Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-seq reads

Martin et al. BMC Genomics 2010

Journal club 24.01.2011

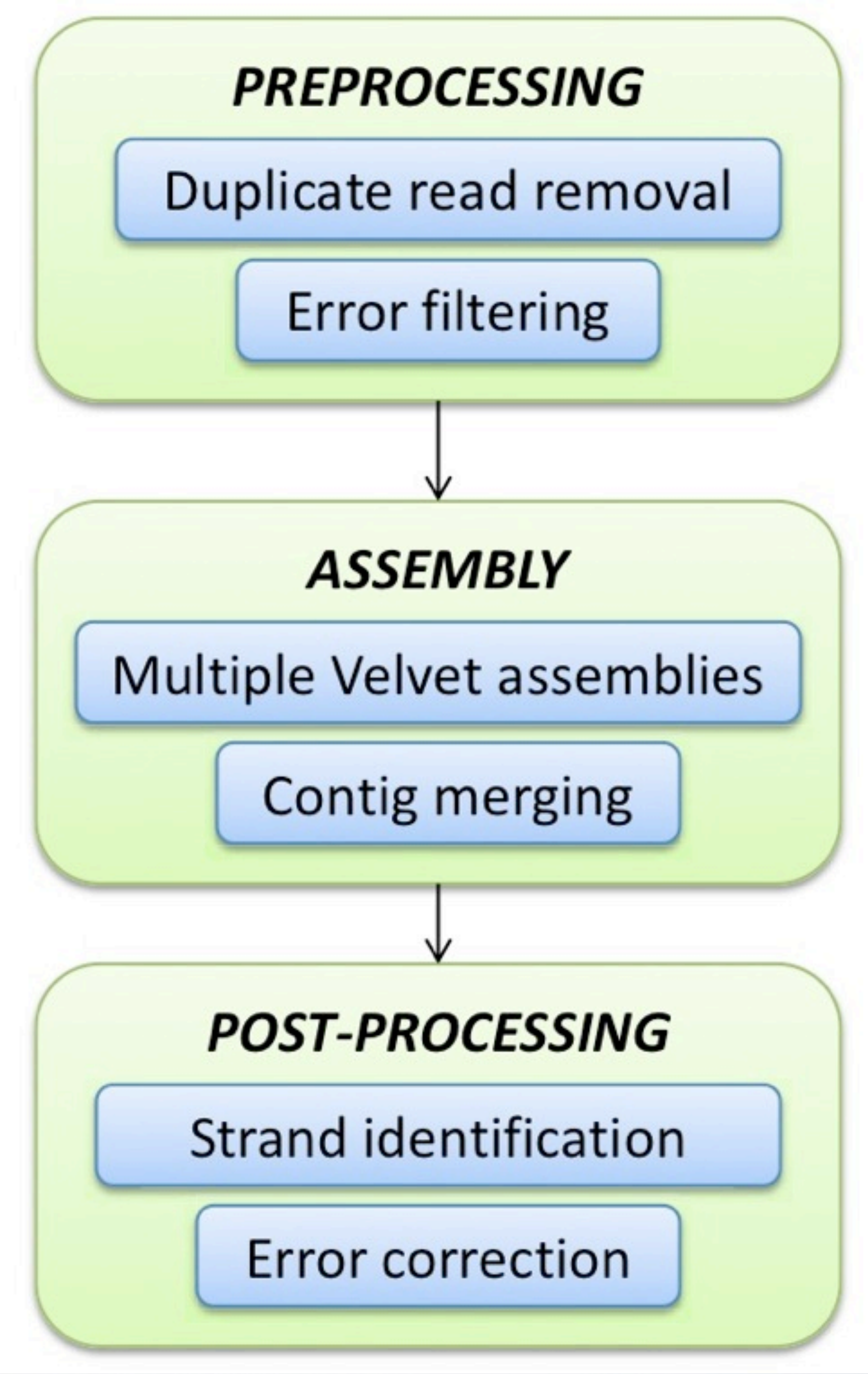
RNA-Seq data analysis - aligning short reads to a reference genome

- TopHat/Cufflinks. TopHat is a fast splice junction mapper for RNA-Seq reads. Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples.
- ERANGE
- Scripture a method for ab initio transcriptome reconstruction from RNA-Seq data

De novo assembly of RNA-Seq reads

- Artifacts from library preparation and sequencing errors
- Very large data set
- Sequencing coverage among transcripts very different

