### HiTSeq 2010: ISMB Special Interest Group on High Throughput Sequencing

#### Day 1: Friday, July 9th

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:45-10:15am:</td>
<td>Session 1</td>
<td>5</td>
</tr>
<tr>
<td>10:15-10:45am:</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>10:45am-12:30pm:</td>
<td>Session 2</td>
<td>6</td>
</tr>
<tr>
<td>12:30-1:45pm:</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>1:45pm-3:30pm:</td>
<td>Session 3</td>
<td>12</td>
</tr>
<tr>
<td>3:30-4:00pm:</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>4:00pm-5:45pm:</td>
<td>Session 4</td>
<td>19</td>
</tr>
</tbody>
</table>

#### Day 2: Friday, July 10th

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30-10:15am:</td>
<td>Session 5</td>
<td>25</td>
</tr>
<tr>
<td>10:15-10:45am:</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>10:45am-12:30pm:</td>
<td>Session 6</td>
<td>35</td>
</tr>
<tr>
<td>12:30-1:45pm:</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>1:45pm-3:30pm:</td>
<td>Session 7</td>
<td>40</td>
</tr>
<tr>
<td>3:30-4:00pm:</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>4:00pm-6:00pm:</td>
<td>Poster Session</td>
<td>44</td>
</tr>
</tbody>
</table>
Day 1: Friday, 9 July 2010

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:45 - 9:00</td>
<td>Welcome</td>
<td>Welcome</td>
</tr>
<tr>
<td>9:00 - 10:15</td>
<td>1.1 David Haussler</td>
<td>Keynote</td>
</tr>
<tr>
<td></td>
<td>1.2 Christian Gilissen</td>
<td>Prioritization of pathogenic variants for monogenic diseases using targeted and exome resequencing data</td>
</tr>
<tr>
<td>10:15 - 10:45</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>10:45 - 12:30</td>
<td>2.1 Vikas Bansal</td>
<td>Accurate genotyping of indels from population-scale short read sequence data</td>
</tr>
<tr>
<td></td>
<td>2.2 Wolfgang Gerlach</td>
<td>Taxonomic Classification of Metagenomic Reads with CARMA3</td>
</tr>
<tr>
<td></td>
<td>2.3 Sharmila Mande</td>
<td>Use of multiple compositional filters for rapid and accurate taxonomic classification of metagenomic sequences</td>
</tr>
<tr>
<td></td>
<td>2.4 Or Zuk</td>
<td>Detection of Rare-Alleles And Their Carriers Using Compressed Se(que)nsing</td>
</tr>
<tr>
<td></td>
<td>2.5 Tianjiao Chu</td>
<td>Statistical model for whole genome sequencing and its application to minimally invasive diagnosis of fetal genetic disease</td>
</tr>
<tr>
<td>12:30 - 1:45</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>1:45 - 3:30</td>
<td>3.1 Steffen Heber</td>
<td>Quantitative RNA-Seq to Investigate the Role of Alternative Splicing in Plant Immune Response.</td>
</tr>
<tr>
<td></td>
<td>3.2 Yin Hu</td>
<td>A Probabilistic Framework for Accurate Detection of Gene Fusion Events with Paired-End RNA-seq Reads</td>
</tr>
<tr>
<td></td>
<td>3.3 Andrew McPherson</td>
<td>Gene fusion discovery in paired end RNA-Seq data with application to ovarian cancer</td>
</tr>
<tr>
<td></td>
<td>3.4 Oliver Stegle</td>
<td>Statistical Tests for Detecting Differential RNA-Transcript Expression from Read Counts</td>
</tr>
<tr>
<td></td>
<td>3.5 Billy Chang</td>
<td>A Regression Approach to Fragmentation Bias Correction for Isoform Quantification in RNA-Seq.</td>
</tr>
<tr>
<td>3:30 - 4:00</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>4:00 - 5:45</td>
<td>4.1 Aaron Klammer</td>
<td>De novo assembly of Pacific Biosciences SMRT DNA sequences</td>
</tr>
<tr>
<td></td>
<td>4.2 Ole Schulz-Trieglaff</td>
<td>Assembling genomic variants and bacterial genomes using Illumina Genome Analyzer data</td>
</tr>
<tr>
<td></td>
<td>4.3 Moran Yassour</td>
<td>De novo assembly of RNAseq for transcriptome reconstruction and characterization from yeasts to human</td>
</tr>
<tr>
<td></td>
<td>4.4 Iman Hajirasouliha</td>
<td>Detection of locus and content of novel sequence insertions using paired-end next-generation sequencing</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td>Title</td>
</tr>
<tr>
<td>------------</td>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>8:30 - 10:15</td>
<td>4.5</td>
<td>Mark Chaisson: Mapping Reads with Insertion and Deletion Error using BLASR</td>
</tr>
<tr>
<td>8:30 - 10:15</td>
<td>5.1</td>
<td>Valentina Boeva: Control-free prediction of copy number alterations in deep-sequencing data using normalization on GC-content</td>
</tr>
<tr>
<td>8:30 - 10:15</td>
<td>5.2</td>
<td>Anamaria Crisan: Informing the inference of single nucleotide variants with copy number alteration annotations reveals novel somatic mutations in resequenced cancer genomes</td>
</tr>
<tr>
<td>8:30 - 10:15</td>
<td>5.3</td>
<td>Anna Ritz: Structural Variation Analysis with Strobe Reads</td>
</tr>
<tr>
<td>8:30 - 10:15</td>
<td>5.4</td>
<td>Jochen Blom: Using GPU programming for short read mapping</td>
</tr>
<tr>
<td>8:30 - 10:15</td>
<td>5.5</td>
<td>Gunnar Ratsch: RGASP Evaluation of RNA-Seq Read Alignment Algorithms</td>
</tr>
<tr>
<td>10:15 - 10:45</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>10:45 - 12:30</td>
<td>6.1</td>
<td>Sven Rahmann: Deep microRNA sequencing reveals differential expression in favorable versus unfavorable neuroblastoma tumors</td>
</tr>
<tr>
<td>10:45 - 12:30</td>
<td>6.3</td>
<td>Irina Abnizova: Reducing the False Positive SNP Rate for Illumina Sequencing: probability of second best call and context dependent errors</td>
</tr>
<tr>
<td>10:45 - 12:30</td>
<td>6.4</td>
<td>Chuck Tweedy: The effect on sequencing accuracy of DNA polymerase kinetics in real-time, single-molecule, sequence-by-synthesis methods (3rd generation)</td>
</tr>
<tr>
<td>10:45 - 12:30</td>
<td>6.5</td>
<td>Leena Salmela: Correction of sequencing errors in a mixed set of reads</td>
</tr>
<tr>
<td>12:30 - 1:45</td>
<td></td>
<td>Lunch</td>
</tr>
<tr>
<td>1:45 - 3:30</td>
<td>7.1</td>
<td>Marc Fiume: Savant: Genome Browser for High Throughput Sequencing Data</td>
</tr>
<tr>
<td>1:45 - 3:30</td>
<td>7.2</td>
<td>Gunnar Ratsch: Online Quantitative Transcriptome Analysis</td>
</tr>
<tr>
<td>1:45 - 3:30</td>
<td>7.3</td>
<td>John Carpten: Keynote</td>
</tr>
<tr>
<td>3:30 - 4:00</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>4:00 - 6:00</td>
<td></td>
<td>Poster Session</td>
</tr>
</tbody>
</table>

Day 2: Saturday, 10 July 2010
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Presenters</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.00pm – 6.00pm</td>
<td>Poster Session</td>
<td>John Archer</td>
<td>The analysis of temporal viral samples using ultra-deep pyrosequencing data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Markus J. Bauer</td>
<td>A pipeline for fast gapped read mapping of Illumina reads</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wayne Clarke</td>
<td>Comparison of SNP calling methods in 454 Sequencing data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nathan Clement</td>
<td>Probabilistic SNP mapping using GNUMAP-SNP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adrian Dalca</td>
<td>VARIID: Variation detection framework for color- and letter-space platforms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Matei David</td>
<td>Sensitive Mapping of Short Reads with SHRiMP2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nilgun Donmez</td>
<td>Assembly of highly polymorphic organisms with hybrid sequencing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Samuli Eldfors</td>
<td>Quantification of systematic bias in ChIP-seq signal profiles</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Angela Gonçalves</td>
<td>ArrayExpressHTS – Distributed Computing for RNA-Seq Data Processing and Quality Assessment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Catherine Grasso</td>
<td>ONCOSEQ: A Cancer Next Generation Sequence Data-Mining Platform</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ramneek Gupta</td>
<td>Targeted sequencing as a cost-effective technique for genotyping custom SNP panels: application on studying response to chemotherapy in childhood acute lymphoblastic leukaemia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kasper Hansen</td>
<td>Biases in Illumina transcriptome sequencing caused by random hexamer priming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adrian Heilbut</td>
<td>GATLING: Tools for Analysis of Structural Variants from Sequencing Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tobias Jakobi</td>
<td>Using GPU programming for short read mapping</td>
</tr>
<tr>
<td></td>
<td></td>
<td>David Kelley</td>
<td>Quality guided correction and filtration of errors in short reads</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Martin Kircher</td>
<td>Increasing the Genome Analyzer’s output using IBIS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Martin Kircher</td>
<td>High-throughput DNA sequencing – concepts and limitations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Klaus Maisinger</td>
<td>Recent advances in data analysis for Illumina sequencers and their application to the detection of somatic mutations in individual cancers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sharmilla Mande</td>
<td>An algorithm for rapid identification of identical and near-identical sequences in pyrosequencing data sets</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sharmilla Mande</td>
<td>An algorithm for <em>in silico</em> identification of host DNA sequences from host-associated metagenomic data sets</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Patrick Marks</td>
<td>Basecalling Models for Single Molecule Real Time DNA Sequencing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Andrew Menzies</td>
<td>Mining for cancer causing mutations in whole genome sequence data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lorena Pantano</td>
<td>User-friendly tool for the analysis of the non-miRNA transcriptome in high-throughput sequencing datasets</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hannes Ponstingl</td>
<td>SMALT – a new mapper for DNA sequencing reads</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shobha Potluri</td>
<td>Next Generation Sequencing of Prebeta-1 HDL Extremes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gunnar Rätsch</td>
<td>PALMapper: Fast and Accurate Spliced Alignments of RNA-seq Reads</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jon Sorenson</td>
<td>Unifying the Analysis of DNA Sequencing Across Platforms: Theory and Applications</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Isheng Jason Tsai</td>
<td>Improving draft assemblies by interactive mapping and assembly of short reads to eliminate gaps</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eric Tsung</td>
<td>A Novel Algorithmic Combined Approach to Accurately Determine Inversions and Large Deletions both in Confidence and in Genomic Location</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ernest Turro</td>
<td>MMSEQ: fast, scalable isoform expression estimation using multi-mapping RNA-seq reads</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Zhengyuan Wang</td>
<td>The challenges of de novo assembly from polymorphic populations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Zheng Zhang</td>
<td>A fast mapping tool for DNA sequence reads from the SOLiD™ System: part of Bioscope™</td>
</tr>
</tbody>
</table>
Prioritization of pathogenic variants for monogenic diseases using targeted and exome resequencing data


Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands

With the development of efficient enrichment strategies for next-generation sequencing it is now possible to either screen specific genomic loci (e.g. linkage regions) or even the entire human exome for genomic variations causing disease. Pinpointing the exact causative mutation is however hampered by the sheer amount of normal variation between individuals at the base pair level and by our limited knowledge on genotype-phenotype correlation. Although some biological hallmarks for pathogenic mutations exist they are not easily available for prioritizing large lists of candidate mutations as revealed by next generation sequencing. For this reason we developed a computational pipeline that prioritizes candidate mutations based upon biological relevant information. We applied this prioritization pipeline to two datasets; one using the NimbleGen array enrichment approach in combination with Roche 454 Titanium sequencing for analysis of all coding sequence in a 40Mb linkage region in two families with Familial Exudative Vitreoretinopathy (FEVR). In addition, we applied exome sequencing using the Agilent SureSelect enrichment in combination with ABI SOLiD 3.5 system in patients with Schinzel-Giedion Syndrome where we identified approximately 20,000 variants per exome. Up to 95% of these variants has previously been reported in SNP databases. Further prioritization of variants based on functional consequences, evolutionary conservation, and on combining data from multiple patients allowed us to rapidly identify TSPAN12 as the causative gene for FEVR and SETBP1 as the causative gene for Schinzel-Giedion Syndrome. These examples demonstrate the possibility of next-generation sequencing when combined with bioinformatics pipelines for the rapid identification of disease genes.


Accurate genotyping of indels from population-scale short read sequence data

Vikas Bansal

Scripps Genomic Medicine, Scripps Translational Science Institute, 3344 N. Torrey Pines Court, La Jolla CA 92037. vbansal@scripps.edu

Small insertions and deletions (1-100 bp) represent the second most frequent class of variation in the human genome after single nucleotide polymorphisms (SNPs). Sequencing of the complete genomes of several individuals using next-generation sequencing platforms has shown that each individual genome harbors ~ 200-400 thousand short indels. Although several tools for detecting indels from short read sequencing data are available, there is a need for new methods for genotyping indels directly from sequence reads. Accurate genotypes for indels can be particularly useful for sequencing-based association studies and population-scale analysis of genomic diversity.

We developed a computational pipeline, PIDGIN, to enable the accurate detection and genotyping of short indels from population-scale sequence data. Candidate indels were identified in each sequenced sample using a combination of gapped-alignment (using BWA), split read alignments and a de novo assembly based approach. Indels across the population of sequenced individuals were merged to generate a set of population indels. For each indel, a consensus sequence representative of all reads that could align to the non-reference allele was generated. Subsequently, the reads for each individual were re-aligned to an "indel-sensitive" reference sequence to determine the allele counts for each indel. Finally, the genotypes for each indel were called using a MCMC clustering algorithm that utilized the allele counts for each indel across the population under a binomial likelihood model. Indels that were artifacts of sequencing or alignment errors were automatically filtered out.

We demonstrated the utility of our method using two population sequencing datasets generated using the Illumina GA: (a) sequencing of 40 LR-PCR amplicons (targeting 190 kb of the human genome) in 289 individuals and (b) whole-genome sequencing of 40 methicillin-resistant Staphylococcus aureus (MRSA) strains. For the first dataset, PIDGIN identified ~ 100 indels including several long insertions and microsatellite variants. Linkage Disequilibrium (LD) and Hardy-Weinberg equilibrium analysis supported the accuracy of genotype calls for common indels. For the MRSA data, genotypes for the 170 indels identified by PIDGIN were consistent with a SNP-based phylogeny for the strains, providing indirect evidence for the accuracy of the genotype calls.
Taxonomic Classification of Metagenomic Reads with CARMA3

Wolfgang Gerlach and Jens Stoye

The vast majority of microbes can not be cultivated in a monoculture and thus can not be sequenced by means of traditional methods. To explore these microbes, they have to be analyzed within complex microbial communities. The new high-throughput sequencing (HTS) technologies like Roche's 454-sequencing, ABI's SOLiD or Illumina's Genome Analyzer make it possible to sequence microbial DNA samples of such communities. Due to the restricted read length currently produced by the different HTS technologies, reconstruction of complete genomic sequences is too difficult. Though, by comparing the metagenomic fragments with sequences of known function, it is possible to analyze the biological diversity and the underlying metabolic pathways in microbial communities. For this task we have developed CARMA [1] as well as WebCARMA [2], a refined version of CARMA available as a web application for the taxonomic and functional classification of unassembled (ultra-)short reads from metagenomic samples.

To infer the taxonomic origin of metagenomic reads, two kinds of methods, composition-based and comparison-based, can be distinguished. The composition-based methods extract sequence features like GC content or k-mer frequencies, and compare them with features computed from reference sequences with known taxonomic origin [3]. The comparison-based methods, in contrast, rely on homology information obtained by database searches.

Probably the most basic method for comparison-based classification is to use BLAST to search for the best hit in a database of sequences with known origin [4]. Since the evolutionary distance between the source organisms of the metagenomic fragment and the database sequence is unknown, a classification result based solely on a best BLAST hit has to be interpreted carefully. In general, such a classification is more reliable on higher taxonomic levels (e.g. Superkingdom or Phylum) than on lower taxonomic levels (e.g. Genus or Species), but it is difficult to decide which taxonomic level is reliable enough, as this strongly varies for each metagenomic fragment.

