Mapping summary

454 Sequencing for UTR annotation

MGP

November 27, 2012

Preprocessing steps

- Small ($< 100nts$) reads were filtered out
- Adaptors were removed

```
5' RACE Adaptor:
CTAATACGACTCACTATAGGGCAAGCAGTGGTATCAACGCAGAGTACT

3' RACE Adaptor:
CTAATACGACTCACTATAGGGCAAGCAGTGGTATCAACGCAGAGTACGCGGG
```

- Low quality ends were trimmed; each read is trimmed starting from the end. Mean quality is calculated for each three nts and bases are progressively removed until that mean raises 15 (Sanger scale).

Raw and clean reads stats

RACE	Tissue	Raw	Reads > 100	% Reads > 100	After trim	% After trim
3'	Kidney	456237	285097	62,5	284430	99,8
	Lung	594709	427119	71,8	424774	99,5
	Liver	402829	255852	63,5	254991	99,7
	Spleen	681557	493471	72,4	492609	99,8
	Heart	668609	483714	72,3	482668	99,8
5'	Kidney	675810	560187	82,9	559124	99,8
	Lung	692302	592014	85,5	590803	99,8
	Liver	741865	612306	82,5	610329	99,7
	Spleen	695740	584436	84,0	584401	100,0
	Heart	629891	523549	83,1	523532	100,0

Mapping approach

- BLAT (v35): avoids hard clipping and is able to handle long gaps. Inchworm wrapper[1] was used to run BLAT and get SAM files.
  - Maximum intron length: 10000
  - Minimum percent identity: 85%
  - Reference fasta: 1000g v37
  - Number of hits per read considered: 1, the best one reported by BLAT.

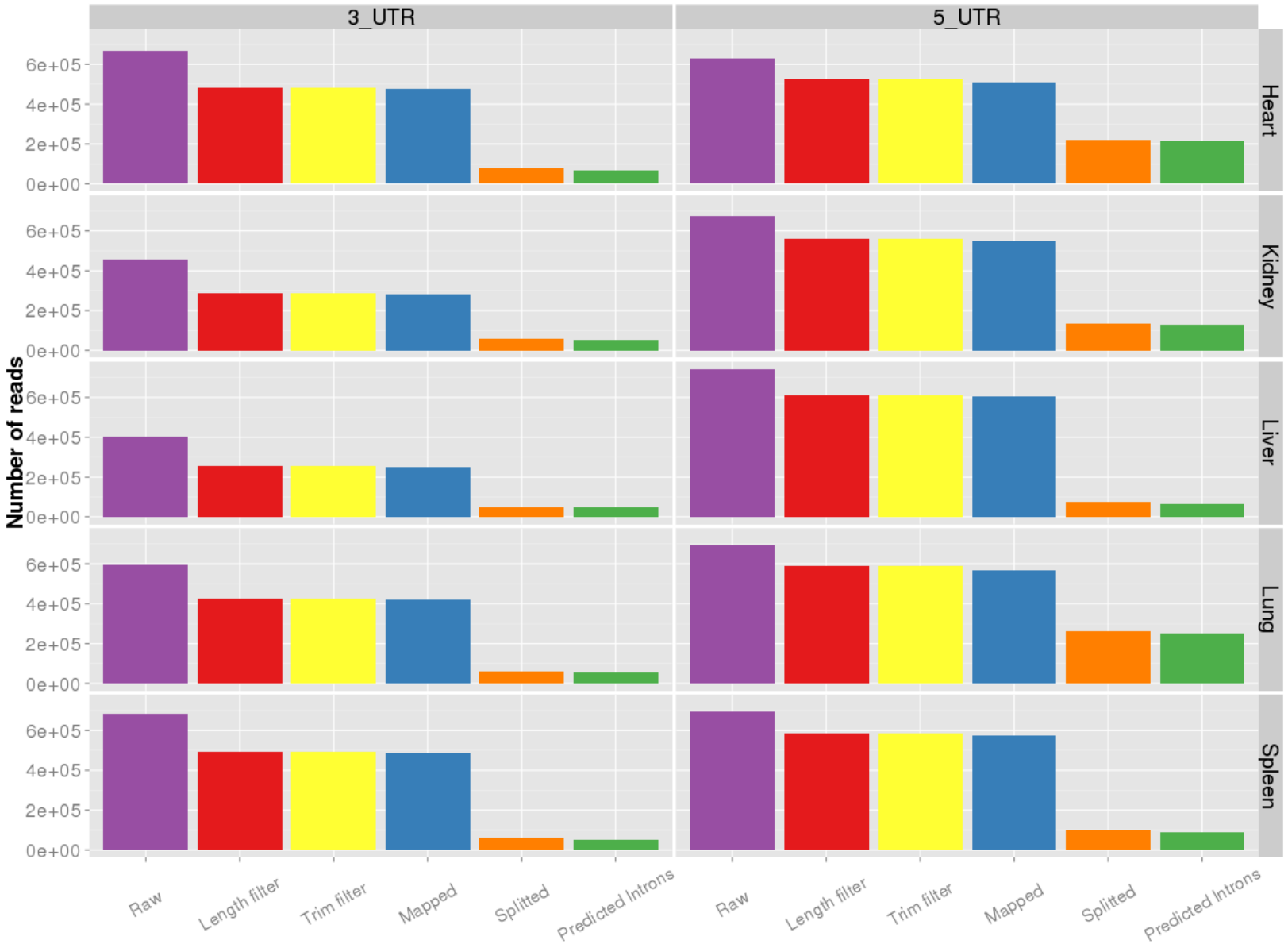
```
run_BLAT_shortReadPipeline.pl --single <fastq> --genome <100gv37fastafile> \
                              --seqType fq -o <outdir> -P 85
```

