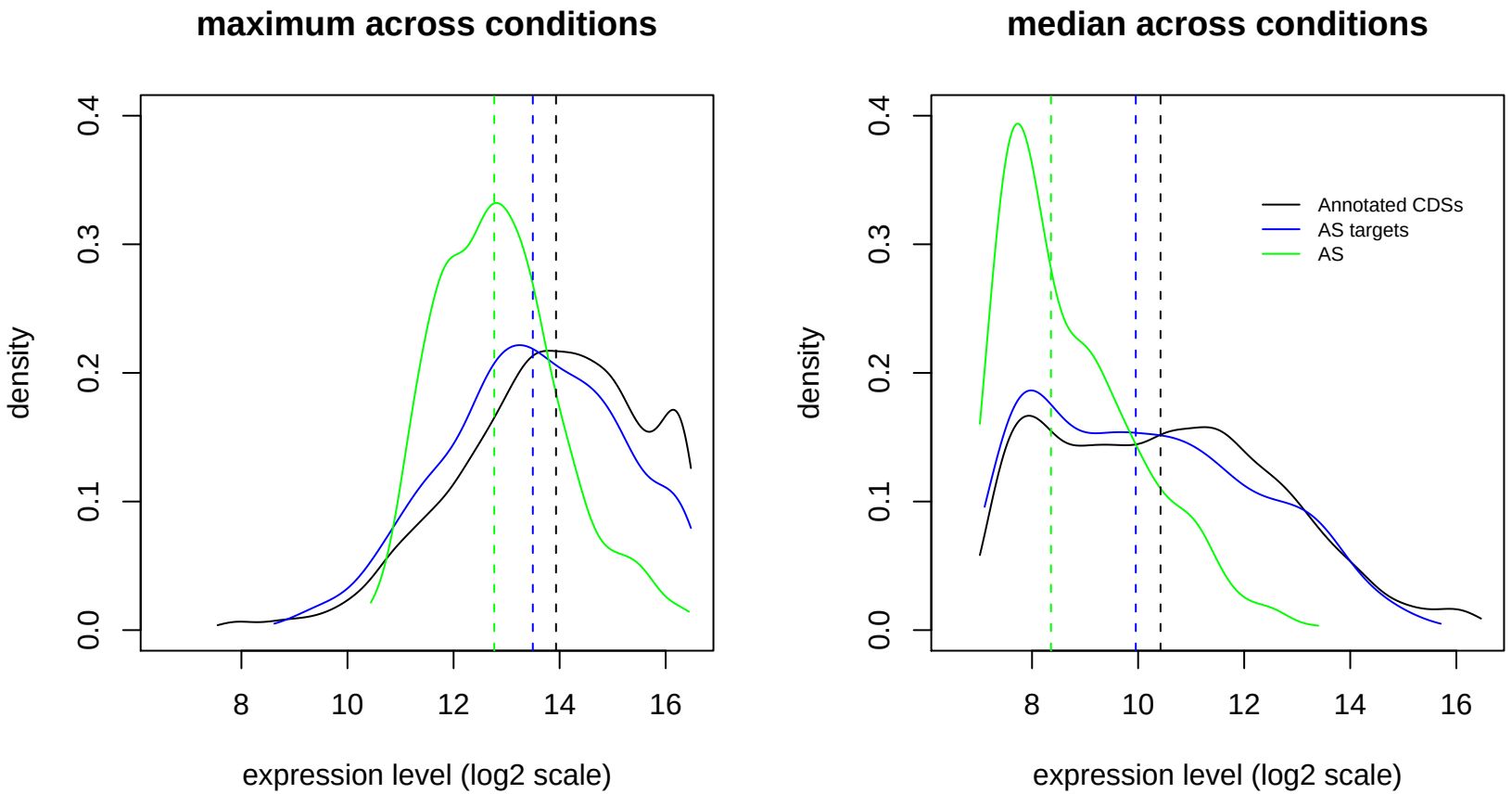
Correlation cannot be directly related to a particular mechanism but is relevant to describe the data and most people would interpret it as an indication of interaction.

Correlation is statistically significant for most (77%) sense-antisense pairs. The correlation is more often negative (47%) than positive (30%).

This however needs to be compared with the expected correlation patterns between random pairs of transcripts . . .



Levels of AS transcription



The maximum and median expression level across conditions tend to be lower for AS segments than for protein coding genes, this is true for SigA and non-SigA dependent AS RNAs. The expression of the CDSs facing AS RNAs is also less likely to reach a very high expression level.

