Evaluation of Read Alignment Submissions

(RGASP.2 BAM File Submissions)

Andre Kahles, Regina Bohnert, Jonas Behr, and Gunnar Rätsch

Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

February 25, 2010

The Wellcome Trust Genome Campus, Hinxton

Submissions

1. *C. elegans*

CI	15G	76,782,527 algnmts.
Lior	19G	86,628,410 algnmts.
Tyl	11G	*157,639,782* algnmts.
Gun	11G	52,894,661 algnmts.

2. *D. melanogaster*

CI	11G	56,941,798 algnmts.
Lior	17G	79,410,759 algnmts.
Tyl	9G	*138,553,978* algnmts.
Gun	13G	64,114,176 algnmts.

3. Human

CI	19G	89,101,033 algnmts.
Lior	26G	118,461,050 algnmts.
Ger	28G	133,403,777 algnmts.
Tyl	19G	*283,872,575* algnmts.
Gun	18G	85,682,837 algnmts.

Legend:

CI : Christian Iseli et al., CH

Lior : Lior Pachter et al., USA

Ger : Mark Gerstein et al., USA

Tyl : Tyler Alioto et al., ES

Gun : Gunnar Rätsch et al., DE

Problems in BAM Submission Files

- **CI** Introns annotated as deletions in cigar strings; used `S` (softclip) which was treated as `M` (mismatch)
- **Lior** OK!
- **Ger** OK!
- **Tyl** All introns were annotated too short by one nucleotide, read and quality information missing
- **Gun** Insertion/deletion mix-up in position calculations

We tried to fix these problems based on the submission files.

Results are based on sanitized file versions.

For Tyl we still had problems and too little time. Some plots are missing or based on unsanitized alignments.

Summarization and Evaluation Strategy

Summaries (Histograms):

- Number of exons per alignment
- Number of mismatches/indels per read position
- Number of introns per read position
- Genome coverage
- Intron coverage

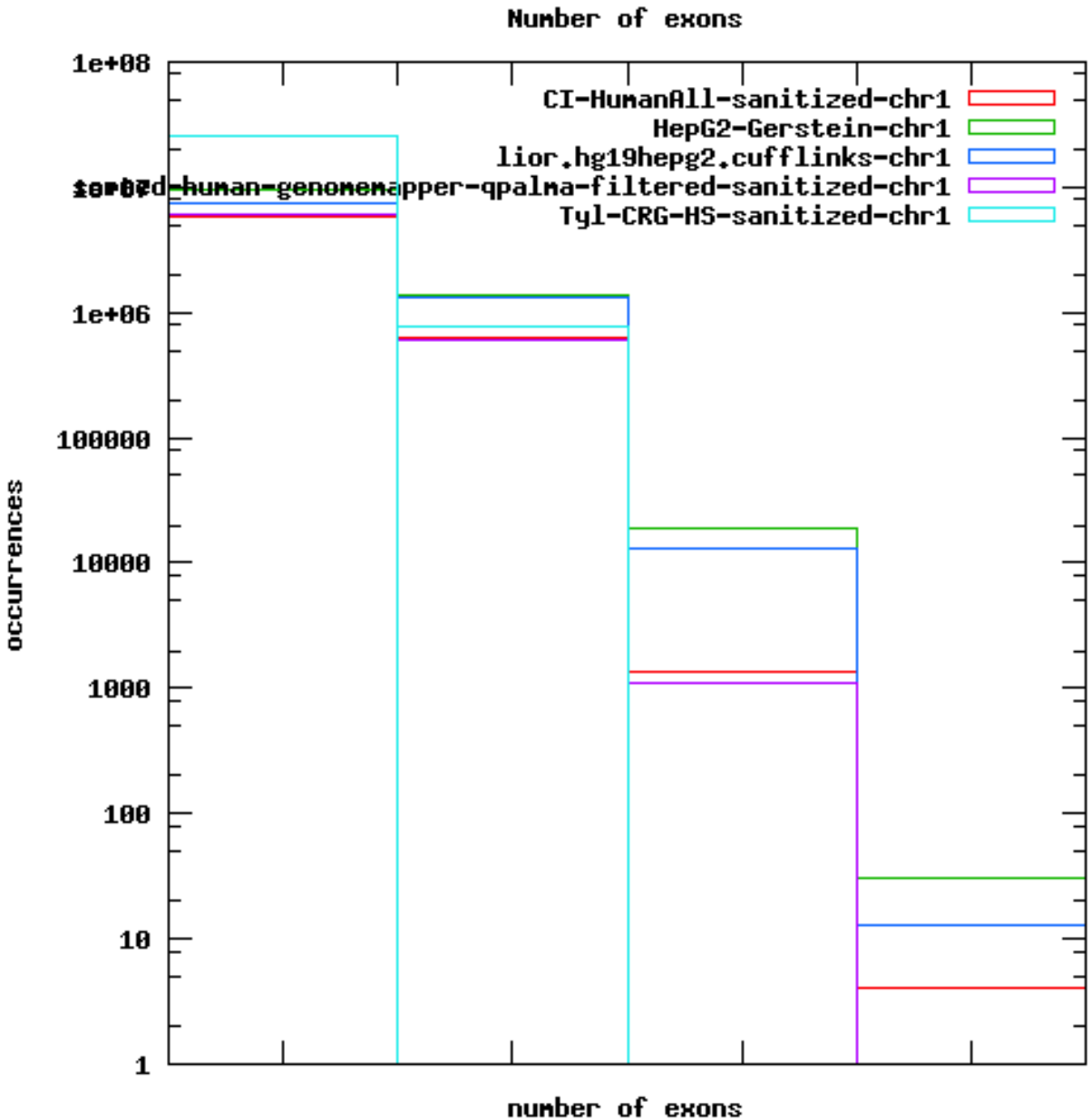
Accuracy Evaluation (based on RGASP genome annotations):

- Sensitivity and Specificity of intron predictions
- Number of reads sticking out of annotated exons

Results are shown for human (chromosome 1 only), where most submissions are available. (similar results for other organisms)

All of these results are very preliminary!

