Computational Methods for High-Throughput Omics Data - Genomics

Seminar Vorbesprechung
18.10.2011
Brief history of DNA sequencing

- Sanger sequencing method developed in early 1970s by Frederick Sanger
- First complete genome: that of a bacteriophage T4 in 1977
- Late 1980s: first automated Sanger sequencing machines
- 1990s: sequencing of small genomes and expressed sequence tags (ESTs)
- 2001: First draft sequence of the human genome published
- Mid 2000s: so-called Next Generation Sequencing (NGS) technologies emerge, e.g. 454 Life Sciences pyrosequencing, Illumina sequencing
- 2008: The 1000 Genomes Project launched, cost approaching $1 per base

**Problem:** NGS produces much shorter reads than Sanger sequencing (50-400bp compared to >1000bp) → mostly reference-guided analyses
Next Generation Sequencing (NGS) has many applications

**RNA-Seq:**
- Find out which genes are expressed in a certain tissue at a certain timepoint
- Quantify gene expression
- Identify alternative splicing events, i.e. Transcript isoforms

**DNA-Seq:**
- Identify differences between individuals/cell types, e.g. Cancer vs. healthy cells
- SNPs, short indels, large indels, inversions, translocations, copy number variations

| genome | AGACTGTCTA CAATCTT-G |
| reads  | GACTGTCTAG ATCTT-GGCT |
| reads  | TGTCTAGCAGA CTT-GGCTTC |
| reads  | GTCTAGCAAT CTAGCAATCA |
| reads  | TATAACCTA-G |

Exon 1  Exon 2  Exon 3
Papers:

Read Mapping:
- RazerS: general Edit/Hamming distance read mapper
- GSNAP: SNP-tolerant (split-)read mapper
- AGE: a dynamic programming algorithm for structural variants

Variant Detection:
- SRMA: refinement of read alignments for SNP/indel detection
- SRiC: Using split-reads for structural variant detection
- VariationHunter: Combinatorial algorithms for structural variant detection with paired-end reads

Applied Papers (incl. methods):
- BreakSeq: classification of variants and their formation mechanisms
- ChimeraScan: detecting gene fusion events in cancer transcriptomes
Read Mapping I

- General purpose read mapper
- approximate string matching with Hamming/edit distance
- filtering & verification algorithm based on q-gram counting
- sensitivity control through lossy filtering (DP recursion)
- need to explain basics of different filtering techniques, mainly q-gram counting

Rating: 7
Read Mapping II

- Specialized read mapping:
  - Reads containing structural variants
  - Transcriptome reads spanning introns
  - filtering with k-mer index and spanning sets

Rating: 6
Read Mapping III

- DP algorithm for alignment of indel-containing reads
- Needleman-Wunsch-like

Rating: 4
Variant Detection I

- Realignment of reads to a variant graph
- Improvement in SNP/indel calling

Rating: 3
Alternative Variant Detection I

- Indel and SNP calling method
- Improvement in SNP calling through realignment of reads to candidate haplotypes

Dindel: Accurate indel calls from short-read data
Cornelis A. Albers, Gerton Lunter, Daniel G. MacArthur, Gilean McVean, Willem H. Ouwehand, and Richard Durbin

Small insertions and deletions (indels) are a common and functionally important type of sequence polymorphism. Most of the focus of studies of sequence variation is on single nucleotide variants (SNVs) and large structural variants. In principle, high-throughput sequencing studies should allow identification of indels just as SNVs. However, inference of indels from next-generation sequence data is challenging, and so far methods for identifying indels lag behind methods for calling SNVs in terms of sensitivity and specificity. We propose a Bayesian method to call indels from short-read sequence data in individuals and populations by realigning reads to candidate haplotypes that represent alternative sequence to the reference. The candidate haplotypes are formed by combining candidate indels and SNVs identified by the read mapper, while allowing for known sequence variants or candidates from other methods to be included. In our probabilistic realignment model we account for base-calling errors, mapping errors, and also, importantly, for increased sequencing error indel rates in long homopolymer runs. We show that our method is sensitive and achieves low false discovery rates on simulated and real data sets, although challenges remain. The algorithm is implemented in the program Dindel, which has been used in the 1000 Genomes Project call sets.

Supplemental material is available for this article. The sequence data from this study have been submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession no. ERAO4258. The program Dindel can be freely downloaded from http://www.sanger.ac.uk/resources/software/dindel/.

