Simultaneous RNA-seq-based Transcript Inference and Quantification Using Mixed Integer Programming

Jonas Behr, Regina Bohnert, Andre Kahles,
Gunnar Rätsch



Friedrich Miescher Laboratory
of the Max Planck Society

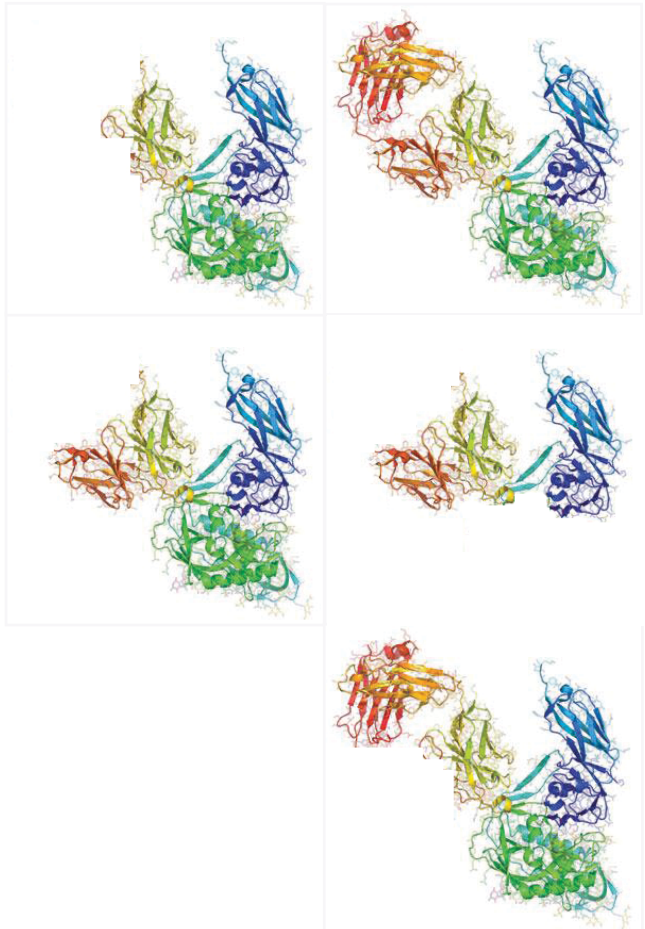
NIPS-MLCB, December, 2011

Motivation: modular proteins

Proteins share/vary in properties like:

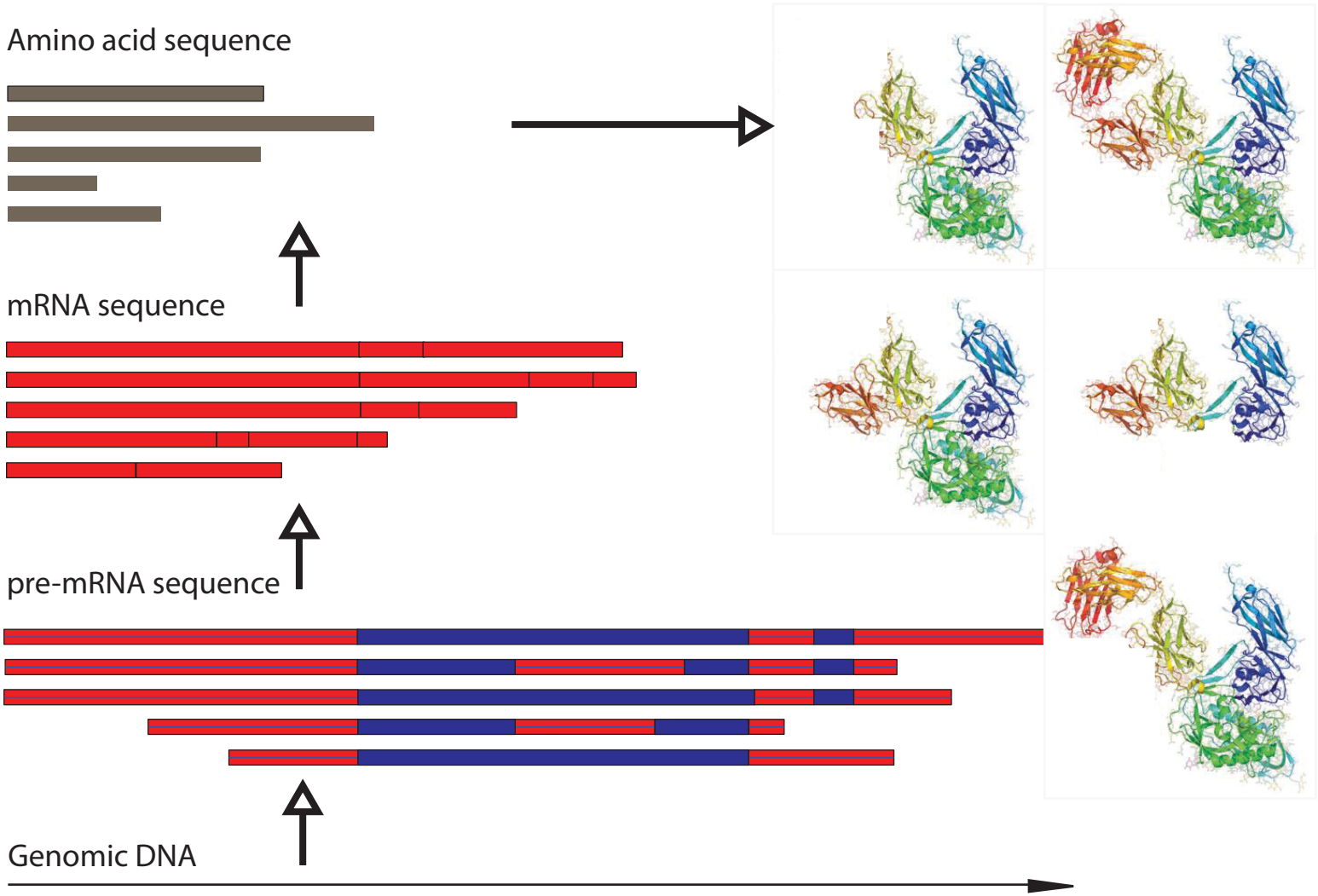
- Function
- Interactions/regulation
- Subcellular location

$\Rightarrow$ We need to know all combinations to understand the system



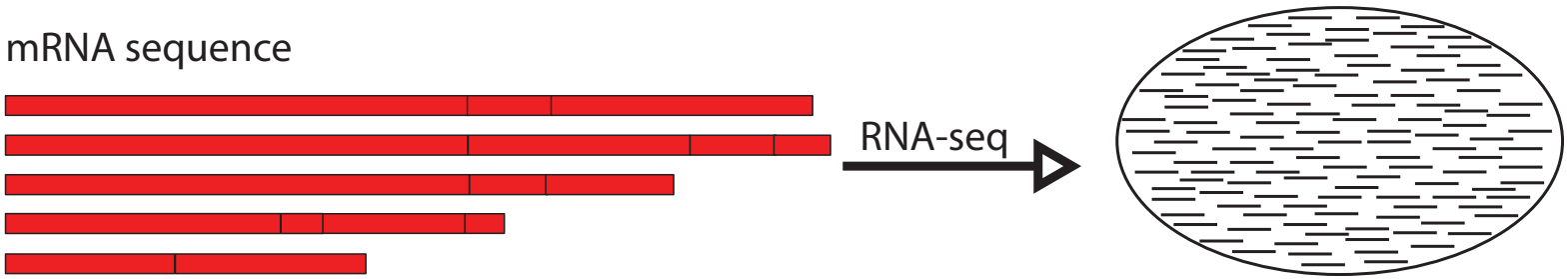
http://www.nature.com/news/2009/090506/images/_tmp_articling-import-20090506091916938530_459024a-i4.0.jpg

