ISMB 2008 Special Interest Group on Algorithms for Short Read Sequencing

July 19, 2008

Schedule:

8:40-10:15am: Session 1: Welcome & Read Mapping
10:15-10:45am: Coffee break
10:45am-12:30pm: Session 2: Imaging, Data Management and SNP calling
12:30-1:45pm: Lunch
1:45pm-3:30pm: Session 3: Structural Variation & Analysis
3:30-4:00pm: Coffee break
4:00pm-6:00pm Session 4: Assembly & Keynote

Organizers:
Michael Brudno, University of Toronto
Pavel Pevzner, University of California – San Diego
Steve Skiena, State University of New York – Stony Brook
Francisco de la Vega, Applied Biosystems

Abstracts:

Read Mapping:

Imaging/Data Management:

SNP calling:

Analysis:

Structural Variations:

Assembly:

Keynote:

John McPherson
Director of Cancer Genomics and Senior Principal Investigator,
Ontario Institute for Cancer Research

PASS: A Fast Algorithm for Illumina/SOLiD Sequence Alignment with Guaranteed Recall

Manhong Dai, Justin Wilson, Stanley Watson and Fan Meng
Psychiatry Department and MBNI, University of Michigan, Ann Arbor, MI 48109, US

Although many sequence alignment programs are used for the analysis of Illumina/SOLiD high throughput sequencing results, there is still significant room for improvement in both speed and recall rate. Traditional sequence alignment programs do not perform well when aligning short sequences generated by deep sequencing techniques since they cannot take advantage of features inherent in deep sequencing results. Many programs designed for high throughput sequence alignment cannot make use of pre-built indices to accelerate future alignments. For example, all of the work used to build the in-memory data structure for a reference genome must be repeated for the next alignment using the same reference genome. The recall rate of popular programs such as Eland is far from satisfactory, too.

These observations led us to create the Pre-sorted Alignment of Short Sequences (PASS) program for sequences from Illumina/SOLiD platforms. PASS capitalizes on the following assumptions (1) query sequences can be preprocessed to a uniform length of $\leq$ 255 bp (2) insertions and deletions can be neglected (3) sequences with mismatches above a threshold (usually $\leq$ 2 bp) can be ignored. (4) the only valid bases are A, T, C and G and the unknown base N that does not match any other base. Furthermore, we assume that time required to index a reference sequence will be amortized over many uses and it is preferable to use data structures that are large on disk but small in memory for better scalability since disk is cheaper than memory.

Aligning sequences using PASS involves three steps. First, the target reference sequences and its reverse complement are permuted to generate every short sequence of the desired length that contains mismatches at a predefined threshold. The generated sequences are indexed alphabetically. Second, the query sequences are indexed alphabetically. Third, the indexed reference sequences and the indexed query sequences are sequentially scanned for perfect matches that indicate an alignment. The indexes generated in the first two steps can be re-used for subsequent alignments. For a given small sequence length and number of short sequences N, the complexity of the indexing procedures is O(N log N). For a given small sequence length, a number of indexed reference sequences M and indexed query sequences N, the complexity of performing the alignment is O(M + N).

Table 1 is a summary of comparative analysis of PASS performance and recall with other popular short sequence alignment programs. Our tests are based on the alignment of 15,000,000 Illumina 1G sequences with the human chromosome 1. Test sequences are downloaded from http://www.bcgsc.ca/data/chipseq (Robertson et al, (2007) Nature Methods 4(8) 651-7).

Table 1: Comparison of High Performance Short Sequence Alignment Programs

	Speed (time in seconds)				Recall (MUMmer and PASS as 100%)			
	Perfect Match		Allow at most 1 bp Mismatch		Perfect Match		Allow at most 1 bp Mismatch	
Program tested	All match	Unique Match	All match	Unique Match	All match	Unique Match	All match	Unique Match
PASS without pre-built index	470	537	14838	13849	1187409 (100%)	1083975 (100%)	1950566 (100%)	1117331 (100%)
PASS with pre-built index	109	102	2971	2668				
MUMmer	366	766	/	/	100%	100%	/	/
Eland *	/	1072*	/	1072*	/	92.39%	/	57.22%
SOAP	8895	/	15728	/	100%	100%	99.47%	99.59%
MosaikAligner	/	675	/	680	/	85.28%	/	54.34%

*Eland allows two mismatches for each query sequence by default.

We believe PASS has significant advantage for large scale sequence alignments requiring near-perfect matches with high recall rates. We are making improvements in PASS for the alignment of longer sequences. The current version of PASS is freely available at: http://brainarray.mbni.med.umich.edu/Brainarray/pass/ .

Optimal Spliced Alignments of Short Sequence Reads

Fabio De Bona*[1], Stephan Ossowski[2], Korbinian Schneeberger[2], Gunnar Rätsch[1]

[1]Friedrich Miescher Laboratory, Max Planck Society, Spemannstr. 39, 72076 Tübingen, Germany
[2]Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany

Email: Fabio De Bona*- fabio@tuebingen.mpg.de;

*Corresponding author

Motivation

Next generation sequencing technologies open exciting new possibilities for genome and transcriptome sequencing. While reads produced by these technologies are relatively short and error prone compared to the Sanger method their throughput is several magnitudes higher. We present a novel approach, called *QPALMA*, for computing accurate spliced alignments which takes advantage of the read's quality information as well as computational splice site predictions. Our method uses a large margin approach similar to support vector machines to estimate its parameters to maximize alignment accuracy. In computational experiments we illustrate that the quality information as well as the splice site predictions help to improve the alignment quality. Our algorithms were optimized and tested using reads produced with the Illumina Genome Analyzer for the model plant *Arabidopsis thaliana*.

