Freie Universität Berlin

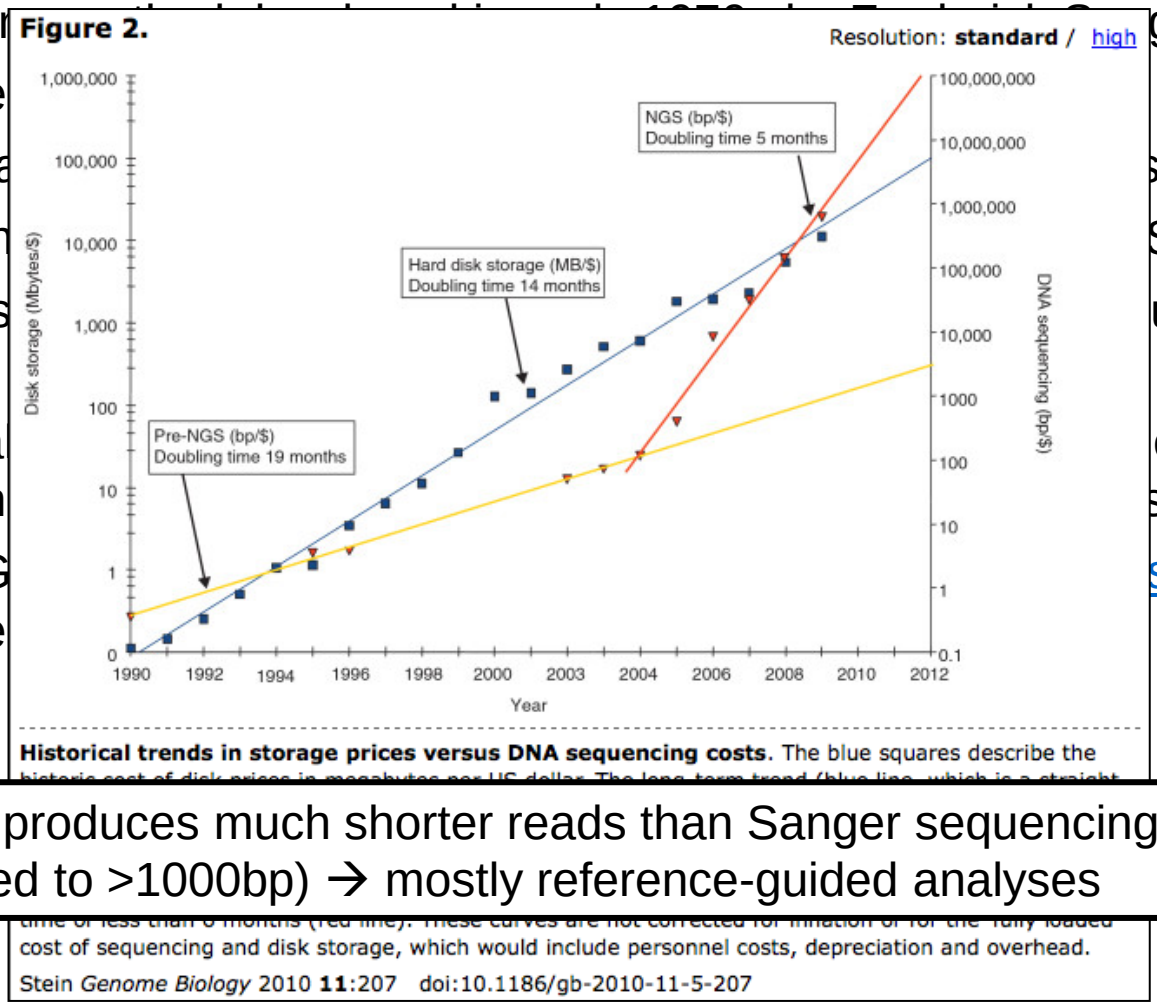
Computional Methods for High-Throughput Omics Data - Genomics