Rnnotator assembly pipeline



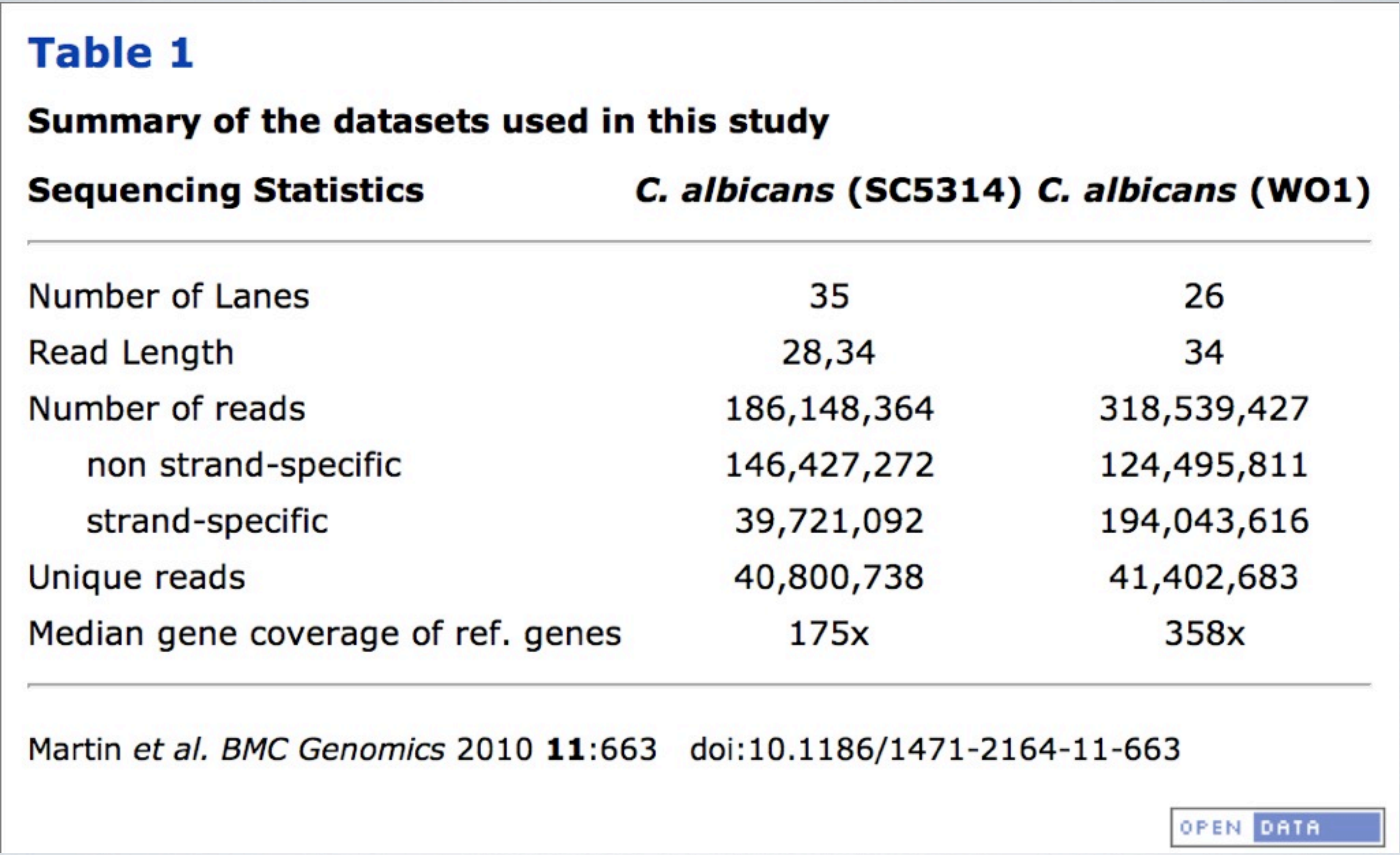
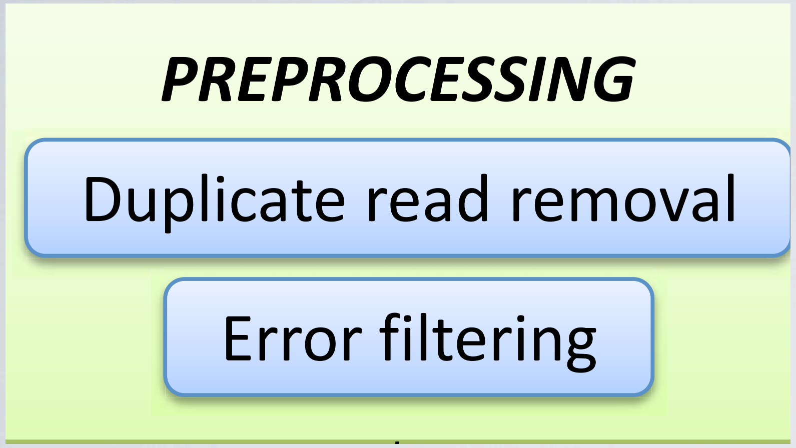