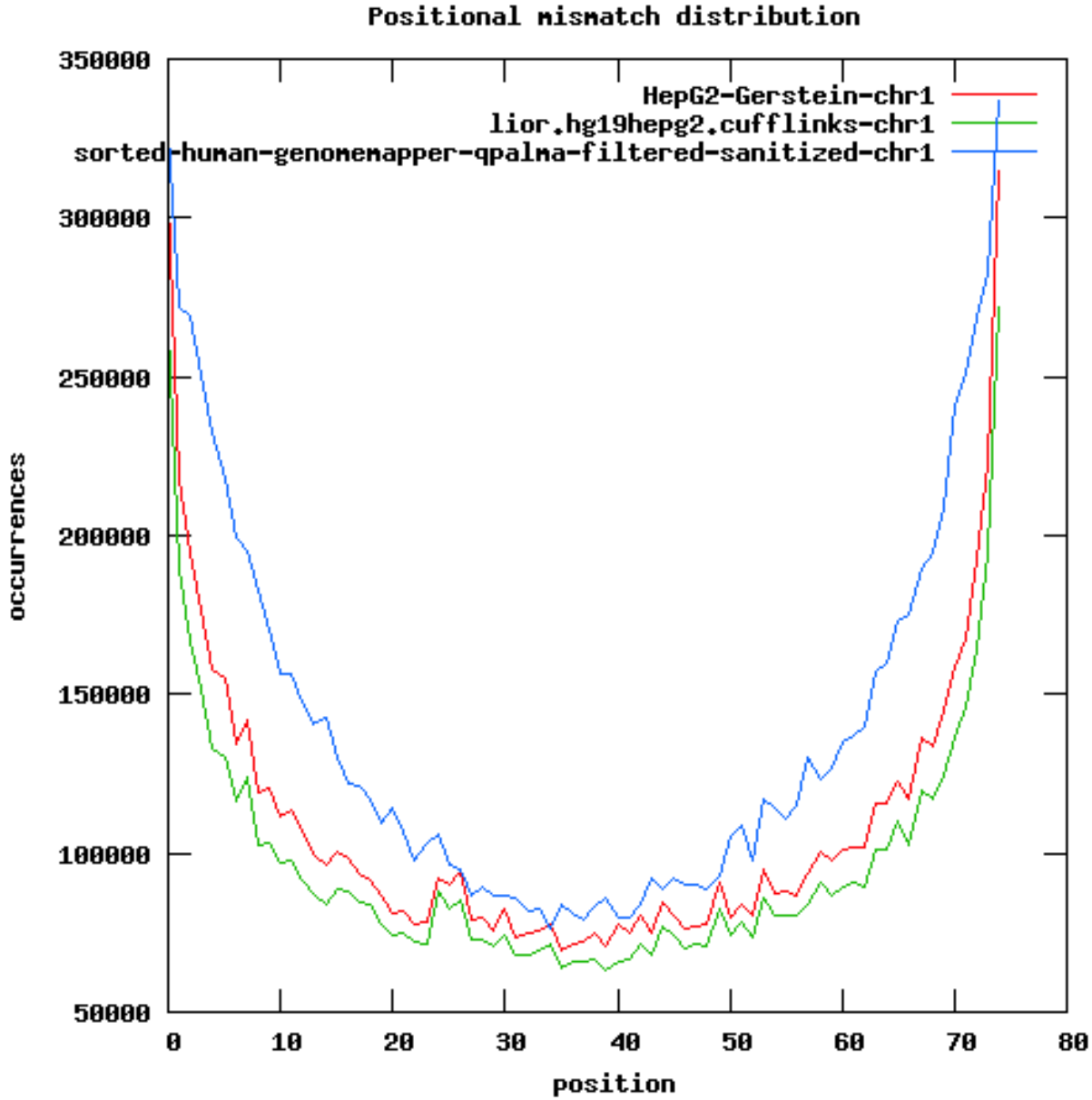
Number of Exons per Alignment



Most unspliced alignments from Tyl.

Fewest alignments from Gun (due to quality filtering prior to submission)

Number of Mismatches per Read Position



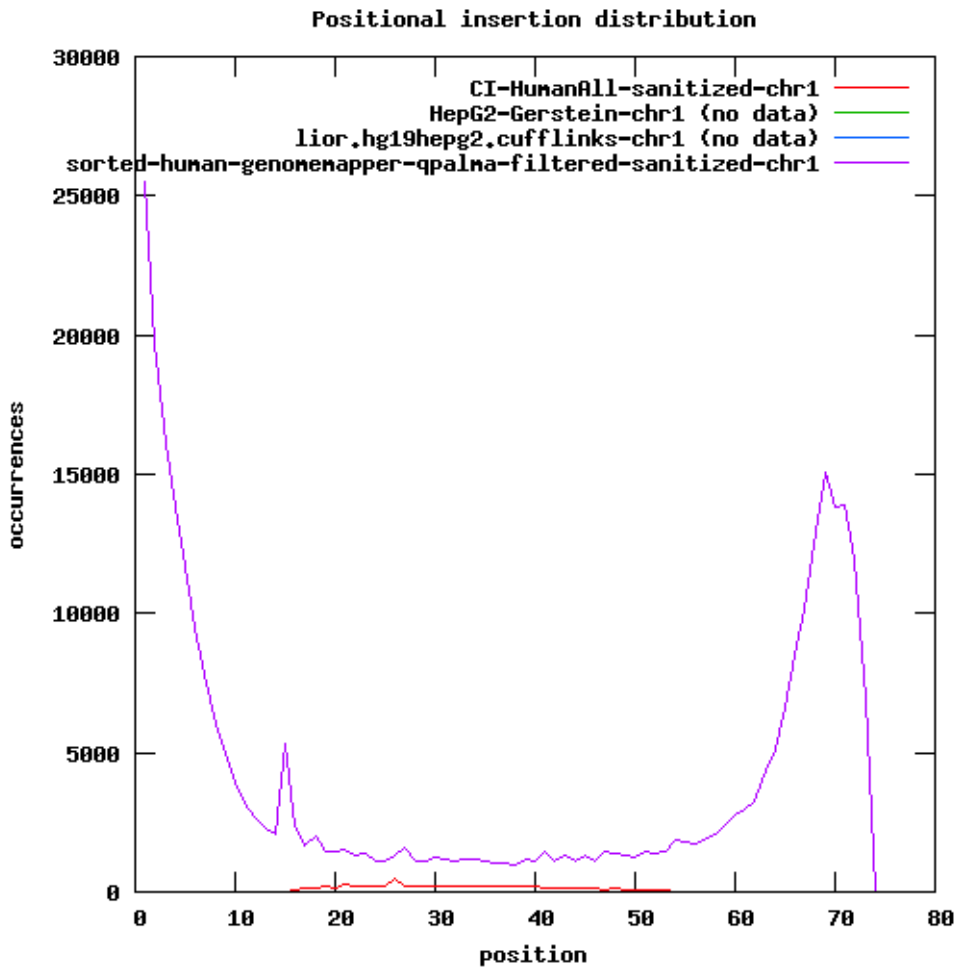
Which method allows for most mismatches?