Seminar Vorbesprechung

18.10.2011

Brief history of DNA sequencing

- Sanger sequenci ger
- First complete ge
- Late 1980s: first a systems)
- 1990s: sequencin STs)
- 2001: First draft s ut $1 per base
- Mid 2000s: so-ca emerge, e.g. 454 Life Scien s
- 2008: The1000 G s.org/), cost approaching $1 pe



Problem: NGS produces much shorter reads than Sanger sequencing (50-400bp compared to >1000bp) → mostly reference-guided analyses

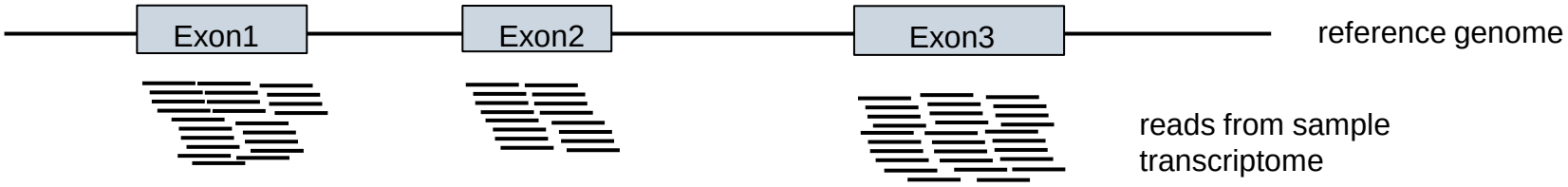
Next Generation Sequencing (NGS) has many applications

RNA-Seq:

→Find out which genes are expressed in a certain tissue at a certain timepoint

→Quantify gene expression

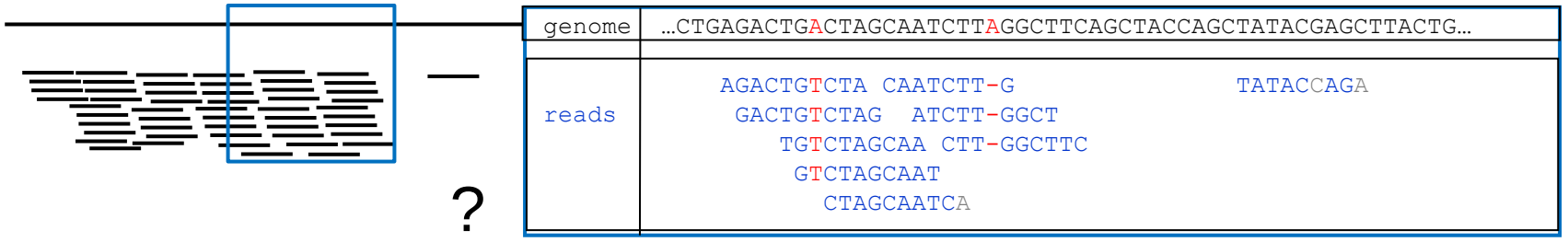
→Identify alternative splicing events, i.e. Transcript isoforms



DNA-Seq:

→Identify differences between individuals/cell types, e.g. Cancer vs. healthy cells

→SNPs, short indels, large indels, inversions, translocations, copy number variations



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: classifcation 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

Resource

RazerS—fast read mapping with sensitivity control

David Weese,[1,3] Anne-Katrin Emde,[1] Tobias Rausch,[2] Andreas Döring,[1] and Knut Reinert[1]

[1]*Department of Computer Science, Free University of Berlin, 14195 Berlin, Germany;* [2]*International Max Planck Research School for Computational Biology and Scientific Computing, 14195 Berlin, Germany*

Second-generation sequencing technologies deliver DNA sequence data at unprecedented high throughput. Common to most biological applications is a mapping of the reads to an almost identical or highly similar reference genome. Due to the large amounts of data, efficient algorithms and implementations are crucial for this task. We present an efficient read mapping tool called RazerS. It allows the user to align sequencing reads of arbitrary length using either the Hamming distance or the edit distance. Our tool can work either lossless or with a user-defined loss rate at higher speeds. Given the loss rate, we present an approach that guarantees not to lose more reads than specified. This enables the user to adapt to the problem at hand and provides a seamless tradeoff between sensitivity and running time.

