SHRiMP: Accurate Mapping of Short Color-space Reads

Stephen M. Rumble, Phil Lacroute, Adrian V. Dalca, Marc Fiume, Arend Sidow, Michael Brudno

Presenter: Billy Chang
CSC 2341
Feb 3, 2010

Features

SHRiMP features:

1. Both Color-Space and Letter-Space reads mapping.
2. Allows insertions and deletions.
3. Read mapping probabilities and statistics.

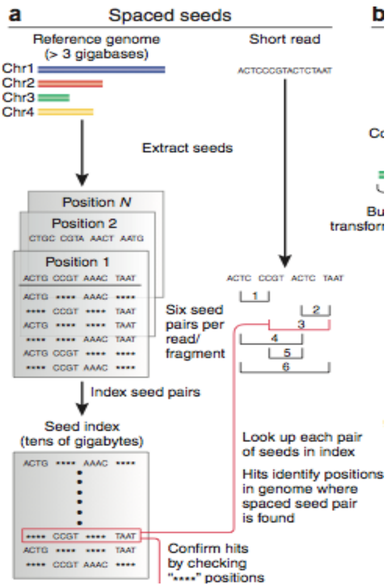
SHRiMP Pipeline

1. Spaced-seed matching
2. Smith-Waterman Algorithm for alignment scores.
3. Alignment probabilities and statistics calculation.

Seed Matching

- Classical approach (Seed and Extend):
  1. extract all k-mers from the reference genome.
  2. for each read, compare all its substrings with the k-mers in step 1.
  3. if a match is found, confirm the read alignment (e.g. by Smith-Waterman Algorithm).
- Problem: 4^k possible k-mers; Storage issues.

SHRiMP uses Spaced Seeds

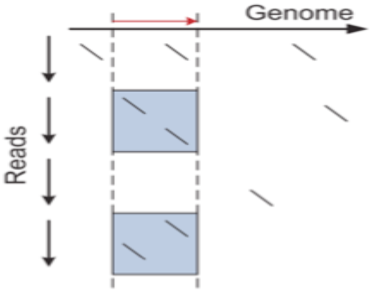


- Only match reads to genome at specified location.
- Spaced Seeds in Figure:
  1111000011110000
  0000111100001111
  1111000000001111
  0000000011111111
  1111111100000000
  0000111111110000
- 1 = must match, 0 = doesn't matter.

Trapnell et al (2009)

Before Alignment

- Spaced seed matching can happen by chance.
- Proceed only with reads that have number of seed matches higher than a specified threshold within a window in the genome.

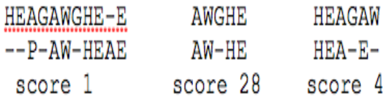


Original Smith-Waterman Algorithm

- Smith-Waterman Algorithm finds the best local alignment of two sequences, subject to a specified substitution matrix and a gap penalty.
- Example (from Durbin et al (2006)):

$$A = HEAGAWGHEE \qquad B = PAWHEAE$$

Using the BLOSUM50 substitution matrix and gap penalty = 8:

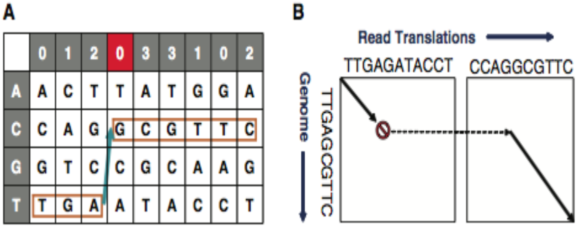


```
HEAGAWGHE-E    AWGHE    HEAGAW
--P-AW-HEAE    AW-HE    HEA-E-
 score 1      score 28  score 4
```

- Problem here: the original version only works on letter-space.

SHRiMP extends Smith-Waterman Algorithm to Color-Space

- Each color space read corresponds to 4 possible letter sequences.
- SHRiMP modifies the original Smith-Waterman Algorithm by also considering transition from one letter space to another during the search for the optimal alignment (with a penalty for space transition).



	0	1	2	0	3	3	1	0	2
A	A	C	T	T	A	T	G	G	A
C	C	A	G	G	C	G	T	T	C
G	G	T	C	C	G	C	A	A	G
T	T	G	A	A	T	A	C	C	T

Before Alignment

- The first run of Smith-Waterman Algorithm is only used to find max alignment scores (i.e. does not store traceback information).
- For each read, retain the (pre-specified) *n* top hits.
- Now run Smith-Waterman Algorithm again for these top hits with traceback to get alignments.

Read Statistics

- For a given read, want to know whether its alignments arise just by chance or are indeed generated by the genome.
- Two probabilities and a alignment score considered.

p_{chance}

p_{chance} gives the probability that an alignment, with number of substitutions and indels equal to the observed alignments, can be aligned to a random genome (equal base frequencies) of length g.

- *pchance*: For an observed alignment of length r:

$$p_{chance} = 1 - (1 - cf(r)\frac{Z}{4^r})^{2g}$$

- Z = # alignments of length r with the same numbers of substitution and indels as the observed alignment.
- $cf(r) = readsize - r + 1$ is a correction factor.

p_{genome}

p_{genome} is the probability that the alignment is generated by the genome, while allowing the observed number of substitution, indels, and errors.

