Mapping summary

454 Sequencing for UTR annotation

MGP

December 4, 2012

Preprocessing steps

- Small ($< 100 nts$) 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
	Testis	504428	460278	91.2	449739	97.7
	Brain	651200	620780	95.3	612989	98.7
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
	Testis	845357	826903	97.8	813418	98.4

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.
  - Minimum percent identity: 95%
  - Reference fasta: 1000g v37
  - Number of hits per read considered: 1, the best one reported by BLAT.

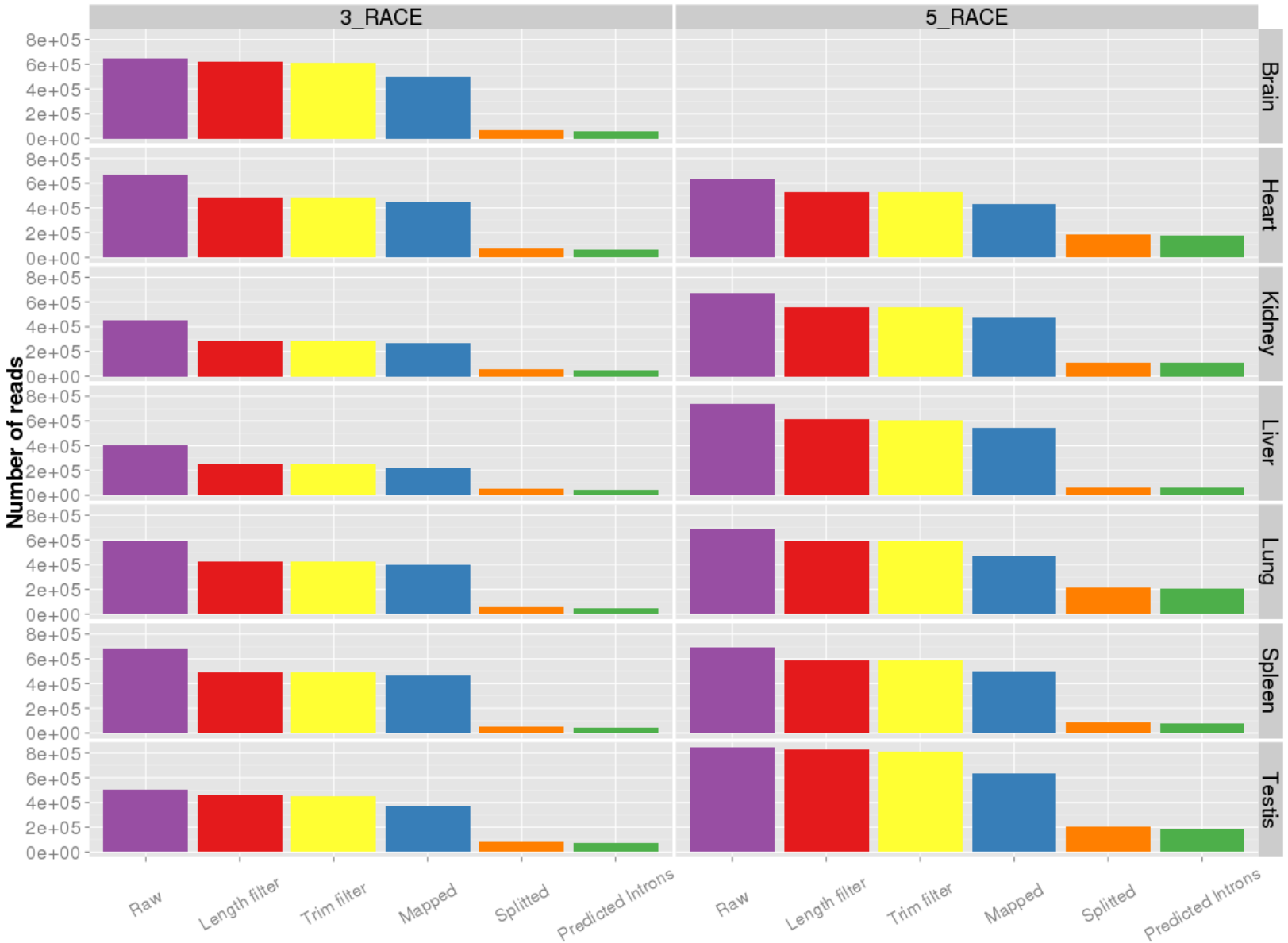
```
run_BLAT_shortReadPipeline.pl --single <fastq> --genome <100gv37fastafile> \
                              --seqType fq -o <outdir> -P 95
```

