Identification of transposon insertion polymorphisms by computational comparative analysis of next generation personal genome data

Xuemei Luo[∗,†], Frank Dehne[∗] and Ping Liang[†,∗∗]

[∗]*School of Computer Science, Carleton University, Ottawa, K1S 5B6, Canada*
[†]*Department of Biological Sciences, Brock University, St. Catharines, L2S 3A1, Canada*
[∗∗]*Corresponding author, Email: pliang@brocku.ca*

Abstract. Structural variations (SVs) in a genome are now known as a prominent and important type of genetic variation. Among all types of SVs, the identification of transposon insertion polymorphisms (TIPs) is more challenging due to the highly repetitive nature of transposon sequences. We developed a computational method, TIP-finder, to identify TIPs through analysis of next generation personal genome data and their extremely large copy numbers. We tested the efficiency of TIP-finder with simulated data and are able to detect about 88% of TIPs with precision of $\geq$91%. Using TIP-finder to analyze the Solexa pair-end sequence data at deep coverage for six genomes representing two trio families, we identified a total of 5569 TIPs, consisting of 4881, 456, 91, and 141 insertions from Alu, L1, SVA and HERV, respectively, representing the most comprehensive analysis of such type of genetic variation.

INTRODUCTION

Structural variations (SVs) in a genome are defined as DNA sequence alternations among individuals including deletion, duplication, insertion, inversion, translocation and transposition of DNA sequences [1, 2]. In recent years, with the development of new sequencing technologies, a vast amount of high throughput sequencing personal genome data has been generated, providing new opportunities for comprehensive analysis of SVs. A number of computational tools have been developed for identification of SVs using these personal genome data [3], mostly without transposon insertions. The identification of transposon insertions is more challenging than other SVs due to the highly repetitive nature of transposon sequences.

Transposons or transposable elements (TEs) are discrete pieces of DNA that can move within a genome. TEs, with several millions of copies classified into many families and subfamilies, account for approximately 45% of the human genome, and they play important roles in the evolution of the genome and regulating gene functions [4]. Transposon insertion polymorphism (TIP) refers to the presence or absence of a transposon insertion at a specific genomic location in populations of a given species. Recent evidence indicates that about 35 to 40 subfamilies of Alu, L1, SVA elements and possibly HERV-K elements remain actively mobile in the human genome [5]. However, only a limited number of TIPs were identified using classical molecular biology techniques, such as locus-specific polymerase chain reactions (PCR), targeted mutation screening, and transposon differential display PCR [6, 7, 8, 9]. With the availability of personal genome data at an unprecedented scale, computational approaches are emerging as valuable tools for the study of TIPs [10, 11].

In this study, we developed a computational method, TIP-finder, to identify TIPs through analysis of next generation personal genome data. TIP-finder uses a greedy algorithm to identify the candidate TIPs loci which are then enhanced by a machine-learning approach.

METHODS

Identification of TIPs_OUT candidate loci

TIP-finder detects TIPs based on pair-end mapping (PEM), which involves the generation of paired end reads that represent the two short sequencing reads of the two ends of a genomic fragment with a known estimated size, i.e. the library size. Alignments of paired-end reads to the reference genome are categorized as concordant and discordant [2]. Paired reads mapped to the reference genome at a distance similar to the insert size and in a correct orientation represent a concordant PEM, otherwise represent a discordant PEM, which indicates a potential SV. In this study, we focus on TIPs representing insertions that are present in the test genome but absent in the reference genome (TIPs_OUT hereafter). For pair-end reads with fragments spanning the boundaries of a TE insertion, one read-mate is expected to map to the regions flanking the insertion with relatively good reliability, while the other read-mate that falls into the transposon insertion will have a random match to a similar transposon sequence in the reference genome. The alignment location of this transposon read is a random pick among all top hits. Sequence alignment data used in this study were obtained using the MAQ software [12]. MAQ labels reads having correct mapping with a flag of 18 (MF18 reads) and those mapped to two different chromosomes with a flag of 32 (MF32 reads). As there are 23 chromosomes in the genome, the possibility of the second read (in the transposon insertion) mapped to the same chromosome as the first read is on average at 4-5%. Therefore, most of these pair-end reads spanning the boundary regions of a TIPs_OUT would have the two reads mapping to different chromosomes (i.e. MF32 reads).

