Quantitatively deconvolving alternative RNA secondary structures

Regina Bohnert and Gunnar Rätsch



Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

July 15, 2011

HiTSeq-SIG, ISMB/ECCB 2011

RNA Secondary Structure and Its Role

- RNA-seq: transcript identification with quantitation
- But: information about structure is missing
- Important role in basic cellular processes

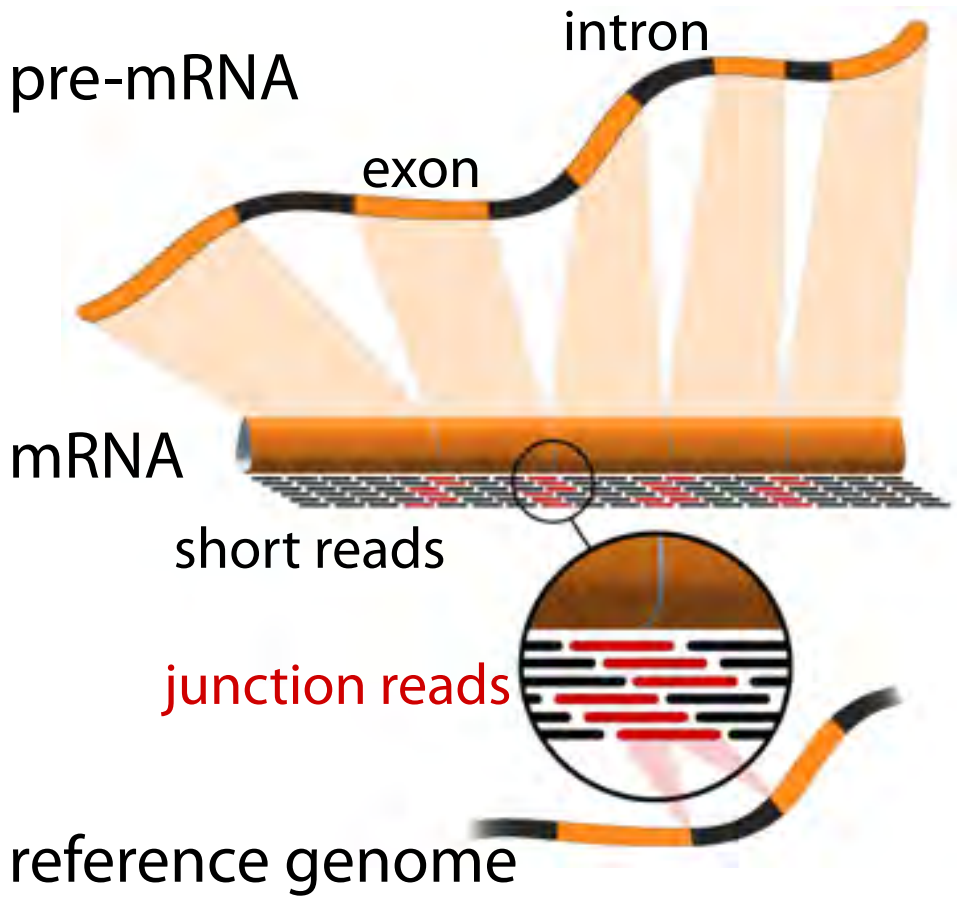
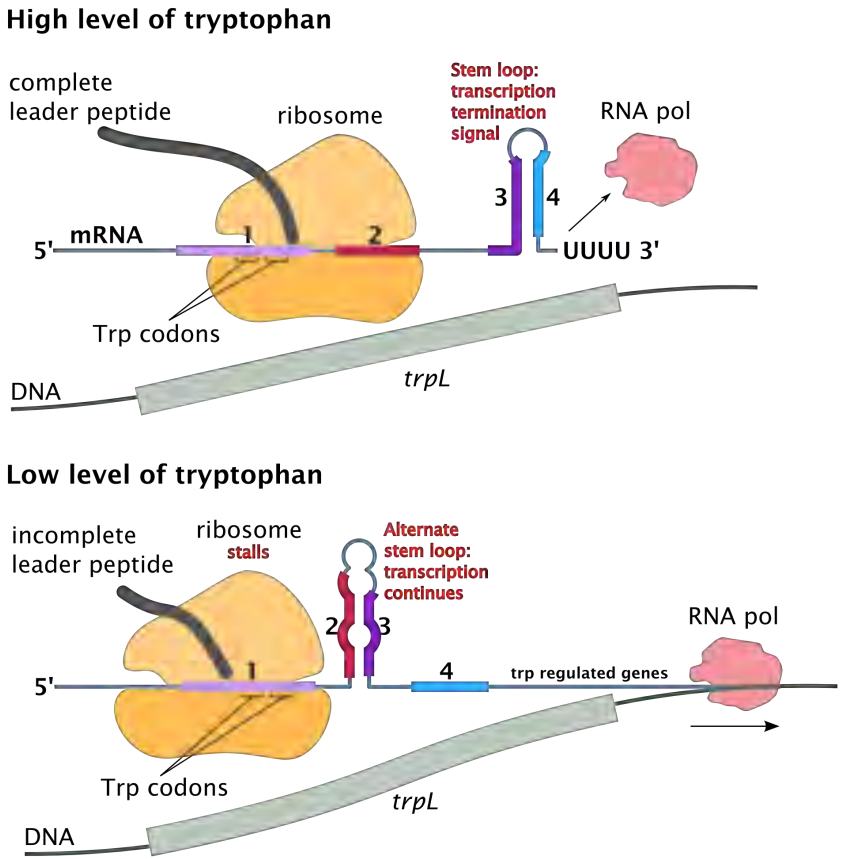


