Bioinformatics pipeline for detection of immunogenic cancer mutations by high throughput mRNA sequencing*

Jorge Duitama[1], Ion Măndoiu[1], and Pramod K. Srivastava[2]

[1] Department of Computer Science & Engineering, University of Connecticut, 371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155, USA
E-mail: {jduitama,ion}@engr.uconn.edu
[2] Center for Immunotherapy of Cancer and Infectious Diseases, University of Connecticut Health Center, 263 Farmington Avenue, Farmington, CT 06030-1601, USA
E-mail: srivastava@uchc.edu

1 Introduction

Immunotherapy is a promising cancer treatment approach that relies on awakening the immune system to the presence of antigens associated with tumor cells. The success of this approach depends on the ability to reliably detect immunogenic cancer mutations, the vast majority of which are expected to be tumor-specific [6]. In this poster we present a bioinformatics pipeline for detecting immunogenic cancer mutations from high throughput mRNA sequencing data. Immunogenic mutations predicted by our pipeline from Illumina mRNA reads generated from a mouse cancer tumor cell line are currently under experimental validation.

2 Analysis pipeline

A schematic representation of our analysis pipeline is given in Figure 1. The pipeline consists of four main stages. First, sequencing reads are mapped separately against a reference transcript library (CCDS) and the reference genome using Maq [4]. Second, reads mapped by the two methods are merged as explained below. Third, merged reads are used to call SNPs, which in this context correspond to positions with enough evidence of the existence of an allele different than the reference. In the last stage, predicted SNPs and tested for immunogenic response using the SYFPEITHI database [5].

The goal of the first step is to map as accurately as possible reads obtained by sequencing mRNA. Mapping mRNA reads against the reference genome using standard mapping programs such as Maq does not require gene annotations but leaves reads spanning exon junctions unmapped. Since spliced alignment methods such as [1] are computationally intensive, we maximize the number of

* Work supported in part by NSF awards IIS-0546457 and DBI-0543365.

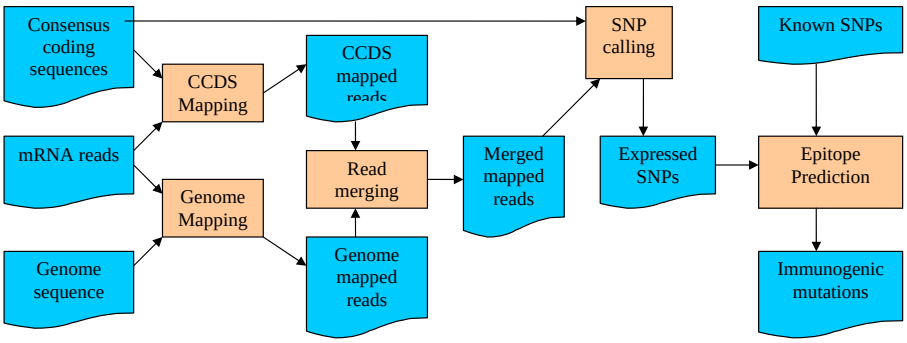

Fig. 1. Analysis pipeline to identify antigenic mutations from mRNA sequencing reads

accurately mapped reads by using Maq to map them both against the reference genome and CCDS transcripts.

For combining read mapping results we implemented two approaches called hard merging and soft merging. Hard merging throws away reads that are mapped uniquely by one procedure and to multiple places by the other while soft merging keeps the unique alignment for these reads. Both merging methods keep reads mapped uniquely to the same place by both mapping procedures and reads mapped by one procedure and not mapped by the other. Both methods throw away reads mapped uniquely to different places by both mapping procedures and reads mapped multiple times by both procedures.

To identify SNPs present in the sample we experimented with the SNP calling method implemented by Maq but found it to be too stringent and implemented two alternative methods. The first method, proposed in [3, 7] for calling SNPs from genomic DNA, uses a binomial test on the two highest allele counts under the null hypothesis that the genotype is heterozygous. The second method starts by calculating the conditional probability of observing the read data given each possible genotype. This probability is computed as a product of read contributions assuming independence between reads. Given a homozygous genotype xx, and a certain position, each base b mapped to this position contributes a term of $1-e_b$ if $x = b$ and $\frac{e_b}{3}$ otherwise. Here, $e_b = 10^{-q_b/10}$ is the probability of of error while sequencing base b, where q_b is its quality score. For a heterozygous genotype xy, each mapped base b contributes a term of $\frac{1-e_b}{2} + \frac{e_b}{6}$ if $x = b$ or $y = b$ and $\frac{e_b}{3}$ otherwise. Maq mapping probabilities are taken into account by raising the corresponding term to the probability that the read is mapped correctly at this location. The posterior probability of each genotype is then evaluated assuming uniform priors. A variant is called in this approach if the genotype with highest posterior probability is different than homozygous reference and exceeds a user specified threshold.

Mapping Method	Mapped Reads	Prob ≥ 0.1	Prob ≥ 0.9	Prob ≥ 0.95	Prob ≥ 0.99	Prob ≥ 0.999
Transcripts	3423706	103065	29719	20067	4165	2818
Genome	4365304	93760	25407	16762	2951	1894
Hard Merge	4557300	100487	28628	19113	3432	2240
Soft Merge	5309877	102781	29660	19949	3896	2627

Table 1. Mapped reads and SNPs called at different posterior probability thresholds from mouse tumor cell line mRNA reads.

