RECOMB-seq 2013

The 3rd Satellite Workshop of RECOMB
on Massively Parallel Sequencing

April 11-12, Tsinghua University, Beijing

RECOMB-seq 2013

Book of Abstracts

Program Committee Chairs:

Haixu Tang, Indiana University, Bloomington

Tao Jiang, University of California, Riverside

Organizing Committee Chairs:

Xuegong Zhang, Tsinghua University

Rui Jiang, Tsinghua University

Bioinformatics Division / Center for Synthetic and Systems Biology

TNLIST, Tsinghua University

Preface

The recent revolution in sequencing technology has opened the door for myriads of new applications and bio-medical discoveries. Projects are under way to sequence thousands of individuals (*e.g.*, the 1000 Genome project), tens of thousands of vertebrate species (*e.g.*, the Genome10K project), and the whole microbial ecosystem that live in our bodies (*e.g.*, the Human Microbiome project). Simultaneously, the novelty and complexity of the data has highlighted the challenges and limitations of current methods. As the technology continues to evolve and approaches its third generation, the challenges facing the community are becoming increasingly computational.

The Annual RECOMB Satellite Workshop on Massively Parallel Sequencing (RECOMB-Seq) is an annual forum for exploring the computational aspect of research, development and novel applications of high-throughput sequencing in life sciences. It brings together researchers, professionals, and industrial practitioners for interaction and exchange of knowledge and ideas. The Third Annual RECOMB Satellite Workshop on Massively Parallel Sequencing, RECOMB-Seq'2013, was held in Beijing on April 11-12, 2013, immediately following the main RECOMB conference (on April 7-10).

A total of 41 papers were submitted to RECOMB-Seq'2013. These submissions came from countries/regions including China, Korea, France, Spain, Finland, Russia, Turkey, Australia, Canada, and USA. We assigned each paper to at least 3 members of the programme committee. Although not all members of the programme committee managed to review all the papers assigned to them, a total of 121 reviews were received. As a result, there were almost 3 reviews per paper on average, and 39 (out of 41) papers received at least 3 reviews.

A total of 23 papers were accepted for oral presentation. Out of these accepted submissions, the authors of 18 papers opted to publish their final papers in a special issue of *BMC Bioinformatics* and the authors of the other 5 papers opted to publish their papers elsewhere. In addition to the contributed papers, the scientific programme of RECOMB-Seq'2013 also included 3 keynote speeches and poster sessions. The abstracts of all accepted papers as well as the keynote talks are included in this book of abstracts.There is no doubt that the presentations covered a broad range of topics on high-throughput sequencing and its applications, and were of very high quality.

Lastly, we wish to express our gratitude to the authors of the submitted papers, the members of the programme committee and their sub-referees, the members of the organizing committee, the keynote speakers, our generous sponsors, and supporting organizations for making RECOMB-Seq'2013 a great success.

Haixu Tang
Tao Jiang
Rui Jiang
Xuegong Zhang
April 2013

Steering Committee

- S. Cenk Şahinalp, Simon Fraser University, Canada
- Michael Brudno, University of Toronto, Canada
- Inanc Birol, BC Genome Sciences Centre, Canada
- Eran Halperin, Tel-Aviv University, Israel
- Ben Raphael, Brown University, USA

Program Committee

- Haixu Tang, Indiana University, Bloomington, USA (chair)
- Tao Jiang, University of California, Riverside, USA (chair)
- Max Alekseyev, University of South Carolina, USA
- Can Alkan, Bilkent University, Turkey
- Vikas Bansal, Scripps Translational Science Institute, USA
- C. Titus Brown, Michigan State University, USA
- Xin Chen, Nanyang Technological University, Singapore
- Francis Chin, University of Hong Kong, China
- Jason Ernst, UCLA, USA
- Peilin Jia, Vanderbilt University, USA
- Rui Jiang, Tsinghua University, China
- Sun Kim, Seoul National University, Korea
- Ben Langmead, John Hopkins University, USA
- Jingchu Luo, Peking University, China
- Paul Medvedev, Pennsylvania State University, USA
- Satoru Miyano, The University of Tokyo, Japan
- Adam Phillippy, National Biodefense Analysis and Countermeasures Center, USA
- Mina Rho, Rosewell Park Cancer Institute, USA
- Jared Simpson, Wellcome Trust Sanger Institute, UK
- Jens Stoye, Bielefeld University, Germany
- Xiao Sun, Southeast University, China
- Glenn Tesler, University of California, San Diego, USA
- Todd Treangen, BNBI/NBACC, USA
- Chaochun Wei, Shanghai Jiao Tong University, China

- Shibu Yooseph, J. Craig Venter Institute, USA
- Alex Zelikovsky, Georgia State University, USA
- Daniel Zerbino, UC Santa Cruz, USA
- Shaojie Zhang, University of Central Florida, USA
- Xuegong Zhang, Tsinghua University, China
- Fangqing Zhao, Beijing Institute of Life Sciences, China
- Zhongming Zhao, Vanderbilt University, USA

Local Organization Committee Chairs

- Xuegong Zhang, Tsinghua University, China
- Rui Jiang, Tsinghua University, China

Table of Contents

RECOMB-seq 2013 Program

	Model	
16:10-16:30	Break	
16:30-17:50	**Session 4. Chair: Eleazar Eskin (UCLA, USA)**	
16:30-16:50	Alexandru I. Tomescu, Anna Kuosmanen, Romeo Rizzi and Veli Makinen *A Novel Min-Cost Flow Method for Estimating Transcript Expression with RNA-Seq*	18
16:50-17:10	Jing Zhang, C.-C. Jay Kuo and Liang Chen *WemIQ: a weighted-log-likelihood expectation maximization method for isoform quantification from RNA-Seq data*	19
17:10-17:30	Yi Li and Xiaohui Xie *A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues*	20
17:30-17:50	**Sponsored Talk** Speaker: **Haisheng Nie**, Illumina China Title: ***Illumina's comprehensive End-to-End solutions for Next-Generation Sequencing***	33
18:30-20:00	**Banquet (UNIS Center Hotel)**	

April 12 (Friday)

8:30-17:00	On-site Registration	
9:00-10:40	**Session 5. Chair: Haixu Tang (Indiana University, USA)**	
9:00-10:00	**Keynote Talk** Speaker: **Inna Dubchak**, DOE Joint Genome Institute and LBNL, USA Title: ***Efficient visualization of Next-Generation Sequencing data - are we there yet?***	8
10:00-10:20	Xi Wang and Murray J. Cairns *Gene Set Enrichment Analysis of RNA-Seq Data: Integrating Differential Expression and Splicing*	21
10:20-10:40	Jie Zhu, Yufang Qin, Taigang Liu, Xiaoqi Zheng and Jun Wang *Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles*	22
10:40-11:00	Break	
11:00-12:20	**Session 6. Chair: Gary Benson (Boston University, USA)**	
11:00-11:20	Seunghak Lee, Aurelie Lozano, Prabhanjan Kambadur and Eric Xing *Detecting SNP-SNP Interactions With Piecewise Independence Screening*	23
11:20-11:40	Yu Zhang *De Novo Inference of Stratification and Local Admixture in Sequencing Studies*	24
11:40-12:00	Eric Bareke, Jean-Francois Spinella, Ramon Vidal, Jasmine Healy, Daniel Sinnett and Miklos Csuros *Joint genotype inference with germline and somatic mutations*	25
12:00-13:30	Lunch Break and Poster Session	
13:30-14:00	**Poster Removal**	

Keynote Speakers

Colin Collins

Professor
Department of Urologic Science
University of British Columbia
Senior Research Scientist and Co-Director, LAGA
Vancouver Prostate Centre
Canada

Of Men, Mice and Prostate Cancer

Prostate cancer is the most common malignancy affecting men in the Western world and in 2012 accounted for 30,000 deaths in North American. In China its incidence is increasing rapidly. Genomics, and sequencing in particular, has been transformative for prostate cancer research. We, and our colleagues, have developed novel tools to interrogate sequence data to reveal high-resolution chromosome copy number profiles, accurate gene expression levels, differential alternative splicing patterns, single nucleotide variants and structural rearrangements including those that lead to the creation of fusion transcripts. Sequencing combined with bioinformatics enables integrated and global analyses in an unbiased manner or to focus, for example, on specific transcriptional complexes and signalling pathways such as the androgen receptor axis. As a consequence molecular subtypes have been defined based on the presence of various fusion genes, the mutational spectrum and copy number profiles. Prognostic biomarkers have been developed and, in some cases key variants such as gene fusions, mutations, and amplifications may be therapeutically actionable. At the same time we have developed a panel of patient derived tumour xenografts spanning most subtypes of prostate cancer. The confluence of these models with sequencing and bioinformatics is allowing us to dissect pathways involved in the progression to the lethal phenotype and to establish a system for drug target discovery and personalized oncology.

Biosketch:

Professor Colin Collins received his BSc degree on Biological Sciences from Western New England College, MA in 1982, his Certificate of Genetics Engineering from San Francisco State University, CA, 1965 and his PhD degree in Medical Genetics, UBC, 1993. Dr. Collins is a professor of Urologic Sciences in the school of medicine at the University of British Columbia and a senior scientist at the Vancouver Prostate Centre where he directs the Laboratory for Advanced Genome Analysis (LAGA). In addition, Dr. Collins is an associate adjunct professor at the University of California San Francisco (UCSF) Helen Diller Family Comprehensive Cancer Center and has visiting Professorships at Fudan University and BGI in China. He has held positions at Lawrence Livermore National Laboratory and Lawrence Berkeley National Laboratory. He holds multiple patents, and has received numerous awards including the California Cancer Research Programs Cornelius L. Hopper Scientific Achievement Award for Innovation.

Bing Ren

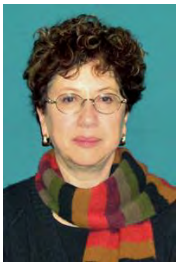
Professor of Cellular and Molecular Medicine
University of California, San Diego, School of Medicine
Member, Ludwig Institute for Cancer Research
USA

Sequencing the 3D Genome

The spatial organization of the genome is intimately linked to its biological function, yet our understanding of higher order genomic structure is still incomplete. In the nucleus of eukaryotic cells, interphase chromosomes occupy distinct chromosome territories, and it is unclear how chromosomes fold within chromosome territories. Recent advances in genomic technologies have led to rapid advances in the study of three-dimensional genome organization. We have investigated the 3D genome organization in a number of human and mouse cell types using next gen DNA sequencing technologies. We identify large, megabase-sized local chromatin interaction domains, which we term "topological domains", as a pervasive structural feature of the genome organization. These domains are stable across different cell types and highly conserved across species, indicating that topological domains are an inherent property of mammalian genomes. I will present evidence that demonstrates an essential role of topological domain structure in mediating long range regulation of gene expression by enhancers.

Biosketch:

Dr. Ren is currently Member of the Ludwig Institute for Cancer Research (LICR) and Professor of Cellular and Molecular Medicine at the UCSD School of Medicine. He leads the San Diego Epigenome Center, one of four NIH-sponsored Reference Epigenome Mapping Centers as part of the Roadmap Epigenomics project. He obtained his Ph.D. from Harvard University in 1998, where he studied mechanisms of transcriptional repression under the guidance of Dr. Tom Maniatis. From 1998 to 2001, he continued to research mechanisms of gene regulation and genomics as a postdoctoral fellow in Dr. Richard Young's laboratory at Whitehead Institute. During this period he developed the ChIP-chip analysis method. At UCSD and LICR, Dr. Ren continued to use genomic approaches to investigate the gene regulatory networks and epigenetic mechanisms in eukaryotic cells. His lab developed high throughput methods for mapping transcription factor binding sites in the human genome, mapped promoters, enhancers, and insulator elements in the human and mouse genomes, and characterized the epigenomic landscapes in pluripotent and lineage-committed human cells. He is a recipient of the Kimmel Scholar award and Young Investigator Award of the Chinese Biological Investigator Society.