We developed a greedy algorithm to identify TIPs_OUT candidate loci. TIPs_OUT candidate reads are defined as paired reads with one read (A_i) in the flank region mapped to the reference genome in a non-transposon position and another (B_i) in newly inserted transposon loci. The location of transposons in the reference genome is based on the RepeatMasker annotation obtained from the UCSC Genome Browser (http://genome.ucsc.edu). We describe below the algorithms and pipelines included in our TIP-finder tool.

Algorithm 1 TIPs_OUTclusterFinding

```
(1) Map all paired-end reads of the test genome to the reference genome
(2) Identify TIPs_OUT candidate reads $(A_i, B_i)$
    (a) Identify and collect all paired MF32 reads.
    (b) Map all MF32 reads to a sorted transposon position list on the reference genome.
    (c) If $A_i$ is not mapped to the transposon positions but $B_i$ is, add $(A_i, B_i)$ into the TIPs_OUT candidate list $L$.
(3) Sort the list $L = [(A_1, B_1), (A_2, B_2), \cdots, (A_{n-1}, B_{n-1}), (A_n, B_n)]$ based on the mapping position of $A_i(PosA_i)$
(4) Cluster $(A_i, B_i)$ based on $PosA_i$. If $|PosA_1\text{-}PosA_2| < 2\times$LibrarySize and $B_1$ and $B_2$ are mapped to the same
    transposon family, group $(A_1, B_1)$ and $(A_2, B_2)$ into one cluster.
(5) Find the minimum (spos) and maximum (epos) positions of $PosA_i$ for each cluster.
```

TIPs_OUT loci filtering

In order to reduce false positives in predicting TIPs_OUT, the predicted TIPs_OUT candidate loci are filtered by imposing restrictions on the number of MF32 reads (numMF32), the ratio between numMF32 and the number of MF18 reads (numMF18), and the percentage of MF32 reads showing reliable mapping in the flank region of TIPs_OUT (percentReliableMF32). To optimize these filtering parameters, a machine learning approach was used based on a simulated genome containing known TIPs_OUT documented in dbRIP database [13].

Human somatic cells have a diploid genome containing two copies of each chromosome, so TIPs_OUT can exist in two ("+/+") or one copy ("+/-"). In simulating the known TIPs_OUT based on the dbRIP database, we inserted the 781 TIPs_OUT from dbRIP into the reference genome at their corresponding locations to generate a simulated diploid genome sequence with one copy containing all of these TIPs_OUT and the other copy containing randomly selected 50% of the TIPs_OUT. Therefore, 50% of these TIPs_OUT are in the "+/+" genotype, while the remaining 50% are in the "+/-" genotype. Based on the sequences of these two copies, we generated paired-end reads with a read length of 35bp (similar to an Illumina Solexa platform) at a size of 260 bp with 60 bp as the standard deviation (*SD*) using the sequence simulation utility included in the MAQ package [12]. These reads were then aligned to the reference genome (UCSC hg18) using MAQ. Different amounts of sequence data were used to simulate genome coverage at 5X, 11X, 22X, 26X, 32X and 46X. For the 781 TIPs_OUT inserted in the simulated genome, we divided them into two disjoint sets: dbRIP_setA containing 391 TIPs_OUT and dbRIP_setB containing 390 TIPs_OUT. dbRIP_setA is used to train our algorithms by identifying the data patterns surround the insertion sites, and it is also used to determine the optimal combinatorial parameters in filtering the noise. dbRIP_setB is used to assess the performance of our algorithms by using the filtering parameters generated from dbRIP_setA.