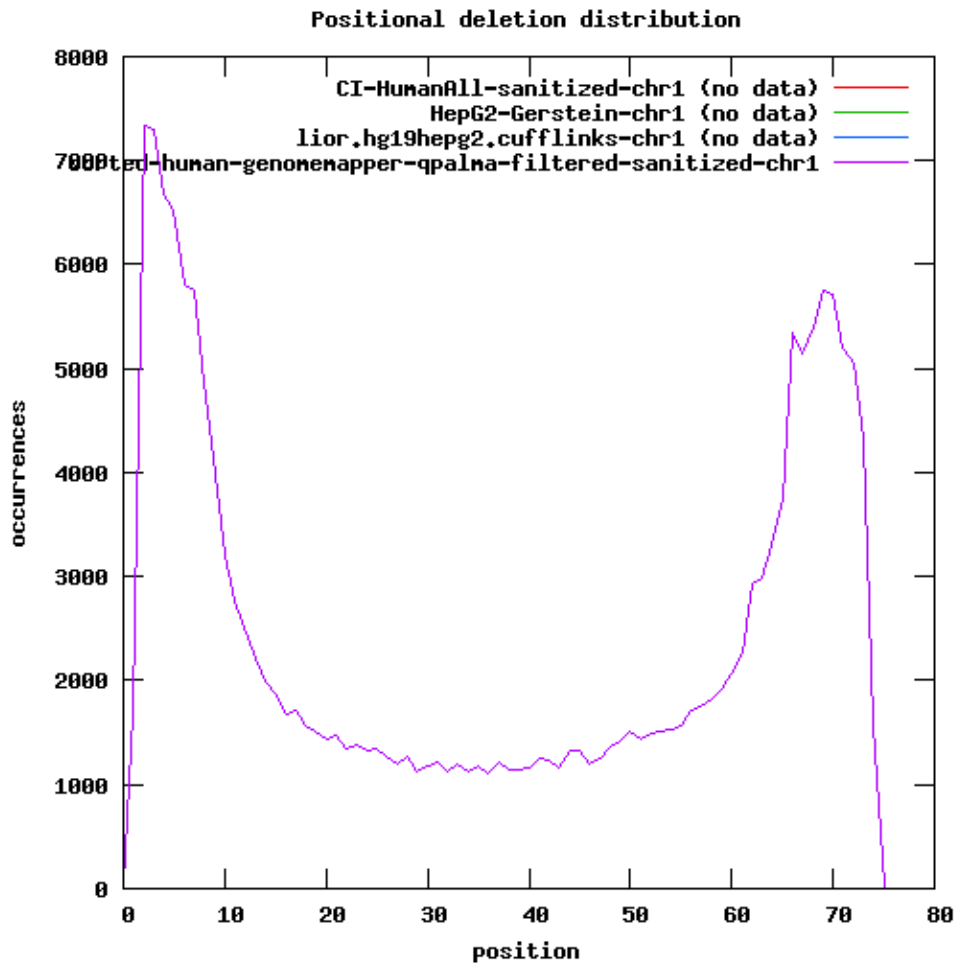
Most mismatches for Gun, fewer for Ger and Lior.

CI evaluation needs to be checked again (much more mismatches).

Tyl not finished in time.