Figure adapted from Wikipedia

RNA Secondary Structure and Its Role

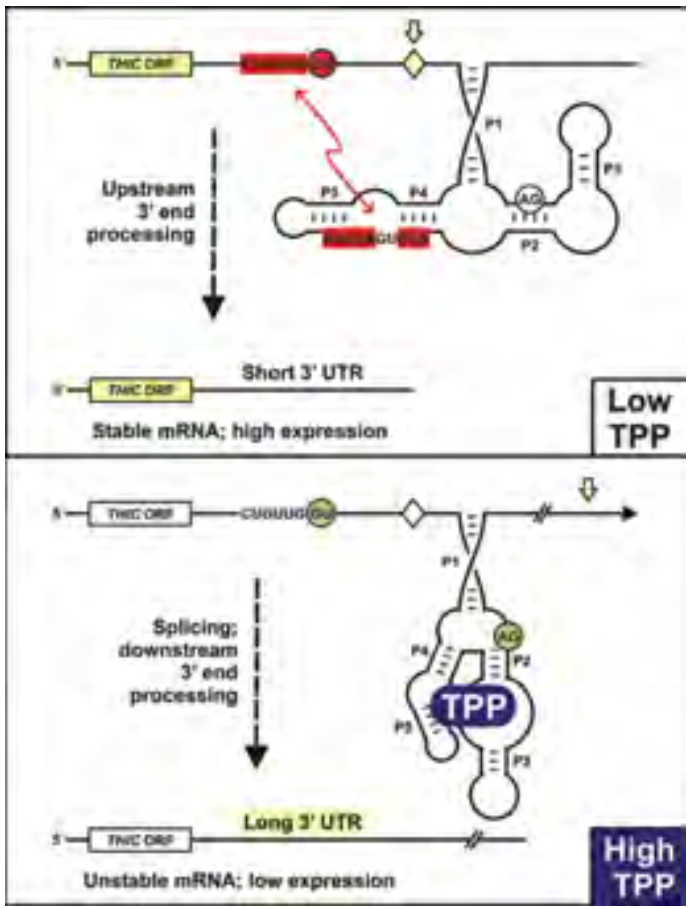
- RNA-seq: transcript identification with quantitation
- But: information about structure is missing
- Important role in basic cellular processes
  - Regulation of gene expression