[1] Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol.* 2011 May 15;29(7):644-52

Mapped reads stats

RACE	Tissue	Mapped %	Splitted %	% Introns
3'	Kidney	98,6	21,5	19,9
	Lung	98,7	15,1	13,8
	Liver	98,5	20,3	19,3
	Spleen	98,6	12,6	11,1
	Heart	98,5	16,3	15,0
5'	Kidney	98,4	24,6	24,8
	Lung	96,3	46,5	45,8
	Liver	99,3	12,4	14,0
	Spleen	98,7	17,7	16,7
	Heart	97,5	42,7	42,7

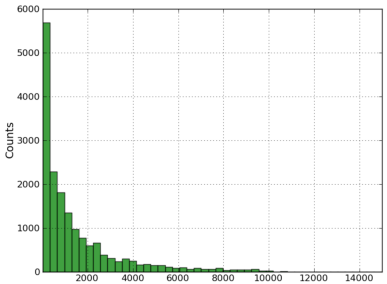
Mapped reads stats

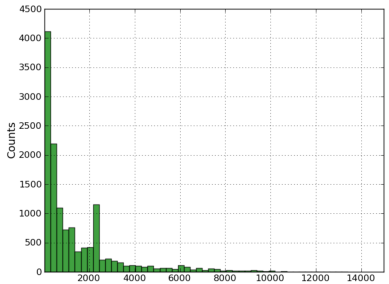


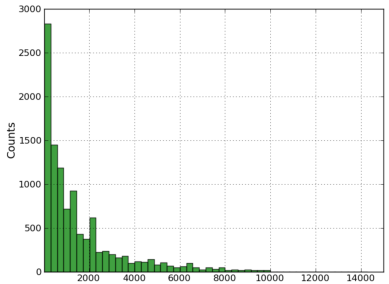
Mapped reads stats

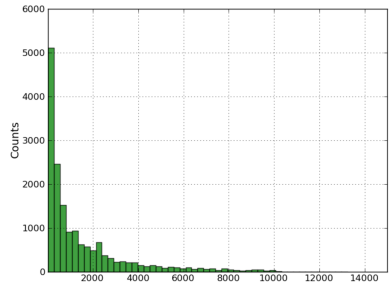
- High ratio of mapped reads
- A mapped read was considered to be splitted if a gap of size > 20 occurred in the alignment.
- Splitted reads represent an important proportion of the mapped reads.
- Intron prediction by Inchworm is based on the presence of splice site consensus seqs in the ends of the gaps (gap size > 20).