[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 %		Split %		% Introns	
		85% ID	95% ID	85% ID	95% ID	85% ID	95% ID
3'	Kidney	98.6	94.6	21.5	20.9	19.9	18.7
	Lung	98.7	94.7	15.1	14.5	13.8	12.5
	Liver	98.5	85.3	20.3	23.5	19.3	18.7
	Spleen	98.6	94.7	12.6	11.9	11.1	10.0
	Heart	98.5	93.8	16.3	15.5	15.0	13.8
	Testis	-	82.5	-	22.1	-	20.6
	Brain	-	81.1	-	12.7	-	11.3
5'	Kidney	98.4	86.2	24.6	23.5	24.8	22.2
	Lung	96.3	79.0	46.5	46.2	45.8	44.6
	Liver	99.3	88.9	12.4	12.0	14.0	10.8
	Spleen	98.7	85.2	17.7	16.7	16.7	15.2
	Heart	97.5	83.0	42.7	42.4	42.7	41.3
	Testis	-	77.7	-	31.9	-	30.3

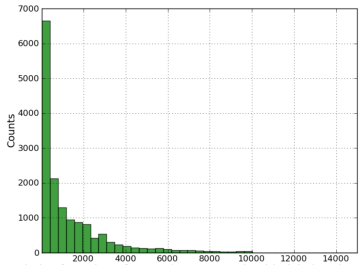
Mapped reads stats

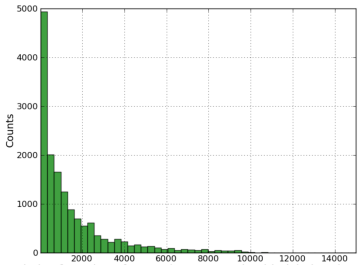


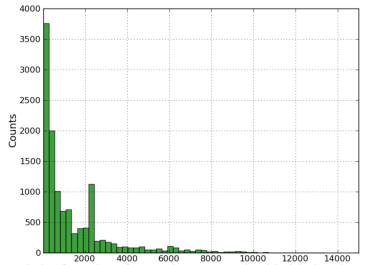
Mapped reads stats

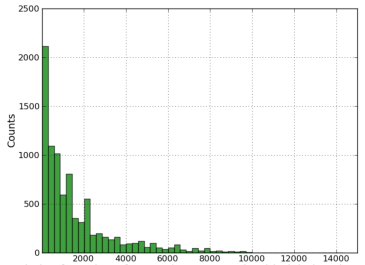
- High ratio of mapped reads
- A mapped read was considered to be split if a gap of size > 20 occurred in the alignment.
- Split reads represent an important proportion of the mapped reads.
- Intron prediction by Inchworm is based on the presence of splice site consensus seqs in the ends of the gaps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

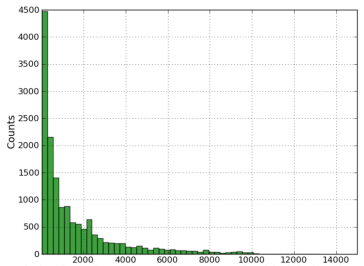
Distribution of the size of the longest gap in a read