Method

In this work we aim to develop a method exploiting all available information to accurately align as many as possible spliced sequence reads to the genome. In a previous work we already proposed a method taking advantage of splice site predictions and an intron length model (*Palma* by Schulze et al). In this work we extend this method to also benefit from the read's quality scores. The algorithm is based on extensions of the well-known Smith-Waterman algorithm using more sophisticated parametrized scoring functions. The idea is to tune the parameters of the scoring functions such that the true alignment does not only achieve a large score (to be "most likely"), but also that all other alignments score lower than the true alignment (to obtain a "large margin between the alignments"). Similar ideas are used in other large margin algorithms such as Support Vector Machines.
In a typical application scenario one needs to align millions of short sequence reads against the genome. In this case the direct application of the extended Smith-Waterman algorithm is not feasible. We therefore propose to combine our method with a fast suffix array based approach to identify a seed for the alignment. This combined strategy will allow us to efficiently align even very large numbers of reads identifying their spliced alignments.

Results

We test our algorithm on 30,000 sequences different from the training set of size 10,000 for an unbiased estimation of *QPALMA*'s accuracy on unspliced reads. We compute the fraction of reads that have been accurately aligned at all four boundaries (start and end of first and second exon) with and without using read quality information, splice site predictions and intron length scoring, respectively. From the results given in the table we can conclude that all three components help to reduce the alignment error rate.

Quality information	Splice site pred.	Intron length	Error rate	Quality information	Splice site pred.	Intron length	Error rate
-	-	-	14.19	-	-	+	9.96 %
+	-	-	13.49	+	-	+	9.68 %
-	+	-	3.16	-	+	+	1.94 %
+	+	-	2.81	+	+	+	1.78%

SHRiMP: The Short Read Mapping Package

Stephen Rumble[1], Vladimir Yanovsky[1], Michael Brudno[1,2]

[1]Department of Computer Science and [2]Banting and Best Dept. of Medical Research
University of Toronto.

The Short Read Mapping Package (SHRiMP) is a method for mapping very short reads to a genome. Our method includes 1) a spaced kmer filtering technique, 2) a very fast, vectorized implementation of the Smith-Waterman algorithm, 3) a separate full color-space, letter-space and multi-pass alignment approaches, and 4) computation of p-values and other statistics for hits.

The algorithm starts with a rapid k-mer hashing step to localize potential areas of similarity between the reads and the genome. All of the spaced k-mers present in the reads are indexed. Then for each k-mer in the genome, all of the matches of that particular k-mer among the reads are found. The approach of indexing the reads, rather than the genome has several advantages including controlling memory usage, as our algorithm never needs memory proportional to the size of the genome, while the large set of short reads can be easily divided between many machines in a compute cluster. If a particular read has as many or more than a specified number of k-mer matches within a given window of the genome, we execute a vectorized Smith-Waterman step to score and validate the similarity. The top *n* highest-scoring regions are filtered through a sensitive alignment algorithm, and output at the end of the program if their final scores meet a specified threshold.

SHRiMP offers three possible options for the final, full backtracking alignment step. These are customized for Illumina/Solexa (regular, letter-space alignment), the AB SOLiD dibase sequencing (color-space alignment), and a method for two-pass sequencing, such as the Heliscope. The AB SOLiD sequencing technology introduced a novel dibase sequencing technique, which reads overlapping pairs of letters and generates one of four colors (typically labeled 0-3) at every stage. The sequencing code can be thought of as a finite state automaton (FSA), in which each previous letter is a state and each color code is a transition to the next letter state. We implement an algorithm for aligning color space reads in letter space. Our key observation is that while a color-space error causes the rest of the sequence to be mistranslated, the genome will match one of the other three possible translations. We adapt the classical dynamic programming algorithm to simultaneously align the genome to all four possible translations of the read, allowing the algorithm to move from one translation to another by paying a "crossover", or sequencing error penalty. Our approach handles not only mismatches, and sequencing errors, but also indels.

A different final alignment mode is available for Single Molecule Sequencing technologies. These technologies suffer from a significantly higher deletion error rate, ameliorated by the ability to sample two reads from the same location. In the SHRiMP tool we combine the Weighted Sequence Graph representation of all optimal and near optimal alignments between the two reads sampled from a piece of DNA, which is then aligned to the reference genome.

The SHRiMP tool is freely available at http://compbio.cs.toronto.edu/shrimp

THE SOLiD SYSTEM RESEQUENCING ALIGNENT SOFTWARE SUITE: MATCHING, PAIRING AND INDEL FINDING

Zheng Zhang, Heather Peckham, and Jingwe Ni
Applied Biosystems, Foster City, CA

Keyword: alignment, short reads, resequencing, paired reads

Applied Biosystems SOLiD System, is a massively parallel sequencing platform that produces hundreds of millions of short sequencing reads every run. It is a challenging to match (align) these rmassive numbers of eads to a reference genome very fast. Due to the 2-base coding scheme that the SOLiD System uses, the existing alignment software generally cannot process them well. In addition, the SOLiD System also provides paired-end reads, whose approximate distance in the sample DNA is known. This information can help in mapping those reads, and finding indels (insertion/deletion) and other structural rearrangements.

We present a software suite for aligning and analyzing SOLiD reads. The suites include the following components: (1) A matching tool that can find matches between short reads and a long reference sequence allowing a fixed number of mismatches; (2) a pairing tool that find, for each mated pair of reads, a pair of hits that are in right orientation, same strand, and within the correct distance decided by library insert size; (3) a rescue/indel finding tool that find alignments allowing for indel and/or more mismatches using paired-end information, to be used when the pairing tool above fails to find a correct pair of hits; (4) a 2-base encoding to base sequence adaptor program that can translate a color sequence assembly (obtained from any de novo assembler algorithm) to base sequence.