Small insertions and deletions (indels) are a common and functionally important type of sequence polymorphism. There have been surveys of genome-wide indel variation (Mills et al. 2006), but many studies focus on single nucleotide variants (SNVs) or large structural variants. The 1000 Genomes Project (The 1000 Genomes Project Consortium 2010; http://www.1000genomes.org) will allow a genome-wide and deep study of indel polymorphisms of frequency $\geq 1\%$ in the population. This will provide an important resource for applications in medical resequencing, as indels have been implicated in a number of diseases (e.g., Mikl et al. 1994; Drapkiniskaja et al. 1999). Here, we present a Bayesian algorithm for calling indels from next-generation sequencing data in informative for detection of novel SNVs, but it is not suitable for detection of large insertions of sequence not present in the reference sequence. However, it is possible to detect large deletions through split-read approaches (Ye et al. 2009) or small insertions using paired-end sequencing and mapping. The approach that we propose starts from the second paradigm, thus requiring reads to be first mapped to a reference genome. However, it also incorporates elements of the first paradigm in considering alternative haplotype sequences to explain the data with a probabilistic model, thereby combining strengths of both.

Accurate inference of indels from short-read data is challenging for a number of reasons. First, compared with SNPs, indels occur at...
Variant Detection II

- Extensive evaluation of accuracy of using split reads for SV detection

Identification of genomic indels and structural variations using split reads

Zhengdong D Zhang1*, Jianguo Du2, Hugo Lam3, Alex Abyzov4, Alexander E Urban3, Michael Snyder5 and Mark Gerstein1,2,3*

Abstract

Background: Recent studies have demonstrated the genetic significance of insertions, deletions, and other more complex structural variants (SVs) in the human population. With the development of the next-generation sequencing technologies, high-throughput surveys of SVs on the whole-genome level have become possible. Here we present split-read identification, calibrated (SRIC), a sequence-based method for SV detection.

Results: We start by mapping each read to the reference genome in standard fashion using gapped alignment. Then to identify SVs, we score each of the many initial mappings with an assessment strategy designed to take into account both sequencing and alignment errors (e.g., scoring more highly events gapped in the center of a read). All current SV calling methods have multilevel biases in their identifications due to both experimental and computational limitations (e.g., calling more deletions than insertions). A key aspect of our approach is that we calibrate all our calls against synthetic data sets generated from simulations of high-throughput sequencing (with realistic error models). This allows us to calculate sensitivity and the positive predictive value under different parameter-value scenarios and for different classes of events (e.g., long deletions vs. short insertions). We run our calculations on representative data from the 1000 Genomes Project. Coupling the observed numbers of events on chromosome 1 with the calibrations gleaned from the simulations (for different length events) allows us to construct a relatively unbiased estimate for the total number of SVs in the human genome across a wide range of length scales. We estimate in particular that an individual genome contains ~60,000 indels/SVs.

Conclusions: Compared with the existing read-depth and read-pair approaches for SV identification, our method can pinpoint the exact breakpoints of SV events, reveal the actual sequence content of insertions, and cover the whole size spectrum for deletions. Moreover, with the advent of the third-generation sequencing technologies that produce longer reads, we expect our method to be even more useful.

Keywords: insertion, deletion, structure variation, split read, high-throughput sequencing

Rating: 5
Variant Detection III

- Detecting SVs based on anomalous reads pairs
- Extensive Combinatorics

Recent studies show that along with single nucleotide polymorphisms and small indels, larger structural variants among human individuals are common. The Human Genome Structural Variation Project aims to identify and classify deletions, insertions, and inversions (>5 Kbp) in a small number of normal individuals with a fosmid-based paired-end sequencing approach using traditional sequencing technologies. The realization of new ultra-high-throughput sequencing platforms now makes it feasible to detect the full spectrum of genomic variation among many individual genomes, including cancer patients and others suffering from diseases of genomic origin. Unfortunately, existing algorithms for identifying structural variation (SV) among individuals have not been designed to handle the short read lengths and the errors implied by the “next-gen” sequencing (NGS) technologies. In this paper, we give combinatorial formulations for the SV detection between a reference genome sequence and a next-gen-based, paired-end, whole genome shotgun-sequenced individual. We describe efficient algorithms for each of the formulations we give, which all turn out to be fast and quite readable: they are also applicable to all next-gen sequencing methods (Illumina, 454 Life Sciences [Roche], ABI SOLID, etc.) and traditional capillary sequencing technology. We apply our algorithms to identify SV among individual genomes very recently sequenced by Illumina technology.

[Supplemental material is available online at www.genome.org. The source code of the algorithm implementations and predicted structural variants are available at http://compbio.cs.ubc.ca/strvar.htm]
Alternative Variant Detection III

- Detecting SVs based on anomalous reads pairs
- Based on clustering and optimization (hill climbing) algorithm

A robust framework for detecting structural variations in a genome
Seunghak Lee1,*, Elango Cheran1,*, and Michael Brudno1,2,*

1Department of Computer Science, 2Banting and Best Department of Medical Research and Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON M5S 3G4, Canada

ABSTRACT
Motivation: Recently, structural genomic variants have come to the forefront as a significant source of variation in the human population, but the identification of these variants in a large genome remains a challenge. The complete sequencing of a human individual is prohibitive at current costs, while current polymorphism detection technologies, such as SNP arrays, are not able to identify many of the large-scale events. One of the promising methods to detect such variants is the computational mapping of clone-end sequences to a reference genome.

