Supplemental Material
Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels

Marcel H. Schulz[1,2,4,*], Daniel R. Zerbino[2,3,*], Martin Vingron[1], and Ewan Birney[2,†]

[1] Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 63, 14195 Berlin, Germany
[2] European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom
[3] Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California, United States of America
[4] Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America
[*] joint first authors, [†] corresponding author

Contents

Dynamic filtering of the scaffold

Oases uses a dynamic filtering approach for paired-end connections that utilizes the average insert size, standard deviation and the coverage of two contigs to compute an expected value for read pair coverage. In genomic assembly, the number of connecting read pairs between contigs A and B is estimated to be:

$$E(X) = \rho_A \left[\sigma \left(\Phi(M) - \Phi(N) - M \int_M^N \Phi \right) + l_B \int_N^O \Phi - \sigma \left(\Phi(O) - \Phi(P) - P \int_O^P \Phi \right) \right]$$

$$\begin{aligned} M &= \frac{D - \mu}{\sigma} \\ N &= \frac{D + l_B - \mu}{\sigma} \\ O &= \frac{D + l_A - \mu}{\sigma} \\ P &= \frac{D + l_A + l_B - \mu}{\sigma}, \end{aligned}$$

where ρ_A is the density of reads in A, l_A and l_B the lengths of A and B (assuming that $l_A \geq l_B$), and ϕ is the standard normal probability distribution function.

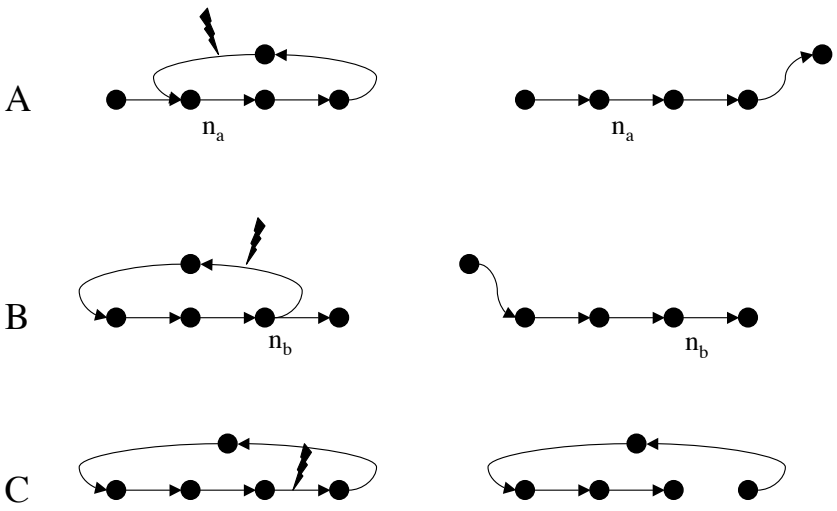
However, this model assumes that the genome is linear, and that the probability of a read pair connecting the two nodes depend solely on read position and insert length. It neglects the fact that a read pair anchored to A can fail to connect to B because it belongs to a transcript from which B is absent. Assuming that at most one of two neighboring exons is involved in any alternative splicing event, the density ρ_A is replaced by the minimum of ρ_A, ρ_B, as it corresponds to the read density on the putative transcript joining A and B. As mentioned in the manuscript, by default 10% of the expected value is used as a cutoff, to remove unlikely connections.

Dynamic programming algorithm

The resolution of complex locus graphs is applied in turn to each locus that do not present a trivial topology, a dynamic programming (DP) algorithm produces a representative set of paths through the graph, giving priority to high coverage paths.

Cycle removal

The DP-algorithm was originally designed to work on acyclic graphs, so that its completion could be garanteed [3]. Cycles can however appear in de Bruijn graphs, and these would block the propagation of the algorithm. Oases distinguishes between three cases, as described in Supplemental Figure



Supplemental Figure 1: **Removal of cycles in de Bruijn Graphs** that block the DP-recursion from n_a. Nodes are displayed as black dots. (A) n_a has a predecessor node which is external to the cycle, so the cycle edge leading to n_a is destroyed. (B) The cycle does not contain any node with predecessors outside of the cycle. However, n_b has a successor node which is outside of the cycle. The cycle edge which goes out of n_b is therefore broken. (C) The locus is circular, an edge is broken at random.

