ISMB 2009 Special Interest Group on Short Read Sequencing Schedule

8:40-10:15am:	**Session 1: Welcome, SNP Discovery & Cancer Genomics (4 talks)**
10:15-10:45am:	***Coffee break***
10:45am-12:30pm:	**Session 2: Variation Discovery (5 talks)**
12:30-1:45pm:	***Lunch***
1:45pm-3:30pm:	**Session 3: RNA Sequencing (5 talks)**
3:30-4:00pm:	***Coffee break***
4:00pm-6:15pm:	**Session 4 Metagenomics, Assembly, Statistics (4 talks) & Keynote**

ISMB 2009 Special Interest Group on Short Read Sequencing List of Talks

SNP Discovery & Cancer Genomics: 8:40-10:15am

- VARiD: Variation Detection in Color-Space and Letter-Space – Adrian V. Dalca and Michael Brudno

- Evaluation of a Bayesian mixture model for detection of single nucleotide variants in ovarian cancer transcriptomes by next generation sequencing – Sohrab P. Shah, Rodrigo Goya, Mark G.F. Sun, Gavin Ha, Ryan Morin, Kim Wiegand, Kevin Murphy, Sam Aparicio, David Huntsman

- Detecting Polymorphisms in Cancer Tumour/Normal Pairs – Dirk Evers

- ISOLATE: A computational strategy for identifying the primary origin of cancers using high throughput sequencing – Gerald Quon and Quaid Morris

Variation Discovery: 10:45am-12:30pm

- Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads – Kai Ye, Marcel H. Schulz, Quan Long, Rolf Apweiler and Zemin Ning

- Next-generation algorithms: detection of SNPs, InDels, and Copy Number Variation in massively parallel short-read oligonucleotide ligation sequencing – Fiona C.L. Hyland, Rajesh Gottimukkala, Ryan Koehler, Susan Tang, Eric Tsung, Heather Peckham, Kevin McKernan, Francisco De La Vega.

- Detecting Copy Number Variation with Mated Short Reads – Paul Medvedev, Marc Fiume, Tim Smith, Adrian Dalca, Seunghak Lee, Michael Brudno

- A method for detecting small scale human microsatellite length- polymorphism using Solexa/Illumina paired-end sequencing data – Weldon Whitener, Avril Coghlan, Li Heng, Richard Durbin

- MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions – Seunghak Lee, Fereydoun Hormozdiari, Can Alkan, Michael Brudno.

RNA Sequencing: 1:45pm-3:30pm

- TopHat: discovering splice junctions with RNA-Seq – Cole Trapnell, Lior Pachter and Steven L. Salzberg

- Quantitative Detection of Alternative Transcripts with RNA-Seq Data – Regina Bohnert, Jonas Behr, and Gunnar Rätsch

- MapSplice: Map RNA-seq Short Reads for Splice Junction Discovery – Jinze Liu, Kai Wang, Zheng Zeng, Stephen J. Coleman, James N. MacLeod , Jan Prins

- *De novo* Transcriptome Assembly with ABySS – İnanç Birol, Shaun D Jackman, Cydney Nielsen, Jenny Q Qian, Richard Varhol, Greg Stazyk, Ryan D Morin, Yongjun Zhao, Martin Hirst, Jacqueline E Schein, Doug E Horsman, Joseph M Connors, Randy D Gascoyne, Marco A Marra and Steven JM Jones

- Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data – Chol-Hee Jung, Martin A. Hansen, Igor V. Makunin, Darren Korbie and John S. Mattick

Metagenomics, Assembly, Statistics: 4:00-6:15pm

- Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly – Bas E. Dutilh, Martijn A. Huynen and Marc Strous

- LOCAS: A new low coverage assembler for short reads – Juliane D. Klein, Stephan Ossowski, Korbinian Schneeberger, Detlef Weigel and Daniel H. Huson

- Poisson Model of Significance for Short Reads Concentrations – Adam Kowalczyk, Thomas Conway, Bryan Beresford-Smith, Sibgat Choudhury, Saraswati Sukumar, Kornelia Polyak and Izhak Haviv

- Design of Association Studies with Pooled Next-Generation Sequencing Data – Su Yeon Kim, Yingrui Li, Yiran Guo, Ruiqiang Li, Torben Hansen, Oluf Pedersen, Jun Wang, and Rasmus Nielsen

Keynote: Edwin Cuppen

Biologically Relevant Advantages and Limitations of Short Read Sequencing Data

VARiD: Variation Detection in Color-Space and Letter-Space

Adrian V. Dalca and Michael Brudno

The recent developments in Next Generation Sequencing (NGS) technologies have transformed the study of genetic variation. While most of these technologies directly sequence the residues of the genome (letter-space sequencing), the Applied Biosystems SOLiD method yields dibase-coded (color-space) sequences. Several algorithms have been developed for both of these sequencing methods to map and align the reads to a draft genome, and many toolsets further facilitate variation studies. However, most tools for single nucleotide polymorphism (SNP) detection in color-space translate each read, or the genome, before applying heuristic consensus statistics. Due to the properties of the color-space system, translating in this manner suffers from several drawbacks.

We present a Hidden Markov Model (HMM) for representing both color-space and letter-space reads together, and a framework (VARiD) for determining variation without direct translation of those reads. This method allows for accurate detection of heterozygous, homozygous, and tri-allelic SNPs, as well as short indels and consecutive SNPs. The VARiD algorithm treats the read colors and read letters at a position as emissions of the HMM, and the correct donor letters as the hidden states. We utilize a Forward-Backward algorithm together with several filters yielding an intuitive yet powerful base caller.

Under simulations, we observe a very strong ROC curve for SNP detection (both heterozygous and homozygous calls). As an example, at an average 10x coverage, VARiD will yield ~95% True Positive and ~0.2% False Positive SNP Rate, and ~85% True Positive and 0% False Positive Indel Rate. We observe only small decreases in accuracy for lower coverage simulations.

We also tested VARiD using sequencing reads from a Human BAC datasets from JCVI, with Sanger-validated SNP calls. Over the entire dataset, we detected 54 out of 61 SNPS (including both heterozygous and homozygous calls), with most of the undetected SNPs having low coverage and badly aligned reads, with a False Positive rate under 0.5%.

Evaluation of a Bayesian mixture model for detection of single nucleotide variants in ovarian cancer transcriptomes by next generation sequencing

Sohrab P. Shah1, Rodrigo Goya1, Mark G.F. Sun1, Gavin Ha1, Ryan Morin2, Kim Wiegand1, Kevin Murphy, Sam Aparicio1, David Huntsman1

1 – British Columbia Cancer Agency, Vancouver BC Canada
2 - Genome Sciences Centre, British Columbia Cancer Research Centre, Vancouver BC Canada
3 – Dept of Computer Science, UBC, Vancouver BC Canada

The advent of next generation sequencing (NGS) has propelled the field of cancer genomics forward such that it is now cost-effective to interrogate entire genomes or transcriptomes of clinical tumor samples for the presence of somatic mutations. This approach generates a massive number of short sequence reads which once aligned to a reference human genome can reveal positions containing single nucleotide variants (SNVs). These are positions in a genome/transcriptome for which at least one allele differs from the reference human genome and can be either germline polymorphisms or somatic mutations. Numerous methods have been developed for short read alignment. However, there is a relative paucity of methods in the literature for inferring SNVs as a post-alignment inference step. Large scale data generation endeavours are now underway using NGS to sequence 1000s of tumours, therefore it will be critical to have in hand robust and accurate analytical tools designed for detection of SNVs from NGS data.