Main facts about ASs and a possible explanation

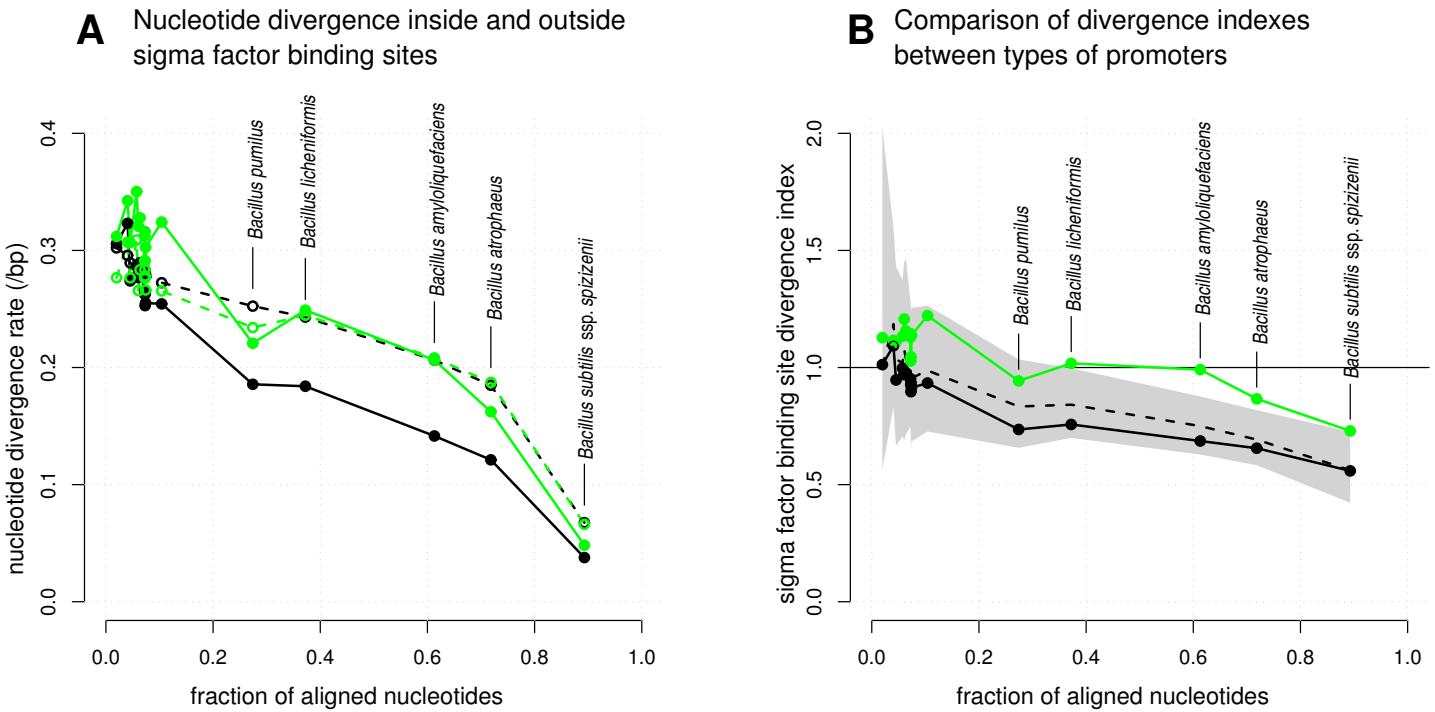
- Most ASs arise apparently from incomplete or missing transcription termination
- Most of the ASs that have their own promoter are not SigA-regulated
- Sense and AS transcripts tend to display a small excess of negative correlation but most of this excess is linked to non-SigA ASs facing SigA transcripts.
- ASs are expressed at lower levels than typical sense transcripts.

All these facts are compatible with the idea that the bulk of AS transcription may arise from imperfect transcriptional control both in 3' and 5' transcript ends.

- 3'-ends: missing and imperfect terminators
- 5'-ends: promoters may appear randomly in the course of evolution. This would be more difficult to avoid and less detrimental (cost and interference) for alternative Sigma factors.

Conservation analysis of AS generating promoters

We tried to find additional data that could support the idea that AS generating promoters (99 that are responsible for “Indep” and “Indep-NT” AS) may not have a biological role.



The imperfect transcriptional control hypothesis

The hypothesis is indeed very difficult to test experimentally.

- It is difficult to show that something does not have a role.
- A number of ASs may even interfere and thus “regulate” the sense transcripts but this would not strictly contradict the hypothesis as long as the roles of these regulations cannot be exhibited.
- The finding that a fraction of the AS may be involved in biologically meaningful transcriptional regulation would not invalidate the hypothesis.

This hypothesis is different from -but not incompatible with- the idea that ASs may also arise from pervasive transcription starting randomly along the genome generating a “background transcriptional noise”. Here ASs are seen expressed above background in particular biological contexts (conditions, promoters, terminators).

Ongoing work and future directions

- analysis of RNA half-lives in order to contribution the of promoter activity and RNA degradation in regulation.
- estimation of an hardware model of gene expression regulation.
- experimental evolution (plug a new Sigma factors and analysis of subsequent adaptations).

The 'hardware' hypothesis

Idea to explain the tuning of transcription levels wrt growth rate :

- Unlikely to rely on specific regulators.
- Indeed we imagine that transcriptional tuning might be hard-coded in gene-specific kinetic parameters governing the rate of synthesis and degradation.

$$
\begin{aligned}
G_i + P &\overset{k_+^i, k_-^i}{\leftrightarrows} (G_i, P) \\
(G_i, P) &\overset{k_s^i}{\rightarrow} M_i \\
M_i &\overset{k_d^i}{\rightarrow} \emptyset
\end{aligned}
$$

core equation of hardware regulation hypothesis

$$
\begin{aligned}
G_i + P \quad &\overset{k^i_+, k^i_-}{\leftrightarrows} \quad (G_i, P) \\
(G_i, P) \quad &\overset{k^i_s}{\rightarrow} \quad M_i \\
M_i \quad &\overset{k^i_d}{\rightarrow} \quad \emptyset
\end{aligned}
$$

In steady-state equilibrium, the amount of messenger i writes as a function of the amount of 'free' polymerase [P]

$$
[M_i] \quad = \quad \frac{k^i_s [G_{i,tot}]}{\frac{k^i_- + k^i_s}{k^i_+} \cdot \frac{1}{[P]} + 1} \cdot \frac{1}{k^i_d}
$$

where $k^i_s [G_{i,tot}]$ 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, and $\frac{1}{k^i_d}$ is proportional to mRNA half-life.

Link between [P] and expression data

This equation allows a variety of profiles and the amount of free polymerase [P] can (at least theoretically) be estimated.