To optimize filtering parameters, at each known transposon insertion site of dbRIP_setA, we collected all MF32 reads and all MF18 reads within a distance of the library size to the insertion site. A percentReliableMF32 is computed based on flag H0 and H1 in the MAQ alignment data which are the number of perfect hits and hits with one difference, respectively. In this study, the mapping of a read is considered as unreliable under any one of the following conditions: 1)$H0 > 4$; 2)$H0 = 0$ and $H1 \geq 10$; 3)$H0 = 0$ and $H1 = 0$. TIPs_OUT are filtered out when percentReliableMF32 is less than 80%, which is the average of $mean - 3 \times SD$ for the training data set at different sequencing coverage. This excludes TIPs_OUT in the repetitive sequence regions which tend to produce a high level of false positive.

The genotype of a TE insertion can be predicted based on the ratio of numMF18/numMF32. By training and analyzing the simulated data at each known transposon insertion site for different genome coverage, the cut off values for defining the genotype are shown in Table 1, which are the means of the optimal cutoff values for different genome coverage. Since the numMF32 in the insertion with genotype "+/+" should be greater than that of genotype "+/-", different cut-off values for numMF32 should be set for genotypes "+/+" and "+/-". In this study, the cut-off values for minimum numMF32 were searched iteratively so that about 95% of the known insertion sites can be correctly identified. The regression functions were further generated to show the relationship between the minimal numMF32 and genome coverage for different genotypes. The final filtering parameters used in TIP-finder are summarized in Table 1.

TABLE 1. Summary of optimal filtering parameters used in TIP-finder

	Filtering criteria
percentReliableMF32	$percentReliableMF32 \geq 80\%$
Genotype "+/+"	$numMF18/numMF32 < 0.8$
Genotype "+/-"	$0.8 \leq numMF18/numMF32 \leq 3$
numMF32 for "+/+"	$numMF32 \geq 1.1 \times coverage + 2.7$
numMF32 for "+/-"	$numMF32 \geq 0.3 \times coverage + 5.8$

RESULTS

TIPs_OUT identification in the simulated genome

With the known insertion locations and the defined genotype, dbRIP_setB allowed us to provide an accurate assessment of the accuracy of TIP-finder in identifying TIPs_OUT in terms of both false positive and false negative rates, as well as the accuracy in predicting insertion genotype. Here, precision is calculated as percentage of true positive among all prediction, i.e. $(TP/(TP+FP))[\%]$, and sensitivity is calculated as percentage of detected true positive among all true TIPs_OUT, i.e. $(TP/(TP+FN))[\%]$, where TP, FN, and FP are the number of true positive, false negative, and false positive, respectively. As shown in Table 2, for genome coverage over 22X, our program is able to detect most of the known TIPs_OUT with a sensitivity $\geq$88% and precision $\geq$91%. At a genome coverage of 10X, TIP-finder is able to detect the known TIPs_OUT with sensitivity $\geq$87% and precision $\geq$93%. In comparison with >85% sensitivity and >90% precision for a simulated genome at coverage 10X reported by VariationHunter [11], the only other tool that handles the TIPs, TIP-finder provides slightly improved sensitivity and precision. The accuracy of genotype prediction is above 80% for all checked levels of genome coverage.

TABLE 2. Summary of TIPs_OUT predictions based on simulation

Genome coverage	**5X**	**10X**	**22X**	**26X**	**32X**	**46X**
TP	328	340	343	345	346	348
TP+FP	355	367	375	378	380	392
TP+FN	390	390	390	390	390	390
Precision	92%	93%	91%	91%	91%	89%
Sensitivity	84%	87%	88%	88%	89%	89%
Genotype match	80%	88%	88%	89%	90%	91%

TIPs_OUT prediction based on six genomes

To identify TIPs_OUT using TIP-finder, we downloaded genome data for the six individuals representing two trio families (each consisting the two parents and a child) that were subjected to deep sequencing by the 1000 Genome Project at a coverage from 24X to 36X (http://1000genomes.org). The specific samples are NA12878 (daughter, 34X), NA12891 (mother, 33X) and NA12892 (father, 28X) for a Utah Caucasian family and NA19238 (mother, 24X), NA19239 (father, 28X) and NA19240 (daughter, 36X) for a Nigerian family. We used the alignment data of sequences generated using the Illumina Solexa platform provided in BAM format by the 1000 genome project using MAQ.