1, depending on whether the cycle has predecessor nodes which are outside of the cycle. In each case a specific rule is applied to break the cycle but preserve the sequence, while keeping the number of paths through the locus to a minimum.

Iterated traversal

After assigning a weight to each node and edge, each node is assigned a predecessor, based on the edge weights.

$$chooseBestPredecessor(j) \ = \arg \max_{\forall i,\ n_i = pred(n_j)} (w_{ij}), \qquad (1)$$

where $pred(n_j)$ denotes a predecessor node of n_j.

The score assigned to each node is the sum of its predecessor's score with the node's weight. After all nodes in the graph have been processed the highest expressed transcript is found by backtracking from the node in the graph with maximum score $\max_i s_i$.

After each iteration of the DP algorithm, the unvisited node with heaviest weight and its connected edges are temporarily upweighted to infinite weight.

The purpose of this approach is to ensure that long and highly expressed nodes which are not part of any predicted transcript from the locus are part of the next generated transcript. Unassigned nodes are sorted by expression and the most highly expressed node n_i is selected.

Oases re-iterates this process till all nodes are explained by a transcript, or the number of transcripts hits a limit (by default 10). Long nodes which were not included as a transcript are then added as singleton assemblies.

Sensitivity and Specificity

Sensitivity and *specificity* are measures of the performance of classification tests. Let TP be the number of *true positive* predictions, FP the number of *false positive* predictions, and FN be the number of *false negative* predictions, then

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TP}{TP + FP} \tag{3}$$

These measures are computed for different levels of the annotation as described in more detail in [2].

$$\text{Nucleotide Sensitivity} = \frac{\text{number of correct bases}}{\text{number of annotated transcriptomic bases}} \tag{4}$$

$$\text{Nucleotide Specificity} = \frac{\text{number of correct bases}}{\text{number of predicted transcriptomic bases}} \tag{5}$$

Nucleotide sensitivity and nucleotide specificity are used to assess the prediction performance against genomic basepairs contained in known transcript annotation.

Computation of expression values

For each data set the reads were aligned onto the complete transcriptome of Ensembl with RazerS [5]. The minimum sequence identity required was 92 % and further the following parameters were set for RazerS *-m 20 -dr 2 -pa -mN -of 1 -s 1111011010001110011 -t 3*. Read counts have been summarized on the gene level in order to compute Reads Per Kilobase per Million mapped reads (RPKM) values for each gene [4]

$$RPKM = \frac{10^9 \cdot Y}{T \cdot L}, \tag{6}$$

H	M	k	edgeFraction cutoff (%)	transfrags (≥100bps)	N-Sens	N-Spec	Cov ≥ 100% #transcripts (#genes)	Cov = 80% #transcripts (#genes)
✓		19	1	70537	19.9	91.71	730(587)	6730(2710)
✓		19	5	70071	19.9	91.77	744(586)	6965(2733)
✓		19	10	67319	20.01	91.67	779(605)	7319(2817)
✓		19	12	66096	20.06	91.49	783(615)	7467(2871)
✓		19	15	64613	20.11	91.4	794(630)	7513(2895)
✓		19	20	63331	20.12	91.25	793(638)	7672(2934)
✓		35	1	34370	7.72	94.91	166(143)	1789(771)
✓		35	5	34104	7.73	94.95	171(148)	1813(776)
✓		35	10	34012	7.74	94.93	171(148)	1828(781)
✓		35	12	33931	7.73	94.94	170(147)	1823(779)
✓		35	15	33872	7.74	95.0	169(146)	1828(777)
✓		35	20	33801	7.73	95.01	171(147)	1841(780)
	✓	21	1	62245	30.12	92.05	890(745)	8399(3922)
	✓	21	5	59033	30.13	91.82	924(765)	8498(3952)
	✓	21	10	57808	30.12	91.73	923(766)	8441(3948)
	✓	21	12	58041	30.13	91.79	896(745)	8348(3904)
	✓	21	15	57705	30.14	91.7	884(732)	8266(3897)
	✓	21	20	57141	30.14	89.09	874(729)	8257(3894)
	✓	35	1	62602	20.87	92.57	209(196)	2330(1398)
	✓	35	5	62173	20.87	92.61	212(198)	2343(1410)
	✓	35	10	61964	20.89	92.73	214(199)	2363(1417)
	✓	35	12	61892	20.89	92.72	212(198)	2368(1418)
	✓	35	15	61851	20.89	92.72	212(198)	2366(1417)
	✓	35	20	61722	20.89	92.81	211(198)	2373(1417)

