ParMap, an Algorithm for the Identification of Complex Genomic Variations in Nextgen Sequencing Data

Hossein Khiabanian[1]*, Pieter Van Vlierberghe[2], Teresa Palomero[2,3], Adolfo A. Ferrando[2,3,4], Raul Rabadan[1]

[1] Department of Biomedical Informatics and Center for Computational Biology and Bioinformatics, [2] Institute for Cancer Genetics, [3] Departments of Pathology and [4] Pediatrics, Columbia University College of Physicians and Surgeons, New York, NY, United States

Running Title: ParMap, Identification of Genomic Variations in Nextgen Sequencing Data

Keywords: NextGen Sequencing, Genomic Variations, Genomic Indels

* Corresponding author. Mailing address: Columbia University College of Physicians and Surgeons, Center for Computational Biology and Bioinformatics, 1130 St. Nicholas Ave., New York, NY 10032, USA. Fax: 212-851-5149. Email: hossein@c2b2.columbia.edu.

Abstract

Next-generation sequencing produces high-throughput data, albeit with greater error and shorter reads than traditional Sanger sequencing methods. This complicates the detection of genomic variations, especially, small insertions and deletions. Here we describe ParMap, a statistical algorithm for the identification of complex genetic variants using partially mapped reads in nextgen sequencing data. We also report ParMap's successful application to the mutation analysis of chromosome X exome-captured leukemia DNA samples.

Introduction

One of the major technological advances in biology in the last few years has been the development of high throughput nextgen sequencing systems that produce gigabases of data in a single run, and allow an unbiased view of the whole genome without relying on prior knowledge about the disease-causing alterations. These ultradeep sequencing technologies produce large amounts of sequence data, which increase the sequencing depth and allow for better statistics in calling various genomic variations. However, they do so at the cost of reducing the read length and increasing the error rate relative to traditional Sanger sequencing. Thus, the development of efficient statistical and computational methods for the high confidence call of genomic variants is needed for the analysis of these high throughput datasets.

At this point, the detection of single mutations and large copy number variations using deep sequencing data is fairly straight forward (Shendure and Ji 2008; McPherson 2009), whereas the identification of small (less than 10 nucleotides) insertions and deletions is more challenging. A few algorithms have been developed for detecting such complex genomic variants using mate-pair or paired-end reads (Medvedev et al. 2009), however, identifying small insertion/deletions in fragment (single-end) data has proved to be very difficult. Mapping algorithms that are designed for very short reads have to assign large penalties for introducing gaps in the middle of the alignment in order to map the majority of the reads efficiently. However, these methods can partially map reads to a reference genome with gaps at the either end, without significantly reducing the alignment score. Although these gaps may be caused by systematic errors in the sequencing and mapping processes, we hypothesized that gaps that appear in multiple reads at a given position on the genome may reflect the presence of a complex genomic variant (e.g. insertion, deletions, multiple base changes).

Following this principle, we aimed to develop a procedure for identifying complex genomic variations with high confidence and built an algorithm (ParMap) capable of producing a list of candidates of small deletions and insertions (along with their nucleotide sequence), through statistical analysis of partially mapped reads (Fig. 1). Specifically, ParMap calculates a measure based on the number of reads that only cover the positions adjacent to a gap without covering their neighboring positions in the direction of the gap, to identify the possible locations of genomic insertions or deletions (Fig. 2 and Methods).

Results and Discussion

To test the ability of this method to detect novel complex genomic variants we applied ParMap to the analysis of SOLiD 3 chromosome X exome sequencing data from 12 T-cell acute lymphoblastic leukemia (T-ALL) DNA samples (P Van Vlierberghe, T Palomero, H Khiabanian, J Van der Meulen, M Castillo, N Van Roy, B De Moerloose, J Philippé, T Taghon, L Zuurbier, et al., submitted). In this experiment leukemia DNA samples were fragmented and ligated to adapters to generate SOLiD sequencing libraries, which were amplified and subsequently enriched in chromosome X exonic sequences usingthe SureSelect Target Enrichment System (Gnirke 2009), a platform which targets 5,217 exonic regions encompassing 3,045,708 nucleotides in the X chromosome. The Chromosome X exome captured DNA samples were sequenced with the Applied

Biosystems SOLiD 3 platform using 1/8th of sequencing slide per sample to produced a total of 105,302,787 fifty-base long fragment reads. The SOLiD platform employs a ligation based chemistry and a two-base encoding system, where each pair of nucleotides is reported with a different color, depending on the first base within the pair. Therefore, to call a single-base change relative to the reference sequence in nucleotide-space, two consecutive color-space mismatches must be observed. Single color-space mismatches solely report errors in the reads (Homer 2009).

