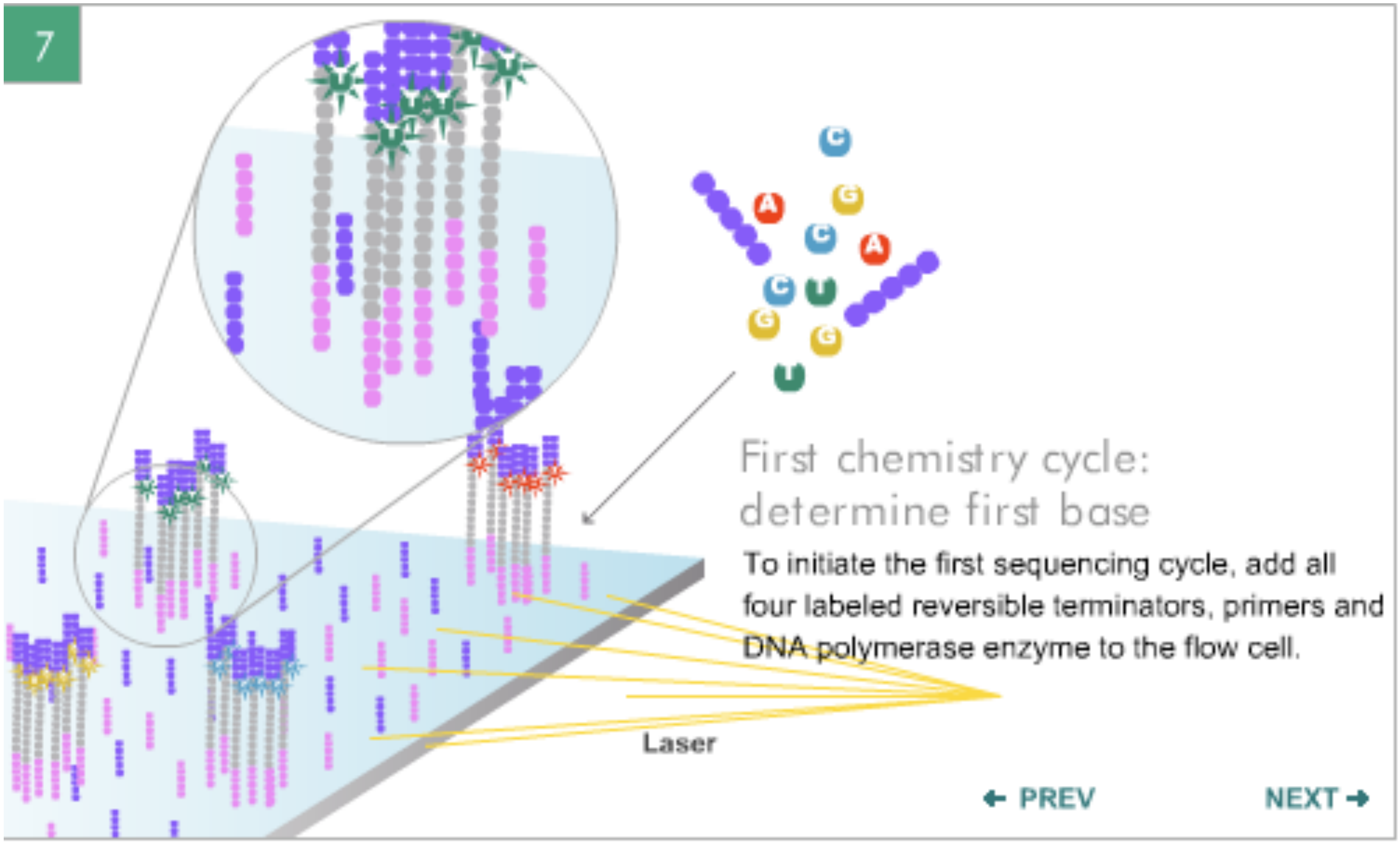
GNUMap: Unbiased Probabilistic Mapping of Next-Generation Sequencing Reads

