HiTSeq 2011 Schedule & Poster Book
Vienna, July 15-16

July 15th

Session 1: 8:45am-10:15am

- Introduction
- Keynote 1: Pamela Hoodless
- Ali Mortazavi: ChIP-seq regulatory analysis using ChIA-PET

Coffee Break: 10:15-10-45am

Session 2: 10:45am-12:30pm

- Darshan Singh: A Graph-based Statistical Method to Detect Differential Transcription from RNA-seq Data
- Lucas Swanson: Browsing assembled RNA for chimera with localized evidence
- Peter Glaus: Estimating differential expression of transcripts with RNA-seq by using Bayesian Inference
- Adam Roberts: A Bayesian online EM algorithm for isoform-level RNA-Seq quantification
- Regina Bohnert: Quantitatively deconvolving alternative RNA secondary structures

Lunch: 12:30pm-1:45pm

Session 3: 1:45pm-3:30pm

- Leena Salmela: Correcting errors in short reads by multiple alignments
- Doron Lipson: An Assembly-based Algorithm for Sensitive Detection of Insertions and Deletions in Targeted Resequencing Data of Clinical Cancer Specimens
- Dan He: Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions
- Veli Makinen: Partitioning the Scaffolding Problem into Small Independent Mixed Integer Programs
- Andrew Roth: JointSNVMix : A Probabilistic Model For Accurate Detection Of Somatic Mutations In Normal/Tumour Paired Sample Sequence Data

Coffee Break: 3:30-4:00pm

Poster Session: 4-6pm

July 16th

Session 4: 8:55am-10:15am

- Keynote 2: Peer Bork
- Henry Leung: A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio

Coffee Break: 10:15-10-45am

Session 5: 10:45am-12:30pm

- Thomas Wu: Improvements to GSNAP and development of the GSTRUCT pipeline for analyzing RNA-Seq data
- Raffaele Calogero: Dissecting a massive parallel sequencing workflow for quantitative miRNA expression analysis
- Jonas Behr: Multiple reference genomes and transcriptomes for Arabidopsis thaliana
- Ryan Giuliany: Auditor: exploring RNA-editing in cancer through simultaneous analysis of tumour genome and transcriptome sequencing data
- Andrew McPherson: Comrad: a novel algorithmic framework for the integrated analysis of RNA-Seq and WGSS data

Lunch: 12:30pm-2:00pm

Session 6: 1:45pm-3:30pm

- Sharmila Mande: Community-Analyzer: A visualization and analysis tool to study the structure and dynamics of microbial communities across metagenomic data sets
- Gavin Ha: Copy number aware Bayesian approach to detect loss of heterozygosity in tumour genome sequence data
- Orion Buske: Variant detection and the Autism Sequencing Project
- Pina F. I. Krell: Sequencing error correction to reliably measure diversity of the human T cell receptor repertoire
- Christian Rödelsperger Identity-By-Descent Filtering of Exome Sequence data for Disease-Gene Identification in Autosomal Recessive Disorders

Coffee Break: 3:30-4:00pm

Session 7: 4:00pm-5:30pm

- Marc Fiume: MedSavant: a platform for identifying causal variants from disease sequencing studies
- Keynote 3: Stefan Schreiber

ChIP-seq regulatory analysis using ChIA-PET

Katherine Fisher[1], Huay Mei Poh[2], Yijun Ruan[2], Barbara Wold[1,3], Ali Mortazavi[1,3]

1). Division of Biology, California Institute of Technology, Pasadena, CA, USA 91125
2). Genome Institute of Singapore, Singapore
3). Beckman Institute, California Institute of Technology, Pasadena, CA, USA 91125

The identification of enhancers and their assignment to their respective promoters remains a significant challenge in deciphering the regulatory logic of mammalian transcription. In particular, long-range interactions that can span many genes are likely to account for a significant fraction of otherwise unexplained transcription factor ChIP-seq signals in gene deserts or at differentially expressed promoters without their expected binding sites. We describe a graph-theoretical method of analyzing ChIP-seq peaks and their long-range interactions using the emerging technique of ChIA-PET (Chromatin Interaction Analysis using Paired-End Tags). We conducted ChIA-PET experiments for RNA Polymerase II and the transfection factor Myogenin in mouse C2C12 cells undergoing skeletal muscle differentiation. Our results show that nearly half of the myogenic binding sites are involved in detectable long-range interactions and that 19% of these involve transcription starts sites that are two or more genes away. We further observe that over 12% of active TSS participate in fully connected promoter-promoter interactions that are likely mediated by co-occupancy at transcription factories and also find large fully connected set of interacting transcription factor ChIP-seq sites interacting cooperatively. Our results suggest that the current ChIP-seq regulatory assignments to the nearest gene are missing a significant fraction of the relevant interactions. We discuss the extension of the analysis to the recently deposited human ENCODE ChIA-PET data.

FDM: A Graph-based Statistical Method to Detect Differential Transcription from RNA-seq Data

Darshan Singh[1], Christian F. Orellana[1], Yin Hu[5], Corbin D. Jones[2], Yufeng Liu[3], Derek Y. Chiang[4], Jinze Liu[5], Jan F. Prins[1]

1). Departments of 1).Computer Science, 2).Biology, 3).Statistics and Operations Research, 4).Genetics – University of North Carolina at Chapel Hill, USA
5).Department of Computer Science – University of Kentucky, USA

Motivation

RNA transcript diversity is achieved through alternative splicing and is important in differentiation of cell function and in response to environmental conditions. Consequently there is a great need for methods to detect differential RNA-transcript expression between samples. For samples obtained using high-throughput short-read sequencing of the transcriptome (RNA-seq), several approaches have emerged: those based on transcript inference and quantification (e.g. Cuffdiff [1]), and those based on coverage differences of read alignments to the genome and without need knowledge of transcript identities (e.g. rdiff using the Maximum Mean Discrepancy kernel [2]). The former approaches are sensitive to errors due to transcript ambiguity and quantification uncertainty, while the latter only provide an indirect measure of alternative splicing. We propose a new method in the second category, based on spliced read alignments, that more directly represents and localizes differential transcription.

Methods

We characterize *differential transcription* between two samples as the difference in relative abundance of the transcripts present in the samples. The difference can be measured by the square root of the Jensen-Shannon Divergence (sqrt(JSD)) of the relative transcript abundances. The use of relative abundance emphasizes differential transcript expression instead of gene expression.

We define a weighted splice-graph representation of RNA-seq data, summarizing in compact form the alignment of RNA-seq reads to a reference genome. The Flow Difference Metric (FDM) is defined between pairs of graphs and identifies regions of differential RNA-transcript expression, without need for an underlying gene model or catalog of transcripts. Using simulated data with controlled JSD, we show that the FDM is highly correlated with the sqrt(JSD) when average RNA-seq coverage of the transcripts is sufficiently deep ($r = 0.82$).

We develop a novel non-parametric statistical test between splice graphs to identify significant differences in transcription between individual samples, and extend the test to operate between sample replicates to show identify significant differences between samples consistent across replicates.

Results

We applied Cuffdiff [1], rdiff (Poisson) and rdiff (Maximum Mean Discrepancy) [2], along with FDM to simulated RNA-seq data. Refseq gene models and empirical transcript expression models were used to create genes in two samples with varying sqrt(JSD). The genes were then sampled to create simulated RNA-seq data. Results are shown in Figure 1.

Using experimental data consisting of four replicates each of cancer cell lines MCF7 and SUM102, FDM identified 1425 genes as significantly different in transcription. Subsequent study of several differential transcription sites using qRT-PCR confirmed significant differences between the samples.

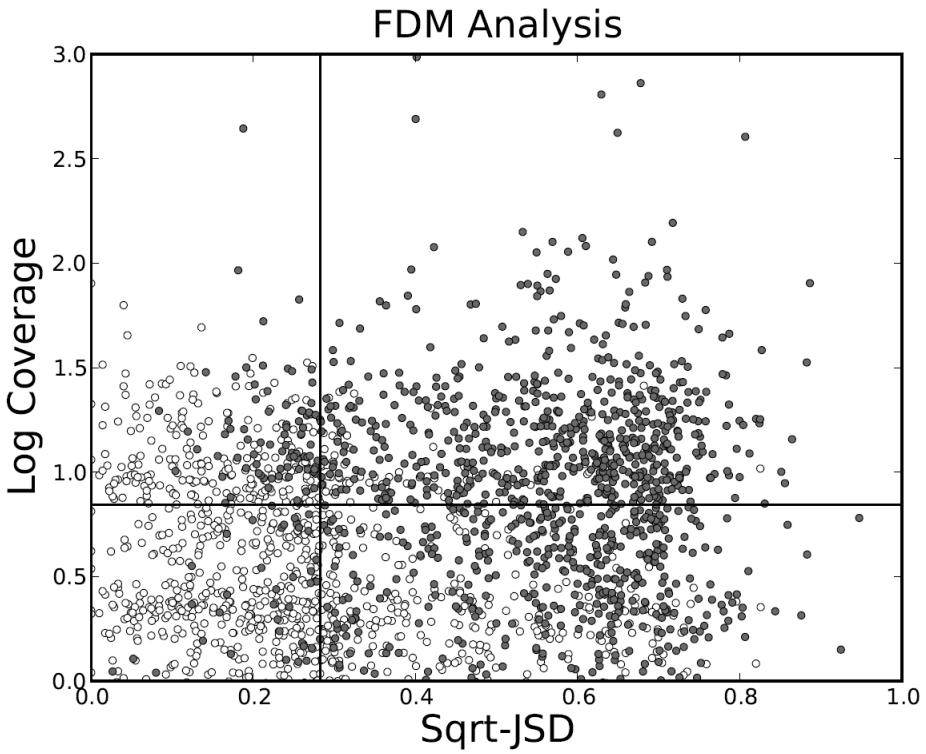


Figure 1: The circles in the scatterplot represent 2100 genes in two samples with varying differential transcription (measured by the sqrt(JSD)) and varying depth of RNA-seq sampling (measured by the average number of reads that cover a transcribed nucleotide). Solid circles correspond to genes with significant differential transcription according to the FDM metric and the statistical test. The FDM consistently identifies differential transcription when coverage is high or the sqrt(JSD) measure is high. For example, for genes with sqrt(JSD) > 0.28 and log(coverage) > 0.85, FDM was able to identify 90% of the genes as differentially transcribed. This represents higher sensitivity than the Cufflinks, rdiff Poisson, and rdiff MMD methods, which identified differential transcription between 24%, 34%, and 49% of the genes in this region, respectively.

[1] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515.

[2] Stegle, Oliver, Drewe, Philipp, Bohnert, Regina, Borgwardt, Karsten, and Rätsch, Gunnar. Statistical Tests for Detecting Differential RNA-Transcript Expression from Read Counts. Available from *Nature Precedings* <http://dx.doi.org/10.1038/npre.2010.4437.1> (2010)

Browsing assembled RNA for chimera with localized evidence

Lucas Swanson[1,2], Karen Mungall[1], Gordon Robertson[1], Readman Chiu[1], Shaun D Jackman[1], Jenny Q Qian[1], Sam Lee[3], Deniz Yorukoglu[2], Rong She[1], Yongjun Zhao[1], Richard Moore[1], Marco A Marra[1], Steven JM Jones[1], Aly Karsan[1], Pamela A Hoodless[3], S Cenk Sahinalp[2], Inanc Birol[1]

[1] Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, B.C., Canada
[2] School of Computing Science, Simon Fraser University, Burnaby, B.C., Canada
[3] Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, B.C., Canada

Through the combinatorial process of alternative splicing, cells are able to produce a complex set of proteins. The discovery of 'chimeric' transcripts that cannot be explained by traditional models of alternative splicing shows that the complexity of protein production is even greater than once believed. Chimeric transcripts can be caused by genomic rearrangements, read-throughs, or trans-splicing. Trans-splicing refers to mRNA splicing that occurs in a non-colinear fashion and may involve multiple distinct transcripts.

Chimeric transcripts due to genomic rearrangements are important in human cancers. For example, in leukemia, fusions between the BCR and ABL genes are commonly detected, and partial tandem duplications in the MLL gene or internal tandem duplications in FLT3 are associated with unfavourable prognoses.

Transcripts due to trans-splicing are functionally important in kinetoplastids, euglenoids and nematodes. Trans-splicing does occur in normal tissues in higher animals, including mammals, albeit at low levels of expression. This makes its functional importance in mammals controversial, but shows that the mechanism is present. We hypothesize that cancer cells could take advantage of deregulation of this mechanism.

Because of the ability to generate complex, non-colinear transcripts and the wide range of expression levels in transcriptome data, it is difficult to efficiently detect and characterize chimeric transcripts using short-read sequencing technologies. Because we expect chimeric transcripts to occur only at low expression levels in normal tissues, any detection tool would have to be very sensitive, yet robust to noise, i.e. very specific. In the past, effort was mainly focused on studying standard colinear cis-splicing: so many existing tools are not optimized for detecting chimeric transcripts. Chimeric transcripts lead to complicated short-read alignments that are difficult to interpret. From our experience analysing transcriptomes using *de novo* assembly, we suggest that the increased alignment specificity of the longer sequences assembled from these reads will yield more interpretable alignments that have predictable signatures, facilitating the characterization of chimeric transcripts.

We have developed a suite of analysis tools, called BARNACLE, for identifying a wide range of structural variation types in *de novo* assemblies of RNA-seq data. In this work, we focus on three event types: fusions, partial tandem duplications (PTDs), and internal tandem duplications (ITDs). We identify a fusion chimera as having parts of transcripts from two distinct genes joined together, a PTD chimera as having an exon or group of exons tandemly repeated within it, and an ITD chimera as having a small portion of an exon tandemly repeated within it.

To predict and characterize chimeric events, BARNACLE integrates evidence from a range of data types: assembled transcriptome contigs, contig alignments to a reference genome, read

alignments to assembled transcriptome contigs and to the reference genome, and gene and repeat annotations. The read-pair-to-genome alignment data is generated with a pipeline that addresses the issues particular to aligning transcriptome reads to a genome.

We have tested BARNACLE using transcriptomes assembled and analyzed with the Trans-ABySS pipeline. To capture a wide range of expression levels, multiple assemblies are performed with appropriate parameter settings, and the contig sets are merged into a meta-assembly of non-redundant contigs. These contigs are then filtered and aligned to the reference genome. The Trans-ABySS outputs that BARNACLE uses are the filtered contig sequences in FASTA format, the contig-to-genome alignments in PSL format, and the read-to-contig alignments in BAM format.

BARNACLE begins by examining the contig-to-genome alignments, identifying the alignment or alignments that best represent each contig, and comparing those alignments to the chimeric event signatures that we have identified. Contigs that match event signatures are identified and grouped by their genomic event locations. Read-pair-to-genome and read-to-contig alignments are then examined to determine the level of support for these putative events, which is used to filter out false positives. At this point, additional annotations are associated with each event: which gene exons the event overlaps, which annotated repeat regions the event overlaps, and whether the genomic breakpoints match annotated exon boundaries.

These events are then filtered based on:

1. support from read-pair-to-genome alignments
2. support from read-to-contig alignments
3. percent identity of the contig-to-genome alignments
4. total fraction of the contig represented by the contig-to-genome alignments
5. number of different events that the contig alignment(s) could represent
6. whether the contig can be mapped to multiple genomic locations
7. whether the event represents a homopolymer sequence

Finally, the events that pass the filters are classified into inferred biological event types based on alignment signatures and sequence properties.

When datasets are available from multiple samples (libraries), BARNACLE can perform across-library post-analysis to identify genes that have events in multiple libraries and prioritize events for manual review.

We have tested BARNACLE on transcriptomes of seven embryonic and adult normal mouse tissues. We predicted fusions involving 164 genes, PTDs involving 49 genes, and ITDs involving 1158 genes. Fusions recurred across libraries in 58 genes, PTDs in 11 genes, and ITDs in 168 genes. Using public EST databases, we have found evidence that a subset of these fusions and PTDs also occur in orthologous genes in humans. We estimated the expression levels of our predicted fusion and PTD chimeric transcripts relative to their corresponding wild-type transcripts. BARNACLE was sensitive enough to detect these chimeras, even though their low relative expression levels suggest that they are unlikely to be functionally important in these normal tissues.

These results on well-annotated normal tissues suggest that *de novo* assembly and BARNACLE will be effective in characterizing chimeric transcripts and mechanisms responsible for such transcripts in complex cancer transcriptomes.

Estimating differential expression of transcripts with RNA-seq by using Bayesian Inference

Peter Glaus[1], Antti Honkela[2] and Magnus Rattray[3]

1). The University of Manchester
2). Helsinki Institute for Information Technology HIIT, University of Helsinki
3). The University of Sheffield

High-throughput sequencing enables expression analysis at the level of individual transcripts.
The analysis of transcriptome expression levels and differential expression estimation requires a probabilistic approach to properly account for the ambiguity caused by shared exons and finite read sampling as well as the intrinsic biological variance of transcript expression. Another important factor are the biological sources of variance, which, as we show in our analysis, can be substantial and may dependent on the transcript expression level. To avoid false positive differential expression calls, one has to anticipate the intrinsic variance of the transcript expression levels using empirical prior knowledge and information from replicates where they exist.

We present a Bayesian approach for estimation of transcript expression level from RNA-seq experiments. Inferred relative expression is in the form of a probability distribution represented by samples of the distribution obtained from a Markov chain Monte Carlo inference method applied to a generative model of the read data. Additionally to implementation of the regular Gibbs sampling algorithm, we provide a comparison with Collapsed Gibbs sampling in which some of the parameters are marginalized in order to obtain faster convergence.