To ensure an optimum mapping of these sequencing results, we created a reference sequence containing all chromosome X capture targeted regions plus adjacent 50 flanking bases using the March 2006 human reference sequence assembly (hg18). We used the SHRiMP algorithm with its default parameters for mapping the reads (Rumble 2009). For further analysis of our dataset, we only included the reads with a maximum number of two color-space mismatches that were uniquely mapped to the reference genome (approximately 31% of the raw reads). An average 90.1% of the reference genome was found covered in the samples, with a mean depth of 42 per base. Less restrictive filtering increased the false positive rate of candidate genomic variants without a significant increase in the coverage.

We created a candidate list of single-base variants for which a minimum of 3 reads (consistent with 1% estimated error rate of this particular run) should map to the candidate’s position, with more than 75% of them calling the nucleotide change. T-ALL samples such as the ones analyzed in this series contain over 80% tumor cells, however a small fraction of contaminating normal cells is expected. Because of the possibility of this contamination, we did not enforce a 100% consensus among the reads. To identify genetic alterations with the most direct impact in gene function we focused on the analysis of non-synonymous changes in the captured exons.

In this analysis we noticed numerous systematic errors that cannot be corrected by increasing the sequencing depth, i.e. genomic variants that are reported systematically beyond the statistical expectations from the estimated error rates. These systematic errors arise from pre-sequencing operations, ligation-based sequencing, mapping, and specific genomic variants in the reference genome. To minimize the number of such systematic errors, we combined the candidate lists from all the samples and only kept the genomic variations that occur in less than 3 samples at a given position. Within the 12 samples analyzed, we identified 66 exonic non-synonymous single-base variant candidates, which were not listed as already known polymorphisms in the human genome (Kuhn 2009). Overall, 61/66 (92%) of these candidates were confirmed via Sanger sequencing (Table 1).

Next, we applied the ParMap algorithm to our dataset to identify possible complex variants such as small insertions and deletions. Following on the selection of candidate variants using ParMap filtering criteria, we selected the ones that were detected in a single sample each. In this analysis, we found a high prevalence of systematic errors in intron-exon boundaries, which may reflect impaired ligation-based sequencing in these positions. Therefore, candidate variants located in intron-exon boundaries were discarded and excluded from further analysis. ParMap identified a total of 7 candidate complex variants (Table 1). Using Sanger sequencing of PCR products encompassing these sequences, we confirmed four indels in four different samples, including two genomic deletions of 3 and 6 nucleotides and two genomic insertions of 5 and 3 nucleotides. Notably, the genomic sequences identified in each of these two insertions matched the predicted sequence variant in the ParMap's results.

In conclusion, we have demonstrated the successful identification of high confidence genomic variants in nextgen sequencing data using a combination of single nucleotide analysis and ParMap. Overall, 89% of our all candidate variants were experimentally validated in this series. ParMap may enhance the identification of elusive complex genetic variants such as small insertions and deletions in nextgen sequencing data, taking advantage of partially mapped reads that might otherwise be discarded.

Methods

In addition to the completely mapped reads and reads reporting single-base changes, the dataset includes partially mapped reads, with the unmatched positions marked as gaps. These reads either start or end with a gap region that is as long as 20% of the length of the read (Fig. 1). ParMap makes use of such reads and for any position p that is adjacent to a gap region and is not the starting or ending position of an exon, calculates the following quantities:

1. ● $N(p)$: The number of reads that cover position p.

2. ● *N*(*p*±1): The number of reads that cover position *p*±1. (Plus, if *p* is the position after the gap and minus, if *p* is the position before the gap, in the direction of the positive strand.)
3. ● *N*(*p* & *p*±1): The number of reads that cover both positions of *p* and *p*±1.

We define

$$r = \frac{N(p \& p \pm 1)}{N(p) + N(p \pm 1) - N(p \& p \pm 1)},$$

which is an inverse measure of the number of reads that only cover the position *p* without covering its neighboring position in the direction of the gap (Fig. 2). Therefore, the smaller the value of *r*, the higher the chance for the gap to be due to a real change in the sequenced genome. Moreover, because the reads in which position *p* is adjacent to a gap region are already collected, referring back to each read prior to the mapping, the genomic sequence of the gap region can be extracted.