The challenge in SNV detection is that the alleles are represented by a stochastic distribution of allelic counts in the aligned reads. To model this distribution and infer the presence of SNVs, we developed a novel Bayesian mixture model called SNVmix. The model assumes the allelic counts at each position were generated from one of three genotypes, each of which 'emit' data with a class conditional Binomial distribution. We implemented an expectation maximization algorithm that simultaneously learns the parameters of the Binomial distributions, the prior distribution over genotypes, and the genotypes themselves for each position. In addition, we implemented a version of the model that can account for uncertainty in the base call and the alignments of the reads. Using this latter model, we examined the effect of excluding and/or probabilistically weighting the contribution of reads on the basis of base quality and alignment quality. We quantitatively compared these models to each other and to a state of the art, commonly used method for calling SNVs called MAQ, which does not fit the model to data but rather sets the parameters of the distribution by hand.

In this study we show the results of SNVmix applied to 16 ovarian cancer data sets derived from clinical samples. To evaluate the model, we generated Affymetrix SNP 6.0 high density genotyping arrays and genotyped the samples. Using positions (mean 9000 from each case and 150000 positions total) from the array at which a genotype could be confidently called as ground truth, we computed receiver operator characteristic curves to evaluate the accuracy of our model. We show that fitting the model to the data using SNVMix outperforms Maq by a statistically significant margin. We show further improvement by probabilistically weighting the contribution of reads as described above. Finally, we show novel, clinically relevant SNVs in ovarian cancer discovered with SNVMix that have been experimentally validated and confirmed as recurrent in an external cohort.

Detecting Polymorphisms in Cancer Tumour/Normal Pairs

Dirk Evers, Illumina Inc.

CASAVA is Illumina's software to call structural variation events on genomes of all size ranges. The presentation will give an overview of the software architecture and illustrate some new algorithms we are testing in the CASAVA Framework to ensure improved sensitivity and specificity of event calling algorithms. Emphasis will be placed on lessons learned on the interplay of calling different classes of events, such as copy number variations and SNPs, as well as specialized algorithms to explore events on different size scales such as short and long insertions/deletions.

In collaboration with the Wellcome Trust Sanger Institute, we have sequenced both tumour and normal genomes from a malignant melanoma cell line, with the aim of identifying somatic variants in the cancer sample. The tumour and normal were sequenced on the Illumina Genome Analyzer II system to 40x and 30x depth respectively, mostly using paired 75bp reads with mean insert size of approx 200bp plus smaller quantities of paired 50bp reads of 2kb, 3kb and 4kb libraries.

SNPs, short insertions/deletions, copy number and structural variants were called on both genomes with our CASAVA software. In order to identify somatic events, we employed strict variant calling criteria in the tumour sample and subtracted a more permissive set of variant calls from the normal. Selected events have been validated by PCR and capillary sequencing at the Wellcome Trust Sanger Institute, among them typical UV radiation induced polymorphisms and other previously known structural variants.

ISOLATE: A computational strategy for identifying the primary origin of cancers using high throughput sequencing

Gerald Quon & Quaid Morris

University of Toronto

Motivation: One of the most deadly cancer diagnoses is the carcinoma of unknown primary origin (CUP). Without knowledge of the site of origin, treatment regimens are limited in their specificity and result in high mortality rates. Classification models based on microarray gene expression data have been previously constructed to predict the site of origin, but they depend on previously classified cancer expression data on which to train, do not account for sample heterogeneity, and rely on noisy microarray technology.

Results: We present ISOLATE, a new statistical model that simultaneously predicts the primary site of origin of cancers and addresses sample heterogeneity, while taking advantage of new high throughput sequencing technology that promises to bring higher accuracy and reproducibility to gene expression profiling experiments. ISOLATE makes predictions de novo, without having seen any training expression profiles of cancers with identified origin. Compared to previous methods, ISOLATE is able to predict the primary site of origin, deconvolve and remove the effect of sample heterogeneity, and identify differentially expressed genes with higher accuracy, across both synthetic and clinical datasets. Models such as ISOLATE are invaluable tools for clinicians faced with carcinomas of unknown origin.

Availability: ISOLATE is available for download at http://www.cs.toronto.edu/~gerald/isolate .

Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads

Kai Ye[1], Marcel H. Schulz [1,3], Quan Long[2], Rolf Apweiler[1] and Zemin Ning[2]

[1]EMBL Outstation European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.
[2]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.
[3]Max Planck Institute for Molecular Genetics and International Max Planck Research School for Computational Biology and Scientific Computing, Berlin, Germany

There is a strong demand in the genomic community to develop effective algorithms to reliably identify genomic variants. Indel detection using next-gen data is difficult and identification of long structural variations is extremely challenging.
We present Pindel, a pattern growth approach, to detect breakpoints of large deletions and medium sized insertions from paired-end short reads. We use both simulated reads and real data to demonstrate the efficiency of the computer program and accuracy of the results.
As far as we know, Pindel is the first program to compute deletion events as large as 10kb with base level precision from 36bp paired-end short reads. Due to Pindel's high performance in sensitivity, specificity and efficiency in memory and speed, it has been proved to be a promising approach to address the structural variations between individuals from next-gen high throughput sequencing.

Next-generation algorithms: detection of SNPs, InDels, and Copy Number Variation in massively parallel short-read oligonucleotide ligation sequencing.

Fiona C.L. Hyland, Rajesh Gottimukkala, Ryan Koehler, Susan Tang, Eric Tsung, Heather Peckham, Kevin McKernan, Francisco De La Vega.

Applied Biosystems, Foster City, CA and Beverly, MA, U.S.A.

With the advent of next-generation sequencing, novel algorithms to detect heterozygous and homozygous variants from massive amounts of short-read data are needed. The dibase-coded oligonuclotides used by the Applied Biosystems SOLiD™ System facilitate removal of sequencing errors, and so allow sensitive and specific heterozygote detection at low coverage. We have developed a Bayesian algorithm, coupled with tunable filters, to detect SNPs on SOLiD data, incorporating a comprehensive error model including prior probabilities of population heterozygosity.

We evaluated the sensitivity and specificity of our SNP algorithm with sequence data from whole-genome sequencing of an African human at 18x coverage. Comparing to HapMap genotypes, we call 90% of heterozygotes at 10x coverage, and more than 99% of heterozygotes at positions with >20x coverage. The heterozygote false discovery rate for HapMap SNPs is 0.1%. Using information about known SNPs could further increase SNP detection at lower coverage while maintaining a very low false positive rate.

We developed an algorithm to detect copy number variation in a single sample. In contrast to array methods, with sequencing, genomic coverage data is available at single base resolution. To detect copy number, we calculate coverage in variable-sized genomic windows that are selected to contain a constant number of mappable positions. Within these windows, we normalize coverage based on predicted mappability and GC content; this is analogous to the array CGH approach of normalizing based on intensity ratio using a matched sample. We then use a hidden markov model for segmentation, and we apply empirically derived filters to the contiguous segments to call copy number variants. Depending on the size of the window and of the CNVs we try to detect, up to 97% of the CNVs we detect are in the Toronto database, suggesting a very high true positive rate.

