Gene-centered, condensed & time-efficient visualization of mRNA Seq Data

Matthias Munz, Dr. German Leparc & Dr. Eric Simon

Target Discovery Research Germany, Boehringer Ingelheim Pharma GmbH & Co KG

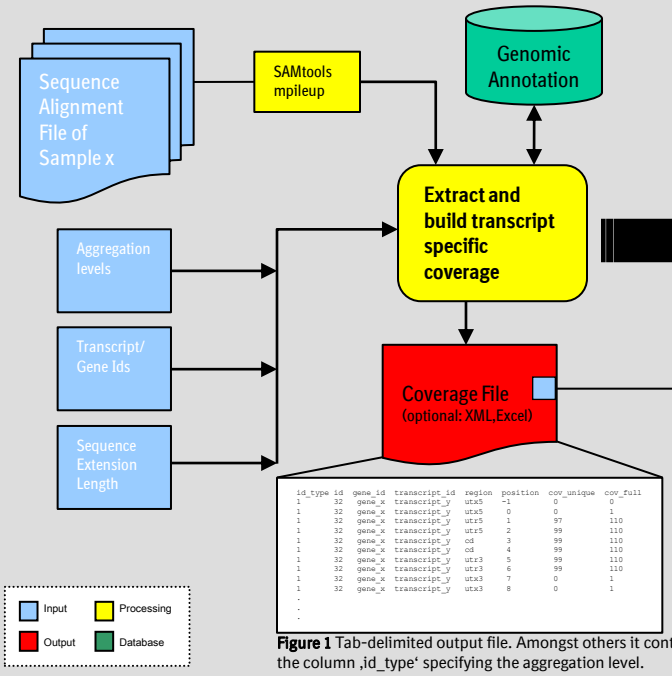
Abstract

The analysis of transcriptomes is important to understand the molecular mechanisms of diseases. mRNA-seq represents an unbiased genome-wide approach to study the expression and isoform distribution of genes in normal and diseased tissuesHowever, there is currently a lack of powerful tools to compute and visualize the coverage of unique and ambiguous reads in a time-efficient gene-centered manner. Many genes can have multiple large introns which are cumbersome to visualize with existing tools like Integrative Genomics Viewer, UCSC Genome Browser or Gbrowse. Also, the aggregation of reads from replicates or other samples is not possible with these tools.

Therefore a novel workflow was implemented to generate visualizations on-the-fly and in reasonable time. Reads were mapped to the genome using the STAR* Aligner which allows for spliced alignment of mRNA-seq reads. For a given input gene, the corresponding transcripts positions are fetched from a local Ensembl database. The genomic coverage is then assembled using SAMtools* mpileup from a single or multiple BAM files, transformed to absolute transcript positions and written into a flat file which is then used for the visualization with ggplot2*.

Our approach displays the transcript specific coverage as a stack of both unique and ambiguous alignments which allows the user to focus on exonic gene regions. Each step in the workflow is parallelized. Pooling of multiple BAM files enables the user to aggregate the coverage on different levels. This allows a flexible extension of the sequencing depth for low abundant genes like GPCRs. An additional feature is given by the optional extension of the transcript start and stop position which allows UTR refinement.

Workflow



id_type	id	gene_id	transcript_id	region	position	cov_unique	cov_full
1	32	gene_x	transcript_y	utx5	-1	0	0
1	32	gene_x	transcript_y	utx5	0	0	1
1	32	gene_x	transcript_y	utr5	1	97	110
1	32	gene_x	transcript_y	utr5	2	99	110
1	32	gene_x	transcript_y	cd	3	99	110
1	32	gene_x	transcript_y	cd	4	99	110
1	32	gene_x	transcript_y	utr3	5	99	110
1	32	gene_x	transcript_y	utr3	6	99	110
1	32	gene_x	transcript_y	utx3	7	0	1
1	32	gene_x	transcript_y	utx3	8	0	1

Figure 1 Tab-delimited output file. Amongst others it contains the column ‚id_type' specifying the aggregation level.

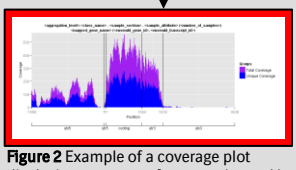


Figure 2 Example of a coverage plot displaying coverage of a transcript and its flanking sequence.

Aggregation Levels

Coverage of samples that semantically belong together can be summed up on three more levels besides the sample level itself.