[RazerS is freely available at http://www.seqan.de/projects/razers.html.]

Second-generation sequencing technologies are revolutionizing the field of DNA sequence analysis, as large amounts of sequencing data can be obtained at increasing rates and dramatically decreasing costs. Biological applications are manifold, including whole-genome resequencing for the detection of genomic variation, e.g., single nucleotide polymorphisms (SNPs) (Bentley et al. 2008; Hillier et al. 2008; Ley et al. 2008; Wang et al. 2008) or large structural variations (Chen et al. 2008), RNA sequencing for small noncoding RNA discovery or expression profiling (Morin et al. 2008), metagenomics applications (Huson et al. 2007), and sequencing of chromatin-immunoprecipitated DNA, e.g., for the identification of DNA binding sites and histone modification patterns (Barski et al. 2007).

Fundamental to all these applications is the problem of mapping all sequenced reads against a reference genome, denoted as the *read mapping problem*. It can be formalized as follows: given a set of read sequences $\mathcal{R}$, a reference sequence G, and a distance $k \in \mathbb{N}$, find all substrings g of G that are within distance k to a read $r \in \mathcal{R}$. The occurrences of g in G are called *matches*. Common distance measures are Hamming distance or edit distance; the former matches in a more time-consuming verification step. In current implementations one has to carefully distinguish whether both steps, the filtration step and the verification step, are adequate for the distance chosen (Hamming or edit distance). Some implementations, for instance, verify matches using base-call qualities, but filter the candidate regions using a fixed Hamming or edit distance (H Li et al. 2008). Filtration methods in use are based on single (Kent 2002; Ma et al. 2002) or multiple seeds (Li et al. 2003; Lin et al. 2008), the pigeonhole principle (Navarro and Raffinot 2002; H Li et al. 2008; R Li et al. 2008; AJ Cox, ELAND: Efficient local alignment of nucleotide data, unpubl.), or based on counting lemmas using (gapped) q-gram (Burkhardt et al. 1999; Rasmussen et al. 2006; Rumble and Brudno 2008). Verification methods encompass semiglobal alignment algorithms for Hamming or edit distance (Levenshtein 1966) or local-alignment algorithms (Smith and Waterman 1981).

BLAT (Kent 2002), as an example of a single seed filter, searches exact occurrences of short fixed sized substrings shared by two sequences. PatternHunter (Ma et al. 2002) was the first to generalize this strategy to gapped seeds (common discontiguous

Read Mapping II

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

Rating: 6

BIOINFORMATICS ORIGINAL PAPER
Vol. 26 no. 7 2010, pages 873–881
doi:10.1093/bioinformatics/btq057

Sequence analysis

Advance Access publication February 10, 2010

Fast and SNP-tolerant detection of complex variants and splicing in short reads

Thomas D. Wu* and Serban Nacu

Department of Bioinformatics, Genentech, Inc., 1 DNA Way, South San Francisco, CA, USA

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Next-generation sequencing captures sequence differences in reads relative to a reference genome or transcriptome, including splicing events and complex variants involving multiple mismatches and long indels. We present computational methods for fast detection of complex variants and splicing in short reads, based on a successively constrained search process of merging and filtering position lists from a genomic index. Our methods are implemented in GSNAP (Genomic Short-read Nucleotide Alignment Program), which can align both single- and paired-end reads as short as 14 nt and of arbitrarily long length. It can detect short- and long-distance splicing, including interchromosomal splicing, in individual reads, using probabilistic models or a database of known splice sites. Our program also permits SNP-tolerant alignment to a reference space of all possible combinations of major and minor alleles, and can align reads from bisulfite-treated DNA for the study of methylation state.