We propose a novel method for differential expression analysis across replicates which propagates uncertainty from the sample-level model while modelling biological variance using an expression-level dependent prior.We demonstrate the advantages of our method using a RNA-seq dataset (Xu G. et al., RNA 2010) with technical and biological replication for both studied conditions.

A Bayesian online EM algorithm for isoform-level RNA-Seq quantification

Adam Roberts and Lior Pachter

Since the advent of RNA-Seq, there has been an explosion of methods for estimating transcript abundances at the isoform-level. Most of these approaches are based on maximizing some likelihood function using a batch EM algorithm. These algorithms have shown excellent results in simulation studies as well as in comparison to alternative technologies such as microarrays and qPCR. However, as the speed of sequencing continues to increase and costs fall, there is a need for quantification methods that scale linearly with the number of sequenced fragments.

Here we present the first online algorithm for transcript abundance estimation using RNA-Seq data. Our method, which is implemented in a software program called CuffExpress, takes advantages of the excellent convergence properties of the online EM algorithm to deconvolute ambiguous read mappings on the fly. Our model has all of the features of the Cufflinks model including bias correction. Due to its speed, CuffExpress can consider many more ambiguous mappings of each fragment and uses an error model to aid in the probabilistic fragment assignments. All parameters including those for bias, error, and fragment length are updated in an online manner along with the transcript abundances, variances, and covariances.

We show that the method is highly competitive with the state-of-the art batch algorithms even at relatively low sequencing depth. Furthermore, we present several possible applications including early sequencing termination using convergence detection, a low-memory web interface, and allele-specific expression estimation.

Quantitatively deconvolving alternative RNA secondary structures

Regina Bohnert[1], Fabio De Bona[1,2,] and Gunnar Rätsch[1]

1). Friedrich Miescher Laboratory of the Max Planck Society, Spemannstraße 39, 72076 Tübingen, Germany
2). Google Switzerland

Secondary structures of RNA transcripts play an important role in basic cellular processes, ranging from aiding in protein synthesis as tRNA to their key role in regulation of alternative splicing of RNA transcripts [1]. For our understanding of their function and their mechanism of action, it is therefore crucial to know the molecule secondary structure. Prediction of RNA secondary structure from the sequence has been one of the early problems of computational biology. The recent development of a high-throughput experimental method has opened new potentials to improve computational prediction of RNA secondary structure [2,3]. The outcome of such experiments can give evidence where paired and unpaired nucleotides are located in the RNA transcript.

To consider this rich set of measurements in computational structure folding, we have developed a novel strategy based on a machine learning algorithm, called RFP. For scoring of RNA secondary structures it combines terms related to the free energy and the read counts evidencing paired and unpaired bases (corresponding to V1 and S1 digested RNA libraries, see e.g. [2]). It is trained using a supervised learning algorithm based on known RNA sequences, their structures and simulated read counts. We will show that by using the read information we can significantly improve the prediction accuracy. The approach is being different from the one proposed in [2] based on constraint folding in that it is able to deal with incomplete and contradicting information to predict the most likely RNA secondary structure.

Often, not only a single structure is folded from a RNA transcript, but different structures of the same transcript may co-exist in a cell. Read counts from a digestion/sequencing experiment then reflect a mixture of the underlying alternative structures, making it difficult to simply read-off the structure from the observed read counts. Assuming that the set of alternative structures is known, we can formulate the problem as a mathematical program. We implemented this idea in the tool sQuant, which infers the abundances of the structures by minimizing the deviation of expected and observed read counts, respectively. The regularization of the abundance variables with the L_1-norm provides a sparse solution. Therefore, such an approach is highly suitable to prune a set of alternative structures, resulting in a small set of structures consistent with the read data. To evaluate our method we assembled a set of simulated read counts. We predicted seven structures per transcript, applying the ViennaRNA RNAsubopt tool [4] on a set of 1,000 transcripts randomly selected from genes of the TAIR10 *A. thaliana* annotation. We randomly assigned abundances to these transcripts and sampled reads from a Poisson distribution at paired and unpaired positions, corresponding to V1 and S1 digested RNA libraries (cf. [2]), respectively. We found that sQuant's predictions very well correlate to the underlying true abundances (Pearson's correlation 0.96); illustrating that RNA secondary structure deconvolution is feasible.

We will discuss strategies to deal with the case where not all alternative structures are not known yet. Together with a structure prediction program such as RFP or RNAsubopt, we can use sQuant to deconvolve and identify a small set of structures underlying a mixture of read measurements. For instance, in an iterative manner,

RFP first predicts a structure candidate, which is then added to the active set of alternative structures. This set is subsequently quantified by sQuant, taking the residue of observed and predicted abundance as an input to RFP in the next iteration. The combination of computational structure prediction and quantitation for RNA entities creates a powerful tool to describe the set of RNA structures and their relative abundance in a cell, enabling further analyses to study RNA function and regulation.

[1] CJ McManus and B. R. Graveley. RNA structure and the mechanisms of alternative splicing. Current Opinion on Genetics & Development, 2011.
[2] M Kertesz et al. Genome-wide measurement of RNA secondary structure in yeast. Nature, 467(7311):103-7, 2010.
[3] JG Underwood et al.. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nature Methods, 7(12):995-1001, 2010.
[4] IL Hofacker et al. Fast Folding and Comparison of RNA Secondary Structures. Monatshefte für Chemie / Chemical Monthly, 125:167-188,1994

Correcting Errors in Short Reads by Multiple Alignments

Leena Salmela[1] and Jan Schroder[2]

1.Department of Computer Science and Helsinki Institute for Information Technology HIIT, University of Helsinki, Finland 2.NICTA Victorian Research Lab, Department of Computer Science and Software Engineering, University of Melbourne, Australia

Motivation: Current sequencing technologies produce a large number of erroneous reads. The sequencing errors present a major challenge in utilising the data in de novo sequencing projects as assemblers have difficulties in dealing with errors.

Results: We present Coral which corrects sequencing errors by forming multiple alignments. Unlike previous tools for error correction Coral can utilise also bases distant from the error in the correction process because the whole read is present in the alignment. Coral is easily adjustable to reads produced by different sequencing technologies like Illumina Genome Analyzer and Roche/454 Life Sciences sequencing platforms because the sequencing error model can be defined by the user. We show that our method is able to reduce the error rate of reads more than previous methods.

Availability: The source code of Coral is freely available at http://www.cs.helsinki.fi/u/lmsalmel/coral/.

Contact: leena.salmela@cs.helsinki.fi

An Assembly-based Algorithm for Sensitive Detection of Insertions and Deletions in Targeted Resequencing Data of Clinical Cancer Specimens

Doron Lipson[1], Michael Schnall-Levin[1], Roman Yelensky[1], Alex Parker[1], Mirna Jarosz[1], Frank S. Juhn[1], Zac Zwirko[1], Kristina Brennan[1], Troy Bloom[1], Sean Downing[1], John Curran[1], Jeffrey Ross[1,2] , Maureen Cronin[1]

1). Foundation Medicine, Inc., Cambridge, MA
[2]). Department of Pathology and Laboratory Medicine, Albany Medical College, Albany, NY

Rapid advancement in the understanding of cancer genomics and the growing number of available targeted therapies provide expanding opportunities for effective cancer treatment based on comprehensive tumor profiling. Although significant progress has been made in experimental and computational approaches for analyzing tumor genomes by next-generation sequencing in the research setting, extending these techniques to the clinical setting poses significant additional challenges. Key among these is the limited purity and heterogeneity of clinical specimens, coupled with requirement to provide high sensitivity for a wide range of potentially clinically-actionable mutations of different types.

Whereas point mutations are typically detected with high sensitivity using widely-adopted short-read mapping pipelines, multi-nucleotide deletions and insertions as well as more complex mutations (such as local inversions) have been notably harder to identify robustly. This is mainly due to the limited ability of many rapid mapping algorithms to correctly position reads in the presence of multiple differences from the reference sequence. While some of these limitations can be countered by more sophisticated mapping techniques, these typically come at the cost of a longer analysis running time which may be prohibitive in a clinical setting, and are generally incapable of handling longer insertions.

For a targeted resequencing application, where a limited number of targets are sequenced to very high depth, an attractive alternative is an assembly-based approach, which minimizes reliance on mapping accuracy. Here we present an algorithm for sensitive detection of indel candidates, based on local assembly where: 1) Paired reads are roughly mapped to targets with a fast aligner, assuming at least one member of the pair is correctly mapped; 2) For each target, the *k*-mer spectrum of the mapped read pairs is compared to the expected *k*-mer spectrum of the reference sequence; 3) All *k*-mers that are unique to the read set are assembled into contigs using a standard de-Bruijn assembly algorithm; 4) Assembled contigs are compared to the known reference sequence to generate mutation candidates.
A simulation study involving variable length insertions and deletions present at different sample fractions with a 250X mean coverage depth demonstrates the superiority of the assembly-based algorithm to a purely mapping-based approach. While the assembly-based approach correctly identified over 95% of deletions and insertions of length 3-25bp present in 10% or more of the sample, a purely mapping-based approach had far lower accuracy and in particular missed the vast majority of longer insertions. Additionally, application of this method to over 80 clinical lung and colon cancer specimens revealed several multi-nucleotide deletions and insertions in cancer-related genes, including an activating 9bp insertion in EGFR which was undetected by a purely mapping-based approach.

Efficient Algorithms for Tandem Copy Number Variation Reconstruction in Repeat-rich Regions

Dan He, Farhad Hormozdiari, Nicholas Furlotte, Eleazar Eskin

Department of Computer Science, University of California Los Angeles.

Motivation: Structural variations and in particular Copy Number Variations (CNV) have dramatic effects of disease and traits. Technologies for identifying CNVs have been an active area of research for over 10 years. The current generation of high-throughput sequencing techniques presents new opportunities for identification of CNVs. Methods that utilize these technologies map sequencing reads to a reference genome and look for signatures which might indicate the presence of a CNV. These methods work well when CNVs lie within unique genomic regions. However, the problem of CNV identification and reconstruction becomes much more challenging when CNVs are in repeat-rich regions, due to the multiple mapping positions of the reads.

Results: In this study, we propose an efficient algorithm to handle these multi-mapping reads such that the CNVs can be reconstructed with high accuracy even for repeat-rich regions. To our knowledge, this is the first attempt to both identify and reconstruct CNVs in repeat-rich regions. Our experiments show that our method is not only computationally efficient but also accurate.

Contact: eeskin@cs.ucla.edu

Partitioning the Scaffolding Problem into Small Independent Mixed Integer Programs

Leena Salmela, Veli Mäkinen, Esko Ukkonen, Niko Välimäki

Department of Computer Science, University of Helsinki

Assembling genomes from short read data has become increasingly popular but the problem remains computationally challenging especially for larger genomes. We study the scaffolding phase of sequence assembly where preassembled contigs are ordered based on mate pair data.

We present MIP Scaffolder which is based on dividing the scaffolding problem into smaller subproblems and solving these with mixed integer programming. We also present a method for filtering mate pair mappings so that only mappings that are consistent with nearby mappings are kept.

The scaffolding problem can be represented as a graph where the nodes represent contigs and edges distance constraints between contigs. Dayarian et al. (2010) showed that the biconnected components of this graph can be solved independently. We present a technique for restricting the size of these subproblems so that they can be solved accurately. We formulate the scaffolding problem as a mixed integer program and solve the subproblems of restricted size using a MIP solver.

Our experiments show that MIP Scaffolder is fast and scales to mammalian genomes. We show that when compared to state-of-the-art methods MIP Scaffolder achieves in most cases longer scaffolds with comparable accuracy.

References:

Dayarian, A., Michael, T. P., and Sengupta, A. M. (2010). SOPRA: scaffolding algorithm for paired reads via statistical optimization. BMC Bioinformatics, 11, 345.

JointSNVMix : A Probabilistic Model For Accurate Detection Of Somatic Mutations In Normal/Tumour Paired Sample Sequence Data

Andrew Roth[1], Jiarui Ding[1,3], Anamaria Crisan[1], Gavin Ha[1], Ryan Giuliany[1], Ali Bashashati[1], Sam Aparicio[1,2] and Sohrab Shah[1,2,3]

[1]Dept of Molecular Oncology, BC Cancer Agency, Vancouver, BC, Canada
[2]Dept of Pathology, University Of British Columbia, Vancouver, BC, Canada
[3]Dept of Computer Science, University Of British Columbia, Vancouver, BC, Canada

Motivation: Next generation sequencing (NGS) is playing an increasingly important role in understanding the biology of cancer. Among the most import problems that can be solved using genomic sequence data is the identification of single nucleotide somatic mutations in primary or metastatic tumours.

Most studies attempting to identify somatic mutations in tumours using NGS data also sequence a matched sample of healthy normal tissue to use as a control. Sequence data from tumour samples can then be compared against this background to identify somatic mutations while screening out germline polymorphisms - a significant source of false positive somatic mutation predictions.

A simple approach to finding somatic mutations involves analyzing data from the normal and tumour independently, then comparing the results. This approach ignores the significant correlation present between the two samples and results in lost signals due to premature thresholding. One consequence of this loss of shared signal is that a significant fraction of predicted somatic mutations are in fact germline polymorphisms.

Results: We have developed a novel probabilistic graphical model, JointSNVMix, to jointly analyze the normal and tumour samples. The key insight of the model is that we jointly model the emissions of allelic counts from both the normal and tumour samples. This allows to borrow statistical strength between samples and share weak signals to better identify sites which are likely to be germline polymorphisms. Accurate identification of germline polymorphisms in turn allows us to separate the germline and somatic positions and thus reduce the false positive rate for somatic mutation predictions.

We compare our joint modelling approach to four other methods which independently analyze the normal and tumour samples, combining predictions post-hoc to find somatic mutations. We analyzed paired normal/tumour sequence data obtained from four triple negative breast cancer patients. The data from these cases were sequenced using two methods. First we obtained exon capture (excap) data run on the Illumina GAII platform using the Agilent SureSelect kit. Second we obtained whole genome shotgun sequencing (WGSS) data run on the ABI SOLiD platform. In both cases the normal and tumour samples were sequenced to 30x coverage.

We observe that in all cases significantly fewer somatic mutations predicted by JointSNVMix are found to be in dbSNP. This suggests that JointSNVMix is better able to capture shared signal between the samples and thus falsely predicts fewer germline polymorphisms to be somatic mutations. Despite the reduction in false positive rate we observe a similar true positive rate on validated ground truth positions when compared to the best independent analysis methods.

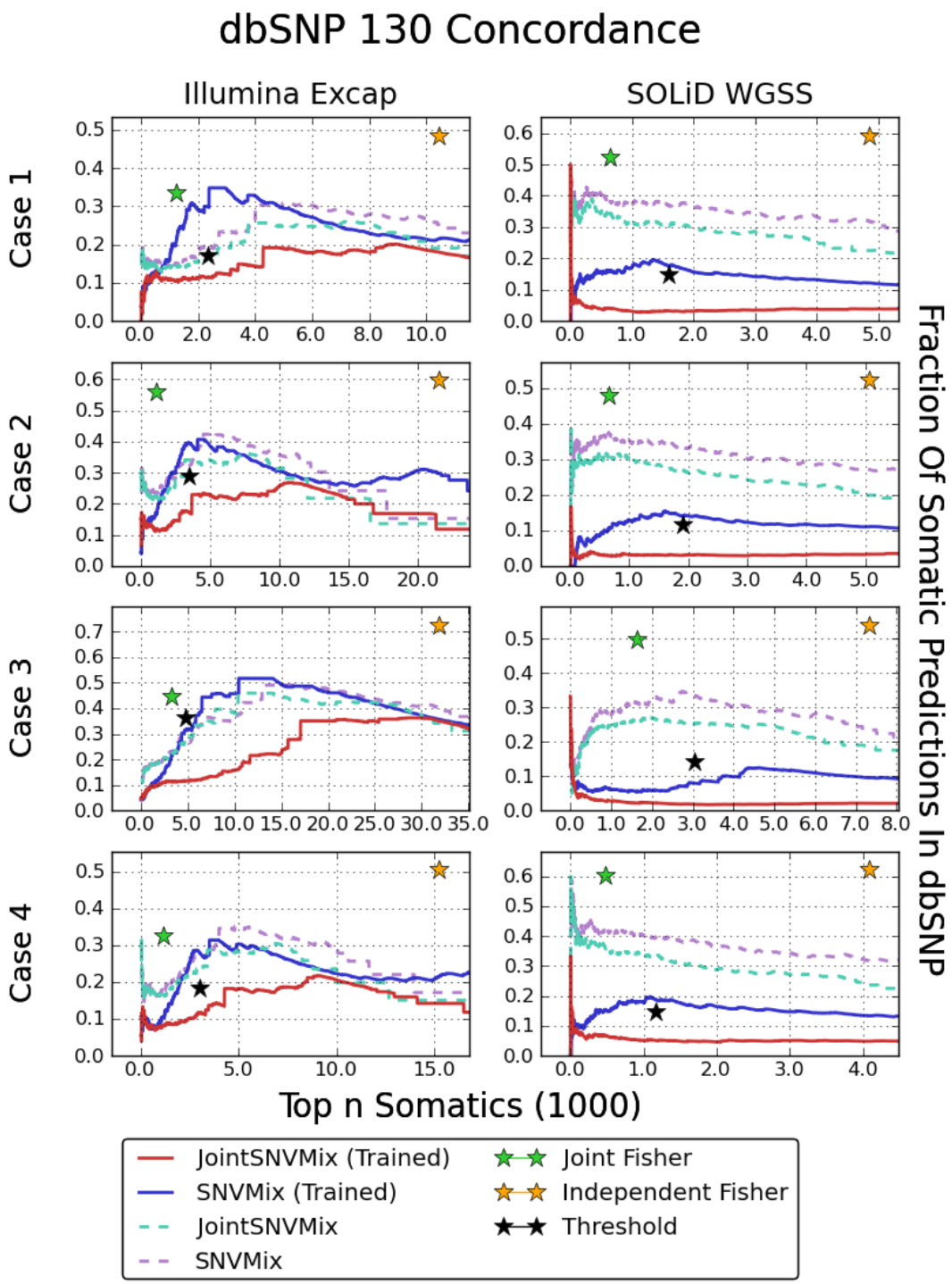


Figure 1: Fraction of somatic mutations predicted to be in dbSNP. On the horizontal axis we show number of somatic mutations made in 1000s. On the vertical axis we show the fraction of somatic mutations found to be ind dbSNP. For methods which assigned a score to their predictions we sorted the predictions by score and plotted a curve by

A Robust and Accurate Binning Algorithm for Metagenomic Sequences with Arbitrary Species Abundance Ratio

Henry C.M. Leung[1], S.M. Yiu[1], Bin Yang[2], Yu Peng[1], Yi Wang[1], Zhihua Liu[1], Jingchi Chen[2], Junjie Qin[3], Ruiqiang Li[3], Francis Y.L. Chin[1]*

1 Department of Computer Science, The University of Hong Kong, Hong Kong SAR, China
2 State Key Laboratory of Bioelectronics, Southeast University, Nanjing, China
3 BGI-Shenzhen, Shenzhen, China

Motivation: With the rapid development of next-generation se- quencing techniques, metagenomics, also known as environmental genomics, has emerged as an exciting research area which enables us to analyze the microbial environment in which we live. An important step for metagenomic data analysis is the identification and taxonomic characterization of DNA fragments (reads or contigs) resulting from sequencing a sample of mixed species. This step is usually referred to as “binning”. Binning algorithms that are based on sequence similarity and sequence composition markers rely heavily on the reference genomes of known microorganisms or phylogenetic markers. Due to the limited availability of reference genomes and the bias and low availability of markers, these algorithms may not be applicable in all cases. Unsupervised binning algorithms which can handle fragments from unknown species provide an alternative ap- proach. However, existing unsupervised binning algorithms only work on datasets either with balanced species abundance ratios or rather different abundance ratios, but not both.