Number of Indels per Read Position

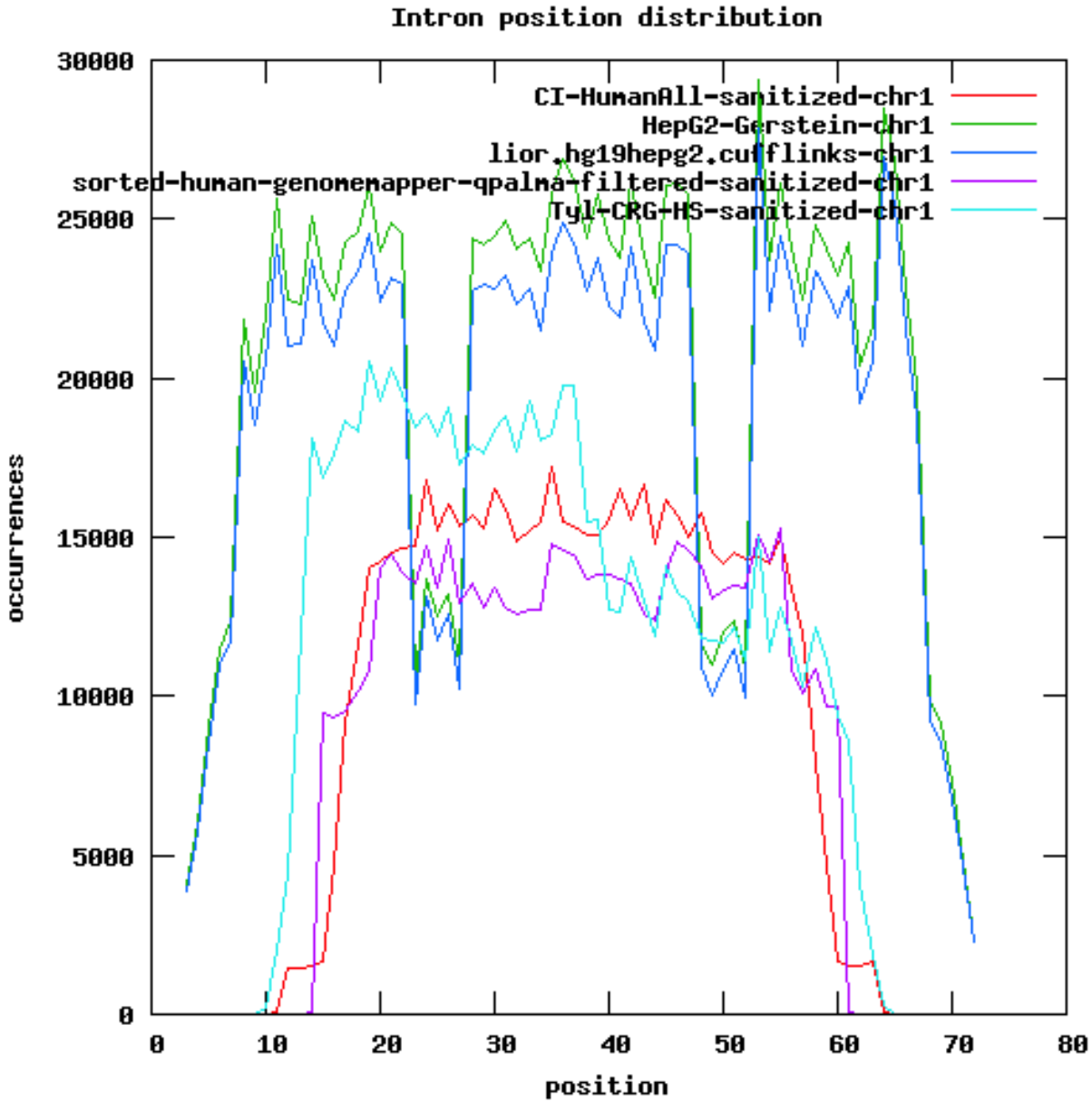




Which method allows for indels?

Alignments typically without insertions(left) and deletions (right), except for Gun (very few insertions also for CI).

Number of Introns per Read Position



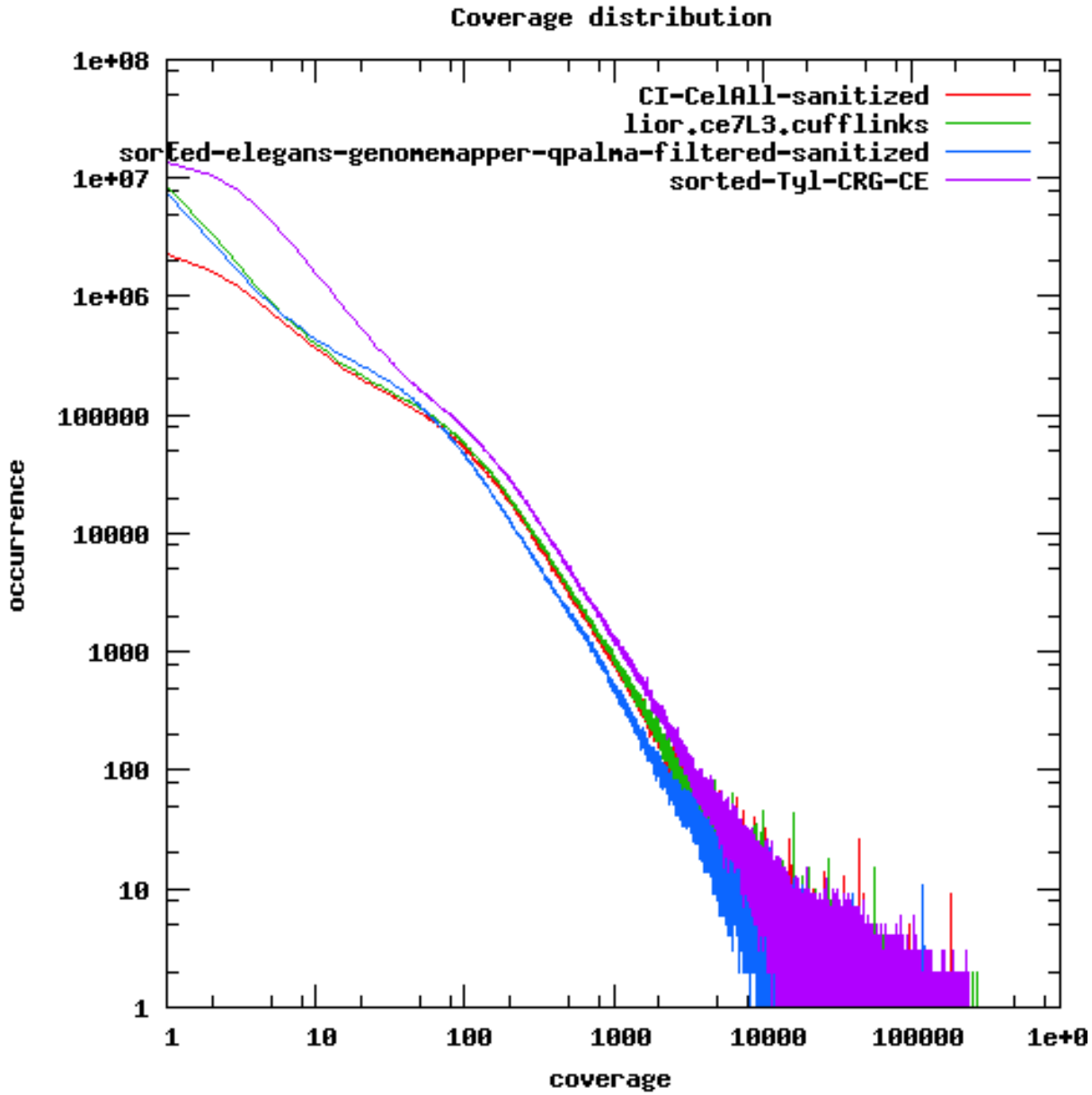
Where are introns relative to read positions?