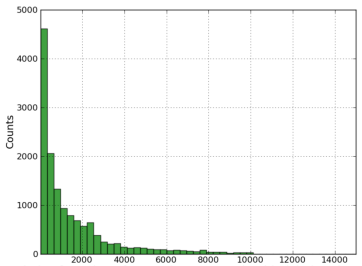


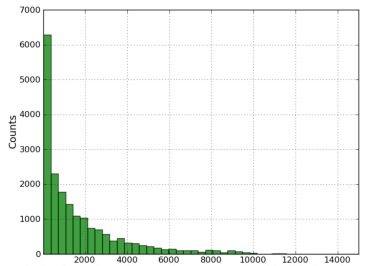


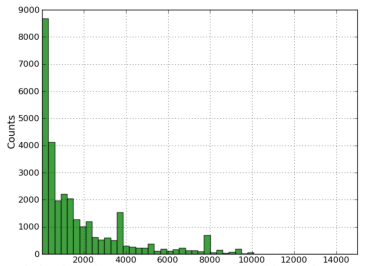


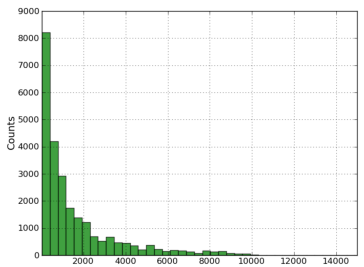


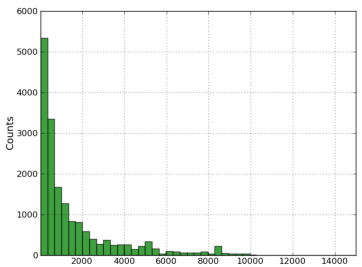


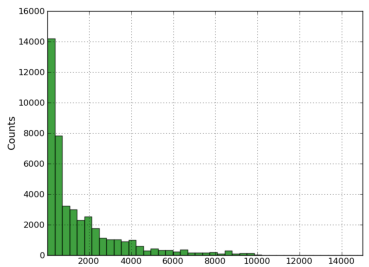


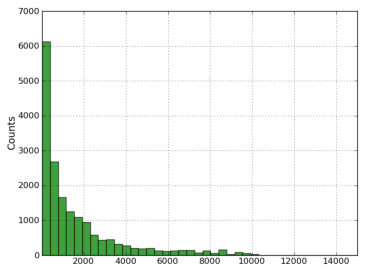


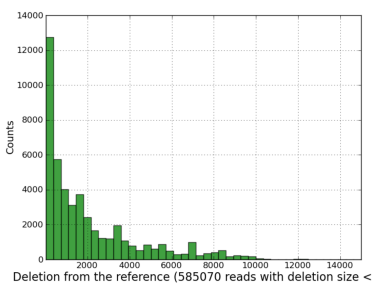




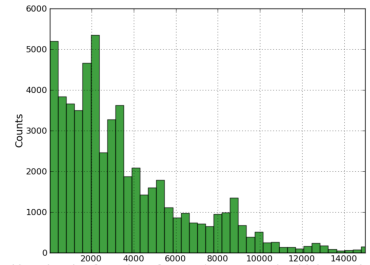




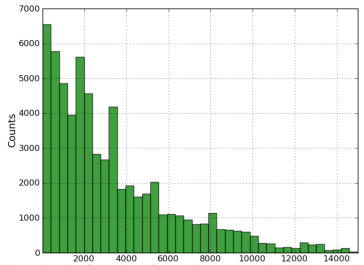




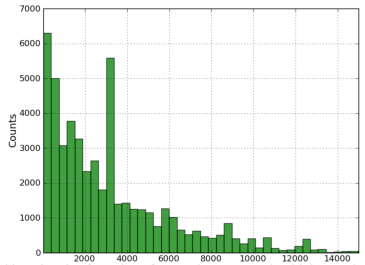
Distribution of sizes of the longest intron in a read



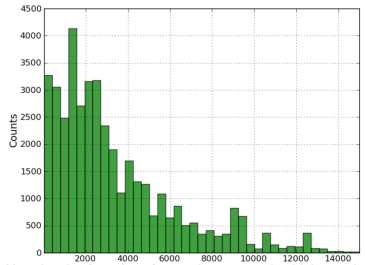
Skipped region from the reference (496841 reads with intron size < 20