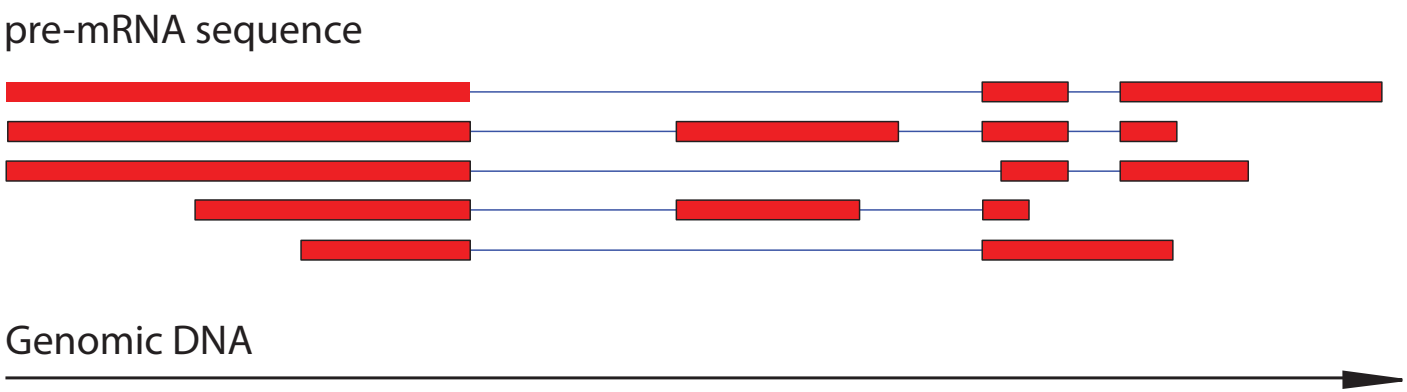
Motivation: modular proteins



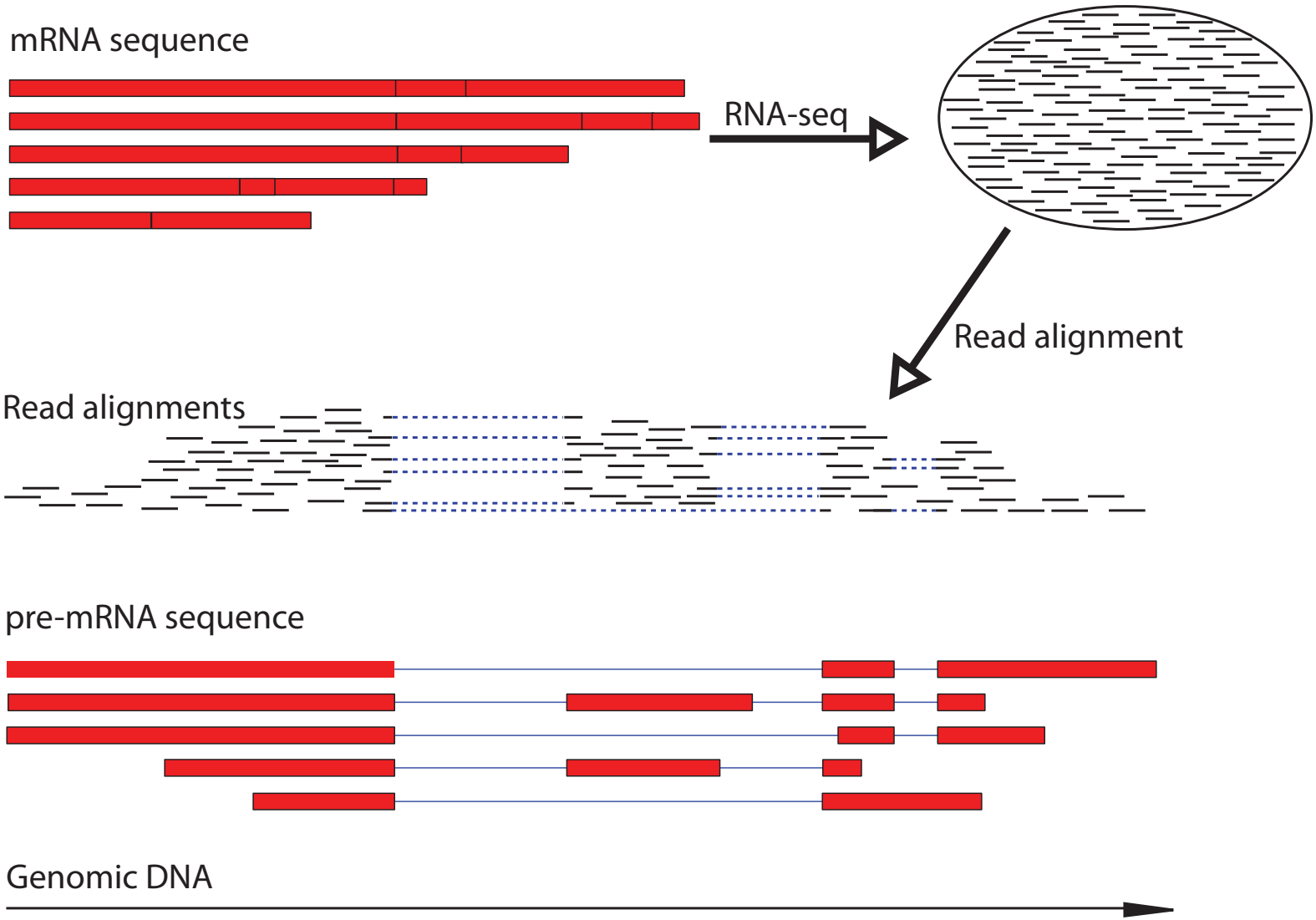
http://www.nature.com/news/2009/090506/images/_tmp_articling-import-20090506091916938530_459024a-i4.0.jpg

