UPS-indel: A better approach for finding indel redundancy

Mohammad Shabbir Hasan
Department of Computer Science
Virginia Tech
Blacksburg, Virginia 24061
Email: shabbir5@cs.vt.edu

Xiaowei Wu
Department of Computer Statistics
Virginia Tech
Blacksburg, Virginia 24061
Email: xwwu@vt.edu

Layne T. Watson
Department of Computer Science,
Department of Mathematics,
Department of Aerospace and Ocean Engineering
Virginia Tech
Blacksburg, Virginia 24061
Email: ltw@cs.vt.edu

Zhiyi Li
Department of Computer Science
Virginia Tech
Blacksburg, Virginia 24061
Email: zli04@cs.vt.edu

Liqing Zhang
Department of Computer Science
Virginia Tech
Blacksburg, Virginia 24061
Email: lqzhang@cs.vt.edu

***Abstract*—Indel which represents the insertion and deletion of base pairs in the sequence of an organism is a very common form of genetic variation that takes place in the human genome. Being responsible for genetic diversity and human disease, indels have been considered as an important area in the genome research community. With progress in Next Generation Sequencing (NGS), a good number of indel calling tools have been developed and different databases store the results of different indel calling tools for future research. Different indels, though differing in allele sequence and position, can be biologically equivalent when they lead to the same altered sequences. Storing these biologically equivalent indels as distinct entries in databases causes data redundancy. Previous research showed that about 10% human indels stored in dbSNP are redundant due to lack of a unified system for identifying and representing equivalent indels. In this paper we describe UPS-indel, a utility tool that creates a universal positioning system for indels so that equivalent indels can be identified easily by a simple comparison of their coordinates generated by the proposed positioning system. Applying UPS-indel, we identify nearly 15% redundant indels in dbSNP (version 142) across all human chromosomes, higher than the previous report. UPS-indel is written in C++ and is freely available at `http://bench.cs.vt.edu/ups-indel`.**

***Keywords*—Indel, Next Generation Sequencing, redundancy.**

I. Introduction

Rapid progress in Next Generation Sequencing (NGS) has created the opportunity to develop numerous indel calling tools. Results obtained from these tools are stored in databases such as dbSNP and Ensembl for future research. Due to the lack of tools for easy and accurate systematic comparison of indels to determine equivalence, these databases store biologically equivalent indels as distinct entries which causes data redundancy. Previous research showed that About 10% of the human indels stored in dbSNP are redundant [1].

This paper described UPS-indel, a user friendly utility tool that converts the positions of all indels stored in a VCF [2] file to our newly proposed UPS-coordinate. UPS-coordinate represents all possible equivalent positions of an indel, which enables the comparison among different sets of indels for redundancies in an easy and convenient way.

II. Methods

UPS-indel is written in C++ and can be executed on Linux, Windows, or Mac operating systems. The input is a reference chromosome sequence, a VCF file containing a list of indels, and an output file name. For example, `./ups_indel ref.fa in.vcf out` produces an output file named `out.uvcf` containing the UPS-coordinates of all the indels listed in the input `in.vcf` file. The UPS-coordinate contains a range of positions in the square brackets which represents the region of equivalence for that indel. Indels containing the same UPS-coordinate are equivalent indels. For example, `./ups_generate_redundant_indel_list in.uvcf redundant_indel_list.txt` produces a list of indel groups containing Ids of redundant indels.

III. Results

We tested UPS-indel on dbSNP (version 142) which contains about 8.9 million indels from human genome. Our result shows that UPS-indel identifies nearly 15% of the indels as redundant which is higher the percentage (10%) reported by Vindel.

References

[1] Z. Li, X. Wu, B. He and L. Zhang, "Vindel: a simple pipeline for checking indel redundancy," *BMC Bioinformatics*, vol. 15, no. 1, p. 1, 2014.

[2] P. Danecek, A. Auton, G. Abecasis, C. Albers, E. Banks, M. DePristo, R. Handsaker, G. Lunter, G. Marth, S. Sherry, G. McVean and R. Durbin, "The variant call format and VCFtools,", *Bioinformatics*, vol. 27, no. 15, p. 2156-2158, 2011.

978-1-5090-4199-2/16/$31.00