Trp operon, figure taken from Wikipedia

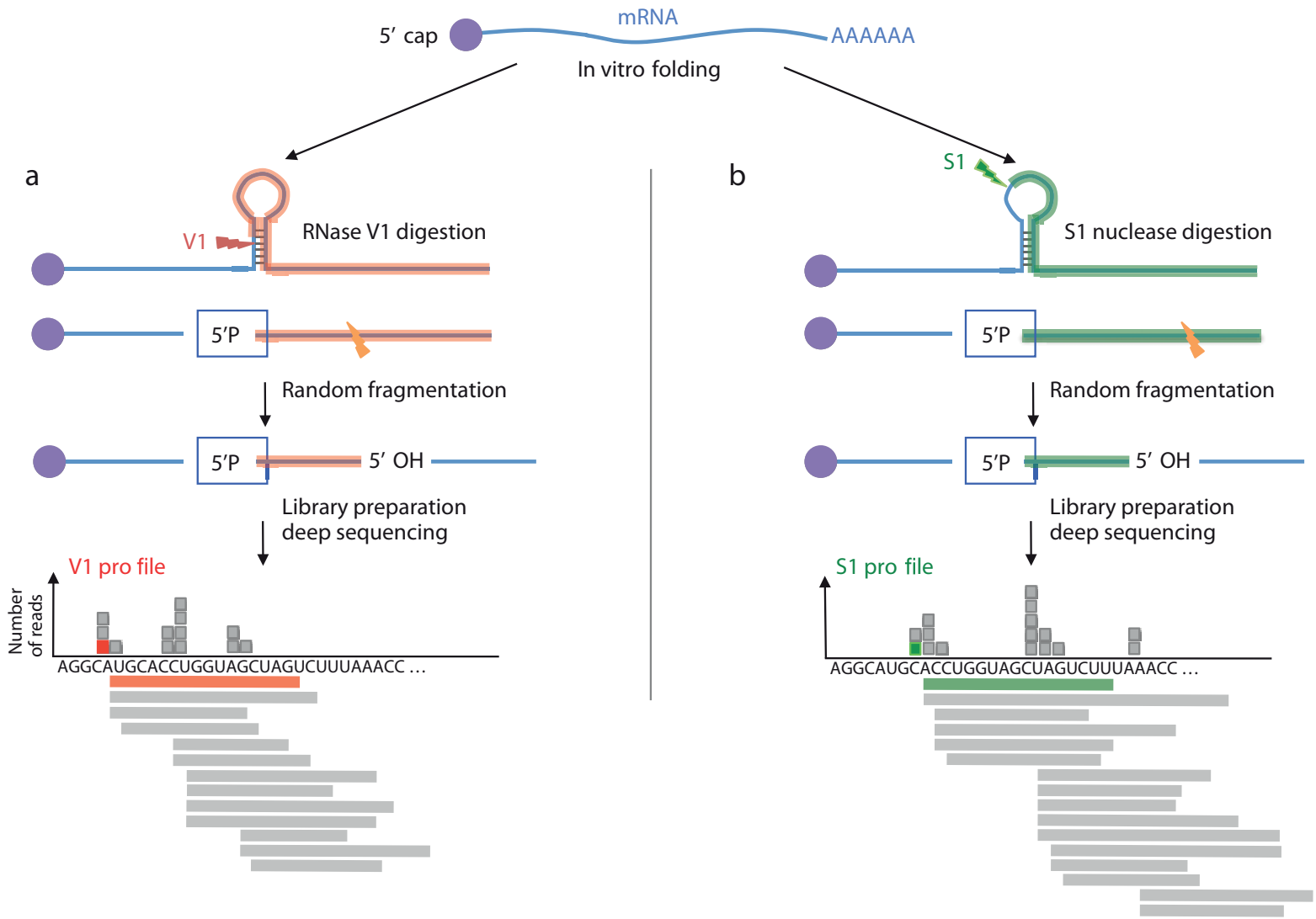
RNA Secondary Structure and Its Role

- RNA-seq: transcript identification with quantitation
- But: information about structure is missing
- Important role in basic cellular processes
  - Regulation of gene expression, alternative splicing
- Mixtures of alternative structures may co-exist



TPP riboswitch, figure taken from Wachter (2010)

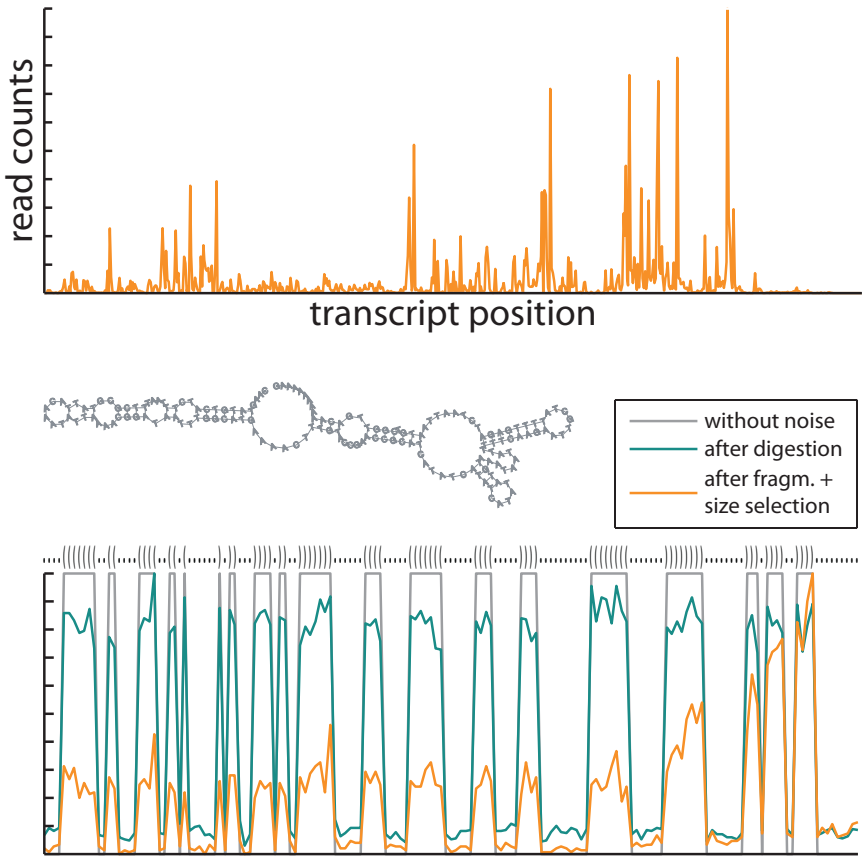
Profiling RNA Secondary Structures



PARS (parallel analysis of RNA structure) protocol, figure taken from (Kertesz et al., 2010)