Results: In this paper, we present MetaCluster 3.0, an integrated binning method based on the unsupervised top-down separation and bottom-up merging strategy, which can bin metagenomic frag ments of species with very balanced abundance ratios (say 1:1) to very different abundance ratios (e.g. 1:24) with consistently higher accuracy than existing methods.

Availability: MetaCluster 3.0 can be downloaded at http://i.cs.hku.hk/~alse/MetaCluster/

Contact: chin@cs.hku.hk

Improvements to GSNAP and development of the GSTRUCT pipeline for analyzing RNA-Seq data

Thomas D. Wu

Abstract: Mapping of short reads from transcriptional sources (RNA-Seq) poses numerous technical challenges, especially when reads span multiple exons or contain combinations of splicing, insertions, and deletions. In previous work, we have reported on our development of GSNAP to be able to quickly align reads containing multiple mismatches and either a single splice, insertion, or deletion. We now report on numerous substantive improvements in our GSNAP algorithm, many of which are intended to facilitate a larger pipeline that we have been developing called GSTRUCT for analyzing RNA-Seq data.

We have improved the ability of GSNAP to use known splicing, either as arbitrary combinations of individual splice sites or specified pairings of splice sites. GSNAP uses prefix tries to rapidly identify splices at the ends of reads, as well as double splicing in a single read. GSNAP will recognize splice ends that are ambiguous and can apply penalties to splice distances that are larger than observed. We have also analyzed different types of solutions for paired-end reads, and determined that the greatest opportunity for improvement involves halfmapping unique and concordant multiple alignments.

To handle halfmapping unique alignments, in which one end aligns uniquely to the genome while the other end does not align, we have integrated the alignment phase of our more general cDNA-genomic alignment program GMAP into GSNAP, which allows it to solve complex alignments involving arbitrary combinations of indels and multiple splicing. To resolve concordant multiple alignments, GSNAP can apply expression coverage information from a previous alignment run to identify the alignment with the greatest support.

GSTRUCT constitutes a pipeline for analyzing RNA-Seq data, that improves individual alignment results by computing over sets of alignments and can make biological inferences about alternate splice forms. In GSTRUCT, we apply GSNAP in an initial alignment run to accumulate evidence for expression coverage and splicing. We then use this expression coverage and splice information, which has been
filtered to remove false positives, to improve GSNAP results in a second alignment run.

Early versions of GSNAP and GSTRUCT have been tested several comparative settings, including the RGASP 2 and 3 contests, and have produced results superior to other programs tested.

Dissecting a massive parallel sequencing workflow for quantitative miRNA expression analysis

Francesca Cordero[1,2], Marco Beccuti [2], Maddalena Arigoni[1], Susanna Donatelli[2], Raffaele A Calogero[1]

1). Dipartimento di Informatica, Università di Torino, C.So Svizzera, 185, Torino 10149, Italy
2). Molecular Biotechnology Center, Università di Torino, Via Nizza 52, Torino, 10126, Italy

Email addresses:
FC: fcordero@di.unito.it
MB: beccuti@di.unito.it
MA: maddalena.arigoni@unito.it
SD: susi@di.unito.it
RAC: raffaele.calogero@unito.it

Abstract

Background

Next Generation Sequencing methods (MPS) can extend and improve the knowledge obtained by conventional microarray technology, both for mRNAs and short non-coding RNAs, e.g. microRNAs. The processing methods used to extract and interpret the information are an important aspect of dealing with the vast amounts of data generated from short read sequencing. The mapping, counting and characterization of the short sequence reads produced by Illumina GA (Genome Analyzer) and Applied Biosystems SOLiD technologies results in a bottle neck in data analysis. Although the number of computational tools for MPS data analysis is constantly growing, their strengths and weaknesses as part of a complex analytical pipe-line have not yet been well investigated.

Results

A benchmark MPS miRNA dataset, resembling a situation in which microRNAs are spiked in biological replication experiments was assembled by merging a publicly available MPS spike-in microRNAs data set with MPS data derived from healthy donor peripheral blood mononuclear cells. Using this data set, the strengths and weaknesses were highlighted of the major steps (i.e. reference sequence data set, alignment to reference and differential expression detection) in the MPS analysis workflow for detection of differential expression of microRNAs .

Conclusions

Since we have already demonstrated the efficacy of semi-synthetic datasets in defining the performances of workflow for high throughput transcription data, by dissecting an exon-level analysis workflow for Affymetrix 1.0 ST arrays [1], we applied a similar approach to the workflow for quantification of microRNAs digital MPS data. Our results indicate that the use of a focused reference data set, i.e. the miRbase microRNA precursor set, guarantees a better microRNA counts detection and requires limited computational resources with respect to the use of all genome. Furthermore, the selection of the alignment software is very important in maximizing the detection rate of microRNAs. From the five alignment tools we tested (Table 1), all specifically devoted to miRNA analysis or having specific setting for miRNA investigation, we got a wide range of detection performances. Our results clearly indicate that SHRiMP and MicroRazerS provide the best results. Concerning the statistical detection of differential expression of digital data we tested all the tools present in Bioconductor release 2.7, baySeq, DESeq, DEGseq, edgeR, RankProd, and we observed that the Negative Binomial model implemented in the baySeq

package was the one that gave the best results. On the basis of our results we can suggest that an optimized workflow for quantitative detection of microRNA differential expression based on digital sequence data need to implement the data and software elements described in Fig. 1.

Table 1: primary mapping tools used to map miRNAs

Name	Version	Reference set	Running time/sample	Spike-in detectio rate
SHRIMP	2.0.1	mir-set	96 sec	96%
miRExpress	2.0.1	mir-set	16 min	91%
miRProf	Web service	mir-set	-	46%
MicroRazerS	1.2	mir-set	14 min	96%
miRanalyzer	Web service	wg-set	-	73%

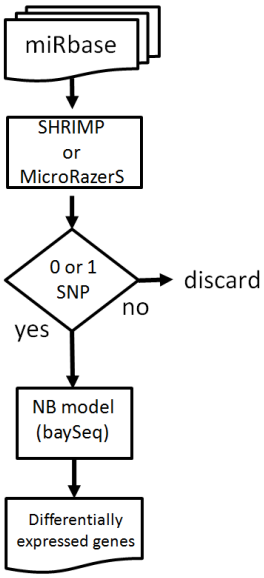


Fig. 1

References

1.Della Beffa C, Cordero F, Calogero RA: Dissecting an alternative splicing analysis workflow for GeneChip Exon 1.0 ST Affymetrix arrays. BMC Genomics 2008, 9:571.

Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*

J Behr[1], R M Clark[3], P Drewe[1], K Hildebrand[4], P Kover[5], R Lyngsoe[6], R Mott[6], E J Osborne[3], G Rätsch[1], S Schultheiss[1], V Sreedharan[1], J Steffen[3], O Stegle[2], C Toomajian[4], G Xiangchao[6]

[1]Friedrich Miescher Laboratory, Max Planck Society**,** Tübingen, 72076, Germany, [2]MPI for Developmental Biology & MPI for Intelligent Systems, Tübingen, 72076, Germany, [3]University of Utah, Biology, Salt Lake City, UT, 84127, [4]Kansas State University, Plant Pathology, Manhattan, KS, 66506, [5] University of Bath, Life Sciences, Manchester, M13 9PT, United Kingdom, [6]University of Oxford, WTCHG, Oxford, OX3 7BN, United Kingdom

We have sequenced the genomes of 18 inbred accessions of *Arabidopsis thaliana* at ~40x coverage using paired-end Illumina sequencing with different insert sizes. We developed an assembly pipeline that uses iterative read mapping and *de novo* assembly to accurately recover genome sequences with an error rate close to 1 in 10kb in single copy regions of the genome, and 1 in 1kb in repetitive or transposon rich loci, as assessed with independent data.

Naive projection of the coordinates of the 27,416 protein coding genes in the reference annotation onto the 18 genomes predicted large effect disruptions in 8,652 (32%), suggestion that *Arabidopsis thaliana* is able to survive disruptions in up to a third of its genes. To shed light onto this high number, we developed a novel pipeline for *de novo* annotation combining computational gene prediction and RNA-seq data from seedlings cultivated in highly controlled environments at ~20x coverage. Using this pipeline, we re-annotated each genome, finding that whilst there is considerable variation in gene structure, compensating changes help to ensure that altered transcripts can retain function. Thus 8,757 genes had at least one additional or modified transcript in at least one accession. In particular, for 8,322/8,757 (96.2%) genes harbouring large effect disruptions in at least one accession, the naively mapped transcripts were replaced by alternative transcripts. The effect of DNA sequence variation and altered gene models can be better understood when investigating the resulting protein sequence. Thus, we analysed how the transcripts' diversity affected their 40,578 inferred protein sequences, finding 3,840 (9.5%) proteins that had less than 50% amino-acid sequence identity with the corresponding TAIR10 proteins. Protein diversity varied across gene models and we found isoforms with severe disruptions to occur with generally low frequency in the accessions.

To complement the genotype-focused analysis, we investigated the quantitative transcriptome variation using the obtained RNA-seq reads. We found 20,963 (78%) of all protein genes to be expressed in at least one strain, with 9,360 (45%) exhibiting significant expression variation between strains. Mapping causal variants affecting gene expression, we identified variants associated with expression polymorphisms near 941 (10%) of differentially expressed genes. These candidate *cis*-eQTLs are tightly mapped, and analysis of the location of eQTLs relative to local gene models revealed an excess of associations in regulatory regions, including the core promoter region and 3'UTR.

This is one of the first studies where multiple genomes from a single species have been assembled, re-annotated and integrated with their transcriptomes to understand and quantify the regulatory role of natural DNA variation on gene structure as well as expression. It may serve as a blue print for forthcoming studies.

Auditor: exploring RNA-editing in cancer through simultaneous analysis of tumour genome and transcriptome sequencing data

Ryan Giuliany[1,3], Andrew Roth[1,3], Sam Aparicio[2,3] and Sohrab Shah[2,3]

1). Bioinformatics Training Program
2). UBC; Dept of Pathology
3). UBC; Dept of Molecular Oncology, BCCA

Motivation:
Post-transcriptional RNA-editing is a normal cellular process that modifies the nucleotide sequence of nascent RNA molecules, resulting in substitutions or indels when compared to the DNA sequence from which they were transcribed. If these edits occur in exons, then the resulting translated protein may, in some cases, have altered, or loss of function due to the change. Until very recently, post-transcriptional RNA-editing have not been considered as a potential oncogenic mechanism. While RNA-edits have been sporadically and anecdotally reported in cancer previously no comprehensive study has been undertaken, and the role of RNA-editing in tumour pathogenesis is unknown.

The advent of next-generation sequencing (NGS) technologies has made transcriptome-wide interrogation of RNA-editing possible. Investigating genome and transcriptome-wide occur-rences of single nucleotide variants (SNVs) using RNA-Seq and whole genome shotgun data obtained from the same tumour have been undertaken and reported over the last several years thus presenting the unprecedented opportunity to simultaneously examine the genome and transcriptome sequence of a tumour.

A simple approach to identify RNA-edits would be to call the SNVs in the transcriptome and genome independently and nominate SNVs that are unique to the transcriptome. We argue in this contribution that this approach suffers from 2 key limitations: i) since the genome and transcriptome from the same tumour are highly correlated, independent analysis is unable to leverage shared signals between them and ii) most SNV callers collapse the allelic state space to a binary representation and therefore non-uniform RNA-editing specific substitution distributions between nucleotides (i.e. A-to-G(I) edits) cannot be modeled in these analyses.

Auditor: model construction, learning and inference (Fig. 1, top)
To address these limitations, we have developed Auditor, a generative probabilistic model based on hierarchical Bayes that assumes observed transcriptome and genome data are generated by two distinct processes: regular transcription and RNA-editing. As such we encode observed nucleotide substitutions between the genome and transcriptome with distinct ‘regular’ and ‘RNA-editing’ substitution matrices.

We represent the status of editing (on or off) with a Bernoulli random variable , regularized by a Beta prior, . Additionally, we assume the observed digital allelic counts of a genome at a given position , denoted , can be represented with eleven possible ‘states’: and similarly for the transcriptome: where the state space represents all true biological states, and an extra “garbage” state, ZZ, for capturing low quality data expected to be enriched for sequencing errors.

Thus, the transition matrices represent the probability of a sample being genotype and transcriptotype for position given editing status . Intuitively, one expects no change

from the genome state to transcriptome state unless editing is present. Therefore we assume one matrix will contain probability mass highly concentrated on the diagonal, while the other matrix has probability mass localized at the most common classes of edit, such as A-to-G (Fig. 1, bottom). [Note that our model allows for inference of allele-specific expression as well, but that is beyond the scope of this contribution].

Given the genotype, we then assume the observed allelic counts at any position in the genome are distributed according to a Multinomial distribution with a state-specific conjugate Dirichlet prior. By evaluating the multinomial probability density function of each , for a set of base counts we obtain the likelihood of the base counts having been generated from that genotype. Additionally, the likelihood of a given genotype is regularized by the overall frequencies, . We employ an analogous approach for the states in , but with transcriptome-specific Dirichlet distributions, allowing the flexibility of modeling the genome and transcriptome patterns independently, and with frequencies instead encoded in .

In order to estimate the parameters of the model and infer editing status, we fit the model to data using expectation maximization (EM). Thus, given the observed allelic counts derived from aligned genome and transcriptome NGS data from the same sample, the Auditor model simultaneously infers the RNA-editing status at each position and the parameters of the distributions described above.