Lior and Ger find less introns in central read positions, more introns overall

CI and Gun ignore alignments at read boundaries

Tyl not finished.

Genome Coverage (*C. elegans*)



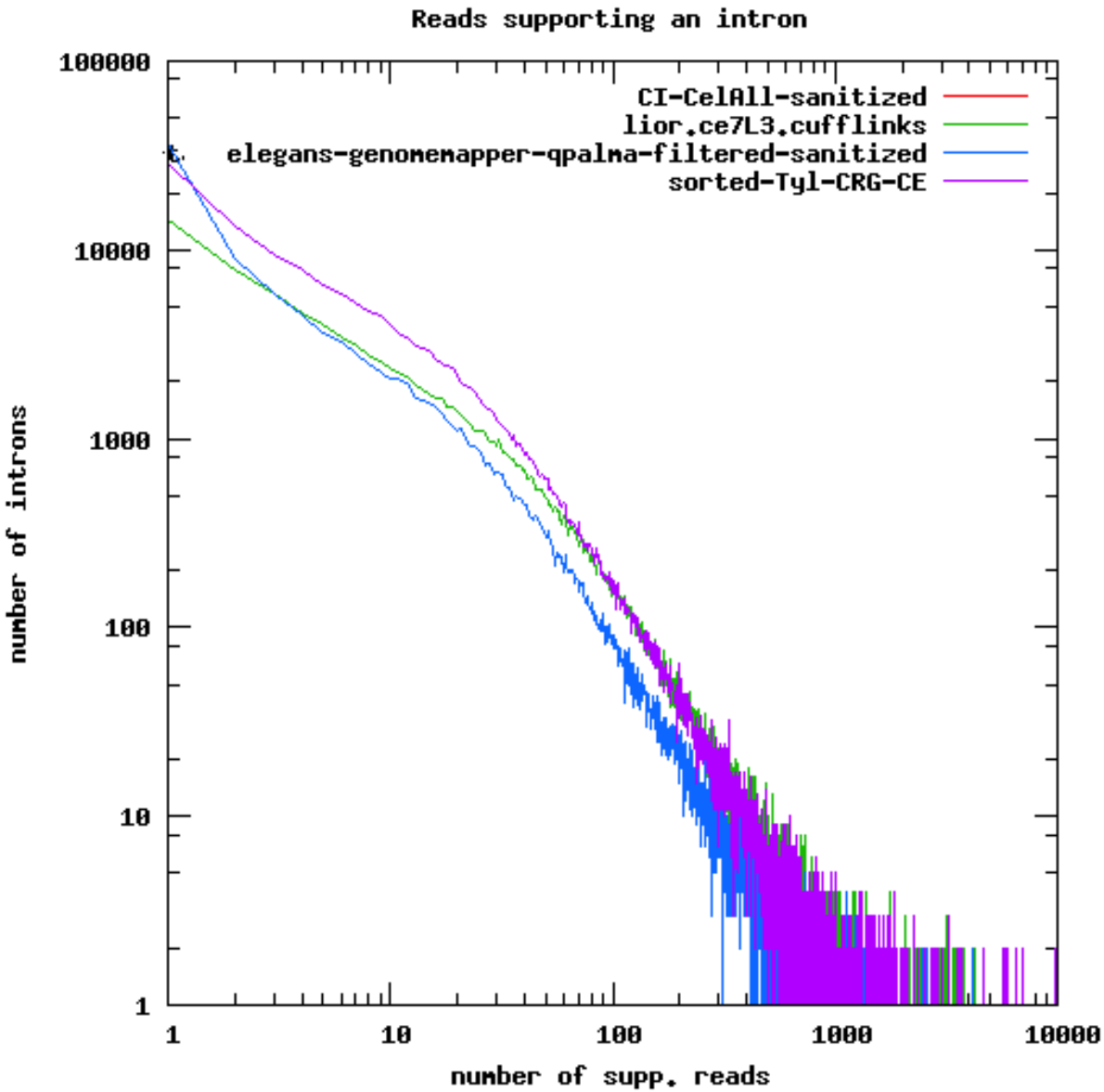
How many nucleotides are covered by how many reads?

Tyl Highest overall coverage

CI Smallest fraction of lowly covered regions

Gun Smallest fraction of highly covered regions

Intron Coverage (*C. elegans*)