Inna Dubchak

Senior Staff Scientist, Genomics Division
Lawrence Berkeley National Laboratory (LBNL) and
DOE Joint Genome Institute
USA

Efficient visualization of Next-Generation Sequencing data - are we there yet?

As the rate of generating sequence data continues to increase, data analysis is becoming a limiting step in genomics studies. Visualization tools to interactively explore and interpret the data at the level of gene, genome, and ecosystem are of critical importance due to the huge volumes and complexity of data produced. As an example, metagenome assemblies are difficult to analyze since they contain sequences from many organisms at different abundance levels, combined into a single file that may contain hundreds of thousands of contig consensus sequences. Contig features such as G+C content and coverage are routinely available, but these provide limited insight into community composition or function. We will talk about strengths and limitations of existing methods, and highlight new challenges in visualization of Next-Generation Sequencing data.

Biosketch:

Inna Dubchak is a senior staff scientist in the Genomics Division at Lawrence Berkeley National Laboratory (LBNL) and the DOE Joint Genome Institute. She has been actively working in computational biology and bioinformatics since 1991, and has authored more than 100 academic articles in the area of analysis and visualization of large volumes of genomic information. Her most significant contributions to date have been leading the development of VISTA, a comprehensive family of tools for comparative genomics that has become one of the major systems used by the biological community; whole-genome comparative analysis of a wide variety of species; building databases and computational tools for investigating transcriptional regulation in bacteria.

Oral Presentations

Evaluating genome architecture of a complex region via generalized bipartite matching

Christine Lo [*1], Sangwoo Kim [1], Shay Zakov [1], Vineet Bafna [1]

[1]: Department of Computer Science and Engineering, University of California, San Diego, CA, USA.

[*]: To whom correspondence should be addressed.

Emails: CL (cylo@eng.ucsd.edu); SK (sak042@eng.ucsd.edu); SZ (szakov@eng.ucsd.edu); VB (vbafna@eng.ucsd.edu)

With the remarkable development in inexpensive sequencing technologies and supporting computational tools, we have the promise of medicine being personalized by knowledge of the individual genome. Current technologies provide high throughput, but short reads. Reconstruction of the donor genome is based either on de novo assembly of the (short) reads, or on mapping donor reads to a standard reference. While such techniques demonstrate high success rates for inferring `simple' genomic segments, they are confounded by segments with complex duplication patterns, including regions of direct medical relevance, like the HLA and the KIR regions.

In this work, we address this problem with a method for assessing the quality of a predicted genome sequence for complex regions of the genome. This method combines two natural types of evidence: sequence similarity of the mapped reads to the predicted donor genome, and distribution of reads across the predicted genome. We define a new scoring function for read-to-genome matchings, which penalizes for sequence dissimilarities and deviations from expected read location distribution, and present an efficient algorithm for finding matchings that minimize the penalty. The algorithm is based on a formal problem, first defined in this paper, called **C***overage* **S***ensitive* many-to-many min-cost bipartite **M***atching* (CSM). This new problem variant generalizes the standard (one-to-one) weighted bipartite matching problem, and can be solved using network flows. The resulting Java-based tool, called SAGE (**S***coring* function for **A***ssembled* **GE***nomes*), is freely available upon request. We demonstrate over simulated data that SAGE can be used to infer correct haplotypes of the highly repetitive KIR region on the Human chromosome 19.

Bellerophon: a hybrid method for detecting interchromosomal rearrangements at base pair resolution using next-generation sequencing data

Matthew Hayes [1], Jing Li [*1]

[1]: Department of Electrical Engineering and Computer Science, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH, USA.

[*]: To whom correspondence should be addressed.

Emails: MH (matthew.hayes@case.edu); JL (jingli@case.edu)

Background: Somatically-acquired translocations may serve as important markers for assessing the cause and nature of diseases like cancer. Algorithms to locate translocations may use next-generation sequencing (NGS) platform data. However, paired-end strategies do not accurately predict precise translocation breakpoints, and "split-read" methods may lose sensitivity if a translocation boundary is not captured by many sequenced reads. To address these challenges, we have developed "Bellerophon", a method that uses discordant read pairs to identify potential translocations, and subsequently uses "soft-clipped" reads to predict the location of the precise breakpoints. Furthermore, for each chimeric breakpoint, our method attempts to classify it as a participant in an unbalanced translocation, balanced translocation, or interchromosomal insertion.

Results: We compared Bellerophon to four previously published algorithms for detecting structural variation (SV). Using two simulated datasets and two prostate cancer datasets, Bellerophon had overall better performance than the other methods. Furthermore, our method accurately predicted the presence of the interchromosomal insertions placed in our simulated dataset, which is an ability that the other SV prediction programs lack.

Conclusions: The combined use of paired reads and soft-clipped reads allows Bellerophon to detect interchromosomal breakpoints with high sensitivity, while also mitigating losses in specificity. This trend is seen across all datasets examined. Because it does not perform assembly on soft-clipped subreads, Bellerophon may be limited in experiments where sequence read lengths are short.

Availability: The program can be downloaded from http://cbc.case.edu/Bellerophon

Keywords: Translocation, structural variation, soft-clipping, interchromosomal, next generation sequencing.

Cloudbreak: A MapReduce Algorithm for Detecting Genomic Structural Variation

Christopher W. Whelan[1] and Kemal Sönmez[1;2]

1 Institute on Development and Disability, Center for Spoken Language Understanding
2 Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA
cwhelan@gmail.com, sonmezk@ohsu.edu

The detection of genomic structural variations remains one of the the most difficult challenges in analyzing high-throughput sequencing data. Recent approaches have demonstrated that considering multiple mappings of all reads, rather than only uniquely mapped discordant fragments, can improve the performance of read-pair based detection methods. However, the computational requirements for storing and processing data sets with multiple mappings can be formidable. Meanwhile, the growing size and number of sequencing data sets have led to intense interest in distributing computation to cloud or commodity servers.

MapReduce, via its Hadoop implementation, is becoming a standard architecture for distributing processing across such compute clusters. In this work we describe a novel conceptual framework for structural variation detection in MapReduce/Hadoop based on computing local features along the genome. Our framework uses Hadoop to take advantage of distributed computing to find all possible read alignments using modern short-read aligners run with sensitive settings. We then provide an architecture to first compute features for each genomic location from the relevant alignments, and then to call structural variants from the set of all features across the genome.

In this framework, we have developed and evaluated a distributed deletion-finding algorithm based on fitting a Gaussian mixture model (GMM) to the distribution of mapped insert sizes spanning each location in the genome. A similar method was used in MoDIL[1]; however, our algorithm and the Hadoop framework drastically reduce the runtime requirements and overall difficulty of using this approach.

On simulated and real data sets of paired-end reads, our algorithm achieves performance similar to or better than a variety of popular structural variation detection algorithms, including read-pair, split-read, and hybrid approaches. Cloudbreak performs well on both small and medium size deletions, and in our simulations has greater sensitivity at most fixed levels of specificity. We also show increased performance in repetitive areas of the genome, identifying more deletions that overlap repeats than other approaches in both simulated and real data.

In addition, our algorithm can accurately genotype heterozygous and homozygous deletions from diploid samples. Using the parameters computed in fitting the GMM and a simple thresholding procedure, we were able to achieve 88.0% and 94.9% accuracy in predicting the genotype of the true positive deletions we detected in simulated and real data sets, respectively.

Finally, we have recently added the ability to detect insertions to Cloudbreak. Our implementation and source code are available at https://github.com/cwhelan/ cloudbreak.

[1] Lee, S. et al., 2009. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat. Methods, 6(7), pp.473474.

Assembling Contigs in Draft Genomes Using Reversals and Block-Interchanges

Chi-Long Li [1], Kun-Tze Chen [1], Chin Lung Lu *[1]

[1]: Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan
*: To whom correspondence should be addressed.
Emails: CL (496470140@ntnu.edu.tw); KC (holystu@gmail.com); CL (cllu@cs.nthu.edu.tw)

The techniques of next generation sequencing allow an increasing number of draft genomes to be produced rapidly in a decreasing cost. However, these draft genomes usually are just partially sequenced as collections of unassembled contigs, which cannot be used directly by currently existing algorithms for studying their genome rearrangements and phylogeny reconstruction. In this work, we study the one-sided block (or contig) ordering problem with weighted reversal and block-interchange distance. Given a partially assembled genome π and a completely assembled genome σ, the problem is to find an optimal ordering to assemble (i.e., order and orient) the contigs of π such that the rearrangement distance measured by reversals and block-interchanges (also called generalized transpositions) with the weight ratio 1:2 between the assembled contigs of π and σ is minimized. In addition to genome rearrangements and phylogeny reconstruction, the one-sided block ordering problem particularly has a useful application in genome resequencing, because its algorithms can be used to assemble the contigs of a draft genome π based on a reference genome σ. By using permutation groups, we design an efficient algorithm to solve this one-sided block ordering problem in $O(\delta n)$ time, where n is the number of genes or markers and δ is the number of used reversals and block-interchanges. We also show that the assembly of the partially assembled genome can be done in $O(n)$ time and its weighted rearrangement distance from the completely assembled genome can be calculated in advance in $O(n)$ time. Finally, we have implemented our algorithm into a program and used some simulated datasets to compare its accuracy performance to a currently existing similar tool, called SIS that was implemented by a heuristic algorithm that considers only reversals, on assembling the contigs in draft genomes based on their reference genomes. Our experimental results have shown that the accuracy performance of our program is better than that of SIS, when the number of reversals and transpositions involved in the rearrangement events between the complete genomes of π and σ is increased. In particular, if there are more transpositions involved in the rearrangement events, then the gap of accuracy performance between our program and SIS is increasing.

Optimal Assembly for High Throughput Shotgun Sequencing

Guy Bresler [1], Ma'ayan Bresler [1], David Tse [1]

[1]: Dept. of EECS, UC Berkeley.

Emails: GB (gbresler@eecs.berkeley.edu); MB (mbresler@ eecs.berkeley.edu); DT (dtse@ eecs.berkeley.edu)

We present a framework for the design of optimal assembly algorithms for shotgun sequencing under the criterion of complete reconstruction. We derive a lower bound on the read length and the coverage depth required for reconstruction in terms of the repeat statistics of the genome. Building on earlier works, we design a de Brujin graph based assembly algorithm which can achieve very close to the lower bound for repeat statistics of a wide range of sequenced genomes, including the GAGE datasets. The results are based on a set of necessary and sufficient conditions on the DNA sequence and the reads for reconstruction. The conditions can be viewed as the shotgun sequencing analogue of Ukkonen-Pevzner's necessary and sufficient conditions for Sequencing by Hybridization.

De Bruijn Superwalk with Multiplicities Problem is NP-hard

Evgeny Kapun [1], Fedor Tsarev [*1]

[1]: St. Petersburg National Research University of Information Technologies, Mechanics and Optics Genome Assembly Algorithms Laboratory, 197101, Kronverksky pr., 49, St. Petersburg, Russia.

[*]: To whom correspondence should be addressed.

Emails: FT (tsarev@rain.ifmo.ru)

De Bruijn Superwalk with Multiplicities Problem is the problem of finding a walk in the de Bruijn graph containing several walks as subwalks and passing through each edge the exactly predefined number of times (equal to the multiplicity of this edge). This problem has been stated in the talk by Paul Medvedev and Michael Brudno on the first RECOMB Satellite Conference on Open Problems in Algorithmic Biology in August 2012. In this paper we show that this problem is NP-hard. Combined with results of previous works it means that all known models for genome assembly are NP-hard.

Keywords: de Bruijn graph, genome assembly, complexity, superwalk.