Table 1

Summary of the datasets used in this study

Sequencing Statistics	***C. albicans* (SC5314)**	***C. albicans* (WO1)**
Number of Lanes	35	26
Read Length	28,34	34
Number of reads	186,148,364	318,539,427
non strand-specific	146,427,272	124,495,811
strand-specific	39,721,092	194,043,616
Unique reads	40,800,738	41,402,683
Median gene coverage of ref. genes	175x	358x

Martin *et al. BMC Genomics* 2010 **11**:663 doi:10.1186/1471-2164-11-663

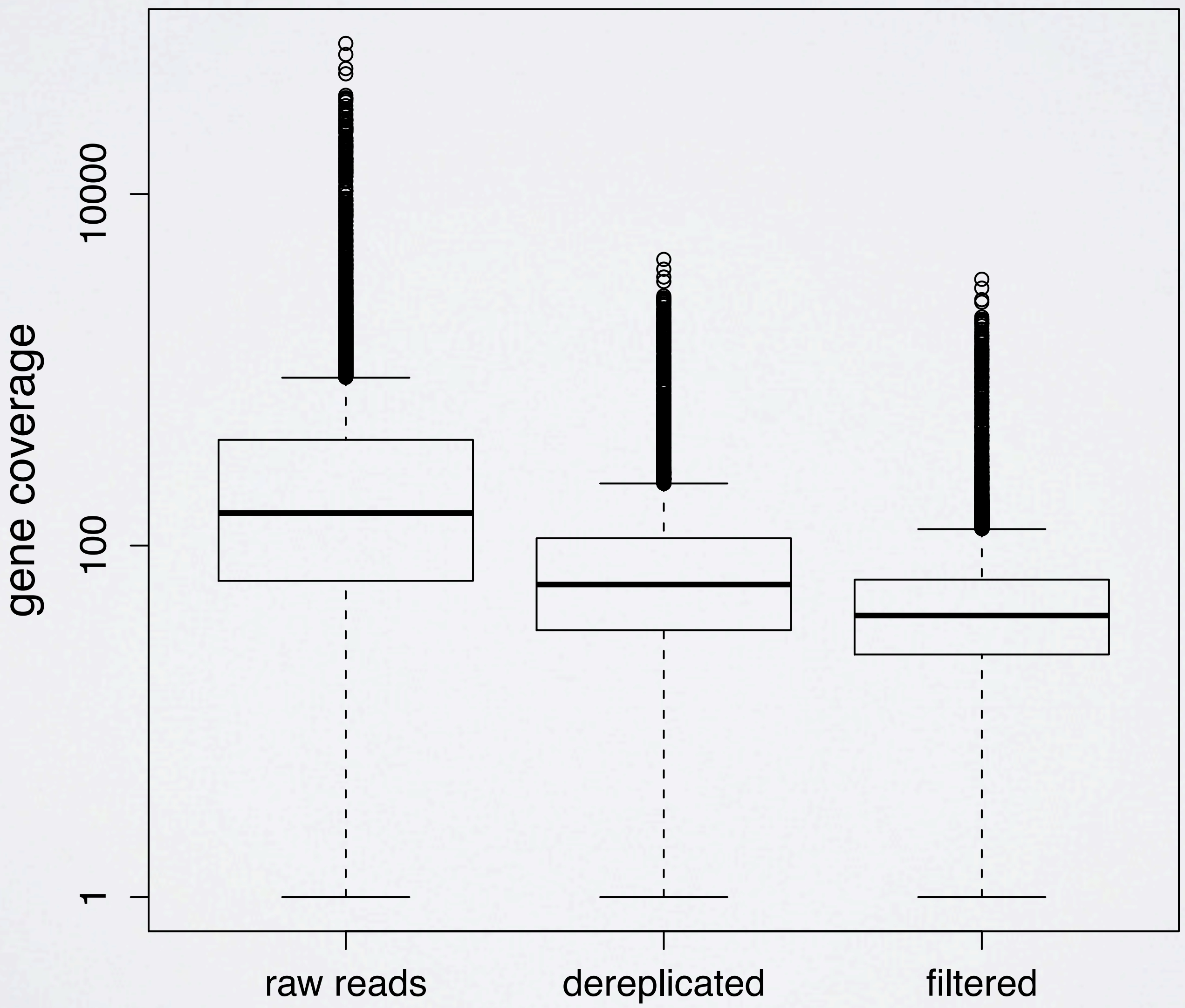


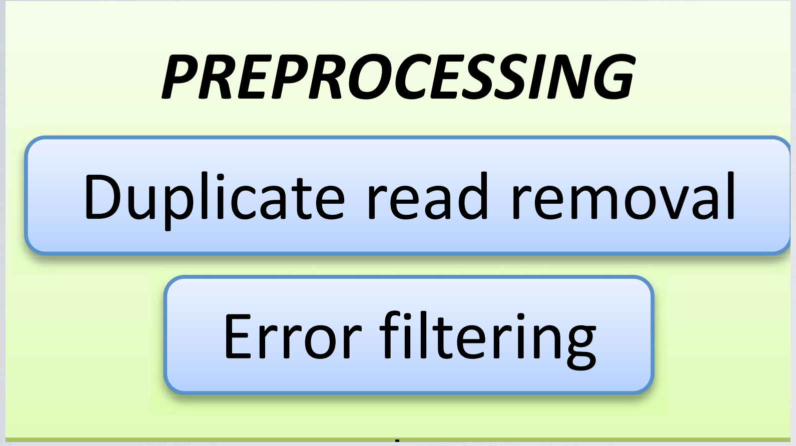
PREPROCESSING

Duplicate read removal

Error filtering

Removal of **identical** reads (dereplication)





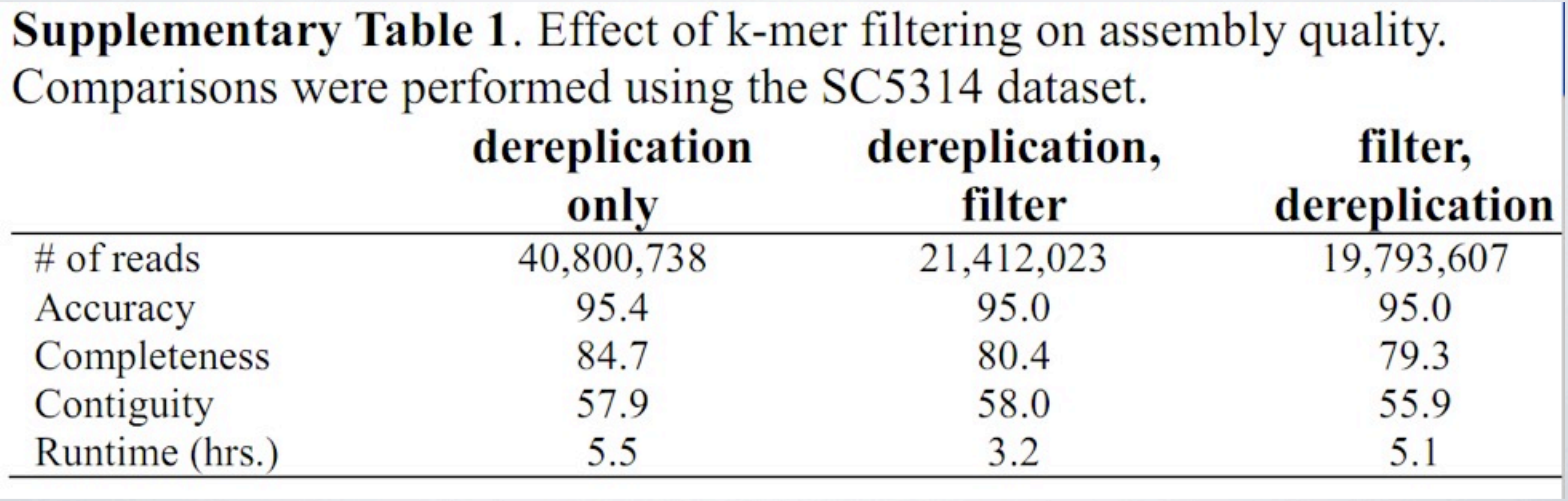
PREPROCESSING

Duplicate read removal

Error filtering

Removal of low quality reads containing sequencing errors using **rare k-mer filtering** approach.

- Frequency of each k-mer was calculated
- Rare k-mers that occurred less than three times in the set of unique reads were not used in the assembly



Supplementary Table 1. Effect of k-mer filtering on assembly quality. Comparisons were performed using the SC5314 dataset.

	dereplication only	**dereplication, filter**	**filter, dereplication**
# of reads	40,800,738	21,412,023	19,793,607
Accuracy	95.4	95.0	95.0
Completeness	84.7	80.4	79.3
Contiguity	57.9	58.0	55.9
Runtime (hrs.)	5.5	3.2	5.1