Each of the components can be run stand alone. The matching tool uses multiple discontinuous word patterns to allow for a very fast search time, using a small internal memory footprint and produces small temporary files, so one can map 10+Gb of data to the human genome reference on a reasonable sized computer farm.

The rescue/indel finding tool can find small to medium size deletions (where there is a stretch of DNA in reference sequence but missed from target sequence), and a mini-assembly tool can find novel insertions up to 200 bp long. We provide statistical analysis and a simulation to analyze the significance of different types of indel alignments.

De novo assembly using short reads is challenging, but more and more tools are becoming available for assembling base sequence short reads. We present an adaptor that enables the user to use any existing short-read assembly algorithm to assemble SOLiD reads. Our approach is to use perform assembly in color space (by transforming color string's 0-3 to ACGT). After assembly it is non-trivial to translate the color assembly to a base sequence. The adaptor we present use the known leading base of each SOLiD read to get a "best" base sequence transformation form any color assembly.

Alta-Cyclic – an improved Solexa base caller for longer and more accurate reads

Yaniv Erlich[1,2], Partha P. Mitra[1], W. Richard McCombie[1],
and Gregory J. Hannon[1,2*]

[1] Watson School of Biological Sciences
[2] Howard Hughes Medical Institute
Cold Spring Harbor Laboratory
1 Bungtown Road
Cold Spring Harbor, NY 11724

The power of next-generation sequencing is limited by high error rates, as compared to conventional sequencing, and short read lengths. Here, we sought improvements in sequence determination from the Illumina Genome Analyzer instrument. We systematically analyzed sources of noise that cause error rates to climb with cycle number. We found de-synchronization of polymerases (phasing), progressive loss of signal (fading) and cycle-depending increases in the fluorescent crosstalk among individual nucleotides. We developed a novel base caller, termed Alta-Cyclic, that allows compensation of these noise sources and is based on machine learning approach. This allows valuable 78-base reads, increases the number of accurate reads by more than 4 folds, and reduces systematic biases that degrade the ability to confidently identify sequence variants. Though the analysis presented here is specific to the Illumina machine, the general strategies may also be applicable to other next generation-platforms.

Swift: Open Source Primary Data Analysis for Next-generation sequencers

Nava Whiteford, Tom Skelly, Irina Abnizova, Clive Brown
Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA

Abstract

Swift is an open source package for primary data analysis on next-gen sequence data "from images to basecalls". Currently Swift is targeted toward Solexa/Illumina sequencing, but is designed to be platform agnostic. In this paper we present post image analysis corrections that have been implemented for Solexa sequencers. Corrections for crosstalk (leakage between the signals of A/C and G/T dyes) and "phasing" (the incorporation of a base on prior of subsequent sequencing cycles) are described that can result in error rates 20% lower than that of the Illumina pipeline with increased yield.

From an informatics perspective Swift performs all analysis on a per tile basis. Because of this each tile can be processed independently making Swift "massively parallel" allowing a complete run to be processed on a compute cluster in minutes rather than hours.

Short Read Archive format and storage solution and SDK

Vladimir Alekseyev
Kurt Rodarmer

NCBI Short Read Archive team

ABSTRACT

The bio-tech industry and research community are rapidly growing in size. With each day the amount and types of data multiply, and need to be managed. Data are stored in a plurality of formats, depending upon particular needs. Sometimes the ability to use text-based tools and editors is desired, but more often we try to optimize for storage or transmission. Any approach to storage will have to wrestle with the problem of keeping data in common representation for data interchange while at the same time supporting custom data for product distinction. NCBI has based its Trace and Assembly archives upon an internally developed format that satisfies most of these needs. In particular, data are stored in a very compact and flexible way that was designed for execution. This means that the format supports rapid random access to any read, and is designed as pluggable data modules that when combined form a complete, addressable and highly scalable archive. New data may be easily added, analyzed and annotated over time; old data may be replaced or removed. The format is capable of representing arbitrary data, combining multiple representations (e.g. custom and common), as well as representing heterogeneous data from different manufacturers. Notable features are its compactness, flexibility, scalability, and its capacity to accommodate as yet undefined/unspecified instruments and tools.

SNP calling by next-generation sequencing on draft genomes

Sebastian Fröhler and Christoph Dieterich
Department of Evolutionary Biology; Max Planck Institute for Developmental Biology
Spemannstr. 35-39; Tübingen, Germany
{sebastian.froehler|christoph.dieterich}@tuebingen.mpg.de

Introduction

Next-generation sequencing technologies generate millions and more short-length reads in a single run. These data sets facilitate genome-wide analysis of several biological processes. We are interested in whole-genome genotyping of small genomes ($\sim 10^8$ bp), which is, for example, easily performed with a single flow cell on Illumina's Solexa platform. Several computational approaches exist to map and assemble short-length reads onto reference genomes. SNP calling is subsequently performed using the genome maps of the previous step. To our knowledge, none of these approaches considers the quality of the reference genome, which is reasonable for well-studied model organisms like *C. elegans* and *D. melanogaster*. Both genome assemblies have undergone "finishing" steps, which involve extensive manual curation. However, the quality of draft genomes may suffer from omitting these finishing steps, which are very costly and therefore often skipped in current projects.

Our solution

A simple yet elegant solution is to include the reference genome quality in a common framework for SNP calling. This is most effectively attained in a Bayesian framework, which utilizes sequence quality scores (e.g. Phred scores) from both, short-reads and the reference genome. We decided to build upon a Bayesian framework proposed by Marth *et.al.* (1999) and extend it to include sequence quality values from the reference genome. This framework calculates the *a-posteriori* probability of a SNP by integrating information on base background frequencies, base call errors and SNP alignment column permutation probabilities. This approach is further optimized by preprocessing filtering steps.