- Given an observed alignment with n_ϵ errors, n_{sub} substitutions, and n_{indel}:

  $$p_{genome} = p_\epsilon p_{sub} p_{indel}$$

  Where evaluations of $p_\epsilon, p_{sub}, p_{indel}$ respectively involves estimated rates of errors, substitution, and indels.
- **Key difference between p_{chance} and p_{genome}:** p_{chance} assumes random reference genome; p_{genome} involves parameters estimated from the data.

Normalized Odds

For all the hits of a read, we have:

$$normodds_{hit} = \frac{pgenome_{hit}/pchance_{hit}}{\sum_{\forall hits} pgenome_{hit}/pchance_{hit}}$$

A hit with high $normodds_{hit}$ will potentially be the true location of alignment from the reference genome.

Real Data Experiment

- 135 million reads of length 35 bp from a single C. savignyi individual.
- Highly polymorphic; SNP heterozygosity 4.5%; even small reads can contain several variants.

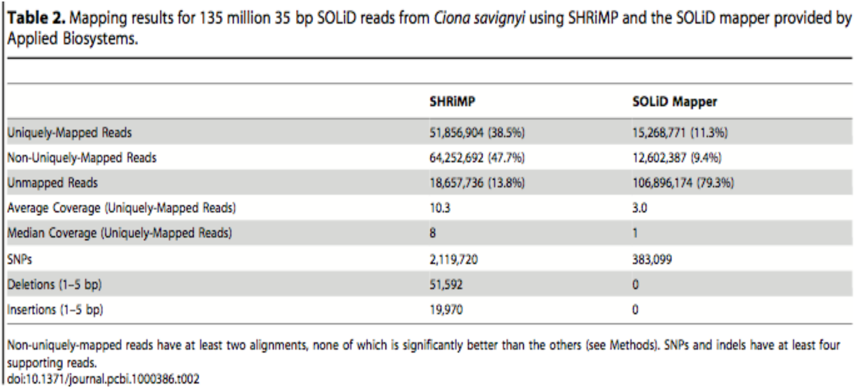


Table 2. Mapping results for 135 million 35 bp SOLiD reads from *Ciona savignyi* using SHRiMP and the SOLiD mapper provided by Applied Biosystems.

	SHRiMP	**SOLiD Mapper**
Uniquely-Mapped Reads	51,856,904 (38.5%)	15,268,771 (11.3%)
Non-Uniquely-Mapped Reads	64,252,692 (47.7%)	12,602,387 (9.4%)
Unmapped Reads	18,657,736 (13.8%)	106,896,174 (79.3%)
Average Coverage (Uniquely-Mapped Reads)	10.3	3.0
Median Coverage (Uniquely-Mapped Reads)	8	1
SNPs	2,119,720	383,099
Deletions (1–5 bp)	51,592	0
Insertions (1–5 bp)	19,970	0

Non-uniquely-mapped reads have at least two alignments, none of which is significantly better than the others (see Methods). SNPs and indels have at least four supporting reads.

doi:10.1371/journal.pcbi.1000386.t002

Simulation Studies

- Design: Introduce SNPs and indels to the C. savignyi genome at random location.
- Generate reads and add sequencing errors.
- Map the reads back to the original genome.

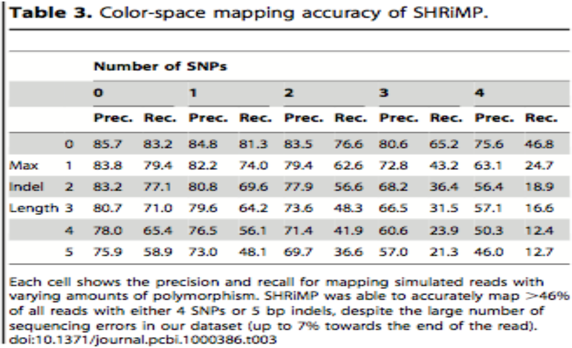


Table 3. Color-space mapping accuracy of SHRiMP.

		Number of SNPs									
		0		1		2		3		4	
		Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
	0	85.7	83.2	84.8	81.3	83.5	76.6	80.6	65.2	75.6	46.8
Max	1	83.8	79.4	82.2	74.0	79.4	62.6	72.8	43.2	63.1	24.7
Indel	2	83.2	77.1	80.8	69.6	77.9	56.6	68.2	36.4	56.4	18.9
Length	3	80.7	71.0	79.6	64.2	73.6	48.3	66.5	31.5	57.1	16.6
	4	78.0	65.4	76.5	56.1	71.4	41.9	60.6	23.9	50.3	12.4
	5	75.9	58.9	73.0	48.1	69.7	36.6	57.0	21.3	46.0	12.7

Each cell shows the precision and recall for mapping simulated reads with varying amounts of polymorphism. SHRiMP was able to accurately map >46% of all reads with either 4 SNPs or 5 bp indels, despite the large number of sequencing errors in our dataset (up to 7% towards the end of the read).
doi:10.1371/journal.pcbi.1000386.t003

- precision - the fraction of reads with correct top hit.
- recall - the fraction of all reads that had a unique, correct hit.

Conclusions

SHRiMP:

- is a color-space read mapper.
- provides alignment quality measures.
- achieves high sensitivity and specificity for SNPs and indels detections.
- can be slow.
- improving alignment quality measures by incorporating read qualities?

Further References

Durbin et. al. (2006) Biological Sequence Analysis, Eleventh Printing. Cambridge University Press.

Trapnell et. al. (2009) How to map billions of short reads onto genomes. Nature Biotechnology, 27:5, p455-457.