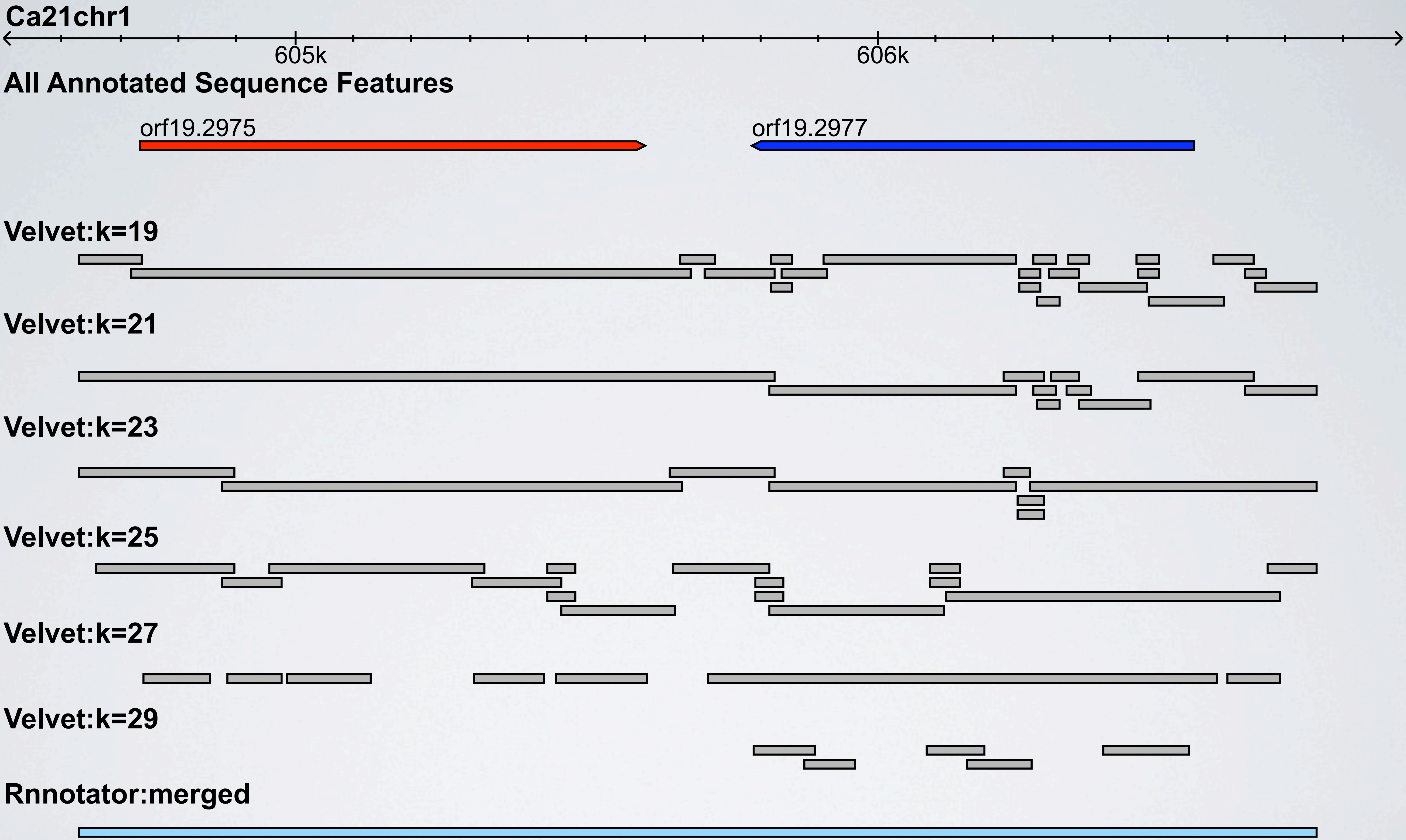
ASSEMBLY

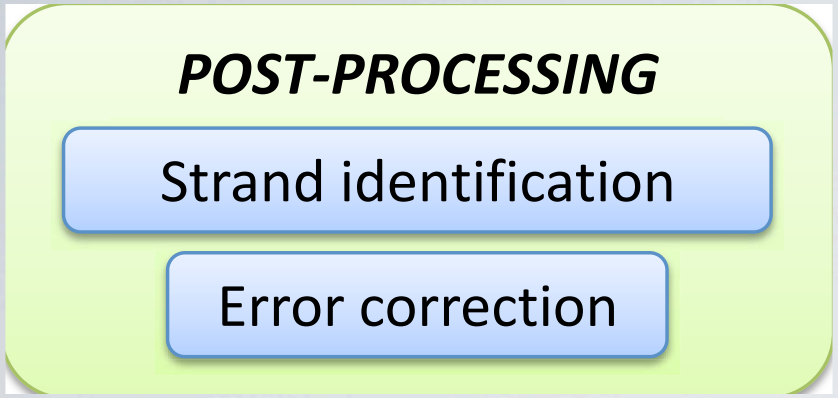
Multiple Velvet assemblies

Contig merging

- No single parameter set can give best results
- Multiple velvet assemblies were done (8 velveth + 8 velvetg)
- Resulting contigs were merged with Minimus2 assembler from AMOS package