Results and Conclusions

SNP calling on draft genomes is error-prone if the reference genome quality is ignored. We propose a simple solution to this problem and show its performance on the newly sequenced genome of *Pristionchus pacificus* a satellite organism to *Caenorhabditis elegans*. We also confirmed our predictions by traditional Sanger sequencing. Our software is implemented in JAVA and freely available upon request.

Using UHTS to detect somatic mutations

Christian Iselli , Stylianos Antonarakis, Jacqui Beckmann, Andy Simpson, Jean-Louis Blouin, Siv Fokstuen, Daniel Robyr, Carlo Rivolta, Jacques Rougemont, Ioannis Xenarios, Brian Stevenson, Victor Jongeneel, Andrew Sharp, Donata Rimoldi, Felix Naef

Ultra-high throughput sequencing (UHTS) technology has been used to assess its ability to detect somatic mutations in clinical samples. The aim is to discover SNPs, short insertions and deletions, CNVs, and chromosomal rearrangements that contribute to the clinical phenotype. Currently, the analysis methodology we have developed allows us to detect SNPs and short insertions and deletions using Solexa reads and chromosomal rearrangements using 454 reads.

Our analysis pipeline has been applied to data derived from cDNA libraries and PCR amplified exons. The cDNA libraries were obtained from total mRNA extracted from HCC1954, a breast cancer cell line. The exon data were obtained by PCR amplification of defined genomic regions from selected patients. The cDNA and amplified DNA fragments were sequenced on a Solexa/Illumina Genome Analyzer using the protocol recommended by the manufacturer. The cDNA was also analyzed on 454/Roche.

Our analysis pipeline comprises two main components: a probabilistic base calling of the Solexa reads, using the image intensity files, and a semi-global alignment of the reads on the genome.

Probabilistic base calling

A significant proportion of the Solexa reads are routinely discarded due to the inability to match them to a reference sequence. We use model-based clustering and probabilistic framework to identify problematic base calls and low-quality reads. For each read, our base calling algorithm proposes an optimal sequence coded in the IUPAC extended alphabet. The length of these sequences is variable accounting for possible uncertain bases towards the end of the reads. We have shown that the method improves both genome coverage and the number of usable reads by an average of 15%, compared to Solexa's data processing pipeline (manuscript submitted for publication).

Semi-global alignment of the reads on the genome

Each short read is aligned using a semi-global alignment method (`align0` [1]) on the whole genome. Reads that map to multiple location (repeats) are discarded. All pairwise alignments are then converted to a multiple sequence alignment using the genomic sequence as the reference template, in which SNPs and short insertions and deletions are readily apparent. We are investigating whether the coverage of each genomic nucleotide might be used to detect CNVs and chromosomal rearrangements. So far, 454 reads give better results to detect chromosomal rearrangements.

The steps used to obtain all semi-global alignments in a reasonable time are:

1. map all reads on the genome using a fast, index-based, program named `fetchGWI` [2]
2. collate genome mapped reads into small regions (usually exons)
3. shred each region into 12-mers and select all remaining reads containing a matching 12-mer
4. use `align0` to align selected reads on the matching region
5. discard reads that find a better match on a genomic location outside of the region
6. prepare multiple sequence alignments using the remaining semi-global alignments

We believe that our analysis pipeline represents a useful improvement over the currently available commercial software both in terms of the efficiency with which the raw data can be interpreted and of the quality of the output.

References

[1] Myers and Miller, CABIOS (1989) 4:11-17
[2] Iseli et al, PLoS ONE, 2007

Taxonomical and Functional Characterization of Ultra-Short Reads from Microbial Metagenomes using the q-Gram Index

Wolfgang Gerlach

Bielefeld University – Faculty of Technology

March 14th, 2008

Abstract

Metagenomics is a new field of research on metagenomes, where natural microbial communities are studied. The new sequencing techniques like 454-sequencing [MEA+05] or Solexa-Illumina sequencing promise new possibilities as they are able to produce huge amounts of data in much shorter time and with less efforts and costs than the traditional Sanger-sequencing. But the data produced comes in even shorter reads (35-50 basepairs with Solexa-Illumina, 100-300 basepairs with 454-sequencing). CARMA [LNA+08] is a new pipeline for the characterization of the species composition and the genetic potential of microbial samples using 454-sequenced reads. The species composition can be described by classifying the reads into the taxonomic groups of organisms they most likely stem from. By assigning the taxonomic origins to the reads, a profile is constructed which characterizes the *taxonomic composition* of the corresponding community. The CARMA pipeline has already been successfully applied to 454-sequenced communities [DPS+08, SSD+08] including the characterization of a plasmid sample isolated from a wastewater treatment plant [Kra07].

Using samples from a biogas plant we examined the applicability of this approach for the ultra-short Solexa-Illumina reads by comparing the results with those obtained by the 454-sequenced sample (in submission). Our first results using 5.4 million 50 bp-reads revealed that this approach indeed produces consistent results. Most differences we have found are in the taxa of higher order, e.g. in the species level, and in general for species with a very low presence.

We do have several plans to improve the accuracy and speed of this method: A preprocessing assembly phase using an adapted q-gram index [Mye94, RSM05] to "increase" read length; Adaption of the pipeline to take the information of mated reads into account including the adaption of the amino acid sequence distance function for the construction of the phylogenetic tree; Implementation of a protein-q-gram index over a multiple alignment for the read-against-Pfam protein family matching.