Problem definition

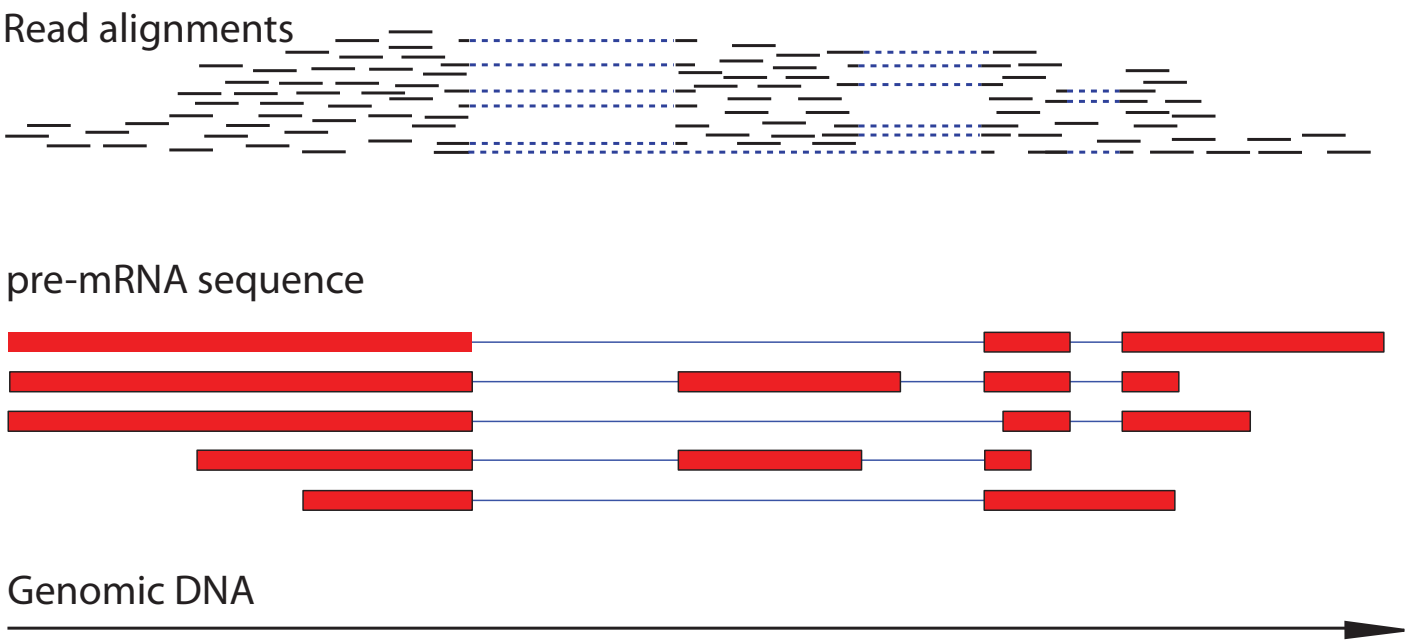




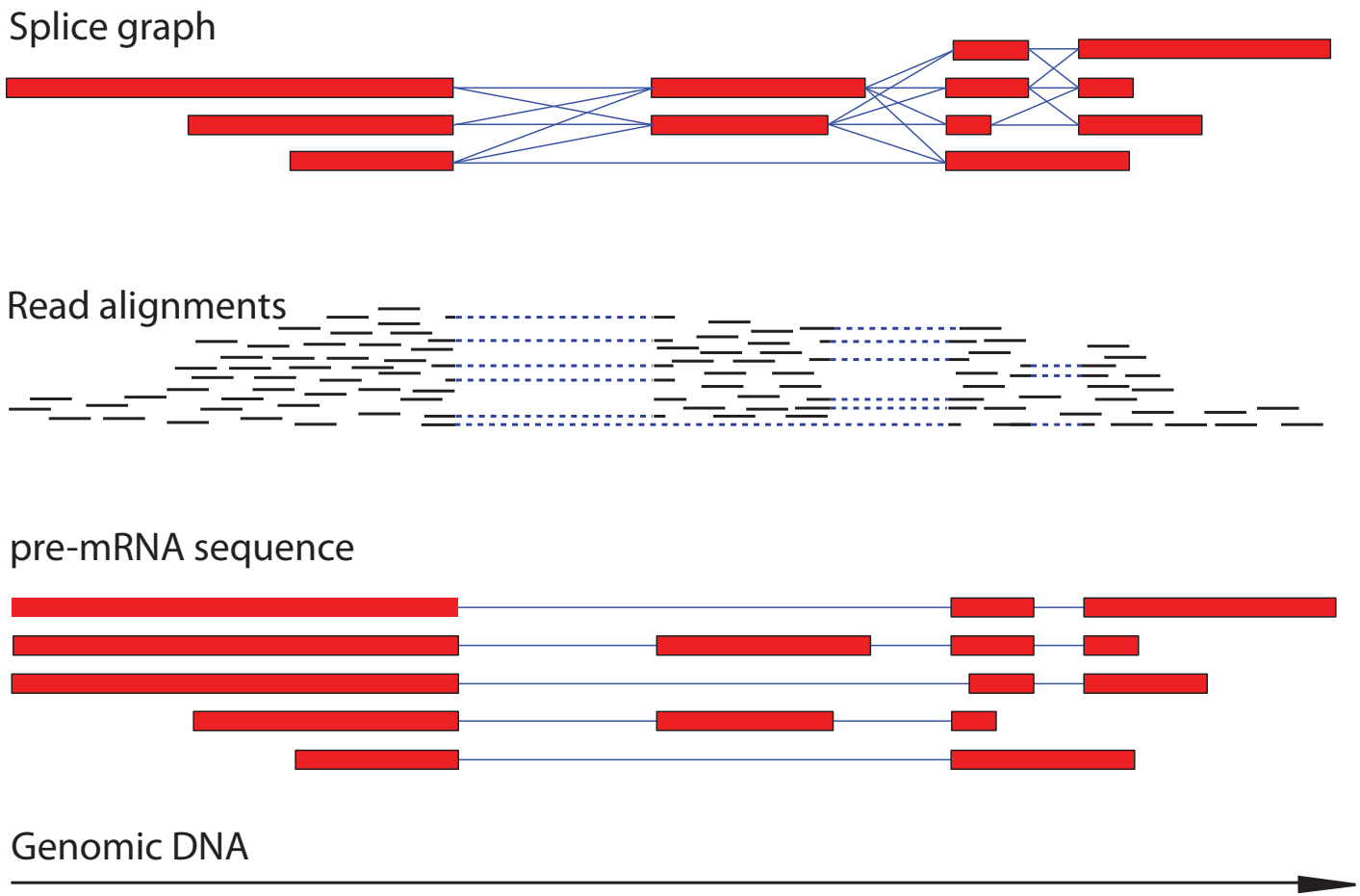
Problem definition



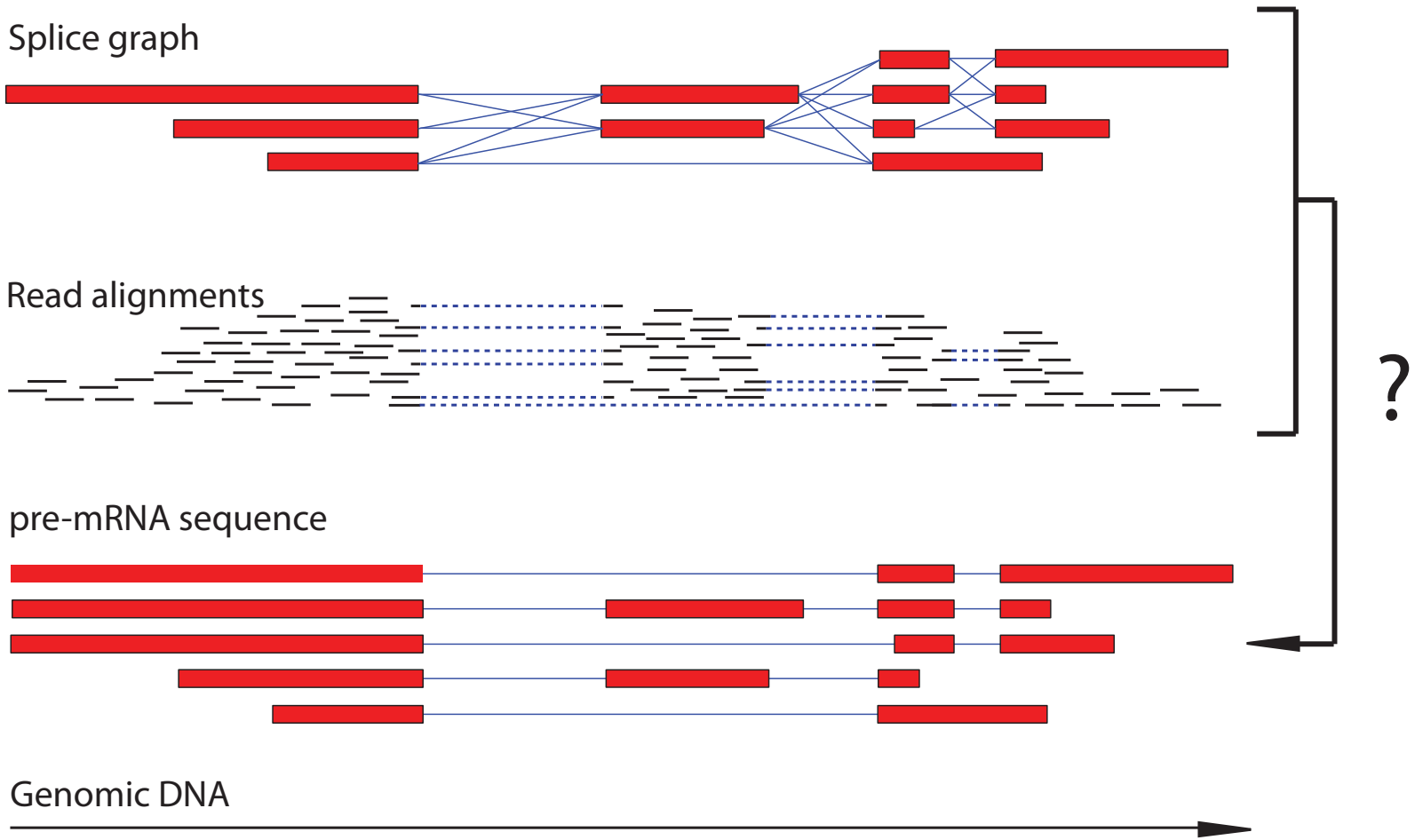
Problem definition



Problem definition



Problem definition



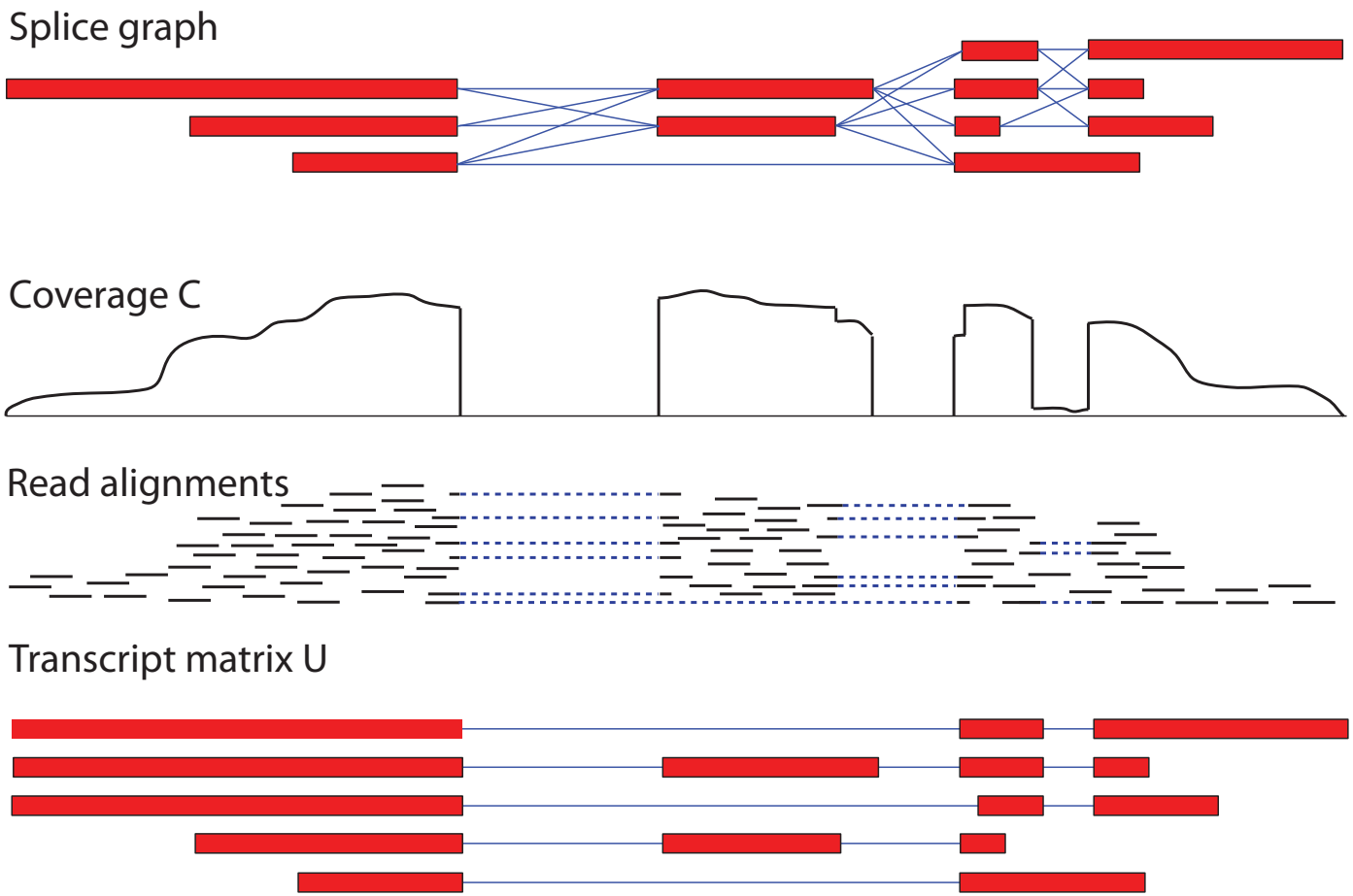
Previous approaches

- Scripture (Guttman et al. [2010]):
  →Report all combinations
