VARUS: Drawing diverse samples from RNA-Seq libraries

Willy Bruhn[1,2], Katharina Hoff[2], Mario Stanke[2]

1) Institute for Medical Biochemistry and Molecular Biology, University Medicine Greifswald, Greifswald, GERMANY
2) Institute for Mathematics and Computer Science, University of Greifswald, Greifswald, GERMANY
Contact: pb133777@uni-greifswald.de, katharina.hoff@uni-greifswald.de

Introduction

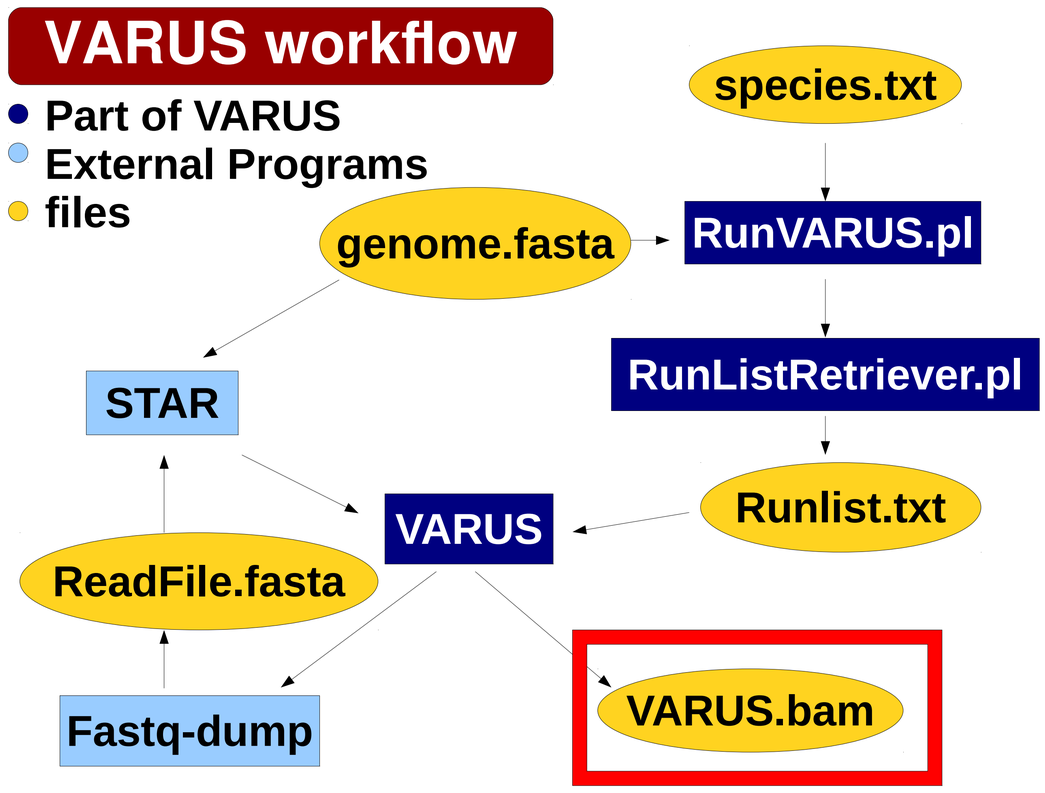
For many tasks involving RNA-Seq data it is desirable that a large fraction of the transcriptome is **sufficiently** covered, rather than a high total coverage that is mainly achieved through a small fraction of very highly expressed transcripts. On the other, hand many freely available RNA-Seq runs, e.g. from the Sequence Read Archive (SRA), have similar expression patterns, so that a naive sampling of runs results in a less complementary coverage of transcripts than possible. This circumstance and the fact that using all available data is often not an option because of its sheer amount, suggested the development of **VARUS**.

VARUS is a software-tool that automates the selection and download of RNA-Seq runs available at the SRA, with regard to a sufficiently high coverage. This is done in a stepwise procedure. An iteration includes:

- selecting a run to download, that is expected to complement previously downloaded reads the most
- download the run with **fastq-dump**
- align the reads with **STAR**
- evaluate the alignment

The key here is that runs are only downloaded partially in each iteration. With these read-samples, estimations of the runs value for further downloads from this run are made. This allows **VARUS** to distinguish good runs from bad runs within the first few downloads, and download more extensively from runs that are likely to contain more reads from yet underrepresented transcripts. **VARUS** is freely available at **https://github.com/WillyBruhn/VARUS.git**.