Evaluation and Results

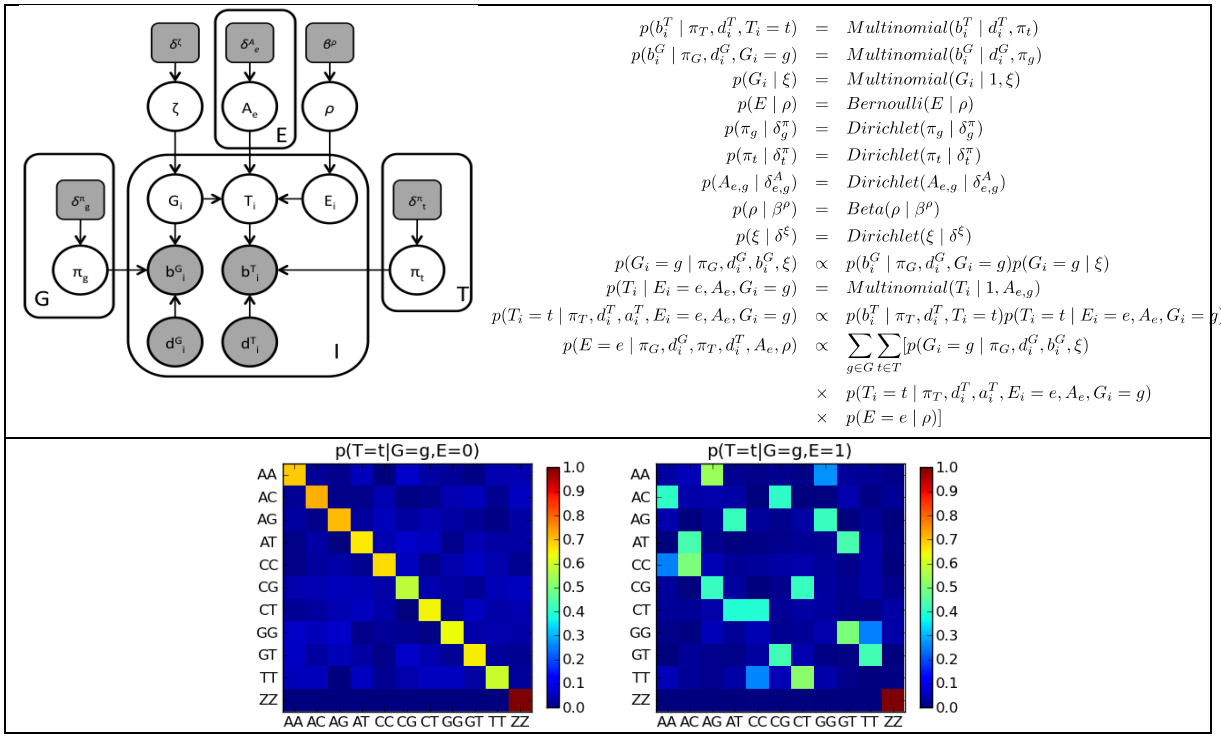
In order to evaluate the performance of Auditor, we simulated two data sets: one of high variance that exaggerates sequencing noise, and low variance that more accurately represents the sequencing error rate. Auditor achieves AUCs of 0.989 and 0.917 on low and high variance data, respectively, for predicting editing status.

For purposes of comparison, we ran SNVMix[1], independently on the simulated DNA and RNA data, and selected positions called as SNVs in the RNA and wildtype in the DNA as RNA-edits. Additionally, we compared the performance of Auditor to that of SNVMix using a validated set of RNA-edits from a previously published lobular breast cancer genome and transcriptome[2]. In all cases, Auditor achieves an increase in specificity with no sacrifice in sensitivity (see Table 1).

Finally, we will present the results of applying Auditor to a novel data set of 25 genome/RNA-seq sequence pairs from breast and ovarian cancers, and show the predicted landscape of RNA-editing across these tumors.

Conclusion

RNA-editing represents an under-explored cellular process that could have significant effects in tumorigenesis and tumor progression. We have developed Auditor, a tool that effectively predicts RNA-edits in paired genome/RNA-seq samples and have begun the process of determining the landscape of RNA-editing in breast and ovarian cancer.



$$
\begin{aligned}
p(b_i^T \mid \pi_T, d_i^T, T_i = t) &= Multinomial(b_i^T \mid d_i^T, \pi_t) \\
p(b_i^G \mid \pi_G, d_i^G, G_i = g) &= Multinomial(b_i^G \mid d_i^G, \pi_g) \\
p(G_i \mid \xi) &= Multinomial(G_i \mid 1, \xi) \\
p(E \mid \rho) &= Bernoulli(E \mid \rho) \\
p(\pi_g \mid \delta_g^\pi) &= Dirichlet(\pi_g \mid \delta_g^\pi) \\
p(\pi_t \mid \delta_t^\pi) &= Dirichlet(\pi_t \mid \delta_t^\pi) \\
p(A_{e,g} \mid \delta_{e,g}^A) &= Dirichlet(A_{e,g} \mid \delta_{e,g}^A) \\
p(\rho \mid \beta^\rho) &= Beta(\rho \mid \beta^\rho) \\
p(\xi \mid \delta^\xi) &= Dirichlet(\xi \mid \delta^\xi) \\
p(G_i = g \mid \pi_G, d_i^G, b_i^G, \xi) &\propto p(b_i^G \mid \pi_G, d_i^G, G_i = g) p(G_i = g \mid \xi) \\
p(T_i \mid E_i = e, A_e, G_i = g) &= Multinomial(T_i \mid 1, A_{e,g}) \\
p(T_i = t \mid \pi_T, d_i^T, a_i^T, E_i = e, A_e, G_i = g) &\propto p(b_i^T \mid \pi_T, d_i^T, T_i = t) p(T_i = t \mid E_i = e, A_e, G_i = g) \\
p(E = e \mid \pi_G, d_i^G, \pi_T, d_i^T, A_e, \rho) &\propto \sum_{g \in G} \sum_{t \in T} [p(G_i = g \mid \pi_G, d_i^G, b_i^G, \xi) \\
&\quad \times p(T_i = t \mid \pi_T, d_i^T, a_i^T, E_i = e, A_e, G_i = g) \\
&\quad \times p(E = e \mid \rho)]
\end{aligned}
$$

Fig. 1 Probabilistic graphical model representation of Auditor (top, left) and the corresponding conditional probability distributions of all random variables (top, right). Edit 'off' (bottom, left) and edit 'on' (bottom, right) substitution matrices. The 'on' matrix represents a hypothetical edit distribution.

Sample	Auditor		SNVMix	
	Sensitivity	Specificity	Sensitivity	Specificity
Low Var. Sim.	0.896	0.953	0.881	0.768
High Var. Sim.	0.798	0.921	0.768	0.701
Lobular B.C.	1.00	0.412	1.00	0.301

Table 1. Comparison of Auditor and SNVMix on three datasets.
[1]Goya *et al.* Bioinformatics. (2010) 26 (6):730-736; [2]Shah *et al.* Nature (2009) 461:809-813

Comrad: a novel algorithmic framework for the integrated analysis of RNA-Seq and WGSS data

Andrew McPherson[1,2,4], Chunxiao Wu[3,4], Iman Hajirasouliha[1,4], Fereydoun Hormozdiari[1,4], Faraz Hach[1], Anna Lapuk[3], Stanislav Volik[3], Sohrab Shah[2], Colin Collins[3,*] and S. Cenk Sahinalp[1,*]

[1]Department of Computing Science, Simon Fraser University, 8888 University Way, Burnaby, BC, V5A 1S6, Canada
[2]BC Cancer Agency, 600th Avenue West, Vancouver, BC, V5Z 4E6, Canada
[3]Vancouver Prostate Centre, 899 12th Avenue West, Vancouver, BC V5Z 1M9, Canada

Motivation: Comrad is a novel algorithmic framework for the integrated analysis of RNA-Seq and Whole Genome Shotgun Sequencing (WGSS) data for the purposes of discovering genomic rearrangements and aberrant transcripts. The Comrad framework leverages the advantages of both RNA-Seq and WGSS data, providing accurate classification of rearrangements as expressed or not expressed and accurate classification of the genomic or non-genomic origin of aberrant transcripts. A major benefit of Comrad is its ability to accurately identify aberrant transcripts and associated rearrangements using low coverage genome data. As a result, a Comrad analysis can be performed at a cost comparable to that of two RNA-Seq experiments, significantly lower than an analysis requiring high coverage genome data.
Results: We have applied Comrad to the discovery of gene fusions and read-throughs in prostate cancer cell line C4-2, a derivative of the LNCaP cell line with androgen-independent characteristics. As a proof of concept we have rediscovered in the C4-2 data 4 of the 6 fusions previously identified in LNCaP. We also identified 6 novel fusion transcripts and associated genomic breakpoints, and verified their existence in LNCaP, suggesting that Comrad may be more sensitive than previous methods that have been applied to fusion discovery in LNCaP. We show that many of the gene fusions discovered using Comrad would be difficult to identify using currently available techniques.
Availability: A C++ and Perl implementation of the method demonstrated in this paper is available at http://compbio.cs.sfu.ca/
Contact: andrew.mcpherson@gmail.com

Community-Analyzer: A visualization and analysis tool to study the structure and dynamics of microbial communities across metagenomic data sets

Bhusan K Kuntal, Tarini Shankar Ghosh and Sharmila S Mande*

Bio-sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services Limited
1 Software Units Layout, Madhapur, Hyderabad – 500083, Andhra Pradesh, INDIA

*Corresponding author: sharmila@atc.tcs.com

The advent of high throughput sequencing technologies and the concurrent emergence of the field of metagenomics has facilitated the rapid extraction and sequencing of the microbial genomic content isolated from the vast and hitherto unexplored biomes round the world. Simultaneously, several computational analysis methods/softwares have been developed that facilitate the rapid analysis of such huge sequence data sets. Using these computational methods, researchers can not only characterize the environment(s) of their interest in functional and taxonomic terms, but can also perform a thorough comparative analysis across several environments under study. Such comparative analyses of the microbial groups present in different environments (or similar environments but exhibiting distinct phenotypic traits) have also led to the identification of key agents probably responsible for conferring a specific phenotypic characteristic to a given environment. These phenotypic traits may include specific disease conditions or a given physiological disorders like obesity.

The currently available comparative metagenomics tools can identify micobial groups that are selectively over- or under-abundant in a given environment and also group metagenomes based on their taxonomic profiles. However, they do not provide any insights as to how the over-abundance or under-abundance of a specific group of organisms is influenced by other co-inhabiting microbial groups. Given the multitude of organisms residing in a given environment, it is expected that a specific phenotypic characteristic may not always be due to the specific presence or absence of a few organisms, but may be the result of the inter-microbial interactions observed between the resident taxonomic groups. In spite of the critical role played by these inter-microbial interactions (in determining the functional and phenotypic traits of the environment), currently no software application is available that, besides performing standard comparative metagenomics analysis, can also provide insights into the interaction dynamics occurring in the given environment.

To address the above issue, we have developed a comparative metagenomic analysis platform called Community-Analyzer. The tool can analyze a given group of metagenomic datasets and identify groups of taxa showing a positive or negative correlation in their occurences. Identification of such groups of taxa can provide insights into the inherent interaction dynamics observed within the analyzed microbial communities. Subsequently, it uses the above information to compare and group metagenomic samples.

The tool generates an interactive graphical layout that can be used to study the community structure of the analyzed metagenomic samples as well as interactively visualize the quantitative and qualitative abundance of the various microbial communities. Besides, the tool also generates statisitical significance matrices, using which the user can judge the (statistical) strength of the association (or inhibition) between the resident taxonomic

groups. In addition, the tool also integrates several standard features of comparative metagenomics analysis (e.g. stacked bar plots, trend plots, etc.). The software is expected to have immense applicability for metagenomics researchers working in diverse areas like healthcare, environmental and industrial biotechnology. Details of the platform will be presented during the conference.

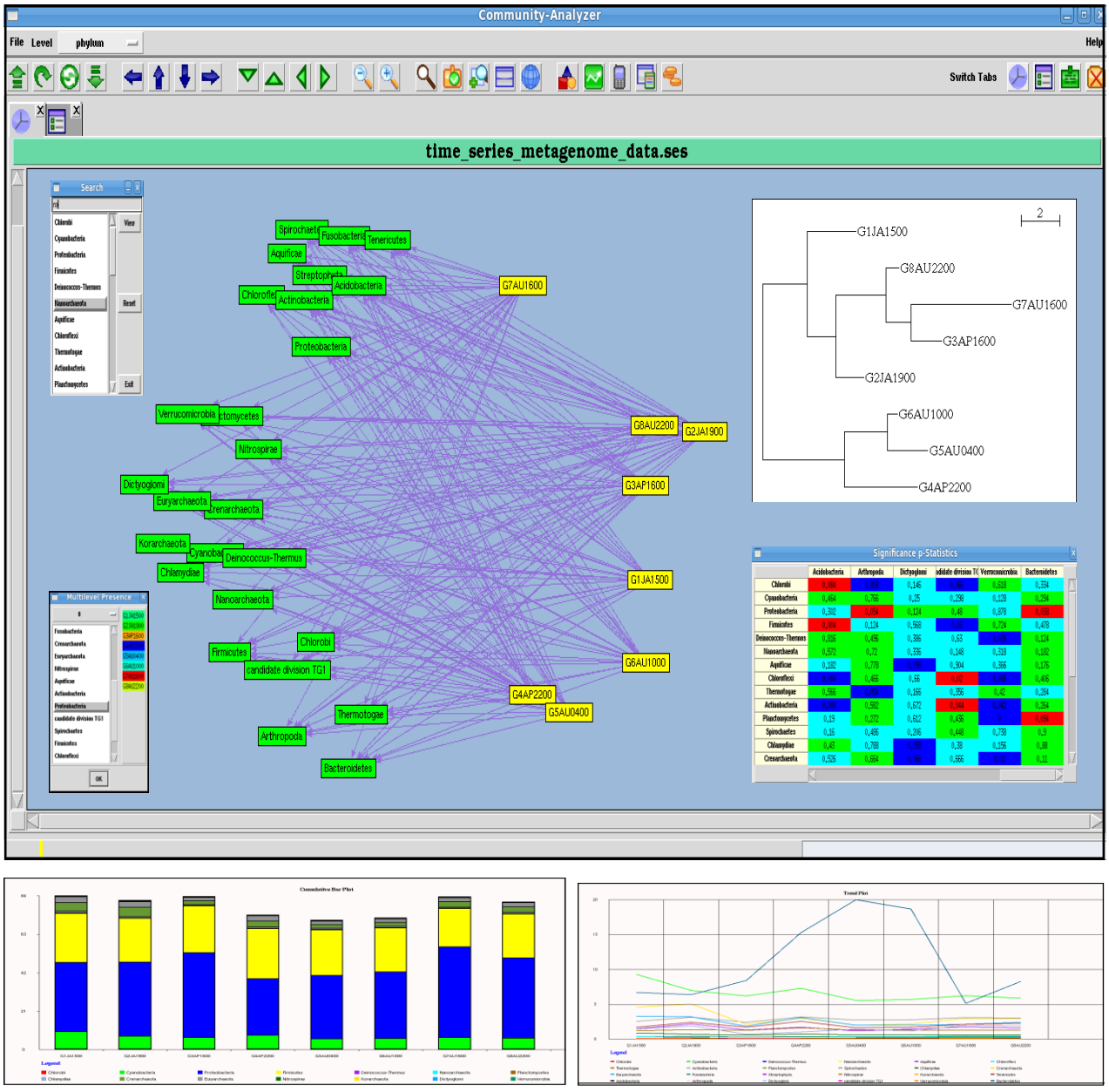


Figure illustrating the range of possible outputs and the different analyses that can be peformed using the Community-Analyzer.

Copy number aware Bayesian approach to detect loss of heterozygosity in tumour genome sequence data

Gavin Ha[1,2], **Samuel Aparicio**[1,2], **Sohrab Shah**[1,2]

1). Department of Molecular Oncology, British Columbia Cancer Agency, Vancouver, Canada
2). Department of Pathology, University of British Columbia, Vancouver, Canada

1. Background

Loss of heterozygosity (LOH) and copy number alteration (CNA) are genomic aberrations that can contribute to tumourigenesis. LOH is the change of genotype from heterozygous to homozygous such that chromosomal regions in the DNA have only one allele present. In breast cancer, LOH has been extensively studied whereby genomic regions involving tumour suppressor genes such as p53, BRCA1, and BRCA2 are affected. LOH can be observed under different copy number changes in cancer cells relative to normal tissue of the same individual. A region of hemizygous deletion in which one copy (i.e. allele) is deleted leaves only a single copy of the region in the cell. Secondary amplification events that duplicate the remaining copy can return the region to two copies (copy neutral) or result in an arbitrary number of copies of the same allele. However, in regions containing both alleles (i.e. heterozygous), events such as mono-allelic copy number amplification can be mistakenly predicted as LOH due to the dominating signal of the amplified allele.

High-throughput sequencing (HTS) technology has enabled the detection of tumour genome mutation events at single base pair resolution; however, methods for predicting regions of allelic imbalance and LOH are yet to be fully explored. Presently, we are not aware of a spatially correlated probabilistic framework that also considers copy number alterations (CNAs) to improve LOH predictions in tumour HTS data.

2. Methodology

We have developed a non-stationary Bayesian hidden Markov model, APOLLOH, which infers the zygosity status of spatial DNA regions in tumour genome sequence data. Our approach draws motivation from existing SNP genotyping array methods used for analyzing allele-specific information; however, it is tailored to HTS whole genome sequence data. APOLLOH infers spatially correlated genotypes from aligned tumour sequence reads at varying levels of copy number to reveal contiguous regions of LOH, allele-specific copy number amplification (ASCNA), and heterozygous (HET) status. The model is informed by copy number status in order to properly distinguish LOH and ASCNA regions. The analysis focuses on tumour loci that have heterozygous genotype in the matched normal sample, helping to reduce the inference problem to a tractable dimensionality. Model parameters are trained independently for each sample using the expectation maximization algorithm to improve modeling of sample-specific signals induced by biological or technical variation.

3. Results

We applied our method to 16 breast and 9 ovarian complete cancer genomes sequenced to 30x coverage, each with matched normal genomes sequenced to the same coverage. Genome-wide landscapes of zygosity profiles were generated, providing a detailed view of allelic imbalance and LOH. Each sample was also analyzed using orthologous genome-wide measurement of LOH by high-density SNP genotyping arrays. We demonstrate that APOLLOH prediction results are comparable to the genotying arrays (Figure 1) while also

providing greater, un-biased coverage of the genome and more precisely defined breakpoints. We also compare against predictions made by another model that assumes identical and independently distributed signals by loci, and clearly demonstrate the benefits of modeling spatial correlation with the HMM.

APOLLOH is a tool that provides the ability to survey the complete landscape of allelic imbalance in tumour genome sequence data and will aid in discovering and understanding novel tumour suppressor genes from inferred LOH profiles.