Nathan Clement
Computational Sciences Laboratory
Brigham Young University
Provo, Utah, USA

GNUMap

Next-Generation Sequencing (Solexa/Illumina)



First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

Computational Sciences Laboratory
Brigham Young University

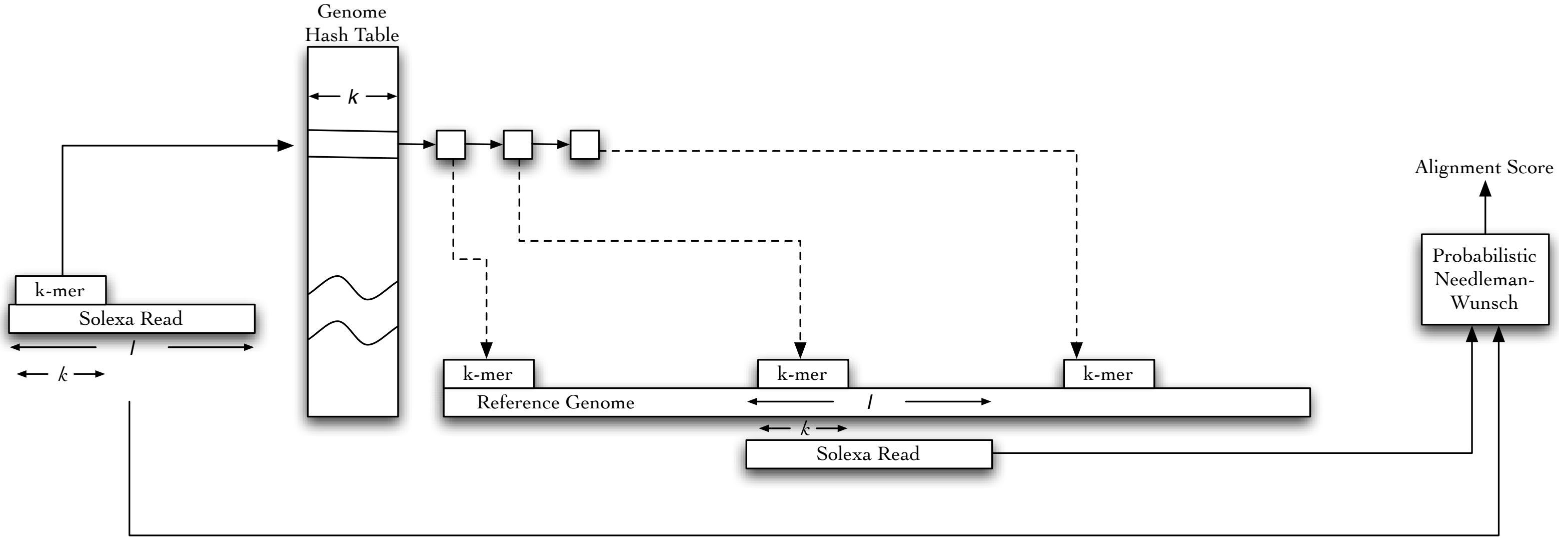
Computational Sciences Laboratory
Brigham Young University

Problem Statement

- Map next-generation sequence reads with variable nucleotide confidence to a model reference genome that may be different from the subject genome.
  - Speed
    - Tens of millions of reads to a 3Gbp genome
  - Accuracy
    - Mismatches included?
    - Repetitive regions
  - Visualization