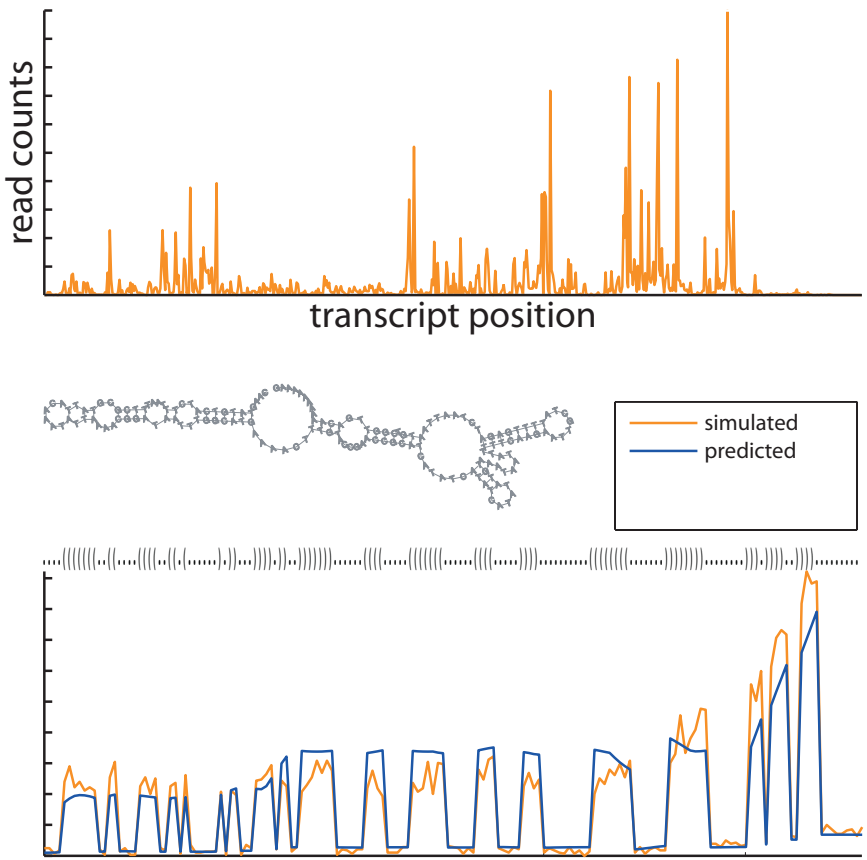
Bias Correction

- Observation: no uniform distribution (Kertesz et al., 2010)
- Data simulation
  - Simulated structures (Hofacker et al., 1994)
  - Generated read counts for ds/ss experiments (PARS-like)
    - Read starts at ds/ss
    - Enzyme errors
    - Fragmentation, size selection



Bias Correction

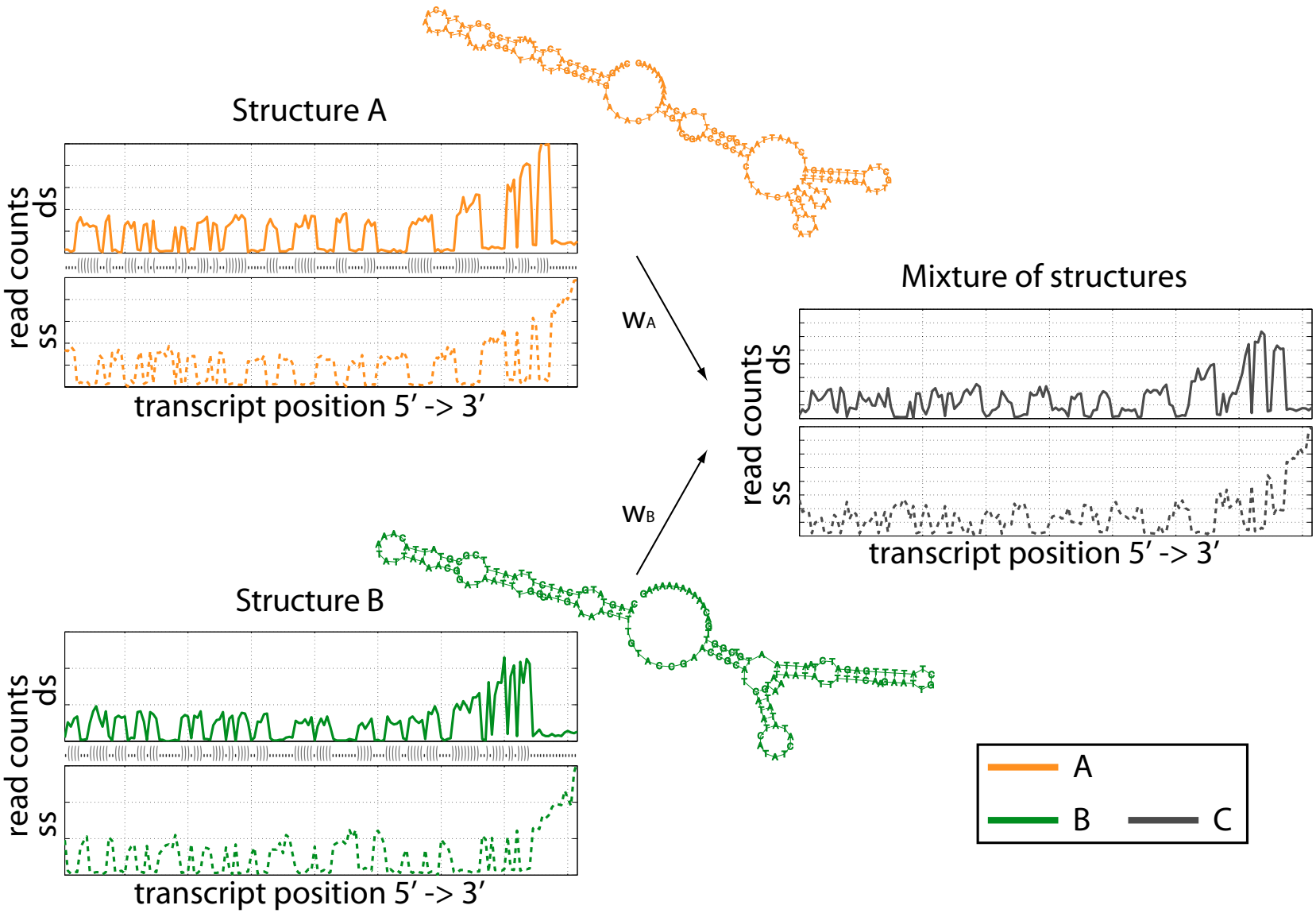
- Observation: no uniform distribution (Kertesz et al., 2010)
- Data simulation
  - Simulated structures (Hofacker et al., 1994)
  - Generated read counts for ds/ss experiments (PARS-like)
- Bias correction: Ridge regression
  - Features: distances to proximal cutting sites
  - Target: expected/observed read counts
  - **Model explains $\approx 75\%$ of the variability**