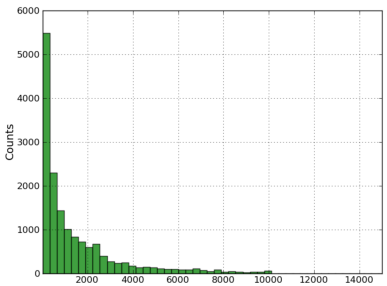
Distribution of the size of the longest gap by BLAT

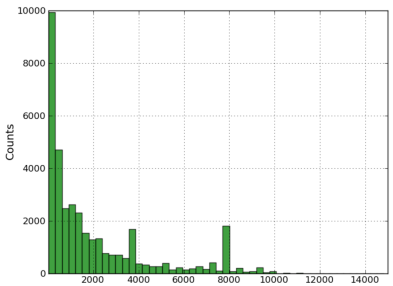


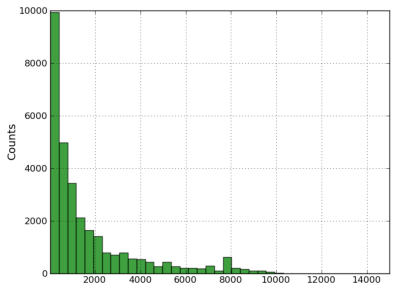


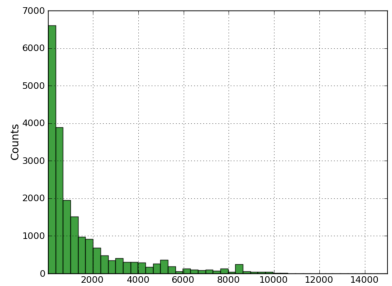


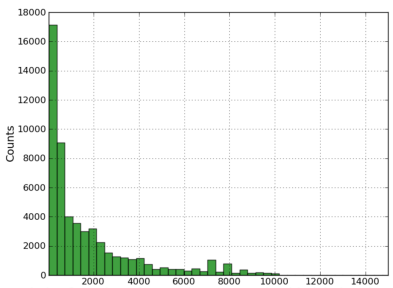


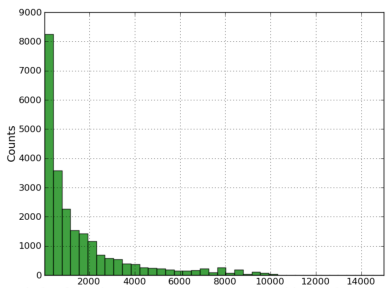




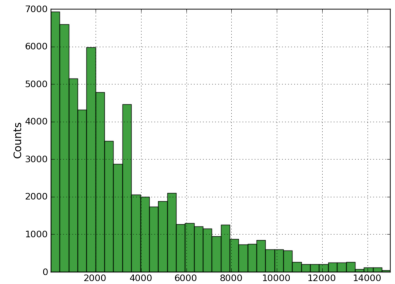




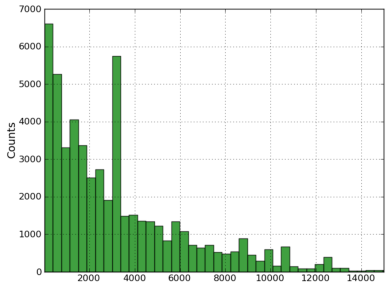




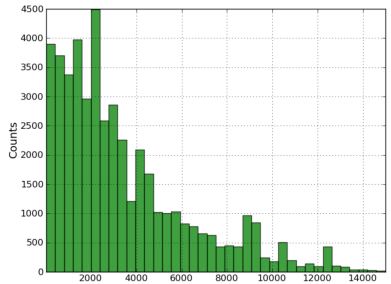
Distribution of sizes of the longest intron by Inchworm