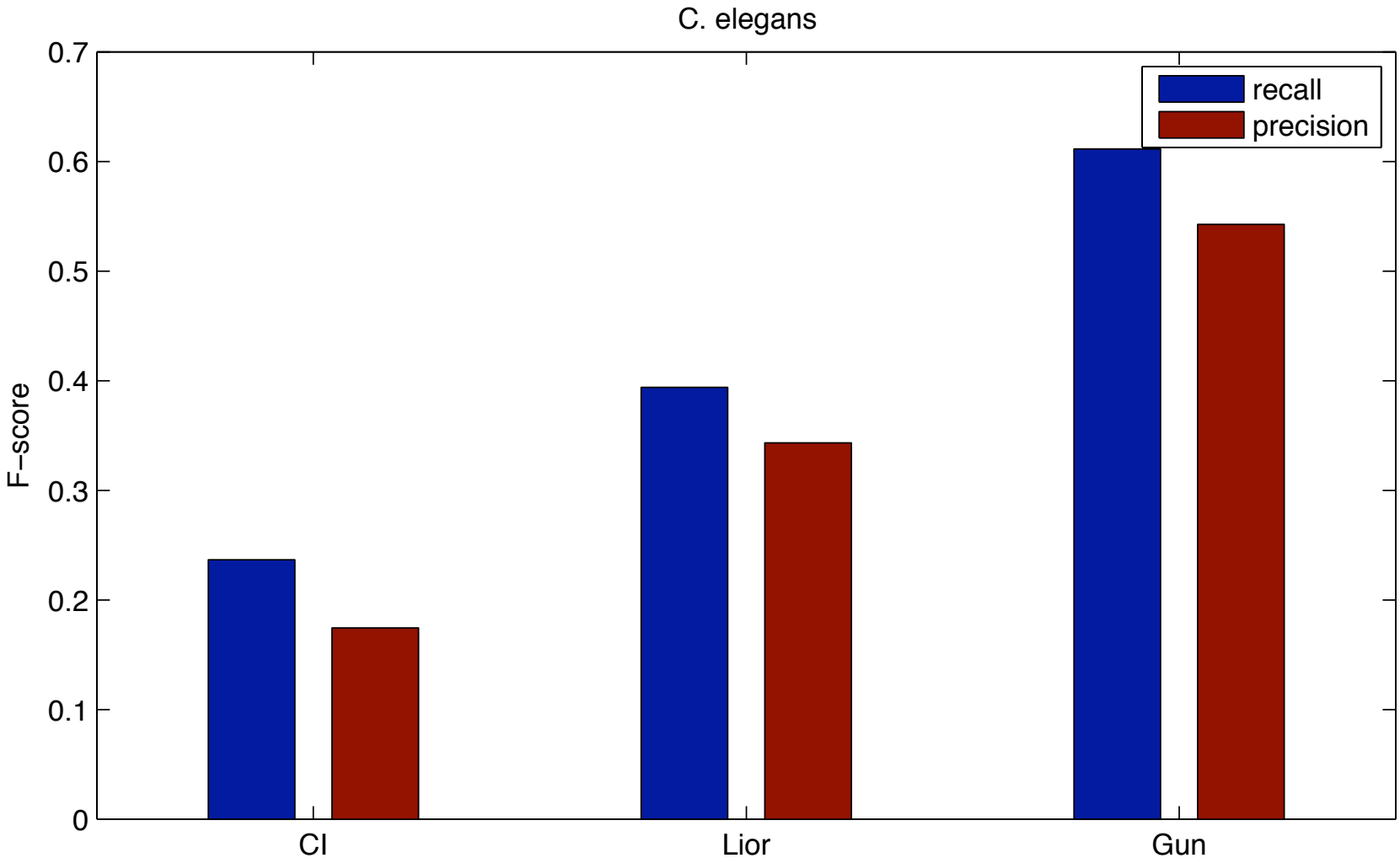
How many introns are confirmed by how many reads?

Tyl Most introns confirmed at high coverage

Lior Fewest introns with low coverage

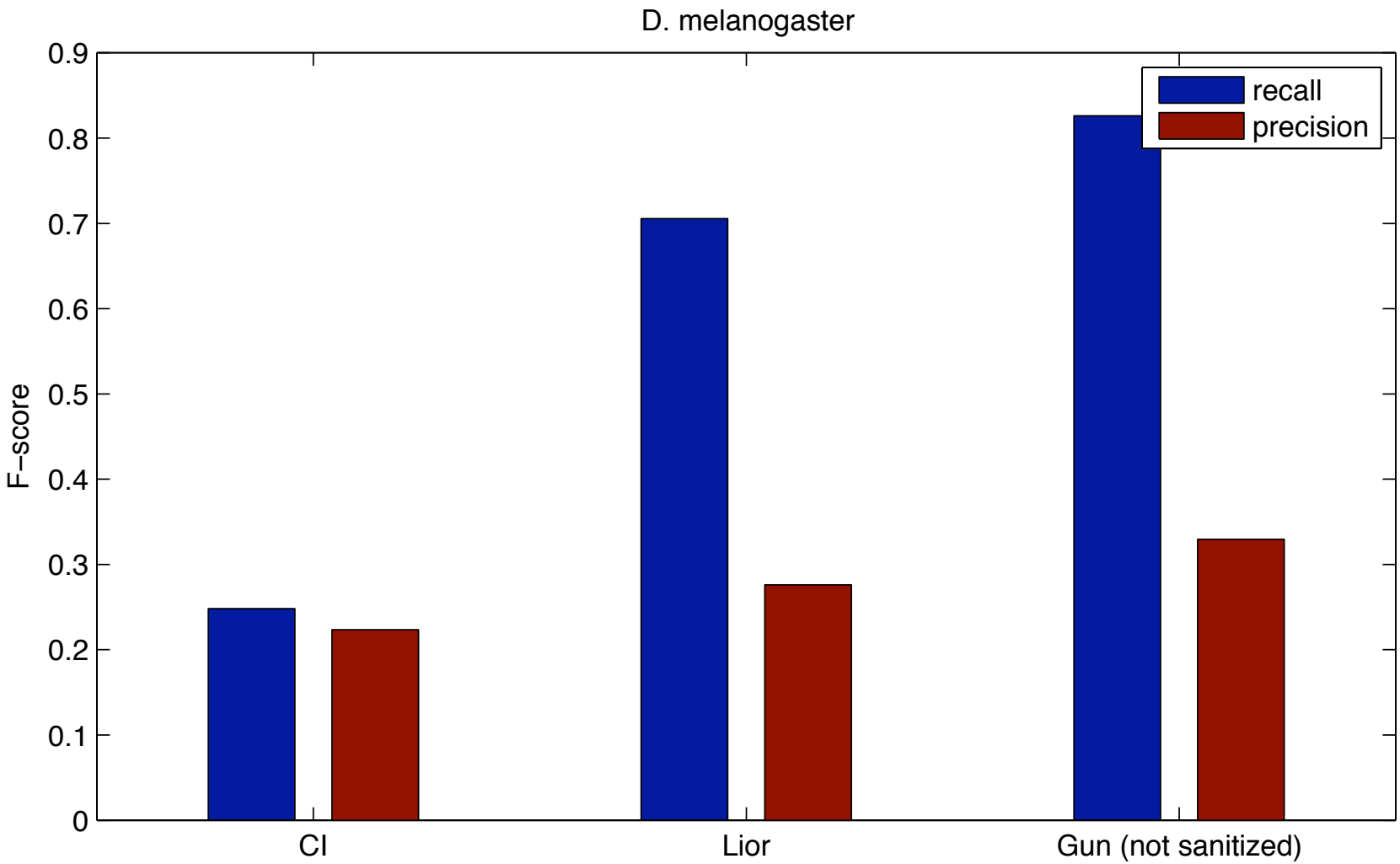
Gun Most introns confirmed at low coverage