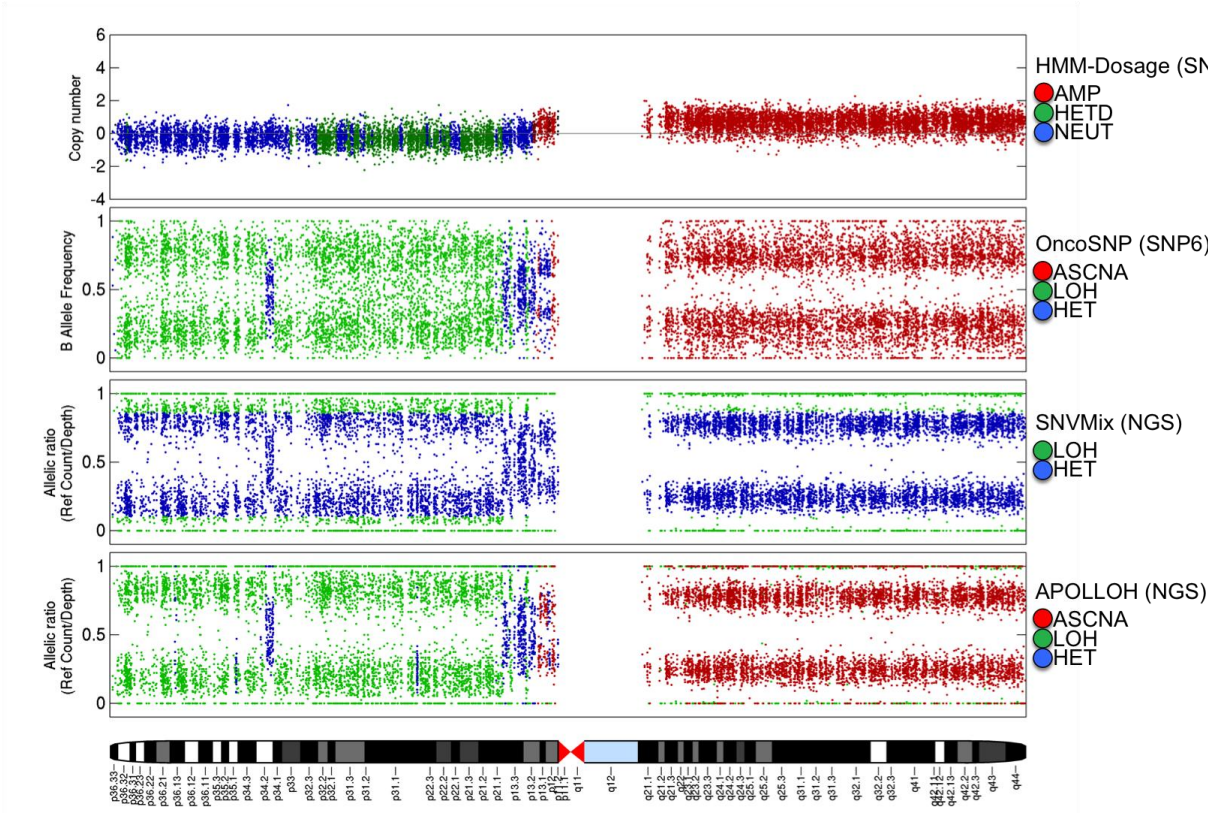


Figure 1. Comparison of allelic imbalance and LOH for chromosome 1 of a breast cancer genome[1]. Panel 1 is the copy number predicted by HMM-Dosage on Affymetrix SNP6 array data. Panel 2 shows the LOH predictions made by OncoSNP[2] on SNP6 array data. Panel 3 shows the results from the independent model, SNVMix[3], on HTS data. Panel 4 shows the results of APOLLOH on HTS data. For comparison between each panel, data points shown are only for heterozygous positions determined from the SNP6 array of the match normal. The results show good concordance with OncoSNP while clearly outperforming the iid model of SNVMix.

4. References

1. Shah, S. P. et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature 461, 809–13 (2009).
2. Yau, C. et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome Biol 11 (2010).
3. Goya, R. et al. Snvmix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 26, 730–736 (2010).

Variant detection and the Autism Sequencing Project

Orion Buske[1], Misko Dzamba[1], Justin Foong[2], Lynette Lau[2], Marc Fiume[1], Christian Marshall[2], Susan Walker[2], Aparna Prasad[2], Stephen Scherer[2], Michael Brudno[1,2,3]

1). Department of Computer Science, University of Toronto, Toronto, Canada
2). The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada
3). Donnelly Centre, University of Toronto, Toronto, Canada

Early detection of Autism Spectrum Disorders (ASD) can improve the quality of life of affected individuals. Qualitative screening methods continue to improve, but still suffer from low sensitivity despite increasing specificity. We are sequencing the exomes of 1,000 individuals with ASD in order to discover genetic variants associated with the disorders.

Sequencing is done on the AB SOLiD4 platform after enrichment with Agilent SureSelect exome capture kits. Sequence reads are then aligned to the human reference (GRCh37) using SHRiMP, a color-space-aware read aligner. To reduce PCR artifacts, duplicate reads are removed using Picard. SRMA is used to locally realign reads in order to refine the alignment by sharing information across reads. Single Nucleotide Variants (SNVs) are then identified from these aligned reads using the Genome Analysis Toolkit (GATK), refined and phased with BEAGLE, annotated with harmfulness predictions from SIFT, and finally filtered with strict thresholds to limit the false positive rate. To consider an SNV high confidence, we require 30 reads to support the call, with three reads on each strand and 15% of the total reads supporting the alternate allele.

SHRiMP was chosen as the read aligner after comparing its performance against BFAST, another popular color-space-aware read aligner, on preliminary exome sequence data. When benchmarked on 45 million read-pairs of SOLiD sequence, SHRiMP aligned the reads in less than half the time of BFAST. Reads were aligned across 5 parallel processes (10 million read-pairs each), taking 97 node-hours with SHRiMP and 208 node-hours with BFAST on Intel Xeon 3.0 GHz processors. After removing duplicate reads, SHRiMP aligned an average of 7% more reads to the genome than BFAST, of which 56% were aligned to targeted exome regions. The quality of the resulting alignments was assessed by measuring the concordance of GATK genotype calls with Illumina 1M array genotypes. SHRiMP alignments resulted in greater concordance with array genotypes, recovering an average of 56% of non-reference array sites with 89% precision.

For the 72 individuals sequenced so far, a median of 6.4 billion bases were sequenced per individual from a median of 151 million reads. On average, 68% of these reads were successfully aligned to the reference genome and 30% of those were removed as duplicates. The remaining 68 million reads per individual provide approximately 30x mean coverage over the target regions.

On these data, GATK calls an average of 65,307 SNVs per individual. Strict filtering reduces these to an average of 5,177 high-confidence SNVs per individual, of which 1,950 are non-synonymous coding mutations, 65 are also novel to the 1,000 Genomes Project and dbSNP 132, and 29 of those are private to a single ASD individual. The median concordance of high-confidence SNVs with array genotypes is 98.7%, and the transition/transversion ratios (Ti/Tv) for private and known SNVs are 2.59 and 3.01, respectively, suggesting high accuracy in this tier of SNV calls. Across all 72 individuals currently sequenced, we identify 2,065 novel non-synonymous SNVs that occur in only

one ASD individual, of which SIFT predicts 652 to be significantly deleterious to protein function.

Additionally, 11 of these private deleterious SNVs occur within 250 genes implicated in ASD and other related neurological disorders, along with two nonsense mutations, one in NRXN1 and one in CEP290. We validated the CEP290 mutation by Sanger sequencing, and validations of other potentially deleterious private variants are currently underway.

Sequencing error correction to reliably measure diversity of the human T cell receptor repertoire

Pina F. I. Krell[*,1,2], Susanne Reuther[1], Michael Gombert[1], Arndt Borkhardt[1], Jens Stoye[2]

1). University Dusseldorf, Medical Faculty, Department of Pediatric Hematology, Oncology, and Immunology, D-40225, Dusseldorf, Germany
2). Genome Informatics, Faculty of Technology, Bielefeld University, Germany
* Corresponding author, contact: pkrell@cebitec.uni-bielefeld.de

Immunological Background

T cell receptor (TR) diversity is characterized through somatic alterations in the complementary determining region three (CDR3) of the human T cell receptor beta chain. Essential to combat attacks of the vast molecular variability of pathogenic microorganisms, diversity in the repertoire of TR beta chains is usually interpreted as the immunological ability of the immune system to defend itself, making TR diversity estimation a major interest in a broad area of clinical and basic immunological questions.

Somatic alterations in the CDR3 are generated through cell-specific, irreversible germline rearrangement, V-D-J recombination, which makes use of three different gene sets encoded in the TR beta chain locus (TRB). Assembling one gene from the variable (TRBV), joining (TRBJ), and diversity (TRBD) gene sets, V-D-J recombination builds the molecular key of antibody recognition, the CDR3, which spans the junction of the three genes. In addition, a process of random, template-independent nucleotide addition and deletion (N-diversity) between the rearranged genes further enhances diversity of the CDR3. Complemented with the TR alpha chain, built in a similar rearrangement process, TR diversity can hypothetically result in $>10^{18}$ different TR molecules [1]. 2×10^7 T cells with unique CDR3s are thought to actually reside in the lymphoid organs and circulation of the human blood [2]. This still extraordinary diversity for long has challenged in depth sequence analysis of TR repertoires so that only small parts of TR diversity were actually captured by sequence analysis.

Measuring TR Diversity using HTS Methods

Emergence of second generation, high-throughput sequencing (HTS) technologies with major capacity progress has made sequencing a useful approach to estimate TR diversity. Despite this improvement, raw data obtained by those technologies are usually error prone [3] and generate new challenges to the reliable estimation of TR repertoire diversity. Sequencing errors, even single nucleotide substitutions, in the CDR3 could have major impact on CDR3 diversity estimation.

However, high-throughput sequencing systems deliver sequencing error estimators by generating quality values for each sequenced nucleotide of a sequence [4]. Using a threshold quality value to filter sequences with inferior quality has been shown to improve overall reliability of sequencing data, but still lacks the ability to distinguish true, unique sequences (clonotypes) which may be rarely represented in a sample from sequences artificially generated through single nucleotide sequencing errors. Especially rare TR variants, being present in low abundance with CDR3 regions that differ only slightly from other variants in the repertoire, are hard to identify as true receptor variants. In addition non-productive TR rearrangements, containing stop codons or out of frame CDR3 would not be apparent but are real and important to the measure of TR diversity.

Quality value-based correction of sequencing errors

We developed a tool able to profile T cell receptor beta chain diversity of 454 sequenced TR repertoires. Other than previous tools, it is able to handle sequencing errors by using system specific quality values not only to pre-process data, but also during alignment-based identification of rearranged TRBV, TRBJ, and TRBD genes as well as CDR3 discrimination. To ensure high-quality results we pre-process raw sequencing data by sliding through a window of bases at the 3' end of the sequence, discarding parts of the sequence that lack a certain average quality threshold, thus removing sequence ends where the sequencing reaction phases out.

Rearranged genes are identified using local alignment incorporating quality values to improve accuracy. Sequence comparison is performed against the known germline genes of the IMGT/GENE-DB reference directory set of the human beta chain germline genes [5]. CDR3 regions, characterized to be flanked by specific amino acid sequence motifs are identified by a quality value-improved alignment-based search in all three reading frames. Sequence-based analysis is further extended by performing hypothetical *in vivo* function prediction, and combinatorial analysis to reflect repertoire diversity in statistical profiles on basis of TRBV, TRBD, and TRBJ segment occurrence frequency, combination frequencies, clonotype count, and CDR3 length polymorphism due to N-diversity. In addition for predictive value in clinical diversity estimation applications, results can be visualized with an independent visualization tool, reflecting CDR3 length polymorphism.

Whereas the rearranged gene segments can be identified using commonly known sequence comparison algorithms, with the lack of a reference sequence the highly variable CDR3, which is of high scientific and clinical interest, demands improvement in bioinformatics TR repertoire diversity estimation. Observing gene specific sequencing errors, occurrence of sequencing errors in the TRBV, TRBJ, and TRBC gene can be used to build an error metric that estimates the sequencing error distribution in a specific TR beta chain sequence. With the additional knowledge that the sequencing reaction usually tends to reach the best quality in the middle of a sequence while the ends accumulate sequencing errors, these criteria can be used to filter sequences with reliably sequenced CDR3.

Our new tool thus allows to identify sequencing error frequencies for each gene in particular and to filter sequences with highly reliable CDR3 sequence. Additionally, error corrected identification of rearranged genes not only allows to build reliable TR diversity measures on the basis of DNA or amino acid CDR3 sequence, but in addition on gene frequencies and gene combination frequencies. This reflects diversity not only as the final product but also exhibits the process a specific immune system is capable of.

References:

[1] Janeway, C.A. (2005) *Immunobiology*, Garland Science, New York.
[2] Arstila, T.P. *et al.* (1999) A direct estimate of the human alphabeta T cell receptor diversity. *Science*, **286**:958-961
[3] Huse, S.M. *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**, R134,doi:10.1186/gb-2007/8/7/R143
[4] Ewing, B., and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using Phred II Error Probabilities. *Genome Res.*, **8**, 186-194.
[5] Giudicelli et al. (2005) IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.*, **34**: D781-4.

Identity-By-Descent Filtering of Exome Sequence data for Disease-Gene Identification in Autosomal Recessive Disorders

Christian Rödelsperger[1,2,3,*], Peter Krawitz[1,2,3,*], Sebastian Bauer[2,*], Jochen Hecht[1,2,3], Abigail W. Bigham[4], Michael Bamshad[4], Birgit Jonske de Condor[1], Michal Schweiger[3] and Peter Robinson[1,2,3,†]

[1]Institute for Medical Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany
[2]Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin, Berlin, Germany
[3]Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany
[4]Department of Pediatrics, University of Washington, Seattle, Washington 98195, USA

Motivation: Next-Generation Sequencing (NGS) and exome-capture technologies are currently revolutionizing the way geneticists screen for disease-causing mutations in rare Mendelian disorders. However, the identification of causal mutations is challenging due to the sheer number of variants that are identified in individual exomes. Although databases such as dbSNP or HapMap can be used to reduce the plethora of candidate genes by filtering out common variants, the remaining set of genes still remains on the order of dozens.
Results: Our algorithm uses a non-homogeneous hidden Markov model (HMM) that employs local recombination rates to identify chromosomal regions that are identical by descent (IBD=2) in children of consanguineous or non-consanguineous parents solely based on genotype data of siblings derived from high-throughput sequencing platforms. Using simulated and real exome sequence data, we show that our algorithm is able to reduce the search space for the causative disease gene to a fifth or a tenth of the entire exome.
Availability: The source code and tutorial are available at http://compbio.charite.de/index.php/ibd2.html.
Contact: peter.robinson@charite.de

MedSavant: a platform for identifying causal variants from disease sequencing studies

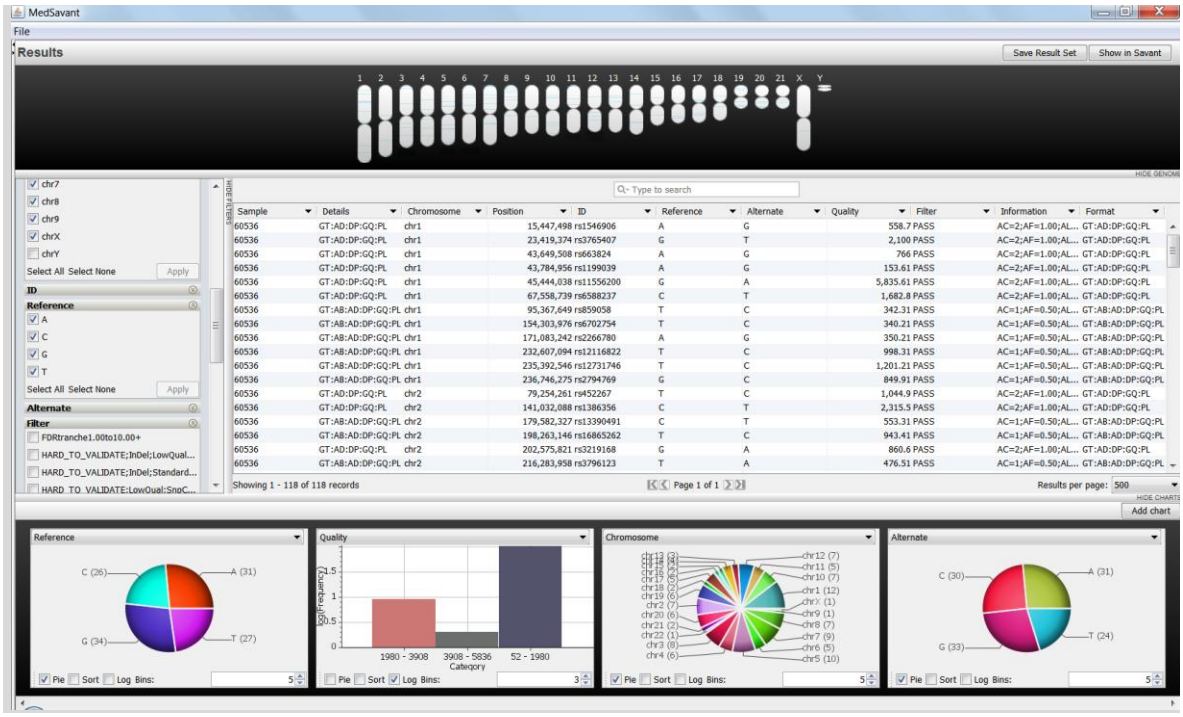
Marc Fiume[1], Nirvana Nursimulu[1], Justin Foong[3], Margie Manker[3], Michael Brudno[1,2,3]

1). Department of Computer Science
2). Donnelly Centre and Banting and Best Department of Medical Research, University of Toronto, Ontario
3). The Center for Applied Genomics, Toronto, Ontario

High Throughput Sequencing (HTS) technologies have revolutionized the speed and economy with which genomic information can be obtained, and are providing a means for deep cataloguing of human variation. The 1000 Genomes Project[1], for example, is using HTS machines to sequence and subsequently genotype a large number of individuals from a wide spectrum of ethnic backgrounds. The catalogue of putative variants obtained from this and similar worldwide studies can be used as a basis of comparison for other studies that use HTS machines to sequence and genotype cohorts of individuals with disease[2,3], with the aim of discovering disease etiology and improving diagnostics. However, many of the candidate variants found in such studies are either not real (due to errors in the prediction pipeline) or are real but have no functional relevance. One of the most challenging problems is thus in identifying those few genetic variants (among potentially millions that are predicted per sequenced individual) that are actually causal in disease.