Questions

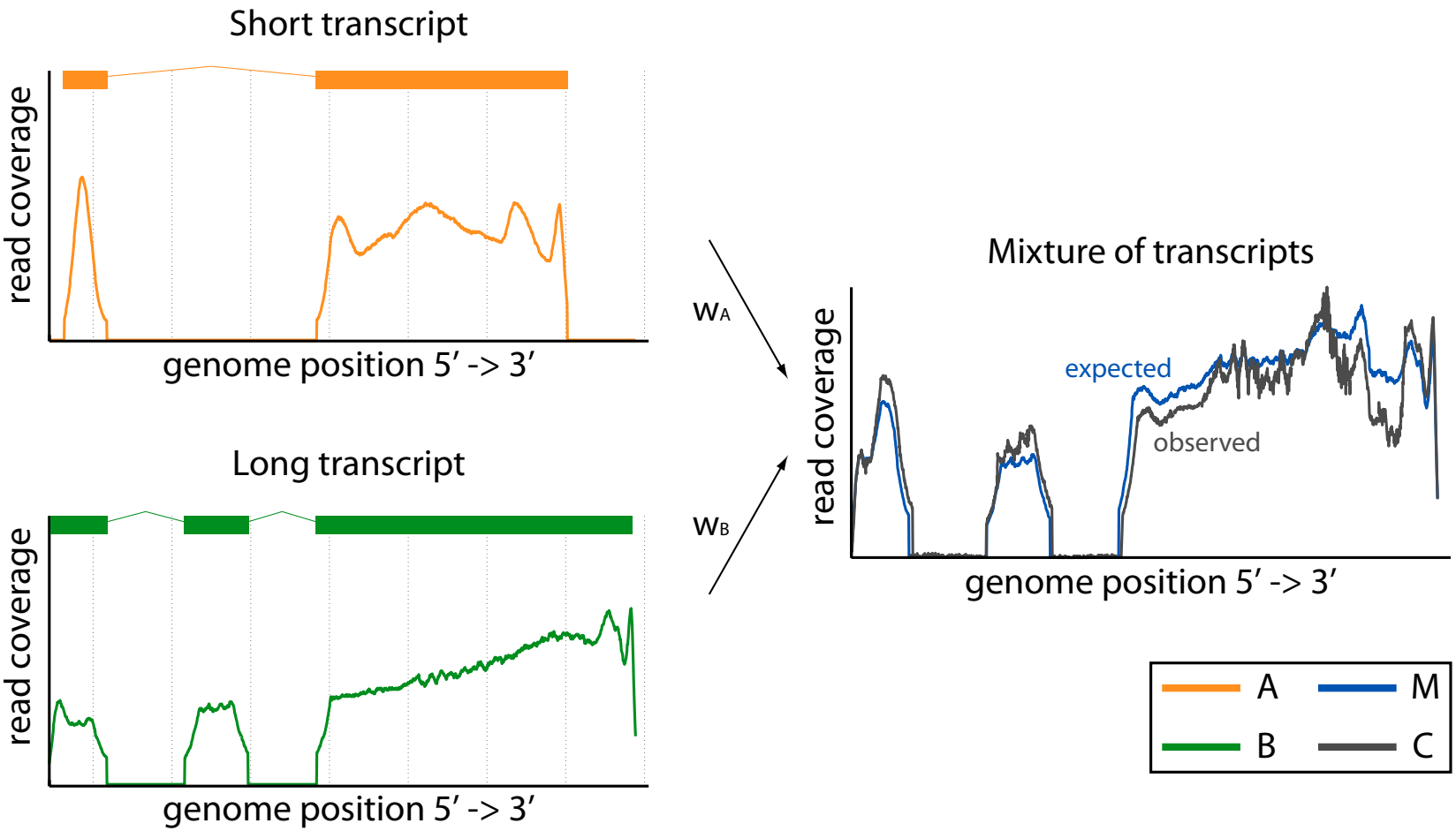
- How can known transcripts with structural information be quantified?
  - Single transcript with multiple possible structures
  - Alternative transcripts with multiple possible structures
- How can structures be inferred from read data?