Workflow

Computational Sciences Laboratory
Brigham Young University



Indexing the genome

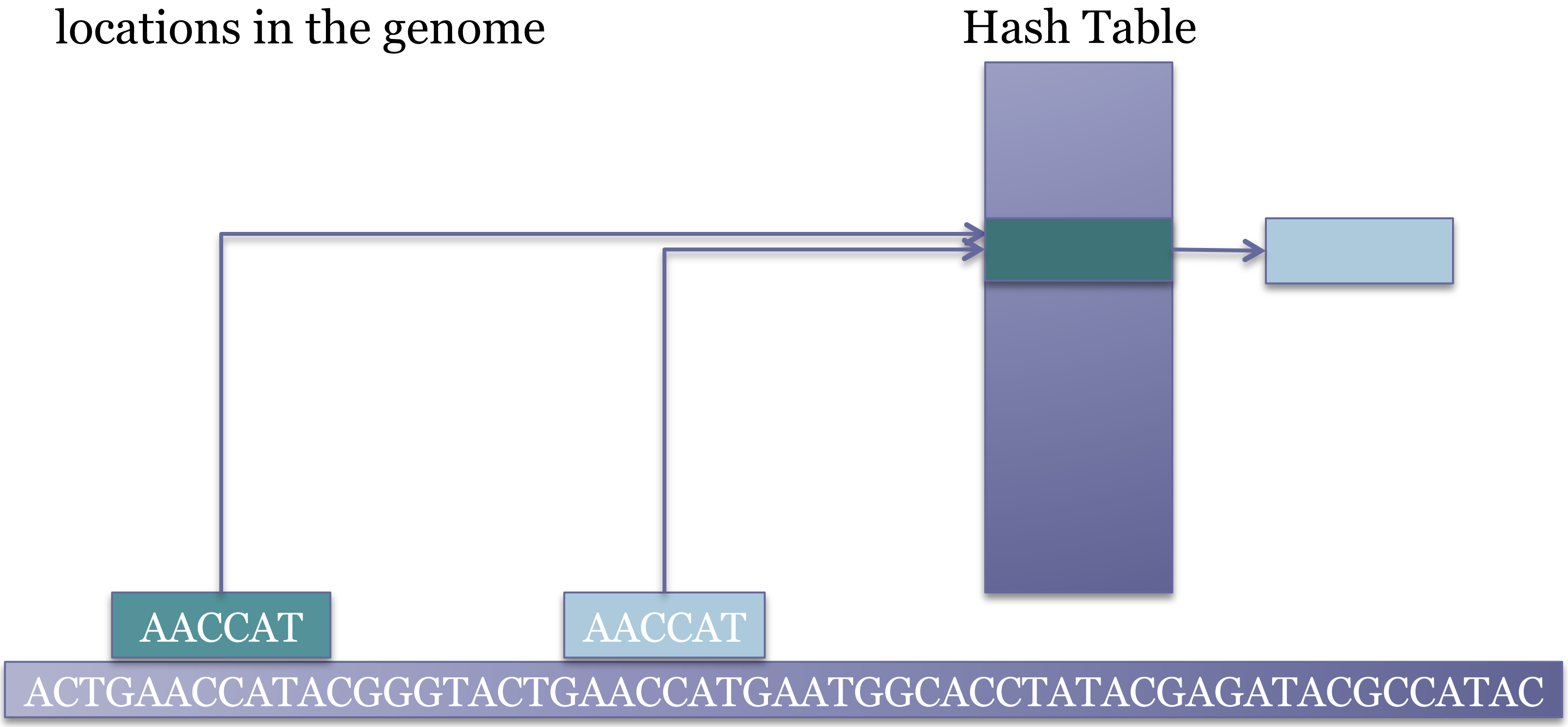
Computational Sciences Laboratory
Brigham Young University

- Fast lookup of possible hit locations for the reads
    - Hashing groups locations in the genome that have similar sequence content
        - k-mer hash of exact matches in genome can be used to narrow down possible match locations for reads
    - Sorting genome locations provides for content addressing of genome
- GNUMAP uses indexing of all 10-mers in the genome as seed points for read mapping