Intron Precision and Recall (*C. elegans*)



No results for Tyl yet (much more alignments).

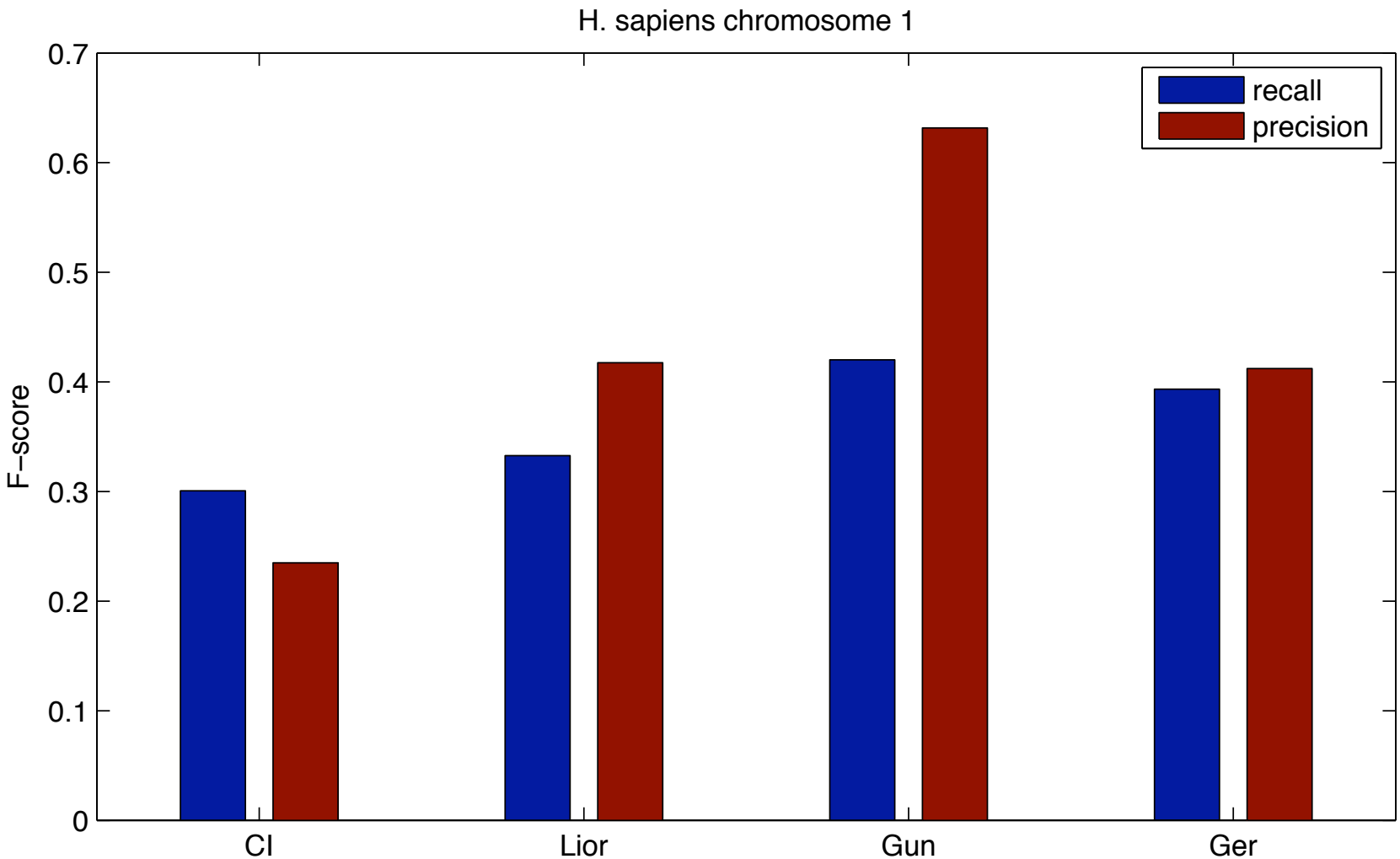
Intron Precision and Recall (*D. melanogaster*)



Very high sensitivity for Lior and Gun.