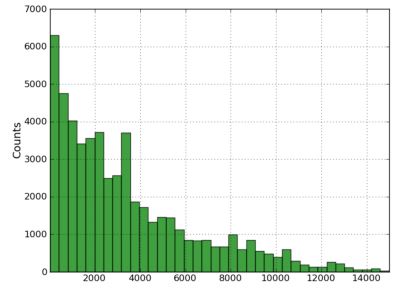
Skipped region from the reference (475300 reads with intron size < 20



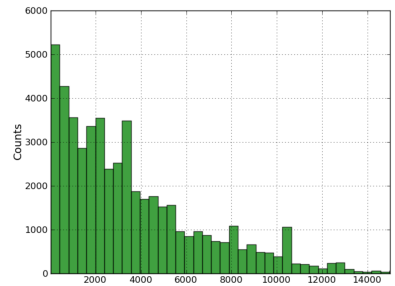
Skipped region from the reference (280580 reads with intron size < 20



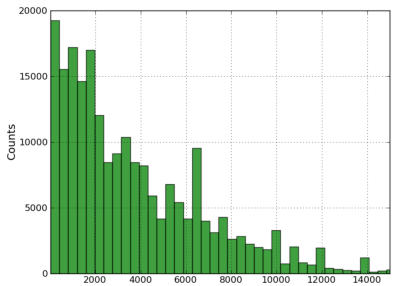
Skipped region from the reference (251186 reads with intron size < 20



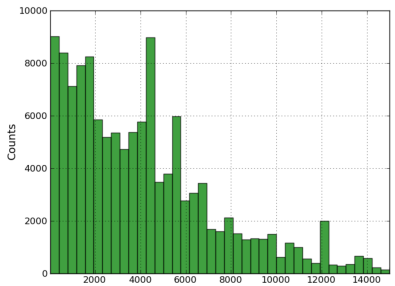
Skipped region from the reference (419112 reads with intron size < 20



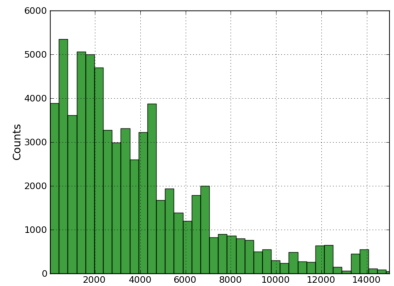
Skipped region from the reference (485630 reads with intron size < 20



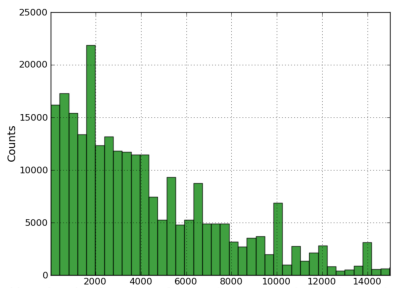
Skipped region from the reference (510661 reads with intron size < 20



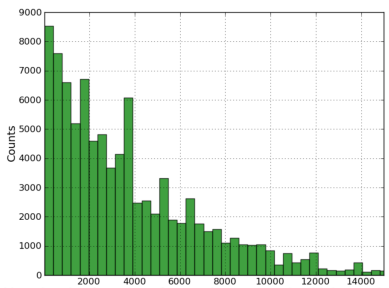
Skipped region from the reference (550008 reads with intron size < 20