HMCan – a tool to detect chromatin modifications in cancer samples using ChIP-seq data

Haitham Ashoor[1], Aurélie Hérault[2,3], François Radvanyi[2,3], Vladimir B. Bajic[1], Emmanuel Barillot[2,4,5,6], ValenDna Boeva[2,4,5,6]

[1] King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBR), Computer, Electrical and Mathematical Sciences and Engineering Division, Thuwal, Saudi Arabia; [2] Institut Curie, Paris, France; 3UMR 144 CNRS; [4] INSERM, U900; [5] Mines Paris Tech, Fontainebleau, France; [6] Equiqe Laebllisee Ligue, Nationale Contre le Cancer

Introduction: Epigenetic changes often play an important role in cancer development. By introducing local or regional epigenetic silencing (RES), cancer cells may limit expression of tumor suppressor genes. One way to study epigenetic silencing in cancer is to explore patterns of tri-methylation of lysine 27 of histone 3 (H3K27me3), associated with Polycomb-mediated Repression. Though several tools were created to enable detection of histone marks in ChIP-seq data from *normal* samples, it was unclear whether these tools can be applied to ChIP-seq data generated from *cancer* samples. The challenge comes form the fact that cancer genomes are often characterized by frequent copy number alterations: gains and losses of large regions of chromosomal material. Copy number alteration may create a substantial statistical bias in the evaluation of histone mark signal enrichment and result in underdetection of the signal in the regions of loss and overdetection of the signal in regions of gain.

Results: We present HMCAN (Histone Modification in Cancer), a tool specially developed to analyze histone modification ChIP-seq data produced from cancer genomes. HMCan corrects for the GC- and copy number bias and then applies Hidden Markov Models (HMMs) to detect the signal in the normalized profile. We showed that HMCan provided a significantly better accuracy of predictions than commonly used tools such as CCAT, MACS, SICER. Of note, MACS and SICER were biased towards regions of copy number gain, while CCAT and HMCan did not demonstrate such a bias.We also generated a ChIP-seq dataset for the repressive histone mark H3K27me3 in a bladder cancer cell line. On this dataset, HMCan was able to recover the expected shape of the H3K27me profile in the vicinity of gene transcription start sites. HMCan predictions included, for example, previously detected H3K27me3 marks on the *DLEC1* gene, which is commonly inactivated in various carcinomas, as well as the *HOXD* gene cluster, which plays a crucial role in normal cell development and proliferation.

Availability: C++ source code is available at http://sourceforge.net/p/hmcan

Discovering and Mapping Chromatin States Using a Tree Hidden Markov Model

Jacob Biesinger [1,3†], Yuanfeng Wang [2†], Xiaohui Xie [*1,3]

[1]: Department of Computer Science, University of California, Irvine.
[2]: Department of Physics and Astronomy, University of California, Irvine.
[3]: Institute for Genomics and Bioinformatics, University of California, Irvine.
[†]: Contributed equally to this work
[*]: To whom correspondence should be addressed.
Emails: JB (jake.biesinger@uci.edu); YW (yuanfenw@uci.edu); XX (xhx@ics.uci.edu)

New biological techniques and technological advances in high-throughput sequencing are paving the way for systematic, comprehensive annotation of many genomes, allowing differences between cell types or between disease/normal tissues to be determined with unprecedented breadth. Epigenetic modifications have been shown to exhibit rich diversity between cell types, correlate tightly with cell-type specific gene expression, and changes in epigenetic modifications have been implicated in several diseases. Previous attempts to understand chromatin state have focused on identifying combinations of epigenetic modification, but in cases of multiple cell types, have not considered the lineage of the cells in question.

We present a Bayesian network that uses epigenetic modifications to simultaneously model 1) chromatin mark combinations that give rise to different chromatin states and 2) propensities for transitions between chromatin states through differentiation or disease progression. We apply our model to a recent dataset of histone modifications, covering nine human cell types with nine epigenetic modifications measured for each. Since exact inference in this model is intractable for all the scale of the datasets, we develop several variational approximations and explore their accuracy. Our method exhibits several desirable features including improved accuracy of inferring chromatin states, improved handling of missing data, and linear scaling with dataset size. The source code for our model is available at http://github.com/xhxielab/TreeHMM.

Keywords: chromatin modifications, hidden Markov model, graphical model, lineage, differentiation

A Novel Min-Cost Flow Method for Estimating Transcript Expression with RNA-Seq

Alexandru I. Tomescu [*1], Anna Kuosmanen [1], Romeo Rizzi [2] Veli Mäkinen [1]

[1]: Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland.
[2]: Department of Computer Science, University of Verona, Italy.
[*]: To whom correspondence should be addressed.
Emails: AT (tomescu@cs.helsinki.fi); AK (aekuosma@cs.helsinki.fi); RR (romeo.rizzi@univr.it); VM (vmakinen@cs.helsinki.fi)

Background: Through transcription and alternative splicing, a gene can be transcribed into different RNA sequences (isoforms), depending on the individual, on the tissue the cell is in, or in response to some stimuli. Recent RNA-Seq technology allows for new high-throughput ways for isoform identification and quantification based on short reads, and various methods have been put forward for this non-trivial problem.

Results: In this paper we propose a novel radically different method based on minimum-cost network flows. This has a two-fold advantage: on the one hand, it translates the problem as an established one in the field of network flows, which can be solved in polynomial time, with different existing solvers; on the other hand, it is general enough to encompass many of the previous proposals under the least sum of squares model. Our method works as follows: in order to find the transcripts which best explain, under a given fitness model, a splicing graph resulting from an RNA-Seq experiment, we find a min-cost flow in an offset flow network, under an equivalent cost model. Under very weak assumptions on the fitness model, the optimal flow can be computed in polynomial time. Parsimoniously splitting the flow back into few path transcripts can be done with any of the heuristics and approximations available from the theory of network flows. In the present implementation, we choose the simple strategy of repeatedly removing the heaviest path.

Conclusions: We proposed a new very general method based on network flows for a multiassembly problem arising from isoform identification and quantification with RNA-Seq. Experimental results on prediction accuracy show that our method is very competitive with popular tools such as Cuffinks and IsoLasso. Our tool, called Traph (Transcrips in gRAPHs), is available at: www.cs.helsinki.fi/gsa/traph/.

Keywords: Isoform Identification, Isoform Quantification, RNA-Seq, Graph Optimization Problem, Network Flow, Min-Cost Flow.

WemIQ: a weighted-log-likelihood expectation maximization method for isoform quantification from RNA-Seq data

Jing Zhang[1], C.-C. Jay Kuo[1], Liang Chen[2,*]

[1]Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089, USA.
[2]Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA.
[*]To whom correspondence should be addressed. Tel.: +1 213 740 2143; E-mail:liang.chen@usc.edu

Isoform-level expression quantification is essential to dissect transcriptomes in higher eukaryotes. However, the deconvolution of isoform expression remains challenging because of the nonuniformity of read sampling in RNA-Seq. As a result, read counts need to be adjusted during gene abundance quantification. Add-on functions have been introduced to handle the overdispersion in the methods of isoform expression estimation. However they either assumed a constant bias factor for each relative position of genes or simply corrected the sequence-specific bias caused by random hexamer priming. These overlooked the fact that the bias is complicated and caused by multiple factors including many unknown factors, and the bias pattern can vary significantly across different genes and different protocols. The quantification of transcript isoforms therefore calls for a more appropriate handling of the bias in RNA-Seq.

Here, we present a weighted-log-likelihood expectation maximization method on isoform quantification (WemIQ). First, WemIQ removes both inter- and intra- gene bias estimated in a data-driven manner through a Generalized Poisson model. Analysis on single-isoform genes in real RNA-seq data demonstrates that the bias correction is effective, but other existing correction methods such as positional and sequence-based bias removal could be problematic. Then, WemIQ distributes reads among isoforms with the fragment length information of paired-end reads. Simulation studies show that, compared with other popular methods such as Cufflinks, RSEM, and SpliceTrap, WemIQ quantifies isoform expression and exon inclusion ratios more accurately. We also simulated genes with low coverage, incomplete annotation, or complicated gene structures, and WemIQ all provides significantly improved estimations over other software packages. In addition, we used the TaqMan qRT-PCR results in the Microarray Quality Control (MAQC) Project as a benchmark for gene expression measurements. WemIQ provides a higher correlation with the qRT-PCR estimates than those of Cufflinks and RSEM, indicating improved expression estimation at the gene level.

Key Words: EM algorithm, RNA-Seq, transcript isoform quantification, weighted log-likelihood

A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues

Yi Li [1], Xiaohui Xie [*1,2,3]

[1]: Department of Computer Science, University of California, Irvine, USA.
[2]: Institute for Genomics and Bioinformatics, University of California, Irvine, USA.
[3]: Center for Machine Learning and Intelligent Systems, University of California, Irvine, USA.
[*]: To whom correspondence should be addressed.
Emails: YL (yil8@uci.edu); XX (xhx@ics.uci.edu)

Background: RNA-seq, a next-generation sequencing based method for transcriptome analysis, is rapidly emerging as the method of choice for comprehensive transcript abundance estimation. The accuracy of RNA-seq can be highly impacted by the purity of samples. A prominent, outstanding problem in RNA-seq is how to estimate transcript abundances in heterogeneous tissues, where a sample is composed of more than one cell type and the inhomo-geneity can substantially confound the transcript abundance estimation of each individual cell type. Although experimental methods have been proposed to dissect multiple distinct cell types, computationally "deconvoluting" heterogeneous tissues provides an attractive alternative, since it keeps the tissue sample as well as the subsequent molecular content yield intact.

Results: Here we propose a probabilistic model-based approach, Transcript Estimation from Mixed Tissue samples (TEMT), to estimate the transcript abundances of each cell type of interest from RNA-seq data of heterogeneous tissue samples. TEMT incorporates positional and sequence-specific biases, and its online EM algorithm only requires a runtime proportional to the data size and a small constant memory. We test the proposed method on both simulation data and recently released ENCODE data, and show that TEMT significantly outperforms current state-of-the-art methods that do not take tissue heterogeneity into account. Currently, TEMT only resolves the tissue heterogeneity resulting from two cell types, but it can be extended to handle tissue heterogeneity resulting from multi cell types. TEMT is written in python, and is freely available at https://github.com/xhxielab/TEMT.

Conclusions: The probabilistic model-based approach proposed here provides a new method for analyzing RNA-seq data from heterogeneous tissue samples. By applying the method to both simulation data and ENCODE data, we show that explicitly accounting for tissue heterogeneity can significantly improve the accuracy of transcript abundance estimation.

Keywords: RNA-seq, Tissue Heterogeneity, Mixture Model, Online Expectation-Maximization, Positional Bias, Sequence-specific Bias, ENCODE

Gene Set Enrichment Analysis of RNA - Seq Data: Integrating Differential Expression and Splicing

Xi Wang [1,2], Murray J. Cairns [*1,2,3]

[1]: School of Biomedical Sciences and Pharmacy, The University of Newcastle, Callaghan, New South Wales, Australia.
[2]: Hunter Medical Research Institute, New Lambton, New South Wales, Australia,
[3]: Schizophrenia Research Institute, Sydney, New South Wales, Australia.
[*]: To whom correspondence should be addressed. School of Biomedical Sciences and Pharmacy, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia; Tel: +61-2-4921-8670, Fax: +61-2-4921-7903, E-Mail: Murray.Cairns@newcastle.edu.au

Emails: XW (Xi.Wang@newcastle.edu.au); MC (Murray.Cairns@newcastle.edu.au)

Background: RNA-Seq has become a key technology in transcriptome studies because it can quantify overall expression levels and the degree of alternative splicing for each gene simultaneously. To interpret high-throughout transcriptome profiling data, functional enrichment analysis is critical. However, existing functional analysis methods can only account for differential expression, leaving differential splicing out altogether.