  →Output files exponential in size of graph

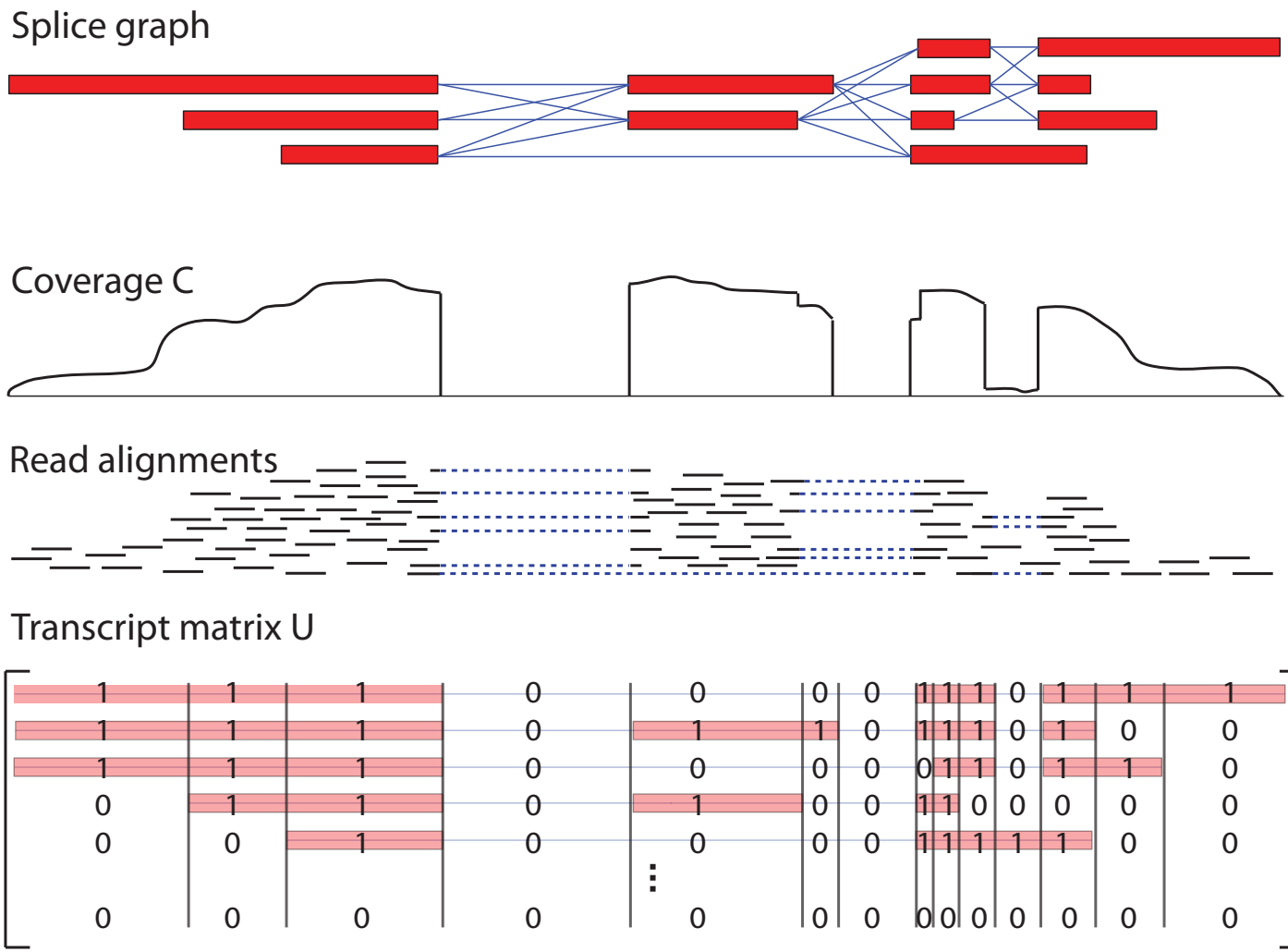
- Cufflinks (Trapnell et al. [2010]):
  →Choose among minimal graph coverages
  →Strong assumption
  →Not robust to errors in the graph

- NSMAP (Xia et al. [2011]):
  → List and quantify all combinations,
  report those with abundance $\geq \epsilon$
  →Optimal results
  →Run time exponential in size of graph