sQuant – Basic Idea



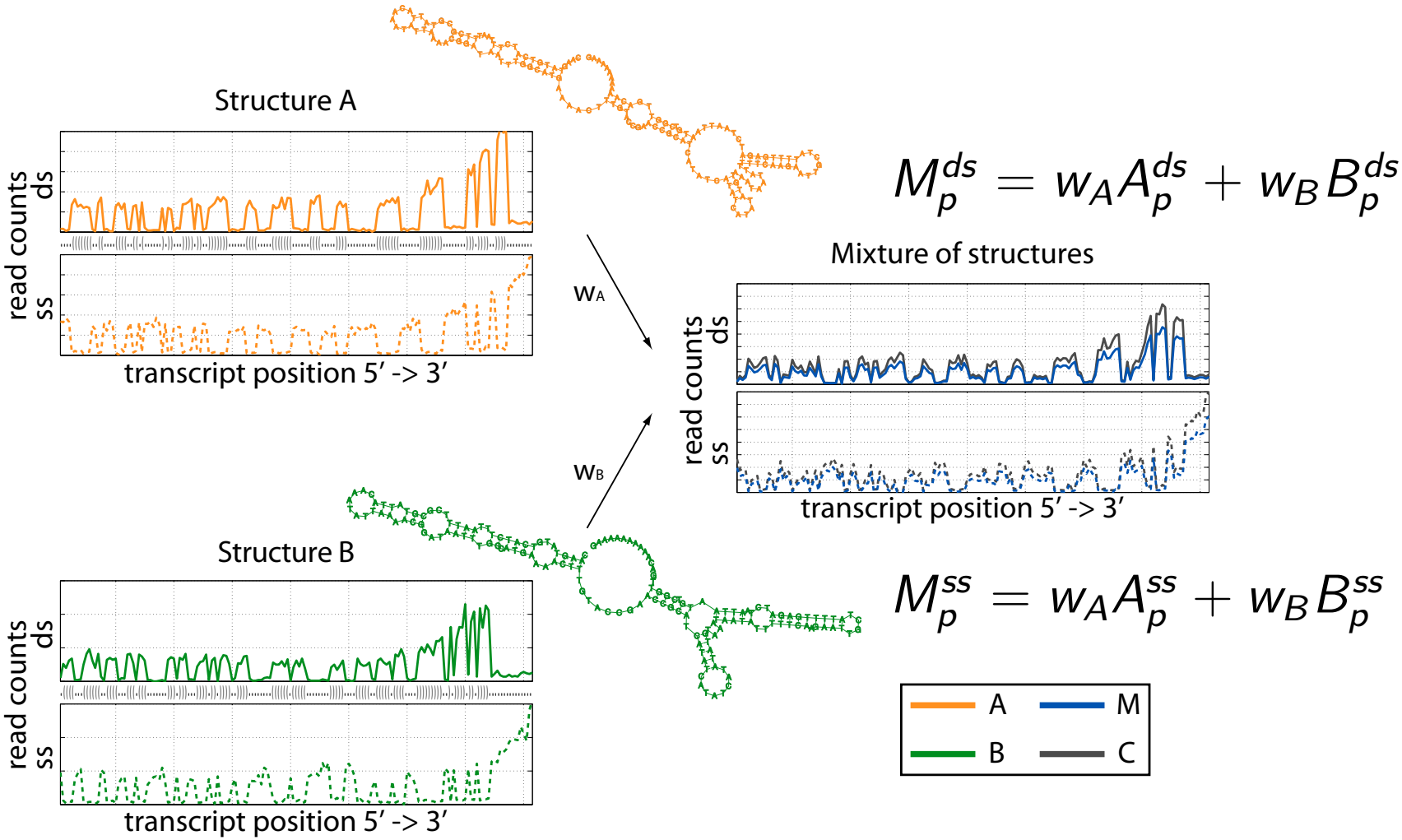
Recap from Transcript Quantitation

rQuant (Bohnert et al., 2009; Bohnert and Rätsch, 2010)



$$M_p = w_A A_p + w_B B_p \quad \Rightarrow \quad \min_{w_A, w_B} \sum_p \ell(M_p, C_p)$$

sQuant – Basic Idea



$$M_p^{ds} = w_A A_p^{ds} + w_B B_p^{ds}$$

$$M_p^{ss} = w_A A_p^{ss} + w_B B_p^{ss}$$

$$\Rightarrow \quad \min_{w_A, w_B} \sum_p \left(\ell\left(M_p^{ds}, C_p^{ds}\right) + \ell\left(M_p^{ss}, C_p^{ss}\right) \right)$$

Results

1. Mixture of structures
   - 500 genes, one transcript per gene
   - Complex mixture of 1-10 known structures
   - Use sQuant to quantify structures
   - Pearson's correlation with ground truth: 90.9%
2. Mixture of structures and transcript isoforms
   - 500 genes, with ≥ 2 alternative transcripts per gene
   - Complex mixture of 1-10 structures
   - Use sQuant to quantify structures
   - Pearson's correlation of with ground truth: 89.8%
3. Advantage of bias correction
   - 1000 genes, 1 structure per transcript
   - Pearson's correlation of with ground truth
     - w/o bias correction: 82.6%
     - w/ bias correction: 83.5%

Does structure help for transcript quantitation?