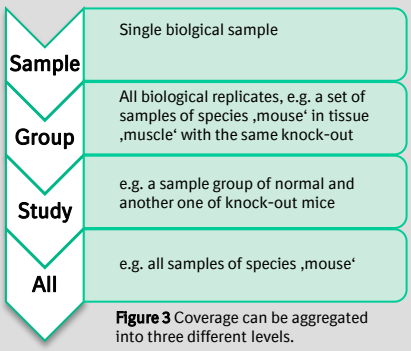


Figure 3 Coverage can be aggregated into three different levels.

Example of Use

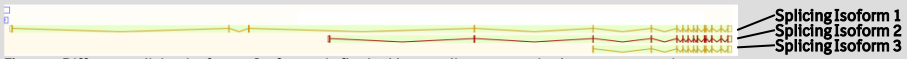
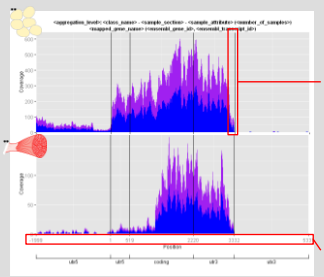


Figure 4 Different splicing isoforms. Isoform 1 is flanked by an adjacent gene in the upstream region.



Vertical lines represent the extended 5'UTR (utx5), the annotated 5'UTR (utr5), the coding region (coding), the 3'UTR (utr3) and the extended 3'UTR (utx3), respectively (see corresponding labels below the plot).

Positions are given relatively to the transcription start site

Figure 5 Coverage of **Splicing Isoform 1** in **adipose tissue (top)** and **muscle**. Coverage in the 5' flanking sequence results from an adjacent gene. Due to low coverage in 5' utr and in parts of the coding sequence in muscle, it can be assumed that an isoform other than 1 is getting expressed there.

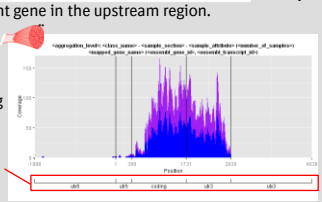


Figure 6 No coverage over the full length of **Splicing Isoform 2** in muscle. This indicates the expression of another isoform.

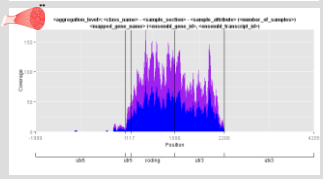


Figure 7 Coverage over the full length of **Splicing Isoform 3** in muscle. Isoform 3 might be expressed here instead of isoform 1.

Performance

The performance measures in table 1 and 2 were derived using the following configurations and environment:

Input Data
- Aggregation level: ‚group'
- 56 mouse samples (11 groups)
- Mappings of 26 Mio. Reads per Sample (80% mapped uniquely, 16% mapped to multiple loci)
- 10 Transcripts of 7 genes with 57 exons in total
- 2000bp flanking sequence extension

Output Data
- 43 MB coverage file (725000 lines)
- 110 Plots

Test Environment
- Red Hat Enterprise Linux Server
- ´x86_64 Architecture
- 64 GB Main Memory
- 8 CPU Cores

SAMtools mpileup			Total Coverage Comp.		Plot Generation		Total Workflow	
Cores	Time (min/sec)		Time (min/sec)		Time (min/sec)		Time (min/sec)	
					6 Cores			
n	μ	σ	μ	σ	μ	σ	μ	σ
1	06:05	26	06:39	20	11:10	5	18:40	73
2	02:55	13	03:44	22	11:03	2	14:56	23
3	01:55	7	02:41	19	11:05	8	13:55	21
4	01:28	4	02:17	19	11:04	4	13:31	21
5	01:13	2	02:05	18	11:08	6	13:22	19
6	01:02	1	01:49	16	11:13	14	13:11	18
7	00:57	2	01:54	15	11:06	3	13:09	18

Table 1 The workflow is divided into two measurement units: the coverage computation and the plot generation. SAMtools* mpileup is part of the coverage computation. The option to additionally create XML or Excel files was disabled. Generating a **single plot** takes about **35 seconds**.

Total Coverage Comp.		Total	
Time (sec)		Time (sec)	
μ	σ	μ	σ
23	6	30	5

Table 2 The measurements above were made using the optional **recycling mode**. This mode allows the re-use of previously computed SAMtools* mpileup data and generated plots. If this option is set, the work space is scanned and already existing data of that kind is selected from all previously computed data and taken for the computation.

*References:
SAMtools Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009). The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9
ggplot2 Wickham, Hadley (2010). ggplot2: Elegant Graphics for Data Analysis. Journal of Statistical Software, 35
STAR Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics

Source: openclipart.org