The TIPs_OUT prediction was first performed for each genome individually and then the 6 lists of TIPs_OUT were combined to generate a non-redundant list of final TIPs_OUT candidate list (see Table 3). TIP-finder identified a total of 5569 TIPs_OUT consisting of 4881, 456, 91, and 141 insertions from Alu, L1, SVA and HERV, respectively. Among the 3 types of TEs that are known to be active in the human genome, Alu has the largest number of TIPs_OUT reflecting its highest level of transposition as expected. In addition to Alu, L1, and SVA, we identified 141 HERV TIPs_OUT, suggesting that they are still active contrasting to our current view. We compared the predicted TIPs_OUT with the 781 TIPs_OUT in dbRIP and found 453 (or 58%) overlapping entries. This is a reasonable number considering the small number of genomes covered in this study. As shown in Table 3, it is worth noting that the numbers of TIPs_OUT from the two families are dramatically different, with the Nigerian family containing much more TIPs_OUT than the Utah Caucasian family. This is expected because the reference genome has mainly a Caucasian origin, thus is more similar to the Utah families than to the Nigerian family. The data suggest that a large number of new TE insertions have occurred in the genomes of the African populations after the migration of ancestors of current non-African populations out of Africa. Experimental validation of TIPs_OUT prediction is underway.

TABLE 3. Summary of TIPs_OUT predictions based on the 6 genomes

Genome	NA19240	NA19238	NA19239	NA12878	NA12891	NA12892	Total(non-redundant)
SVA	65	25	45	8	5	2	91
L1	224	145	211	101	16	27	456
HERV	58	26	57	33	12	19	141
Alu	2379	1865	2220	799	174	188	4881

CONCLUSIONS

TIP is a type of structural variation difficult to analyze. This study demonstrates the high efficiency of TIP-finder for the computational identification of TIPs and represents one of the few comprehensive analyses of TIPs performed so far. Despite the limited number of genomes analyzed, our data revealed an unexpectedly high level of transposon associated genetic polymorphisms in humans, and thus we expect to identify much more of such genetic variation by analyzing additional personal genome data, allowing us to explore the contribution of such genetic diversity to phenotype variations

REFERENCES

1. E. Eichler, D. Nickerson, and et al., *Nature* **447**, 161 – 165 (2007).
2. E. Tuzun, A. Sharp, and et al., *Nature genetics* **37**, 727 – 732 (2005).
3. P. Medvedev, M. Stanciu, and M. Brudno, *Nature methods* **6**, S13 – 20 (2008).
4. R. Cordaux, and M. Batzer, *Nat Rev Genet* **10**, 691 – 703 (2009).
5. R. Mills, E. Bennett, R. Iskow, and S. Devine, *Trends Genet* **23**, 183 – 191 (2007).
6. J. Xing, Y. Zhang, and et al., *Genome research* **19**, 1516 – 1526 (2009).
7. S. Arcot, T. Shaikh, and et al., *Gene* **163**, 273 – 278 (1995).
8. M. Batzer, S. Arcot, and et al., *Journal of molecular evolution* **42**, 22 – 29 (1996).
9. A. Roy, M. Carroll, and et al., *Genetica* **107**, 149 – 161 (1999).
10. J. Wang, L. Song, and et al., *Gene* **365**, 11 – 20 (2006).
11. F. Hormozdiari, I. Hajirasouliha, and et al., *Bioinformatics* **26**, i350 – 357 (2010).
12. H. Li, J. Ruan, and R. Durbin, *Genome research* **18**, 1851 – 1858 (2008).
13. J. Wang, L. Song, and et al., *Human mutation* **27**, 323 – 329 (2006).