The program MEGAN [5] is based on the lowest common ancestor (LCA) method. A BLAST search is performed and all BLAST hits that have a bit score close to the bit score of the best hit (top-percent filter, e.g. 5 percent) are collected. The metagenomic fragment is then classified by computing the lowest common ancestor of all species in this set. One of the reasons that classifications based on this approach are more accurate, is that fragments with ambiguous hits are assigned at higher taxonomic levels.

The SOrt-ITEMS [6] algorithm extends the LCA method. A first initial BLAST search for the metagenomic fragment Q is performed and hits \{A,B,C,...\} are collected (similar to MEGAN). Then a new BLAST database containing \{Q,A,B,C,...\} is built. The best BLAST hit A from the initial BLAST search is now used as a query sequence to search in the new database. From this new BLAST result all hits that yield a score higher or equal to Q are collected. Finally, the LCA from this set is computed as usual to classify Q.

Assuming a model of evolution where different gene families have different rates of mutations, but within each family this rate does not change too much, the effect of the reciprocal method is the
following: All hits that have a better BLAST match to A than Q, are likely to be evolutionarily closer related to A than Q is. Thus, the LCA of these hits indicates the last point in the evolution where the ancestor of metagenomic fragment Q should have separated from the ancestor of A. The problem is that Q might also have been separated earlier from the ancestor of A, but the LCA provides only a lower bound. Reciprocal hits that have a worse BLAST match than Q could provide additional information, but are ignored by this method.

Inspired by this observation, we developed a new algorithm based on the reciprocal method. We first assign each reciprocal hit to the lowest taxonomic rank at which it shares an ancestor with A. Then we measure for each taxonomic rank (Superkingdom, Phylum, Class, Order, Family, Genus and Species) the range (or interval) of scores of the assigned hits. When a taxonomic rank has no hits assigned, but some higher rank has, missing scores can be filled up using linear interpolation. To classify Q, the best fitting taxonomic rank is in principle found by choosing the lowest taxonomic rank whose scores are still all below that of Q.

The algorithm is part of CARMA 3 and has been adapted to work with BLAST as well as with HMMER3 for fast and sensitive detection of protein encoding DNA fragments. For the HMMER3 variant, Pfam multiple alignments are used to compute the reciprocal scores. CARMA 3 provides fast and accurate functional and taxonomic classifications of metagenomic fragments. On publication it will also be made accessible via WebCARMA (http://webcarma.cebitec.uni-bielefeld.de).

References:


Use of multiple compositional filters for rapid and accurate taxonomic classification of metagenomic sequences

Bio-Sciences Division, Innovation Labs, Tata Consultancy Services Limited, Hyderabad, India

Taxonomic classification of metagenomic sequences is the first and most crucial step in any metagenomic analysis. The current taxonomic binning approaches can be broadly divided into two major categories namely, similarity-based and composition-based. Similarity-based binning approaches, though accurate and specific, are slow as they involve generating alignments of all query sequences with each sequence in the reference database. Given that current metagenomic sequencing projects generate millions of sequences, for a research group having modest computational resources, taxonomic classification using similarity based approaches is virtually infeasible. In this study, we present a novel composition-based (consequently alignment-free) binning approach that is not only rapid in execution, but also has an accuracy and specificity of binning comparable to existing similarity-based binning approaches.

The current method uses a unique work-flow which incorporates two layers of composition-based filters. The first layer identifies a set of organisms, the genomes of which have an oligonucleotide composition similar to the composition of the query sequence. Subsequently, the second layer scores the query against the precomputed 'Mixed Memory Markov Models' [1] generated for the identified set of organisms. The query sequence is finally assigned to the organism having the highest score. In cases, where comparable scores are obtained with models corresponding to multiple organisms, the query is assigned to a taxonomic clade using a 'most common phylogeny' approach.

Validation results using data sets/database variants that simulate typical metagenomic scenarios indicate that the accuracy and specificity obtained using this approach is comparable to existing alignment-based approaches. In addition, the binning time is observed to be an order of magnitude lower than existing alignment-based approaches.

With the availability of faster and cheaper sequencing technologies, the volume of sequence data generated from metagenomic sequencing projects is expected to drastically increase. Performing taxonomic classification of these huge metagenomic data sets in a reasonable amount of time will require huge computing resources, thereby limiting the scope of metagenomic studies to a few 'select' labs having the requisite computational infrastructure. Given the observations in this study, coupled with the fact that the current algorithm can be executed on a simple desktop with modest hardware specifications, the given algorithm is expected to be of immense utility for metagenomic research groups having limited computational resources.

References:

Detection of Rare-Alleles And Their Carriers Using Compressed Se(que)nsing

Noam Shental, Amnon Amir, Or Zuk

Detection of rare variants by resequencing is important for the identification of individuals carrying disease alleles. Rapid sequencing by new technologies enables low-cost resequencing of target regions, although it is still prohibitive to test more than a few individuals. In order to improve cost trade-offs, it has recently been suggested to apply pooling designs which enable the detection of carriers of rare alleles in groups of individuals. However, this was shown to hold only for a relatively low number of individuals in a pool, and requires the design of pooling schemes for particular cases. We propose a novel pooling design, based on a compressed sensing approach, which is both general, simple and efficient. We model the experimental procedure and show via computer simulations that it enables the recovery of rare allele carriers out of larger groups than were possible before. Our approach can also be combined with barcoding techniques to enhance performance and provide a feasible solution based on current resequencing costs. For example, when targeting a small enough genomic region (~100 base-pairs) and using only ~10 sequencing lanes and ~10 distinct barcodes per lane, one recovers the identity of rare allele carriers out of a population of over 4000 individuals.
Statistical model for whole genome sequencing and its application to minimally invasive diagnosis of fetal genetic disease

Tianjiao Chu 1,2, Kimberly Bunce 1,2, W. Allen Hogge 1,2 and David G. Peters 1,2

1Department of Obstetrics, Gynecology and Reproductive Sciences, University of Pittsburgh and 2Center for Fetal Medicine, Magee-Womens Research Institute, Pittsburgh, PA 15213, USA

There is currently great interest in the development of methods for the minimally invasive diagnosis of fetal genetic disease using cell-free DNA from maternal plasma samples obtained in the first trimester of pregnancy. With the rapid development of high-throughput sequencing technology, the possibility of detecting the presence of trisomy fetal genomes in the maternal plasma DNA sample has recently been explored. The major concern of this whole genome sequencing approach is that, while detecting the karyotype of the fetal genome from the maternal plasma requires extremely high accuracy of copy number estimation, the majority of the available high-throughput sequencing technologies require polymerase chain reaction (PCR) and are subject to the substantial bias that is inherent to the PCR process. We introduce a novel and sophisticated statistical model for the whole genome sequencing data, and based on this model, develop a highly sensitive method of Minimally Invasive Karyotyping (MINK) for the diagnosis of the fetal genetic disease. Specifically we demonstrate, by applying our statistical method to ultra high-throughput whole sequencing data, that trisomy 21 can be detected in a minor (‘fetal’) genome when it is mixed into a major (‘maternal’) background genome at frequencies as low as 5%. This observation provides additional proof of concept and justification for the further development of this method towards its eventual clinical application. Here, we describe the statistical and experimental methods that illustrate this approach and discuss future directions for technical development and potential clinical applications.
Quantitative RNA-Seq to Investigate the Role of Alternative Splicing in Plant Immune Response

Steffen Heber1,2, Brian Howard2, Xiaping Tan3, Paola Veronese3

1Department of Computer Science, 2Bioinformatics Research Center, 3Center for Integrated Fungal Research, North Carolina State University, Raleigh, NC

Alternative splicing (AS) plays a crucial role in the animal immune system [1], and several lines of evidence also link AS to the plant defense machinery [2,3]. The goal of this NSF funded project is to assess the extent and functional significance of AS in the plant immune response. Using paired-end short-read transcriptome sequencing in a time course experiment, we are investigating interactions of Arabidopsis thaliana with virulent and avirulent strains of the bacterial pathogen Pseudomonas syringae pv tomato. The results will provide a first approximation of the extent of pathogen-induced AS and its impact on proteins associated with host disease resistance. A crucial component of our project is developing methods for the accurate estimation of splice-isoform expression levels from RNA-Seq data. Our isoform quantification method uses a linear models framework to estimate the ratios of known transcripts in a sample. The method is flexible enough to accommodate a variety of sequencing technologies, including those that incorporate paired reads. A key advantage of our approach is that it takes into account the non-uniformity of RNA-Seq read positions along the targeted transcripts. We anticipate that the ability to reliably compute quantitative isoform expression values will help us to separate true alternative splicing events from spurious transcripts originating from single mis-spliced transcripts -- a major problem in genome-wide AS studies. Based on recent studies [4,5] and our own observations derived from publicly available Arabidopsis RNA-Seq datasets, isoform expression level changes frequently involve a continuum of isoform ratios, in addition to all-or-nothing expression patterns. Furthermore, in the datasets we have examined, isoform expression level changes appear to be independent of gene expression changes. This suggests the existence of a so far uninvestigated, dynamic layer of the transcriptome. Therefore, we believe that, in the future, researchers will prioritize genes for functional analysis based not only on observed changes in gene expression levels, but also on changes in alternative splicing.

References

A Probabilistic Framework for Accurate Detection of Gene Fusion Events with Paired-End RNA-seq Reads

Yin Hu1, Kai Wang1, Xiaping He3, Derek Y. Chiang3, Jan F. Prins2, Jinze Liu1.
1Department of Computer Science, University of Kentucky
2Department of Computer Science, University of North Carolina at Chapel Hill
3Department of Genetics, University of North Carolina at Chapel Hill

High-throughput sequencing technologies are providing unprecedented visibility into the mRNA transcriptome of a cell. In cancer, alternative splicing and gene fusion events are common changes observed in the mRNA transcriptome. Cancer specific splicing events are promising biomarkers and targets for diagnosis, prognosis, and treatment purposes.

Figure 1. A set of gene fusion events discovered by RNA-seq paired end reads generated from two cancer cell lines: MCF-7 and SUM-102. Red links refer to the gene fusion events specific to MCF-7 validated by Maher et al. (PNAS 2009). Blue links refer to two additional gene fusion events detected in MCF-7. Green links refer to the predicted gene fusion events in SUM102. The sequence similarity between the two genes involved in any predicted events is less than 35%.

Our work focuses on the alignment of RNA-seq reads generated from paired-end protocols to the reference genome for novel splice or fusion junction discovery. The larger implicit length of pair-end reads (PERs) compared to single reads makes them more likely to reveal accurate long range alternative splicing or gene fusion events. However, these PERs are typically discarded by existing
alignment methods (TopHat, SpliceMap) due to the difficulties in resolving the correct pairing of end reads that are highly separated when aligned to a reference genome.

We propose a novel probabilistic framework for aligning RNA-seq PERs to a reference genome, featured by identifying not only sequenced alignment of PER end reads but also unsequenced alignment between PER end reads. The goal is to achieve accurate alignment of a PER fragment regardless of the distance between the mapped locations of the end reads. The alignment, in turn, predicts both gene fusion events as well as regular local splicing events. Our approach sets up a graphical model to represent all putative splicing and fusion events derived from individual end read alignments. The alignment of PER fragments is hence turned into the problem of determining accurate paths within the graph. An expectation-maximization algorithm is then applied to identify the alignment of each PER, which maximizes the likelihood of all sampled data. Our approach was benchmarked on two RNA-seq datasets of 2x35bp PER reads from MCF-7 and SUM-102, two well known breast cancer cell lines. The inferred PER alignments increased the original single end coverage by an average of three-fold. The accuracy of the alignments was validated by qRT-PCR experiments on 16 exon skipping events. We also used these PERs to validate a set of putative fusion junctions identified by a set of 75bp RNA-seq data from the same samples. Our results demonstrate high specificity of gene fusion detection by comparison to confirmed cell-line specific fusion events.
Gene fusion discovery in paired end RNA-Seq data with application to ovarian cancer

Andrew McPherson(1,2), Sohrab Shah(1), Cenk Sahinalp(2), David Huntsman(1)
1. BC Cancer Agency, 2. Department of Computer Science, Simon Fraser University

Background: Gene fusions created by somatic genome rearrangements are known to play an important role in the onset and development of some cancers. The RNA products of these gene fusions often reside in the transcriptomes of tumor cells and are thus detectable by the newly developed paired end RNA-Seq platform. However, the effective discovery of gene fusions in RNA-Seq data requires the explicit consideration of several confounding factors such as alternative splicing and repetitive sequence. The scale of RNA-Seq data is such that naive approaches often produce an unmanageable number of predictions, many of them likely to be false positives. To date there has not been a comprehensive attempt to deal with these problems, and provide computational solutions with explicit performance guarantees.

Methods: We have developed deFuse, a multistep method for gene fusion discovery in RNA-Seq data that deals with the confounding factors of alternative splicing and repetitive sequence. The method we propose considers all alignments of paired end reads to cDNA and unspliced gene regions. We begin by using the principle of maximum parsimony to assign each discordant paired end read to a single alignment location implying a single fusion event. Applying the maximum parsimony principle has the benefit of resolving, simultaneously, the two genes involved in a gene fusion and the splice variant for each of those genes. We show that using all alignments and resolving the actual alignment location by maximum parsimony results in a more sensitive method when compared to methods that use only single best-hit alignments. For the next step, clusters of paired end reads supporting the same fusion event are then used to restrict the search space of a dynamic programming based split read analysis for fusion boundary prediction. We show that our split read approach is sensitive to fusion boundaries at non-canonical splice sites unlike approaches that align to combinations of known exons. Finally, we apply novel filtering metrics to the results of our algorithms to reduce the false discovery rate. We show that our approach has a high specificity for detecting real fusion events.

Results: We have applied deFuse to 45 RNA-Seq datasets derived from 41 ovarian tumour samples, an ovarian cancer cell line, an epithelial sarcoma cell line, a hodgkins lymphoma cell line, and an unclassifiable sarcoma tumor sample. An initial set of validations by PCR was performed on 23 fusion predictions including 1 putative promotor exchange and 11 open reading frame preserving fusions. All 23 were validated by PCR. Re-running the analysis using only uniquely aligning reads resulted in a 34% reduction in evidence for the 23 validated fusions, with 3 of those fusions, all open reading frame preserving, filtered due to lack of evidence. Of the 23 validated fusions, 3 of those fusions do not have fusion boundaries that coincide with canonical exon boundaries and are therefore not discoverable using a more simplistic split read analysis that involves aligning reads to combinations of known exons.

Conclusions: Our results show that discarding non-unique alignments will eliminate valuable evidence of real gene fusions. We also show that a more sophisticated split read alignment algorithm will increase the sensitivity of fusion prediction over methods that simply align reads to combinations of known exons.
Statistical Tests for Detecting Differential RNA-Transcript Expression from Read Counts

Oliver Stegle 1,2,*, Philipp Drewe 3,*, Regina Bohnert 3, Karsten Borgwardt 1,2, Gunnar Rätsch 3
1 Max Planck Institute for Developmental Biology, Spemannstraße 35, 72076 Tübingen, Germany
2 Max Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany
3 Friedrich Miescher Laboratory of the Max Planck Society, Spemannstraße 39, 72076 Tübingen, Germany

Motivation
As a fruit of the current revolution in sequencing technology, transcriptomes can now be analyzed at an unprecedented level of detail. These technological advances have been exploited in diverse ways. Examples of such include the detection of differentially expressed genes across biological samples, and the quantification of the abundances of various RNA transcripts within single genes. A natural next step is now to extend the detection of differential abundance, focusing on individual transcripts within one gene. However, explicit strategies to solve this problem have not yet been defined. This task is particularly challenging, if the transcript annotation is incomplete or incorrect.