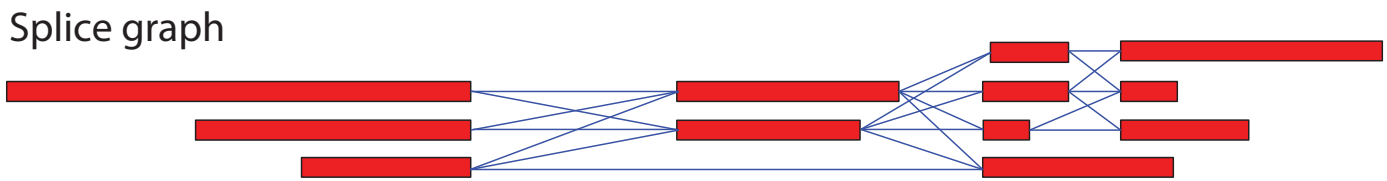
Our model

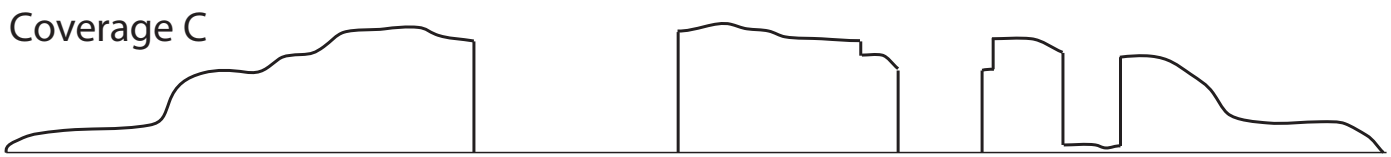

Splice graph
Coverage C
Read alignments
Transcript matrix U

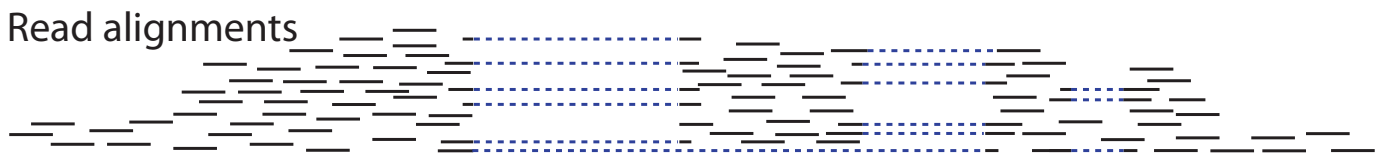
Our model


Splice graph
Coverage C
Read alignments
Transcript matrix U
1
2
3
4
5
k

Our model







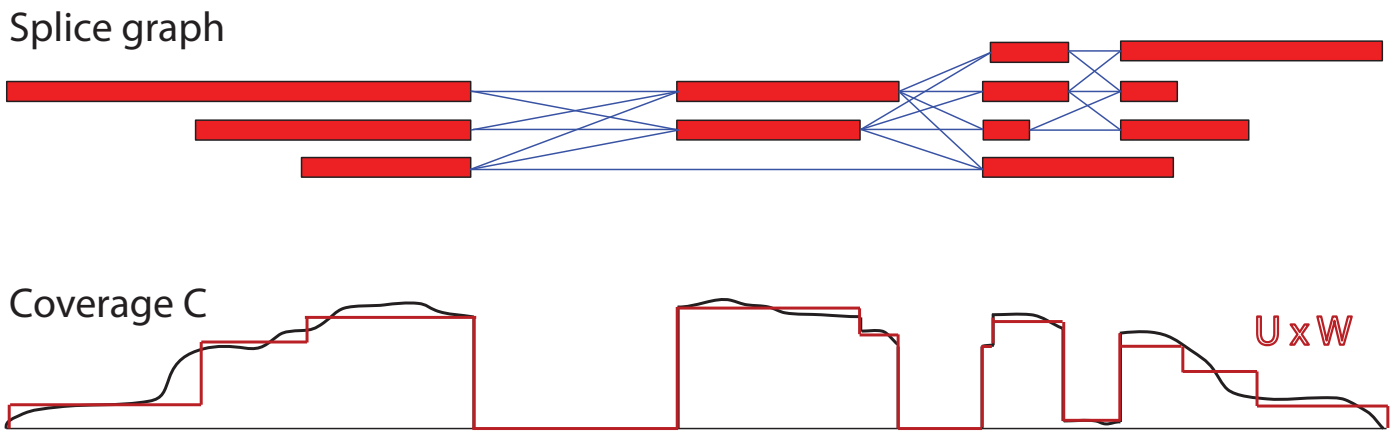
Transcript matrix U

1	1	1	1	0	0	0	0	1	1	1	0	1	1	1								
2	1	1	1	0	1	1	0	1	1	1	0	1	0	0								
3	1	1	1	0	0	0	0	0	1	1	0	1	1	0								
4	0	1	1	0	1	0	0	1	1	0	0	0	0	0								
5	0	0	1	0	0	0	0	1	1	1	1	1	0	0								
⋮					⋮																	
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0								

Abundance W

0.25
0.10
0.45
0.03
0.17
⋮
0.00

Our model


Splice graph
Coverage C
U x W
Read alignments
Transcript matrix U^T
Abundance W
1
2
3
4
5
k
0.25
0.10
0.45
0.03
0.17
0.00

Mixed integer programming (naïve)

$$\min_{U \in \{0,1\}^{n \times k}, W \in [0,1]^k} \| \underbrace{U \times W}_{\text{expected cov.}} - \underbrace{C}_{\text{observed cov.}} \|_2^2 \quad , C \in [0,1]^n$$

How can we solve this?

Substitute:

$$E_{i,j} = U_{i,j} * W_i$$

Add constraints:

$$E_{i,j} \leq U_{i,j}$$

$$E_{i,j} \leq W_i - U_{i,j} + 1$$

$$-E_{i,j} \leq -W_i - U_{i,j} + 1$$

Mixed integer programming (naïve)

$$\min_{U \in \{0,1\}^{n \times k}, W \in [0,1]^k} \| \underbrace{U \times W}_{\text{expected cov.}} - \underbrace{C}_{\text{observed cov.}} \|_2^2 \quad , C \in [0,1]^n$$

How can we solve this?

Substitute:

$$E_{i,j} = U_{i,j} * W_i$$

Add constraints:

$$E_{i,j} \leq U_{i,j}$$

$$E_{i,j} \leq W_i - U_{i,j} + 1$$

$$-E_{i,j} \leq -W_i - U_{i,j} + 1$$

Mixed integer programming (naïve)

$$\min_{U \in \{0,1\}^{n \times k}, W \in [0,1]^k} \| \underbrace{U \times W}_{\text{expected cov.}} - \underbrace{C}_{\text{observed cov.}} \|_2^2 \quad , C \in [0,1]^n$$

How can we solve this?

Substitute:

$$E_{i,j} = U_{i,j} * W_i$$

Add constraints:

$$E_{i,j} \leq U_{i,j}$$

$$E_{i,j} \leq W_i - U_{i,j} + 1$$

$$-E_{i,j} \leq -W_i - U_{i,j} + 1$$

Mixed integer programming (naïve)

$$\min_{U \in \{0,1\}^{n \times k}, W \in [0,1]^k} \| \underbrace{U \times W}_{\text{expected cov.}} - \underbrace{C}_{\text{observed cov.}} \|_2^2 \quad , C \in [0,1]^n$$

How can we solve this?

Substitute:

$$E_{i,j} = U_{i,j} * W_i$$

Add constraints:

$$E_{i,j} \leq U_{i,j}$$

$$E_{i,j} \leq W_i - U_{i,j} + 1$$

$$-E_{i,j} \leq -W_i - U_{i,j} + 1$$

MIP

$$U \times W \approx C$$

- Splice graph not fully connected: valid paths only
- Sparsity:

$$\sum_{i=1}^{k} I(W_i > 0)$$

- Multiple samples

$$U \times W^1 \approx C^1$$
$$\vdots$$
$$U \times W^S \approx C^S$$

MIP

$$U \times W \approx C$$

- Splice graph not fully connected: valid paths only
- Sparsity:

$$\sum_{i=1}^{k} I(W_i > 0)$$

- Multiple samples

$$U \times W^1 \approx C^1$$
$$\vdots$$
$$U \times W^S \approx C^S$$

MIP

$$U \times W \approx C$$

- Splice graph not fully connected: valid paths only
- Sparsity:

$$\sum_{i=1}^{k} I(W_i > 0)$$

- Multiple samples

$$U \times W^1 \approx C^1$$
$$\vdots$$
$$U \times W^S \approx C^S$$

MIP

$$U \times W \approx C$$

- Splice graph not fully connected: valid paths only
- Sparsity:

$$\sum_{i=1}^{k} I(W_i > 0)$$

- Multiple samples

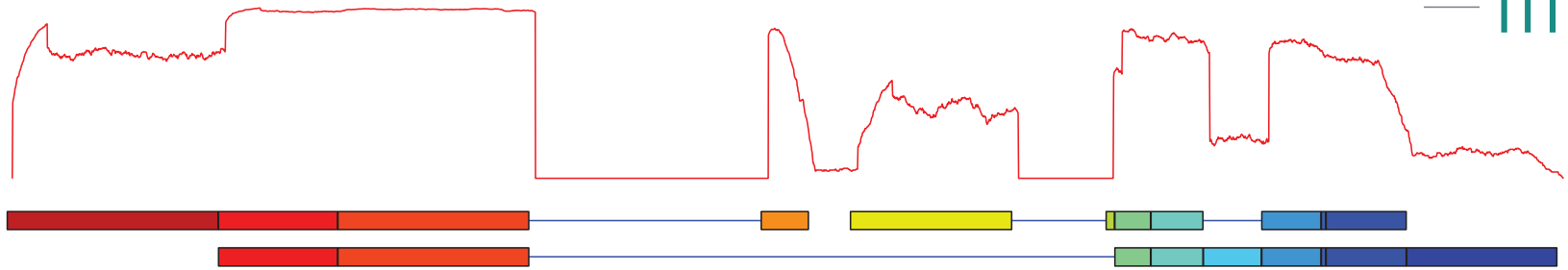
$$U \times W^1 \approx C^1$$
$$\vdots$$
$$U \times W^S \approx C^S$$