Results: In comparison testing, GSNAP has speeds comparable to existing programs, especially in reads of ≥70 nt and is fastest in detecting complex variants with four or more mismatches or insertions of 1–9 nt and deletions of 1–30 nt. Although SNP tolerance does not increase alignment yield substantially, it affects alignment results in 7–8% of transcriptional reads, typically by revealing alternate genomic mappings for a read. Simulations of bisulfite-converted DNA show a decrease in identifying genomic positions uniquely in 6% of 36 nt reads and 3% of 70 nt reads.

Availability: Source code in C and utility programs in Perl are freely available for download as part of the GMAP package at

programs, such as Bowtie (Langmead *et al.*, 2009), BWA (Li and Durbin, 2009) and SOAP2 (Li *et al.*, 2009), have shown how suffix arrays (Manber and Myers, 1993), compressed using a Burrows–Wheeler Transform (BWT; Burrows and Wheeler, 1994), can rapidly map reads that are exact matches or have a few mismatches or short insertions or deletions (indels) relative to the reference.

In addition to speed, it is also important to broaden the range of possible variants that can be detected in reads, since interesting biology is likely to be revealed not merely as single nucleotide polymorphisms (SNPs) or mutations from the reference, but also as more complex phenomena, such as multiple mismatches, long indels and combinations thereof. Such complex variants represent a substantial source of genetic diversity. For example, indels represent 7–8% of human polymorphisms, with 25% of coding indels being longer than 3 nt (Bhangale *et al.*, 2005; Weber *et al.*, 2002). Long indels that affect multiple amino acids may have significant biological consequences. Moreover, as reads continue to lengthen, from their original ~30 nt to their current 75–100 nt, they are more likely to have multiple or complex differences from the reference, making detection of complex variants even more critical.

Other important biological phenomena arise from splicing events, which provide insights into gene structure, alternative splicing, gene fusions and chromosomal rearrangements. Although splicing can be determined readily in long EST and cDNA sequences using general-purpose genomic mapping and alignment programs such as BLAT (Kent, 2002) or GMAP (Wu and Watanabe, 2005), short reads pose a challenge because they often align to numerous places in a genome, and because they often lack insufficient sequence specificity on one

Read Mapping III

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

Rating: 4

BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 5 2011, pages 595–603
doi:10.1093/bioinformatics/btq713

Genome analysis
Advance Access publication January 13, 2011

AGE: defining breakpoints of genomic structural variants at single-nucleoti with gap excisi

A.Abyzov and M.Gerstein

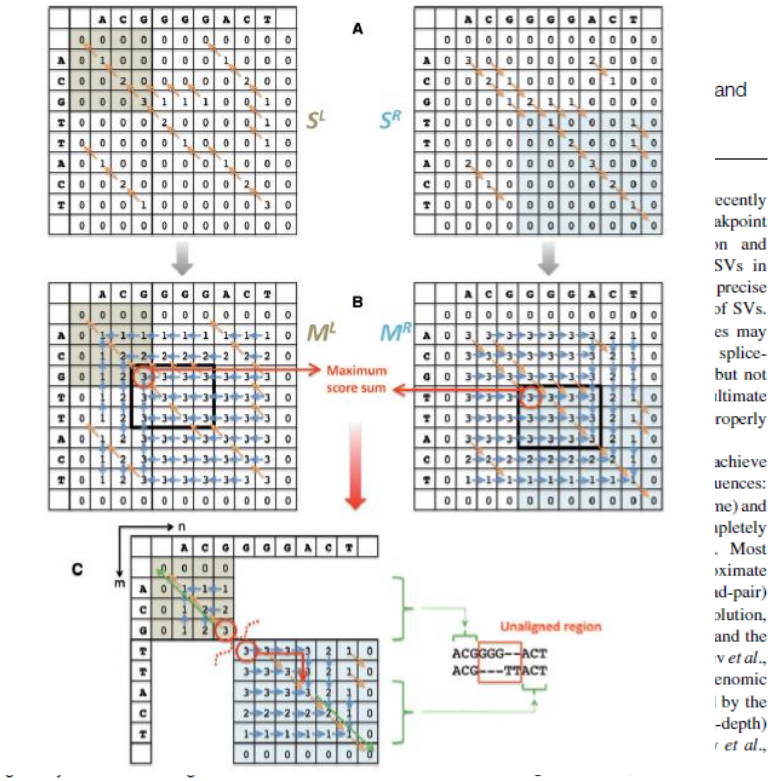
Alexej Abyzov[1,2,*] a
[1]Program in Computation
[3]Department of Computer