Results
In this work, we present two novel statistical tests to address this important methodological gap: a ‘gene-structure-sensitive’ Poisson test for detecting differential expression when the transcript structure of the gene is known, and a non-parametric kernel-based test, called Maximum Mean Discrepancy (MMD), when the transcripts expressed in a region are yet unknown. Both approaches directly use the read alignments and do not rely on previous quantifications of transcripts that are often inaccurate. We analyzed both proposed methods on simulated read data from two artificial samples as well as on factual reads generated by the Illumina Genome Analyzer for two C. elegans samples [2]. Our analysis shows that the Poisson test identifies genes with differential transcript expression considerably better than approaches based on RNA transcript quantification (Figure 1). Even more striking, the MMD test is able to detect a large fraction (75%) of true differential cases in the absence of any knowledge of the annotated transcripts. This method is therefore very well suited to analyze RNA-Seq experiments where other approaches are doomed to fail, namely when the genome annotations are incomplete, false or entirely missing.

It is current work to extend both the Poisson model and the non-parametric MMD test to cope with multiple replicates in order to better account for technical and biological variation. We will expand on our preliminary analysis of a recent C. elegans development study [2], comprehensively investigating differential transcript expression in settings where current methodology cannot be applied.
Figure 1: Receiver operator curve for detecting differential transcript expression (left) and test sensitivity as a function of the number of differential reads (right) for alternative testing approaches.

References


A Regression Approach to Fragmentation Bias Correction for Isoform Quantification in RNA-Seq.

Billy Chang (billy.chang@utoronto.ca, Dalla Lana School of Public Health), Rafal Kustra (r.kustra@utoronto.ca, Dalla Lana School of Public Health), Quaid Morris (quaid.morris@utoronto.ca, Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research)

University of Toronto, Toronto, Ontario, Canada, M5S 1A1

Isoform quantification, a crucial step towards any isoform expression analysis, remains a challenge even in the advent of cutting-edge transcriptome analysis protocol such as RNA-Seq due to reads which can be ambiguously aligned to multiple isoforms. The problem is further complicated by the non-uniformity of within-isoform read distribution due to the isoform fragmentation procedure prior to sequencing. To account for fragmentation bias, one must estimate the within-isoform read distribution, a challenging task due to alternative splicing, low read coverages for lowly expressed isoforms, and varying read distributions for isoforms with different lengths.

We will demonstrate a new parametric regression approach for estimation of the fragmentation pattern. By regressing the within-isoform read alignment location onto the isoform's length, we can provide isoform-length specific fragmentation estimate. An added benefit is that of borrowing strength across isoforms: fragmentation estimation for lowly expressed isoform can utilize information from other isoforms with similar length. This parametric fragmentation estimator is then embedded into a mixture-model based isoform quantification algorithm which also accounts for ambiguously aligned reads and read errors. Results on simulated data based on reads generated from the Ensembl reference isoforms demonstrates the improved quantification precision on our approach. In particular, abundances for isoforms with strongly skewed fragmentation pattern tend to be more accurately estimated. Results based on real data will also be discussed.
**De novo assembly of Pacific Biosciences SMRT™ DNA sequences**

*Aaron Klammer1, Susan Tang1, Ali Bashir1, Mark Chaisson1, Eric Schadt1, Sam Phattarasukol2, Carrie Harwood2, Jon Sorenson1*

1Pacific Biosciences, Inc. Menlo Park, CA  
2Department of Microbiology, University of Washington, Seattle, WA

We present the hybrid assembly of a bacterial genome using reads from Pacific Biosciences’ single-molecule real-time (SMRT™) DNA sequencing technology. Our approach is based on the AMOS assembly software package and employs several novel algorithms tailored to Pacific Biosciences’ long reads.

For overlap detection we use the algorithm BLASR (see related poster) that provides substantial flexibility and sensitivity in overlapping reads of widely disparate lengths and error patterns (e.g. hybrid of PacBio and Illumina reads). For layout, we demonstrate a parallelized version of the AMOS layout algorithm tigger, and for consensus we present the algorithm EviCons which employs a probabilistic graphical model to represent the error characteristics of PacBio reads.

Using this suite of algorithms we are able to produce a hybrid assembly of a strain of *Rhodopseudomonas palustris* using 30X sequencing coverage with a significantly longer contig N50 than a comparable assembly using Illumina reads alone.

In addition, we scaffold the *R. palustris* assembled contigs using sequence generated by the PacBio strobe sequencing technology—a sequencing protocol which allows the linkage of multiple reads across large distances, in a fashion similar to mate-pair sequencing. Using these reads we are able to significantly increase the scaffold N50 for the *R. palustris* genome.

The use of PacBio long reads and strobe reads promises to greatly simplify the completion of draft and finished bacterial genomes.
Assembling genomic variants and bacterial genomes using Illumina Genome Analyzer data

Ole Schulz-Trieglaff 1, Bret Barnes 2, Jonathan Boutell1, Anthony J. Cox 1, R. Keira Cheetham 1, Russell Grocock 1, Sean Humphray 1, Lisa J. Murray 1, Roberto Rigatti 1, Christopher T. Saunders 2, Wendy Wong 1, Geoff Smith 1 and Dirk J. Evers 1

The Illumina Genome Analyzer (GA) platform currently generates paired reads with a length of 100 bp and above. Recent publications demonstrate the versatility of these reads for the de-novo assembly of large and complex genomes, including human [2, 4], giant panda [3] and the cucumber [1]. Assembly is a computationally demanding task and a challenging theoretical problem which has created much interest in the scientific community.

In this presentation, we highlight two applications of de-novo assembly: the reconstruction of inserted and deleted sequences (indels) in resequencing experiments and the high-quality assembly of difficult, GC-rich bacterial genomes such as Mycobacterium tuberculosis using GA sequencing reads.

We present an algorithm called GROUPER, for Guided Reassembly Of Unaligned Paired-End Reads, which can detect insertions and deletions of between 1 and 100bp in length between a reference sequence and a set of paired-end DNA reads. This size range addresses a gap that exists in the size spectrum of genetic variants between those detectable by a gapped alignment of individual reads and those detectable by methods based on anomalous alignment of paired-end reads. GROUPER incorporates information from reads spanning the position of the variant and groups them into a cluster. The algorithm assembles the reads in each cluster using a de Bruijn graph approach and compares the reconstructed sequences to the reference and thus determines type, position and sequence of the variant.

To characterize the specificity and sensitivity of GROUPER, we applied the algorithm to simulated reads from human chromosome 20, showing it compares favorably to other published methods. To test performance on real human genomic data, we run our algorithm on published datasets of 35 and 100 bp paired reads from the NA18507 Yoruba individual. GROUPER is part of Illumina's CASAVA (Consensus Assessment of Sequence and Variation) [5] software and is freely available on request.

The second part of the presentation focuses on the de-novo assembly of bacterial genomes. As an example, we present high-quality assemblies of the Mycobacterium tuberculosis (MTB) H37Rv strain. MTB is the causative agent of most cases of tuberculosis in Human. Its genome consists of more than 4 million base pairs and has an average GC content of 65% with some genomic regions reaching a GC content of more than 80%. We sequenced strain H37Rv using paired 100bp reads and assembled the genome using the Velvet assembler [6]. The resulting assembly is close to an optimal assembly obtained from simulations and covers the non-repetitive parts of the genome with high quality. We evaluate contig quality and coverage of the reference genome and highlight applications of the assembly for genotyping and probing of drug-resistance genes.


1Illumina Cambridge, Chesterford Research Park, Cambridge CB10 1XL, United Kingdom.  
2Illumina San Diego, 9885 Towne Centre Drive, CA-92121 San Diego, United States.
De novo assembly of RNAseq for transcriptome reconstruction and characterization from yeasts to human

Moran Yassour1,2,3, Manfred Grabherr1, Joshua Levin1, Mike Berger1, Pamela Russell1, Jessica Alfoldi1, Andi Gnirke1, Federica di Palma1, Kerstin Lindblad-Toh1,4, Nir Friedman3,5 and Aviv Regev1,2,6

1 Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, 2 Department of Biology, Massachusetts Institute of Technology, 3 School of Engineering and Computer Science, Hebrew University, Jerusalem, Israel, 4 Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden 5 Alexander Silberman Institute of Life Sciences, Hebrew University, Jerusalem, Israel, 6 Howard Hughes Medical Institute.

Experimentally defining the complete transcriptome of eukaryotic organisms has traditionally been a challenging task, but advances in sequencing RNA (RNAseq) offer new and powerful approaches to the study of transcriptomes. Recent studies have used RNAseq to quantify the expression levels of known genes, identify splice isoforms and refine gene boundaries. However, many studies depend on existing annotation, limiting the ability of discovering novel transcripts, and most require mapping to an available genome sequence, limiting their applicability to organisms without a sequenced genome, complex environmental samples, and cancer.

Here, we present a novel approach for de novo assembly of a transcript catalog from read data alone. At the heart of our approach is a novel algorithm that takes read data and generates a host of assembly graphs, each one ideally corresponding to a single transcript. Our algorithm then extracts from each graph one or more transcript isoforms, quantifies their levels, and scores their confidence. These transcripts can then be mapped to a reference genome, from the same organism or a related species.

We show how these approaches scale to organisms from yeasts to vertebrates, helping in genome annotation of newly discovered organisms from the Schizosaccharomyces clade, transcriptome analysis in the Anolis grahami lizard for which the genome sequence is not available, and for the discovery of novel fusion transcripts in human cancers.
Detection and characterization of novel sequence insertions using paired-end next-generation sequencing

Iman Hajirasouliha 1, Fereydoun Hormozdiari 1, Can Alkan 2,3, Jeffrey M. Kidd 4, Inanc Birol 1,5, Evan E. Eichler 2,3,* and S. Cenk Sahinalp 1,*

1 Lab for Computational Biology, Simon Fraser University, Burnaby, BC, Canada, 2 Department of Genome Sciences, University of Washington, 3 Howard Hughes Medical Institute, Seattle, WA, 4 Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA and 5 BC Cancer Agency, Genome Science Center, Vancouver, BC, Canada

Motivation: In the past few years, human genome structural variation discovery has enjoyed increased attention from the genomics research community. Many studies were published to characterize short insertions, deletions, duplications and inversions, and associate copy number variants (CNVs) with disease. Detection of new sequence insertions requires sequence data, however, the ‘detectable’ sequence length with read-pair analysis is limited by the insert size. Thus, longer sequence insertions that contribute to our genetic makeup are not extensively researched.

Results: We present NovelSeq: a computational framework to discover the content and location of long novel sequence insertions using paired-end sequencing data generated by the next-generation sequencing platforms. Our framework can be built as part of a general sequence analysis pipeline to discover multiple types of genetic variation (SNPs, structural variation, etc.), thus it requires significantly less-computational resources than de novo sequence assembly. We apply our methods to detect novel sequence insertions in the genome of an anonymous donor and validate our results by comparing with the insertions discovered in the same genome using various sources of sequence data.

Availability: The implementation of the NovelSeq pipeline is available at http://compbio.cs.sfu.ca/strvar.htm
Mapping Reads with Insertion and Deletion Error using BLASR

Mark Chaisson, Jon Sorenson
Pacific Biosciences, Inc. Menlo Park, CA

An important application of Single Molecule Real Time (SMRT™) Sequencing is whole genome shotgun resequencing. The first step in such a project is to map every read to the homologous position in a reference genome from which it is sampled, typically done by selecting the highest-scoring local alignment of a read and the genome. The reads differ from the reference genome by both mutations in the sample genome and sequencing errors. Since an exhaustive search across all local alignments is prohibitively expensive, it is important that a heuristic for local alignment be sensitive to both modes of mutation and error. SMRT Sequencing produces variable length reads with insertions and deletions as the dominant mode of sequencing error. Because current mapping routines have focused on alignment of sequences with only mismatches, we have developed an alignment routine, Branching Local Alignment with Successive Refinement (BLASR) for aligning SMRT Sequences to a reference genome. A branching search on a suffix array (for optimal speed) or Burrows Wheeler transform (for minimal space) is used to generate a set of candidate positions to map a read. These positions are scored using global chaining methods developed for whole-genome alignment, then refined by sparse dynamic programming before finally producing detailed alignments. BLASR is demonstrated to be sensitive to aligning reads across a wide range of insertion and deletion error. We demonstrate alignments of Human whole genome shotgun SMRT sequences.
Control-free prediction of copy number alterations in deep sequencing data using normalization on GC-content.
Valentina Boeva$^{1,2,3,4}$, Andrei Zinovyev$^{1,2,3}$, Kevin Bleakley$^{1,2,3}$, Jean-Philippe Vert Vert$^{1,2,3}$, Isabelle Janoueix-Lerosey$^{1,4}$, Olivier Delattre$^{1,4}$ and Emmanuel Barillot$^{1,2,3}$


Short summary:
Here we present a tool for control-free copy number alteration (CNA) detection using deep-sequencing data. It automatically computes, normalizes, segments copy number profiles (CNPs) and calls CNAs. To normalize raw CNPs, the user can provide a control dataset if available; otherwise GC-content is used. We demonstrate that GC-content normalization provides smooth profiles which can be further segmented and analyzed in order to predict copy number alterations.

Introduction:
In any study that applies deep sequencing to predict structural variants in cancer genomes, one has to calculate copy number profiles (CNPs) and predict regions of gain and loss. For various reasons, sequencing of an appropriate control sample is not always possible. There is therefore a need for a bioinformatics tool able to automatically detect copy number alterations (CNAs) without use of a control dataset. Several programs have been published that allow automatic calculation and analysis of CNPs using moving window approaches [1, 2]. However, both CNV-seq [1] and SegSeq [2] need datasets for the given tumor and its paired normal DNA. Moreover, CNV-seq predicts gains and losses without providing information about how many copies were lost or gained. An interesting approach for predicting copy number variants was suggested in [3], where GC-content is used to normalize data. However, to estimate the "normal" copy number, they rely on the assumption that there are similar percentages of amplified and deleted regions, which is not true in general.

Here we propose an algorithm to call CNAs with or without a control sample. The algorithm is implemented in the C++ program FREEC (control-FREE Copy number caller). FREEC uses a moving window approach to calculate read depth (RD) in non-overlapping windows (raw CNP). Then, if a control sample is available, the program normalizes raw CNP using the profile calculated for the control dataset. Otherwise, the program calculates GC-content in the same set of windows and performs normalization by GC-content. Since this removes a major source of variability in raw CNPs [2-4], the resulting normalized profile becomes sufficiently smooth to apply segmentation. This is followed by the analysis of predicted regions of gains and losses in order to assign copy numbers to these regions.

Methods:
The algorithm includes several steps. First, it calculates the raw CNP by counting reads in non-overlapping windows. If not provided by user, window size can be automatically selected using depth of coverage information to optimize accuracy of CNA prediction. The second step is profile normalization. If a control is not provided by the user, we compute the GC-content profile. To normalize the CNP we calculate the ratio between RD values and either fitted GC-content or the control profile if this is available (explained below). The third step is segmentation of the normalized CNP. To do this we implemented a LASSO-based algorithm suggested in [5]. The segmentation profiles provided by this algorithm are robust against outliers, which makes it suitable for segmentation of deep sequencing CNPs. The last step involves analysis of segmented profiles. This includes identification of regions of genomic gains and losses, prediction of copy number changes in these regions, and calculation of frequencies of gains and losses in the whole genome.
To normalize a raw CNP we fit the observed RD by the GC-content (or the control RD if it is available). We base our fitted model on several assumptions: (i) the sample ploidy $P$ is given, (ii) the observed RD in $P$-copy regions (i.e., regions with copy number equal to $P$) can be modeled as a quadratic polynomial of GC-content (or control RD), and (iii) the observed RD in a region with altered copy number is linearly proportional to the RD in $P$-copy regions. We make a reasonable guess to initialize the polynomial’s parameters, and then we optimize the polynomial’s parameters by iteratively selecting data points related to $P$-copy regions and making least-square fit on them only. The resulting polynomial is then used to normalize the CNP (Fig. 1).

**Results:**

We applied the method to predict CNAs in mate pair datasets for two neuroblastoma (NB) samples (near-diploid and near-tetraploid cell lines, unpublished data) and a COLO-829 cell line (near-triploid cell line) [6]. All three samples were sequenced using the Illumina Genome Analyzer platform. For the COLO-829 cell line we used only two lanes of a 2kb-insert size mate pair library. The sequences were aligned to the reference genome using Bowtie [7]. The number of uniquely mapped mate-pairs in each sample was around 20 million.