Skipped region from the reference (605937 reads with intron size < 20



Skipped region from the reference (569063 reads with intron size < 20

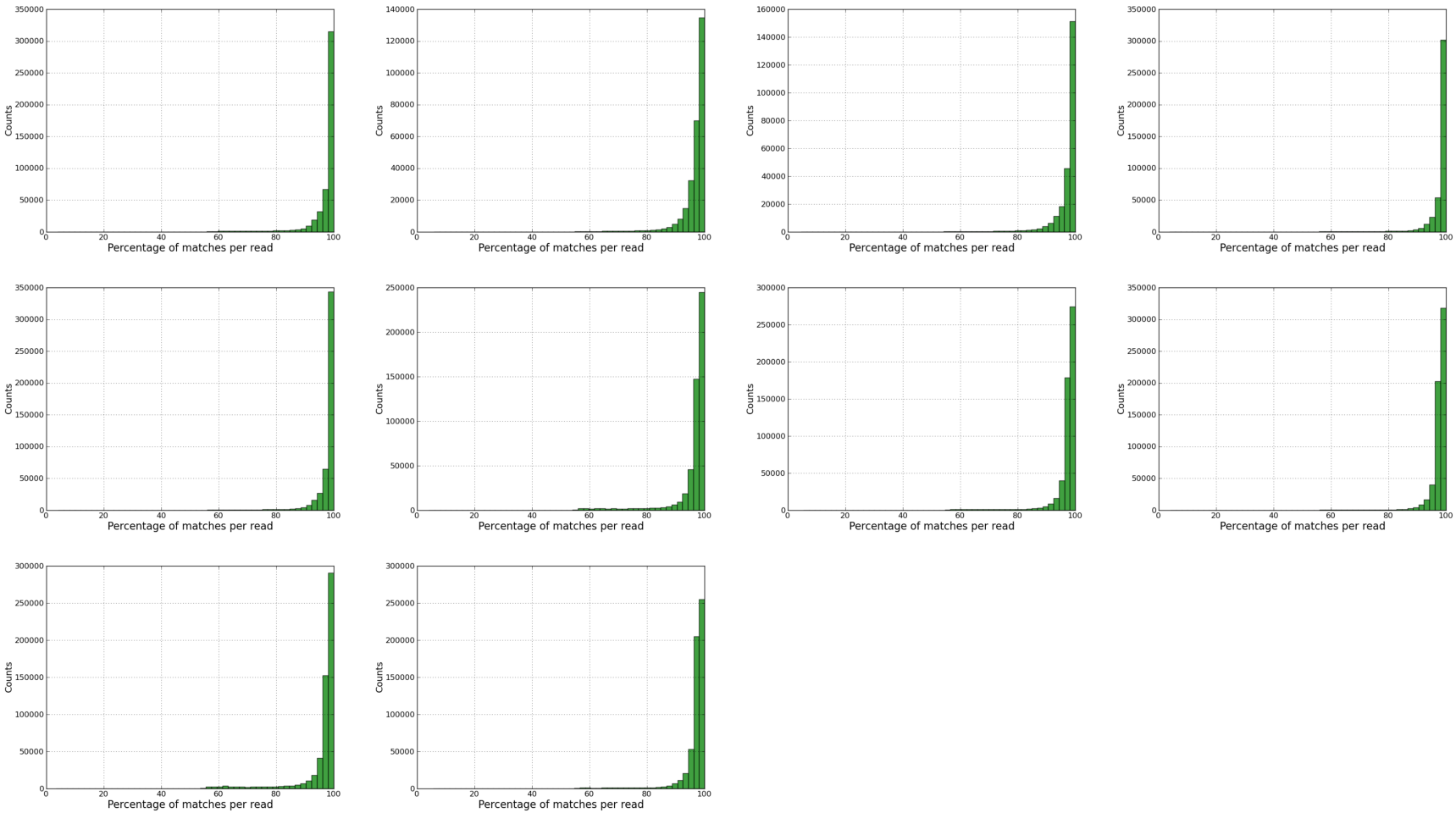


Skipped region from the reference (576617 reads with intron size < 20

Long gaps

- BLAT enables the identification of gaps of several thousands of nts.
- Parsing of the BLAT output by Inchworm produces predicted introns (gaps with splice site consensus seqs in the ends) and gaps without any splice site consensus seqs.