References

[DPS+08] Elizabeth A. Dinsdale, Olga Pantos, Steven Smriga, Robert A. Edwards, Florent Angly, Linda Wegley, Mark Hatay, Dana Hall, Elysa Brown, Matthew Haynes, Lutz Krause, Enric Sala, Stuart A. Sandin, Rebecca Vega Thurber, Bette L. Willis, Farooq Azam, Nancy Knowlton, and Forest Rohwer. Microbial ecology of four coral atolls in the northern line islands. *PLoS ONE*, 3(2):e1584, Feb 2008.

[Kra07] Lutz Krause. *From single Genomes to natural microbial Communities: Novel Methods for the high-throughput Analysis of genomic Sequences*. PhD thesis, Bielefeld University, 2007.

[LNA+08] Krause L., Diaz N., Goesmann A., Kelly S., T. W. Nattkemper, Rohwer F., Edwards R., and Stoye J. Phylogenetic classification of short environmental dna fragments. *Nucleic Acids Research (NAR)*, 2008. Advance Access published online on February 19, 2008.

[MEA+05] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B. Dewell, Lei Du, Joseph M. Fierro, Xavier V. Gomes, Brian C. Godwin, Wen He, Scott Helgesen, Chun H. Ho, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. I. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum Kim, James R. Knight, Janna R. Lanza, John H. Leamon, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L. Lohman, Hong Lu, Vinod B. Makhijani, Keith E. Mcdade, Michael P. Mckenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan F. Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, Pengguang Yu, Richard F. Begley, and Jonathan M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, July 2005.

[Mye94] E. W. Myers. A sublinear algorithm for approximate keyword searching. *Algorithmica*, V12(4):345–374, October 1994.

[RSM05] Kim R. Rasmussen, Jens Stoye, and Eugene W. Myers. Efficient q-gram filters for finding all epsilon-matches over a given length. In Satoru Miyano, Jill P. Mesirov, Simon Kasif, Sorin Istrail, Pavel A. Pevzner, and Michael S. Waterman, editors, *RECOMB*, volume 3500 of *Lecture Notes in Computer Science*, pages 189–203. Springer, 2005.

[SSD+08] Stuart A. Sandin, Jennifer E. Smith, Edward E. DeMartini, Elizabeth A. Dinsdale, Simon D. Donner, Alan M. Friedlander, Talina Konotchick, Machel Malay, James E. Maragos, David Obura, Olga Pantos, Gustav Paulay, Morgan Richie, Forest Rohwer, Robert E. Schroeder, Sheila Walsh, Jeremy B. C. Jackson, Nancy Knowlton, and Enric Sala. Baselines and degradation of coral reefs in the northern line islands. *PLoS ONE*, 3(2):e1548, Feb 2008.

Statistically-corrected Counting of Short Reads for Digital Gene Expression Profiling

Doron Lipson, Tal Raz, Alix Kieu, Marie Causey, Ed Thayer
Helicos Biosciences, Cambridge, MA

Digital Gene Expression (DGE) profiling by next-generation sequencing (NGS) technologies has the potential of becoming the leading method for quantitative analysis of complete transcriptomes. While similar approaches have been implemented in the past by methods such as SAGE [Velculescu] and MPSS [Brenner], one of the key advantages of NGS technologies is an extremely high throughput which enables accurately quantifying very low-abundance transcripts [Kim].

Here we assume a sample preparation method that can generate reads from any part of the transcript sequence (e.g. [Nagalakashmi]). The task of transcriptome profiling by DGE generally consists of two steps: First, each read is aligned with the complete reference set using a pairwise sequence aligner, and the highest scoring alignments are determined. Then, significant alignments can be used to assign reads to transcripts and produce a count for each transcript. However, the typical short length and non-negligible error-rate of reads produced by NGS technologies inherently lower read specificity and may limit the detection limit of the measurement due to read misassignment to low-abundance transcripts.

We use simulation to demonstrate the effect of read misassignment on the specificity of DGE profiling, and introduce a statistical method for correcting this effect named **Read-Misassignment Corrected** counting (RMC counting). The essence of this method is to replace *hard* assignments of reads to transcripts with *soft* probability-based assignments (a concept that has previously been employed in different computational settings, e.g. for clustering [Zhang]). Our method considers all significant alignments between a given read and the entire reference library, and distributes the vote of each read between the transcripts based on both alignment p-values and the transcripts' abundance. Since the latter values are initially unknown the calculation is applied iteratively until it converges on a stable transcript distribution.

We apply the new method to both simulated and real DGE measurements of the transcriptome of *S. cerevisiae*, generated on a single-molecule sequencing platform [Harris]. While the naïve approach to read assignment (based on the highest-scoring alignment) results in significant over-counting of low-abundance transcripts, the application of the RMC counting method adequately resolves read misassignments, thereby significantly improving the detection limit of the technology and reducing the rate of false positive transcript detection.

References

Brenner S. et al., *Nat Biotechnol*. 18(6):597-8, 2000
Harris, T.D. et al, *Science* 320, 106-109, 2008
Kim J. B., *Science* 316(5830): 1481-4, 2007
Nagalakshmi, U. et al, *Science*, 2008
Velculescu V.E, *Science*, 270(5235):484-7, 1995
Zhang, B., Proceedings of the 1st SIAM ICDM, 2001

Personal Genomics Using the Next Generation Sequencing Technologies

Can Alkan[*], Fereydoun Hormozdiari[†], Gozde Aksay[*], Jeffrey M. Kidd[*],
Onur Mutlu[‡], S. Cenk Sahinalp[†], Evan E. Eichler[*]

There are currently multiple new sequencing technologies under development such as *pyrosequencing* (454), *sequencing-by-synthesis* (Illumina), and *sequencing-by-ligation* (SOLiD). These new, high-throughput sequencing strategies produce short reads, but also increases the coverage redundancy by 20-fold, or more.