The polynomial fit by GC-content explained 84%, 93% and 90% of observed RD in $P$-copy regions of near-diploid NB ($P = 2$), near-tetraploid NB ($P = 4$) cell lines and COLO-829 ($P = 3$), respectively (Fig. 1A, B and C). We identified regions of gains and losses in the three samples (Fig. 1D, E and F). The total lengths of these regions are shown in Table 1. Comparison of CNA regions between the two NB samples provided information about deleted or duplicated genes in both cell lines. As expected (some specific chromosomal regions are recurrently gained or deleted in NB samples), the overlap was statistically significant (Table 2). Surprisingly, the overlap in deleted or duplicated genes was also significant between COLO-829 and either of the two NB samples, except for the question of common gains between COLO-829 and the near-diploid NB cell line.

**Discussion:**

We have developed an algorithm to predict copy number alterations in deep sequencing data without needing a control experiment. The main steps of the algorithm are (1) normalization of the copy number sequencing profile using GC-content (or control CNP if available), (2) segmentation of normalized profiles, (3) assignment of copy number changes to losses and gains, and (4) calculation of frequencies of gain and loss for each copy number.

Resolution of CNA identification provided by deep sequencing is much higher than that provided by current microarray technology, even if one has a relatively small amount of reads (about 20 million mate-pair reads in the tested samples). Thus, it is important to be able to efficiently analyze sequencing data to predict CNAs. Here we present a tool for automatic detection of CNAs and calculation of CNA frequencies. Our tool is especially useful when no control sample is available, since it can perform CNP normalization using GC-content. This can serve as a first step in the analysis of mate-pair/paired-end sequencing data, followed by detection of genomic structural variants using the orientation and order of abnormally mapped reads.

**Availability:**

The program is available at http://bioinfo-out.curie.fr/projects/freec/.
Since read depth (RD) is highly dependent on GC-content, it is possible to normalize copy number profile using only information about average GC-content in a window. (A,B,C) GC-content vs RD in 50kb windows for near-diploid NB, near-tetraploid NB cell lines and COLO-829, respectively. The result of the expectation maximization splitting scheme is shown in black. (D,E,F) GC-content normalized copy number profiles for chromosome 1 for near-diploid NB, near-tetraploid NB cell lines and COLO-829, respectively. The result of segmentation is shown is red. Automatically predicted copy numbers are shown in black.

Table 1. The total lengths of losses and gains in the near-diploid NB (NB2), near-tetraploid NB (NB4) cell lines and COLO-829 (excluding X and Y chromosomes).

<table>
<thead>
<tr>
<th></th>
<th>losses</th>
<th>gains</th>
<th>#gene in losses</th>
<th>#genes in gains</th>
<th>in</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB2</td>
<td>338.7Mb</td>
<td>254.75Mb</td>
<td>2731</td>
<td>2332</td>
<td></td>
</tr>
<tr>
<td>NB4</td>
<td>465Mb</td>
<td>446Mb</td>
<td>4464</td>
<td>3633</td>
<td></td>
</tr>
<tr>
<td>COLO-829</td>
<td>934.6Mb</td>
<td>880.5Mb</td>
<td>7289</td>
<td>6877</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Genes in regions of common losses and gains for near-diploid NB (NB2), near-tetraploid NB (NB4) cell lines and COLO-829 (excluding X and Y chromosomes) and the corresponding Fisher's exact test p-values.

<table>
<thead>
<tr>
<th></th>
<th>#genes in losses (p-value)</th>
<th>#genes in gains (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB2 &amp; NB4</td>
<td>1245 (&lt;10^{-16})</td>
<td>1113 (&lt;10^{-16})</td>
</tr>
<tr>
<td>NB2 &amp; COLO-829</td>
<td>2197 (&lt;10^{-16})</td>
<td>451 (1)</td>
</tr>
<tr>
<td>NB4 &amp; COLO-829</td>
<td>1879 (&lt;10^{-16})</td>
<td>2231 (&lt;10^{-16})</td>
</tr>
</tbody>
</table>

References:
Informing the inference of single nucleotide variants with copy number alteration annotations reveals novel somatic mutations in resequenced cancer genomes

Anamaria Crisan1, Rodrigo Goya1,2, Gavin Ha1, Leah Prentice1, Arusha Oloumi1, Janine Senz3,4, Thomas Zheng2, Kane Tse2, Allen Delaney2, Marco Marra2, David Huntsman3,4, Martin Hirst2, Sam Aparicio1, Sohrab Shah1,3,4

1Department of Molecular Oncology Breast Cancer Research Program, British Columbia Cancer Research Centre, 2 Genome Sciences Centre, British Columbia Cancer Agency, 3 Centre for Translational and Applied Genomics of British Columbia Cancer Agency, 4 Provincial Health Services Authority Laboratories

Background: Cancer arises from a mosaic of mutations that confer a selective advantage for growth and proliferation. Assembling a complete catalogue of mutations existing in a tumour is important for understanding aetiology, tumourigenesis and evolution. Next generation sequencing (NGS) technologies and accompanying bioinformatics algorithms have critical roles in the effort to identify the cancer mutational landscape. Several cancer genomes have been deeply sequenced with NGS and analyzed for both copy number alterations (CNAs) and single nucleotide variants (SNVs). These studies have revealed novel mutations in acute myeloid leukaemia (Mardis 2009); breast cancer (Shah 2009), ovarian cancer (Shah(b) 2009), lung cancer, lymphoma and melanoma (Pleasance 2010). Although studies by Pleasance et al. (2010), Chiang et al. (2009) and our own work (Shah, 2009) demonstrate that CNAs can be inferred from sequence data, none of these studies integrate CNAs in the inference of SNVs. We propose that allele-specific CNAs will alter the distribution of alleles in the NGS data such that some SNVs will be undetectable by conventional tools that assume diploidy. We assumed that modelling induced genotypes due to CNAs would result in more sensitive algorithms for SNV detection. In this study, we show with validated somatic mutations, that if the genome is to be fully characterized by mutational profiling with NGS, the consideration of copy number changes in the detection of SNVs is essential.

Methods: We developed a novel method called CoNAn-SNV (Copy Number Annotated –SNV) capable of adjusting to the structural alterations found in tumour genomes. We introduce $C_i \in \{\text{LOSS}, \text{NEUT}, \text{GAIN}, \text{AMP}, \text{HLAMP}\}$ as the copy number state at position $i$, where LOSS corresponds to a deletion, NEUT is copy number neutral, GAIN is a low level duplication, AMP is an amplification and HLAMP is a high-level amplification. We assume $C_i$ is known at run time and is given as input by the user. They key intuition in the CoNAn-SNV model is that $C_i = c$ informs the state space of possible genotypes $G_i^c = k$ at position $i$ as follows:

$$G_i^c = k, k \in \begin{cases} 
\{aa,ab,bb\} & \text{if } c \in \{\text{LOSS}, \text{NEUT}\} \\
\{aaa,aab,abb,bbb\} & \text{if } c = \text{GAIN} \\
\{aaaa,aabbb,aaabb,abb,bbb\} & \text{if } c = \text{AMP} \\
\{aaaaa,aaaaa,aaabb,aaabb,abb,bbb,bbb\} & \text{if } c = \text{HLAMP}
\end{cases}$$

Using a Binomial mixture model framework, we fit the model to data using expectation maximization and predict the most likely genotype given the model parameters, the distribution of alleles and copy number annotations.
Results: In a re-analysis of a previously published oestrogen-receptor positive lobular carcinoma we successfully validated an additional 24 non-synonymous somatic mutations that were undetectable by MAQ (Li 2008) and SNVMix1 (Goya 2010) and were consequently not reported in our original analysis of this tumour (which reported 32 somatic mutations). We show that CoNAn-SNV calls more variants in the whole genome shotgun sequencing and exon capture lobular carcinoma compared with existing tools without compromising accuracy and that the CoNAn-specific variants are significantly enriched in copy number amplified regions of the genome. Importantly, we show that without appropriate consideration of amplification status, these events may have been considered loss of heterozygosity events, when in fact the imbalance results from allele-specific amplifications. We also show how mutations exhibiting a low allelic frequency are more likely to be associated with allele-specific CNA rather than mutational heterogeneity or rare clonal populations.

Conclusion: Our results suggest that the consideration of CNAs in the interpretation of allelic distribution in genomically unstable cancers is critical for a more comprehensive identification of somatic mutations and for the interpretation of mutational heterogeneity and tumour evolution in resequenced cancer genomes.

References:
Structural variation analysis with strobe reads

Anna Ritz, Ali Bashir, and Benjamin J. Raphael.

Motivation: Structural variation including deletions, duplications and rearrangements of DNA sequence are an important contributor to genome variation in many organisms. In human, many structural variants are found in complex and highly repetitive regions of the genome making their identification difficult. A new sequencing technology called strobe sequencing generates strobe reads containing multiple subreads from a single contiguous fragment of DNA. Strobe reads thus generalize the concept of paired reads, or mate pairs, that have been routinely used for structural variant detection. Strobe sequencing holds promise for unraveling complex variants that have been difficult to characterize with current sequencing technologies.

Results: We introduce an algorithm for identification of structural variants using strobe sequencing data. We consider strobe reads from a test genome that have multiple possible alignments to a reference genome due to sequencing errors and/or repetitive sequences in the reference. We formulate the combinatorial optimization problem of finding the minimum number of structural variants in the test genome that are consistent with these alignments. We solve this problem using an integer linear program. Using simulated strobe sequencing data, we show that our algorithm has better sensitivity and specificity than paired read approaches for structural variation identification.
Using GPU programming for short read mapping

J. Blom1*, T. Jakobi1, D. Doppmeier2, S. Jaenicke1, J. Stoye3, A. Goesmann4

1Computational Genomics, Center for Biotechnology (CeBiTec), Bielefeld University, 33594 Bielefeld, Germany.
2Bioinformatics of Signaling Networks, CeBiTec, Bielefeld University, 33594 Bielefeld, Germany.
3Genome Informatics, Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany.
4Bioinformatics Resource Facility, CeBiTec, Bielefeld University, 33594 Bielefeld, Germany.
*Corresponding author, contact: jblom@cebitec.uni-bielefeld.de

Since the introduction of next generation sequencing technologies like Solexa, 454 and SOLiD the amount of generated data rises with each technology upgrade. As the application scenarios especially of the short read techniques include the re-sequencing of known genomes or sequencing of closely related strains, new software tools are needed for the fast mapping of sequencing reads against a reference.

Currently, there are several tools available, but most of them are limited either in speed or accuracy. Our goal was to develop an exact and complete mapping algorithm with an adequate runtime. The result is SARUMAN (Semiglobal Alignment of Short Reads using CUDA and Needleman-Wunsch). SARUMAN uses a q-gram based filter algorithm followed by a modified Needleman-Wunsch alignment. To speed up the normally time-consuming alignment step all alignments are processed on an NVIDIA graphics card to exploit the massively parallel architecture of new GPUs. As a result of this alignment strategy SARUMAN detects not only mismatches but also allows for insertions and deletions. The mapping algorithm is exact and complete, it identifies all possible matching positions for a given error threshold and always returns the optimal local alignment.

To facilitate quick exploration and evaluation of mapped reads we also developed VAMP - Visualization and Analyses of mapped sequences. VAMP offers different views on mapped data, allowing for an easy examination of position specific coverage information. VAMP is also able to detect all possible SNPs based on the minimum percentage of variation for a given position and an absolute threshold value, automatically.
RGASP Evaluation of RNA-Seq Read Alignment Algorithms

A Kahles1, J Behr, R. Bohnert, J. Harrow1, F. Kokocinski1, J Abril3, G Williams1, T Steijger2, M Salmon2, A Mortazavi7, G Raetsch4, M. Gerstein6, A. Reymond8, T Gingeras9, B Wold7, R. Guigó5, T. Hubbard1
1 Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. 2 EBI, Wellcome Trust Genome Campus Hinxton, Cambridge, UK. 3 Facultat de Biologia Universitat de Barcelona, Spain. 4 Friedrich Miescher Laboratory, Max Planck Society, Tübingen, Germany. 5 Centre for Genomic Regulation, Barcelona, Spain. 6 Department of Molecular Biophys. and Biochem., Yale University, New Haven, USA. 7 California Institute of Technology, Pasadena, USA. 8 Center for Integrative Genomics, University of Lausanne, Switzerland. 9 Cold Spring Harbour Laboratory, Cold Spring Harbour, NY, USA.

High throughput sequencing (HTS) data volumes are rapidly growing and the need for their fast and accurate analysis becomes increasingly important. To accommodate this demand, a wide range of read alignment algorithms has been developed over the past years. Inside this spectrum are algorithms capable of aligning reads from RNA-seq experiments, which form a particularly interesting subgroup of methods. These techniques are able to align reads over exon/intron boundaries and their results are very valuable for downstream transcriptome analyses. Although many algorithms are in use, most of the original publications were not accompanied by a comparison of alignment performance and result quality.

Following the successful format of the EGASP workshop in 2005 (Guigo et al., Genome Biology, 2006), the RNASeq Genome Annotation Assessment Project (RGASP) was launched to assess the current progress of automatic gene building using RNA-Seq as its primary dataset. The goals of this community effort are to assess the success of computational methods to correctly map RNA-Seq data onto the genome, assemble transcripts and quantify their abundance in particular datasets. The input data originated from different sequencing platforms (Illumina, SOLiD and Helicos) from Human, Drosophila and C. elegans. For these three organisms, which are also analyzed as part of the (mod)ENCODE project, high quality genome annotation is available which served as the references for the analysis.

As part of this effort, participants also submitted their alignments using a diversity of different methods, including blat, tophat, palmapper, and gsnap. First, we compared the submitted alignments among each other. Figure 1 shows a pair-wise comparison of sanitized alignments for reads aligned to human chromosome 1, with respect to the agreement of the complete read alignment and with respect to the agreement of intron predictions inferred from spliced alignments. Moreover, we compared the submitted alignments with respect to different descriptive statistical criteria as, for instance, sensitivity and precision of intron recognition, mismatch and indel distribution, and the overlap to annotated transcript and exon boundaries. In Figure 2 we show the sensitivity and specificity of intron predictions inferred from spliced alignments for human (chromosome 1) where we used the ENCODE genome annotation as ground truth. For a selection of algorithms we carried out further analyses, based on the re-alignment of artificially generated reads. This approach has two advantages: the availability of the theoretically optimal alignment and the controllability of the error distribution in the generated data.
Our comparisons showed a great diversity in the behavior of the different alignment strategies. In particular, there is surprisingly small agreement between a subset of methods. We can show that the largest differences are the result of different alignment filtering strategies, which can, for instance, drastically increase the precision of intron predictions. The evaluations of the transcript annotations derived from these alignments performed within RGASP additionally allow us to correlate the alignment accuracy with the preciseness of exon, transcript, and gene prediction. We will discuss specific features of the different alignment strategies that most influence the success of subsequent analysis steps.