Supplemental Table 1: Analysis of single k Oases transcriptome assemblies for different values of the edgeFraction cutoff parameter applied to the human CD4 and mouse C2C12 data set. N-Sens and N-Spec denote nucleotide sensitivity and specificity. The last two columns show the number of reconstructed Ensembl transcripts to full length or at least 80%. The number in brackets denotes the corresponding genes.

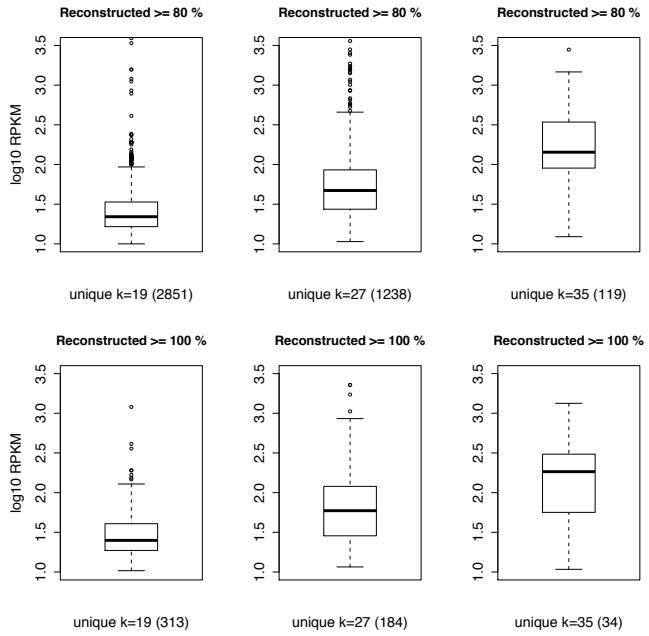
where Y is the observed number of reads in the gene, T the total number of mapped reads in the experiment, and L the gene length. Matches of non-unique reads were equally weighted with $\frac{1}{\delta}$, where δ represents the number of different mapping positions in the transcriptome.

Influence of local edge removal

We illustrate the importance of dynamic filters to the de Bruijn graph with an analysis of the local edge removal parameter (termed *edgeFraction cutoff* in the Oases software). In Supplemental Table 1 the effect of varying the edgeFraction cutoff can be observed. As a default value Oases uses 10% which shows good performance for both k-values tested in both of the data sets.

Influence of parameter k

In the absence of coverage gaps or errors, a large k is preferable over a smaller k, in order to simplify the locus topology. But the common problem of de Bruijn graph based sequence assemblers is the susceptibility to sequencing errors. Each sequencing error can destroy up to k k-mers and therefore the



Supplemental Figure 2: **Comparison of uniquely reconstructed Ensembl transcripts between different Oases single k assemblies.** Each box plot shows the gene expression levels (log10 RPKM, y-axis) of all uniquely reconstructed transcripts for a particular Oases single k assembly, where k=19, 27, 35 in the *left, middle and right column*, respectively. Uniquely means that none of the other two single-k assemblies considered here has reconstructed the transcript to the analyzed length of at least 80% *(top row)* or 100% *(bottom row)*. The number in brackets in each x-axis label denotes the total number of unique reconstructions. Unique transcripts for large k emanate from genes with a high mean expression level with RPKM > 100.