Results: In this work, we present a novel approach to derive biological insight by integrating differential expression and splicing from RNA-Seq data with functional gene set analysis. This approach designated SeqGSEA, uses count data modelling with negative binomial distributions to first score differential expression and splicing in each gene, respectively, followed by two strategies to combine the two scores for integrated gene set enrichment analysis. Method comparison results and biological insight analysis on an artificial data set and three real RNA-Seq data sets indicate that our approach outperforms alternative analysis pipelines and can detect biological meaningful gene sets with high confidence, and that it has the ability to determine if transcription or splicing is their predominant regulatory mechanism.

Conclusions: By integrating differential expression and splicing, the proposed method SeqGSEA is particularly useful for efficiently translating RNA-Seq data to biological discoveries.

Keywords: RNA-Seq; Gene set enrichment analysis; Differential expression; Differential splicing; Data integration

Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles

Jie Zhu [1], Yufang Qin [2], Taigang Liu [2], Jun Wang [1,3], Xiaoqi Zheng [*1,3]

[1]: Department of Mathematics, Shanghai Normal University, Shanghai 200034, China.
[2]: College of Information Technology, Shanghai Ocean University, Shanghai 201306, China.
[3]: Scientific Computing Key Laboratory of Shanghai Universities, Shanghai 200234, China.
[*]: To whom correspondence should be addressed.
Emails: JZ (jzhu@shnu.edu.cn); YQ (yfqin@shou.edu.cn); TL (aaaltg@126.com); JW (jwang@shnu.edu.cn); XZ (xqzheng@shnu.edu.cn)

Background: Identification of gene-phenotype relationships is a fundamental challenge in human health clinic. Based on the observation that genes causing the same or similar phenotypes tend to correlate with each other in the protein-protein interaction network, a lot of network-based approaches were proposed based on different underlying models. A recent comparative study showed that diffusion-based methods achieve the state-of-the-art predictive performance.

Results: In this paper, a new diffusion-based method was proposed to prioritize candidate disease genes. Diffusion profile of a disease was defined as the stationary distribution of candidate genes given a random walk with restart where similarities between phenotypes are incorporated. Then, candidate disease genes are prioritized by comparing their diffusion profiles with that of the disease. Finally, the effectiveness of our method was demonstrated through the leave-one-out cross-validation against control genes from artificial linkage intervals and randomly chosen genes. Comparative study showed that our method achieves improved performance compared to some classical diffusion-based methods. To further illustrate our method, we used our algorithm to predict new causing genes of 16 multifactorial diseases including Prostate cancer and Alzheimer's disease, and the top predictions were in good consistent with literature reports.

Conclusions: Our study indicates that integration of multiple information sources, especially the phenotype similarity profile data, and introduction of global similarity measure between disease and gene diffusion profiles are helpful for prioritizing candidate disease genes.

Availability: Programs and data are available upon request.

Keywords: PPI network, Random walk, Diffusion profile, Prostate cancer, Alzheimer's disease.

Detecting SNP-SNP Interactions With Piecewise Independence Screening

Seunghak Lee [1], Aur´elie Lozano [2], Prabhanjan Kambadur [3], Eric P. Xing [4]

[1]School of Computer Science, Carnegie Mellon University. E-mail: seunghak@cs.cmu.edu
[2]IBM T. J. Watson Research Center. E-mail: aclozano@us.ibm.com
[3]IBM T. J. Watson Research Center. E-mail: pkambadu@us.ibm.com
[4]School of Computer Science, Carnegie Mellon University. E-mail: epxing@cs.cmu.edu

Interactions between genetic variants are key to understanding the genetic effects on phenotypic traits. However, detecting interaction effects is an ultra-high dimensional problem, and therefore, it is both statistically and computationally challenging. Despite recent breakthroughs in detecting interaction effects, several problems remain including: (1) processing the large number of genetic interactions without compromising for the sake of scalability, (2) solving the multiple testing problem posed by the ultra-high dimensionality of the problem, (3) accounting for strong correlations that exist between SNPs and SNP-SNP pairs, (4) identifying non-linear relationships between genotypes and phenotypes. We present a principled and scalable framework to address these problems in a unified way. Our framework consists of three steps: a screening procedure with piecewise linear model to account for non-linear relationships between genotypes and phenotypes, a procedure for penalized multivariate regression, and a procedure for p-value computation with a correction for multiple testing. The screening procedure is employed to handle the extremely large number of candidate pairs, while the penalized multivariate regression and p-value computation allow the selection of statistically significant SNP pairs. We demonstrate the effectiveness and scalability of the proposed framework on simulated and real-world datasets. Our results on simulated data show that our framework exhibits higher accuracy vis-`a-vis recovering the true associated SNPs and SNP-SNP pairs when compared to existing methods. Those on Alzheimer's data demonstrate that our framework is able to uncover biologically meaningful associations, some as of yet unreported. We also present a high-performance implementation of our screening method, which is highly scalable and is able to screen O(109) SNP-SNP pairs in only a few hours.

De Novo Inference of Stratification and Local Admixture in Sequencing Studies

Yu Zhang [*1]

[1]: Department of Statistics, The Pennsylvania State University. 326 Thomas Building, University Park, PA 16802

[*]: To whom correspondence should be addressed.

Emails: YZ (yzz2@psu.edu)

Analysis of population structures and genome local ancestry has become increasingly important in population and disease genetics. With the advance of next generation sequencing technologies, complete genetic variants in individuals' genomes are quickly generated, providing unprecedented opportunities for learning population evolution histories and identifying local genetic signatures at the SNP resolution. The successes of those studies critically rely on accurate and powerful computational tools that can fully utilize the sequencing information. Although many algorithms have been developed for population structure inference and admixture mapping, many of them only work for independent SNPs in genotype or haplotype format, and require a large panel of reference individuals. In this paper, we propose a novel probabilistic method for detecting population structure and local admixture. The method takes input of sequencing data, genotype data and haplotype data. The method characterizes the dependence of genetic variants via haplotype segmentation, such that all variants detected in a sequencing study can be fully utilized for inference. The method further utilizes a infinite-state Bayesian Markov model to perform de novo stratification and admixture inference. Using simulated datasets from HapMapII and 1000Genomes, we show that our method performs superior than several existing algorithms, particularly when limited or no reference individuals are available. Our method is applicable to not only human studies but also studies of other species of interests, for which little reference information is available.

Software Availability: http://stat.psu.edu/~yuzhang/software/dbm.tar

Joint genotype inference with germline and somatic mutations

Eric Barake [1], Virginie Saillour [1], Jean-François Spinella [1], Ramon Vidal [1], Jasmine Healy [1], Daniel Sinnett [1,2], Miklós Csűrös[3]

[1]: Division of Hematology-Oncology, Sainte-Justine University Hospital Centre, Montréal, QC, Canada.

[2]: Department of Pediatrics, University of Montréal, Sainte Justine UHC Research Centre, QC, Canada.

[3]: Department of Computer Science and Operations Research, University of Montréal, QC, Canada.

[*]: To whom correspondence should be addressed.

Emails: MC (csuros@iro.umontreal.ca)

The joint sequencing of related genomes has become an important means to discover rare variants. Normal-tumor genome pairs are routinely sequenced together to find somatic mutations and their associations with different cancers. Parental and sibling genomes reveal *de novo* germline mutations, and inheritance patterns related to Mendelian diseases.

Acute lymphoblastic leukemia (ALL) is the most common paediatric cancer and the leading cause of cancer-related death among children. With the aim of uncovering the full spectrum of germline and somatic genetic alterations in childhood ALL genomes, we conducted whole-exome re-sequencing on a unique cohort of over 120 exomes of childhood ALL quartets, each comprising a patient's tumor and matched-normal material, and DNA from both parents. We developed a general probabilistic model for such quartet sequencing reads mapped to the reference human genome. The model is used to infer joint genotypes at homologous loci across a normal-tumor genome pair and two parental genomes.

We describe the algorithms and data structures for genotype inference, model parameter training. We implemented the methods in an open-source software package (QUADGT) that uses the standard file formats of the 1000 Genomes Project. Our method's utility is illustrated on quartets from the ALL cohort.

A Random-Permutations-Based Approach to Fast Read Alignment

Roy Lederman [*1]

[1]: Applied Mathematics Program, Yale University, 51 Prospect st. New Haven, CT 06511, USA

[*]: To whom correspondence should be addressed.

Emails: RL (roy.lederman@yale.edu)

Background: Read alignment is a computational bottleneck in some sequencing projects. Most of the existing software packages for read alignment are based on two algorithmic approaches: prefix-trees and hash-tables. We propose a new approach to read alignment using random permutations of strings.

Results: We present a prototype implementation and experiments performed with simulated and real reads of human DNA. Our experiments indicate that this permutations-based prototype is several times faster than comparable programs for fast read alignment and that it aligns more reads correctly.

Conclusions: This approach may lead to improved speed, sensitivity, and accuracy in read alignment. The algorithm can also be used for specialized alignment applications and it can be extended to other related problems, such as assembly.

More information: http://alignment.commons.yale.edu

Keywords: Sequencing, Read mapping, Read alignment, Random permutations.

metaBEETL: high-throughput analysis of heterogeneous microbial populations from shotgun DNA sequences

Christina Ander [1], Ole B. Schulz-Trieglaff [2], Jens Stoye [1], Anthony J. Cox [*2]

[1]: Genome Informatics, Faculty of Technology and CeBiTec, Bielefeld University, Bielefeld, Germany.
[2]: Computational Biology Group, Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex CB10 1XL, United Kingdom.
[*]: To whom correspondence should be addressed.
Emails: AC (acox@illumina.com)

Environmental shotgun sequencing (ESS) has potential to give greater insight into microbial communities than targeted sequencing of 16S regions, but requires much higher sequence coverage. The advent of next-generation sequencing has made it feasible for the Human Microbiome Project and other initiatives to generate ESS data on a large scale, but computationally efficient methods for analysing such data sets are needed.

Here we present metaBEETL, a fast taxonomic classifier for environmental shotgun sequences. It uses a Burrows-Wheeler Transform (BWT) index of the sequencing reads and an indexed database of microbial reference sequences. Unlike other BWT-based tools, our method does not have an upper limit on the number or the total size of references which is important in a metagenomic setting. By capturing sequence relationships between strains, the reference index also allows us to classify reads which are not unique to an individual strain but only unique at the level of some higher phylogenetic order.

We tested our classifier with data sets of known taxonomic composition and compared the results to other similarity-based tools. Its performance is comparable to existing tools while metaBEETL requires less RAM and thus scales better to large data sets. Due to normalization steps which other classifiers lack, the taxonomic profile computed by metaBEETL closely matches the true environmental profile. We also evaluate metaBEETL on a real data set from the Human Microbiome Project.

Code to construct the BWT indexed database and for the taxonomic classification is part of the BEETL library, available as a github repository at git@github.com:BEETL/BEETL.git.

Keywords: metagenomics, Burrows-Wheeler Transform, indexing, backward search.

Gene Prediction in Metagenomic Fragments Based on the SVM Algorithm

Yongchu Liu [1,2], Jiangtao Guo [1,2], Gangqing Hu [1,2,4], Huaiqiu Zhu [*1,2,3]

[1]: State Key Laboratory for Turbulence and Complex Systems and Department of Biomedical Engineering, College of Engineering, Peking University, Beijing 100871, China.

[2]: Center for Theoretical Biology, Peking University, Beijing 100871, China.

[3]: Center for Protein Science, Peking University, Beijing 100871, China.

[4]: Laboratory of Molecular Immunology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

[*]: To whom correspondence should be addressed.

Emails: YL (liuyc@ctb.pku.edu.cn); JG (jtguo@ctb.pku.edu.cn); GH (Gangqing.Hu@nih.gov); HZ (hqzhu@pku.edu.cn)

Background: Metagenomic sequencing is becoming a powerful technology for exploring micro-ogranisms from various environments, such as human body, without isolation and cultivation. Accurately identifying genes from metagenomic fragments is one of the most fundamental issues.