VARUS workflow



Running VARUS

Input: - *species.txt* containing genus and species name and genome.fasta, the corresponding genome. Format example:
`Schizosaccharomyces pombe; genome.fasta`

Output: - *VARUS.bam*, resulting RNA-Seq alignment-file

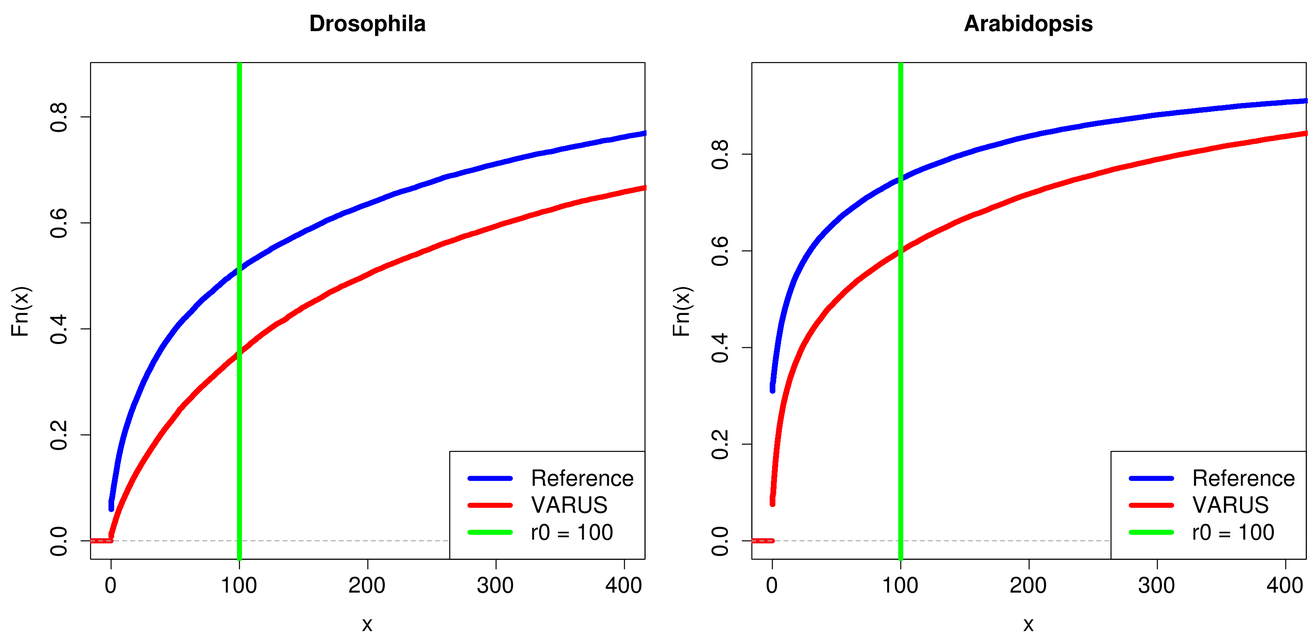
Call: `runVarus.pl`
(assuming that *species.txt* is located in `cwd`)

Summary

- **VARUS** downloads RNA-Seq runs from the **SRA** such that a large fraction of the transcriptome is sufficiently covered.
- Tests with **Kallisto** showed that the coverage could be improved for the species *D. melanogaster* and *A. thaliana* used in (1).
- One application scenario is using **VARUS** generated RNA-Seq alignment files for gene prediction with **BRAKER**.

Freely available at **https://github.com/WillyBruhn/VARUS.git**

Kallisto



X: number of reads
Fn(x): portion of the transcripts that have x or less reads mapping to them

Plots show the empiric distribution of the estimated read counts per transcript from aligning VARUS retrieved reads against the annotated transcriptome with **Kallisto**. Reference refers to STAR alignments created with “hand picked” libraries from (1).

Drosophila

	Reference	VARUS
Fn(0)	0.0594	0.0084
Fn(100)	0.5126	0.3549

Arabidopsis

	Reference	VARUS
Fn(0)	0.3099	0.0753
Fn(100)	0.7484	0.6000

The portion of the transcripts that have less than 100 reads mapping to them is smaller for the reads downloaded with **VARUS.** In other words: the reads downloaded with **VARUS** are more evenly distributed among the transcriptome and hence suggest that **VARUS** could be usefull for retrieving input-reads for RNA-Seq incorporating gene prediction tools.

Accuracy of BRAKER with VARUS

	A. thaliana		*D. melanogaster*	
	Reference	**VARUS**	**Reference**	**VARUS**
Gene F1	0.60	0.52	0.64	0.58
Exon F1	0.83	0.79	0.79	0.75

Possible current issues:
- setting sampling depth
- quality of libraries in SRA

We expect improvements.

References

[1] Katharina J. Hoff, Simone Lange, Alexandre Lomsadze, Mark Borodovsky, Mario Stanke; BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS, Bioinformatics, Volume 32, Issue 5, 1 March 2016, Pages 767–769
[2] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1):15–21, 2013
[3] Mario Stanke, Mark Diekhans, Robert Baertsch, and David Haussler. Using Native and Syntenically Mapped Cdna Alignments to Improve De Novo Gene Finding. Bioinformatics, 24(5):637–644, 2008.

Funding
The financial support by the Deutsche Forschungsgemeinschaft (SPP1710 Li 984/3-2; GRK1974 - A1; SPP 1927 Li984/4-1) is gratefully acknowledged.

Many thanks to the Group '**Redox control of cell function**' for the support

AG Lillig
redox control of cell function