Mixed integer optimizaton (simplified)

$$\min_{U,W} \underbrace{\sum_{s=1}^{S} \| U \times W^s - C^s \|_2^2}_{\text{loss}} + \gamma \underbrace{\sum_{i=1}^{k} I(\sum_{s=1}^{S} W_i^s > 0)}_{\text{sparsity}}$$

s.t. 'transcripts are valid paths through the splice graph'

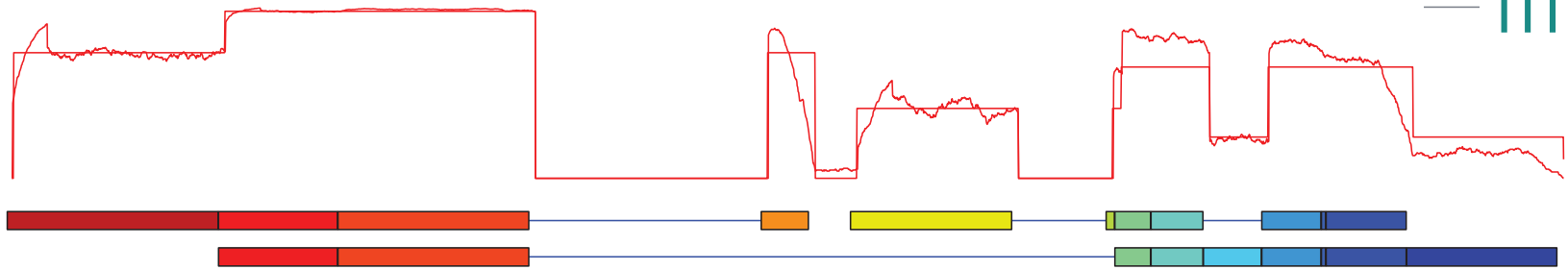
Human genes: artificial reads



1 sample

Annotation

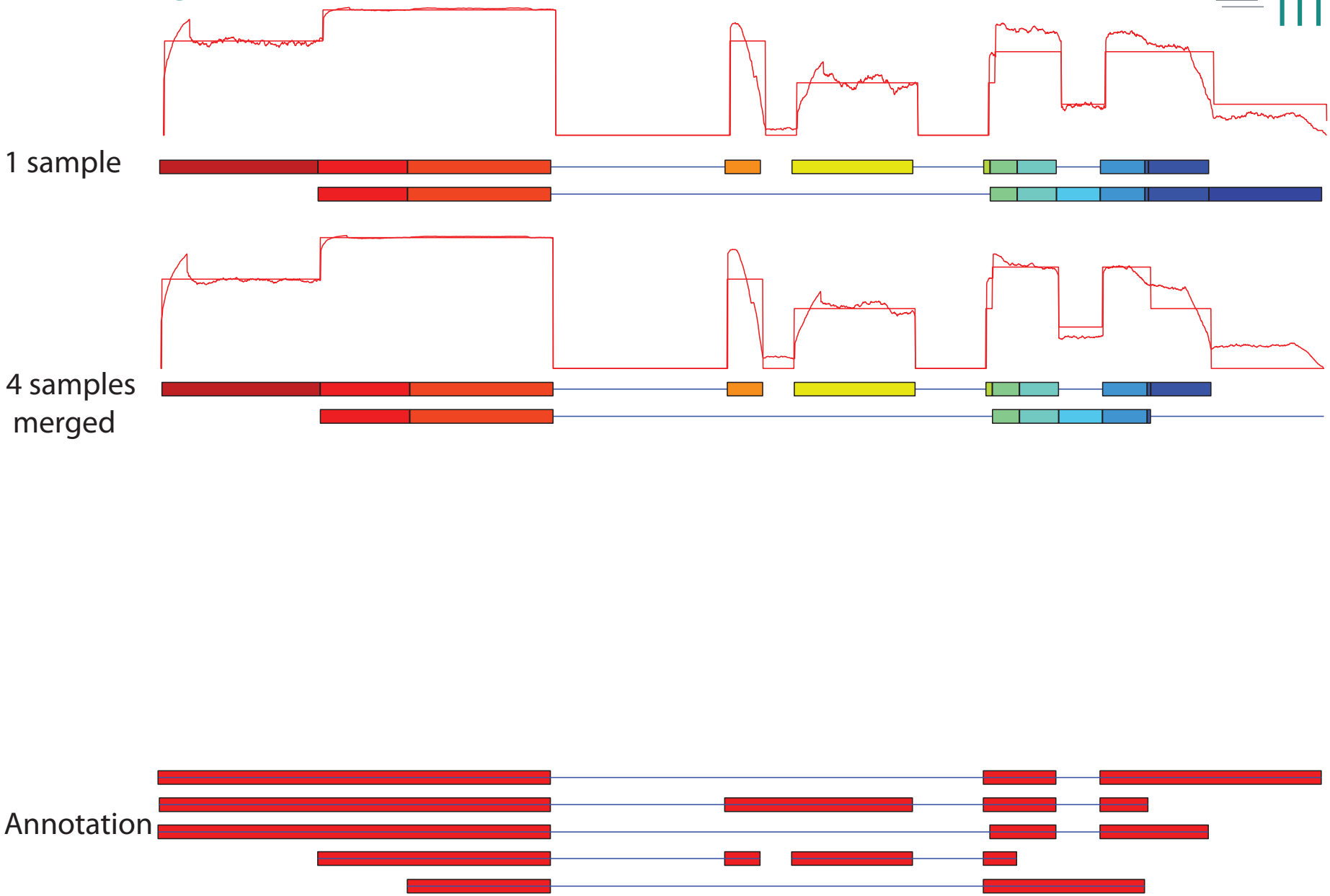
Human genes: artificial reads



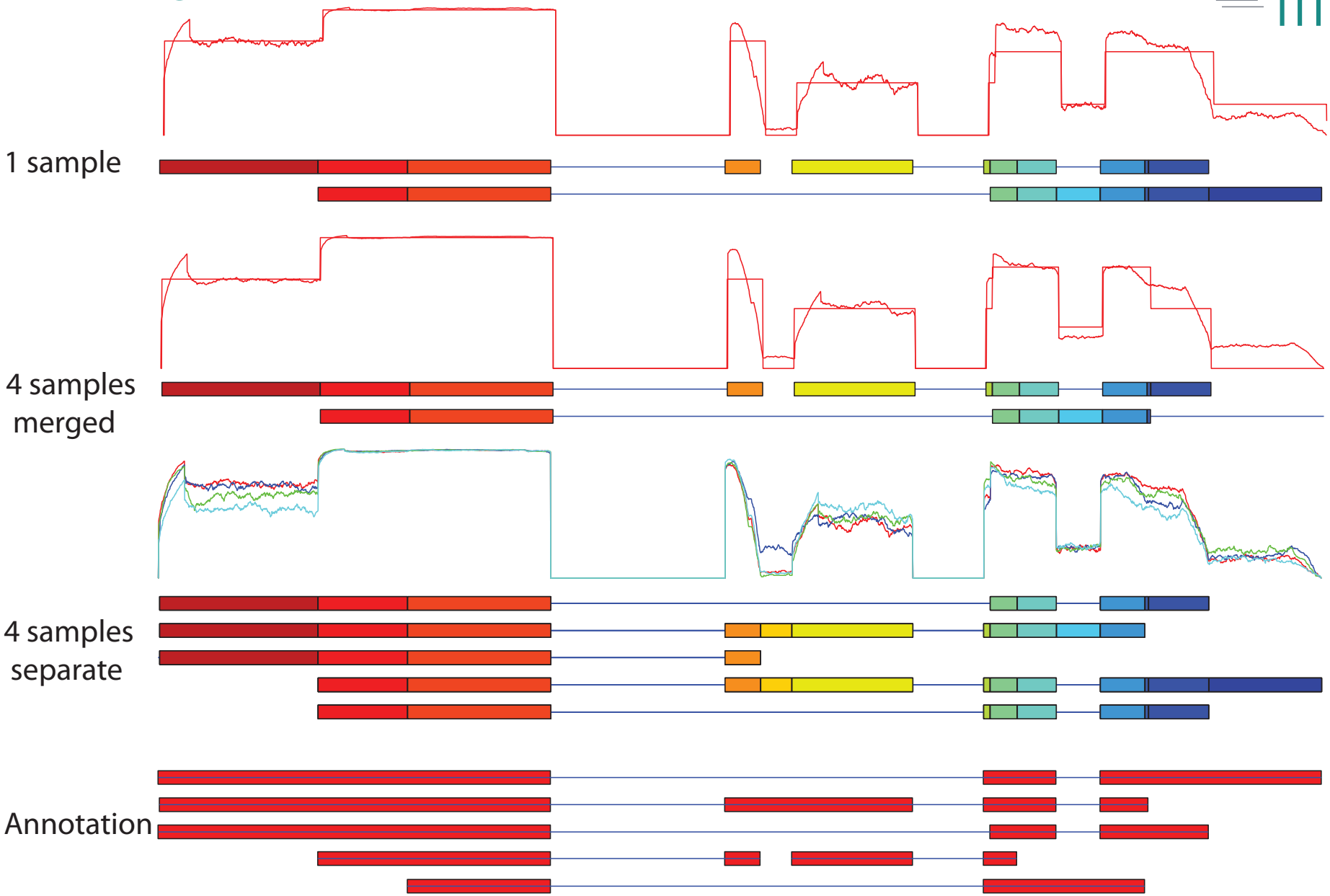
1 sample

Annotation

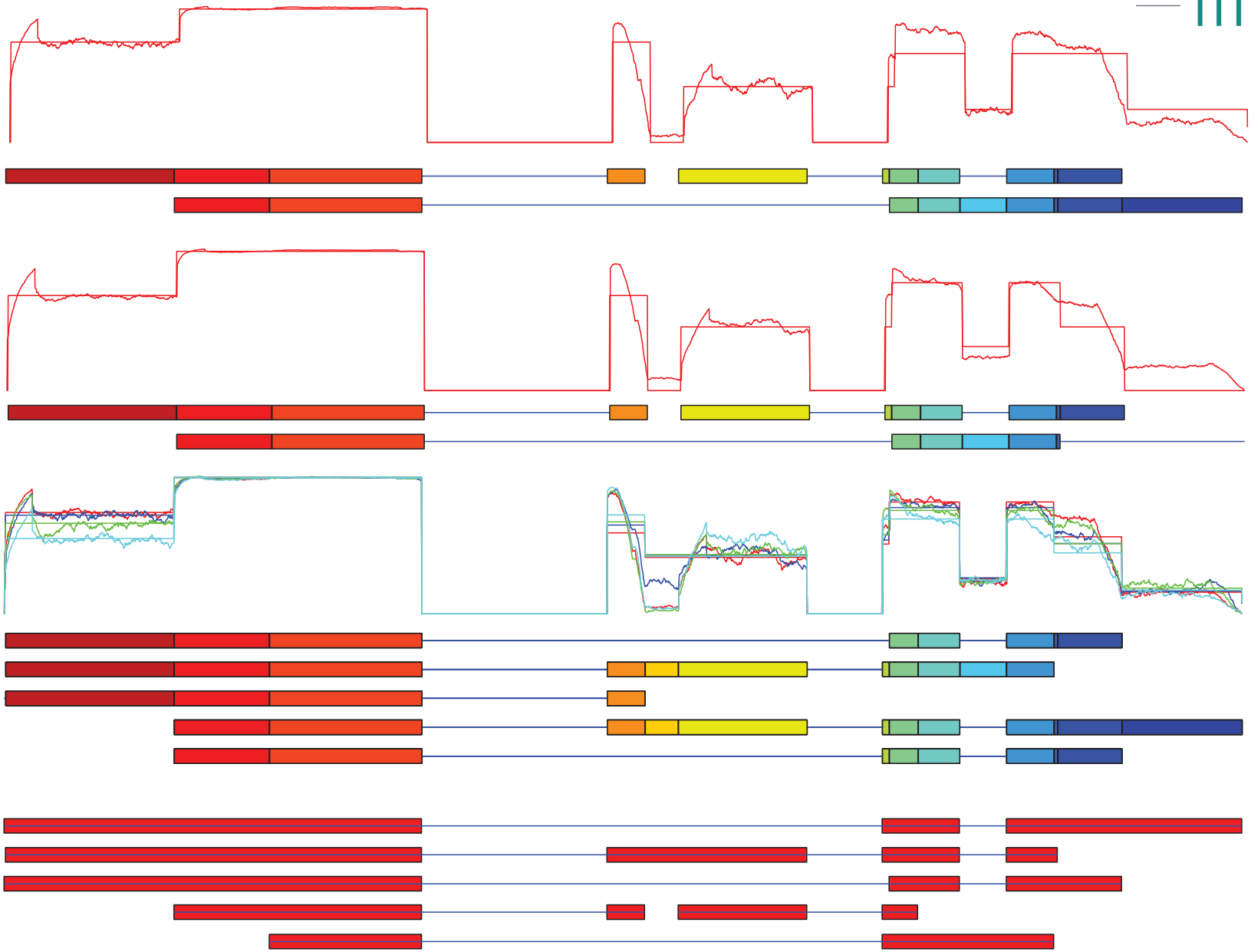
Human genes: artificial reads



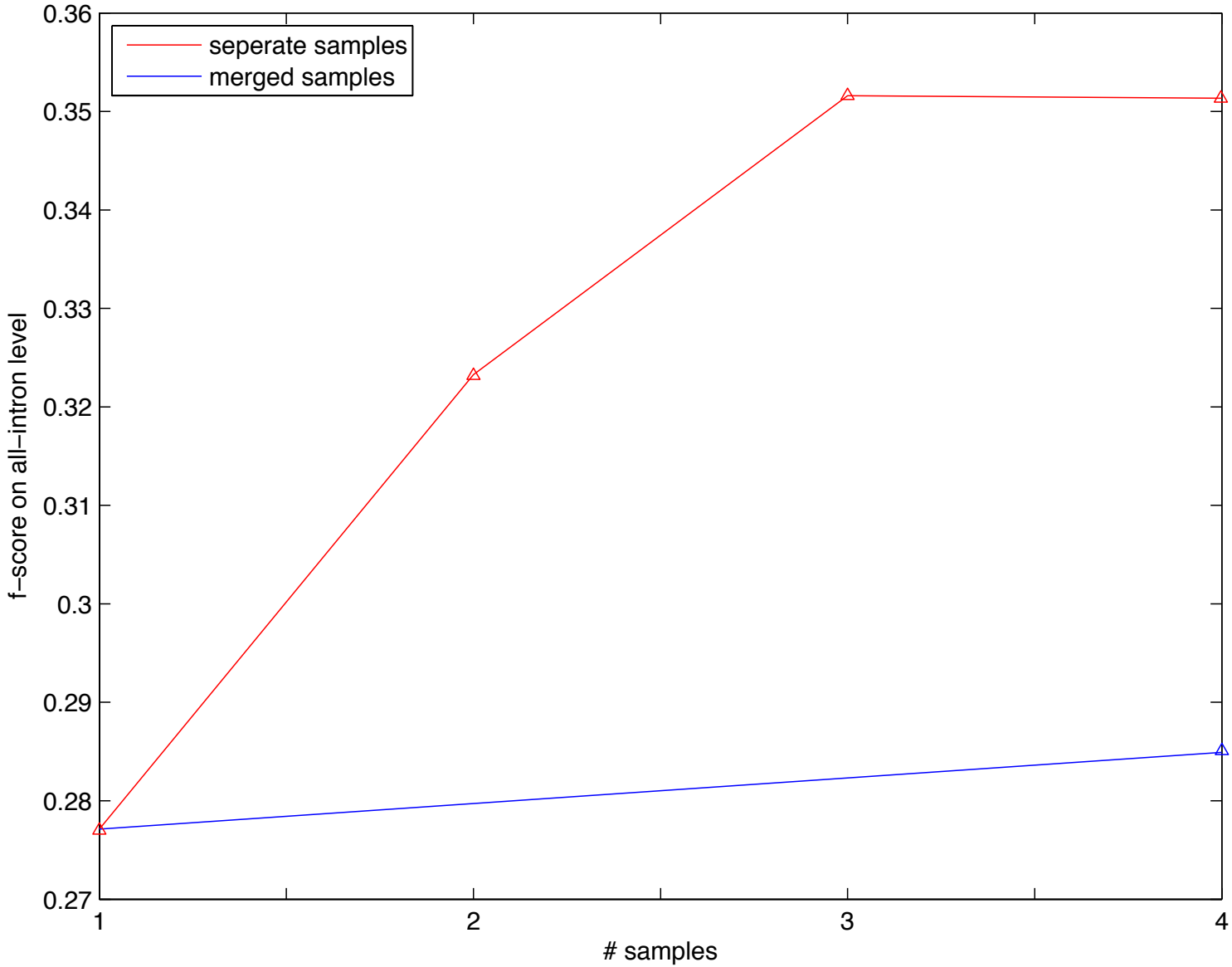
Human genes: artificial reads



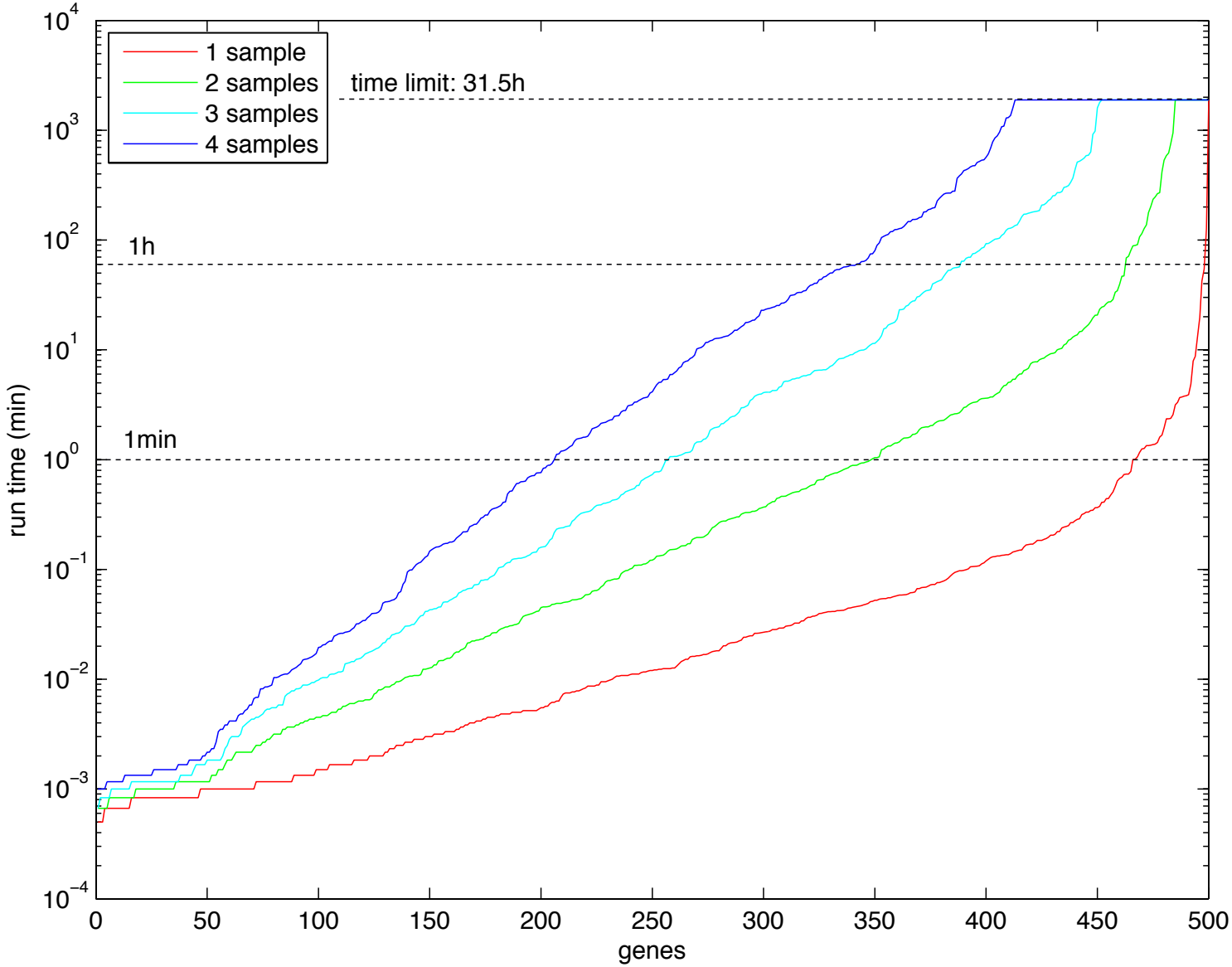
Human genes: artificial reads



Performace: 500 human genes



Run time


1 sample
2 samples
3 samples
4 samples
time limit: 31.5h
1h
1min
run time (min)
genes
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
50
100
150
200
250
300
350
400
450
500

Conclusion

Transcript inference and quantification using MIP:

- Close a gap in RNA-seq analysis which is so far filled with heuristics/inefficient solutions
- Combine quantification and transcript recognition to resolve long range dependencies