Results: In this article, we present a novel gene prediction method named MetaGUN for metagenomic fragments based on a machine learning approach of SVM. It implements in a three-stage strategy to predict genes. Firstly, it classifies input fragments into phylogenetic groups by a k-mer based sequence binning method. Then, protein-coding sequences are identified for each group independently with SVM classifiers that integrate entropy density profiles (EDP) of codon usage, translation initiation site (TIS) scores and open reading frame (ORF) length as input patterns. Finally, the TISs are adjusted by employing a modified version of MetaTISA. To identify protein-coding sequences, MetaGun builds the universal module and the novel module. The former is based on a set of representative species, while the latter is designed to find potential functionary DNA sequences with conserved domains.

Conclusions: Comparisons on artificial shotgun fragments with multiple current metagenomic gene finders show that MetaGUN predicts better results on both 3' and 5' ends of genes with fragments of various lengths. Especially, it makes the most reliable predictions among these methods. As an application, MetaGUN was used to predict genes for two samples of human gut microbiome. It identifies thousands of additional genes with significant evidences. Further analysis indicates that MetaGUN tends to predict more potential novel genes than other current metagenomic gene finders

Discovering Motifs that Induce Sequencing Errors

Manuel Allhoff [1,2,3,4,5], Alexander Schönhuth [2], Marcel Martin [3], Ivan G. Costa [4], Sven Rahmann [5,3], Tobias Marschall [*2]

[1]: Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Germany.

[2]: Life Sciences Group, Centrum Wiskunde & Informatica, Amsterdam, Netherlands.

[3]: Bioinformatics, Computer Science XI, TU Dortmund, Germany.

[4]: Interdisciplinary Centre for Clinical Research (IZKF) & Institute for Biomedical Engineering, RWTH University Medical School, Aachen, Germany

[5]: Genome Informatics, Faculty of Medicine, University of Duisburg-Essen, Germany

[*]: To whom correspondence should be addressed.

Emails: MA (allhoff@aices.rwth-aachen.de); TM (T.Marschall@cwi.nl)

Elevated sequencing error rates are the most predominant obstacle in NGS-based single-nucleotide polymorphism (SNP) detection, which is a major goal in the bulk of current NGS-based studies. Beyond routinely handled generic sources of errors, certain base calling errors relate to specific sequence patterns. Statistically principled ways to associate sequence patterns with base calling errors have not been previously described. Extant approaches either incur decisive losses in power, due to relating errors with individual genomic positions rather than motifs, or do not properly distinguish between motif-induced and sequence-unspecific sources of errors. Here, for the first time, we describe a statistically rigorous framework for the discovery of motifs that induce sequencing errors. We apply our method to several datasets from Illumina GA IIx, HiSeq 2000, and MiSeq sequencers. We confirm previously known error-causing sequence contexts and report new more specific ones. To facilitate filtering of sets of putative SNPs, we provide tracks of error-prone genomic positions (in BED format).

CLASS: Constrained Transcript Assembly of RNA-seq Reads

Li Song [1], Liliana Florea [*2]

[1]: Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA 21218.

[2]: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA 21205.

[*]: To whom correspondence should be addressed.

Emails: LS (lsong10@jhu.edu); LF (florea@jhu.edu)

Background: RNA-seq has revolutionized our ability to survey the cellular transcriptome in great detail. However, while several approaches have been developed, the problem of assembling the short reads into full-length transcripts remains challenging.

Results: We developed a novel algorithm and software tool, CLASS (Constraint-based Local Assembly and Selection of Splice variants), for accurately assembling splice variants using local read coverage patterns of RNA-seq reads, contiguity constraints from read pairs and spliced reads, and optionally information about gene structure extracted from cDNA sequence databases. The algorithmic underpinnings of CLASS are: i) a linear program to infer exons, ii) a compact splice graph representation of a gene and its splice variants, and iii) a transcript selection scheme that takes into account contiguity constraints and, where available, knowledge about gene structure.

Conclusions: In comparisons against leading transcript assembly programs, CLASS is more accurate on both simulated and real reads and produces results that are easier to interpret when applied to large scale real data, and therefore is a promising analysis tool for next generation sequencing data.

Availability: CLASS is available from http://sourceforge.net/projects/splicebox.

VNTRseek - A Computational Pipeline to Detect Tandem Repeat Variants in Next-Generation Sequencing Data: Analysis of the 454 Watson Genome

Yevgeniy Gelfand[1], Joshua Loving[1;2], Yozen Hernandez[1;2] and Gary Benson[*1;2;3]

[1]Laboratory for Biocomputing and Informatics, Boston University, Boston, MA
[2]Graduate Program in Bioinformatics, Boston University, Boston, MA
[3]Department of Computer Science, Boston University, Boston, MA
Email: YG - ygelfand@bu.edu; JL - jloving@bu.edu; YH - yhernand@bu.edu; GB - gbenson@bu.edu;
[*] Corresponding author

DNA tandem repeats (TRs) are ubiquitous genome features which consist of two or more adjacent copies of an underlying pattern sequence. The copies may be identical or approximate. Tandem repeats can be highly mutable with respect to the number of adjacent copies, with mutation rates orders of magnitude higher than those of SNPs. Tandem repeats are known to directly cause more than a dozen human neurological syndromes, are associated with other diseases, may have regulatory functions, and are useful genetic markers due to their high mutation rates.

Whole genome sequencing provides the raw material for identifying and studying, for the first time on a genome-wide scale, those repeats which are inherently variable. In this paper we describe our VNTRseek pipeline for the discovery of minisatellite tandem repeat variants, and its application to the 454 next-generation sequencing data from the Watson genome. A minisatellite tandem repeat has a pattern size $\geqslant$7 nucleotides. AVNTR, or Variable Number of Tandem Repeats, is defined as a tandem repeat locus for which more than one allele has been observed, with each allele associated with a particular number of adjacent pattern copies in the tandem array. Our pipeline maps reads which contain tandem repeats to a set of reference TRs and then identifies those references which appear to be VNTRs based on the number of pattern copies in the mapped reads. For the Watson genome, we have identified 792 VNTRs, with pattern sizes ranging from 7 to 64 nucleotides. Coverage of the Watson genome by the 454 data is quite low. We expect that thousands of undiscovered VNTRs will be detected in higher coverage data. This is, to the best of our knowledge, the first software for genome wide detection of larger patterned VNTRs.

An optimized algorithm for detecting and annotating regional differential methylation

Sheng Li [1,2], Francine E. Garrett-Bakelman [3], Altuna Akalin [1,2], Paul Zumbo [1,2], Ross Levine [4], Bik L. To [5], Ian D. Lewis [5], Anna L. Brown [6], Richard J. D'Andrea [6,7], Ari Melnick [3,8], Christopher E. Mason [*1,2]

[1]: Department of Physiology and Biophysics, 1305 York Ave., Weill Cornell Medical College, New York, NY 10065, USA
[2]: The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, 1305 York Ave., Weill Cornell Medical College, New York, NY 10065, USA
[3]: Department of Medicine, Division of Hematology/Oncology, 1300 York Ave., Weill Cornell Medical College, New York, NY 10065, USA.
[4]: Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, Box 20, New York, NY 10065, USA
[5]: Directorate of Haematology, SA Pathology and Department of Haematology, Royal Adelaide Hospital, Adelaide, South Australia
[6]: Directorate of Haematology and Centre for Cancer Biology SA Pathology, The Queen Elizabeth Hospital, Woodville, South Australia
[7]: Department of Haematology and Oncology, The Queen Elizabeth Hospital, Woodville, South Australia
[8]: Department of Pharmacology, 1300 York Ave., Weill Cornell Medical College, New York, NY 10065, USA
[*]: To whom correspondence should be addressed.
Emails: SL (shl2018@med.cornell.edu); FEG-B (frg9015@med.cornell.edu); AA (ala2027@med.cornell.edu); PZ (paz2005@med.cornell.edu); RL (leviner@mskcc.org); BLT (Bik.To@health.sa.gov.au); IDL (ian.lewis@health.sa.gov.au); ALB (anna.brown@health.sa.gov.au); RJD (richard.dandrea@health.sa.gov.au); AM (amm2014@med.cornell.edu); CEM (chm2042@med.cornell.edu)

Background: DNA methylation profiling reveals important differentially methylated regions (DMRs) of the genome that are altered during development or that are perturbed by disease. To date, few programs exist for regional analysis of enriched or whole-genome bisulfate conversion sequencing data, even though such data are increasingly common. Here, we describe an opensource, optimized method for determining empirically based DMRs (eDMR) from highthroughput sequence data that is applicable to enriched whole-genome methylation profiling datasets, as well as other globally enriched epigenetic modification data.

Results: Here we show that our bimodal distribution model and weighted cost function for optimized regional methylation analysis provides accurate boundaries of regions harboring significant epigenetic modifications. Our algorithm takes the spatial distribution of CpGs into account for the enrichment assay, allowing for optimization of the definition of empirical regions for differential methylation. Combined with the dependent adjustment for regional p-value combination and DMR annotation, we provide a method that may be applied to a variety of datasets for rapid DMR analysis. Our method classifies both the directionality of DMRs and their genome-wide distribution, and we have observed that shows clinical relevance through correct stratification of two Acute Myeloid Leukemia (AML) tumor sub-types.

Conclusions: Our weighted optimization algorithm eDMR for calling DMRs extends an established DMR R pipeline (methylKit) and provides a needed resource in epigenomics. Our method enables an accurate and scalable way of finding DMRs in high-throughput methylation sequencing experiments. eDMR is available for download at http://code.google.com/p/edmr/.

Keywords: Differentially methylated regions, DNA methylation, epigenetics, DMR.

Sponsored Talk

Title:

Illumina's comprehensive End-to-End solutions for Next-Generation Sequencing

Presenter:

Haisheng Nie Ph.D.
Sequencing Specialist
Illumina China

Abstract

Simplifying and accelerating the sample prep process is fundamental to improving sequencing workflows and decreasing turnaround time, an important factor for sequencing to be adopted in broader markets. As the leading NGS company, Illumina recently released several innovations in sample preparation technology, including TruSeq DNA PCR-Free Kits, TruSeq Stranded-RNA Kits, TruSeq Targeted RNA Kits, Nextera Rapid Capture Kits, and Nextera Mate-Pair Sample Preparation Kits. Illumina also highlighted a number of core sequencing platform enhancements in the near future that will increase the throughput, read lengths, and speed of existing systems, as well as decrease running costs. These improvements will help better adaptation of NGS in both research and applied markets in the very near future.