The major limitation of these new sequencing methods is that they produce huge amounts of short sequence data of lower quality. New algorithms must be developed that take these sequence properties into consideration. In complex genomes, the presence of repeat sequences complicates the mapping and assembly of more than 50% of these short reads. Solutions to this problem exploit the fact that *micro reads* (reads of length $25 - 50$bp) can be acquired from both ends of an insert where the insert length follows a known distribution. Although the read-length is short, the overall throughput is enormous, since each run produces up to 60 million individual reads and yields $>$ 1Gb of sequence data. Given the massive volume of data being produced by the new platforms, data management and analysis becomes a major undertaking for those adopting the new platforms. Furthermore, genome rearrangement inference is possible with next generation sequencing technology, but new algorithms that can utilize the information provided by paired reads are required. Algorithms for mapping short reads to the genome are still in their infancy and are not yet optimized.

Here we present a set of computational methods we developed to study human genetic variation using the next-generation sequencing technologies. We first developed a tool to efficiently map micro reads to *all* possible locations in the human genome. We then used our tool to map reads generated by Illumina 1G Analyzer from CEPH trio samples as part of the 1000 Genomes Project (NA12878, NA12891, and NA12892), and an African sample (NA18507) to build maps of segmental duplication, single-nucleotide polymorphisms (SNPs), micro indels ($1 - 2bp$), and structural variation. We compare the segmental duplication maps with the similar maps we constructed from the Venter and Watson genomes; and the SNP, indel, and structural variation with experimentally validated sites for two individuals (NA18507 and NA12878). Our initial analyses suggest that the detection of SNPs, and structural and copy-number variants, using short-read double-ended sequences is feasible, fast, and reliable. We estimate that the paired-end approach coupled with the 445/Illumina/SOLiD approaches offers a significant enhancement in coverage. We plan to re-apply our algorithms and methods to reads generated with next-generation technologies from other individuals, and use experimental validation to guide us to further improve our techniques. If successful, this research can open the way for rapid and low-cost detection of structural variation and segmental duplication events, and in turn, their relationship to individual disease.

[*]Department of Genome Sciences, University of Washington, USA
[†]School of Computing Science, Simon Fraser University, Canada
[‡]Microsoft Research, USA

Fine-scale mapping of copy number alterations with next generation sequencing

Derek Y. Chiang*, Gad Getz*, David Jaffe, Xiaojun Zhao, Carsten Russ,
Chad Nusbaum, Matthew Meyerson, Eric S. Lander
*** These authors contributed equally to this work**.

Several major classes of somatic alterations in tumor genomes can instigate cancer, including mutations, copy number alterations and structural rearrangements. Recent advances in sequencing technologies have enhanced both the throughput and resolution of characterizing these somatic alterations. In this study, we demonstrate the accuracy of inferring copy number alterations from the shotgun sequencing of tumor genomes. We benchmarked the Illumina/Solexa 1G GenomeAnalyzer on three pairs of tumor cell lines, as well as their matched normals. For each cell line, between 360.4 Mb to 636.6 Mb were unambiguously mapped to the human genome, corresponding to between 0.17x and 0.29x effective coverage. We derived a statistical framework to evaluate differences in read counts between tumors and normals in genomic windows of arbitrary size. We also developed an agglomerative segmentation algorithm for mapping chromosomal breakpoints that demarcate regions of copy number gain and loss. We compared measurements of copy number alterations between next generation sequencing and Affymetrix 238K Sty arrays. Sequencing methods provided more accurate measurements of high-level amplifications, due to the saturation of copy number estimates on single nucleotide polymorphism arrays. The high density of sequence reads also enabled more precise mapping of the breakpoints of copy number alterations. In particular, sequencing at 0.23x coverage enabled breakpoint mapping of a confirmed homozygous deletion within 260 bp. Thus, massively parallel shotgun sequencing of genomic DNA detects copy number alterations with comparable sensitivities, with higher accuracy and precision.

Identifying Structural Variation Using Next Generation Sequencing Data

Seunghak Lee, Elango Cheran, Michael Brudno
University of Toronto

Recently, structural genomic variants have come to the forefront as a significant source of variation in the human population, however 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 for the detection of structural variants is the clone-end sequencing approach, pioneered in [1,2]. We expand on these by building a probabilistic framework for the automated identification of structural variants using clone-end sequencing. We are able to compute likelihood that matepairs explain the same structural variant [4]. Using hierarchical clustering with the likelihood as linkage affinity, we cluster matepairs such that matepairs in the same cluster support the same structural variant. In order to compute the likelihood for insertion and deletion clusters, we need to estimate the optimal size of insertion or deletion, which is achieved by gradient descent algorithm with Kullback–Leibler divergence. Also, we can use Kullback–Leibler divergence as an alternative measure of likelihood for indel clusters. Because each structural variant is supported by a number of clones we can assign p-values (the probability any given structural variant has been observed due to variation in the mate pair size) to each variant, and control our global false positives via False Discovery Rate.

We have applied our method toward the detection of structural variants in a donor provided by AB-SOLiD. The methodology of the original study could only detect homozygous structural variants, but in our analysis we use the individual clones from the donor, rather than the assembled genome, to detect both homozygous and heterozygous events. We identified (FDR 0.05) 2357 indel structural variants (376 insertions, ranging from 682bp to 1130bp and 1961 deletions, ranging from 1021bp to 1megabases), as well as 140 inversions between the donor and the reference human genome.