To detect small InDels, we identify paired-end reads for which one of the reads aligns and the other does not. We realign these pairs using the anchored pair as a seed and performed a more aggressive alignment with the other tag in a several kb window (depending on the insert size of the library) around the anchored mate. Using the unanchored tag we start aligning both ends of the read until the number of allowed mismatches occurs. Disallowing for indels within 3 bases of either end of the read, we identify if we are able to piece together both ends only allowing for a single gap of a maximum size inserted or deleted (maximum size is determined by the read length). We subsequently apply additional filters to refine the set of candidate InDels. Validation of InDels with Sanger sequencing suggests a true positive rate of above 95%.

Detecting Copy Number Variation with Mated Short Reads

Paul Medvedev[1], Marc Fiume[1], Tim Smith[1], Adrian Dalca[1], Seunghak Lee[1], Michael Brudno[1,2]

[1] Department of Computer Science, University of Toronto, Canada
[2] Banting and Best Department of Medical Research, University of Toronto, Canada

1 Background

Methods for CNV detection have until recently been based on whole-genome array comparative genome hybridization (aCGH), which tests the relative frequencies of a probe DNA segment between two genomes (Kallioniemi et al. 1992, Lucito et al. 2003). Due to the inherently noisy nature of microarray experiments, they require multiple probes within a CNV before the variant can be called. While computational methods based on aCGH data have been successfully used to identify CNVs, their power is limited. The number of probes, and hence the resolution (width) of any prediction is limited by the density of the array: the size of predicted variations from a single array is at least 40Kbp, while actual CNV regions can be much shorter.

The advent of NGS technologies has spurred new methods for CNV discovery. Recent studies of cancer cells have demonstrated that by considering the depth-of-coverage of mapped reads along a reference genome, it is possible to detect changes in copy number (Campbell et al. 2008; Chiang et al. 2009). Alternatively, paired-end mapping (PEM) techniques have been used to detect structural variants with NGS data (Korbel et al. 2007; Lee et al. 2008; Bentley et al. 2008; Hormozdiari et al. 2009, Lee et al. 2009). In this approach, two paired reads (called *matepairs*) are generated at an approximately known distance in the donor genome. The reads are mapped to a reference genome, and matepairs mapping at a distance significantly different from the expected length (termed *discordant*) suggest structural variants.

2 Methods

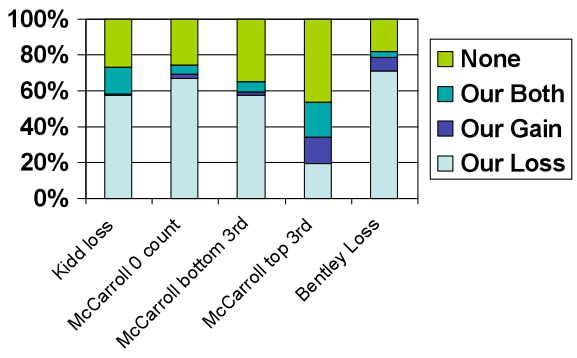
We build upon the earlier work of Chiang et al. (2009) and Campbell et al. (2008) by developing a CNV detection method that supplements depth-of-coverage with paired-end mapping (PEM) information. To do this, we unify and extend the previous PEM approaches of Tuzun et al. (2005), Cooper et al. (2008), and Korbel et al. (2007) into a method of mining discordant matepairs which specifically targets CNV detection. The depth-of-coverage, PEM, and reference genome are all integrated within a novel computational framework, which we term the "donor graph"—an extension of the repeat graphs commonly used for genome assembly and alignment (Pevzner et al. 2004, Raphael et al. 2004, Paten et al. 2008). By using techniques from the theory of network flows, we are able to search the donor graph for a genome which closely matches both the observed discordant matepairs and the depth-of-coverage, on a genome-wide scale.

Our approach is able to overcome some of the obstacles which have previously limited the use of PEM for CNV prediction. First, the insert size does not limit the size of the variants we can detect, as it has in many previous approaches (Tuzun et al. 2005, Kidd et al. 2008, Bentley et al.

2008). Additionally, since our algorithm requires a combination of support from coverage depth and discordant matepairs for making a call, we do not rely on having uniquely best read mappings, a limitation that has made it difficult for previous PEM techniques to detect variation within regions of segmental duplications (Tuzun et al. 2005, Korbel et al. 2007, Kidd et al. 2008, Bentley et al. 2008). Our approach is able to use multiple good mappings, since spurious discordant matepairs mappings without support from depth-of-coverage are ignored by our algorithm. Finally, unlike previous depth-of-coverage approaches, we do not partition the genome into small fixed size windows. Instead, we use the breakpoints detected by PEM analysis to delineate the windows used for calculating coverage depth. This allows us to have much larger windows, and therefore mitigate the sequencing biases that cause uneven local coverage.

3 Experimental Results

We use our method to detect CNVs within an individual (NA18507) using a dataset with ~36bp long reads, an insert size of ~208bp, and ~41x coverage. We make 4114 gain calls, and 5795 loss calls (9909 total). Most of the calls (87%) coincide with previously known variants within the human population (via the Database of Genomic Variants). To verify the sensitivity, we compared against Kidd et al.'s losses (146 calls), McCarroll et al.'s annotations with zero copy-count in NA18507 (39 calls), those for which it is strictly in the bottom third percentile of the 256 samples (94 calls) or strictly in the top third (26 calls), and Bentley et al.'s loss calls (2289). The chart shows the percentage of given calls that overlap just our loss, just our gain, both our loss and gain, and neither our loss and gain.



References

David R. Bentley et al. 2008. Nature 456: 53-59. Chiang et al. 2009. *Nat. Methods* 6: 99-103.
Fereydoun Hormozdiari, Can Alkan, Evan E. Eichler, S. Cenk Sahinalp, RECOMB 2009
Campbell, P. et al. 2008. *Nat. Genet*. advanced online publication: 722-729.
Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E., and Nickerson, D. A. 2008. *Nat. Genet.* **40:** 1199-1203.
Kallioniemi, A., Kallioniemi, O. P., Sudar, et al. 1992. *Science* **258:** 818-821.
Kidd, J. M., Cooper, G. M., Donahue, et al. 2008. *Nature* **453:** 56-64.
Korbel, J. O., Urban, A. E., Affourtit, J. P., et al. 2007. *Science* **318:** 420-426.
Lee, S., Cheran, E., and Brudno, M. 2008. *Bioinformatics* **24:** i59-67.
Lee, S., Hormozdiari, F., Alkan, C., and Brudno, M. 2009. *Nat Meth* advanced online publication
Lucito, R., Healy, J., Alexander, et al. 2003. *Genome Res.* **13:** 2291-2305.
Paten, B. et al. 2008. *Genome Research* 18: 1814-1828.
McCarroll, S. A., Kuruvilla, F. G., Korn, et al. 2008 *Nat. Genet.* **40:** 1166-1174.
Pevzner, P. A. et al. 2004. *Genome Res*. 14: 1786-1796.
Raphael, B., Zhi, D., Tang, H., and Pevzner, P. 2004. Genome Research 14: 2336-2346.
Tuzun, E., Sharp, A. J., Bailey, J. A., et al. 2005. *Nat. Genet.* **37:** 727-732.