Poster Presentations

Accepted Posters

ID	Authors and Titles	page
45	Shaoping Ling, Jiahan Liu, Lingtong Hao, Longhui Yin, Lili Dong, Lihua Cao, Wei Zou, Fen Xiao, Junsuo Zhao, Chung-I Wu and Xuemei Lu. *CASmap: Splitting Short Reads Alignment with FPGA-based Streamline Optimization*	38
46	Ling Shaoping, Dong Lili, Cao Lihua, Jia Caiyan, Lu Xuemei and Wu Chung-I. *GRiG: A PPV-sensitive method for predicting somatic SNVs from cancer-normal paired sequencing data with greedy rule induction algorithm*	39
47	Yan Guo, Jiang Li, Yu Shyr and David Samuels. *MitoSeek: Extracting Mitochondria Information and Performing High Throughput Mitochondria Sequencing Analysis*	40
48	Yumin Nie, Hongde Liu and Xiao Sun. *Histone modifications around transcription factor binding sites in human lymphoblastoid cell lines*	41
49	Jie Xiong and Tong Zhou *Structure Identification for Gene Regulatory Networks via Linearization and Robust State Estimation*	42
50	Ning Leng, John Dawson, James Thomson, Victor Ruotti, Anna Rissman, Bart Smits, Jill Haag, Michael Gould, Ron Stewart and Christina Kendziorski. *EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments*	43
51	Xavier Rogé and Xuegong Zhang. *RNAseqViewer: A New Software Program for RNA-seq Data Visualization*	44
52	Wang Lei, Zhou Dequan and Lu Zuhong. *Megakaryocytic/Erythroid differentiation of K562 cell by RNA-Seq*	45
53	Jintao Meng, Bingqiang Wang, Yanjie Wei, Shengzhong Feng, Jiefeng Cheng and Pavan Balaji. *SWAP-Assembler: A Scalable De Bruijn Graph Based Assembler for Massive Genome Data*	46
54	Zhiyi Qin, Weichen Wang and Xuegong Zhang. *DSGseq: A software for detecting differential splicing genes between two groups of samples*	47
56	Dequan Zhou, Wang Lei and Zuhong Lu. *Identification of Egr-1 transcriptional targets using ChIP-Seq*	48
57	Byunghan Lee and Sungroh Yoon. *Preprocessing methods to enhance the quality of diversity estimation for pyrosequenced amplicon samples*	49
58	Roy Lederman. *A Random-Permutations-Based Approach to Fast Read Processing.*	50
59	Antti Ylipää, Kati Waltering, Matti Annala, Kimmo Kartasalo, Leena Latonen, Simo-Pekka Leppänen, Mauro Scaravilli, Wei Zhang, Tapio Visakorpi and Matti Nykter. *Novel prostate cancer specific transcripts identified using RNA-seq*	51
61	Zhihui Liu, Brendan Mumey, Kelly Spendlove, Binhai Zhu and Peiqiang Liu. *An Exact Algorithm for Reconstructing Genomic Scaffolds*	52

62	Qian Xiang. *Impact of DNA Structure on Functional Regulatory Motifs*	53
63	Alexandru I. Tomescu, Anna Kuosmanen, Romeo Rizzi and Veli Mäkinen. *A Novel Combinatorial Method for Estimating Transcript Expression with RNA-Seq: Finding a Bounded Number of Paths*	54
64	Weizhong Li, Sitao Wu and Limin Fu. *Effective computational tools for next generation microbiome sequence analysis*	55
70	Miaomiao Zhao, Guoqin Mai, Zhao Zhang, Youxi Luo and Fengfeng Zhou. *Echo: Evolutionary CHaracterization of SREBP binding mOtifs*	56
86	Xueya Zhou, Suying Bao, Youqiang Song and Xuegong Zhang. *A simple method for detecting cross-sample contamination in deep exome sequencing data*	57
87	Jin Jen, Jin Sung Jang, Karl Oles, Ana Robles, Jaime Davila, Bruce Eckloff, Curtis Harris and Eric Wieben. *High Throughput Mutation Screening of the TP53 Gene in Lung Cancer Using Single Molecule Real Time (SMRT) Sequencing*	58
89	Hongmei Jiang, Lingling An, Zhenyu Zhao, Issac Jenkins and Naruekamol Pookhao. *Statistical methods for comparative metagenomic analysis*	59
90	Jifeng Tang, Erwin Datema, Rui Peng Wang, Alexander Wittenberg, Rolf Mank, Rudie Antonise, Rik Op Den Camp, Peter van Dijk and Antoine Janssen. *PacBio RS long Read Applications in Plant Genomics*	60
91	Layla Oesper, Ahmad Mahmoody and Ben Raphael. *Inferring Intra-Tumor Heterogeneity from Copy Number Aberrations*	61
92	Jeffrey S. Ross, John Curran, Philip J. Stephens, Doron Lipson and Roman Yelensky. *Bringing next generation sequencing to the clinic: Analytical validation and initial deployment of a comprehensive cancer genomic profiling test*	62
93	Tingting Zhu, Eviatar Nevo, Dongfa Sun, Junhua Peng and Xuan Li. *Phylogenetic analyses unravel the evolutionary history of NAC proteins in plants*	63

CASmap: Splitting Short Reads Alignment with FPGA-based Streamline Optimization

Shaoping Ling [1], Jiahan Liu [2], Lingtong Hao [1], Longhui Yin [3], Lili Dong [1], Lihua Cao [1], Wei Zou [1], Fen Xiao [3], Junsuo Zhao [2], Chung-I Wu [1,*], Xuemei Lu [1,*]

[1]: Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China.

[2]: Institute of Software, Chinese Academy of Sciences, Beijing, China.

[3]: Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education, Xiangtan University, Xiangtan, China.

[*]: To whom correspondence should be addressed.

Emails: SL (spling@big.ac.cn); JL (liujiahan@gmail.com); LH (haolt@big.ac.cn) ; LY (yinlonghui.big@gmail.com) ; LD (donglili@big.ac.cn) ; LC (caolh@big.ac.cn) ; WZ (zouwei@big.ac.cn) ; FX (xiaof@xtu.edu.cn) ; JZ (junsuo@iscas.ac.cn) ; CW (wuci@big.ac.cn) ; XL (luxm@big.ac.cn)

Short reads alignment, as a core computational issue in HTS (High-Throughput Sequencing) data analysis, it is a bottle-neck in HTS data real-time clinical applications. We created a new alignment system (CASmap) which implemented BWT-based alignment algorithm in a customized desktop reconfigurable computer based on FPGA reconfigurable platform. It accelerated ~30X and ~2X higher than BWA (one thread) and SOAP3 in alignment in suffix array with the power of FPGA-based streamline optimization. Multi-threading parallelization of smith-waterman algorithm was implemented in multi-core host of CPU to verify the location of reads. CASmap achieved the high speed of ~310Gbp/day,cpu in aligning real human whole genome sequencing Hiseq pair-end reads (4 mismaches/read, 2×100bp) with the low power consumption of ~30w/Gb and high accuracy of 99%. CASmap, as an efficient reconfigurable heterogeneous computing system for short read alignment, provided a new green computing framework for HTS genome re-sequencing projects and a solution for real-time HTS data application.

References

1. Li H. and Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics, Epub.* (2010)
2. C. Liu, T. Wong, E. Wu, R. Luo, S. Yiu, Y. Li, B. Wang,C. Yu, X. Chu, K. Zhao, R. Li, and T. Lam, SOAP3:Ultra-fast GPU-based parallel alignment tool for short reads, *Bioinformatics*. (Jan. 2012)

GRiG: A PPV-sensitive method for predicting somatic SNVs from cancer-normal paired sequencing data with greedy rule induction algorithm

Shaoping Ling[1,$], Lili Dong [1,$], Lihua Cao[1], Caiyan Jia[2,3], Xuemei Lu[1*], Chung-I Wu[1,4*]

[1]: Beijing Institute of Genomics, Chinese Academy of Scineces, Beijing, China.

[2]: School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China.

[3]: Department of Bioengineering/Bioinformatics, University of Illinois at Chicago, Chicago, IL 60612, USA.

[4]: Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA.

[$]: These authors contributed equally to this work.

[*]: To whom correspondence should be addressed.

Emails: SL (spling@big.ac.cn); LD (donglili@big.ac.cn); LC (caolh@big.ac.cn); CJ (caiyan.jia@gmail.com); XL (luxm@big.ac.cn); CW (wuci@big.ac.cn)

Predicting somatic SNVs from cancer-normal paired sequencing is a key computational issue in high-throughput sequencing-driven cancer genomics. Classic methods based on statistical inference (SI) have been developed and become standard pipeline in human variation detection. However, they can not provide enough high positive prediction value (PPV) for further experimental validation and function analysis. We presented a Greedy Rule Induction alGorithm (GRiG) for predicting somatic SNVs in cancer-normal paired sequencing data, which integrates feature selection and rule inference into a machine learning frame work. We evaluated the performance of GRiG on public datasets which consist of two candidate somatic SNVs datasets from 48 breast exome capture sequencing (ECS) datasets and 4 whole genome sequencing (WGS) datasets for training and testing respectively. GRiG always achieved the better performance in ECS training dataset with 10x cross-validation and WGS testing dataset than both Samtools and GATK and presented comparable performance with four statistical learning algorithms including random forest, Bayesian additive regression tree, support vector machine and logistic regression in ECS training dataset with 10x cross-validation and WGS testing dataset. Especially, it always achieved better PPV than these four classifiers.

References

1. Ding,J. et al. (2012) Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics*, 28, 167–175.
2. McKenna,A. et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20, 1297–1303.
3. Li,H. et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.
4. Greenman,C. et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature*.

MitoSeek: Extracting Mitochondria Information and Performing High Throughput Mitochondria Sequencing Analysis

Yan Guo[1*], Jiang Li[1], Yu, Shyr[1], David C. Samuels[2]

1. Vanderbilt Ingram Cancer Center, Center for Quantitative Sciences, Nashville, TN
2. Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN

ABSTRACT

Motivation: Exome capture kits have capture efficiencies that range from 40% to 60%. A significant amount of off-target reads are from the mitochondrial genome. These unintentionally sequenced mitochondrial reads provide unique opportunities to study the mitochondria genome.

Results: MitoSeek is an open-source software tool which can reliably and easily extract mitochondrial genome information from exome and whole genome sequencing data. MitoSeek evaluates mitochondrial genome alignment quality, estimates relative mitochondrial copy numbers, and detects heteroplasmy, somatic mutation, and structural variants of the mitochondrial genome. MitoSeek can be set up to run in parallel or serial on large exome sequencing datasets.

Availability: https://github.com/riverlee/MitoSeek

Histone modifications around transcription factor binding sites in human lymphoblastoid cell lines

Yumin Nie[1], Hongde Liu[1], Xiao Sun[1,*]

[1]: State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China

[*]: To whom correspondence should be addressed.

Emails: YN (nieyum@126.com); HL (liuhongde@seu.edu.cn); XS (xsun@seu.edu.cn)

Transcription factor (TF) binding at specific DNA sequences is the fundamental step in transcriptional regulation and is highly dependent on the chromatin structure context, which may be affected by specific histone modifications and variants. Previous studies have focused mainly on histone marks at binding sites for several specific TFs. We therefore studied 11 histone marks around binding sites for 164 and 34 TFs, which were determined by CENTIPEDE algorithm and ChIP-seq technology, respectively, in human lymphoblastoid cell lines. For H2A.Z, methylation of H3K4, and acetylation of H3K27 and H3K9, the mark patterns exhibited bimodal distributions and strong pairwise correlations in the 600-bp region around enriched sites, suggesting that these marks mainly coexist within the two nucleosomes proximal to the TF sites. Mark H3K79me2 showed a unimodal distribution on one side of binding sites and the signals extended up to 4000 bp, indicating a longer-distance pattern. Interestingly, H4K20me1, H3K27me3, H3K36me3 and H3K9me3, which were more diffuse and less enriched surrounding the binding sites, showed unimodal distributions around the enriched sites, suggesting that some TFs may bind to nucleosomal DNA. In conclusion, this study demonstrated the ranges of histone marks associated with TF binding, and the common features of these marks around the binding sites. These findings have epigenetic implications for future analysis of regulatory elements.