Building the Hash Table

Computational Sciences Laboratory
Brigham Young University

Sliding window indexes all locations in the genome



Alignment

- Given a possible genome match location, determine the quality of the match
- If you call bases in the read
  - Every base gets the same weight in the alignment, no matter what the quality
  - Later bases in the read that have lower quality have equal weight in the alignment with high quality bases at the start of the read
- GNUMap uses a Probabilistic Needleman-Wunsch to align reads found with seed points from the genome hash

Probabilistic Needleman Wunsch

Computational Sciences Laboratory
Brigham Young University

j	0	1	2	3	4	5
PWM	*A*	0.059	0.000	0.172	0.271	0.300
	C	0.108	0.320	0.136	0.209	0.330
	G	0.305	0.317	0.317	0.164	0.045
	T	0.526	0.578	0.375	0.356	0.325
NW		T	T	T	T	C
	0	-2	-4	-6	-8	-10
T	-2	**0.052**	-1.948	-3.948	-5.948	-7.948
T	-4	-1.844	**0.208**	-1.792	-3.792	-5.792
C	-6	-3.844	-1.792	**-0.520**	-2.448	-4.448
A	-8	-5.844	-3.792	-2.374	**-0.978**	-2.978
C	-10	-7.844	-5.792	-4.131	-2.774	**-1.318**

- Uses PWM in calculation of alignment score
- Allows for probabilistic mismatches and gaps
- Greater ability to map reads of variable confidence

Computational Sciences Laboratory
Brigham Young University

Assignment

- Given a read that has matches to possibly multiple locations in the genome, assign the read to locations where it matches
  - Repeat Masking– Discard reads that match to repeat regions.
    - Half of the human genome contains repeat regions, so you are not able to map to those regions
    - Many regulatory regions are repeated in the genome
  - Map to all locations – Repeat regions will be over-represented since one read will generate multiple hits
  - Pick a random location – Biased if there are small numbers of reads
- GNUMap uses probabilistic mapping to allocate a share of the read to matching locations in the genome according to the quality of the match

Equation for probabilistic mapping

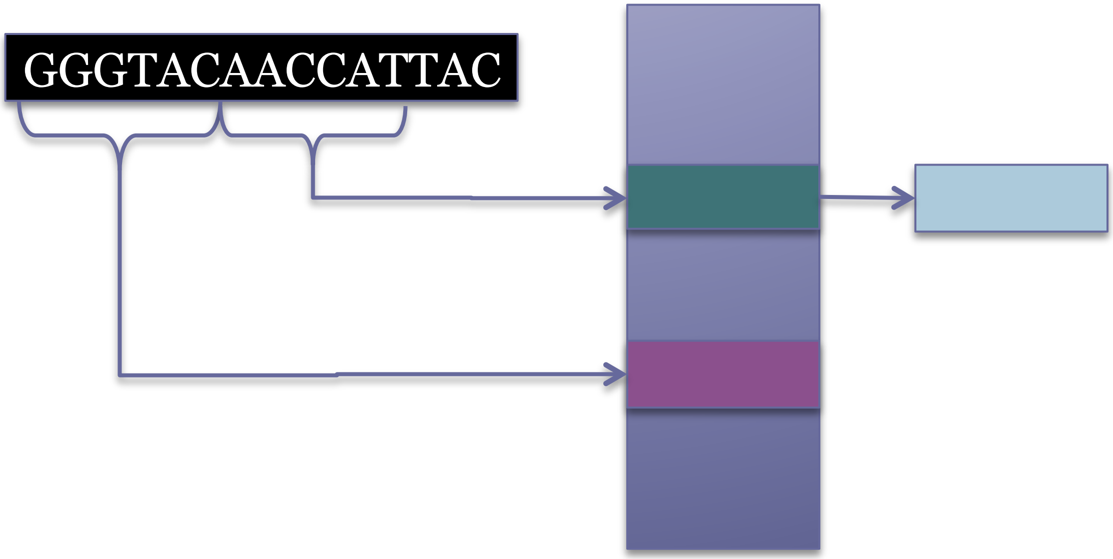
Computational Sciences Laboratory
Brigham Young University

$$G_{M_j} = \frac{Q_{M_j}}{n_{M_j} Q_{M_j} + \sum_{k \neq j}^{n} n_{M_k} Q_{M_k}}$$