$k_{MIN} - k_{MAX}$	transfrags (≥100bps)	N-Sens	N-Spec	Cov ≥ 100% #transcripts (#genes)	Cov = 80% #transcripts (#genes)
19-23	106468	21.06	91.78	1229(922)	10619(3683)
19-25	142964	21.48	91.49	1278(961)	10804(3743)
19-27	157207	21.54	91.56	1368(1016)	11028(3793)
19-29	167700	21.68	91.68	1412(1050)	11107(3819)
19-31	176662	21.7	91.78	1445(1070)	11139(3839)
19-33	185119	21.78	91.89	1455(1075)	11156(3840)
19-35	192320	21.79	92.0	1455(1073)	11159(3837)
21-25	112716	20.58	91.5	1130(857)	9803(3471)
21-27	127626	20.61	91.65	1218(910)	10071(3537)
21-29	138449	20.64	91.81	1259(941)	10171(3565)
21-31	147567	20.62	91.89	1300(965)	10220(3583)
21-33	156122	20.65	92.04	1317(976)	10246(3587)
21-35	163299	20.69	92.17	1315(973)	10262(3586)
23-27	106638	19.51	91.54	1010(773)	8648(3152)
23-29	117754	19.48	91.73	1073(814)	8805(3192)
23-31	126849	19.48	91.9	1114(839)	8876(3212)
23-33	135424	19.48	92.09	1126(848)	8904(3213)
23-35	142525	19.51	92.28	1125(847)	8929(3213)
25-29	100038	18.54	91.81	905(688)	7394(2768)
25-31	109062	18.49	91.98	948(714)	7505(2790)
25-33	117479	18.51	92.24	967(729)	7571(2801)
25-35	124657	18.45	92.42	973(731)	7604(2803)

Supplemental Table 2: Analysis of Oases-M merged transcriptome assemblies for different values of the de Bruijn graph parameter for the lowest and highest k on the human CD4 data set. All Oases-M assemblies use $k_{MERGE} = 27$. N-Sens and N-Spec denote nucleotide sensitivity and specificity. The last two columns show the number of reconstructed Ensembl transcripts to full length or at least 80%. The number in brackets denotes the corresponding genes.

influence of sequencing errors has to be balanced against the gain in repeat resolution when choosing k. In RNA-seq data an additional difficulty is that a major fraction of the expressed transcripts is represented with very few sequence reads as often only 5% of the expressed genes contribute $\geq 50\%$ of all reads [1]. In consequence, a large k will favor highly expressed transcripts.

Single k assemblies

In order to observe these oppositional influences single k assemblies of Oases for $k = 21$, 27, and 35 on the human CD4 data set have been compared. Uniquely reconstructed transcripts were compared between the single k assemblies. The box plots in Supplemental Figure 2 show the expression distribution of uniquely reconstructed transcripts. The reconstruction of highly expressed genes improves from small to high k values and the benefit of using large k values is clearly to reconstruct highly expressed genes.

Analysis of k range for merged assembly

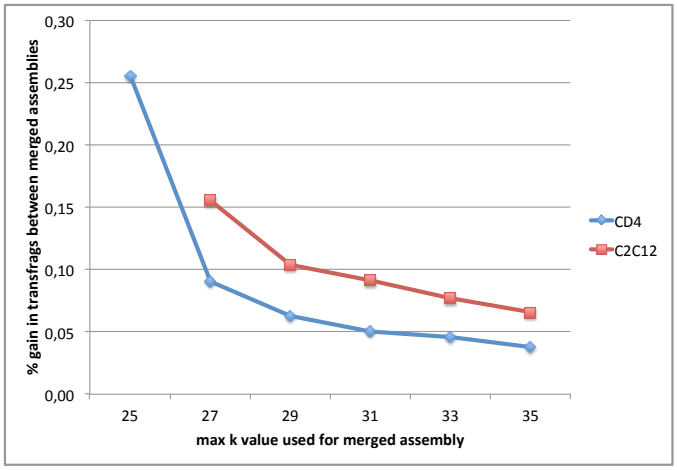
Given the observations in the paper and the previous subsection it is obvious that the range of k values used for the Merged assembly influences the reconstruction accuracy. Supplementary Tables 2 and 3 show that it is im-