The tools developed for this analysis have been incorporated into the publicly accessible Galaxy Webservice (http://galaxy.fml.mpg.de). Moreover, the evaluation code and details of the evaluation will be available from http://www.fml.mpg.de/raetsch/suppl/srm-eval.
Deep microRNA sequencing reveals differential expression in favorable versus unfavorable neuroblastoma tumors

Sven Rahmann (TU Dortmund), Tobias Marschall (TU Dortmund), Marcel Martin (TU Dortmund), Alexander Schramm (University Hospital Essen), Johannes Schulte (University Hospital Essen)

Small non-coding RNAs, in particular microRNAs, regulate fine-tuning of gene expression and can act as oncogenes or tumor suppressor genes. Differential microRNA expression has been reported to be of functional relevance for tumor biology. Using next-generation sequencing, the unbiased and absolute quantification of the small RNA transcriptome is now feasible. Neuroblastoma is an embryonal tumor with highly variable clinical course. We analyzed the small RNA transcriptomes of five favorable and five unfavorable neuroblastomas using SOLiD next-generation sequencing, generating a total of more than 188,000,000 reads. MicroRNA expression profiles obtained by deep sequencing correlated well with real-time PCR data. Cluster analysis differentiated between favorable and unfavorable neuroblastomas, and the microRNA transcriptomes of these two groups were significantly different. In-depth sequence analysis revealed extensive post-transcriptional miRNA editing. Of 13 identified novel miRNAs, three were further analyzed, and expression could be confirmed in a cohort of 70 neuroblastomas.
Analysis Methods for Directly Detecting DNA Methylation Through Single-Molecule Real-Time Sequencing

Dale Webster, Benjamin Flusberg, Jason Chin, Jon Sorenson, Eric Schadt, Stephen Turner
Pacific Biosciences, Menlo Park, CA 94025

The analysis of alterations to the standard 4-base DNA alphabet such as methylation, lesions, and alternative bases is challenging in 1st and 2nd generation sequencing technologies. Common approaches involve time-consuming and specialized sample preparation which converts the chemical entity of interest back to a base which can be readily read using conventional sequencing technologies. The SMRTTM sequencing technology allows real-time monitoring of DNA polymerase as it incorporates nucleotides matching a native DNA template. Polymerase kinetics are reproducibly affected by the presence of non-native bases, and we show here that these differences can be predictive of methylation state of the template DNA at per-base resolution. In conjunction with circular consensus sequencing, which allows the attainment of very high accuracies from a single molecule of DNA, this method can produce high-confidence determination of methylation status. We develop PCA and statistical classification methods for predicting methylation state and successfully apply these algorithms to the detection of methyl-cytosine in short templates and methyl-adenosine in bacterial samples. The method is expected to be generally applicable to other DNA modifications, and we show results suggesting that kinetic analysis of DNA synthesis will give rise to a new class of biomarkers for DNA-phenotype association studies.
Reducing the False Positive SNP Rate for Illumina Sequencing: probability of second best call and context dependent errors

Irina Abnizova*, Steven Leonard, Tom Skelly and Tony Cox.
Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Email: ia1@sanger.ac.uk

The new generation of short read sequencing technologies still requires both accuracy of data processing methods and reliable measures of that accuracy and data quality. Such measures are especially important for variant calling. However, in the particular case of SNP calling, a great number of False Positive SNPs (FP SNP) may be obtained. Reliable methods are required to distinguish putative SNPs from sequencing- or other errors.

We found that not only the probability of sequencing errors (i.e. the quality value) is important to distinguish a FP SNP, but also the conditional probability of ‘correcting’ this error (the ‘second best call’ probability, conditional on that of the first call). Surprisingly, around 70% of errors can be ‘corrected’ with this second call.

Another way to reduce the rate of FP SNPs is to retrieve DNA motifs which seem to be prone to sequencing errors. Attempts have been made to account for this by attaching a quality value to a called base conditional on the nucleotide motif preceding it [1,2]. However, just declaring a short motif to be error prone is not enough, and leads to over-training. We found a very strong negative association between the error motif context (related to read coverage of regions) and the occurrence of (single-stranded) systematic errors.

We incorporated the estimation of “second best call probability” and the error-proneness of motifs (based on the context of the region in which they occur) in a simple and reliable way to predict quality values for Illumina sequencing.

We illustrate our method with data from the phiX genome and human mitochondrial DNA. We implemented the new calibration algorithm in the Sanger Institute Production line. Our method produces reliable quality values with a predictive power of at least 80% and is freely available from the Sanger Institute repository.

The effect on sequencing accuracy of DNA polymerase kinetics in real-time, single-molecule, sequence-by-synthesis methods (3rd generation)

Chuck Tweedy, Chunhong Zhou, Michael Previte, Guobin Luo,

Life Technologies

Third generation sequencing technologies present a new generation of unique data characteristics resulting from the realities of single-molecule reactions. The stochastic nature of these reactions give rise to primary base-calls with quality variations that must be accounted for in downstream analysis. For this discussion, we look specifically at the challenges that are presented due to the kinetics of the DNA polymerase. Variations in nucleotide incorporation time due to polymerase catalytic kinetics result in sequencing deletions when incorporation occurs too quickly for the system to detect. Opposing these deletions are sequencing insertions caused by polymerase-nucleotide binding kinetics due to so-called non-productive binding events. An analytical model of these competing processes is presented, and its application to real data is explored. This subject will be of increasing interest to bioinformatics scientists in the near future as data from 3rd Gen machines starts to become available.
Correction of sequencing errors in a mixed set of reads

Leena Salmela

Department of Computer Science, PO Box 68 (Gustaf Hällströmin katu 2b), FI-00014 University of Helsinki, Finland

Motivation: High-throughput sequencing technologies produce large sets of short reads that may contain errors. These sequencing errors make de novo assembly challenging. Error correction aims to reduce the error rate prior assembly. Many de novo sequencing projects use reads from several sequencing technologies to get the benefits of all used technologies and to alleviate their shortcomings. However, combining such a mixed set of reads is problematic as many tools are specific to one sequencing platform. The SOLiD sequencing platform is especially problematic in this regard because of the two base color coding of the reads. Therefore, new tools for working with mixed read sets are needed.

Results: We present an error correction tool for correcting substitutions, insertions and deletions in a mixed set of reads produced by various sequencing platforms. We first develop a method for correcting reads from any sequencing technology producing base space reads such as the SOLEXA/Illumina and Roche/454 Life Sciences sequencing platforms. We then further refine the algorithm to correct the color space reads from the Applied Biosystems SOLiD sequencing platform together with normal base space reads. Our new tool is based on the SHREC program that is aimed at correcting SOLEXA/Illumina reads. Our experiments show that we can detect errors with 99% sensitivity and >98% specificity if the combined sequencing coverage of the sets is at least 12. We also show that the error rate of the reads is greatly reduced.

Availability: The JAVA source code is freely available at http://www.cs.helsinki.fi/u/lmsalmel/hybrid-shrec/

Contact: leena.salmela@cs.helsinki.fi
Savant: Genome Browser for High Throughput Sequencing Data

Marc Fiume1, Vanessa Williams1, Michael Brudno1,2

1Department of Computer Science, University of Toronto, Toronto, Ontario, Canada; 2Banting and Best Department of Medical Research, Toronto, Ontario, Canada

HTS machines (e.g. Solexa, SOLiD, 454) are capable of sequencing billions of nucleotide bases per week, requiring many gigabytes of data to be stored on locally accessible hard disks. The computational analysis of the large volumes of data generated by the new sequencing machines remains a challenge. Servers-based browsers, such as the UCSC and ENSEMBL browsers, have been developed to display a variety of genomic data, however, their use for visualization of HTS data is undesirable because a huge amount of data must be uploaded over the internet, they are typically slow and not interactive, and they can neither be customized nor extended. We introduce Savant, the Sequence Annotation Visualization and ANalysis Tool, a powerful desktop platform for visualization and analysis of genomic data. Savant enables fast visualization of huge datasets, has a simple and intuitive user interface, and is extensible via a plugin framework. Savant is the first, to our knowledge, extensible genome browser specifically designed for HTS data.

In addition to displaying HTS mappings, Savant can be used to visualize any genome-based character, point, interval, or continuous data, corresponding for example to sequence, SNP, gene, and per-base coverage or conservation data, respectively. The Savant UI allows users to easily select, zoom, and pan regions of the genome. It optimizes both nucleotide- and genome-scale visualization, via dynamic resolution adjustments in response to changes in the size of displayed regions. This feature improves the clarity of visualization, in addition to reducing Savant’s memory footprint while improving overall speed. Using Savant, users are also able to view and export visualized data in a tabular format, perform real-time annotations, export images for publications, and more. Finally, developers can use the plugin framework to programmatically access tracks, their data, and aspects of the UI in order to customize Savant and to incorporate analyses such as real-time genomic variation detection.

Savant is freely available at http://compbio.cs.toronto.edu/savant/.
Online Quantitative Transcriptome Analysis

Jonas Behr1, Regina Bohnert1, Philipp Drewe1, Géraldine Jean1, André Kahles1, Georg Zeller1, Oliver Stegle2,3, Gunnar Rätsch1*
1 Friedrich Miescher Laboratory of the Max Planck Society, Spemannstraße 39, 72076 Tübingen, Germany
2 Max Planck Institute for Developmental Biology, Spemannstraße 35, 72076 Tübingen, Germany
3 Max Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany
*To whom correspondence should be addressed: Gunnar.Raetsch@tuebingen.mpg.de

Introduction
The current revolution in sequencing technologies allows us to obtain a much more detailed picture of transcriptomes. Studying them under different conditions or in mutants will lead to a considerably improved understanding of the underlying mechanisms of gene expression and processing. An important prerequisite is to be able to accurately determine the full complement of RNA transcripts and to infer their abundance in a sample. However, the analysis is made considerably more difficult by various limitations and biases in next-generation sequencing (NGS) technologies. We present the first integrative platform for quantitatively analyzing RNA-Seq experiments. It is based on the interactive Galaxy-framework [1] and builds on recently developed methods for NGS sequence analysis: read mapping, transcript reconstruction, quantitation, and tests for differential expression (see Figure 1).

Read mapping
We provide PALMapper [8], a new RNA-seq read mapper that combines an extended version of the alignment method QPALMA [2] with the GenomeMapper short read aligner [3] to efficiently align both spliced and unspliced reads with high accuracy. While the QPALMA training algorithm takes advantage of each read quality information and computational splice site predictions to compute an extended alignment scoring model, GenomeMapper quickly carries out an initial read mapping to the reference genome which will then guide a banded Smith-Waterman-like algorithm that allows for long gaps that correspond to introns.

Transcript reconstruction
We advanced two independent methods to de novo predict transcripts de novo based on NGS data. Firstly, mTiM predicts transcripts based on the RNA-seq read alignments and computational splice site predictions. It relies on only a few assumptions and is therefore well suited for annotating coding as well as non-coding genes. Secondly, we extended the gene finding system mGene [4] to additionally model RNA-Seq read alignments. It also takes advantage of additional genomic sequence features and can therefore better bridge missing read information for lowly expressed coding genes. We developed a strategy that enables us to annotate a genome without prior annotation information, i.e., purely based on RNA-seq reads and the genome sequence.

Quantitation
For accurate quantitation of RNA transcripts it appears essential to account for sequencing of biases and other technical limitations, inherent in library preparation, sequencing, and read mapping, into account. For this purpose, we developed a method, called rQuant that implements this idea by simultaneously estimating the effect of biases as well as the abundance of RNA transcripts [5]. It
illustrated that the non-uniform read-coverage is well predictable. rQuant can tackle multiple transcripts per gene locus and is therefore particularly well-suited to quantify alternative isoforms.

**Testing differential expression**

Finally, we developed statistical tests that identify significant differences between two RNA-Seq experiments [9] to find differentially expressed regions or transcripts (with or without knowledge of the transcript annotation).

![Galaxy pipeline for quantitative transcriptome analysis.](image)

The platform can be used for many purposes, including 1) to (re-)annotate genomes while profiting from NGS transcriptome data; 2) to identify novel transcripts that are only expressed under certain conditions; and 3) to identify regions or transcripts that are the target of regulated RNA processing. We have tested the system with data from three different model organisms (*C. elegans*, *D. melanogaster*, and *A. thaliana*) and also for data from human (alignments were done offline). Moreover, we have participated in the RGASP competition [6] for an external evaluation.

The Galaxy instance also offers tools like the SAM [7], other alignment tools as well as sequence analysis tools. It is publicly available at [http://galaxy.fml.mpg.de](http://galaxy.fml.mpg.de). Moreover, we are currently preparing a virtual machine of the Galaxy installation that can be downloaded and run locally to perform the analyses.

**References**


Abstracts for the Poster Session
The analysis of temporal viral samples using ultra-deep pyrosequencing data

John Archer\(^1\), Marilyn Lewis\(^2\), Andrew Rambaut\(^3\), and David L. Robertson\(^1\)

\(^1\)Faculty of Life Sciences, University of Manchester, Manchester, UK; \(^2\)Pfizer Global Research and Development, Sandwich, UK; \(^3\)Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.

Parallel sequencing technology based on pyrosequencing produces unprecedented quantities of sequence data. Current mapping software is inadequate for dealing with the levels of variation present within such data when generated from viral populations. Thus, studies using such data depend on the development of bespoke mapping pipelines. As a consequence of platform and viral dependent intricacies mapping is rarely adequately optimized; resulting in data loss. This loss often occurs non-randomly and has potential to introduce bias to the population structure. This is especially relevant in drug resistance studies where ultra-deep sequencing is used and non-biased mapping required. In order to apply 454 Life Sciences’ pyrosequencing system to the study of viral data, we have developed software to overcome this mapping problem. We apply the software to HIV-1 intra-patient data, detecting minority CXCR4-using variants that may limit the response to a maraviroc-containing treatment regimen. Five time points were available from two patients from a clinical study of maraviroc. Genotypic algorithms that detect the presence of CXCR4-using virus were used to compare reads obtained from pre- and post-treatment samples. In both patients distinct clusters of low frequency CXCR4-using variants were detected prior to treatment. Reads were then represented phylogenetically in order to characterize emergence of probable maraviroc-insensitive forms. The software, which will also have utility for studying the within-host diversity of other fast evolving viruses, is available from http://www.bioinf.manchester.ac.uk/segminator/.
A pipeline for fast gapped read mapping of Illumina reads

Markus J. Bauer\textsuperscript{1}, Anthony J. Cox\textsuperscript{1}, Dirk J. Evers\textsuperscript{1}

The current chemistry of Illumina’s sequencing by synthesis \cite{1} approach allows read lengths of up to 150bp and the latest generation of Illumina sequencers produces over one billion reads per sequencing run. Mapping reads against a reference, i.e. determining the position from which the read originated, is a critical step in almost any subsequent analysis. Longer reads and paired reads, however, lead to changes in how the problem is tackled algorithmically.

We present the latest improvements to GERALD, the mapping module of the CASAVA package (Consensus Assessment of Sequence And VAriation) which performs fast gapped read mapping. Typical scenarios comprise the mapping of reads against a well-defined reference for resequencing projects to call variants (small insertions and deletions or SNPs), determining expression levels in RNA-Seq experiments, or mapping the reads against contigs or scaffolds as a quality check for de novo assembly. Traditionally, read mappers build an index on either the reads or the reference. Well established programs that build the index on the reads include MAQ \cite{2} and RazerS \cite{6}. Recently, alignment programs like BWA \cite{3} or Bowtie \cite{4}, which are based on the Burrows-Wheeler transform (BWT) and the FM-index \cite{5}, have also been developed. In the following, we will present GERALD, the mapping module within CASAVA, which allows fast gapped mapping of longer reads.

The GERALD module performs the read mapping in three stages: (1) given a read pair, ELANDv2 determines a set of potential mapping positions for both the left and the right read, (2) we align read pairs where one of the reads acts as an anchor, and finally (3) the mappings are scored by taking mismatches and the insert size of the mapped positions into account.

In stage (1), ELANDv2 splits up the reads into substrings of length 32, called seeds, and builds a lookup table of those seeds. Given that lookup table, we scan the reference genome and identify positions where any of the seeds match, with at most two mismatches. For all positions found, we perform a banded global-local alignment between the genomic region and the read to account for small insertions or deletions. In stage (2), localized alignments are performed to resolve read pairs having a unique alignment of one read but an ambiguous alignment of the second read. Finally, (3) estimates the insert size distribution using uniquely aligned read pairs and scores read pairs according to their insert size and the number of mismatches to the reference.

To assess the sensitivity and specificity of our approach, we used simulated paired-end reads from the 1000G aligner comparison \cite{7}. We calculated sensitivity and specificity of the mappings, and the comparison to other published methods shows that our approach performs favourably. Assessing the results for our pipeline and Bowtie, for instance, shows that we achieve higher sensitivity and specificity at a much lower memory footprint and a highly competitive running time. Finally, since we map each read independently, we can use that mapping information to detect insertions, deletions, and larger structural variants.
References

¹Illumina Cambridge, Chesterford Research Park, Cambridge CB10 1XL, United Kingdom.
Comparison of SNP calling methods in 454 Sequencing data

Wayne Clarke

Next-generation sequencing technologies, such as those from Roche and Illumina, are being utilized for high-throughput Single Nucleotide Polymorphism (SNP) discovery with increasing frequency. Consequently, software to analyze such data and call high quality SNPs is being sought by researchers. This poster details the evaluation of a variety of methods to identify reliable SNPs within 454 sequence data from two diverse plant species. The data was analyzed using four SNP calling methods and the efficiency of each was compared based on several criteria.