Structure Identification for Gene Regulatory Networks via Linearization and Robust State Estimation

Jie Xiong[1], Tong Zhou[1,2]

1. Department of Automation, Tsinghua University, Beijing, 100084, P. R. China
E-mail: xiongj08@mails.tsinghua.edu.cn

2. Tsinghua National Laboratory for Information Science and Technology, Beijing, 100084, P. R. China
E-mail: tzhou@mail.tsinghua.edu.cn

Abstract: Inferring causal relationships among numerous cellular components is one of the fundamental problems in understanding biological behaviors. The gene regulatory network is widely considered as a nonlinear dynamic stochastic model that consists of the gene measurement equation and the gene regulation equation, in which the extended Kalman filter (EKF) is sometimes used for estimating both the model parameters and the actual value of gene expression levels. However, first-order linearization usually results in modeling errors, but the EKF based method does not take either unmodelled or parametric uncertainty into account. As a result, the estimation performance of the EKF based method may not be satisfactory, such as slow convergence speed and low estimation accuracy. To overcome these problems, a sensitivity penalization based robust state estimator is suggested for reconstructing the structure of a gene regulatory network. The suggested method has been used to identify some parameters of a nonlinear state-space system and recovery an artificially gene regulatory network. Compared with the widely adopted EKF based method, computation results show that parametric estimation accuracy can be significantly increased and false positive errors can be greatly reduced.

Key Words: Gene Regulatory Networks, Nonlinear State-Space Model, Extended Kalman Filter, Sensitivity Penalization Based Robust State Estimation

EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq

Ning Leng [1,*], John Dawson [1], James Thomson [2], Victor Ruotti [2], Anna Rissman [3], Bart Smits [3], Jill Haag [3], Michael Gould [3], Ron Stewart [2], Christina Kendziorski [4]

[1]: Department of Statistics, University of Wisconsin, Madison, WI

[2]: Morgridge Institute for Research, Madison, WI

[3]: McArdle Laboratory for Cancer Research, Department of Oncology, University of Wisconsin, Madison, WI

[4]: Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI

[*]: To whom correspondence should be addressed.

Emails: NL (nleng@wisc.edu); JD (jadawson@wisc.edu); JT (jthomson@morgridgeinstitute.org); VR (ruotti@wisc.edu); AR (rissman@wisc.edu); BS (bsmits@wisc.edu); JH (jhaag@facstaff.wisc.edu); MG (gould@oncology.wisc.edu); RS (rstewart@morgridgeinstutute.org); CK(kendzior@biostat.wisc.edu)

Motivation: Messenger RNA expression is important in normal development and differentiation, as well as in manifestation of disease. RNA-seq experiments allow for the identification of differentially expressed (DE) genes and their corresponding isoforms on a genome-wide scale. However, statistical methods are required to ensure that accurate identifications are made. A number of methods exist for identifying DE genes, but far fewer are available for identifying DE isoforms. When isoform DE is of interest, investigators often apply gene-level (count-based) methods directly to estimates of isoform counts. Doing so is not recommended. In short, estimating isoform expression is relatively straightforward for some groups of isoforms, but more challenging for others. This results in estimation uncertainty that varies across isoform groups. Count-based methods were not designed to accommodate this varying uncertainty and consequently application of them for isoform inference results in reduced power for some classes of isoforms and increased false discoveries for others.
Results: Taking advantage of the merits of empirical Bayesian methods, we have developed EBSeq for identifying DE isoforms in an RNA-seq experiment comparing two or more biological conditions. Results demonstrate substantially improved power and performance of EBSeq for identifying DE isoforms. EBSeq also proves to be a robust approach for identifying DE genes.
Availability: An R package containing examples and sample data sets is available at http://www.biostat.wisc.edu/~kendzior/EBSEQ/

References

N. Leng, J.A. Dawson, J.A. Thomson, V. Ruotti, A.I. Rissman, B.M.G. Smits, J.D. Haag, M.N. Gould, R.M. Stewart, and C. Kendziorski. EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 2013

RNAseqViewer: A New Software Program for RNA-seq Data Visualization

Xavier Rogé [1], Xuegong Zhang [1,2,*]

[1]: MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China.

[2]: School of Life Sciences, Tsinghua University, Beijing 100084, China.

[*]: To whom correspondence should be addressed.

Emails: XR (xavier.roge@gmail.com); XZ (zhangxg@tsinghua.edu.cn)

New advances in RNA sequencing have opened up new horizons in the field of transcriptomics and given access to new extensive data. The analysis of these data needs effective visualization tools, so as scientists can gain an insight into the data and are able to review the results of the computational tools. We developed a new software program, RNAseqViewer, to visualize the various data from the RNA-Seq analyzing process for single or multiple samples. By focusing on expression of genes and transcript isoforms, the program offers innovative ways to present the transcriptome data in a quantitative and interactive manner.

RNAseqViewer currently supports 7 types of data: read alignments (SAM/BAM format) and junction reads (BED), which can be provided by RNA-Seq mappers like TopHat; transcripts (GTF), which can be computed by tools like Cufflinks; numeric data (Wiggle); reference sequences (FASTA) and annotations (RefSeq); and generic BED tracks. Different types of view for each data set allow the visualization of different levels of information, including heatmap-like views for informative and yet very compact tracks, making possible to visualize dozens of samples simultaneously. Special attention has been given to the user interface, so that the data can be explored in a fast and intuitive way, and to the memory management, so that very large data sets can be visualized without exceeding memory limits nor affecting the fluidity of the user interface.

The software is a handy tool for scientists who use RNA-Seq data to compare multiple transcriptomes, for example, to compare gene expression and alternative splicing of cancer samples or of different development stages.

Megakaryocytic/Erythroid differentiation of K562 cell by RNA-Seq

Lei Wang[1], Dequan Zhou[1], Zuhong Lu[1]*
[1]State Key Lab for Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China
* Corresponding authors: zhlu@seu.edu.cn

The human erythroleukemia cell line K562, derived from a chronic myelogenous leukemia patient, resembles a bipotent MEP. K562 cell therefore has been used as an important model of common progenitor of erythroblasts and magakaryocytes and can be differentiated into erythroid or megakaryocytic lineages by hemin/PMA, respectively[1, 2]. Erythroid and megakaryocytic lineages represent two of the terminally differentiated cell types stemed from hematopoietic stem cells. Erythroid lineages eventually produce a biconcave discoid structure cell dedicated to the delivery of oxygen to the tissues. On the other hand, megakaryocytes ultimately release platelets which subsequently mediate hemostasis and thrombosis. Elucidation of differentiation mechanism could lead to development of therapeutic agents for treatment of related erythroblast and megakaryocytes disease, such as sickle cell disease (SCD) and the β-thalassemia syndromes[3].

Analysis of gene expression profile in erythro-megakaryocytic lineage divergence is necessary for an understanding gene regulation network of hematopoietic differentiation and would also provide insights into the consequences of gene expression change in hematopoietic diseases. Gene expression change can be impacted by DNA methylation, miRNA, and transcription factor, gene fusion, and so on, whose crosstalk also mediate erythro-megakaryocytic differentiation. In our previous studies, we have obtained ChIP-Seq data of the transcription factor ERG1 and C-Jun in PMA induced K562 cell, and we detected nearly hundreds of promote regions of miRNA genes harboring EGR1 binding sites. MeDIP-seq can reveal the DNA methylation information on the whole genome level. We have obtained the DNA methylation spectrum in K562 cell. In order to understand the whole stories of the gene regulation inside a cell, integration of the omics data-sets from different platforms is necessary.

To elucidate the mechanisms involved in terminal erythroid and megakaryocytic differentiation, we used two series dynamic change RNA-Seq data by PMA/hemin to compare the difference of regulated genes, the long noncoding RNAs, alternative splicing and gene fusion by combining genomic information, at several stages of erythroid and megakaryocytic differentiation of K562 cells. Our goal is to build the gene regulation networks of several important cellular process of K562 cells related to blood cells differentiation, provide useful clues for understanding the mechanisms of erythroid and megakaryocytic differentiation and developing therapy strategies of the diseases related to hemopoiesis differentiation.

References

[1] J. K. Limb, S. Yoon, K. E. Lee, *et al*. Regulation of megakaryocytic differentiation of k562 cells by fosb, a member of the fos family of ap-1 transcription factors [J]. *Cell Mol Life Sci*, 66(11-12): 1962-1973，2009.

[2] S. Addya, M. A. Keller, K. Delgrosso, *et al*. Erythroid-induced commitment of k562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements [J]. *Physiol Genomics*, 19(1): 117-130,2004.

[3] M. Brand, J. A. Ranish, N. T. Kummer, *et al*. Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics [J]. *Nature structural & molecular biology,* 11(1): 73-80,2003.

SWAP-Assembler: A Scalable De Bruijn Graph Based Assembler for Massive Genome Data

(Submitted for RECOMB-SEQ Poster Track)

Jintao Meng
Shenzhen Institutes of Advanced Technology, CAS
Shenzhen, P.R. China
jt.meng@siat.ac.cn

Bingqiang Wang
Beijing Genomics Institute
Shenzhen, P.R. China
wangbingqiang@genomics.cn

Yanjie Wei*, Shengzhong Feng*, Jiefeng Cheng
Shenzhen Institutes of Advanced Technology, CAS
Shenzhen, P.R. China
{yj.wei,sz.feng, jf.cheng}@siat.ac.cn

Pavan Balaji
Mathematics and Computer Science Division
Argonne National Laboratory
balaji@mcs.anl.gov

Abstract

Sequencing species with large genome can produce Tara bytes data, and the de bruijn graph constructed from these data - in some cases having ten billions of vertices and edges - poses challenges to genome assembly problem. This paper presents a multi-step bi-directed graph (MSG) to abstract the standard genome assembly (SGA) problem. With MSG, SGA can be decomposed into several edge merging operations, and this operation and the multi-step semi-extended edges are proved to be a semi-group. Afterwards a small world asynchronous parallel model (SWAP), which can automatically detect and make use of the locality of computation and communication in semi-group to maximize potential parallelism, is proposed for this type of computation. With MSG and SWAP, SWAP-assembler is developed, the scalability test shows that it can scale up to 1024 cores with improved performance, the 2008 Asian (YanHuang) genome can be assembled in 2 hours, which is 6 times faster than SOAPdenovo on one server with 32 cores, and about 24 times faster than ABySS with 1024 cores.

Keywords: De Bruijn Graph, Parallel Computing, Genome Assembly, Semi-group

DSGseq: A software for detecting differential splicing genes between two groups of samples

Zhiyi Qin[1], Weichen Wang [1,2], Xuegong Zhang[1,3,*]

[1]: MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China

[2]: Department of Operational Research and Financial Engineering, Princeton University, NJ 08544, USA

[3]: School of Life Sciences, Tsinghua University, Beijing 100084, China

[*]: To whom correspondence should be addressed.

Emails: QZ (qzy06@mails.tsinghua.edu.cn); WW (weichenw@princeton.edu); ZX (zhangxg@tsinghua.edu.cn)

Recent study revealed that most human genes have alternative splicing through RNA-seq. As differences in the relative abundance of the isoforms of a gene can have significant biological consequences, identifying genes that are differentially spliced between two groups of samples is having been an important task in the study of transcriptomes with next-generation sequencing technology. There have been several methods aimed to detect differential splicing genes, but most of them were designed for comparing two individual samples. Recently we studied the question of identifying genes that are differentially spliced between two groups of samples and proposed one exon-based method used an NB-statistic with the negative binomial model to detect differential splicing [1]. This was a new route to study alternative splicing quantitatively in an exon-centric manner.

We implemented this method named DSGseq, which can detect differentially spliced genes between two groups of samples. It is designed for comparing two groups of RNA-seq samples and does not need to infer isoform structure or to estimate isoform expression. With counting exons' covered reads and then applying the software DSGseq, one can identify differentially spliced genes between two groups of samples, as well as the exons that contribute the most to the differential splicing. Simulation experiments showed that the proposed method performs well on both detecting differentially spliced genes and identifying the alternative exons. Experiments on real RNA-seq data of human kidney and liver samples illustrated the method's good performance and applicability. DSGseq is written in R and can run on all major computer platforms running Windows or Unix/Linux. The software tool is available at: http://bioinfo.au.tsinghua.edu.cn/software/DSGseq for free academic use.

References

1. Weichen Wang, Zhiyi Qin, Zhixing Feng, Xi Wang and Xuegong Zhang, 2012. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene*, http://dx.doi.org/10.1016/j.gene.2012.11.045.

