The Dependence of Computational Assignment of Putative SNPs on Choice of Alignment Tool

Deepti Anand, Lucie N. Hutchins, Joel H. Graber

The Jackson Laboratory , Bar Harbor, Maine 04609 USA

Abstract

Computational discovery of Single Nucleotide Polymorphism (SNP) involves aligning and scanning homologous sequences for mismatches. While this analysis is conceptually straightforward, we find that the results depend significantly on the choice of alignment tool. We used three different alignment tools (BLAT, BLAST and sssahaSNP) to compute candidate SNPs between two different strains of mice - *Mus musculus molossinus* (MSM) and C57BL/6J (B6), using 337,492 MSM sequences and B6 genome. The overlap amongst the three prediction sets was significantly worse than expected; therefore we repeated all three analyses in house, using a randomly selected subset of 10,000 MSM sequences. We report on the frequency and specific types of disagreement among the three analyses, and the benefits of merging multiple approaches for high confidence predictions.

Introduction

SNPs

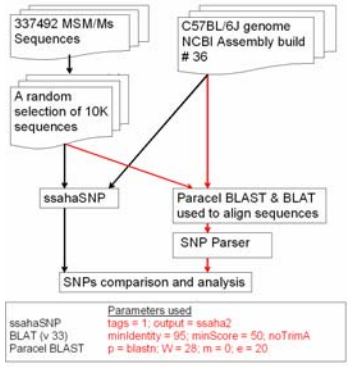
A Single nucleotide polymorphism (SNP) is a DNA sequence variation that occurs when a single nucleotide (A,T,C,or G) in the genome differs between members of a species. For example, a base pair variation between sequences from different strains of mice (MSM and C57BL/6J) could be a Putative SNP.

sample	genotype
1	...TACGTTTTAACGGGGATT...
2	...TACGTTTTAATGGGGATT...

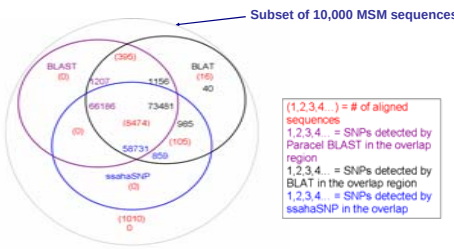
Background on Mouse Strains

- Inbred mice are generated through 20 or more generations of sister X brother matings.
- All members of an inbred "strain" have identical genomes, and within each member two copies of each autosome are identical to each other.
- Most research is performed on "laboratory strains" of mouse, which were derived from a limited number of "fancy mice" over the last ~80 years.
- The "mouse genome" was generated from the laboratory strain C57BL/6J (B6)
- More recently, new "wild-derived" strains have been generated to introduce new diversity into available inbred strains of mice, enabling expanded studies of evolution and systems biology.
- Comparison of sequences from the wild-derived strains with the reference mouse genome enables identification of putative SNPs.
- The strain MSM/Ms was derived from wild mice captured in Asia.

Methods



Alignment distribution with ssahaSNP, Paracel BLAST and BLAT



SNP Statistics for sequences aligned by ssahaSNP, Paracel BLAST and BLAT

Partial overlap between 3 alignments – 8360/ (8474)

Tools	SNP Count	SNP Count (150-300 bp)
SSAHA	57290	5422
BLAST	64435	9313
BLAT	72393	10571

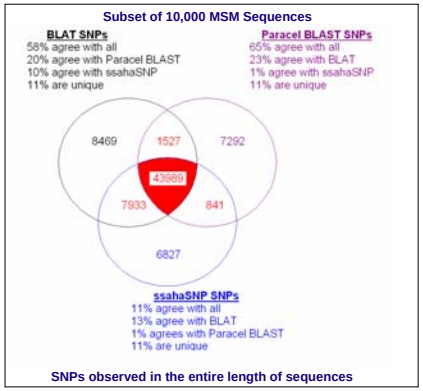
Partial overlap between 2 alignments – 102/ (8474)

Tool	SNP Count	SNP Count (150-300 bp)
SSAHA	1160	108
BLAST	1448	341
BLAT	994	190

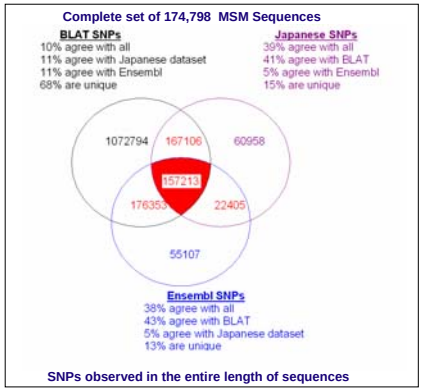
No overlap between alignments - 12/ (8474)

Tool	SNP Count	SNP Count (150-300 bp)
SSAHA	281	49
BLAST	303	83
BLAT	94	21

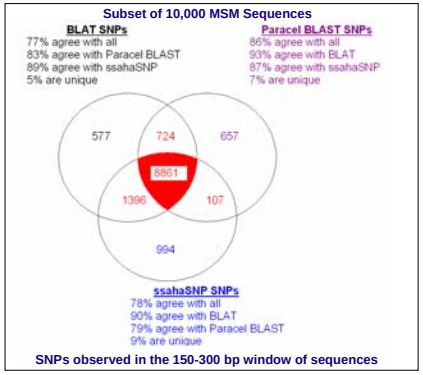
Putative SNPs in the subset and the complete set of MSM Sequences



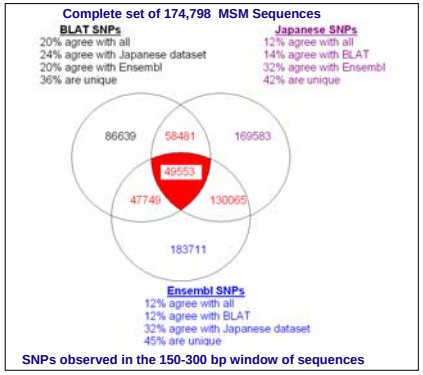
SNPs observed in the entire length of sequences



SNPs observed in the entire length of sequences

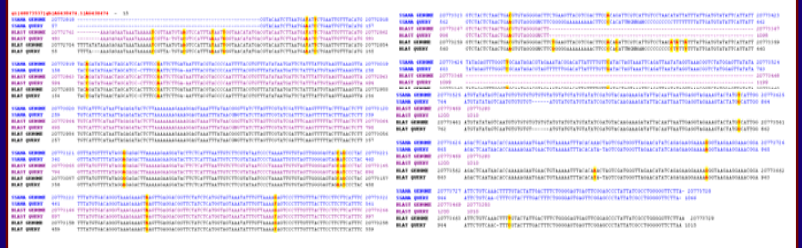


SNPs observed in the 150-300 bp window of sequences



SNPs observed in the 150-300 bp window of sequences

Sample output illustrating the alignments from ssahaSNP, Paracel BLAST and BLAT



Conclusion

1. Alignment tools like SSAHA, BLAST and BLAT use different scoring schemes and algorithms. Our study shows that these differences can lead to significant differences in the assignment of putative SNPs.
2. Nucleic Acid Alignment tools are built to align "near perfect" matches. Aligning related species (like the two strains used here, which are genetically diverse) challenges these assumptions and this calls for optimizing the parameters of these alignment tools, in order to fit one's requirements.
3. A good SNP set could be the overlap of these sets.