A method for detecting small scale human microsatellite length- polymorphism using Solexa/Illumina paired-end sequencing data

Weldon Whitener[1], Avril Coghlan[1,2], Li Heng[1], Richard Durbin[1]

[1]Wellcome Trust Sanger Institute, Cambridge UK, [2]University of Cork, Ireland

Microsatellites are common repetitive sequence motifs in the human genome that are known to be subject to high mutation rates; however, previous studies have only considered a small minority of microsatellites in the human genome. To address this, we have developed a method to use deep Solexa/Illumina sequencing data to identify microsatellites that differ in length between an individual's genome and the human reference genome.

Solexa/Illumina sequencing technology produces short reads that often do not span larger microsatellites, making direct calls on microsatellite length unfeasible. However, the paired-end reads mapping information can be used as the reads are separated by a known distance determined by the insert size of the sequencing library (IS). Sequenced reads are mapped using MAQ (Li, Ruan and Durbin, Genome Research 2008, http://maq.sourceforge.net) to the human reference. For each microsatellite in the reference genome, mapping distances (MD) of spanning paired-end reads are stored. Reference alleles will have a matching distribution for MD compared to IS, while deletion alleles will tend to have a right shifted distribution for MD compared to IS, and insertion alleles a left shifted distribution for MD compared to IS. Based on this, we have developed a Bayesian approach to determine whether the distribution of MD of alleles at a particular microsatellite locus is statistically different than the distribution of IS.

We used this approach to analyse publicly available deep whole-genome Solexa/Illumina sequencing data for a single individual, HapMap individual NA18507 (Bentley et al, Nature 2008). Our algorithm successfully detected repeat loci that are known (using an independent approach based on Sanger sequencing) to be longer or shorter compared to the reference genome. We are in the process of applying the method to data from the 1000 Genomes Project, and are also exploring extending the power of the method by using information from "hanging" reads; reads where one end maps to flanking unique sequence and the other into the repetitive microsatellite sequence.

MoDIL: Detecting Small INDELs from Clone-end Sequencing with Mixtures of Distributions

Seunghak Lee[1], Fereydoun Hormozdiari[2], Can Alkan[3], Michael Brudno[1,4]

Human genetic variation comes in a wide range of sizes - from SNPs and very small (single nucleotide) insertions/deletions (indels) to large-scale "structural" variations, where kilo base pairs of the genome are inserted, deleted, inverted, or duplicated. Several methods for the identification of both small scale variants (SNPs and insertions/deletions < 10bp) [2, 3, 8] and large scale ones (inversions and large insertions/deletions) [10, 6, 7, 5, 4] have been developed, and their discovery and cataloguing is well underway. Simultaneously, one would expect there should also be a large amount of "medium-sized" variation: insertions and deletions of 10 to 50 nucleotides. Currently, however, there are no methods, either computational or wet lab, for high-throughput detection of these medium sized polymorphisms. In this work we develop a method to find these variants (10-50bp) by relying on the high clone coverage of many NGS datasets. While deviation by several of these clones from the expected insert size is to be expected, a deviation by a very large number (even by a small amount) would be indicative of an insertion or deletion. Here, we show a rigorous method for identifying and analyzing indels, especially medium sized indels using this intuition.

Algorithm for Indel Detection

Our algorithm is given a list of mapping of all matepairs to the reference genome. Each matepair consists of two reads, and the distance between them is referred to as the mapped distance. Unlike previous methods [10, 7, 4], which identify structural variations based on a pre-specified number of matepairs with a mapped distance significantly different from the insert size, our algorithm operates on the whole distribution of the observed mapped distances, rather than the outliers. Fig. 1 illustrates the essence of our method. We take advantage of the distribution of insert sizes in the sequenced library (we call this distribution $p(Y)$, and estimate it from all mapped distances, genome wide). Given a particular genomic location i, we define the corresponding cluster C_i as all of the matepairs that overlap the location (the left read of the pair is to the left of the position, and the right read is to the right). We will call the distribution of the observed mapped distances in a cluster $p(C_i)$. A cluster that comes from a location that has no indel polymorphisms will have mapped distances that follow the same distribution as the size of the clones in the library ($p(C_i)$ and $p(Y)$ will be identically distributed). This is illustrated (for our dataset) in Fig. 1a. If, on the other hand, there has been a homozygous insertion or deletion at this location, the distribution $p(C_i)$ will shift (Fig. 1b). In case of heterozygous indel, we will observe that $p(C_i)$ consists of two distributions, shifted and non-shifted $p(Y)$s (Fig. 1c). Furthermore, the size of the indel event can be estimated with high confidence; its expected size follows a Gaussian distribution with mean $\mu = \mu_{p(y)} - \mu_{p(C_i)}$ and $\sigma = \sigma_{p(y)} / \sqrt{n}$, where n is the number of matepairs in the cluster. Note that expected size of indel is normally distributed regardless of observed distribution of mapped

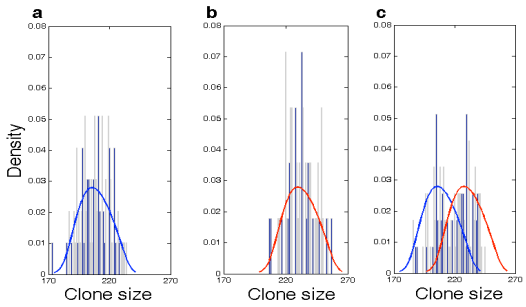


Figure 1: Observed distribution of mapped distances within a cluster for cases of (a) no indel, (b) homozygous deletion and (c) heterozygous deletion, superimposed on distribution of insert sizes, $p(Y)$

[1] Department of Computer Science, University of Toronto, Canada.
[2] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada.
[3] Department of Genome Sciences, University of Washington and the Howard Hughes Medical Institute, Seattle, WA, USA.
[4] Banting and Best Dept. of Medical Research, University of Toronto, Canada.

distances according to the central limit theorem. As the number of matepairs in each cluster grows, our confidence in the size of the indel will increase, allowing for the prediction of progressively smaller indels with higher coverage.

In order to estimate indel sizes, we model the random variable of the expected size of indel with two random variables, one for each haplotype. We call this probabilistic model *Mixture of Distributions* (MoD) since the observed distribution will be a mixture of distributions of two random variables. Given a cluster, we identify the two distributions which have the fixed shape of $p(Y)$ and arbitrary means that best fit the observed data using the Kolmogorov-Smirnov goodness of fit test [9]. The means of the two distributions are found using the Expectation-Maximization algorithm, while appropriate Bayesian priors are used to prevent over-fitting.

Results

We have applied our model to the whole genome shotgun reads generated by Illumina for the Yoruban HapMap individual NA18507 [1]. This data provided 40x read coverage and 120x clone coverage for the NCBI reference genome build 35. The dataset was mapped to the reference NCBI human genome with mrFAST alignment tool [4]. The resulting dataset had 292,190,651 templates, giving us 20x clone coverage of the genome. We required each cluster to have at least 20 matepairs to minimize false positives. Using our approach we discovered 2,529 insertions 7,716 deletions in the Yoruban individual genome relative to the NCBI reference genome (False Discovery Rate < 0.0001). These ranged in size from 5 to 119 nucleotides for insertions, and 5 to 334,646 nucleotides for deletions. In validation of our results, as shown in Table 1, we observed significant overlap between our indel calls and the short indels discovered by Kidd et al. with the same individual and Mills et al. with 36 human genomes. Our experimental results demonstrate that MoDIL can identify, with high sensitivity, indels ≥ 20bp, while accurately estimating the true size of the variants. The size correlation of overlapping indels (20-500bp) between Mills et al. results and our indel calls was very strong ($r^2 = 0.96$).