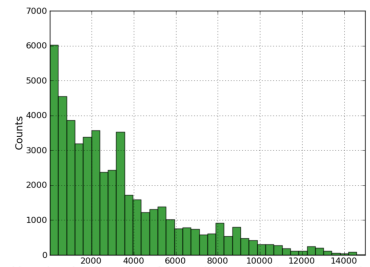
Skipped region from the reference (452807 reads with intron size < 20



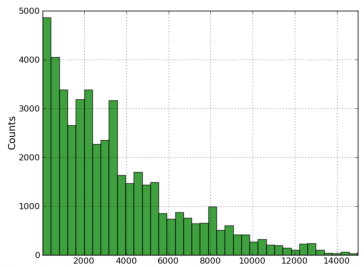
Skipped region from the reference (269083 reads with intron size < 20



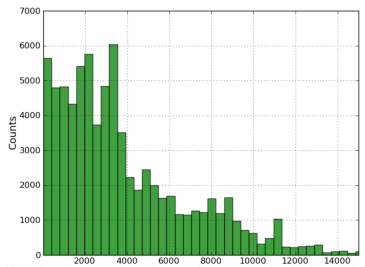
Skipped region from the reference (217409 reads with intron size < 20



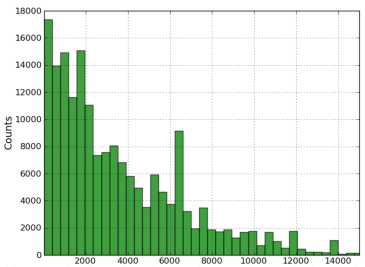
Skipped region from the reference (402061 reads with intron size < 20



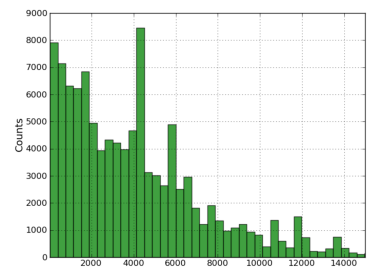
Skipped region from the reference (466363 reads with intron size < 20



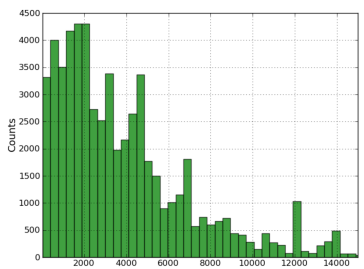
Skipped region from the reference (371106 reads with intron size < 20



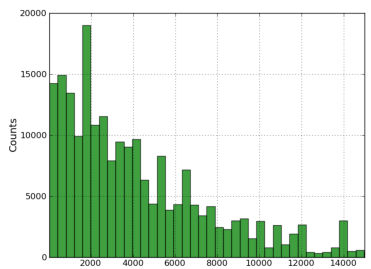
Skipped region from the reference (434678 reads with intron size < 20



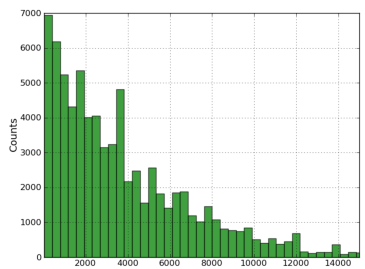
Skipped region from the reference (482098 reads with intron size < 20



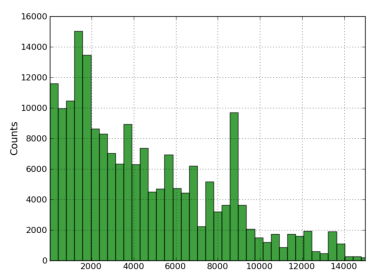
Skipped region from the reference (542400 reads with intron size < 20