Results: Here, we present a probabilistic framework for the identification of structural variants using clone-end sequencing. Unlike previous methods, our approach does not rely on an a priori determined mapping of all reads to the reference. Instead, we build a framework for finding the most probable assignment of sequence clones to potential structural variants based on the other clones. We compare our predictions with the structural variants identified in three previous studies. While there is a statistically significant correlation between the predictions, we also find a significant number of previously uncharacterized structural variants. Furthermore, we identify a number of putative cross-chromosomal events, primarily located proximally to the centromeres of the chromosomes.

Availability: Our dataset, results and source code are available at http://compbio.cs.toronto.edu/structvar/
Applied Papers I

- Mapping of short reads onto library of breakpoint junctions
- Inference of ancestral state of structural variants through comparison with other primate genomes
- Classification of SVs based on sequence features

Rating: 7

Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library

Hugo Y K Lam1,13, Xiameng Jasmine Mu1,2,13, Adrian M Stutz3, Andrea Tanzer4, Philip D Cayting5, Michael Snyder2,12, Philip M Kim6,9, Jan O Korbel3,10,12 & Mark B Gerstein1,5,11

Structural variants (SVs) are a major source of human genomic variation; however, characterizing them at nucleotide resolution remains challenging. Here we assemble a library of breakpoints at nucleotide resolution from collating and standardizing ~2,000 published SVs. For each breakpoint, we infer its ancestral state (through comparison to primate genomes) and its mechanism of formation (e.g., nonallelic homologous recombination, NAHR). We characterize breakpoint sequences with respect to genomic landmarks, chromosomal location, sequence motifs and physical properties, finding that the occurrence of insertions and deletions is more balanced than previously reported and that NAHR-formed breakpoints are associated with relatively rigid, stable DNA helices. Finally, we demonstrate an approach, BreakSeq, for scanning the reads from short-read sequenced genomes against our breakpoint library to accurately identify previously overlooked SVs, which we then validate by PCR. As new data become available, we expect our BreakSeq approach will become more sensitive and facilitate rapid SV genotyping of personal genomes.

Structural variation of large segments (>1 kb), including copy-number variation and unbalanced inversion events, is widespread in human genomes, with ~20,000 SVs presently reported in the Database of Genomic Variants (DGV). These SVs have considerable impact on genomic variation by causing more nucleotide differences between individuals than single-nucleotide polymorphisms (SNPs). In several genomic loci, rates of SV formation could even be orders of magnitude higher than rates of single nucleotide substitution. To measure the influence on human phenotypes of common SVs (that is, those present at substantial allele frequencies in populations) and de novo formed SVs, several studies have mapped SVs across individuals. They reported associations of SVs with normal traits and with a range of diseases, including cancer, HIV, developmental disorders and autoimmune diseases. Although most SVs listed in DGV are presumably common, de novo SV formation is believed to occur constantly in the germline and several mutational mechanisms have been proposed. Nevertheless, so far our understanding of SVs and the way we analyze SV maps is limited by the limited resolution of most recent surveys, such as those solely based on microarrays, which have not revealed the precise
Applied Papers II

- Mapping of cancer transcriptome reads (short and long reads)
- Identification of gene fusions

Transcriptome sequencing to detect gene fusions in cancer

ChimeraScan: a tool for identifying chimeric transcription in sequencing data
Matthew K. Iyer$^{1,2}$, Arul M. Chinnaiyan$^{1,2,3,4,5}$ and Christopher A. Maher$^{1,2,3,4}$
$^1$Michigan Center for Translational Pathology, $^2$Center for Computational Medicine and Biology, $^3$Department of Pathology, $^4$Howard Hughes Medical Institute and $^5$Department of Urology, University of Michigan Medical School, Ann Arbor, MI 48109, USA
Associate Editor: Alfonso Valencia

**ABSTRACT**

**Summary:** Next generation sequencing (NGS) technologies have enabled de novo gene fusion discovery that could reveal candidates with therapeutic significance in cancer. Here we present an open-source software package, ChimeraScan, for the discovery of chimeric transcription between two independent transcripts in high-throughput transcriptome sequencing data.

**Availability:** http://chimerascan.googlecode.com
**Contact:** cmaher@dom.wustl.edu
**Supplementary Information:** Supplementary data are available at Bioinformatics online.

Received on March 4, 2011; revised on July 26, 2011; accepted on August 3, 2011

**INTRODUCTION**

High-throughput transcriptome sequencing (RNA-Seq) facilitates detection of aberrant, chimeric RNAs (Maher et al., 2009a;...