Associate Editor: Alfonso Valenci

ABSTRACT

Motivation: Defining the preci (SVs) at single-nucleotide bre problem, as it is a prerequisite functional impact and reconstru Given approximate breakpoint or split read, the problem esse sequence alignment. Classical generalizations guarantee findi global or local alignment of tw generally be applied to findin of genomic sequences contai simultaneously span the SV (e precise local alignments at the

Results: Here, we formulate problem and describe a dyna solution. Specifically, our algori Gap Excision, finds the optimal the 5′ and 3′ ends of two g 'large-gap jump' between the the total alignment score. We al application of AGE to tandem d events involving two large ga implementation of AGE (allowi make it available as a downloa applied AGE for breakpoint dete 1000 Genomes Project by align



and

ecently
akpoint
on and
SVs in
precise
of SVs.
es may
splice-
but not
ultimate
roperly

achieve
uences:
ne) and
upletely
. Most
oximate
ad-pair)
olution,
and the
v et al.,
enomic
by the
-depth)
et al.,

Variant Detection I

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

Rating: 3

Homer and Nelson *Genome Biology* 2010, **11**:R99
http://genomebiology.com/2010/11/10/R99

Genome Biology

METHOD Open Access

Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA

Nils Homer[1,2,3*], Stanley F Nelson[2]

Abstract

A primary component of next-generation sequencing analysis is to align short reads to a reference genome, with each read aligned independently. However, reads that observe the same non-reference DNA sequence are highly correlated and can be used to better model the true variation in the target genome. A novel short-read micro re-aligner, SRMA, that leverages this correlation to better resolve a consensus of the underlying DNA sequence of the targeted genome is described here.

Background

Whole-genome human re-sequencing is now feasible using next generation sequencing technology. Technologies such as those produced by Illumina, Life, and Roche 454 produce millions to billions of short DNA sequences that can be used to reconstruct the diploid sequence of a human genome. Ideally, such data alone could be used to *de novo* assemble the genome in question [1-6]. However, the short read lengths (25 to 125 bases), the size and repetitive nature of the human genome (3.2×10^9 bases), as well as the modest error rates (approximately 1% per base) make such *de novo* assembly of mammalian genomes intractable. Instead, short-read sequence alignment algorithms have been developed to compare each short sequence to a reference genome [7-12]. Observing multiple reads that differ combination of sequencing error, equivalent positions of the variant being equally likely, and adjacent variants or nearby errors driving misalignment of the local sequence. These local misalignments lead to false positive variant detection, especially at apparent heterozygous positions. For example, insertions and deletions towards the ends of reads are difficult to anchor and resolve without the use of multiple reads. In some cases, strict quality and filtering thresholds are used to overcome the false detection of variants, at the cost of reducing power [13]. Since each read represents an independent observation of only one of two possible haplotypes (assuming a diploid genome), multiple read observations could significantly reduce false-positive detection of variants. Algorithms to solve the multiple sequence alignment problems typically compare multiple

Alternative Variant Detection I

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

Rating: 7

Resource

Dindel: Accurate indel calls from short-read data

Cornelis A. Albers,[1,2,5] Gerton Lunter,[3] Daniel G. MacArthur,[1] Gilean McVean,[4] Willem H. Ouwehand,[1,2] and Richard Durbin[1]

[1]*Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1HH, United Kingdom;* [2]*Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge CB2 1TN, United Kingdom;* [3]*Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, United Kingdom;* [4]*Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom*