For each identified non-synonymous SNP, reference and alternative aminoacid sequences are generated using CCDS transcript annotations. Both peptides are then tested by querying the SYFPEITHI database [5]. The final result of the analysis is the list of SNPs for which the mutated peptide exceeds a binding affinity threshold while the wild-type peptide does not.

3 Results

We tested the performance of implemented methods on publicly available Illumina mRNA reads generated from blood cell tissue of Hapmap individual NA12878 [2] (NCBI SRA database accession number SRX000566). We included in evaluation Hapmap SNPs in known exons for which there was at least one mapped read by any method. A total of $22,364$ homozygous reference SNPs and $7,888$ heterozygous or homozygous non reference SNPs were considered. We defined as true positive a correctly called heterozygous or non-reference homozygous SNP. Conversely, we defined as false positive a called SNP for which NA12878 is homozygous reference according to Hapmap genotypes. Figure 2 gives the number of true positives against false positives for the implemented read mapping and SNP calling methods. SNP calling based on posterior probabilities dominates the other methods except at very low false positive rates. Among the mapping strategies, the two approaches that merge genome and transcript mapping results performed slightly better. The number of detected SNPs monotonically increases with the number of reads within the tested sequencing depth range, showing no signs of saturation up to 22 million mapped mRNA reads.

We also ran our analysis pipeline on a set of 6.75 million Illumina reads from mRNA isolated from a mouse cancer tumor cell line. The number of mapped reads and identified SNPs at different posterior probability thresholds are given in Table 1. A total of 15 identified SNPs are currently under experimental validation.

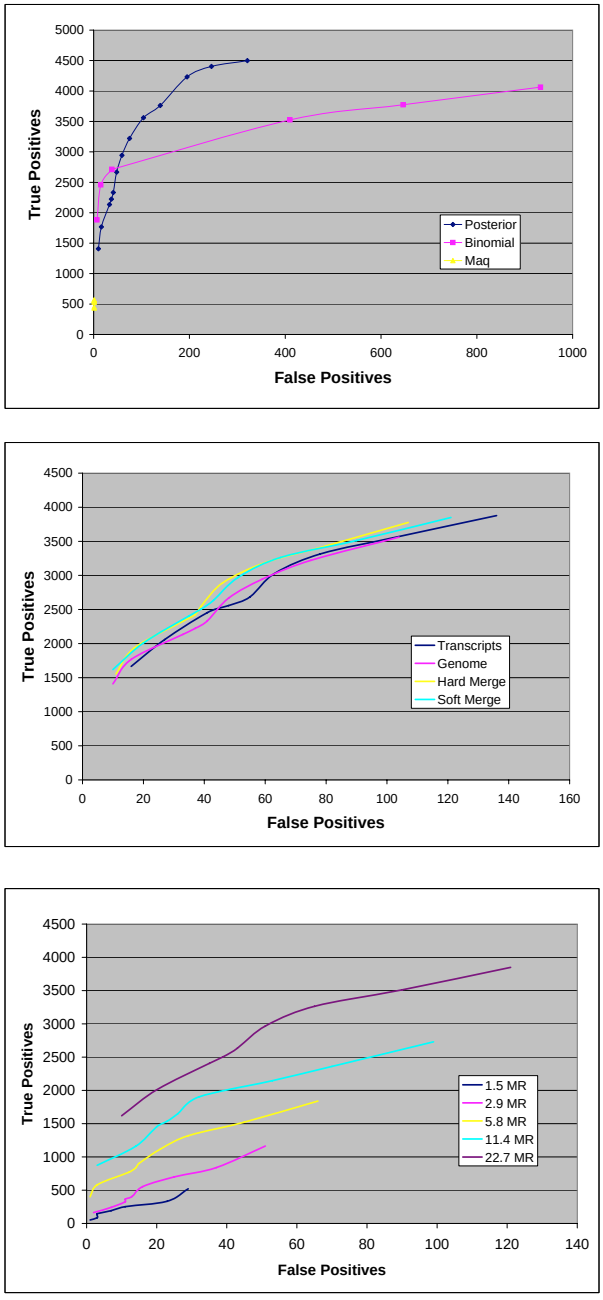

Fig. 2. True positives against false positives for three different SNP calling methods (top), four different read mapping strategies (middle), and varying mRNA sequencing depth (bottom).

References

1. F.D. Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):174–180, 2008.
2. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(18):851–861, 2007.
3. S. Levy *et al.* The diploid genome sequence of an individual human. *PLoS Biology*, 5(10):e254+, 2007.
4. H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(1):1851–1858, 2008.
5. H. Rammensee, J. Bachmann, N.P. Emmerich, O.A. Bachor, and S. Stevanovic. SYFPEITHI: database for mhc ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, 1999.
6. H. Rammensee, T. Weinschenk, C. Gouttefangeas, and S. Stevanovic. Towards patient-specific tumor antigen selection for vaccination. *Immunological Reviews*, 188(1):164–176, 2002.
7. D.A. Wheeler *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452:872–876, 2008.