The methods evaluated included that described by Barbazuk et al., which uses BLAST and cross_match for sequence alignment and the Polybayes software for SNP calling. The second method, a variation of the first, utilizes Polybayes to analyze sequence assemblies generated in the ACE file format. The third method uses a combination of the Mosiak sequence aligner and Gigabayes for SNP calling. Gigabayes is a newer version of Polybayes tuned specifically for short read sequences. Finally, a method utilizing commercially available software from DNASTAR, the NGen assembler and the SeqMan Pro analysis suite, was evaluated. Evaluation of each of the above methods suggests that for high-throughput SNP identification, an alternative approach is necessary, especially for complex genomes.
Probabilistic SNP mapping using GNUMAP-SNP

Nathan L. Clement, Quinn Snell, Mark J. Clement, and W. Evan Johnson
Brigham Young University, Provo, UT, 84602. May 10, 2010

The massive amounts of data generated from each next generation sequencing instrument run have presented researchers with unique analytical challenges which require innovative, computationally efficient statistical solutions. There are several existing approaches for calling bases and identifying SNPs in next generation sequencing data. MAQ (Li et al. 2008) and SOAP-SNP (Li et al. 2009) rely on ad hoc cut-offs and lack comparisons with background noise, leading to many false-positive results. These methods also remove reads with low-quality bases and either remove or randomly assigns reads that map to multiple locations which may lead to biased results. Finally, these methods rely on the single, although most plausible, mapping of the read to the genome and therefore do not use all the available information in the data. These other highly plausible local alignments can provide researchers with probabilistic confidence in SNP calls.

GNUMAP-SNP uses a probabilistic Pair-Hidden Markov Model for base calling and SNP detection in next generation sequencing data. Our results show that GNUMAP-SNP has both high sensitivity and high specificity throughout the genome, which is especially true in repeat regions or in areas with low read coverage. In addition, our approach applies a likelihood ratio test that provides researchers with straight-forward SNP calling cut-offs based on a p-value cut-off or a false discovery control. This framework makes GNUMAP-SNP highly sensitive and specific and more appropriate for the noisy data generated by current next generation sequencing platforms. GNUMAP-SNP is available as a module in the GNUMAP probabilistic read mapping software. GNUMAP is freely available for download at: http://dna.cs.byu.edu/gnumap/. An initial test was performed using data generated by the metagenomic simulator Metasim (Richter et al., 2008). The C. elegans chromosome I reference sequence was mutated every 1000 bases, from which Metasim generated a simulated 15x coverage of Illumina reads. The SNP callers for SOAP, MAQ and GNUMAP-SNP were evaluated based on this known set of SNPs.

Table 1: Experimental results for simulated data (15k actual SNPs) Program Time True Positives False Positives False Negatives
SOAP-SNP 41m 14309 18208 764
MAQ 13m 12193 1173 2876
GNUMAP-SNP 23m 14047 57 1026

GNUMAP was able to achieve similar false negative rates while nearly eliminating false positives. Future work will focus on cutting down the execution time through parallelization and optimized code paths. The low false positive rate that GNUMAP-SNP produces and the p-value metric to determine which SNPs are most significant will allow researchers to focus their expensive wet lab research on sites that are most promising.
VARiD: A variation detection framework for color-space and letter-space platforms

Adrian V. Dalca, Stephen M. Rumble, Misko Dzamba, Samuel Levy and Michael Brudno

Recent development in High-Throughput Sequencing (HTS) technologies are enabling much cheaper and faster study of genomic variation. However, the different HTS platforms available can suffer from different sequencing biases, error types and error rates. The sequencing reads themselves also differ among the platforms: while most of these technologies sequence the nucleotides of the genome directly, generating base calls for each position (letter-space), the Applied Biosystem’s SOLiD platform produces di-base sequences (color space). Several algorithms have been developed for both read types to map and align the reads to a draft genome, and many toolsets further facilitate variation studies for letter-space sequences. However, to date there are only a few tools that can identify variants from color space data, and while combining data from the various platforms should increase the accuracy of variation detection, there are no tools that can analyze color space and regular (letter space) data together.

In this poster we describe VARiD, a Hidden Markov Model (HMM) that uses the forward backward algorithm to accurately detect heterozygous, homozygous and tri-allelic SNPs, as well as micro-indels. This new probabilistic method for variation detection can work with both letter-space and color-space reads simultaneously, and by combining these different data sources, VARiD exploits their strengths to improve on the accuracy of the variation calls. Our analysis shows that on color-space only data VARiD yields more accurate variation detection than AB’s Corona Lite pipeline, and on letter-space only data can perform on par with the GigaBayes software. More importantly, we also show that VARiD’s ability to combine letter-space and color-space data dramatically improves the calls when using a mixed dataset.
Sensitive Mapping of Short Reads with SHRiMP2

Matei David¹, Misko Dzamba², Daniel Lister², Stephen Rumble³, Michael Brudno²

¹Department of Computer Science, Princeton University
²Department of Computer Science, University of Toronto
³Department of Computer Science, Stanford University

SHRiMP2 is a software package for mapping Illumina and AB SOLiD reads from a donor genome against a target (reference) genome. The main features of SHRiMP2 include fasta and fastq input formats; SAM output format; Full support (including indels) for both letter space (Illumina/Solexa) and colour space (AB SOLiD) reads; Paired mapping mode that fully utilizes mate-pair/pair-end information, and parallel computation, fully utilizing the power of modern shared-memory multi-core architectures.

SHRiMP2 uses two techniques to match reads to a genome: we first use the q-gram filtering technique, utilizing multiple spaced seeds, to rapidly identify candidate mapping locations for each read. Subsequently, these locations are thoroughly investigated by a vectored implementation of the standard dynamic programming (Smith-Waterman) alignment algorithm, allowing for the accurate identification of mismatches (SNPs), micro-indels, and crossovers (colour space sequencing errors.) SHRiMP2 primarily targets sensitivity, but it has come a long way to achieve it at a reasonable speed. SHRiMP2 achieves 95% precision and 79% recall when mapping simulated colour space reads of 50 base pairs (bp) each containing 1 SNP, 1 indel of size up to 5bp, and 4% per-color error rate to the Ciona savignyi genome (180Mb). SHRiMP2 on a single 3.0GHz core with 16GB of RAM can map 36bp colour-space reads against the reference human genome (hg18) at the rate of 320,000 reads per hour in mate-pair/pair-end mode. SHRiMP2 is fully core-parallelizable: a cluster of 4 machines, each with 2 quad-core CPUs and 16GB of RAM, will map 4x2x4=32 times faster, or approximately 10 million paired reads per hour.
Assembly of highly polymorphic organisms with hybrid sequencing

Nilgun Donmez1 and Michael Brudno1, 2

1Department of Computer Science and 2Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto

In the last decade the sequencing and assembly of genomes have become routine biological experiments. Currently available assembly tools are optimized for the assembly of large mammalian genomes using Sanger style reads, or smaller bacterial genomes from High Throughput Sequencing (HTS) data. While many of these assembly tools also account for the possibility of polymorphisms in the sequenced individuals, the low frequency of SNPs and other genetic variants in these genomes makes addressing these polymorphisms a relatively tractable problem. Several known organisms, such as sea squirt and sea urchin, show extremely high polymorphism rates. Highly polymorphic genomes complicate the assembly process since allelic differences are often hard to distinguish from imperfect repeats, and may also be mistaken for sequencing errors. Moreover, extreme heterozygosity means that sequence coverage will be lower than would otherwise be expected for most regions. These challenges call for higher coverage than typically available in Sanger sequencing projects. Simultaneously, short reads from HTS are quite limited in their ability to resolve repeats and polymorphic sequences.

By mixing a small number of long reads with a moderate coverage of short reads, the challenges mentioned above may be mitigated. Furthermore, Sanger style sequencing enables the use of mate pair sequencing with very long insert libraries (~10-40kb). The use of such libraries makes it possible to resolve long repeats and link contigs that may be tens of thousands of base pairs apart.

Our assembly pipeline includes a probabilistic error correction procedure that accounts for polymorphism and uses mate pairs to resolve repeats and heterozygous regions. We evaluate the performance of our method with a mix of simulated Sanger and Illumina style reads sampled from a 16mb region of the C. savignyi (sea squirt) genome and experiment with various coverage levels and ratios of Sanger to Illumina reads. The results show that the contiguity of the assembly can be greatly improved with this hybrid approach, while keeping the cost of sequencing relatively low.
Quantification of systematic bias in ChIP-seq signal profiles

Samuli Eldfors, Petri Törönen, Paul Bertone, Liisa Holm

ChIP-seq signal profiles have been shown to contain systematic bias where high signal levels are observed irrespective of the experimental conditions. This type of signal bias can confound the inference of protein-DNA binding sites if not appropriately accounted for. Firstly, we present a novel method for identifying genomic regions exhibiting such systematic bias. Secondly, we present the results from applying this method to the analysis of a set of experimentally independent human ChIP-seq data sets released by the ENCODE consortium. Our method is based on the analysis of a set of experimentally independent signal profiles. In our method ChIP-seq reads are first aligned to the reference genome and from this read mapping a sequence read coverage signal profile is computed for each data set. The resulting signal profiles are combined into a matrix where rows represent signal profiles and columns correspond to genomic positions. A summary statistic indicative of systematic bias is then computed from this matrix for each genomic position, allowing for the quantification of the level of bias along the genome and identification of positions exhibiting consistently high signal levels in all data sets regardless of experimental conditions. Our analysis represents a novel approach for identifying genomic regions likely to yield false positive peak calls and our results highlight the need for careful background modeling in the analysis of ChIP-seq data. Interestingly, our results indicate that the distribution and severity of systematic bias appears depend to large extent on the read mapping strategy employed. Lastly, we expect this approach to be generalizable for the quantification of systematic bias in sequence read coverage from other methods based on selective enrichment of DNA fragments, such as in MeDIP-seq and Exome sequencing.
We present **ArrayExpressHTS**, an R/Bioconductor-based pipeline for pre-processing and data quality assessment of high throughput sequencing (HTS) transcriptional profiling datasets (so called RNA-seq) submitted to the ArrayExpress Archive of Functional Genomics Data at the EBI. Starting from an ArrayExpress accession number of an RNA-seq experiment, all the heavy computation takes place on the EBI infrastructure, using the **EBI R CLOUD**: first, raw short reads are retrieved from the European Nucleotide Archive, then alignment and assembly steps for each lane are performed in parallel, producing raw and processed data quality reports.

The pipeline produces Bioconductor **ExpressionSet** objects containing transcript-level counts, grouping samples by factors specified in ArrayExpress, ready for downstream gene expression data analysis, e.g., using **edgeR** or **DEseq**. **ArrayExpressHTS** provides a variety of configuration options, allowing the user to select from **BWA** and **TopHat** as aligners, with reference genomes/transcriptomes available readily for all major organisms, including human, mouse and rat. Mappings can be made to the transcriptome or the genome, or both sequentially. The three approaches are compared over datasets in ArrayExpress. **Cufflinks** is used for transcript assembly, mapping the reads to transcript isoforms as annotated by the latest Ensembl genome build. The user can specify a number of filtering options, including cut-offs based on average base quality, polyX head/tail runs, low complexity read score, as well as reads mapping to specified chromosomes or genomic regions and verify the effect of the filter at each step.

For each **ArrayExpressHTS** run, a data quality assessment report is produced, consisting of diagnostic plots on raw and processed reads (per lane), as well as several between-lane comparison graphics. For raw reads, the plots improve upon those in the **ShortRead** package, namely base call and average base quality per cycle, and introduce several additional diagnostic graphics, such as the "dusty" score and the cumulative density function (CDF) plot of uncalled bases per read, helping to identify poor quality datasets. For processed reads, a bar chart is provided of the proportions of aligned reads, a mapping quality distribution plot and a CDF plot of polyX run quantities.

We tested the **ArrayExpressHTS** pipeline on several HTS datasets in ArrayExpress. Initial results of the quality assessment as well as all processed transcript counts are available for download through the ArrayExpress website. **ArrayExpressHTS** will be offered as a service, accessible via the EBI R CLOUD Workbench as well as via standard R.

In preparation for inclusion in the pipeline is support for paired-end reads, differential treatment of stranded/unstranded libraries, automatic scoring of the dataset quality, quantification of features outside annotated transcripts and integration of data from SNP calling (using Samtools) to give allele-specific estimates.

**Availability:**
ArrayExpressHTS will be available via the EBI R CLOUD Workbench at http://www.ebi.ac.uk/tools/rcloud as well as a standalone R package in Bioconductor.
**ONCOSEQ: A Cancer Next Generation Sequence Data-Mining Platform**

*Catherine Grasso, Michael Quist, Robert Lonigro, Shanker Kalyana-Sundaram, Xuhong Cao, Terrence Barrette, & Arul Chinnaiyan*

*Michigan Center for Translational Pathology, University of Michigan Medical School, Ann Arbor, Michigan*

Next generation sequencing technology has led to a marked increase in tumor-derived transcriptome sequencing, generating a wealth of data with the potential to uncover complex relationships between somatic mutations, gene expression, and transcript editing. The large variety of analyses made possible by transcriptome sequencing presents great opportunities, as well as great challenges, for researchers and the results of transcriptome studies are often underutilized due in part to the lack of effective tools for mining RNA-seq data. Here, we present ONCOSEQ, a web-based data-mining platform that is conceptually based on the ONCOMINE microarray data-mining platform and is aimed at facilitating discoveries in cancer transcriptome data. Like ONCOMINE, ONCOSEQ focuses the data mining process on isoforms or mutations that appear to be highly correlated with cancer progression. Once features of interest have been identified by the user, ONCOSEQ can be used to view relevant supporting data: position-specific coverage maps showing read-level details have the potential to reveal single nucleotide variants (SNVs), while whole gene coverage maps have the potential to reveal novel isoforms. ONCOSEQ can also be used to relate transcript expression data to data from other sequencing based experiments, such as ChIP-seq, allowing the user to explore the relationship between transcript expression patterns and regulation. As in ONCOMINE, the core functionality of ONCOSEQ is tethered to box plots and bar plots that summarize differences associated with cancer progression, making it an effective tool to aid in the identification of somatic mutations and transcript biomarkers that can be used as diagnostic and therapeutic targets for cancer treatment.
Targeted sequencing as a cost-effective technique for genotyping custom SNP panels: application on studying response to chemotherapy in childhood acute lymphoblastic leukaemia.

Agata Wesolowska, Marlene D. Dalgaard, Louise Borst, Laurent Gautier, Mads Bak, Nils Weinhold, Betina F. Nielsen, Jacob Nersting, Niels Tommerup, Soren Brunak, Thomas Sicheritz Ponten, Henrik Leffers, Kjeld Schmiegelow, Ramneek Gupta*

Center for Biological Sequence Analysis, Technical University of Denmark
Pediatric Oncology labs and Department of Growth and Reproduction, Rigshospitalet

Several SNPs are known to be key determinators for interindividual differences in treatment resistance and toxic side effects. As most childhood acute lymphoblastic leukaemia treatment protocols include up to 13 different chemotherapeutic agents, the impact of individual SNPs has been difficult to evaluate. We are designing a high-throughput assay for ca. 20,000 targeted SNPs to explore combined gene-dosage effects. Such a multiple-SNP assay would allow the investigation of effects mediated via pathways where one of several SNPs (in the same or different genes) may lead to very similar clinical phenotypes due to related molecular mechanisms. In selection of genes, we have included those involved in pharmacology, immunology, DNA repair mechanisms, mitosis activity genes, and genes that affect apoptosis, neurotoxicity, and thrombosis.