We compare our predictions with the structural variants identified in three previous studies on different donors: [1,2] used a similar approach while [3] analyzed through a different approach. There is a clear correlation between our results and the three previously available sets of structural variation (e.g. we find 36 indel variants that overlap the 241 indels found by [1], 274 of our deletions overlap one of 742 deletions identified by [2], and 136 of our indel variants overlap one of the 663 indels from [3], all p-values < 0.001 by permutations). Despite the strong correlation with previous results, the large majority of the variants we identify have not been characterized previously: only ~13% of the structural variants found by us overlap any variant annotated in the Database of Genomic Variants. While the overall amount of structural variation found by us is consistent with previous work, the fact that 87% of the events we find are novel suggests that the overall mosaic of structural variants varies widely between individuals.

[1] Tuzun et al, Nature Genetics 2005
[2] Korbel et al, Science 2007
[3] Levy et al, PLoS Biology 2007
[4] Lee et al, Bioinformatics 2008

A Scalable Short Read Assembler – AbySS

Inanc Birol, Jared Simpson, Shaun Jackman, Kim Wong and Steven Jones
Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada

This work describes a de novo assembly algorithm called ABySS, standing for Assembly By Short Sequences. Like most of the recent short read de novo assembly algorithms, ABySS is based on a de Bruin graph representation of sub-reads, but unlike available algorithms, it offers a scalable solution to the de novo assembly problem on commodity hardware and is designed to assemble large genomes. It achieves the feat through a novel partitioning of the sub-read space that optimizes contig assembly over a cluster of CPUs using a message passing interface (MPI) protocol.

We tested ABySS on Linux cluster nodes with AMD Opteron x86 64bit multi-core CPUs with 2GB of RAM per CPU. For 30-fold coverage simulated human chromosome 1 single-end 36bp read data, the assembly took about 5 hours on 20 CPUs, resulting in a maximum contig size of 10kbp and an n50 of 1kbp. Using 2.5 million Illumina 37bp paired-end Streptococcus pneumoniae reads, we assembled a maximum contig size of 53kbp and an n50 of 10kbp on a single CPU in less than an hour. Currently, we are attempting a human genome assembly with 30-fold coverage experimental data from the 1000 Genomes Project. We expect to assemble the data set, which contains 25bp, 36bp and 37bp single- and paired-end SOLiD and Illumina reads, in less than a day on about 200 CPUs.

We anticipate ABySS to fill in an important niche.

Short Read Fragment Assembly with EULER-SR

Mark Chaisson[1], Dumitru Brinza[2], and Pavel Pevzner[2]

1 Bioinformatics Program, University of California, San Diego.
2 Department of Computer Science, University of California, San Diego.

The field of Short Read Fragment Assembly has expanded recently, enabled by the availability of high-throughput short read sequencers, and motivated by biologists who are eager to assemble their new datasets. The conventional methods used to assemble long Sanger sequences have proved difficult to adapt to short reads. Fortunately, algorithms for short read fragment assembly were first considered nearly two decades ago in the context of Eulerian assembly for Sequencing by Hybridization. Many current short read assemblers, including EULER-SR, have been shown to work on short reads using the Eulerian method for fragment assembly.

The basis of Eulerian assembly is to find an Eulerian path through the de Bruijn graph constructed on a set of reads. However, it is difficult to construct the correct de Bruijn graph due to sequencing errors, and it is difficult to find the right Eulerian path even on the correct graph.

We will discuss the methods we use in EULER-SR to correct errors in reads prior to any assembly, and to detect and correct spurious vertices and edges in the de Bruijn graph once it is created. We have tuned a dynamic programming method called Spectral Alignment that uses high-frequency words to detect and fix errors in 100 nt. 454 reads, and an iterative method to fix errors in shorter, 25-35 nt. Illumina reads prior to assembly. After reads are corrected, we construct a de Bruijn graph. We remove spurious vertices and edges using two techniques called Erosion and Bulge Removal that identify and remove erroneous terminal edges and undirected cycles caused by remaining errors in reads.

Furthermore, contrary to the belief that low quality portions of reads should be discarded prior to assembly, we will show that it is possible to use extended, yet low quality reads in order to define paths in the de Bruijn graph to improve assembly quality. Most high-throughput short read sequencers produce high quality reads for several bases and then degrade in read quality along the length of a read. Our results show that as long as there is a short high quality prefix, reads may be sequenced to a length "past their prime", where error rates as high as 20% may be tolerated and used to improve final assembly nearly as much as error-free reads of the same length.

FuzzyPath – a hybrid de novo assembler using mixed Solexa and 454 short reads

Zemin Ning

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus,
Hinxton, Cambridge CB10 1SA UK

Pair-ending data from new sequencing platforms provides challenging, but exciting prospects for de novo assemblies for bacterial genomes. However, the wide size distribution of inserts in read pairs often brings unexpected consequences – shorter contigs and mis-assembly errors. We present a new computational strategy for a hybrid de novo assembler using mixed Solexa and 454 reads. The process consists of two distinctive steps: read extension and whole genome assembly. Firstly, raw sequencing Solexa and 454 reads are extended into segments of consensus sequences each with a maximum length of 2 kb. Sequence extension starts from kmer seeds which are randomly sampled to ensure overlaps among extended segments. To obtain genome assemblies, the new data set of the extended sequences with 10-15X coverage is processed and assembled using the Phusion assembler [1] - the previously developed capillary read assembly pipeline.

Before read extension, a hash table is constructed in a way that for each kmer node, values such as read index, kmer direction at 3' or 5', kmer offset, and link to the next kmer are stored. For paired-end reads, pair information is also processed for further use. Starting from selected kmer seeds only from Solexa data, extension proceeds one base at a time until these conditions are met: (1) multiple path; (2) no kmers for selection on the next move; and (3) extension length reaches 2 kb. In the case of multiple path, read pairs, 454 kmer links whenever available are used and path made by bases at kmer junctions with quality <= Q30 is ignored to guide the walk. To minimize the effect of homopolymers from 454 sequencing, kmer links from 454 reads are only used in repeat junctions where there is no unique path for Solexa kmers. When all the sampled kmer seeds are extended, walk is also carried out on the neighbourhood node from both forward and reverse directions to cover the junction nodes. These extra segments around the repetitive nodes are likely to close sequence gaps due to repeats, base errors, or polymorphisms.