Table 1: Overlap between the predictions of the MoDIL algorithm and the short indels discovered by Kidd et al. and Mills et al. The fraction of Kidd et al. indels present in our results indicates a low False Negative Rate (FNR) for our algorithm for indels ≥ 20bp, and lower sensitivity for shorter indels. The large amount of overlap between our results and the Mills et al. data over all indel sizes indicates a strong correlation between them.

		MoDIL	Kidd et al.			Mills et al.		
Length	Type	Total	Total	Found	FNR	Total	Overlapping	% Overlap
>=20bp	INS	1,607	101	91	0.10	6,240	260	0.16
	DEL	3,646	244	231	0.05	10,742	701	0.19
15 – 19bp	INS	592	124	57	0.54	3,096	115	0.19
	DEL	2,584	183	109	0.40	3,698	162	0.06
10 – 14bp	INS	257	373	56	0.85	8,615	125	0.49
	DEL	1,157	601	171	0.72	9,990	237	0.20

References

[1] D. R. Bentley et al. Nature, 456(7218):53–59, November 2008.
[2] K. Chen et al. Genome Res, 17(5):659–666, May 2007.
[3] L. W. Hillier et al. Nature Methods, 5(2):183, 2008.
[4] F. Hormozdiari, C. Alkan, E. E. Eichler, and S. Sahinalp. RECOMB 2009
[5] J. M. Kidd et al. Nature, 453(7191):56–64, 2008.
[6] J. O. Korbel et al. Science, 318(5849):420, 2007.
[7] S. Lee, E. Cheran, and M. Brudno. Bioinformatics, 24(13):i59–i67, 2008.
[8] H. Li, J. Ruan, and R. Durbin. Genome Research, 18(11):1851, 2008.
[9] F.J. Massey. Journal of the American Statistical Association, 46(253):68–78, 1951.
[10] E. Tuzun et al. Nature Genetics, 37:727–732, 2005.

TopHat: discovering splice junctions with RNA-Seq

Cole Trapnell[1], Lior Pachter[2] and Steven L. Salzberg[1]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742 and [2]Department of Mathematics, University of California, Berkeley, CA 94720, USA

Motivation: A new protocol for sequencing the messenger RNA in a cell, known as RNA-Seq, generates millions of short sequence fragments in a single run. These fragments, or 'reads', can be used to measure levels of gene expression and to identify novel splice variants of genes. However, current software for aligning RNA-Seq data to a genome relies on known splice junctions and cannot identify novel ones. TopHat is an efficient read-mapping algorithm designed to align reads from an RNA-Seq experiment to a reference genome without relying on known splice sites.

Results: We mapped the RNA-Seq reads from a recent mammalian RNA-Seq experiment and recovered more than 72% of the splice junctions reported by the annotation-based software from that study, along with nearly 20 000 previously unreported junctions. The TopHat pipeline is much faster than previous systems, mapping nearly 2.2 million reads per CPU hour, which is sufficient to process an entire RNA-Seq experiment in less than a day on a standard desktop computer. We describe several challenges unique to *ab initio* splice site discovery from RNA-Seq reads that will require further algorithm development.

Availability: TopHat is free, open-source software available from http://tophat.cbcb.umd.edu

Quantitative Detection of Alternative Transcripts with RNA-Seq Data

Regina Bohnert[1], Jonas Behr[1], and Gunnar Rätsch[1*]

[1] Friedrich Miescher Laboratory of the Max Planck Society, Spemannstraße 39, 72076 Tübingen, Germany

*To whom correspondence should be addressed: Gunnar.Raetsch@tuebingen.mpg.de

Background

Novel high-throughput sequencing technologies open exciting new approaches to profiling of transcriptomes harboring alternative transcripts. Sequencing transcript populations of interest, e.g. from different tissues or variable stress conditions, with RNA sequencing (RNA-Seq) [1] generates millions of short reads. Accurately aligned to a reference genome, they provide digital counts and thus facilitate transcript quantification. As the observed read counts only provide the summation of all expressed sequences at one locus, the inference of the underlying transcript abundances is crucial for further quantitative analyses and for the identification of alternative transcripts along with their relative expression level.

Results

To approach this problem, we have developed a new technique based on linear programming (LP). Given an annotation of (alternative) transcripts and position-wise exon/intron read coverages from read alignments, we determine the abundances for each annotated transcript by minimizing an appropriate loss function. It penalizes the deviation of the observed from the expected read coverage within a segment or at each covered nucleotide given the transcript weights. The observed read coverage is typically non-uniformly distributed over the transcript due to several biases in the processing steps to obtain the sequencing libraries and the sequencing. This leads to distortions of the transcript abundances, if not corrected properly. We therefore extended our approach to jointly optimize transcript profiles in order to model the coverage deviations depending on the position in the transcript (cf. Figure 1). Our method can be applied without knowledge of the underlying transcript abundances and equally benefits from loci with and without alternative transcripts.
To quantitatively evaluate the quality of our abundance predictions, we used a set of simulated reads from transcripts with known expression as a benchmark set. It was generated using the Flux Simulator [2] modeling biases in RNA-Seq as well as preparation experiments. Table 1 shows preliminary results with segment and position based loss as well as with and without the transcript profiles. Our results indicate that the position-based modeling together with transcript profiles allows us to accurately infer the underlying expression of single transcripts as well as of multiple isoforms of one gene locus, only given a transcript annotation.

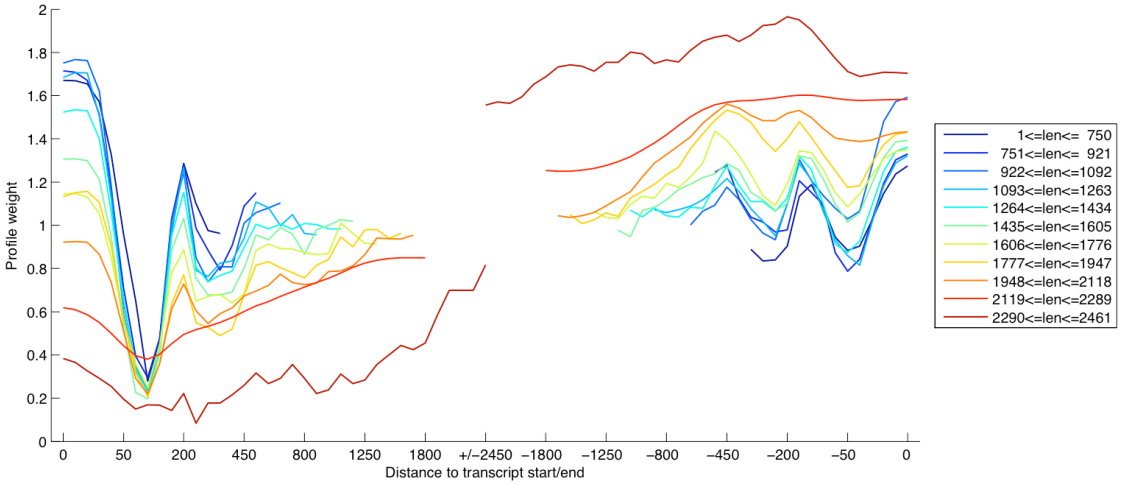


Figure 1: Expression profiles obtained using our approach for different transcript lengths. The x-axis shows the absolute distances to the start and end of the transcript, respectively. We used 11 bins pooling transcripts of similar lengths; color encoding as indicated in the panel to the right. The read data was generated based on the gene annotation TAIR 7 of the model plant organism *A. thaliana* [3], simulating reverse transcription of the RNA population, poly-dT priming and subsequent physical shearing using the Flux Simulator [2].