For this purpose we introduce MedSavant: a software platform for accelerating the identification of disease-causing genetic variants found in population sequencing studies by enabling complex and dynamic querying of patient data. The platform is comprised of two parts: a graphical interface and a backend database. The database is designed to securely store patient data across three main axes: (1) basic patient data: e.g. age, sex, and pedigree (2) phenotype data: e.g. disease, signs, and symptoms (3) genotype data: e.g. candidate variants, their types, and genomic locations. It is being engineered to handle huge volumes of data that can be updated frequently while still being efficiently searchable. The client-side interface enables users to dynamically visualize global trends in the data, construct complex queries, and analyze the results. The framework allows one to easily design a complex query that returns, for example, only variants found in male patients (filtering on basic data) in a specific disease cohort (filtering on phenotype data) who share a rare point mutation with high predicted confidence (filtering on genotype data). MedSavant also supports filters that are generated from external datasources, such as whether or not the variation has been discovered before (using dbSNP[4] data), is predicted to be damaging (using SIFT[5] or Polyphen[6] annotations), is found in genes having a pertinent function (using the GO ontology[7]), or has been associated with a related disease (using OMIM[8] data). Furthermore, MedSavant can be integrated with the Savant Genome Browser[9], for manual inspection of the read alignment data supporting the most likely causal variants found in the filtration process. One could use this additional information to confirm the validity of variant predictions and to design region-specific primers for wetlab validation, for example.

MedSavant will be made available at http://genomesavant.com/med.



Screenshot of MedSavant: Top panel shows distribution of variants across the genome. Middle panel shows variant filters and results. Bottom panel contains charts showing global trends

1 The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature, 467(7319):1061–1073, October 2010.
2 Brian J O'Roak et al. Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. Nature Genetics, Advance online publication, May 2011.
3 K Inaki et al. Transcriptional consequences of genomic structural aberrations in breast cancer. Genome Research, 21(5):676–687, May 2011.
4 S. T. Sherry, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research, 29(1):308–311, January 2001.
5 P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. Genome research, 11(5):863–874, May 2001.
6 Ivan A. Adzhubei et al. A method and server for predicting damaging missense mutations. Nature methods, 7(4):248–249, April 2010.
7 M. Ashburner et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature genetics, 25(1):25–29, May 2000.
8 A. Hamosh et al. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research, 33(Database issue), January 2005.
9 M. Fiume, et al. Savant: genome browser for high-throughput sequencing data. Bioinformatics, 26(16):1938–1944, August 2010.

Poster Abstracts

Digital Signal Processing of Quantitative Genomic Data

Davide Cittaro[1], Heiko Muller[1], Gabriele Bucci[1]

1). Istituto Italiano di Tecnologia, Center of Genomic Science of IIT@SEMM, Milano

Acquisition of quantitative genomic data (QGD), that is to say a measure of a defined event associated with a position or interval on the chromosome, has been limited to few phenomena gathered from the chromosome sequence as resulted of whole shotgun sequencing studies (i.e. GC content, gene density, interspecies conservation…). Recent and unprecedented flow of sequencing data, provided by high throughput techniques, allowed for the analysis of a wider collection of features including, but not limited to, epigenetic rearrangements and transcription regulation machinery.

Many efforts have been done to develop ad hoc models and software able to analyze each specific case. Such tools are suited for the purposed context but do not perform well outside it: typically the very same model can't be applied seamlessly to the analysis of features spreading over few hundreds of bases or spanning over megabases.

To tackle this problem we reasoned that QGD could be configured as signals in the chromosomal domain, regardless their source. As a consequence, we could borrow digital signal processing (DSP) techniques, exploiting their robustness, flexibility and availability. Some attempts in this direction have been successfully conducted in the past, as in the case of nucleosome positioning studies, using wavelet analysis. In order to show the capabilities of this approach, we chose two very distant biological problems: 1) replication fork progression analysis with Repli-seq and 2) degraded ChIP-seq data (H3K4me3).

The first case is representative of a class of sequencing experiments in which the enrichment of wide functional domains are concealed by background noise; many histone and DNA modification assays fall in the same class. We show that low-pass filters increase the signal-to-noise ratio; consequently, we obtain a better signal segmentation and feature recognition. The second case may happen in presence of degraded biological samples (e.g. FFPE tissues) or poorly performing antibodies in IP experiments. Standard analysis in these conditions could result in scattered signals hampering detection of enrichments. We show that using an appropriate set of operators we can reconstruct the degraded ChIP-seq profile.

We developed dspchip, a python application that implements a simplified interface to a number of DSP procedures ready for the analysis of quantitative genomic data. dspchip supports most of the file formats adopted by popular genomic browsers.

Source code is available for download at http://code.google.com/p/dspchip/.

Analysis of DNA methylation rates with GSNAP, Goby and IGV

Fabien Campagne[1,2]* (fac2003@med.cornell.edu), Nyasha Chambwe[1,2], Ji-eun Oh[3], Thomas D. Wu[4], Jim T. Robinson[5], Kevin C. Dorff[1] and Miklos Toth[3]

1). The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine; Weill Medical College of Cornell University, New York, NY 10021
2).Department of Physiology and Biophysics;
3).Department of Pharmacology; Weill Medical College of Cornell University, New York, NY 10021;
3Genentech, Inc., South San Francisco, CA 94080, USA;
4). Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA.
*To whom correspondence should be addressed.

Several next-generation sequencing protocols have been designed to assay DNA methylation. Among these methods, methyl-Seq provides base-level resolution estimates of methylation rates in biological samples. Methyl-Seq requires sequencing capacity similar to that needed for whole genome sequencing and is therefore currently limited to profiling small number of samples [1]. A more cost effective protocol is the Reduced Representation Bisulfite Sequencing (RRBS) protocol. By focusing on a subset of CpG sites available in a genome, RRBS can yield estimates of methylation in a sample with only tens of million reads. Further protocol developments are expected to yield approaches that enable the estimation of DNA methylation rates in focused regions of the genome, for a large number of biological or clinical samples. We propose that taking advantage of these protocols will require efficient algorithms and tools that can reliably estimate methylation in a variety of genomic contexts, are robust to sequencing or mapping artifacts and that support comparisons across groups of samples. This presentation will describe our progress in developing interoperable analysis tools that satisfy these requirements.

Goby. The Goby framework (http://goby.campagnelab.org/) is a set of APIs, high performance algorithms and implementations that support a variety of data analyses for Next-Generation Sequencing (NGS) projects. The framework offers structured and adaptable file formats that can evolve with the needs of analysis pipelines without disrupting previously written software.
GSNAP. The GSNAP aligner can map reads with indels, or which span exon-exon junctions, perform SNP-tolerant mapping and align bisulfite treated reads [2]. We have recently extended GSNAP to interoperate with the Goby framework. This extension enables efficient parallel alignment on a grid of servers. Using tools implemented with the Goby framework, we estimate methylation rates at observed sites of the genome and write the information in Variant Calling Format (VCF).
IGV. The Integrated Genomic Viewer (IGV) is a widely used genome viewer that supports data integration across modalities [3]. A recent version of IGV includes a VCF track designed to view genomic variations called from NGS data. We have extended the VCF track to make it possible to view methylation rates stored with Goby in VCF files. This extension makes it possible to view DNA methylation rates in the context of gene structure (Figure 1).

Various approaches have been developed to map bisulfite converted reads to the genome and determine the methylation state of read bases, including Bismark, BSMAP, BSeeker or GSNAP. When an experimental design compares samples grouped by treatment or condition of interest the output of these tools typically requires significant additional

processing to extract actionable information. The developments that we present greatly facilitate this process and provide (i) the ability to call sites whose methylation rate differs significantly between groups, (ii) the ability to identify genes whose genomic boundaries contain several sites significantly different between groups, and (iii) the ability to visualize arbitrary regions of the genome and view methylation rates in the IGV genome browser. This presentation will describe the results of simulations and the application of the tools to the analysis of RRBS datasets.

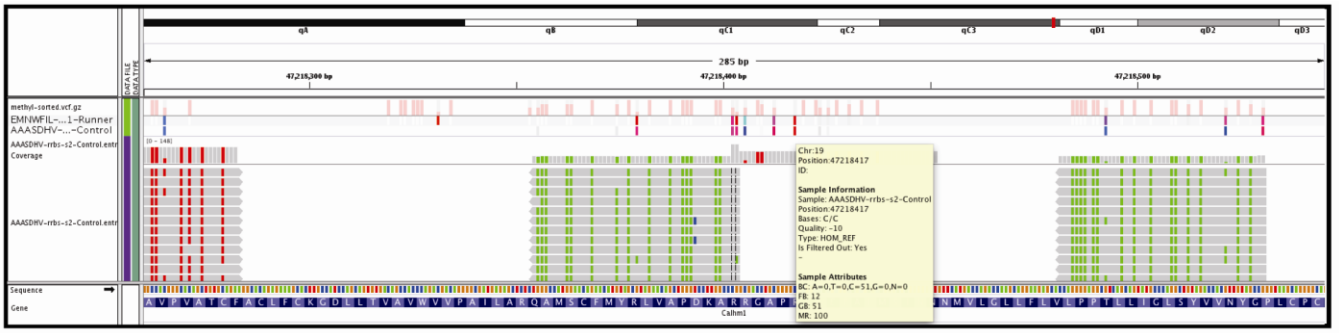


Figure 1. Visualizing DNA methylation for two samples. DNA was obtained from microdissected hippocampal neurons and subjected to RRBS. About 50 million reads per sample were aligned against the mouse genome with GSNAP and GobyWeb. Methylation rates were estimated and written to VCF format with Goby. The VCF file is viewed with IGV early release 2.0. Top track shows color coded methylation rates, second track shows Goby alignment file with aligned reads.

Citations. 1. Lister, R., et al., Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 2009. 462(7271): p. 315-322. 2. Wu, T.D. and S. Nacu, Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 26(7): p. 873-881.3. Robinson, J.T., et al., Integrative genomics viewer. Nat Biotechnol. 29(1): p. 24-6.

A novel pipeline for the assembly of complex genomes from next-generation sequencing data.

Bryan R. Downie, Philipp Koch, Niels Jahn, Jens Schumacher, and Matthias Platzer

With the advent of next generation sequencing technologies (NGS), whole genome sequencing (WGS) can be conducted for much reduced prices and timescales. However, increased error rate and short read length in NGS compared to classical sequencing technologies such as Sanger sequencing introduces new difficulties for de novo whole-genome assemblies of complex and repeat-rich genomes as well as for mapping NGS reads to a reference.

To this end, we have developed the novel pipeline Kilape (K-masking and Iterative Local Assembly of Paired Ends) as a universal method for genome scaffolding and finishing using paired end data. It improves existing assemblies by using only unique (non-repetitive) reads to scaffold contigs together and performing local assemblies of paired end data to perform gap filling. We will present an overview of the pipeline as well as demonstrate improvements of a human de novo assembly using data downloaded from the short read archive.

Multifactorial analysis of digital gene expression data from a large human cohort

Peter A.C. 't Hoen[1], Davis McCarthy[2], Yunshun Chen[2,3], Eco J.C. de Geus[4], Dorret I. Boomsma[4], Brenda W.J.H. Penninx[5], Gertjan B. van Ommen[1] and Gordon K. Smyth[2,3]

1). Center for Human and Clinical Genetics and Leiden Genome Technology Center, Leiden University Medical Center, Leiden, Netherlands;
2). Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia;
3). The University of Melbourne, Melbourne, Australia;
4). Department of Biological Psychology, Netherlands Twin Registry, VU University, Amsterdam, Netherlands;
5). Department of Psychiatry, Netherlands Study of Depression and Anxiety, VU University Medical Center, Amsterdam, Netherlands

With rapidly decreasing sequencing cost, sequencing-based gene expression profiling becomes an attractive alternative over array-based studies. The digital nature of the data asks for new models for statistical analysis and is best modeled with the negative binomial distribution. So far, efforts have focused on comparison between two groups. In the current study, we faced several confounding factors, requiring the implementation of multifactorial statistical models. To this end, we fitted generalized linear models (GLM) on negative binomial distributed count data and estimated the overdispersion (squared biological coefficient of variation) with a Cox-Reid adjusted profile likelihood method. Subsequently, we applied a likelihood ratio test to identify differentially expressed genes. The algorithms have been implemented in the R/Bioconductor package edgeR.

In this study we used deepSAGE, creating one tag per transcript, to analyze gene expression levels in total blood of 94 subjects. We obtained 14 ± 6 million SAGE tags per sample. Despite the high abundance of reticulocyte-derived hemoglobin mRNAs (20-80% of reads), the reliable quantification of mRNAs derived from around 10,000 genes with an expression level of >1 transcript per million. To identify possible risk factors for type 2 diabetes, we were interested to find differences in gene expression between individuals with high and low fasting glucose levels. Gender and body mass index were included as confounders in the GLM. An overview of the results of the fitting of the GLM and the subsequent likelihood ratio test is given in Figure 1. Among the differentially expressed genes, there were many genes in the PKA and MAPK signaling pathways with higher expression in subjects with low fasting glucose levels, confirming the importance of these pathways for the maintenance of glucose homeostasis.

Filtering for low abundant genes (Figure 1B) did not affect differential expression analysis to a considerable extent, stressing that the statistical model is able to cope with stochastic effects in the low abundant genes. A comparison with a microarray-like analysis workflow involving robust normalization, square root transformation followed by limma showed that results were very similar for high abundant genes, but somewhat less consistent for low to medium expressed genes. This indicates that the application of microarray-type analysis methods to count data may lead to false positive identifications due to stochastic events.

In conclusion, we have developed a robust and versatile framework for multifactorial analysis of digital gene expression data. Validation of the framework and identification of additional confounding factors is ongoing. In a next step, we will take differences in the

relative contribution of different blood cell types into account to increase the power for detection of differentially expressed genes.

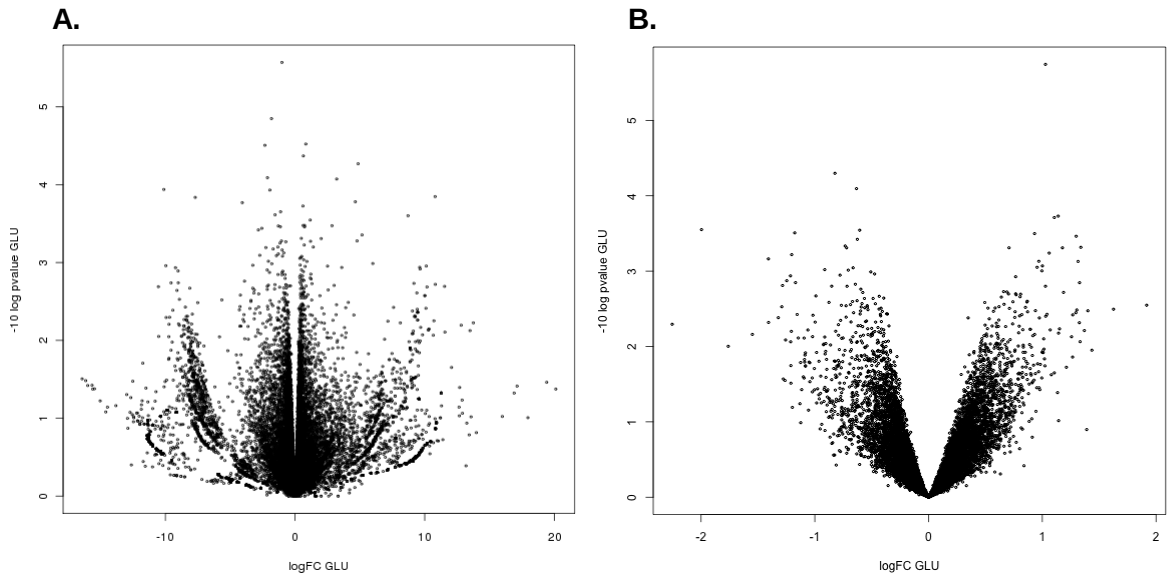


Figure: Volcano plot showing the differences in levels of 30,470 transcripts between subjects with high (N=53) and low (N=41) fasting glucose levels. The x-axis displays the fold-change (logarithmic scale with base 2, after robust normalization with edgeR), while the y-axis displays the inverse log10 of the p-value from a log-likelihood ratio test after fitting a generalized linear model including gender and BMI as confounders. **A.** Results without filtering for low abundant genes. The extreme (but non-significant) fold-changes and lines of points in the plot are due to discrete, low values (0, 1, 2, 3,…) in one of the groups; **B.** Results after filtering for low abundant transcripts, keeping only transcripts with an abundance of at least one transcript per million in 25% of the samples (14,316 out of 30,470). The lines of points have disappeared and the fold changes are in a more realistic range.