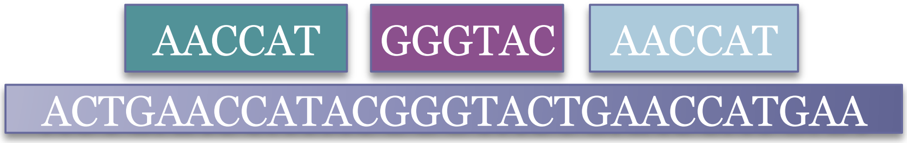
- Allows for multiple sequences of different matching quality.
- Includes probability of each read coming from any genomic position.

Alignment

Read from sequencer



Read is added to both repeat regions proportionally to their match quality

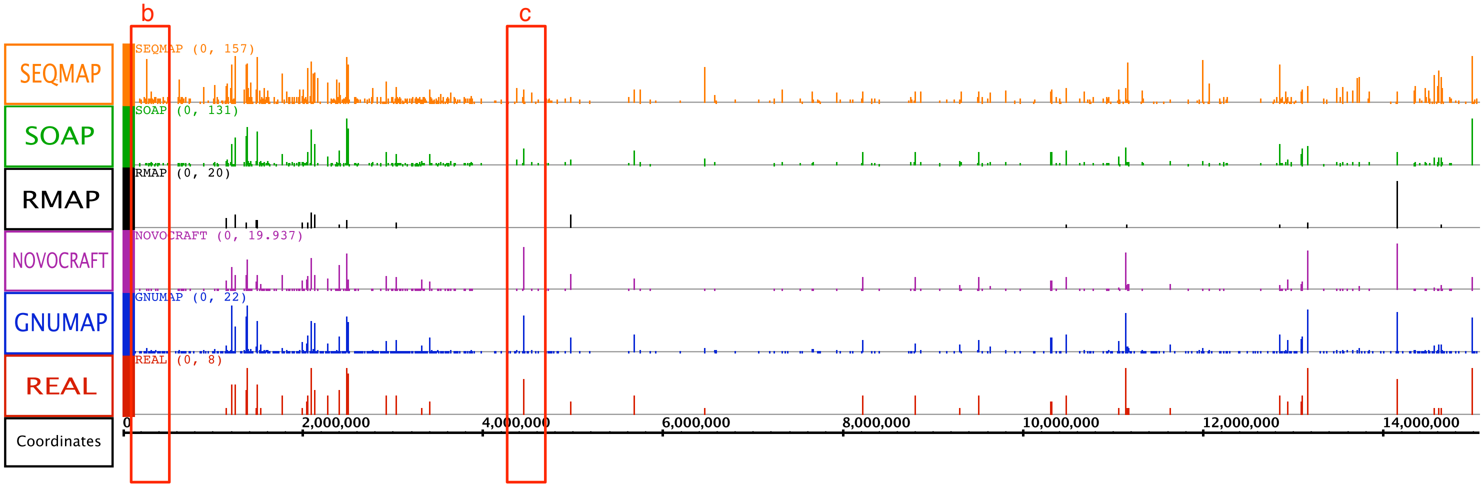


Which Program to Use?

Computational Sciences Laboratory
Brigham Young University

- Many different programs. How do they relate?
  - ELAND (included with Solexa 1G machine)
  - RMAP (Smith et al., BMC Bioinformatics 2008)
  - SOAP (Li et al., Bioinformatics 2008)
  - SeqMap (Jiang et al., Bioinformatics 2008)
  - Slider (Malhis et al., Bioinformatics 2008)
  - MAQ (Unpublished, http://maq.sourceforge.net/)
  - Novocraft (Unpublished, http://www.novocraft.com)
  - Zoom (Lin et al., Bioinformatics 2008)
  - Bowtie (Langmead et al., Genome Biology 2009)
  - …