The short read assembly pipeline has been tested for a variety of real data as well as simulation data. For a Prokaryotic genome of Streptococcus Suis (~2.0 Mb), single end reads of ~40X sequenced at the Sanger Institute, produced an assembly of 380 contigs with contig N50 at 8.55 kb. Adding 10X 454 data, the contig number is reduced to 73 and contig N50 is at 64 kb. The results of de novo transcriptome assemblies are also reported for *Plasmodium falciparum* and *Caenorhabditis elegans*.

[1] http://www.sanger.ac.uk/Software/production/phusion/

Consensus Computation on Segments

Tobias Rausch[1,2], Anne-Katrin Emde[1,2] and Knut Reinert[2]

1. International Max Planck Research School for Computational Biology and Scientific Computing, Ihnestr. 63 - 73, 14195 Berlin, Germany

2. Algorithmische Bioinformatik, Institut für Informatik, Takustr. 9, 14195 Berlin, Germany

Introduction: High-throughput sequencing technologies with short read data pose a new challenge to the current three-phase assembly methodology: Overlap-Phase, Layout-Phase, and Consensus-Phase. We describe a new consensus method that is robust in the face of high coverage, shorter reads, and genomic variation.

Methods: Given an initial layout of the reads, we generate a consensus sequence and a multi-read alignment with the following protocol: (1) Computation of all necessary (with respect to the layout) pairwise overlap alignments. (2) Extraction of all gapless alignment segments and generation of a segment-based weighted overlap graph (see Fig. 1). Conflicts between segment matches are resolved using a novel multiple segment match refinement algorithm [3]. (3) An adjustment of the edge-weights using a variant of the triplet extension pioneered in the T-Coffee package [2]. By means of the triplet extension we increase the weights of clique-edges within the overlap graph and thus, these edges are more likely to be chosen in the subsequent progressive alignment stage. (4) A progressive graph-based alignment of the reads using the heaviest common subsequence algorithm and a guide tree computed from the pairwise alignment scores. Note that our algorithm does not align single nucleotides but the segments identified in the overlap alignment phase. This ensures that columns with genetic variation (e.g., SNPs) are preserved. (5) Output of the multi-read alignment, the gapped consensus and all positioned reads with their respective deltas. We also support the AMOS message file *.afg format to visualize results with Hawkeye [4].

Results: We used a read simulator and real data from the NCBI trace archive to evaluate our consensus tool. The main parameters of the read simulator are the source sequence length, the average read length, the number of reads and the error rate per base call. In addition, multiple haplotypes can be simulated. Two further parameters, namely the number of SNPs and the number of indels, specify the genetic variation randomly introduced into these haplotypes. We performed two experiments: (1) Given a source sequence length of 10000, we simulated reads under different settings outlined in Fig. 2 and evaluated the consensus quality. (2) Given two haplotypes each of length 10000 with 100 SNPs and 5 Indels, we simulated reads of length 200, coverage 20 and 4% error rate. We then manually inspected the multi-read alignment with Hawkeye to evaluate the consensus in case of genetic variation. As an example Fig. 3 shows a region with two SNPs.

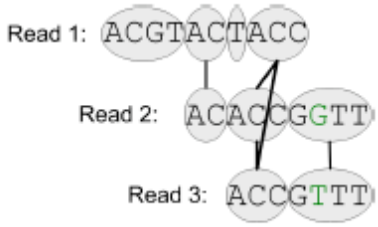

Figure 1: A segment-based alignment graph of three reads. The green-colored SNP is embedded in a segment and a clique is highlighted in bold font.

	2%	4%
35, 50x	**9994**/9994	**9996**/9996
200, 20x	**9954**/9954	**9928**/9928

Figure 2: Results on simulated data in various settings. Row names indicate the average read length and the coverage. Column names indicate the error rate. Each cell indicates how many consensus nucleotides were correctly identified out of all source sequence nucleotides covered by at least three reads (coverage > 2).

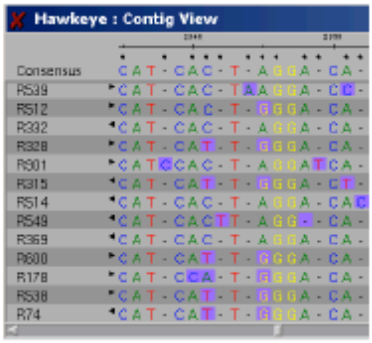

Figure 3: Clipped multi-read alignment view with 2 SNPs.

Conclusion: The results on simulated data are encouraging and preliminary results on real data show that our consensus quality is comparable to other tools. It remains to be shown that our program outperforms other tools in difficult settings, namely high coverage and short, error-prone read data. The consensus tool is part of the SeqAn library [1] and the read simulator is available on request: rausch@inf.fu-berlin.de.

References

[1] A. Döring, D. Weese, T. Rausch, and K. Reinert. SeqAn - An efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9:11, Jan 2008.

[2] C. Notredame, D.G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302:205–217, 2000.

[3] T. Rausch, A.-K. Emde, D. Weese, A. Döring, C. Notredame, and K. Reinert. Segment-based multiple sequence alignment. *Accepted for Publication*, 2008.

[4] M. Schatz, A. Phillippy, B. Shneiderman, and S. Salzberg. Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome Biology*, 8(3):R34, 2007.