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. ERA014258. 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 ≥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., Miki et al. 1994; Draptchinskaia et al. 1999). Here, we present a Bayesian algorithm for calling indels from next-generation sequencing data in in-

powerful 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

Rating: 5

Zhang *et al. BMC Genomics* 2011, **12**:375
http://www.biomedcentral.com/1471-2164/12/375

BMC Genomics

METHODOLOGY ARTICLE Open Access

Identification of genomic indels and structural variations using split reads

Zhengdong D Zhang[1*], Jiang Du[2], Hugo Lam[3], Alex Abyzov[1], Alexander E Urban[4], Michael Snyder[5] and Mark Gerstein[1,2,3*]

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 ~670,000 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

Variant Detection III

- Detecting SVs based on anomalous reads pairs
- Extensive Combinatorics

Rating: 8

Methods

Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes

Fereydoun Hormozdiari,[1,4] Can Alkan,[2,3,4] Evan E. Eichler,[2,3,5] and S. Cenk Sahinalp[1,5]

[1]School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6; [2]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; [3]Howard Hughes Medical Institute, Seattle, Washington 98195, USA

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 reliable; 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.sfu.ca/strvar.htm.]

Recent introduction of the next-generation sequencing technologies has significantly changed how genomics research is conducted (Mardis 2008). High-throughput, low-cost sequencing technologies such as pyrosequencing (454 Life Sciences [Roche]), sequencing-by-synthesis (Illumina and Helicos), and sequencing-by-ligation (ABI SOLiD) methods produce shorter reads than the traditional capillary sequencing, but they also increase the redundancy by 10- to 100-fold or more (Shendure et al. 2004; Mardis 2008). With the arrival of these new sequencing technologies, along with the capability of sequencing paired-ends (or "mate-pairs") of a clone insert that follows a tight length distribution (Raphael et al. 2003; Volik et al. 2003; Dew et al. 2005; Tuzun et al. 2005; Korbel et al. 2007; Bashir et al. 2008; Kidd et al. 2008; Lee et al. 2008), it is becoming feasible to perform detailed and comprehensive genome variation and rearrangement studies.

duplications, inversions, and translocations (Feuk et al. 2006; Sharp et al. 2006) (see Supplemental material for details on types of SV).

End-sequence profiling (ESP) was first presented by Volik et al. (2003) and Raphael et al. (2003) to discover SV events using bacterial artificial chromosome (BAC) end sequences to map structural rearrangements in cancer cell line genomes, and it was used by Tuzun et al. (2005) to systematically discover structural variants in the genome of a human individual. Several other genome-wide studies (Iafrate et al. 2004; Sebat et al. 2004; Redon et al. 2006; Cooper et al. 2007; Korbel et al. 2007) demonstrated that SV among normal individuals is common and ubiquitous. More recently, Kidd et al. (2008) detected, experimentally validated, and sequenced SV from eight different individuals. The ESP method was also utilized by Dew et al. (2005) to evaluate and

Alternative Variant Detection III

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

Rating: 8

BIOINFORMATICS

Vol. 24 ISMB 2008, pages i59–i67
doi:10.1093/bioinformatics/btn176

A robust framework for detecting structural variations in a genome

Seunghak Lee[1,*], Elango Cheran[1,*] and Michael Brudno[1,2,*]

[1]Department of Computer Science, [2]Banting 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 most 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 sequenced 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/

a change in the abundance of DNA that matches any probe, such as inversions and translocations. Recently, the completion of the diploid genome of an individual (Levy *et al.*, 2007) has, for the first time, made it possible to directly compare two complete human genomes, enabling us to begin to understand the variety of genotypes present in the human population. This fully assembled genome, however, is quite different from the data that will become available in the near future. The National Human Genome Research Institute is planning to sequence the genomes of 1000 human individuals in the next few years using next generation sequencing (NGS) technologies. While the NGS technologies will drastically reduce the cost of resequencing an individual human, it is currently unclear to what extent these platforms can be used to identify structural variations.