$k_{MIN} - k_{MAX}$	transfrags (≥100bps)	N-Sens	N-Spec	Cov ≥ 100% #transcripts (#genes)	Cov = 80% #transcripts (#genes)
21-25	104328	30.69	91.89	1179(958)	9617(4356)
21-27	123497	30.69	88.74	1224(997)	9746(4401)
21-29	137774	30.71	89.05	1266(1029)	9814(4428)
21-31	151567	30.8	89.38	1294(1053)	9880(4453)
21-33	164305	30.78	89.62	1312(1068)	9896(4464)
21-35	175906	30.84	89.08	1329(1083)	9891(4469)
23-27	100669	29.34	87.37	945(789)	8523(3979)
23-29	116108	29.35	88.08	999(822)	8665(4029)
23-31	130284	29.36	91.03	1029(850)	8734(4053)
23-33	143465	29.36	91.88	1052(868)	8749(4062)
23-35	155052	29.39	91.99	1062(878)	8768(4075)
25-29	97716	28.16	91.74	813(693)	7361(3551)
25-31	112672	28.12	91.83	837(715)	7450(3589)
25-33	126055	28.08	91.87	856(734)	7502(3609)
25-35	137777	28.03	91.93	870(740)	7532(3626)
27-31	95729	26.71	82.57	652(573)	6208(3113)
27-33	109760	26.66	84.11	676(594)	6281(3142)
27-35	121852	26.63	87.98	697(607)	6330(3170)

Supplemental Table 3: Analysis of Oases-M merged transcriptome assemblies for different values of the de Bruijn graph parameter for the lowest and highest k on the mouse C2C12 data set. All Oases-M assemblies use $k_{MERGE} = 27$. N-Sens and N-Spec denote nucleotide sensitivity and specificity. The last two columns show the number of reconstructed Ensembl transcripts to full length or at least 80%. The number in brackets denotes the corresponding genes.

portant to start with a low k to achieve high sensitivity but that extending to a higher k also results in a higher number of reconstructed transcripts for the mouse and human data. However, there are some differences. In Supplementary Table 2 it can be seen that adding the assembly with k=35 only leads to an improvement of reconstructed transcripts for k_{MIN}=25. For smaller values of k_{MIN} there is no difference. In contrast, for the mouse data in Supplementary Table 3 newly reconstructed transcripts are added up to k_{MAX}=35. One may conclude that adding a higher k_{MAX} might be beneficial for the mouse dataset, and in turn using k_{MAX}=33 for human might suffice.

It is interesting to consider if such a question can be answered without having annotation to compare to, namely in a pure *de novo* setting. In Fig. 3 the relative gain (%) in number of transfrags for a stepwise increment of the k_{MAX} value is analyzed. For both datasets the % of added transfrags decays rapidly. For the human dataset less than 5% new transfrags are added after k_{MAX}=33. From the analysis before it is suggestive to use 5% relative gain as a reasonable cutoff to use in practice to stop the Merged assembly if no or limited annotation is available.



Supplemental Figure 3: **Relative increment of assembled transfrags for merged assemblies in a *de novo* setup.** It is explored how addition of a larger single k assembly benefits the merged assembly. The y-axis shows the relative gain (%) in the number of transfrags produced by the Oases-M merged assembly with the max k value tested (x-axis). Merged assemblies use a fixed k_{MIN} value for the human CD4 (blue) and mouse C2C12 (red) data sets with k_{MIN}=19 and k_{MIN}=21, respectively.

k_{MERGE}	transfrags (≥100bps)	N-Sens	N-Spec	Cov ≥ 100% #transcripts (#genes)	Cov = 80% #transcripts (#genes)	95% aligned
19	185946	21.62	92.51	1441(1055)	11090(3815)	75
25	174444	21.44	92.37	1461(1069)	11148(3833)	78
27	174469	21.44	92.35	1463(1070)	11169(3837)	79
29	174437	21.41	92.34	1456(1068)	11156(3836)	79
31	174667	21.42	92.34	1456(1069)	11139(3831)	79
35	175009	21.42	92.34	1457(1070)	11148(3834)	79

Supplemental Table 4: **Analysis of Oases-M merged transcriptome assemblies with different k_{MERGE} values** used on the CD4 data set. The range for the merged assembly is the same with k_{MIN}=19 and k_{MAX}=35. N-Sens and N-Spec denote nucleotide sensitivity and specificity. The 5th and 6th column show the number of reconstructed Ensembl transcripts to full length or at least 80%, respectively. The number in brackets denotes the corresponding genes. The last column denotes the percentage of transfrags that aligned to the genome with at least 95% of length.