Skipped region from the reference (467009 reads with intron size < 20



Skipped region from the reference (497850 reads with intron size < 20

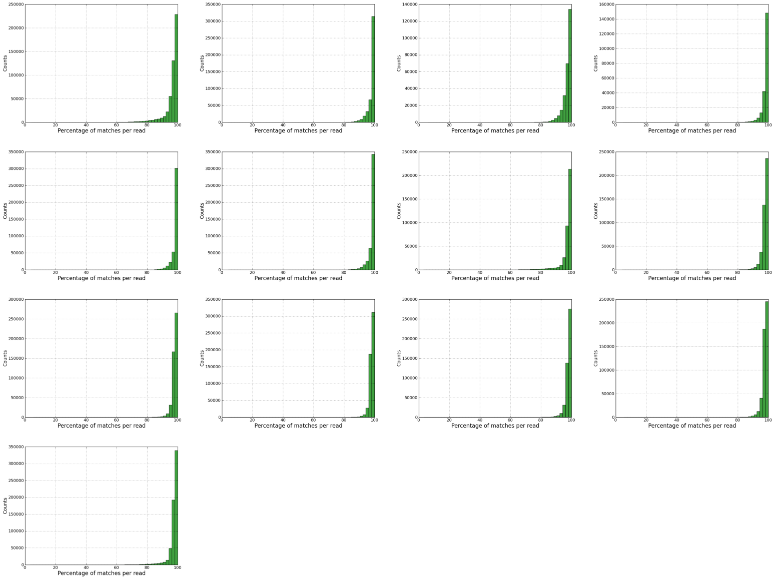


Skipped region from the reference (632212 reads with intron size < 20

Split reads

- 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.