To assay a custom SNP panel, we demonstrate multiplexed targeted sequencing as a cost-effective technique for genotyping. This poster will present some pilot data.
Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen, Steven E. Brenner and Sandrine Dudoit

Generation of cDNA using random hexamer priming induces biases in the nucleotide composition at the beginning of transcriptome sequencing reads from the Illumina Genome Analyzer. The bias is independent of organism and laboratory and impacts the uniformity of the reads along the transcriptome. We provide a read count re-weighting scheme, based on the nucleotide frequencies of the reads, that mitigates the impact of the bias.
Remarkable technological progress in recent years has enabled rapid survey of common variation and copy number polymorphisms at a genome-wide level. What remains unclear is the contribution of balanced genomic rearrangements to the overall genetic variance in complex disorders. Translocations and inversions that do not involve a net gain or loss in genetic material are not detectable by genome-wide array based methods and traditional molecular approaches are inadequate to rapidly identify breakpoints to base pair resolution. Next-generation sequencing offers the unique potential to rapidly and precisely characterize all structural variants, including balanced rearrangements. Numerous tools exist for initially mapping short reads to the genome, yet downstream analyses to identify structural variants remains non-standardized and tedious. As sequencing technology, protocols, and alignment software are all rapidly evolving, the ability to monitor instrument and algorithm performance over time is critical. We are currently developing GATLING (Genome Analysis Tools with Lightweight INtegrated Graphics), a suite of python and C++ programs to facilitate discovery and analysis of structural variation from sequencing data. GATLING tracks samples and analysis runs, calculates and stores QC and alignment statistics, and makes all information available through a web interface. To identify probable translocation events, GATLING efficiently clusters and scores anomalous read pairs. Once putative breakpoints have been discovered, an iterative analysis can be employed to search for additional evidence using slower but more rigorous alignment algorithms. To aid in interpretation of short read results, GATLING also draws diagrams showing alignments in the context of the reference sequence and putative variants. We are applying these tools to characterize structural variants in patients with autism spectrum disorders and other neurodevelopmental phenotypes.
Using GPU programming for short read mapping

Tobias Jakobi¹, Jochen Blom¹, Daniel Doppmeier¹, Sebastian Jaenicke¹, Alexander Goesmann²

¹Computational Genomics, Bielefeld University, Bielefeld, Germany
²Bioinformatics Resource Facility (BRF), Bielefeld University, Bielefeld, Germany

Since the introduction of next generation sequencing technologies like Solexa, 454, and SOLiD the amount of generated data rises with each new technology upgrade. As the application scenarios especially of the short read techniques include the re-sequencing of known genomes or sequencing of closely related strains, new software tools are needed for the fast mapping of sequencing reads against a reference genome.

Currently, there are several tools available, but most of them are limited either in speed or accuracy. Limitations in accuracy lead to non detected mappings, which could become important in post processing steps like SNP calling. Because of those limitations our goal was to develop an exact and complete mapping algorithm with equivalent running time compared to available heuristic implementations.

The result is SARUMAN (Semiglobal Alignment of Short Reads using CUDA and Needleman-Wunsch). SARUMAN uses a qgram index based filter algorithm followed by a modified Needleman-Wunsch alignment. To speed up the normally time-consuming alignment step all alignments are processed on a NVIDIA graphics card to exploit the massive parallel architecture of new graphics processing units (GPUs). Based on this technique, depending on the input read length, SARUMAN is able to process hundreds of thousands of alignments just a few seconds.

As a result of this alignment strategy SARUMAN not only detects mismatches, but also allows for insertions and deletions. The mapping algorithm is exact and complete, it identifies all possible matching positions for a given error threshold and always returns the optimal local alignment.

To facilitate quick exploration and evaluation of mapped reads we also developed the VAMP software (Visualization and Analyses of mapped sequences). VAMP is a platform independent tool implemented in Java and offers different views on mapped data, allowing for an easy examination of position specific coverage information. The software supports the import annotated reference genomes in GenBank format and thus gives an overview of all known features in given sequence space. Together with mapping information from the SARUMAN tool, a detailed analysis of sequencing runs is possible. VAMP is also able to detect all possible SNPs based on the minimum percentage of variation for a given position and an absolute threshold value automatically.
Quality guided correction and filtration of errors in short reads

David R. Kelley, Michael C. Schatz, and Steven L. Salzberg
Center for Bioinformatics and Computational Biology, Department of Computer Science, Institute for Advanced Computer Studies
University of Maryland, College Park

High-throughput sequencing technologies such as that offered by Illumina have permeated nearly all areas of biological research. Illumina’s technology produces sequencing reads of 35-125 bp that may have base calling errors at rates as high as 1-2%. These errors create difficulties for downstream sequence analysis tasks, such as detecting overlaps for genome assembly or aligning reads to a reference genome for SNP detection.

Due to its lower cost, deep coverage of the genome is generally possible using Illumina sequencing. Past work has shown that errors can be identified in a set of reads with deep coverage by first counting k-mers, and then considering k-mers with coverage less than some threshold to be artifacts of sequencing errors. Previous methods to correct reads with errors have searched for a minimal set of edits to the read that ensure all k-mers have sufficient coverage. We demonstrate that due to biases with respect to where errors occur in the read and the likelihood of specific nucleotide to nucleotide errors, such approaches are prone to mistakes that introduce “corrected” sequencing reads that fail to represent any true genomic fragment.

We introduce a program named Quake that corrects errors in reads using a more robust model of sequencing errors than previous approaches. By using read quality values and learning the rates at which nucleotides are called as errors to different nucleotides, Quake achieves near perfect accuracy on simulated data. We also demonstrate the role of error correction with Quake in improving assembly and SNP detection.
Increasing the Genome Analyzer’s output using IBIS

Martin Kircher, Udo Stenzel and Janet Kelso

Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

Email: ibis@eva.mpg.de

Base calling, the conversion of intensity values acquired by the Genome Analyzer into bases, is hampered by a number of error sources. Different model-based approaches have been proposed (1-4), which are sensitive to updates to the instrument and chemistry. We have shown previously, that given a training data set, fast and accurate base calling can be performed using a statistical learner without the need to specifically model the sequencing process (5).

IBIS significantly reduces the error rate and thereby increases the output of usable reads. IBIS handled all available chemistries and all technical versions (GAI, GAII/Ix) without requiring modification. Further, in contrast to other publicly available base calling packages, we support all versions of GA Pipeline/OLB/RTA output for single read, paired end read and multiplex sequencing runs.

We demonstrate how an indexed control DNA library can be spiked into each of the eight lanes. The resulting 2% of control reads in each lane allow for lane-specific quality measurement and provide sufficient training data for IBIS. We suggest that even for whole runs of libraries with unbalanced base composition a set up omitting the dedicated control lane is feasible. From our experience, the spike-in can be done even when no index read is performed, given that there is a high sequence divergence of the control from the organism studied.

2. W. C. Kao, K. Stevens, Y. S. Song, Genome Res, (Sep 1, 2009).
High-throughput DNA sequencing – concepts and limitations

Martin Kircher and Janet Kelso
Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany
Email: martin.kircher@eva.mpg.de  kelso@eva.mpg.de

Short 50 word version:
Recent advances in DNA sequencing make it possible to generate huge amounts of sequence data very rapidly and at substantially lower cost. We review sequencing technologies currently available and show how vast increases in throughput are associated with both new and old types of errors in the resulting sequence data.

Long 250 word version:
Recent advances in DNA sequencing have revolutionized the field of genomics, making it possible for even single research groups to generate large amounts of sequence data very rapidly and at substantially lower cost. These high-throughput sequencing technologies make deep transcriptome sequencing and transcript quantification, whole genome sequencing and resequencing available to many more researchers and projects. However, while the cost and time have been greatly reduced, the error profiles and limitations of the new platforms differ significantly from those of previous sequencing technologies. The selection of an appropriate sequencing platform for particular types of experiments is an important consideration, and requires a detailed understanding of the technologies available; including sources of error, error rate, as well as the speed and cost of sequencing. We review the five sequencing technologies currently available on the market (capillary sequencing, pyrosequencing, reversible terminator chemistry, sequencing-by-ligation, and virtual terminator chemistry), and discuss the intrinsic limitations of each. We explain how the vast increases in throughput are associated with both new and old types of problems in the resulting sequence data, and how these limit the potential applications and pose challenges for data processing and analysis.
Recent advances in data analysis for Illumina sequencers and their application to the detection of somatic mutations in individual cancers

Klaus Maisinger
Associate Director Scientific Programming
Illumina Cambridge, Chesterford Research Park, Cambridge CB10 1XL, United Kingdom

We give an overview of the current status of data produced by Illumina sequencers such as GAIIx and HiSeq2000. Recent advances in throughput and data densities resulting from improvements to image analysis algorithms are discussed. We also characterize the data quality and accuracy obtained from the current SBS sequencing chemistry. As described elsewhere, the introduction of a new fast and sensitive gapped aligner, Elandv2, and a new indel finder module, GROUPER, as part of the CASAVA software package allow the detection of small indels in a size range from a single base pair to the full DNA insert size. In addition to these publicly released programs, we have developed software tools for the detection of larger structural variants and CNVs. The resulting analysis pipeline has been applied to the investigation of tumour-normal pairs from melanoma and ovarian cancer cell lines.
An algorithm for rapid identification of identical and near-identical sequences in pyrosequencing data sets

Monzoorul Haque M, Tarini S. Ghosh, Dinakar, Sudha Chandran, and Sharmila S. Mande

Bio-Sciences Division, Innovation Labs, Tata Consultancy Services Limited, Hyderabad, India

A significant proportion (4-44%) of sequences in genomic/metagenomic sequence data sets (generated using pyrosequencing technologies) are identical or near identical [1]. These redundant sequences are either a result of amplification artifacts (generated during sequencing library preparation) or are a result of over-abundance of an organism in a metagenomic sample. Identification of such redundant sequences is important, since they have a direct impact on the accuracy/performance of several downstream applications such as assembly, read mapping, variation analysis, taxonomic diversity estimates, etc. Few methods are currently available that can identify identical and/or near identical sequences from pyrosequencing data sets. In spite of having high sensitivity of detection, these methods (relying on an 'all-versus-all' sequence comparison approach) take enormous amounts of computing time for processing data sets having millions of sequences. Further, an exponential increase in time is observed with increasing size of the data set.

The aim of the current study was to develop an algorithm that can rapidly identify identical and near identical sequences in pyrosequencing data sets with high sensitivity. The goal was also to ensure that the execution time of the developed algorithm scales linearly with the size of the data set. With this motivation, we have developed an algorithm which rapidly processes a million sequences in a few minutes, and identifies redundant sequences with high sensitivity. The performance of the developed algorithm has been exhaustively validated with both simulated data sets as well as real metagenomic data sets. The performance of the present algorithm was compared with the recently published cd-hit-454 method [1].

On a desktop with just 1GB RAM memory, the developed algorithm identified identical and near-identical sequences from the mouse metagenome data set (having approximately 1.6 million sequences of read lengths approximately 100 bp) in less than 10 minutes as compared to more than 35 hours taken by the cd-hit-454 method. Positive validation results, coupled with the fact that the time taken by the developed algorithm increases linearly with the increasing size of the dataset highlight the immense utility of the proposed algorithm.

Details of the algorithm and validation results with various datasets will be presented during the conference.

References:
An algorithm for *in silico* identification of host DNA sequences from host-associated metagenomic data sets

Monzoorul Haque M, Dinakar Komanduri, Sudha Chandran, Tarini S. Ghosh and Sharmila S. Mande

Bio-Sciences Division, Innovation Labs, Tata Consultancy Services Limited, Hyderabad, India

The metagenomics approach employs high throughput techniques to simultaneously sequence and analyze the DNA content of millions of microbes obtained directly from environmental samples. However, in cases where DNA is obtained from microbes living in close association with a host (e.g. microbes living in human gut), microbial DNA gets contaminated with DNA of the host cells. It is crucial to identify and remove the contaminating host DNA sequences, since this contamination not only leads to incorrect estimates of microbial diversity but affects the accuracy of several other downstream analyses. Currently, few computational methods are used for identifying the contaminating host DNA sequences from typical metagenomic data sets (having millions of DNA sequences). These methods, being alignment-based, are slow, inefficient and require huge computing resources for performing the given task.

The aim of the current study was to develop a novel alignment-free method that can quickly and efficiently discriminate between DNA sequences originating from microbes and those originating from host cells. Besides, the goal was to develop a method, which end-users should be able to execute on a simple desktop having modest hardware specifications. With this motivation, we have developed an algorithm which can rapidly process a million sequences in a few minutes, and can identify contaminating host sequences with high sensitivity.

The developed algorithm has been exhaustively validated with both simulated data sets as well as real metagenomic data sets. Additionally, the performance of the method (in terms of sensitivity and time taken for identifying host sequences) was compared with corresponding results obtained using an alignment based method – megaBLAST [1]. On a desktop with 2 GB RAM memory, the developed algorithm was able to rapidly identify host DNA sequences (from a real metagenomic data set having millions of sequences) with high sensitivity. Results also indicate that the current method is at least 40-80 times faster than megaBLAST.

Validation results, coupled with the fact that the developed algorithm can be executed on a simple desktop with modest hardware specifications highlight the immense utility of the proposed algorithm for rapidly identifying host DNA sequences in metagenomic data sets with high sensitivity.

Details of the developed method, and validation results obtained with both simulated and real metagenomic data sets will be presented in the conference.

References:
Basecalling Models for Single Molecule Real Time DNA Sequencing

Patrick Marks, James Labrenz, Phillip McClurg, Jon Sorenson
Pacific Biosciences

Single Molecule Real Time (SMRT) DNA sequencing provides a real-time view of the template-directed synthesis of a new strand of DNA. Each base observation (and spurious error events) carries temporal, spectral, and intensity information governed by interactions of stochastic enzyme kinetics, template sequence context, fluorophore photophysics, and detector noise. Effective base calling for SMRT sequencing data requires techniques compatible with high dimensional, structured observation data. We have developed a gradient-boosted conditional random field that generates a model of the form $P(\text{dna sequence} \mid \text{pulse observations})$, that models indel and substitution errors, and exploits the rich set of pulse features available in SMRT data. This approach has proven useful in improving basecalling accuracy, providing quality values on basecalls, and finding the consensus sequence of multiple reads from a single circular molecule. Our results demonstrate that circular consensus sequencing leads to highly accurate sequences more rapidly than approaches unable to consider the full set of observed features.
Mining for cancer causing mutations in whole genome sequence data.


Abstract: The Cancer Genome Project at the Wellcome Trust Sanger Institute, UK (http://www.sanger.ac.uk/genetics/CGP/) is using high throughput sequencing technologies to systematically screen exome and genome sequence for diverse categories of mutations. The rapid developments in new sequencing technologies have allowed many new approaches to be applied to the detection and characterisation of a wide range of mutation types. We employ a variety of automated and semi-automated techniques to identify and classify copy number variation, chromosomal rearrangements, small insertion/deletion and substitution mutations in both cell lines and primary tumours.

Large scale genomic rearrangements can be discovered by mapping paired reads to a reference genome. High quality read mappings that produce unexpected pair sizes are used to identify putative rearrangement events. These can be confirmed either by local de novo sequence assembly and PCR, or by capillary sequencing. Copy number can be calculated by applying segmentation methods to correctly mapping read-pairs.

Smaller scale mutations can be detected using local read coverage to predict the genotypic and allelic content of samples. The detection techniques can be used on both whole genome and exome pull-down data. Substitutions are detected using a novel algorithm based on expectation maximisation (Cancer Variants from Expectation Maximisation, or CaVEMan) written in-house and confirmed using capillary sequencing. Small scale insertions/deletions are found using a combination of Pindel (http://www.ncbi.nlm.nih.gov/pubmed/19561018) and other methods.