H	M	c	n	transfrags (≥100bps)	N-Sens	N-Spec	Cov ≥ 100% #transcripts (#genes)	Cov = 80% #transcripts (#genes)	95% aligned
✓		2	10	73729	20.26	91.43	1119(844)	9644(3454)	84
✓		2	5	143833	21.47	91.74	1275(980)	10993(3867)	70
✓		2	15	147254	20.65	91.65	1125(863)	9634(3464)	70
✓		3	10	100127	19.65	92.16	1358(997)	10992(3767)	79
	✓	2	10	140855	14.52	93.31	148(138)	2031(1204)	75
	✓	2	5	248296	33.66	92.69	1028(839)	9147(4100)	85
	✓	2	15	253176	32.98	92.56	778(647)	7087(3422)	85
	✓	3	10	174744	30.66	92.79	1149(932)	9376(4173)	88

Supplemental Table 5: **Analysis of parameter influence for Trans-ABySS** for the human CD4 (H) and mouse C2C12 (M) dataset. The parameter c (coverage cutoff) and n (number of required read pairs) were varied (default is $c = 2$, $n = 10$). Similar to Oases-M the range of k values is k_{MIN}=19 and k_{MAX}=35 for human and k_{MIN}=21 and k_{MAX}=35 for mouse. N-Sens and N-Spec denote nucleotide sensitivity and specificity. The 5th and 6th column show the number of reconstructed Ensembl transcripts to full length or at least 80%, respectively. The number in brackets denotes the corresponding genes. The last column denotes the percentage of transfrags that aligned to the genome with at least 95% of length. The default parameters perform bad on the mouse dataset.

Trans-ABySS parameter optimization

In order to allow the reader to analyze the performance of the different parameters tested for trans-ABySS we show the assembly results in Supplementary Table 5. We analysed the influence of parameter c (coverage cutoff) and n (number of required read pairs) for Trans-ABySS for the human CD4 (H) and mouse C2C12 (M) dataset. For a fair comparison, Trans-AbySS uses the same range of k values as Oases-M, with k_{MIN}=19 and k_{MAX}=35 for human and k_{MIN}=21 and k_{MAX}=35 for mouse. For both datasets, the parameter set ($c = 3$ and $n = 10$) performed best and assembled the greatest number of full length transcripts, and was therefore used in the comparison. We found that the default parametrization ($c = 2$ and $n = 10$) recommended by the authors performed generally worse, especially for the mouse dataset. Note that for each parameter set all single k assemblies from $k_{MIN}, \ldots, k_{MAX}$ were computed with ABySS and given to the Trans-ABySS module to create the merged assembly.

H	M	assembler	transfrags (≥100bps)	95% aligned
✓		Oases-M	174469	79
✓		Trans-ABySS	100127	79
✓		Trinity	76232	84
	✓	Oases-M	175914	83
	✓	Trans-ABySS	174744	88
	✓	Trinity	92810	87

Supplemental Table 6: **Analysis of misassemblies** for the human CD4 (H) and mouse C2C12 (M) dataset and the *de novo* assemblers quantified by the percentage of transfrags (4th column) that map to at least 95% of their length to the genome (last column).

H	M	k	transfrags (≥100bps)	95% aligned
✓		19	67319	81
✓		25	53504	85
✓		29	50936	90
✓		35	34012	90
	✓	21	57095	86
	✓	25	56473	88
	✓	29	59503	88
	✓	35	61939	90

Supplemental Table 7: **Analysis of misassemblies** for the human CD4 (H) and mouse C2C12 (M) dataset and single k assemblies by Oases quantified by the percentage of transfrags (4th column) that map to at least 95% of their length to the genome (last column).

Misassemblies in *de novo* transcriptome assemblies

In Supplementary Table 6 the percentage of transfrags that are potentially misassemblies, because they do not align to at least 95% of their length to the genome, are compared between the different *de novo* assemblers. Further in Supplementary Table 7 we show that the misassembly rate is closely related to the k value used for the single k assemblies and for longer reads misassemblies may be avoided by using a higher k_{MIN} value for the Merged assembly.

References

[1] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differen-

tial expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11:94, 2010.

[2] M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, Jun 1996.

[3] Christopher Lee. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics*, 19(8):999–1008, May 2003.

[4] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*, May 2008.

[5] David Weese, Anne-Katrin Emde, Tobias Rausch, Andreas Döring, and Knut Reinert. RazerS–fast read mapping with sensitivity control. *Genome Res*, 19(9):1646–1654, Sep 2009.