We apply the following criteria to produce a list of candidates: the value of *r* should be less that 0.35 and at least 5 reads should map to *p* adjacent to a gap, reporting a consensus sequence for it. To reduce the systematic errors due to mapping artifacts, we remove the candidates whose gap regions cover the already known polymorphisms of the human genome. We experimentally observed that less restrictive criteria increased the number of false positives.

Acknowledgments

This work was supported by awards from the Northeast Biodefence Center (U54-AI057158 to R.R.), the National Library of Medicine (1R01LM010140-01 to R.R.), the National Institutes of Health (R01CA120196 and R01CA129382 to A.F.); the Rally Across America Foundation (A.F); the Swim Across America Foundation (A.F.) and the Golfers Against Cancer Foundation (A.F.). P.V.V. is supported by a postdoc grant from the Fund for Scientific Research (FWO) Flanders. T.P. is a recipient of a Young Investigator Award from the Alex's Lemonade Stand Foundation, and A.F. is a Leukemia & Lymphoma Society Scholar. We are especially grateful to Vladimir Trifonov for many insightful discussions.

Figure Legends

Figure 1: ParMap employs reads that are partially mapped to a reference genome to identify genomic variations. These variations include small insertions and deletions of less than 10 nucleotides. When the sequenced read (line 2) is mapped to the reference genome (line 1), the unmatched bases are marked as gaps (line 3), adjacent to position *p* (Methods).

Figure 2: A measure based on the number of reads that only cover position *p* without covering its neighboring position (*p*±1) in the direction of the gap is calculated. In other words, we find the ratio of the intersection (orange area) and the union (yellow and red areas) of the two sets of reads that cover *p* or *p*±1 (Fig. 1). Here, *N*(*p*), *N*(*p*±1), and *N*(*p* & *p*±1) are the number of reads that cover position *p,* position *p*±1, and both positions, respectively (Methods).

Figures

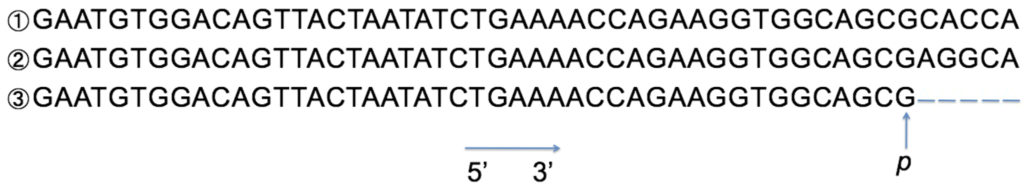
①GAATGTGGACAGTTACTAATATCTGAAAACCAGAAGGTGGCAGCGCACCA
②GAATGTGGACAGTTACTAATATCTGAAAACCAGAAGGTGGCAGCGAGGCA
③GAATGTGGACAGTTACTAATATCTGAAAACCAGAAGGTGGCAGCG
5'
3'
p

Figure 1: ParMap employs reads that are partially mapped to a reference genome to identify genomic variations.

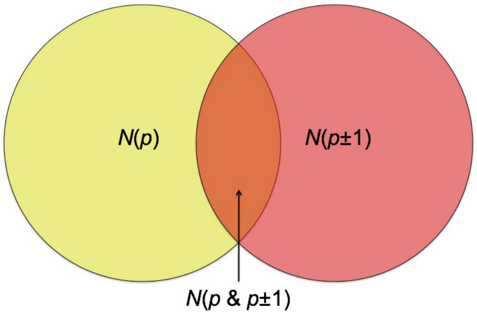

Figure 2: A measure based on the number of reads that only cover the position *p* without covering its neighboring position (*p*±1) in the direction of the gap is calculated.

Tables

Table 1: Summary of the efficiency of our identification methods. In total, 89% of our candidates were confirmed via Sanger sequencing.

	Number of Genomic Variation Candidates	Number of Confirmed Candidates	Percentage
Singe-base Analysis	66	61	**92%**
ParMap	7	4	**57%**
Total	73	65	**89%**

References

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**: 1135-45.

McPherson JD. 2009. Next-generation gap. *Nat. Methods* **6**: S2-S5.

Medvedev P, Stanciu M, Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6**: S13-S20.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**: 182-189.

Homer N, Merriman B, Nelson SF. 2009. Local alignment of two-base encoded DNA sequence. *BMC Bioinformatics* **10:**175.

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. 2009. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5**: e1000386.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al. 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.* **37**: D755-D761.