Table 1: Correlation of underlying expression level and inferred abundances for different approaches. Our proposed approach, which infers transcript abundances from read data at each position, is compared against a segment-based approach, which uses averages of averaged read counts at shared transcript segments. The Spearman correlation between true and inferred abundance was determined across all annotated transcripts; for alternatively annotated genes, the average of correlation within transcripts of each gene was calculated. We compare against not optimizing the transcript profiles (i.e. uniform profiles). We use the same data as in Figure 1.

Approach	Spearman correlation	
	Across genes	Within genes (mean)
Position-wise inference with transcript profiles	0.820	0.635
Segment-wise inference with transcript profiles	0.693	0.488
Position-wise inference without transcript profiles	0.684	0.540
Segment-wise inference without transcript profiles	0.580	0.367

Discussion/Conclusions

Our preliminary results show that modeling the transcript profiles can significantly improve the accuracy of transcript abundance estimates from RNA-Seq data. However, the described and other recent approaches [4] for transcript quantification with RNA-Seq rely on annotated gene structures. As most genome annotations are incomplete, they cannot reveal and quantify novel and also (novel) alternative transcripts. Nevertheless, our method can be extended to simultaneously predict the structures of genes and their transcript abundances by combining it with a recent extension [5] of the gene finding system mGene [6]. Here, we repeatedly call mGene to suggest refined transcript structures that fit to the observed read counts. Using this approach we can simultaneously predict sets of transcripts including their abundances that jointly explain the observed read coverages.

Revealing and quantifying novel alternative transcripts with the powerful tool of RNA-Seq will be a fundamental step towards a deeper understanding of RNA transcript regulation.

References

1. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature Reviews Genetics* 2009, 10:57-63.
2. Sammeth M: **Flux Simulator.** 2009 [http://flux.sammeth.net/simulator.html].
3. **TAIR 7.** 2009 [http://www.arabidopsis.org/]
4. Jiang H, Wong WH: **Statistical Inferences for Isoform Expression in RNA-Seq.** *Bioinformatics* 2009.
5. Behr J, Schweikert G, Cao J, Bona FD, Zeller G, Laubinger S, Ossowski S, Schneeberger K, Weigel D, Rätsch G: **RNA-Seq and Tiling Arrays For Improved Gene Finding.** *Genome Informatics Talk* 2008.
6. Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong C, Philips P, De Bona F, Hartmann L, Bohlen A, Krüger N, Sonnenburg S, Rätsch G: **mGene: Accurate SVM-based Gene Finding with an Application to Nematode Genomes.** Submitted to *Genome Research* 2009.

MapSplice: Map RNA-seq Short Reads for Splice Junction Discovery

Jinze Liu[1], Kai Wang[1], Zheng Zeng[1], Stephen J. Coleman[2], James N. MacLeod[2] , Jan Prins[3]
[1]Department of Computer Science, University of Kentucky, Lexington, KY, 40506, USA
[2]Maxwell H. Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, KY, 40546, USA
[3]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-3175.

Next-generation deep sequencing technologies provide new opportunities to characterize the transcriptome with unprecedented resolution. However, mapping hundreds of millions of shotgun cDNA reads to a reference genome becomes a challenging computational problem. Our specific interest is the mapping of mRNA splice junctions. Reads crossing splice junctions are critical in the identification of transcript isoforms, but their mapping is more problematic than exonic reads since their alignment to the genome requires a gap. We propose a novel and efficient algorithm MapSplice for this task. The algorithm requires only RNA-Seq reads and a reference genome as input. It scans the genome in a sliding window fashion, detecting prefix and corresponding suffix matches in separated locations along the genome for all reads simultaneously. MapSplice performs approximate matching to tolerate errors. Splice junctions are determined according to the distribution of reads at the putative prefix/suffix sequence boundaries flanking a gap. At each splice junction, relative expression level is reflected by the number of the tags with the specific gap identity. MapSplice is completely unsupervised and data driven, unlike QPalma (Bona et.al. 2008) which incorporates a supervised splice site prediction technique to bias the alignment towards sites that are more likely to be exon boundaries. MapSplice does not rely on EST/cDNA databases or on known splice databases, including synthetic junctions generated from existing genome annotations. Compared with Tophat (Trapnell et.al. 2009) which exhaustively enumerates candidate splice junctions by identifying putative introns based on the canonical GT-AG dinucleotide sequence at the potential splice donor and splice acceptor sites, MapSplice can also find novel splice sites with non-canonical nucleotide sequences at intron ends. In initial applications, MapSplice performed approximately 40 times faster than Tophat with typical parameter setting. Performance scales linearly with genome size, number of reads, and increasing maximum splice length.

De novo Transcriptome Assembly with ABySS

İnanç Birol[1,*], Shaun D Jackman[1], Cydney Nielsen[1], Jenny Q Qian[1], Richard Varhol[1], Greg Stazyk[1], Ryan D Morin[1], Yongjun Zhao[1], Martin Hirst[1], Jacqueline E Schein[1], Doug E Horsman[2], Joseph M Connors[2], Randy D Gascoyne[2], Marco A Marra[1] and Steven JM Jones[1]

[1] Genome Sciences Centre, 100-570 W 7th Avenue, Vancouver BC V5Z 4S6, Canada, www.bcgsc.ca

[2] British Columbia Cancer Agency, 600 West 10th Avenue, Vancouver, BC V5Z 4E6, Canada, www.bccancer.ca

ABSTRACT

Motivation: Whole transcriptome shotgun sequencing data from non-normalized samples offer unique opportunities to study the metabolic states of organisms. One can deduce gene expression levels using sequence coverage as a surrogate, identify coding changes or discover novel isoforms or transcripts. Especially for discovery of novel events, *de novo* assembly of transcriptomes is desirable.

Results: Transcriptome from tumor tissue of a patient with follicular lymphoma was sequenced with 36 base-pair (bp) single- and paired-end reads on the Illumina Genome Analyzer II platform. We assembled approximately 194 million reads using ABySS into 66,921 contigs 100bp or longer, with a maximum contig length of 10,951bp, representing over 30 million base pairs of unique transcriptome sequence, or roughly 1% of the genome.

Availability and Implementation: Source code and binaries of ABySS are freely available for download at http://www.bcgsc.ca/platform/bioinfo/software/abyss. Assembler tool is implemented in C++. The parallel version uses open mpi. Explorer tool is implemented in Java using the Java universal network/graph framework.