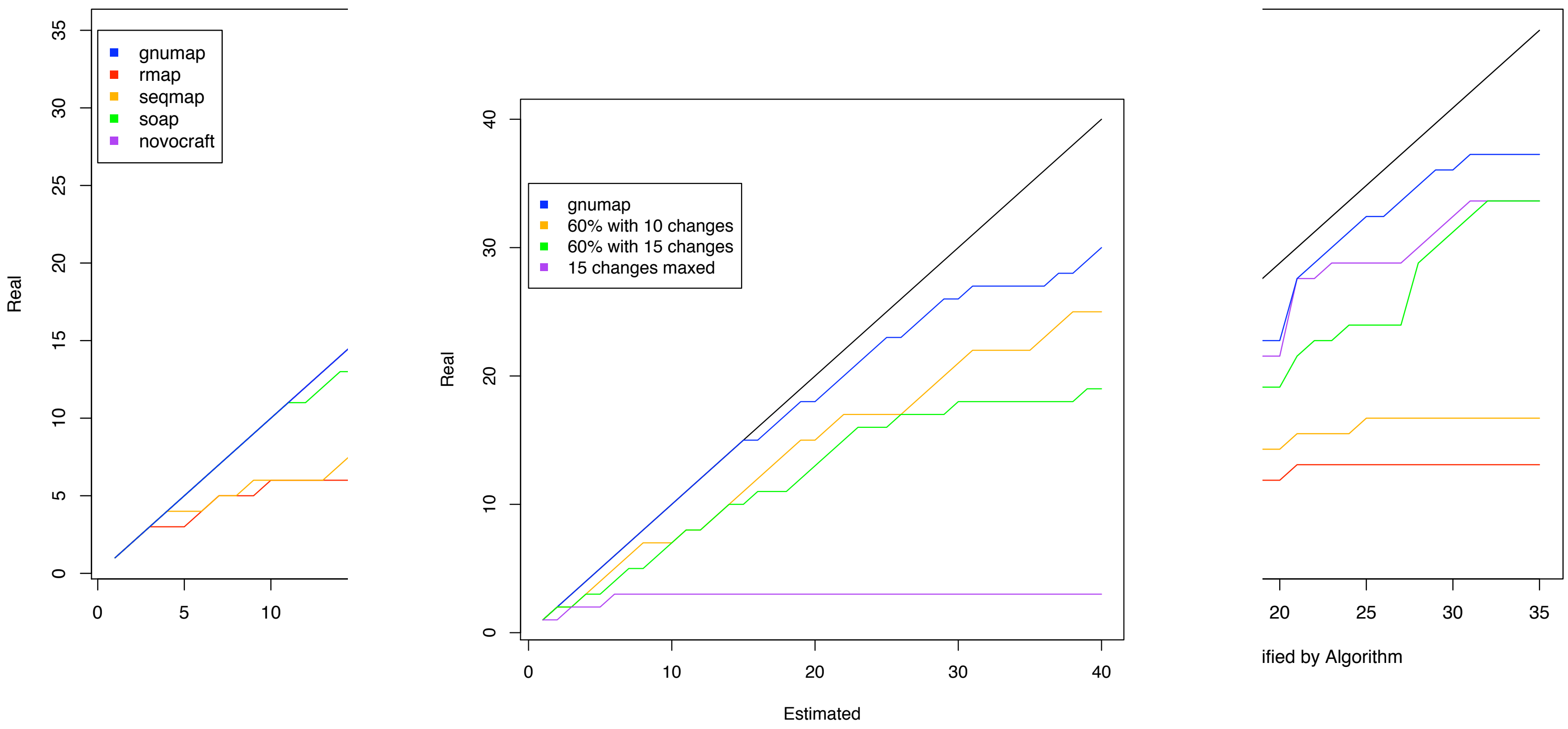
Simulation Studies



- Ambiguous reads cause:
  1. Missed (unmapped) regions
  2. Too many mapped regions (noise)