A


Ca21chr1
605k
606k
All Annotated Sequence Features
orf19.2975
orf19.2977
Velvet:k=19
Velvet:k=21
Velvet:k=23
Velvet:k=25
Velvet:k=27
Velvet:k=29
Rnnotator:merged

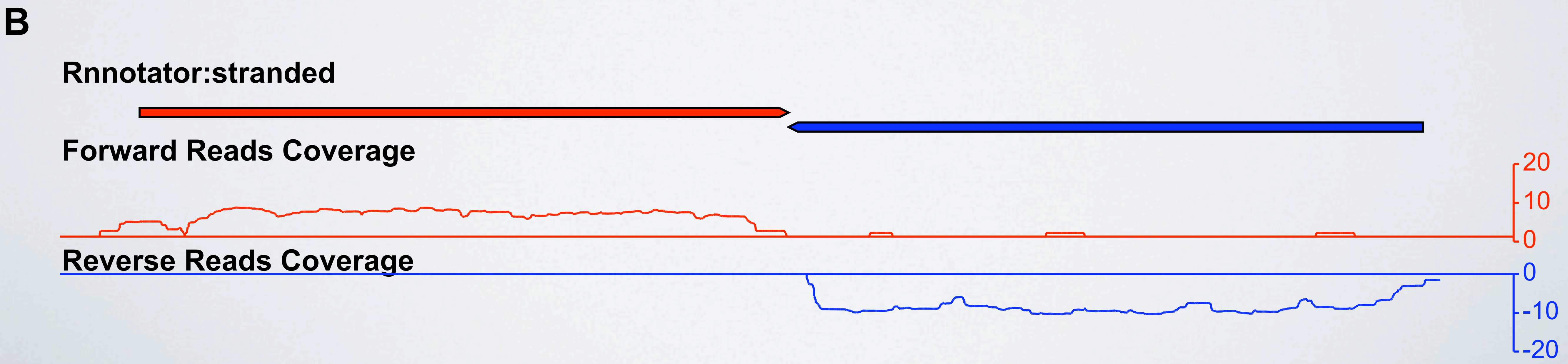


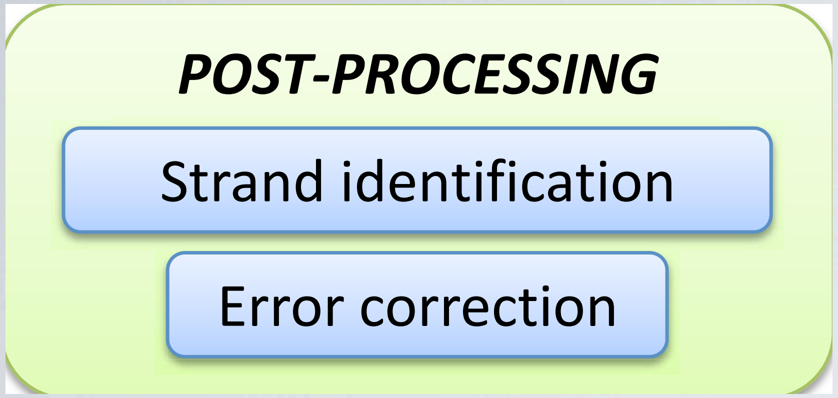
POST-PROCESSING

Strand identification

Error correction

- Special consideration of the direction of transcription
- strand-specific RNA-Seq reads were aligned to each contig and then the contigs were split at the strandness transition point





- Single base errors in the assembled contigs were corrected by aligning the reads back to each contig to generate consensus nucleotide sequence

Evaluation of Rnnotator's performance

- **Accuracy** - correctness of the assembly estimated by aligning each contig to the reference genome

- **Completeness** - degree to which the transcriptome is covered by assembled contigs. Estimated by calculating the percentage of genes in the annotated gene catalog that are covered at > 80% of the gene length.

- **Contiguity** - likelihood that a full-length transcript is represented as a single contig. Calculating the percentage of complete genes covered by a single contig to > 80% of the gene length

- **Gene fusions** - the number of contigs which contain two genes assembled into a single contig.