RNA-seq-based transcript quantitation is difficult:



when

- reads are short
- alternative events are far apart
- difficult to determine which exon combination is present

Does structure help for transcript quantitation?

RNA-seq-based transcript quantitation is difficult:



Different exon combinations lead to different RNA structures!

```
Transcript 1   UGGCACAAGC   GCCGCAAAAC   CGGGGGCAAGCGGGGGCAAG   AGAAAGCCCU   GAGUUGGGGC
MFE structure  ..((.(..((   (((.(.....   ..).)))..))..).))...   .....(((((   .....)))))   -17.30 kcal/mol

Transcript 2   UGGCACAAGC                CGGGGGCAAGCGGGGGCAAG   AGAAAGCCCU   GAGUUGGGGC
MFE structure  .(((....))                ).....((.(((.((((((...   .....)))))   ..)))...))   -16.00 kcal/mol

Transcript 3   UGGCACAAGC   GCCGCAAAAC   CGGGGGCAAGCGGGGGCAAG                GAGUUGGGGC
MFE structure  ..........   (((.(((..(   (....((........)).)..)              ).)))).)))   -12.50 kcal/mol

Transcript 4   UGGCACAAGC                CGGGGGCAAGCGGGGGCAAG                GAGUUGGGGC
MFE structure  ..((.(((.(                (....((........)).)..)              ).)))..))   -6.10 kcal/mol
```

Comparison: PARS-seq & sQuant *vs.* RNA-seq & rQuant

- 500 genes with ≥ 2 alternative transcripts
- Ideal situation: one known structure per transcript
- Compare quantitation of RNA transcripts/structures
  sQuant: 85.1% rQuant: 64.9%

Questions

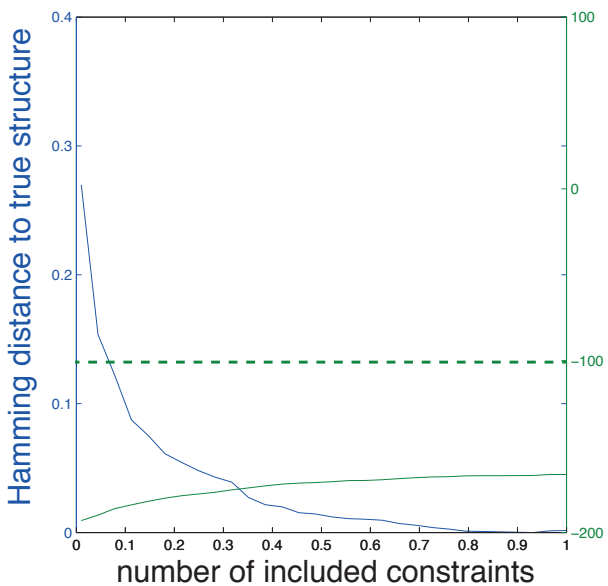
- How can known transcripts with structural information be quantified?
- How can structures be inferred from read data?

De Novo Structure Predictions

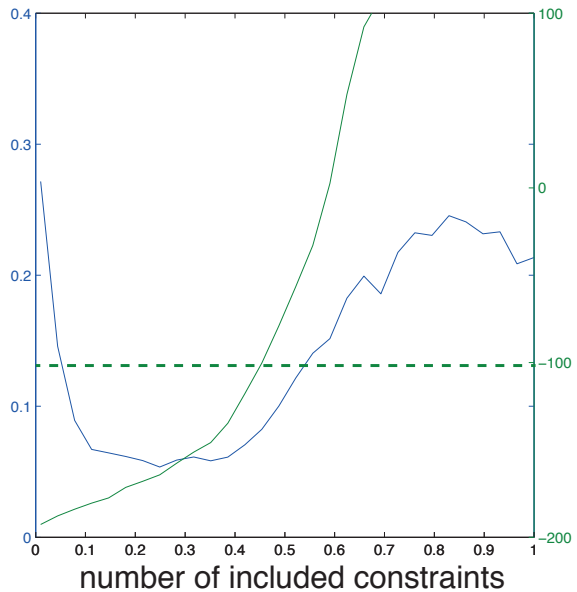
Idea: Use most confident positions as contraints for paired and unpaired bases (τ = pseudo count)