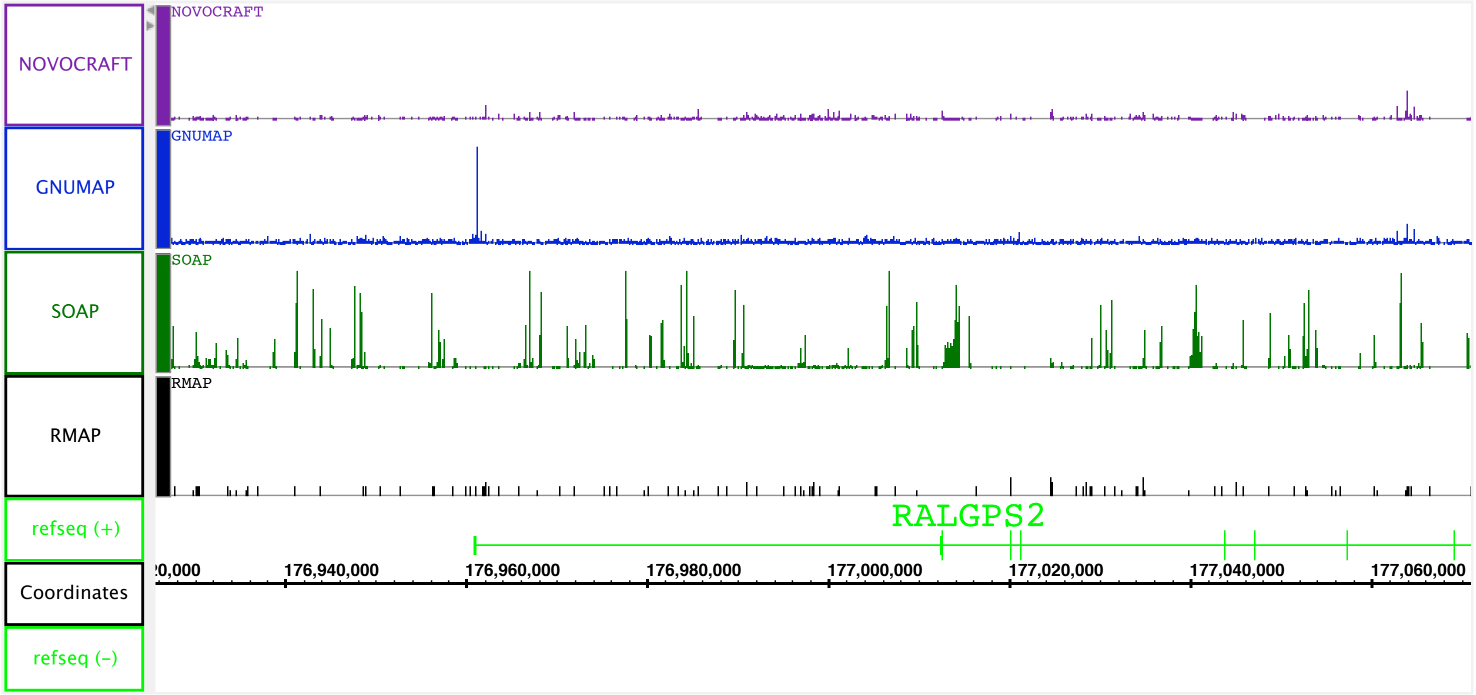
Computational Sciences Laboratory
Brigham Young University

Simulation Studies



Computational Sciences Laboratory
Brigham Young University

Actual Data



- ETS1 binding domain
- Repetitive region

Future Plans

Computational Sciences Laboratory
Brigham Young University

- Removal of adaptor sequences
- Methylation analysis
- Paired-end reads
- SOLiD color space

Acknowledgements

Evan Johnson
Quinn Snell
Mark Clement
Huntsman Cancer Institute