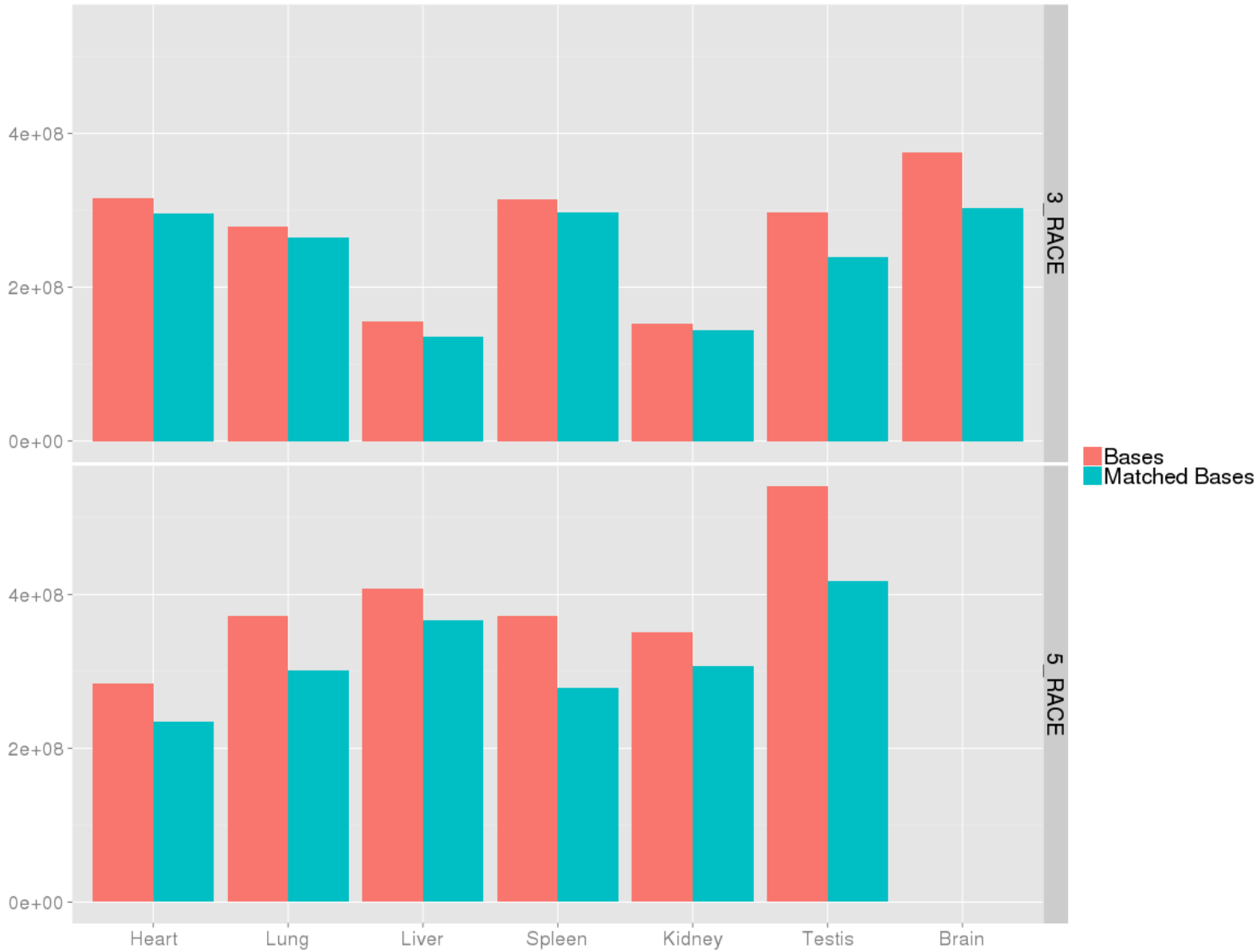
% Mapped bases per read



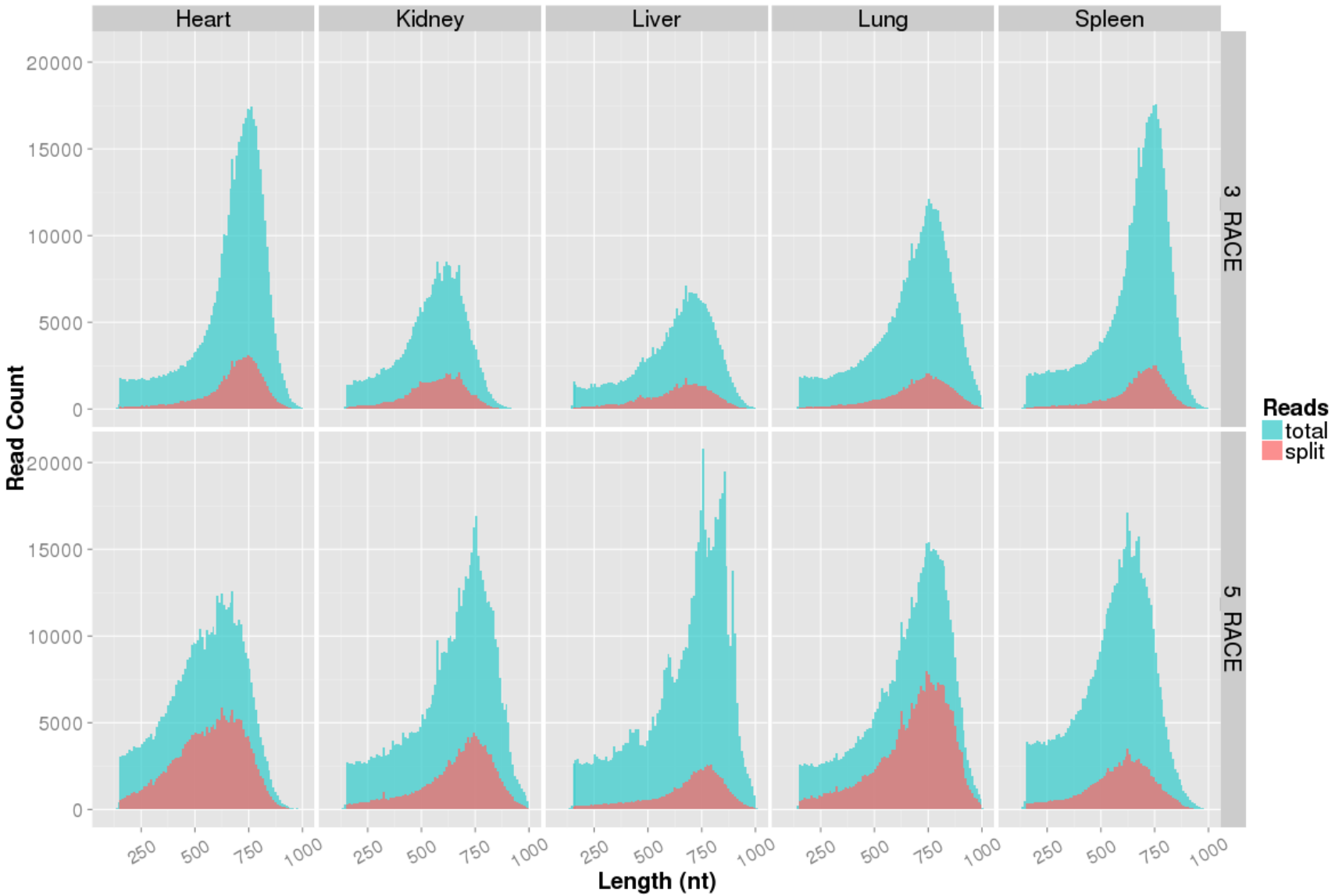
Overall bases stats

RACE	Tissue	Total Bases	Total Mapped Bases %	
			Blat 95% ID	Blat 85% ID
3'	Kidney	153,216,297	93.9	96.8
	Lung	278,933,711	94.7	97.7
	Liver	155,550,987	86.8	97.2
	Spleen	314,204,143	94.8	97.6
	Heart	315,978,690	93.8	97.1