Contact: Software help: abyss@bcgsc.ca, authors {ibirol, sjackman, cydneyn, jqian, sjones}@bcgsc.ca

Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data

Chol-Hee Jung, Martin A. Hansen, Igor V. Makunin, Darren Korbie and John S. Mattick

Institute for Molecular Bioscience, University of Queensland, St Lucia QLD 4072, Australia

ABSTRACT

The increasing interest in small non-coding RNAs (ncRNAs) such as microRNAs (miRNAs), small interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs) and recent advances in sequencing technology have yielded large numbers of short (18~32nt) RNA sequences from different organisms, some of which are derived from small nucleolar RNAs (snoRNAs) and transfer RNAs (tRNAs). We observed that these short ncRNAs frequently cover the entire length of annotated snoRNAs or tRNAs, which suggests that other loci specifying similar ncRNAs can be identified by clusters of short RNA sequences.

We combined publicly available datasets of tens of millions of short RNA sequence tags from *Drosophila melanogaster*, and mapped them to the *Drosophila* genome. Approximately 6 million perfectly mapping sequence tags were then assembled into 521,302 tag-contigs (TCs) based on tag overlap. Most transposon-derived sequences, exons and annotated miRNAs, tRNAs and snoRNAs are detected by TCs, which show distinct patterns of length and tag-depth for different categories. The typical length and tag-depth of snoRNA-derived TCs was used to predict 8 previously unrecognized box H/ACA and 26 box C/D snoRNA candidates. We also identified 86 loci with a high number of tags that are yet to be annotated, 7 of which have a particular 18mer motif and are located in introns of genes involved in development. A subset of new snoRNA candidates and putative ncRNA candidates was verified by Northern blot.

In this study, we have introduced a new approach to identify new members of known classes of ncRNAs based on the features of TCs corresponding to known ncRNAs. We also used this approach for the identification of putative novel ncRNAs. A large number of the identified TCs are yet to be examined experimentally suggesting that many more novel ncRNAs remain to be discovered.

Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly

Bas E. Dutilh[1], Martijn A. Huynen[1] and Marc Strous[2]

[1]Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Center, Geert Grooteplein 28, 6525 GA, Nijmegen, The Netherlands
[2]Department of Microbiology, Radboud University Nijmegen, Heyendaalsweg 135, 6525 AJ, Nijmegen, The Netherlands; MPI for Marine Microbiology, Celsiusstr. 1 D-28359, Bremen, Germany; Centre for Biotechnology, University of Bielefeld, Germany

Motivation: Most microbial species can not be cultured in the lab. Metagenomic sequencing may still yield a complete genome if the sequenced community is enriched and the sequencing coverage is high. However, the complexity in a natural population may cause the enrichment culture to contain multiple related strains. This diversity can confound existing strict assembly programs and lead to a fragmented assembly, which is unnecessary if we have a related reference genome available that can function as a scaffold.

Results: Here, we map short metagenomic sequencing reads from a population of strains to a related reference genome, and compose a genome that captures the consensus of the population's sequences. We show that by iteration of the mapping and assembly procedure, the coverage increases while the similarity with the reference genome decreases. This indicates that the assembly becomes less dependent on the reference genome and approaches the consensus genome of the multi-strain population.

LOCAS: A new low coverage assembler for short reads

Juliane D. Klein[1], Stephan Ossowski[2], Korbinian Schneeberger[2], Detlef Weigel[2] and Daniel H. Huson[1]
[1] University of Tübingen, ZBIT Center of Bioinformatics [2] Max-Planck Institute for Developmental Biology, Tübingen

With the continuing improvement of second generation sequencing technologies, the possible applications in the area of resequencing and polymorphism detection are steadily increasing. The new technologies deliver ever increasing amounts of sequence, albeit at short read lengths. To address the arising computational questions, existing algorithms are being adjusted and new algorithms are being developed.

In a typical resequencing project reads are mapped onto a closely related reference genome. Then, a consensus from the mapped reads is calculated as an approximation of the new genome sequence. This consensus often covers much of the new sequence, even with low coverage data. Consequently, in resequencing projects often low coverage, less than 20x, is used.
To additionally cover highly polymorphic regions or insert sites of the new sequence it is necessary to incorporate unmapped reads. Therefore, an assembly tool is required that reassembles mapped regions taking unmapped reads into account. Here, existing short read assemblers such as VELVET [1] can be used to a degree. However, such tools are not designed for low coverage data. Also, the incorporation of left-over reads, which are reads where both mates do not map to the reference, is computationally expensive with these assemblers. Thus, a short read assembler for low coverage is needed, which provides a time efficient solution to incorporate left over reads.

We have developed a new assembly tool called LOCAS (LOw Coverage Assembly Software). It explicitly handles low coverage data by allowing mismatches in the overlap alignment of reads. Further, it provides some additional features for resequencing projects. It takes advantage of given mapping positions by either validating or shifting these positions to improve the mapping. An additional module of LOCAS can also incorporate left-over reads in a time efficient manner. LOCAS operates in the following steps:

1. Detect overlaps
2. Build overlap alignment graph
3. Reduction, solving sequencing errors and repeats
4. Path extraction and consensus determination

The module for left-over reads builds an alignment graph for all left-over reads in a preprocessing step. Then the above mentioned workflow is started for each pair of consecutively mapped regions and the program attempts to merge a subgraph of the left-over graph into the local assemblies.
LOCAS is being developed in the context of the "1001 Genomes" resequencing project on *Arabidopsis thaliana* and is part of the resequencing pipeline SHORE [2].

To study the performance of LOCAS in comparison to VELVET, we simulated reads from the first chromosome of *Arabidopsis thaliana*, using an error model of the Illumina sequencing technique. We randomly chose 500 regions of length 20,000 and for each we simulated reads at coverages from 5x to 15x. For each region the corresponding reads were given as input to the two assembly tools. For very low coverage data sets, 5x, 7.5x and 12x, LOCAS covered substantially more of the original sequence than VELVET, at the same level of accuracy. For higher coverages, 12.5x and 15x, LOCAS had a longer mean contig size than Velvet, also at the same level of accuracy.

1 Velvet: Algorithms for de novo short read assembly using de Bruijn graphs, Zerbino, D.R. and Birney, E., Genome Research 18(821), 2008
2 Sequencing of natural strains of Arabidopsis thaliana with short reads, Ossowski, S. and Schneeberger, K. and Clark, R.M. and Lanz, C. and Warthmann, N. and Weigel, D., Genome Research 18(2024), 2008

Poisson Model of Significance for Short Reads Concentrations

Adam Kowalczyk [1,2] , Thomas Conway [1,3] , Bryan Beresford-Smith [1,2] , Sibgat Choudhury [5] , Saraswati Sukumar [6] , Kornelia Polyak [5] and Izhak Haviv [4,7]
[1] NICTA, Victoria Research Laboratory & [2] Department of Electrical and Electronic Engineering & [3] Department of Computer Science and Software Engineering & [4] Department of Biochemistry, School of Medicine, The University of Melbourne, Parkville, Victoria, Australia; [5] Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115; [6] Department of Oncology, Johns Hopkins University, Baltimore, MD 21231; [7] Baker IDI, Melbourne, Australia.
Contact : adam.kowalczyk@nicta.coma.au