$$score(p) = \log\left(\frac{ds(p) + \tau}{ss(p) + \tau}\right)$$

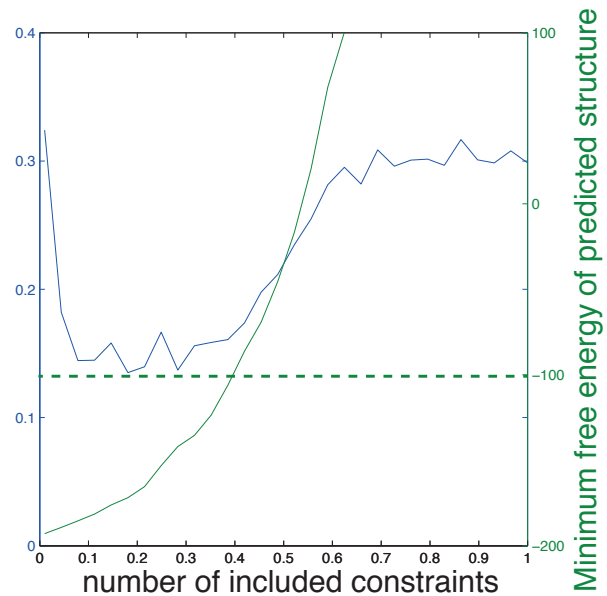
Single structure w/o bias



Single structure with bias



Two structures with bias



Predicting Multiple Structures

Idea:

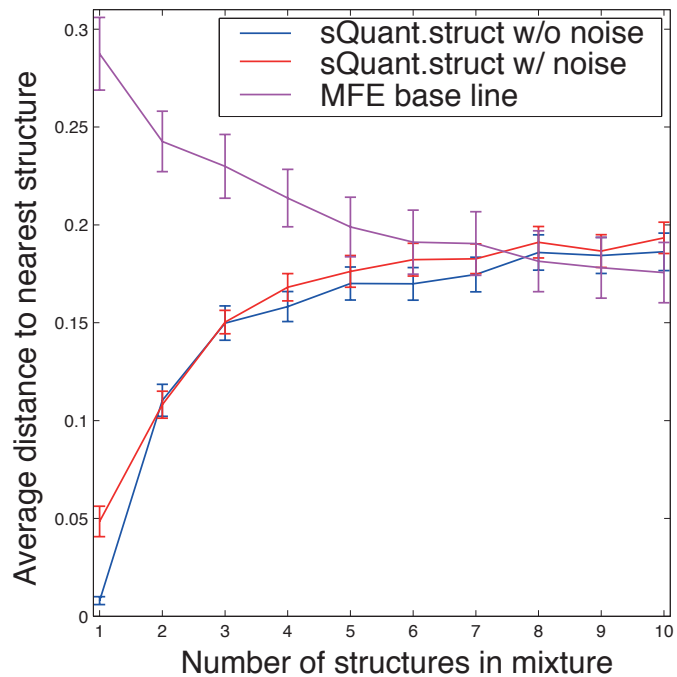
1. Predict most abundant structure s based on M_p^{ds} & M_p^{ss}
2. Quantify structure with sQuant $\Rightarrow w_s$
3. Adjust read counts based on predicted profiles A_p^{ds} & A_p^{ss}

$$M_{p,new}^{ds} = M_p^{ds} - w_s A_p^{ds} \qquad M_{p,new}^{ss} = M_p^{ss} - w_s A_p^{ss}$$

4. Go to 1 or stop

Preliminary results:

- 50 genes with 1-10 structures
- Read counts w/o & w/ biases
- Run sQuant.denovo to identify structures
- Measure distance to nearest prediction



Summary



- Prediction of single structures appears feasible
- Quantitation of multiple know structures works well
  - PARS-seq may have an advantage over RNA-seq for transcript quantitation
- Enzyme digestion leads to predictable, non-uniform read coverage
- Predicting multiple mixed structures appears difficult
- So far based on simulations, need more real data ...

- More information on `http://www.fml.mpg.de/raetsch/suppl/squant`
- The slides can be found on `http://www.fml.mpg.de/raetsch/lectures`

Thank you for your attention!

References

Regina Bohnert and Gunnar Rätsch. rQuant.web: a tool for RNA-Seq-based transcript quantitation. Nucleic Acids Research, 38 (Web Server issue):W348–51, Jul 2010. doi: 10.1093/nar/gkq448.

Regina Bohnert, Jonas Behr, and Gunnar Rätsch. Transcript quantification with RNA-Seq data. BMC Bioinformatics, 10(Suppl 13):P5, Oct 2009.

Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, L. Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast Folding and Comparison of RNA Secondary Structures. Monatshefte für Chemie / Chemical Monthly, 125:167–188, 1994.

Michael Kertesz, Yue Wan, Elad Mazor, John L Rinn, Robert C Nutter, Howard Y Chang, and Eran Segal. Genome-wide measurement of RNA secondary structure in yeast. Nature, 467(7311):103–7, Sep 2010. doi: 10.1038/nature09322.

Andreas Wachter. Riboswitch-mediated control of gene expression in eukaryotes. RNA Biology, 7(1):67–76, 2010.