This research was funded through EC's FP7 Programme (FP7/2007-2013) ENGAGE, grant agreement HEALTH-F4-2007-201413.

A Robust Method for Transcript Quantification with RNA-seq Data

Yan Huang[1], Yin Hu[1], Matthew S Hestand[2], Corbin D. Jones[3], James N. MacLeod[2], Derek Chiang[4], Yufeng Liu[5], Jan F. Prins[6], Jinze Liu[1]

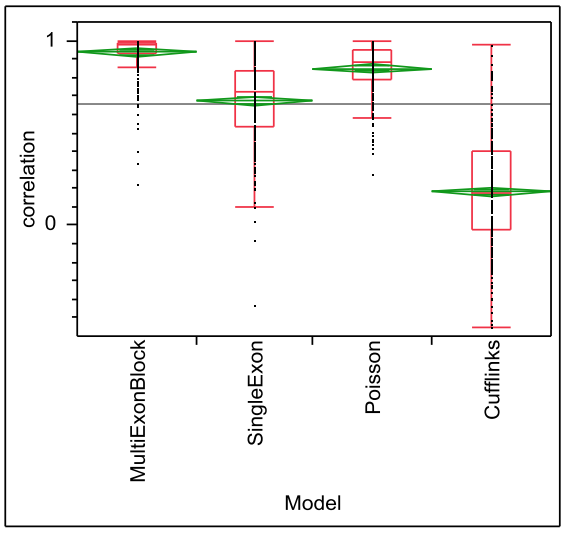
1). Department of Computer Science, 2). Department of Veterinary Science, University of Kentucky. 3). Department of Biology, 4). Department of Genetics, 5). Department of Statistics and Operations Research, 6). Department of Computer Science, University of North Carolina at Chapel Hill.

Motivation

Recent studies have estimated that as much as 95% of multi-exon genes are alternatively spliced, resulting in multiple transcripts per gene[1]. Determining the abundance of each transcript based on short reads sampled from the transcriptome (RNA-seq) is a key step for downstream mRNA transcriptome analysis. Assuming we start from a list of annotated transcripts, several issues complicate the ability to achieve accurate quantitative assessments. First, disambiguating multiple overlapping transcripts from short reads may yield a problem without a unique solution[2,3]. Second, short reads are not sampled uniformly from the transcriptome for a variety of reasons. Recent efforts to better model these effects may ameliorate this problem[4]. Finally, it is common that only a subset of the annotated transcripts for a given gene will be expressed concurrently in a cell. However, existing analytical approaches tend to assign positive expression values to every candidate transcript provided, thereby creating a situation in which large errors in abundance estimation are introduced for transcript variants that in reality are barely expressed. To circumvent these problems, we have developed a robust model for transcript quantification that leverages discriminative features in spliced reads to ameliorate the issue of identifiability.

Methods

In practice, long reads and paired-end reads may connect multiple exons, suggesting that these exons appear together within a single transcript. These multi-exon blocks are more specific in determining the identity of the transcripts than single exons, thereby improving identifiability. In our new method, we model the relationship between the observed read coverage on both single exons and multi-exon blocks and their expected coverage in one linear system. The expected coverage of a multi-exon block is computed by its probability of being covered by a read and the total number of transcripts. By solving the linear regression problem with least absolute shrinkage and selection operator (LASSO[5]), our approach is able to reach a parsimonious set of transcripts, effectively eliminating spurious non-expressed transcripts.



Results

We report the following two key results using simulated RNA-seq reads from transcripts documented in the current human hg19 gene models available in UCSC. Provided sufficient sequencing depth and uniform sampling with synthetic data we find:

1. Genes that are identifiable with the new model have significantly higher accuracy in transcript abundance estimation than the exon-only model. In 308 genes that become identifiable with the new model using 100bp reads, the correlation of the estimated abundance with the ground truth is close to 0.92, much higher than the exon-only model, Poisson model[8], and Cufflinks (version 1.0.1) [11], as shown in Figure 1.

Figure 2: Box plots of correlations between known abundance of 308 genes and their estimated abundance by four different methods.

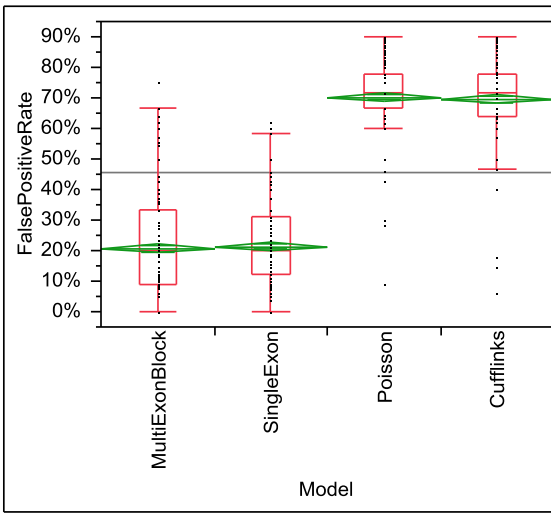


Figure 2A: Box plots of false positive rate of non-expressed transcripts with estimated expression > 0

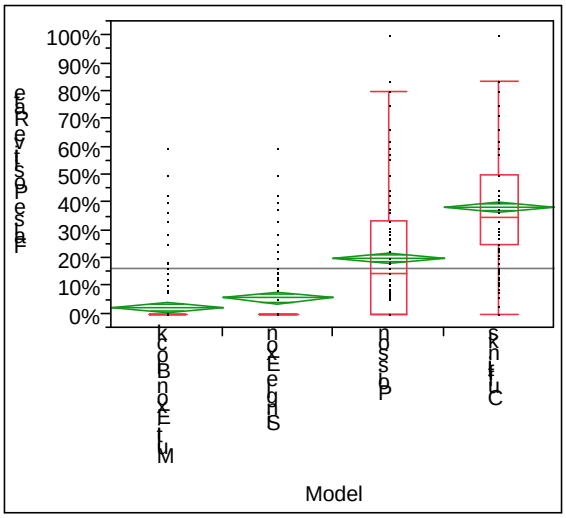


Figure 2B: Box plots of false positive rate of non-expressed transcripts with estimated expression > 10% of total expression

2. Our method also has higher specificity in inferring transcript presence than other approaches. For each gene with more than two transcripts, we simulate RNA-seq reads by sampling from at most two randomly chosen transcripts. The false positive rate of transcripts that were not expressed in the simulated data but were estimated with positive expression is shown in Figure 2A. When transcripts with low estimated abundance (less than 10% total expression) were eliminated, the false discovery rate of all approaches improves. Our proposed approach achieves less than 3% false discovery rate on average, while the Poisson and Cufflinks methods average more than 20% and 30% respectively as shown in Figure 2B.

The newly proposed model also takes advantage of longer reads in transcript inference. Hence, we estimated the read length to infer all transcripts. Of 14,530 multi-transcript human in UCSC, we can identify 95%, 97%, and 99% of the transcript isoforms with read lengths of 50bp, 250bp, and 1000bp respectively. At a read length of 1000bp we can identify genes' isoforms for 99% in mouse, 96% in worm, and 94% in fly of all mRNA transcripts. At a read length of 5000bp, we can identify 99% or greater of the mRNA transcripts for all four species. Current 2nd generation machines generate reads that are up to several hundred bps in length, but the emergence of 3rd generation sequencing will increase the average read lengths to thousands of bps, which makes this set-up a reality.

Conclusion:

We propose a new approach for addressing the identifiability issue of transcripts abundance inference by introducing multi-exon blocks covered by spliced reads. Results from simulated data have shown that we both improve the estimation accuracy and reach a parsimonious set of dominant isoforms present.

Reference:

1. Pan, Q., Shai, O., Lee, L.J., Frey, D.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413-1415 (2008).
2. Lacroix, V., Sammeth, M., Guigo, R. & Bergeron, A. Exact Transcriptome Reconstruction from Short Sequence Reads. *Proceedings of the 8th international workshop on Algorithms in Bioinformatics* (2008).
3. Hiller, D., Jiang, H., Xu, W. & Wong, W.H. Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics* **25**, 3056-3059 (2009).
4. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. & Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol*. **12**, R22 (2011).
5. Tibshirani, R. Regression shrinkage and selection via the lasso*. Journal of Royal Statistical Society Series B (Methodological)* **58**, 267–288 (1996).
6. Jiang, H. & Wong, W.H. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25**, 1026–1032 (2009).
7. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
8. Richard, H. et al. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res.* **38**, e112 (2010).
9. Srivastava S. & Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.* **38**, e170 (2010).
10. Bohnert, R. & Rätsch, G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.* **38**, W348-51 (2010).
11. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
12. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).

The CLC bio integrated framework for identification and comparison of genomic variants

Holger Karas[1], Anika Joecker[1], Jacob Grydholt[1], Bjarne Knudsen[1], Martin Bundgaard[1], Morten Vaerum[1], Roald Forsberg[1]

1). Research & Development department, CLC bio Company, Finlandsgade 10-12, Katrinebjerg, 8200 Aarhus N

The substantial decrease in data production costs has led to the sequencing of complete human genomes becoming a feasible and attractive diagnostic tool in the clinic. The inspection of genomic variants like Single Nucleotide Polymorphisms (SNPs), Copy Number Variations (CNVs) as well as Structural Variations (SVs) like insertions, deletions, duplications and inversions could be used to guide the clinician in choosing the right treatment for patients.

However, a major bottleneck in the movement of these technologies into a clinical setting is the bioinformatic handling and analysis of the data, and the visual inspection and comparison of the results.

For this reason, we have designed the CLC Genomics Workbench which is a user-friendly interface to powerful bioinformatics analyses, offering biomedical researchers a “swiss army knife” for high-throughput sequence data analysis.

The Workbench has already proven itself as a useful, intuitive and flexible tool for the analysis of high-throughput sequencing data, and is in use in many institutes and companies around the world for this purpose.

In this work we will show how new features in the Workbench can be used to analyze a complete human genome accurately and quickly, to identify relevant variants. As an example we will use an Illumina paired-end read data set from an human individual, which is a sample from the HapMap project. Structural variations as well as SNPs are well described for this type of data, which enables the measurement of the performance of the analysis tools. Here we will step through the whole re-sequencing workflow, from mapping of the reads to the reference sequence, tor the identification of SNPs, copy number variations and structural variations, and finishing with the filtering of common variants and comparison of the results with available gene annotations using the new integrated annotation comparison plugin.

We will demonstrate that the new Workbench functionality gives accurate results while being faster than commonly used tools like BreakDancer and MAQ. Furthermore, we will show that the newly integrated annotation comparison plugin enables filtering, as well as comparison of all analysis results with external data sources, in an intuitive way.

QuasR : Quantify and Annotate Short Reads in R

Anita Lerch, Dimosthenis Gaidatzis, Florian Hahne, Michael Stadler

Deep sequencing technology, due to its high throughput and low cost, has become a powerful research tool in a wide range of applications, such as RNA-seq and ChIP-seq. In the last years there have been many efforts in the bioinformatics community to provide software in R/Bioconductor to simplify the processing and the biological interpretation of such large data sets. However until now, there is no integrated start-to-end analysis solution within R that sufficiently abstracts the technical details and would be suitable for use by biologists. In particular, alignments need to be performed outside of R and genome annotation information must be manually incorporated. Here we outline the deep sequencing analysis package QuasR, a further development of the FMI deep sequencing pipeline, built to make efficient use of available hardware resources and to simplify analysis of next generation sequencing data.

Statistical strategies for de-noising RNA sequencing coverage data.

Anna Lesniewska[1,2], Martin Ryan[1], Michal J. Okoniewski[1]

(1) Functional Genomics Center UNI ETH Zurich, Winterthurerstrasse 190, CH-8057,Zurich, Switzerland
(2) Poznan University of Technology, Institute of Computer Science, ul.Piotrowo 2, 60-965 Poznan, Poland

RNA sequencing [1,2] is one of the applications of next-generation sequencers, such as Illumina High-seq or ABI SOLiD. The initial data are fragments of RNA sequences that are then mapped to the reference genome. In this way a coverage function is obtained, which is defined as the number of counts mapped to every nucleotide. The coverage function may then be viewed in genome browsers such as IGV [3] or GBrowse [4] to reveal the landscape of RNA expression in the biological sample.

Due to various sources of noise and factors contributing to the mappability [4] of reads, the coverage function has a number of artifacts such as peaks or sudden drops within well-defined exons. The package rnaSeqMap [5] includes tools for processing RNAseq expression profiles defined as the coverage function. Among those tools are the Aumann-Lindell algorithm [6] for discovering regions with high coverage and new applications of classic loess local regression for smoothing the profiles.

We performed several tests to prove the utility of these tools to de-noise the coverage data and to distinguish the artifacts. As a result we present strategies which may lead to more accurate read counts and comparisons with the mappability track from [4]. Understanding the artifacts of RNA sequencing will lead to the development of more complex operations for the coverage profiles.

[1] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 2008, 5(7)

[2] Oshlack A, Robinson MD, Young MD, From RNA-seq reads to differential expression results, Genome Biology 2010, 11:220

[2] Integrated Genome Viewer, [http://www.broadinstitute.org/igv].

[3] GBrowse, [http://gmod.org/wiki/GBrowse]

[4] Koehler R, Issac H, Cloonan N, Grimmond SM, The uniqueome: a mappability resource for short-tag sequencing Bioinformatics (2011) 27(2): 272-274

[5] Lesniewska A, Okoniewski MJ. rnaSeqMap: a Bioconductor package for RNA sequencing data exploration, BMC Bioinformatics. 2011 May 25;12(1):200

[6] Aumann Y, Lindell Y: A Statistical Theory for Quantitative Association Rules. J. Intell. Inf. Syst. 2003, 20(3)

Mutalyzer 2: Improved Sequence Variant Descriptions from next generation sequencing data and gene variant databases

Jeroen F.J. Laros, Martijn Vermaat, Gerben R. Stouten, Gerard C.P. Schaafsma, Johan T. den Dunnen, and Peter E.M. Taschner

Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, Nederland
P.Taschner@lumc.nl

Most sequence variants identified in next generation sequencing data are described using chromosomal position numbering or dbSNP rsIDs. These have to be converted to transcript-oriented positions for storage in gene variant databases (locus-specific databases, LSDBs) or to query these databases for phenotypic and functional consequence data. Mutalyzer 2 facilitates batch-wise conversion from dbSNP rsIDs or chromosomal position numbering to transcript position numbering for all genes and their corresponding transcripts. Mutalyzer is also used to quickly check new variant submissions in LSDBs based on LOVD software (www.LOVD.nl/) to prevent mistakes and uncertainties leading to undesired errors in clinical diagnosis. The Mutalyzer sequence variation nomenclature checker names all sequence variants following the Human Genome Variation Society sequence variant nomenclature recommendations (www.hgvs.org/mutnomen), using a GenBank or Locus Region Genomic (LRG) accession number, a HGCN gene symbol and the variant description as input. Mutalyzer generates an output containing a description of the sequence variant at DNA level, the effect on all annotated transcripts, its deduced outcome at protein level and gains or losses of restriction enzyme recognition sites, as well as checking of sequence variants in locus-specific sequence variation databases (LSDBs). In addition, LOVD2 uses Mutalyzer 2 to map variants to genomic positions for visualization in the Ensembl and UCSC genome browsers and the NCBI Sequence Viewer. In LOVD3, Mutalyzer will support the description of variants on multiple transcripts of a specific gene. The new Name Generator can be used to train your self to generate correct HGVS descriptions. New webservices also support the use of Mutalyzer's functionality from other computer programs. Mutalyzer 2 is accessible at www.mutalyzer.nl.

Funded in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 200754 - the GEN2PHEN project.

Functional k-means clustering and principal component analysis of ChIP-Seq data

Pedro Madrigal, Pawel Krajewski

Institute of Plant Genetics, Polish Academy of Sciences, Poznan, Poland
pmad@igr.poznan.pl, pkra@igr.poznan.pl

The advent of next generation sequencing (NGS) technologies entails new computational challenges in the analysis of high throughput -omic data, like locating the genome-wide protein bound regions in transcription factor binding experiments. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) allows to isolate and map specific DNA sites directly interacting with transcription factors and other chromatin-associated proteins. Moreover, identifying protein-DNA interactions has been proved essential to establish the rules of the regulatory transcriptional networks in higher organisms. The foundations of the majority of peak detection algorithms are based on the search of local peaks over the coverage counts of the mapped reads along the genome, assuming a certain type of statistical distribution [1]. Our aim is to introduce a new method based on functional data analysis [2], which moves the classic paradigm of multivariate statistical analysis into a set of underlying smoothed functions encapsulating a model of the biological process under study. So far, first we propose smoothing splines as a representation of a set of candidate binding sites. Based on this approach, we perform k-means clustering with the objective to divide the enrichment signals into clusters or groups according to different binding profile patterns. Secondly, a functional principal component analysis highlights the way in which a set of ChIP-Seq functional data varies from its mean and quantifies the discrepancy from the mean of each candidate transcription factor binding site in terms of modes of variability (principal components). It enables us to robustly recognize the most significant transcription factor binding sites according to the global variation within the ChIP-Seq dataset. Analyses of data of the model plant Arabidopsis coming from experiments involved in flower development will be presented using the methodology proposed.