The bulk of the currently known structural variants have been determined by mapping either individual reads (Mills *et al.*, 2006) or clone-ends (Korbel *et al.*, 2007; Tuzun *et al.*, 2005) from donor individuals to a reference genome. Many sequencing techniques allow for the generation of reads from the two ends of a DNA fragment simultaneously. Because the size of a DNA fragment can be determined, e.g. by running it on a gel, this allows for the generation of paired reads, positioned at a known distance (insert size) from each other in a genome. Such pairs of reads are known as clone-ends, or matepairs. Using a known genomic sequence as a

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

_computational BIOLOGY

RESOURCE

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

Hugo Y K Lam[1,13], Xinmeng Jasmine Mu[1,2,13], Adrian M Stütz[3], Andrea Tanzer[4], Philip D Cayting[5], Michael Snyder[2,12], Philip M Kim[6–9], Jan O Korbel[3,10,13] & Mark B Gerstein[1,5,11]

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[1–6] with ~20,000 SVs presently reported in the Database of Genomic Variants (DGV)[2]. These SVs have considerabe impact on genomic variation by causing more nucleotide differences between individuals than single-nucleotide polymorphisms[4–6] (SNPs). In several genomic loci, rates of SV formation could even be orders of magnitude higher than rates of single nucleotide substitution[7,8]. 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[9–14]. 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[15].

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 as whether common SVs emerged initially as insertions or deletions at ancestral genomic loci. Instead, operational definitions have been applied for classifying common SVs into gains, losses, insertions and deletions based on either allele frequency measurements, or the 'human reference genome' (hereafter also referred as the reference genome) that was originally derived from a mixed pool of individuals[17]. Thus, inference of the ancestral state of an SV locus is crucial for relating SV surveys to primate genome evolution and population genetics.

The lack of data at nucleotide resolution has also limited the number of SVs for which the likely mutational mechanisms of origin have been inferred. These mechanisms are thought to include (i) NAHR involving homology-mediated recombination between paralogous sequence blocks; (ii) nonhomologous recombination (NHR) associated with the repair of DNA double-strand breaks (that is, nonhomologous end-joining) or with the rescue of DNA replication-fork stalling events (that is, fork stalling and template switching[18]); (iii) variable number of tandem repeats (VNTRs) resulting from expansion or contraction of simple tandem repeat

Applied Papers II

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

Rating: 6

Vol 458|5 March 2009|doi:10.1038/nature07638 nature

LETTERS

Transcriptome sequencing to detect gene fusions in cancer

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 20 2011, pages 2903–2904 doi:10.1093/bioinformatics/btr467

a-Sundaram[1,3], Bo Han[1,3], Chinnaiyan[1,2,3,4,5]

Genome analysis

Advance Access publication August 11, 2011

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,*]

[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

1 INTRODUCTION

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

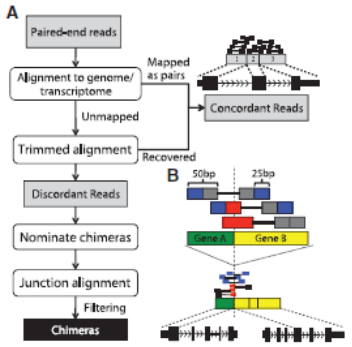


Fig. 1. ChimeraScan flowchart. (**A**) Paired-end reads failing an initial

detection, using short-read sequen-
sequence data and long-read tech-
platform) to provide reference
fusion genes.
criptome sequencing was whether
ripts in the background of highly
at is, would complementary DNA
ed). To address this, we compared
non-normalized cDNA libraries of
P, which harbours the gene fusion
Table 1). Overall, the normalized
3.6-fold reduction in the total num-
hermore, although we expected the
or the *TMPRSS2–ERG* gene fusion,
–ERG chimaeras, which suggested
ormalization in our analyses.
ng massively parallel transcriptome
sions, we generated non-normalized
cancer cell lines VCaP and LNCaP,
cell line RWPE. As a first step, using
ated 551,912 VCaP, 244,984 LNCaP