- Integrate various expression measurements
- Exploit variation across samples

Conclusion

Transcript inference and quantification using MIP:

- Close a gap in RNA-seq analysis which is so far filled with heuristics/inefficient solutions
- Combine quantification and transcript recognition to resolve long range dependencies
- Integrate various expression measurements
- Exploit variation across samples

Conclusion

Transcript inference and quantification using MIP:

- Close a gap in RNA-seq analysis which is so far filled with heuristics/inefficient solutions
- Combine quantification and transcript recognition to resolve long range dependencies
- Integrate various expression measurements
- Exploit variation across samples

Conclusion

Transcript inference and quantification using MIP:

- Close a gap in RNA-seq analysis which is so far filled with heuristics/inefficient solutions
- Combine quantification and transcript recognition to resolve long range dependencies
- Integrate various expression measurements
- Exploit variation across samples

Acknowledgments

Read simulation:	**Michael Sammeth, Philipp Drewe**
Quantification:	**Regina Bohnert**
Alignments:	**André Kahles, Géraldine Jean, Gunnar Rätsch**
Programming:	**Vipin T Sreedharan**
Shogun:	**Sören Sonnenburg**
Discussions:	**Marcel Schulz, Georg Zeller, Christian Widmer, Nico Görnitz**
Supervision, Ideas:	**Gunnar Rätsch**
Funding:	**Max Planck Society**

References

Jonas Behr, Regina Bohnert, Andre Kahles, and Gunnar Rätsch. Simultaneous rna-seq-based transcript inference and quantification using mixed integer programming. i.p.

Mitchell Guttman, Manuel Garber, Joshua Z. Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J. Koziol, Andreas Gnirke, Chad Nusbaum, John L. Rinn, Eric S. Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5):503–510, May 2010. ISSN 1087-0156. doi: 10.1038/nbt.1633. URL `http://dx.doi.org/10.1038/nbt.1633`.

Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, May 2010. ISSN 1087-0156. doi: 10.1038/nbt.1621. URL `http://dx.doi.org/10.1038/nbt.1621`.

Zheng Xia, Jianguo Wen, Chung C. Chang, and Xiaobo Zhou. NSMAP: A method for spliced isoforms identification and quantification from RNA-Seq. *BMC Bioinformatics*, 12(1):162+, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-162. URL `http://dx.doi.org/10.1186/1471-2105-12-162`.