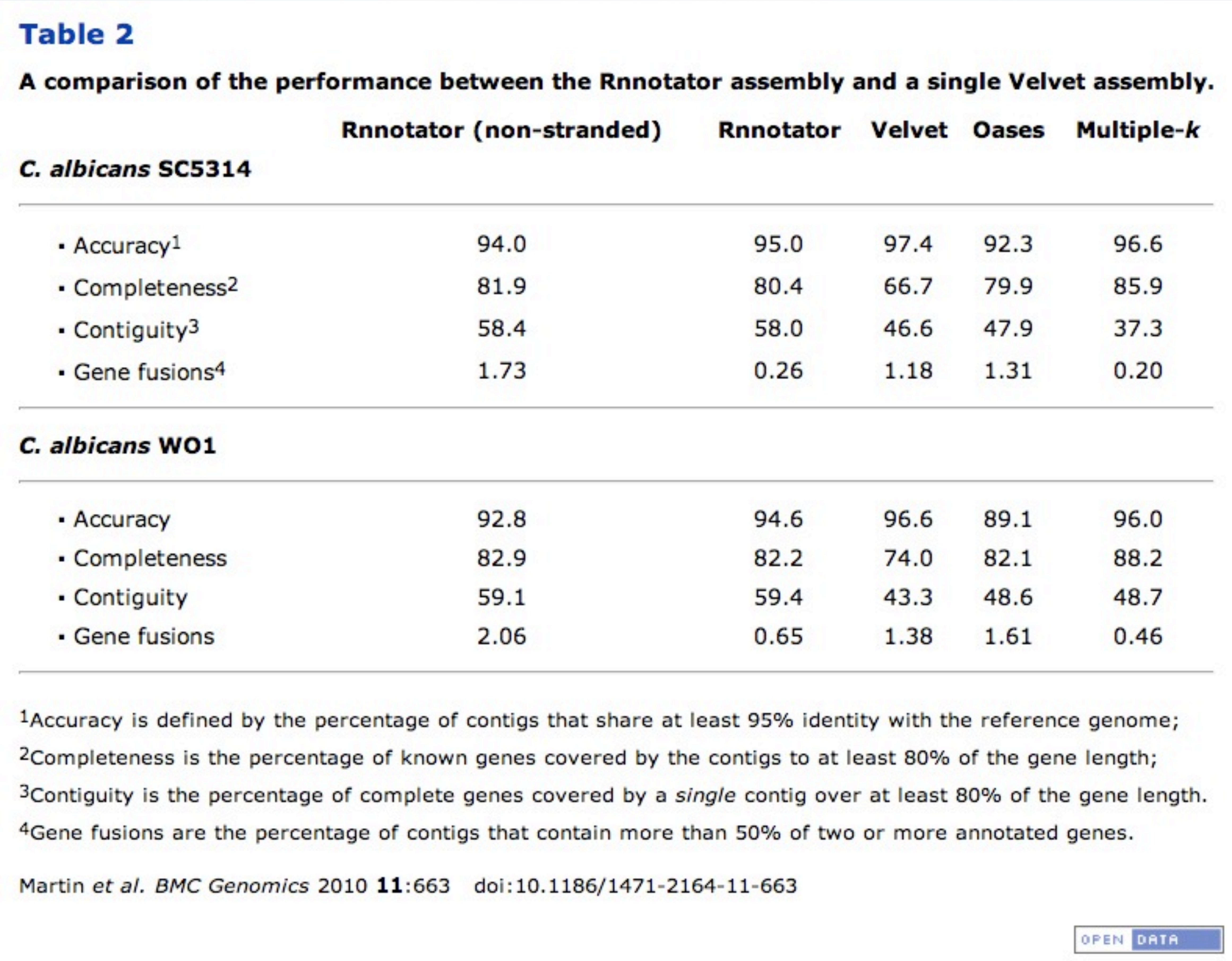


Table 2

A comparison of the performance between the Rnnotator assembly and a single Velvet assembly.

	Rnnotator (non-stranded)	**Rnnotator**	**Velvet**	**Oases**	**Multiple-*k***
***C. albicans* SC5314**					
▪ Accuracy[1]	94.0	95.0	97.4	92.3	96.6
▪ Completeness[2]	81.9	80.4	66.7	79.9	85.9
▪ Contiguity[3]	58.4	58.0	46.6	47.9	37.3
▪ Gene fusions[4]	1.73	0.26	1.18	1.31	0.20
***C. albicans* WO1**					
▪ Accuracy	92.8	94.6	96.6	89.1	96.0
▪ Completeness	82.9	82.2	74.0	82.1	88.2
▪ Contiguity	59.1	59.4	43.3	48.6	48.7
▪ Gene fusions	2.06	0.65	1.38	1.61	0.46