Putative mutations of all types are compared to public resources (eg Ensembl, dbSNP) to infer genomic context and possible deleterious consequences. Once confirmed, mutations are annotated and submitted to COSMIC (http://www.sanger.ac.uk/cosmic/).
User-friendly tool for the analysis of the non-miRNA transcriptome in high-throughput sequencing datasets

Lorena Pantano\textsuperscript{1,2}, Xavier Estivill\textsuperscript{1,2,3} and Eulàlia Martí\textsuperscript{1,2}

\textsuperscript{1}Genetic Causes of Disease Group, Genes and Disease Program, Centre for Genomic Regulation (CRG), \textsuperscript{2}Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP) and \textsuperscript{3}Experimental and Health Sciences Department, Pompeu Fabra University, Barcelona, Catalonia, Spain

Small RNAs are short RNA molecules with regulatory functions on gene expression ranging from heterochromatin formation to mRNA destabilization. Therefore, small RNAs are involved in almost every biological process. According to their biogenesis and the gene regulation mechanism, small non-coding RNAs are classified in miRNAs (micro-RNAs), piRNAs (Piwi-interacting RNAs) and endo-siRNAs (endogenous small interfering RNAs). Recent progress in high-throughput sequencing has elucidated a remarkable landscape of small RNAs including novel non-canonical small RNAs derived from non-coding RNA (nasRNAs), repeats elements (rasRNAs), and coding regions (pasRNA) according to the deepBase database classification (http://www.deepbase.com). In addition to all these classes of small RNAs, in HeLa and HepG2 human cells there have been identified small RNAs that map onto exon-exon junctions of protein coding genes. At present, most analysis tools of high-throughput sequencing data are focused on miRNA characterization. However, no tools are prepared to cover a complete analysis of all types of small RNAs, and no visual user interface is provided, making difficult the inspection of these data.

Here we present a tool for the characterization of non-miRNA small RNA transcriptome integrated in the tool SeqBuster (http://estivill_lab.crg.es/seqbusterHiTSeq_2010_abstracts-1.doc). We have applied our tool to characterize human brain small-RNA sequencing data and also to public sequencing datasets.
SMALT - a new mapper for DNA sequencing reads

Hannes Ponstingl and Zemin Ning

Affiliation: Sequencing Informatics Division, The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.

Abstract: A computer program is presented that facilitates the efficient and accurate mapping of DNA sequencing reads onto genomic reference sequences. Reads from most types of sequencing platforms, for example Illumina, Roche/454 or ABI-Sanger, can be processed.

The software employs a hash index of short words, less than 15 nucleotides long, sampled at equidistant steps along the genomic reference sequences. For each read, potentially matching segments in the reference are identified from seed matches in the index and subsequently aligned with the read using a banded Smith-Waterman algorithm.

Mismatches, insertions and deletions are reported for the best gapped alignments of each read. A score for the reliability of the mapping is assigned to the best match. The user can adjust the trade-off between sensitivity and speed by tuning the length and spacing of the hashed words. A range of different output formats are supported (SAM, PSL, CIGAR etc.) for downstream analysis. A mode for the detection of split (chimeric) reads is provided. Multi-threaded execution is supported.

Speed, sensitivity and error rates were assessed on paired sequencing reads generated computationally from the sequence of the entire human genome. Single base changes and short insertions and deletions of up to 10 nucleotides were introduced randomly at a rate of 2% with every 5th variation an insertion or deletion. A total of 10^6 read-pairs were generated each read of a pair having a length of 100 nucleotides. These simulated reads were mapped on a single Intel E5420 2.5 GHz CPU. The average speed was 6x10^5 read-pairs per hour using 3.3 GB memory. 97% of the reads were confidently mapped (mapping score > 20) with an error rate of less than 0.1%.

The speed combined with high sensitivity and low error rates makes the program very useful for a wide range of genomic re-sequencing projects.

The software is available via FTP from ftp://ftp.sanger.ac.uk/pub/hp3/smalt.tgz
Next Generation Sequencing of Prebeta-1 HDL Extremes

Pfizer: Shobha Potluri, Lin Guey, Jason Laramie, Mike Schaffer, Steve Pitts, Purnima Sundar, Gabriella Huerta, Gautam Mehta, Dil Telman, Dong Liu, Steven Bentivegna, Jason Hughes, Jan Berka, Omar Francone, Albert Seymour
UCSF: Clive Pullinger, Brian Ishida, John Kane, Mary J. Molloy

Abstract:
Prebeta-1 high-density lipoprotein (HDL) is a molecular species of plasma HDL that participates in a cyclic process involved in the retrieval of cholesterol from peripheral tissues. The goal of the current study was to identify rare variants in genes that influence prebeta-1 HDL levels using deep sequencing. The hypothesis being tested is whether biological candidate genes which modulate HDL levels and/or participate in the reverse cholesterol transport process also modulate prebeta-1 HDL levels. Two hundred individuals were selected from the extremes of the preBeta-1 HDL distribution in a cohort of patients collected by our collaborators at UCSF. Seven candidate genes (PCOLCE, PCOLCE2, APOM, APOF, BMP1, WWOX, and CD93) were sequenced in the 200 individuals using 454 sequencing technology. Eighty percent of targeted bases were sequenced at least 25x coverage for 6 of the 7 genes. A total of 313 variants were found across the 200 samples. We found one nonsense mutation in BMP1 in a sample with low prebeta-HDL and a few frameshift mutations in WWOX. None of the variants identified were significantly associated with prebeta-1 HDL levels when examined individually or in an aggregate across the gene.
PALMapper: Fast and Accurate Spliced Alignments of RNA-seq Reads

Gunnar Rätsch\textsuperscript{1*}, Géraldine Jean\textsuperscript{1}, André Kahles\textsuperscript{1}, Sören Sonnenburg\textsuperscript{3}, Fabio De Bona\textsuperscript{1}, Korbinian Schneeberger\textsuperscript{2}, Jörg Hagmann\textsuperscript{2}, and Detlef Weigel\textsuperscript{2}

\textsuperscript{1} Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany
\textsuperscript{2} Max Planck Institute for Developmental Biology, Tübingen, Germany
\textsuperscript{3} Machine Learning Group, Berlin Institute of Technology, Berlin, Germany

We propose a combination of the spliced alignment method QPALMA [1] with the short read alignment tool GenomeMapper [2]. The resulting method, called PALMapper, efficiently computes both spliced and unspliced alignments at high accuracy while taking advantage of base quality information and (optionally) splice site predictions. QPALMA relies on a machine learning strategy and is highly sensitive and adaptive to the error characteristics in the reads. It, however, suffers from its time consumption in the Smith-Waterman alignment step. To speed this up and thus to improve efficiency, we combined it with GenomeMapper that quickly carries out an initial read mapping. Resulting partial alignments, information from previously mapped reads, and (optionally) computational splice site predictions then guide a very efficient \textit{banded} Smith-Waterman-like algorithm that allows for long gaps that correspond to introns. PALMapper considerably reduced time consumption without decreasing accuracy compared to QPALMA. In fact, it runs around 50 times faster and allows to align around 7 million reads per hour on a single AMD CPU core (similar speed as, for instance, TopHat [3]). PALMapper only requires a few matching seeds to trigger an alignment that may have arbitrarily many mismatches or gaps. It is therefore also very well suited for reads from other technologies (e.g., Pacific Biosciences), which exhibit many deletions, where many other mapping approaches are deemed to fail. We illustrate that PALMapper indeed benefits from computational splice site predictions and present a strategy to easily train a splice predictor: the predictor training takes advantage of very confidently aligned reads obtained during a first iteration of PALMapper without splice site predictions to then improve the accuracy of more difficult cases during a second iteration integrating splice site information. PALMapper can therefore be used to align reads to arbitrary genomes and at the same time build a splice site predictor applicable to the whole genome.

We performed an extensive comparison with other spliced alignment methods. Our study shows that PALMapper predicts introns with very high sensitivity (72\%) and specificity (82\%) when using the annotation as ground truth. PALMapper is considerably more sensitive than, for instance, TopHat (47\% and 81\%, respectively).

PALMapper is open source and available from http://fml.mpg.de/raetsch/suppl/palmapper. Moreover, it can be used in the Galaxy instance available at http://galaxy.fml.mpg.de in combination with other tools for transcriptome reconstruction and quantitation.

\begin{thebibliography}{9}
\end{thebibliography}
The foundation for probabilistic treatments of DNA sequence information was set forth in the 1990s by developments such as the Phred model for predicting per-base error probabilities (Ewing & Green, 1998) and the Bayesian treatment of consensus (Churchill & Waterman, 1991). With the advent of next-generation sequencing technologies, the number of sequencing error models has increased in proportion to the number of new sequencing approaches. Our models for treating sequence error have not kept up, and the result is that information is lost as we attempt to squeeze new types of sequence data into older representations. Recently several groups have rallied around an “event-modeling” approach to DNA sequence representation which aims to compactly describe the full amount of information obtained from any sequencing experiment, in a technology-agnostic format. In this work we show how information theoretic approaches to DNA sequencing help quantify what is or isn’t lost when trading off different error types and readlengths. We back up these theoretical notions with simulated and empirical results. We also describe the theory and software development which layers upon the event-modeling representation of DNA information and offers access to fundamental analyses like genotyping, mapping significance, and consensus calling in a platform-independent fashion. These approaches are expected to remain relevant far into the future even as sequencing platforms evolve past 2nd and 3rd generation technologies.
Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps

Isheng Jason Tsai1, Thomas Otto1, and Matthew Berriman1

1 Parasite Genomics, Wellcome Trust Sanger Institute

With the recent advances in second-generation sequencing technologies (454 pyrosequencing, Illumina, SOLiD, and Helicos), genome projects have seen an explosion of sequence data production at a fraction of the cost per-base. However, the assembled data frequently are highly fragmented with many gaps that have also made the relative cost of manual finishing become prohibitive. As a result most genomes are being released as draft only. These draft sequences have important shortcomings: they vary greatly in their qualities and are interrupted by many gaps.

Our group is committed to producing high quality reference genomes and as part of that commitment we have developed an approach termed Iterative Mapping and Assembly for Gap Elimination (IMAGE), to drive the quality of draft assemblies towards “finished” but without manual intervention. The approach utilises the large number of sequences that an Illumina Genome Analyzer produces. Reads that correspond to gaps or questionable regions are identified and reassembled locally before being incorporated back into the final assembly. An advantage of a local assembly as opposed to a de novo one is that the number of reads used is only a fraction of total available reads. This reduces the complexity of regions to be assembled as well as the time and computing memory required. We show IMAGE can greatly improve draft assemblies for several ongoing genome projects.

The complete genome sequence of an organism provides an invaluable resource to wider research community and is the foundation for comparative and evolutionary genomics studies. We believe that this is a simple, practical, and effective approach and will therefore be of wide interest to anyone producing genome data.
A Novel Algorithmic Combined Approach to Accurately Determine Inversions and Large Deletions both in Confidence and in Genomic Location

Yutao Fu*, Caleb Kennedy*, Eric Tsung*, Stephen McLaughlin, Heather Peckham, Aaron Kitzmiller, Clarence Lee

* These authors contributed equally.

Advancing algorithms enable short read technologies to determine structural variations with ever increasing accuracy. While a single algorithmic approach gives evidence towards these events, they often have inherent weaknesses in accuracy of the call itself or the positioning of breakpoints. What we present here is how we combine multiple approaches, using different alignment types, to strengthen our confidence in both of these areas.

To detect inverted sequences, a search is done for pairs of reads that align to inverted orientation relative to normal pairs. For large deletions, paired tags are isolated by combining evidence from discordant pairs: sets of reads with insert sizes that deviate more than the expected average insert size of the entire library. Both these techniques provide some confidence of the event within a range of genomic locations, but exact breakpoints can be hard to determine. Precision of these locations and increased confidence of the call benefits from accurately determining split reads which we do by effectively aligning disjoint parts of the read. As a first round of validation, we present how we compare with existing databases of these structural variants.
MMSEQ: fast, scalable isoform expression estimation using multi-mapping RNA-seq reads.

Ernest Turro, Sylvia Richardson, Alex Lewin
Dept of Epidemiology & Biostatistics, Imperial College London

Abstract:
Second-generation sequencing of RNA samples (RNA-seq) allows researchers to quantify the abundance of transcripts in biological samples with greater sensitivity than microarrays, which suffer from non-specific hybridisation and saturation biases. RNA-seq, however, poses new statistical and computational challenges. Sequence reads are in general much shorter than transcripts and may therefore map to multiple isoforms of the same gene or even to multiple genes sharing a subsequence. Moreover, paired-end reads and reads that straddle one or more exon junctions complicate the mapping process. There is a need for methods that deal with isoform-level quantitation using all mappable reads in a principled, scalable and user-friendly way.

We present a fast and scalable method for isoform and gene expression estimation called MMSEQ, freely available from http://bgx.org.uk. Each read is mapped to a set of reference transcripts and the total number of reads mapping to each set is counted. The contribution from multi-mapping reads to the expression signals is then disaggregated using expectation-maximisation and Gibbs sampling to provide expression estimates and accompanying standard errors for each isoform and gene. We assess our method using a simulated data set and a real data set consisting of 61 HapMap individuals. We show that MMSEQ scales well with increasing coverage and provides better expression estimates than the current state of the art.
The challenges of de novo assembly from polymorphic populations

Zhengyuan Wang, Liang Ye, and Makedonka Mitreva
The Genome Center
Washington University School of Medicine

Genetic polymorphism is a challenge for de novo assembly of population-based genome sequencing using next generation technology, such as those of nematodes and microbes. We studied the effect of polymorphism on de novo assembly quality via a series of simulations. We generated two types of populations of C. elegans chromosome I, one with only SNP, the other with both SNP and indel polymorphism. Each type had two populations with two levels of polymorphism rates, 0.01 and 0.025, measured by theta. Fragment and paired end 454 sequences were simulated to different coverages from these populations. Simulated reads were assembled using Newbler. Assembled contigs were mapped back to C. elegans chromosome I to evaluate the quality of assemblies. Our result showed that high polymorphism, especially from indels, significantly reduced assembly quality measured by assembly contiguity and source chromosome coverage. Increasing sequencing depth or including paired end reads could not significantly improve assembly quality, but using two chromosomes only (i.e. an individual organism) could improve the assembly a lot, probably because of the reduction of segregating sites in the sequences. Chromosome regions with low complexity had a low coverage no matter if the polymorphism rate was high or low, either with or without indel polymorphism. Our simulation revealed new insights into the challenges facing de novo assembly of polymorphic genomes, and suggested that new methodologies are necessary for better de novo assemblies.
A fast mapping tool for DNA sequence reads from the SOLiD™ System: part of Bioscope™

Zheng Zhang, Jing Zhai, Danwei Guo, Yuandan Lou and Asim Siddiqui
Life Technologies, 850 Lincoln Centre Dr., Foster City, CA

The wide adoption of next generation sequencing technologies demands a fast mapping tool that can map billions reads against genomes of human size. Currently, most of the existing short read mapping tools solve an approximate string matching problem allowing $k$ mismatches. Hashes and other implementations of suffix trees are the two main methods used. As read length increases, the error rate difference between the beginning and end of the reads increases. In order to achieve good sensitivity, a large number of mismatches needs to be allowed. This slows down the mapping significantly, especially for suffix based methods. At a time when the throughput of the technologies continues to grow rapidly, more efficient approaches are imperative.

To address this problem, we adopt an anchored mapping strategy. We first pick a region in the read as anchor (typically the first 25 bases of the read). Then, we perform approximate string matching on the anchored region to the reference. For every match to the anchored region, we find the best ungapped local alignment extended from the anchor match. Formally, we define a scoring function where a matched base has a score of 1 and a mismatch, a user defined negative penalty. The match to the anchored part of read can be extended to any substring of the read, and each extended alignment will have a score dependant on the number of matches and mismatches. The local alignment process picks the substring with the highest score.

For current next generation sequencing technologies, there exist a number $N$ where the first $N$ bases of any read are of high quality, and from the $N$th base on, it become worse. The main advantage of our approach is that we can use the high quality bases as an anchor region, and allow relatively few mismatches in that region. This makes the matching of the anchored region very fast. The local extension part is a simple computation and will be performed only after a high quality hit is found. Hence, it will be performed relatively infrequently, leading to an overall running time that is much faster than trying to align the full read.

To further generalize the approach, we allow the user to specify any part of the reads as the anchor, and use an iterative approach to improve sensitivity. For example, the user can specify the first 25 bases as the anchor in a first round of mapping. Then, for reads that do not map in the first round, they can perform a second round of mapping using the next 25 bases, etc. In Bioscope, we provide a framework that allows the user the flexibility to design their own iterative strategy. In addition, this feature is very useful in whole transcriptome experiments where reads often map across exon boundaries.