	Testis	296,789,772	80.6	-
	Brain	375,839,920	80.6	-
5'	Kidney	351,560,008	87.4	96.9
	Lung	371,977,463	81.0	95.8
	Liver	407,366,100	89.8	97.6
	Spleen	322,864,564	74.9	96.9
	Heart	278,842,254	82.5	96.1
	Testis	540,479,622	77.4	-

Total mapped bases by BLAT



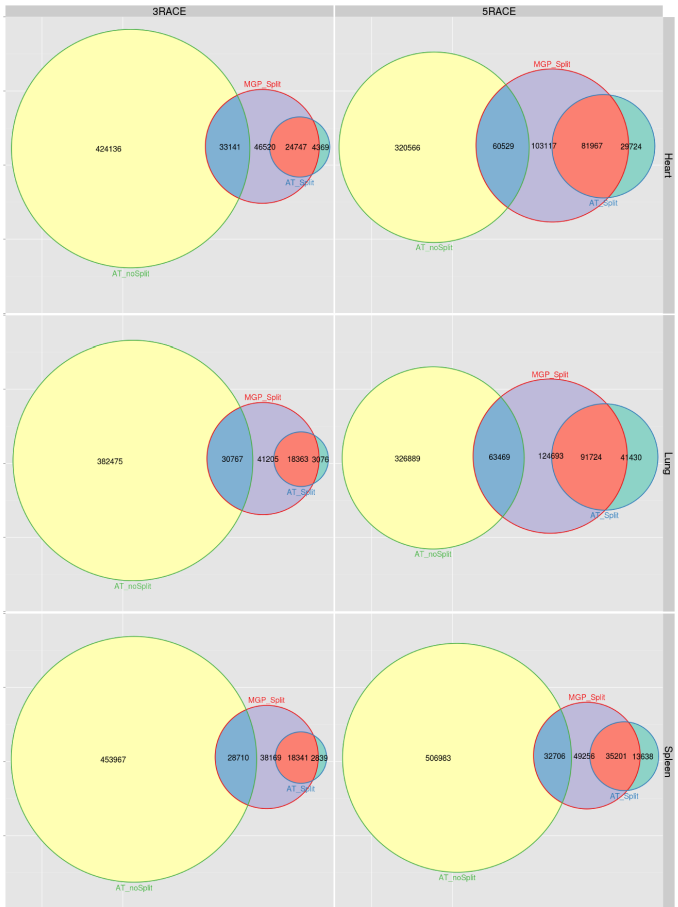
No bias split/read-length (Blat 85% ID)