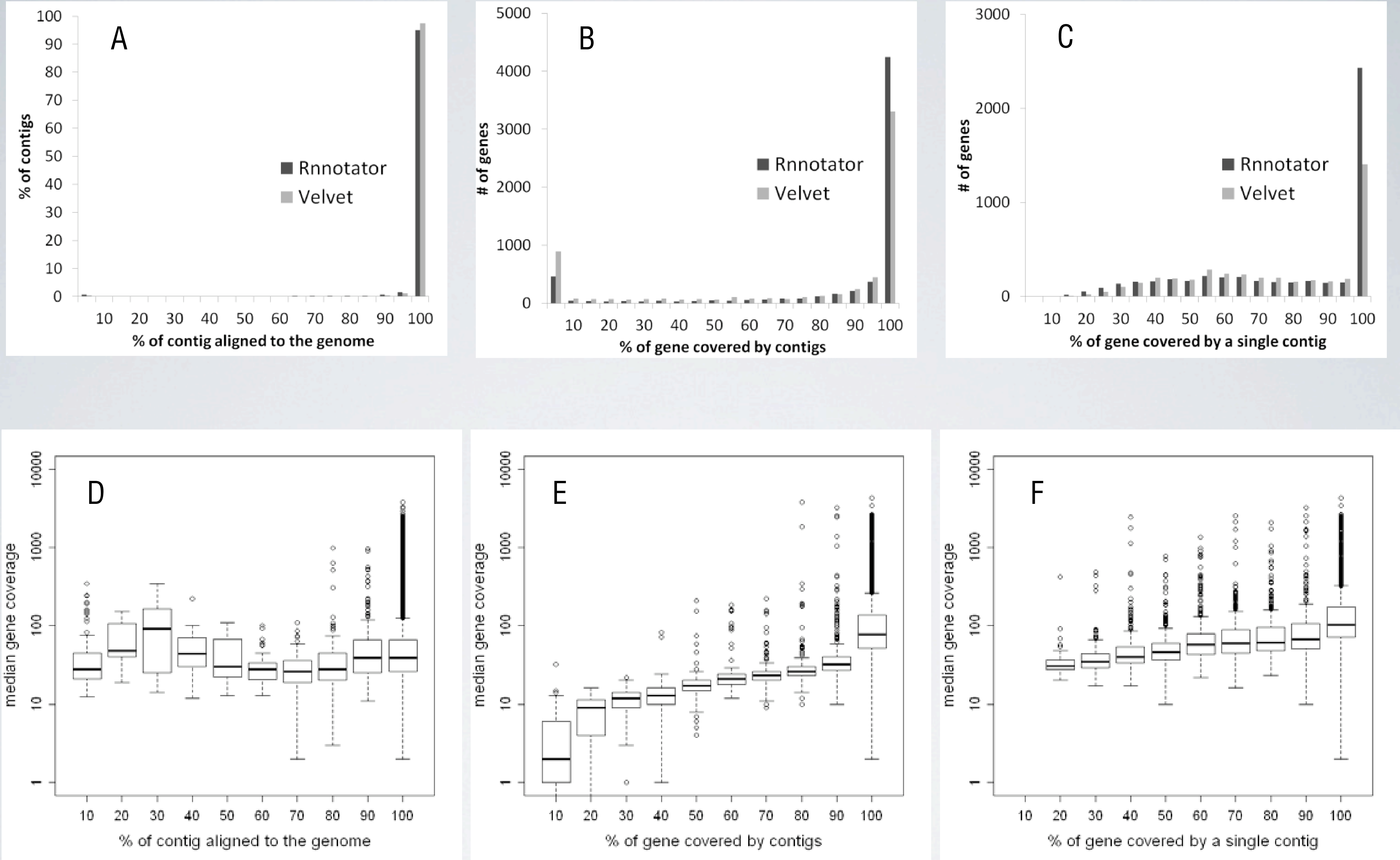
[1] Accuracy is defined by the percentage of contigs that share at least 95% identity with the reference genome;

[2] Completeness is the percentage of known genes covered by the contigs to at least 80% of the gene length;

[3] Contiguity is the percentage of complete genes covered by a *single* contig over at least 80% of the gene length.

[4] Gene fusions are the percentage of contigs that contain more than 50% of two or more annotated genes.

Martin *et al. BMC Genomics* 2010 **11**:663 doi:10.1186/1471-2164-11-663



Accuracy, completeness, and contiguity of assembled transcripts for *Candida albicans* SC5314 are shown in panels (A,D), (B,E), and (C,F), respectively. For contiguity only genes with > 80% completeness are shown. In panels D), E), and F) a box plot of median gene coverage by unique reads is shown for genes falling into each bin. Open circles above each boxplot depict outliers in the coverage distribution.

Martin *et al. BMC Genomics* 2010 **11**:663 doi:10.1186/1471-2164-11-663

http://www.scivee.tv/node/19174