The main challenge in the analysis of modern genomic data stems from very high dimensionality and very small sample size (number of observations). Although such challenges have been recognised since at least the 1960's (see the works of J. Tuckey [12] and more recently the account of mathematical challenges for the 21st century by D. Donoho [3]), the practical and theoretical solutions for answering it are still in their infancy. The immediate impact on genomic data analysis is felt in terms of severe limitation on the statistical significance which can be claimed for detected subsets of features, in particular the differential features between phenotypes of interest. In the statistical setting the problem is cast as multiple hypothesis testing, and has been studied by statisticians for a long time. However, the relentless emergence of novel genomic wet-lab technologies in the last ten years pushes the dimensionality bar higher and higher. The problem is well documented for the analysis of microarray data (see for instance the overview [5] and [4]). In the case of next generation sequencing (NGS) the problem becomes even worse as the dimensionality in the number of features (peaks) to choose from is measured not in the thousands but in the millions or more. More specifically, a number of algorithms for controlling the Family-wise Error Rate (FWER), such as Bonferroni , ˘Sid ´ak , Holm [8], Simes [11] procedures or principled algorithms for controlling the false discovery rate of Benjamini & Hochberg [1] or Benjamini & Yekutieli & [2], rely on a series of univariate significance tests and adjusted p-values. In order to provide non-vacuous results, the extreme raw univariate p-values in the family of tests have to be of order 1 divided by the number of hypotheses tested (Bonferroni threshold), i.e. of the order of the inverse of the dimensionality of the data space. Under such a constraint, the raw p-values for NGS experiments have to be very low, of the order 10_{-6} . There is virtually no hope of satisfying such stringent requirements if classical univariate feature selection tests, such as the t-test or its various modifications, are applied to the sample sizes typically feasible in current NGS pilot studies (e.g. of order 10).
One way around this obstacle could be to develop "more practical" algorithms such as the local empirical Bayes approach of Efron [6, 7, 4] which resort to comparisons to more or less ad hoc "null" distributions. However, there is an alternative possibility, based on the usage of improved univariate test statistics utilising the digital nature of NGS signals. Examples of the successful application of such statistics can be found in recent publications [10, 9], where the fair coin tossing model is used, in a rather "ad hoc" fashion, to allocate p-values for pairs of counts. In this paper we develop a more principled approach using the following rationale. The mechanisms by which NG sequence data are generated are amenable to natural statistical modeling as sampling from a binomial distribution (sample dependent) or its Poisson approximation, as has already been observed in [10, 9]. The Poisson distribution is dependent on a single parameter (recall that the mean and variance of a Poisson distribution are equal), hence two samples from two distributions allow a meaningful test for the order of generating means. Furthermore, the unobservable Poisson rates involved in such a test can be eliminated by considering the worst case scenario, i.e. the most unfavorable values of rates which would generate the highest p-value. As it turns out, in the special case of comparison of two concentrations our method is numerically equivalent to tests used in [9, 10]. As has been shown in those papers, in some practical experiments of special interest to us, some of the estimated p-values are so low that the most conservative Bonferroni control of FWER [1, 5, 4] provides non-vacuous results. Additionally, we have developed an extension

of our algorithms to the general case of discrimination between two phenotypes using multiple samples. An experimental validation to date (using alternative PCR measurements) has confirmed that peaks selected by our techniques were indeed differential between phenotypes of interest represented by multiple samples.

Another advantage of Poisson models is the availability of closed-form expressions for the p-values. This allows for a principled analysis of the impact of the depth of sequencing on significance of the final results, with important practical consequences. Most NGS machines partition their sequencing capacity so multiple independent samples can be processed in a single run of the machine (e.g. “lanes” in the Solexa machines). By doing principled analysis, we can determine how many lanes are necessary to achieve a desired level of statistical significance. If we can demonstrate analytically that fewer lanes can be used, then the cost of the experiment can be reduced, or more samples can be included.

References

1. Y. Benjamini and Y. Hochberg, Control ling the false discovery rate: a practical and powerful approach to multiple testing, J. Roy. Statist. Soc. Ser. B 57 (1995), 289–300.
2. Y. Benjamini and D. Yekutieli, The control of the false discovery rate in multiple hypothesis testing under dependency, annals of statistics, Ann. of Statist. 29 (2001), 1165–1188.
3. D.L. Donoho, Aide-Memoire. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality., 2000, http://www-stat.stanford.edu/ donoho/Lectures/AMS2000/Curses.pdf.
4. S. Dudoit, H.N. Gilbert, and M.J. van der Laan, Resampling-based empirical bayes multiple testing procedures for control ling generalized tail probability and expected value error rates: Focus on the false discovery rate and simulation study, Biometrical Journal 50 (2008), 716–744.
5. S. Dudoit, J.P. Juliet Popper Shaffer, and J.C. Boldrick, Multiple hypothesis testing in microarray experiments, Statistical Science 18 (2003), 71 – 103.
6. B. Efron, Local False Discovery Rates., 2005, http://www-stat.stanford.edu/ ckirby/brad/papers/2005LocalFDR.pdf.
7. , Size, Power, and False Discovery Rates., 2006, http://www-stat.stanford.edu/ ckirby/brad/papers/2006Size.pdf.
8. S. Holm, A simple sequential ly rejective multiple test procedure, Scand. J. Statist. 6 (1979), 65–70.
9. D.A. Nix, S.J. Courdy, and K.M. Boucher, Empirical methods for control ling false positives and estimating confidence in chip-seq peaks, BMC Bioinformatics 9 (2008), 523.
10. J. Rozowsky, G. Euskirchen, R.K. Auerbach, Z.D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M.B. Gerstein, Peakseq enables systematic scoring of chip-seq experiments relative to controls, Nature Biotechnology 27 (2009), 6675.
11. R. J. Simes, An improved bonferroni procedure for multiple tests of significance, Biometrika 73 (1986), 751–4.
12. J.W. Tukey, Exploratory data analysis, Addison-Wesley, ISBN 0-201-07616-0, 1977.

Design of Association Studies with Pooled Next-Generation Sequencing Data

Su Yeon Kim, Yingrui Li, Yiran Guo, Ruiqiang Li, Torben Hansen, Oluf Pedersen, Jun Wang, and Rasmus Nielsen

Most of the common hereditary diseases in humans are complex and multifactorial. Large scale genome-wide association studies based on SNP genotyping, have only identified a small fraction of the heritable variation of these diseases. One explanation may be that many rare variants (a minor allele frequency, MAF > 5%), which are not included in the common genotyping platforms, may contribute a substantial portion of the genetic variation of these diseases. Next-generation sequencing, which would allow the analysis of rare variants, is now becoming so cheap that it provides a viable alternative to SNP genotyping. In this paper, we present cost-effective protocols for using next-generation sequencing in association mapping studies based on pooled and un-pooled samples, and identify optimal designs with respect to total number of individuals, number of individuals per pool, and the sequencing coverage. We perform a small empirical study to evaluate the pooling variance in a realistic setting where pooling is combined with exon capturing. To test for associations, we have developed a likelihood ratio statistic that accounts for the high error rate of next-generation sequencing data. We also perform extensive simulations to determine the power and accuracy of this method. Overall, our findings suggest that with a fixed cost, sequencing many individuals at a more shallow depth with larger pool size achieves higher power than sequencing a small number of individuals in higher depth with smaller pool size, even in the presence of high error rates. Our results provide guidelines for researchers who are developing association mapping studies based on next-generation sequencing.