[1] Wilbanks E.G., Facciotti M.T. (2010). Evaluation of algorithm performance in ChIP-Seq peak detection. PLoS ONE 5(7): e11471.
[2] Ramsay, J.O., Silverman, B.W. (2005). Functional Data Analysis (Second edition). Springer Series in Statistics, New York.

A software platform for an end-to-end analysis of metagenomics data sets

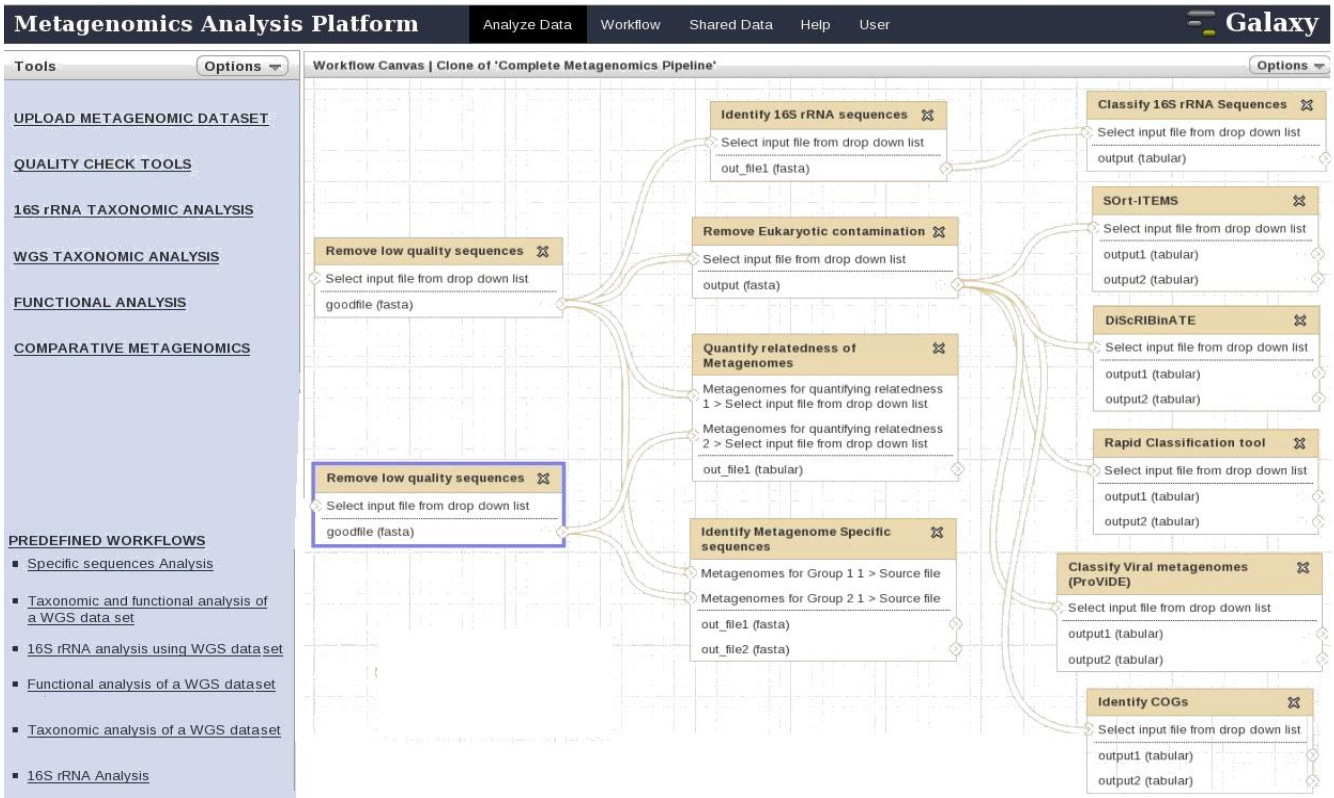
Monzoorul Haque Mohammed, Sudha Chadaram, C. V. S. K. Reddy, and Sharmila S Mande*

Bio-sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services Limited
1 Software Units Layout, Madhapur, Hyderabad – 500083, Andhra Pradesh, INDIA
*Corresponding author: sharmila@atc.tcs.com

Two recent developments have significantly impacted the focus of life-sciences and health-care R&D. First is the emergence of low cost and high throughput Next Generation Sequencing (NGS) technologies. Second is the concomitant emergence of a new research area called 'Metagenomics'. By performing operations on specially fabricated micro/nanochips, NGS technologies have enabled rapid and simultaneous sequencing of millions of DNA fragments in a cost effective manner. Metagenomics has its roots in the following observation. Majority of microorganisms present in natural ecosystems cannot be cultured in experimental labs. The metagenomics approach bypasses the culturing step and analyzes DNA obtained directly from microbes inhabiting a given environmental niche. This approach thus enables the direct exploration, characterization, and beneficial exploitation of the unexplored diversity of microorganisms. Results of several ongoing studies have indicated the tremendous potential of metagenomics in unearthing thousands of novel genes and proteins, some of which have commercial potential. Given that genomic content of a multitude of microbes (living in a particular environment) are sequenced and analyzed in a typical metagenomics project, NGS technologies have played a perfect complementary role in furthering the development of Metagenomics.

Despite the tremendous potential of NGS technologies and metagenomics, the lack of efficient tools, algorithms, and analysis platforms that can perform an accurate and meaningful end-to-end analysis of generated data remains a prime concern of researchers in life sciences and health care sectors. Besides specialized algorithms, significant compute power and infrastructure are also required for analyzing metagenomics data. In this study, we present a comprehensive analysis platform that implements a suite of algorithms specialized for metagenomic sequence data sets. The platform includes algorithms for data pre-processing, decontamination of host associated sequences, detection and taxonomic characterization of 16S rDNA sequences, functional and taxonomic characterization of the entire metagenomic content, comparative analyses of metagenomic data sets, as well as, tools for the detection of habitat specific genes/sequences. Apart from hosting several pre-defined work-flows, the platform allows creation of customized project-specific work-flows that enables the end users to perform an end-to-end analysis of metagenomic data sets. Details of this metagenomics analysis platform will be presented during the conference.

(Diagram on next page)



A screen-shot depicting the various functionalities, predefined work-flows and the interface which allows users to create customized project-specific work-flows.

A Distributed and High-Throughput Short Read Processing Suite

Luca Pireddu*, Simone Leo and Gianluigi Zanetti

CRS4, Polaris, Ed. 1, I-09010 Pula, Italy
luca.pireddu@crs4.it

Seal is a suite of open source tools for short read processing designed for high-throughput sequencing operations. The suite currently includes scalable, distributed tools for: demultiplexing output from Illumina multiplexed sequencing runs; read filtering and format conversion; read mapping (based on the popular BWA aligner); duplicate read identification and removal; sorting read mappings.

These tools are designed following the MapReduce scalable computing model and have a throughput that scales linearly with the number of computing nodes, providing a solution that can grow in capacity with the amount of data to be processed. Seal leverages the Hadoop open source MapReduce distributed computing platform to provide resilience to node failures and transient events such as peaks in cluster load. In conclusion, Seal provides tools that can harness all available computational resources to efficiently process large amounts of data with a limited amount of operator effort.

The Seal suite is currently used to implement most of the production pipeline at the CRS4 Sequencing and Genotyping Platform, currently processing data from 6 Illumina sequencing machines. Seal is available online at http://biodoop-seal.sourceforge.net/

Model-based bayesian clustering of chromatin ChIP-Seq data coupled with visualization by graph representation

Romain PONCET, Denis MESTIVIER, Thierry GRANGE

Institut Jacques Monod, UMR CNRS, UniversitÈ Paris Diderot
15 rue HÈlËne Brion, Paris, France
romain.poncet@ijm.univ-paris-diderot.fr

Clustering approaches are a useful first step when trying to interpret a large data set, especially in the context of high-throughput sequencing methods. Clustering using K-means has been used to reveal patterns of positioning of the histone variant H2A.Z at human gene promoters, including classes of positioning patterns correlating with nucleosome shifting during passage from interphase to mitosis [4]. We sought to apply sensitive clustering techniques with this data to empower identification of factors correlating with more precisely described classes.

Among the wealth of available clustering methods, an unsupervised Bayesian classification system, ´†AutoClass†ª (Ames Research Center, NASA), has key advantages [1,2]. It separates the data into an automatically determined optimal number of classes. Furthermore, it can handle missing values, and it can cluster heterogeneous data: discrete and continuous valued data. AutoClass† has been exploited to classify exon donor and acceptor sites, and yeast gene expression microarray data [1,2,3]. It is freely available via a web server : Autoclass@IJM [3] at: http://ytat2.ijm.univ-paris-diderot.fr/AutoclassAtIJM.html.

Using multiple Autoclass runs until convergence on H2A.Z ChIP-seq enrichment profiles at human promoters from [4] reveals that the determined optimal number of classes is above 100, whereas the original authors used K-means with 12 classes. A bigger number of classes could augment our ability to discern patterns in the data.

However, such a large number of classes compromises the possibility to generate practical testable models, and renders the interpretation less robust. In order to exploit the discriminating power of a clustering method based on finding many classes, we explored methods to determine the relationships between the classes, and ways to group classes into ìsuperclassesî according to reproducible criteria. Such presentations of the data offer greater potential for generating robust models of the underlying phenomena.

Our approach consists in building a graph with classes as nodes (from an AutoClass clustering), and using a measure of distances between two classes classes as the edge weights. Subsequently, incorporation of additional criteria enables correlation between various features at promoters with their chromatin profiles. By selecting vertices according to chosen criteria, such as selecting only those between two classes with similar average transcriptional activity, each connected component of the new graph represents a set of genes that are similar in terms of enrichment profile, and share additional characteristics.

Ultimately, we aim to integrate AutoClass, and our auxiliary utilities (chromatin profile pre-processing, graph generation), into a workflow engine (such as Galaxy) to permit non-specialists to transform raw NGS data into testable models with a few mouse clicks.

[1] Cheeseman,P., Kelly,J., Self,M., Stutz,J., Taylor,W. And Freeman,D., AutoClass: A Bayesian Classification System, Proceedings of the Fifth International Conference on Machine Learning. Morgan Kaufmann Publishers, San Francisco (1988).

[2] Cheeseman,P. and Stutz,J. ,Fayyad,U., Piatelsky, Bayesian Classi?cation (AutoClass): Theory and Results. In Advances in Knowledge Discovery and Data Mining, AAAI Press/MIT Press, Cambridge (1996)

[3] Achcar F, Camadro JM, Mestivier D., AutoClass@IJM: a powerful tool for Bayesian classification of heterogeneous data in biology. , Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W63-7. Epub 2009 May 27.

[4] Kelly TK, Miranda TB, Liang G, Berman BP, Lin JC, Tanay A, Jones PA., H2A.Z maintenance during mitosis reveals nucleosome shifting on mitotically silenced genes., Mol Cell. 2010 Sep 24;39(6):901-11.

SlideSort: Fast and exact algorithm for Next Generation Sequencing data analysis

Kana Shimizu and Koji Tsuda

Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology
Contact: Shimizu-kana@aist.go.jp

Next Generation Sequencing (NGS) technology calls for fast and accurate algorithms that can evaluate sequence similarity for a huge amount data. In this study, we designed and implemented exact algorithm SlideSort that finds all similar pairs whose edit-distance does not exceed a given threshold from NGS data, which helps many important analyses, such as de novo genome assembly, identification of frequently appearing sequence patterns and accurate clustering.

Using an efficient pattern growth algorithm, *SlideSort* discovers chains of common k-mers to narrow down the search. Compared to existing methods based on single k-mer, our method is more effective in reducing the number of edit-distance calculations. In comparison to state-of-the-art methods, our method is much faster in finding remote matches, scaling easily to tens of millions of sequences. Our software has an additional function of single link clustering, which is useful in summarizing NGS data for further processing.

CAMERA: Next-Generation Sequencing (NGS) and Metagenomics

Shulei Sun, Jing Chen, Weizhong Li, Ilkay Altinatas, Abel Lin, Steve Peltier, Karen Stocks, Eric Allen, Mark Ellisman, John Wooley and Jeffrey Grethe

High throughput NGS technologies, such as Illumina, provide tremendous flexibility for cost-effective genomic investigations that can be scaled to suit diverse project goals. In terms of sequence coverage alone, NGS platforms represent a dramatic advance over capillary-based methods, however they also present significant challenges related to data handling, data archiving and analysis. NGS data sets create further data quality challenges due to short read lengths, method-specific sequencing errors, and the absence of physical clones.

CAMERA, the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis, is a single system for depositing, locating, analyzing, visualizing and sharing sequencing data and metadata about microbial biology. Its computational resources include rich and distinctive bioinformatics tools and a flexible collaborative workflow environment that makes it possible for researchers to analyze large and complex sequencing data generated by NGS technologies.

CAMERA organizes data analysis tools through collaborative, user selected scientific workflows. At the core of this environment is the Kepler scientific workflow platform that supports the integration of data and provides capabilities for provenance, associated with the processing of data. This data-oriented view of an analysis captured via end-to-end provenance of a workflow enables the communication of what has occurred within a workflow to collaborators and allows for the exchange and reproducibility of the computation itself. Through the CAMERA portal, users can create, share, retrieve and run the processing workflows specific to their own experiment without having to install special software. In addition, this workflow-based environment reduces the cost of moving from a stand-alone scientific application to a workflow-based community resource by (i) designing and publishing workflows based on application services; (ii) executing workflows based on local or online data; (iii) saving and querying workflow results; (iv) saving and viewing data and process provenance; (v) creating ad-hoc collaborations and project spaces; and (vi) publishing uploaded workflows or sharing workflows with workspace members. Using this platform, CAMERA's workflow system currently makes the following metagenomic analyses available to researchers: a QC filter for sequencing raw reads quality control, identification of duplicates reads, assembly, gene prediction, clustering (DNA and protein), and functional annotation (including COG, KOG, Pfam, TIGRFAM, protein clustering database), BLAST searches, KEGG pathway analyses, community diversity analyses, and more ongoing workflows for NGS.

Expandable de novo genome assembler for short-read sequence data.

Nikolay Vyahhi, Sergey Nurk, Anton Bankevich, Max Alekseyev, Pavel Pevzner.

Algorithmic Biology Laboratory, St. Petersburg Academic University, Russian Academy of Sciences, St. Petersburg, Russia.

De novo genome sequence assembly is the essential step to reveal genomic sequences of different species world-wide. Currently there exists various genome assemblers for short-read NGS data, such as Velvet, SOAPdenovo, ALLPATH, ABySS and others. We present new open-source de Bruijn graph-based assembler currently in development on C++, which uses novel algorithmic ideas such as context-free graph approach and also have agile and expandable software architecture. It requires affordable amount of memory and computations while giving high quality results. It provides solid basis for single-cell and mammalian assemblers in the near future.

Conflicting deletion calls in matched normal/tumor genomes

Roland Wittler[1,2,*], Cedric Chauve[2]

1) Technische Fakult‰t, Universit‰t Bielefeld, Germany
2) Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada
*) corresponding author

Background: Structural variations in human genomes, such as insertions, deletion, or rearrangements, play an important role in cancer development [1, 2]. Next-Generation Sequencing technologies have been central in providing ways to detect such variations. In recent years, many software tools have been developed that follow this approach to identify putative structural variations (reviewed in [3, 4]). Most existing methods however are limited to the analysis of a single genome, and this is only recently that the comparison of closely related genomes has been considered. In particular, a few recent works considered the analysis of data sets obtained by sequencing both tumor and healthy tissues of the same cancer patient. In that context, the goal is to detect variations that are specific to exactly one of the genomes, for example to differentiate between patient-specific and tumor-specific variations. This is a difficult task, especially when facing the additional challenge of the possible contamination of healthy tissues by tumor cells and conversely.

Results: In the current work, we analyzed a data set of mate-pair short-reads, one obtained by sequencing tumor tissues and one obtained by sequencing healthy tissues, both from the same cancer patient. Based on a combinatorial notion of conflict between deletions, we show that in the tumor data, more deletions are predicted than there could actually be in a diploid genome. In contrast, the predictions for the data from normal tissues are almost conflict-free. We designed and applied a method, specific to the analysis to such pooled and contaminated data sets, to detect potential tumor-specific deletions. Our method takes the deletion calls from both data sets and assigns reads from the mixed tumor/normal data to the normal one with the goal to minimize the number of reads that need to be discarded to obtain a set of conflict-free deletion clusters. We observed that, on the specific data set we analyze, only a very small fraction of the reads needs to be discarded to obtain a set of consistent deletions.

Conclusions: We present a framework based on a rigorous definition of consistency between deletions and the assumption that the tumor sample also contains normal cells. A combined analysis of both data sets based on this model allowed a consistent explanation of almost all data, providing a detailed picture of candidate patient- and tumor-specific deletions.

References

[1] Mardis E: Cancer genomics identifies determinants of tumor biology. Genome Biol. 2010, 11(5):211.
[2] Pleasance E, Cheetham R, Stephens P, McBride D, Humphray S, Greenman C, Varela I, Lin M, nez GR O, GR B, et al.: A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010, 463(7278):191≠196.
[3] Medvedev P, Stanciu M, Brudno M: Computational methods for discovering structural variation with next-generation sequencing. Nature Methods 2009, 6:S13≠S20.
[4] Dalca A, Brudno M: Genome variation discovery with high-throughput sequencing data. Brief. Bioinform 2010, 11:3≠14.