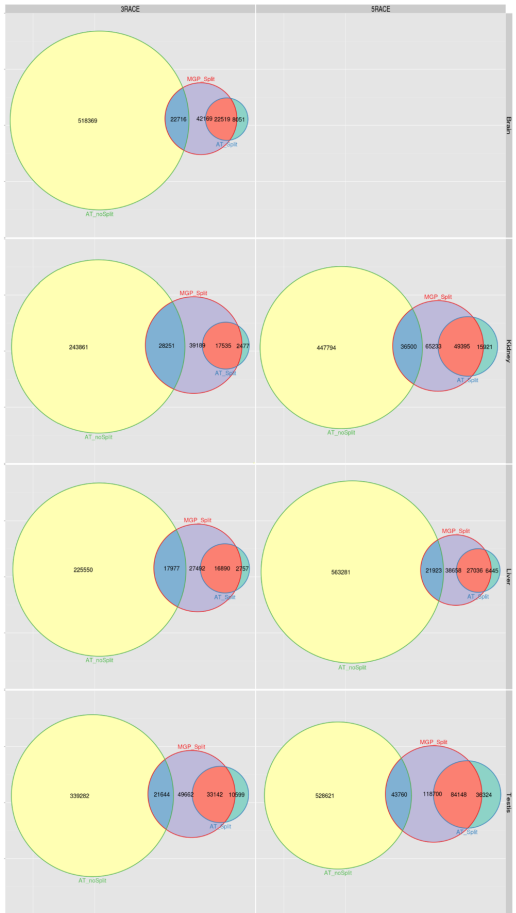


No bias split/read-length (Blat 85% ID)

- Read size distribution of split reads shows no bias towards long/short reads.

Mapping approaches comparison of split reads





Mapping approaches comparison of split reads

	AT_noSpl-MGP_Spl /AT_noSpl	AT_noSpl-MGP_Spl /MGP_Spl	AT_Spl-MGP_Spl /MGP_Spl	AT_Spl-MGP_Spl /AT_Spl
Heart_3'	7.25%	31.74%	23.70%	84.99%
Heart_5'	15.88%	24.64%	33.37%	73.39%
Lung_3'	7.45%	34.06%	20.33%	85.65%
Lung_5'	16.26%	22.68%	32.77%	68.89%
Spleen_3'	5.95%	33.69%	21.52%	86.60%
Spleen_5'	6.06%	27.91%	30.04%	72.08%
Brain_3'	-	-	-	-
Brain_3'	4.20%	25.99%	25.76%	73.66%
Kidney_3'	10.38%	33.25%	20.64%	87.62%
Kidney_5'	7.54%	24.15%	32.68%	75.62%
Liver_3'	7.38%	28.83%	27.09%	85.97%
Liver_5'	3.75%	25.02%	30.86%	80.75%
Testis_3'	6.00%	20.72%	31.73%	75.77%
Testis_5'	7.65%	17.74%	34.12%	69.85%

Mapping approaches comparison of split reads

- Most of split reads (69% – 87%) by the first approach are also split by the Blat (95% identity) approach.
- 4% – 16% of the reads that were not split by the first approach are split by Blat.
- More split reads are mapped by using 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.
- 95% identity mapping results in a slight decrease of the number of mapped reads. No significant changes are observed in the rest of mapping stats when compared to 85% identity.
- Comparison with the AT mapping approach shows that Blat maps more reads and is able to capture most of the previous split reads.
- Good mapping performance is obtained by using BLAT.