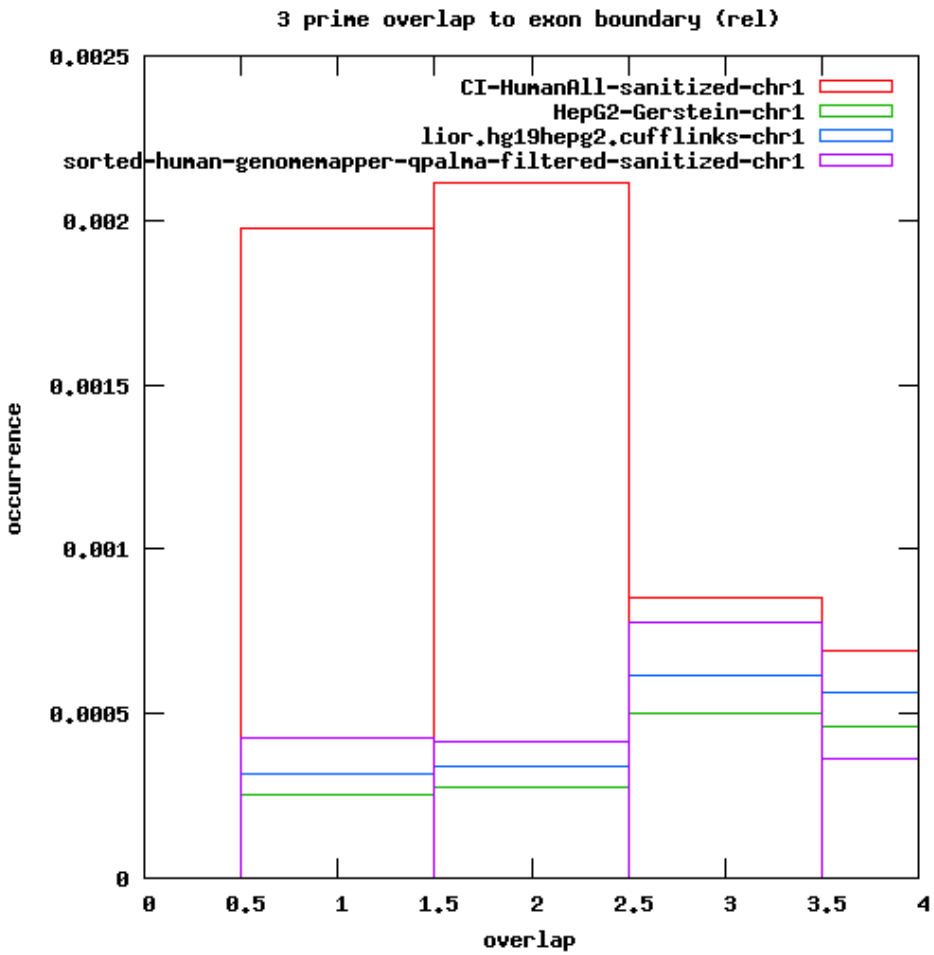
No results for Tyl yet. Alignments from Gun not sanitized (+10%?)

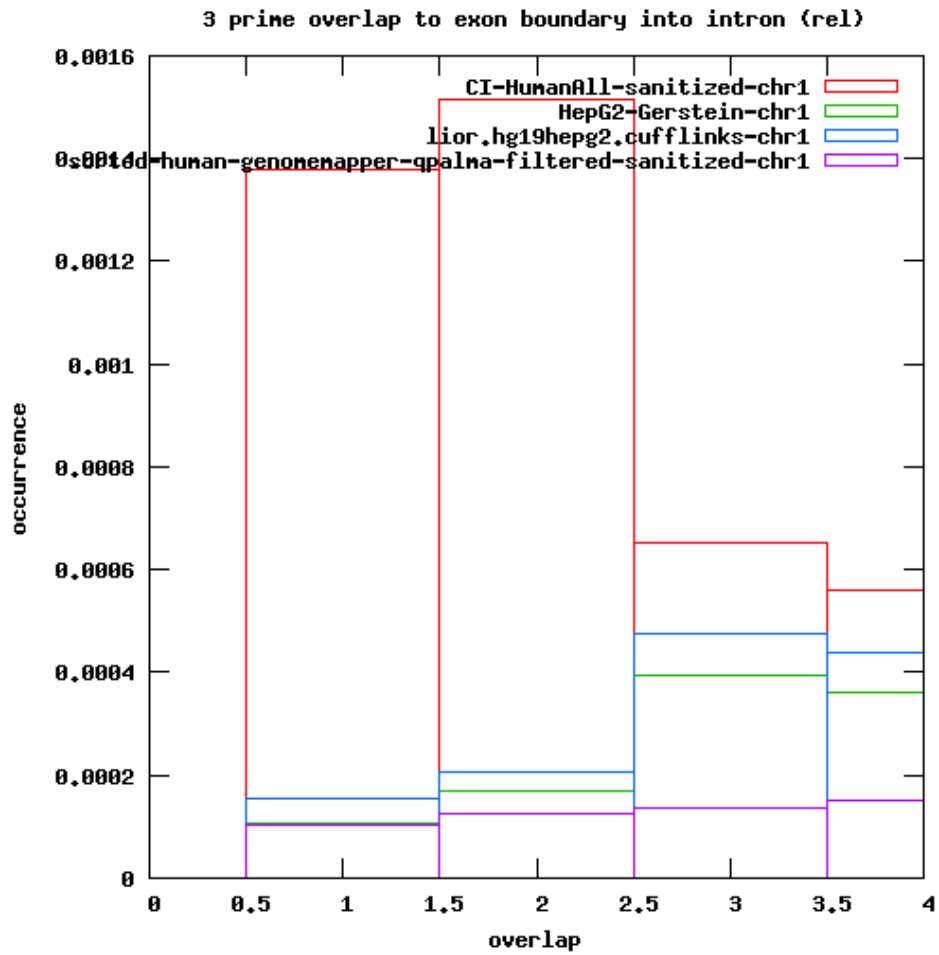
Intron Precision and Recall (human)



Lior and Ger presumably both used TopHat/Cufflinks (but with different settings). No results yet for Tyl.

Alignments over Exon Boundaries





How many alignments go over the exon boundaries (here: 3' end).
Precaution: Same plots for 5' ends look quite different (buggy?).

Remarks

- So far considered all alignments, but some submissions had multiple alignments
- Some algorithms filtered their alignment sets, should one try to unify the filtering to make results comparable?
- Restrict analyses only to expressed transcripts/genes?
  - will increase recall
  - may decrease precision
- We will provide
  - Figures for all organisms
  - Evaluation code (python) to anybody who wants to reproduce the evaluation results (after some cleanup)