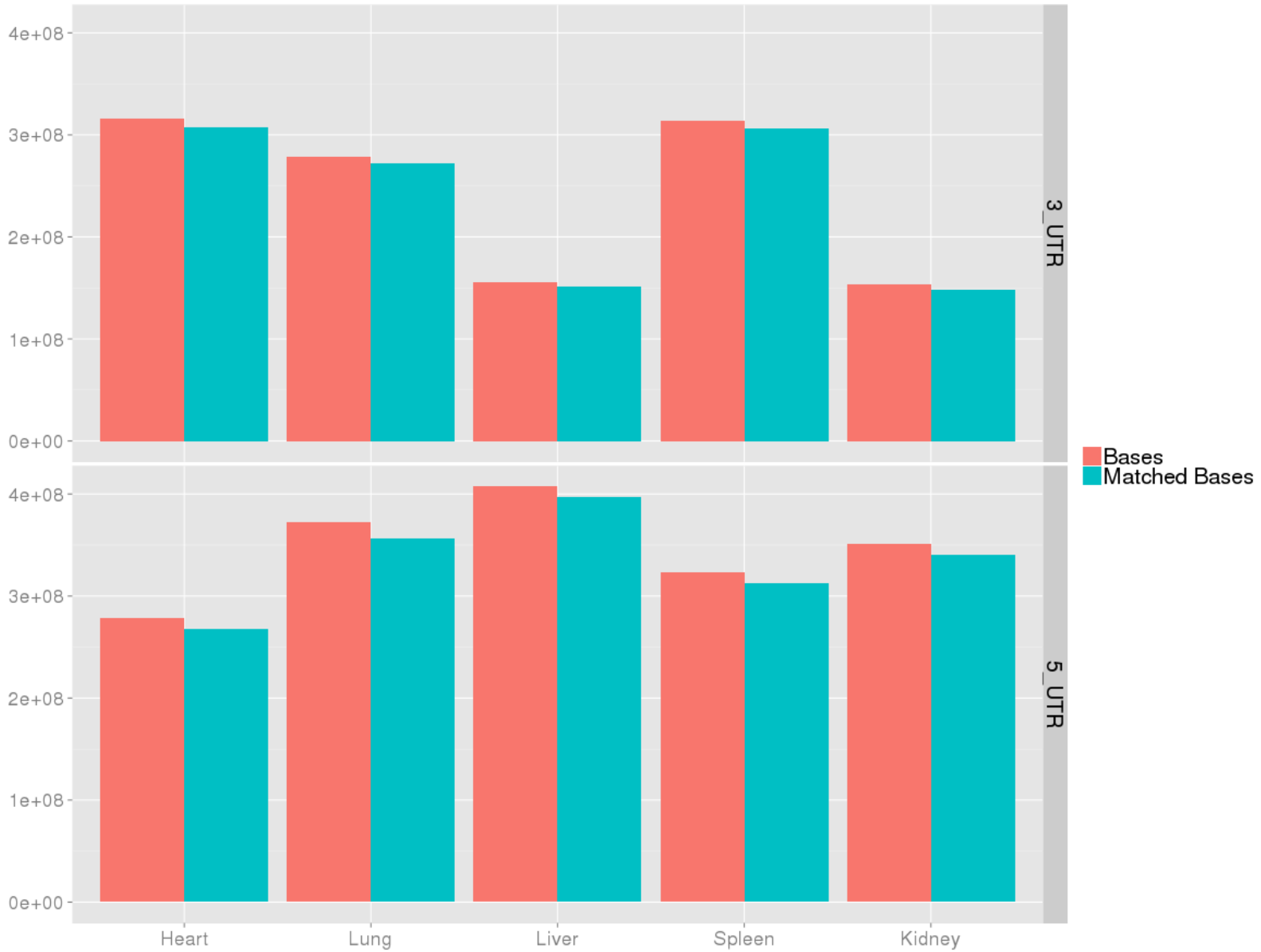
% Mapped bases per read



Overall bases stats

RACE	Tissue	Total Bases	Total Mapped Bases %
3'	Kidney	153216297	96,8
	Lung	278933711	97,7
	Liver	155550987	97,2
	Spleen	314204143	97,6
	Heart	315978690	97,1
5'	Kidney	351560008	96,9
	Lung	371977463	95,8
	Liver	407366100	97,6
	Spleen	322864564	96,9
	Heart	278842254	96,1

Total mapped bases by BLAT



Conclusions

- BLAT results reveal a significant proportion of reads with splitting events (gaps > 20).
- Parsing of the BLAT output by Inchworm produces predicted introns (gaps with splice site consensus seqs in the ends) and gaps without any splice site consensus seqs.
- Good mapping performance is obtained by using BLAT.