Identification of Egr-1 transcriptional targets using ChIP-Seq

Dequan Zhou, Xiaolong Shi, Lei Wang, Zuhong Lu*
State Key Laboratory of Bioelectronics, Southeast University, Nanjing, 210096, P. R. China
* To whom correspondence should be addressed: zhlu@seu.edu.cn

Early growth response gene 1 (Egr-1) has been implicated in megakaryocyte differentiation induced by phorbolester [1]. But the molecular mechanism of Egr-1 in this process has not been widely investigated. The identification of direct Egr-1 target genes in genomic level is critical for our understanding of how Egr-1 exerts effects globally. The human erythroleukemia cell line K562, resembles the bipotent MEP, has been used as an important transcriptional model of common progenitor of erythroblasts and magakaryocytes and can be induced to differentiate into erythroid or megakaryocytic lineages by hemin or PMA, respectively [2]. In this study, we first provide a global survey on the binding location of Egr-1 in K562 cells using chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP–Seq) [3]. Over 14 000 highly confident *in vivo* Egr-1 binding sites in PMA-treated K562 cells were identified. More than 70% of the genomic areas associated with Egr-1 binding were located around annotated gene regions. We also investigated the relationship between Egr-1 binding and gene expression in K562 cells treated with PMA by overlapping Egr1-binding sites with genes expression profile, and found that 1163 of the 4428 genes displaying differential expression in the differentiated K562 cells were associated with Egr-1 binding, suggesting that Egr-1 is a key regulator in K562 cells differentiation induced by PMA.

Of the target genes revealed by genome-wide ChIP-seq analysis, *Sox5*, *FOXA3*, *Smad7*, and a member of E2F family, *E2F5* were then selected as potential transcriptional targets of Egr-1. Detailed analysis of the promoters defined the Egr-1-binding sites by EMSA and ChIP. Additionally, serial deletion of each promoter and site-mutation of Egr-1-binding sequence resulted in reduced or enhanced stimulation by Egr-1. Taken together, the results first suggest that *Sox5*, *FOXA3*, *Smad7* and *E2F5* are transcriptional targets of Egr-1, further implicating that Egr-1 may act as a key regulator in regulation networks of K562 differentiation.

References

1. Cheng T., et al., Transcription factor EGR1 is involved in phorbol 12-myristate 13-acetate-induced megakaryocytic differentiation of K562 cells. *J Biol Chem*, 1994. **269**(49): 30848-30853.

2. Green AR., et al., Induced myeloid differentiation of K562 cells with downregulation of erythroid and megakaryocytic transcription factors: a novel experimental model for hemopoietic lineage restriction. *Exp Hematol*, 1993, **21**(4): 525-531.

3. Tang C., et al., Global analysis of in vivo EGR1-binding sites in erythroleukemia cell using chromatin immunoprecipitation and massively parallel sequencing. *Electrophoresis*, 2010, **31**(17): 2936-2943.

Preprocessing methods to enhance the quality of diversity estimation for pyrosequenced amplicon samples

Byunghan Lee [1] and Sungroh Yoon [1,2,*]

[1]: Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Korea.

[2]: Bioinformatics Institute, Seoul National University, Seoul 151-747, Korea.

[*]: To whom correspondence should be addressed.

Emails: B. L. (styxkr@snu.ac.kr); S. Y. (sryoon@snu.ac.kr)

To analyze pyrosequenced amplicon data in metagenomics, we need to filter or denoise the sequencing data before estimating the diversity appearing in a given sample. Raw sequenced data often contain problematic sequences, which may lead to overestimation [1]. To avoid that, there exist computational tools for preprocessing (*i.e.*, filtering or denoising) pyrosequenced data. Most of these tools utilize nucleotide sequence data, whereas some techniques rely on flow data. In preprocessing, removing erroneous reads and filtering out duplicates can affect the diversity estimation. In addition to preprocessing, the procedure used to identify operational taxonomic units (OTUs) may also bias the estimation, but here we only consider the effect of preprocessing. We compare existing preprocessing approaches [2,3,4] and examine their effectiveness in terms of accuracy and efficiency.

References

1. Victor Kunin, Anna Engelbrektson, Howard Ochman, Philip Hugenholtz. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, 12: 118-123, 2010.
2. Susan M. Huse, Julie A. Huber, Hilary G. Morrison, Mitchell L. Sogin, David M. Welch. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8(7): R143, 2007.
3. Susanne Balzer, Ketil Malde, Markus A. Grohme, Inge Jonassen. Filtering duplicate reads from 454 pyrosequencing data. *Bioinformatics*, in press.
4. Weizhong Li, Limin Fu, Beifang Niu, Sitao Wu, John Wooley. Ultrafast Clustering Algorithms for Metagenomic Sequence Analysis. *Briefings in Bioinformatics*, 13(6): 656-668, 2012.

A Random-Permutations-Based Approach to Fast Read Processing

Roy Lederman [1]

[1]: Applied Mathematics Program, Yale University.

Email: roy.lederman@yale.edu

Read alignment and assembly are computationally expensive steps in the processing of NGS reads. Existing read alignment programs use prefix-tree algorithms and hash-table algorithms.

We present a new approach to read alignment which uses random permutations of strings. This randomized approach is flexible, accurate and fast. We present experimental results to demonstrate that random-permutations-based algorithms can successfully align significantly more reads than comparable programs in significantly shorter run times.

We demonstrate the flexibility of permutations-based algorithms by extending them to other applications, such as assembly.

We also describe "homopolymer-length-filters," a separate method of read processing which allows random-permutations-based algorithms to also process 454/IonTorrent reads rapidly and accurately.

Our paper "A Random-Permutations-Based Approach to Fast Read Alignment" will be presented at RECOMB-seq.
Technical reports and more information: http://alignment.commons.yale.edu.

Novel prostate cancer specific transcripts identified using RNA-seq

Antti Ylipää[1,*], Kati K Waltering[1], Matti J Annala[1], Kimmo Kartasalo[1], Leena Latonen[2], Simo-Pekka Leppänen[1], Mauro Scaravilli[2], Wei Zhang[3], Tapio Visakorpi[2], Matti Nykter[2,*]

[1]: Tampere University of Technology, Tampere, Finland.

[2]: University of Tampere, Tampere, Finland.

[3]: The University of Texas, MD Anderson Cancer Center, Houston, Texas, USA.

[*]: To whom correspondence should be addressed.

Emails: AY (antti.ylipaa@tut.fi); KKW (kati.waltering@tut.fi); MJA (matti.annala@tut.fi); KK (kimmo.kartasalo@tut.fi); LL (leena.latonen@uta.fi); S-PL (simo-pekka.leppanen@tut.fi); MS (mauro.scaravilli@uta.fi); WZ (wzhang@mdanderson.org); TV (tapio.visakorpi@uta.fi); MN (matti.nykter@uta.fi)

The poster abstract should include the Prostate cancer is the third most common cause of male cancer deaths in developed countries, with castration resistance being the most challenging clinical problem. Here we report an investigation into novel transcripts in primary prostate cancers (PCs), and castration resistant prostate cancers (CRPCs) in particular. We characterized 28 PCs, 13 CRPCs, and 12 benign prostatic hyperplasias (BPH) using deep transcriptome sequencing (RNA-seq). Reference-based transcriptome assembly uncovered 145 previously unannotated intergenic PC associated transcripts or isoforms. The expression patterns of the transcripts were confirmed in two previously published independent cohorts of primary tumors (n=30 and n=34), 21 PC cell lines, and 25 normal tissues or cell lines. By integrating publicly available ChIP-sequencing data and transcription factor (TF)-transcript expression correlations, we identified a transcript that positively correlated with ERG expression and exhibited an ERG binding event in a PC cell line coinciding with the canonical ETS-family TF binding motif at its proximal promoter region. Enrichment of histone modification H3K4me3 and PolII at the promoter of the transcript in the cell line provided further evidence of open chromatin and active transcription. We downregulated the expression of the transcript with siRNAs in the cell line and observed a decrease in cell growth and reduced migration, invasion and colony formation. Annexin V assay indicated increased rate of apoptosis in the cells. Pathway analysis indicated that cell cycle, mitosis and apoptosis were the most extensively affected cellular processes. These results suggested that the transcript significantly affects tumor growth in ETS-positive prostate cancers.

An Exact Algorithm for Reconstructing Genomic Scaffolds

Zhihui Liu [1], Brendan Mumey [2], Kelly Spendlove [3], Binhai Zhu [2,*] Peiqiang Liu [1]

[1]: School of Mathematical and Information Sciences, Shandong Institute of Business and Technology, Yantai, China.

[2]: Department of Computer Science, Montana State University, Bozeman, MT 59717, USA.

[3]: Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717, USA.

[*]: To whom correspondence should be addressed.

Emails: ZL (dane.zhihui.liu@gmail.com); BM (mumey@cs.montana.edu);

KS (kelly.spendlove@gmail.com); BZ (bhz@cs.montana.edu); PL (liupq@126.com)

Current de novo genome sequence assemblers use paired-end read data from high throughput next-generation technology in order to reconstruct genomes. These assemblers produce longer contiguous sequences known as contigs. Using the paired-end read data it becomes possible to assess the relative order and orientation of the contigs; a process known as scaffolding. As scaffolding is a complex combinatorial problem, most modern assemblers use heuristic scaffolding subroutines within larger implementations. This research, in contrast to the heuristic approaches, concerns producing an optimal solution to the scaffolding problem. We first show that the scaffolding problem is NP-hard. We then prove certain cases of the problem are in FPT. Finally, we present an algorithm for the general problem which exploits the fixed parameter tractability of these certain inputs. With the real datasets, the actual running time of this algorithm is still too slow to be practical. So, at last, we point out several possible ways to tune the algorithm to work for real datasets.

This research is supported by NSF of China (grant No. 60928006) and by the Natural Science Foundation of Shandong Province of China (grant No. ZR2011FL004).

Impact of DNA Structure on Functional Regulatory Motifs

Qian Xiang
School of Information Science and Technology, Sun Yat-Sen University
Guangzhou, PR China

QX(xiangq@mail.sysu.edu.cn)

The three-dimensional structure of DNA has been proposed to be a major determinant for functional TF-DNA interaction, as it is a critical feature recognized by the regulatory machinery within a cell. Here we use hydroxyl radical cleavage pattern as a measure of local DNA structure, and the regulatory protein may recognize a variety of divergent nucleotide sequences which adopt the same local structures. We compared the conservation between DNA sequence and structure in terms of information content and attempted to assess the functional implications of DNA structures on regulatory motifs. We used statistical methods to evaluate the structural divergence of substituting a single position within a binding site and applied them to a collection of putative regulatory motifs. The following are our major observations: (i) We observed more information in structural alignment than the corresponding nucleotide sequence alignment for most of the transcriptional factors; (ii) For each TF, majority of positions have more information in the structural alignment as compared to the nucleotide sequence alignment; (iii) We further defined a DNA structural divergence score (SD-score) for each wild-type and mutant pair that are distinguished by single base mutation. The SD-score for benign mutations are significantly lower than that of switch mutations. This indicates structural conservation is important for TFBS to be functional. Based on these findings, we speculate that some of the functional information in the TFBS is conferred by DNA structure as well as by nucleotide sequence. DNA structures will provide previously unappreciated information for TF to realize the binding specificity. Our results should facilitate the prediction of the divergent functional TF regulatory interaction with binding site variations by altering the DNA structure.

A Novel Combinatorial Method for Estimating Transcript Expression with RNA-Seq: Finding a Bounded Number of Paths

Alexandru I. Tomescu [1,*], Anna Kuosmanen [1], Romeo Rizzi [2], Veli Mäkinen [1]

[1]: Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland.

[2]: Department of Computer Science, University of Verona, Italy.

[*]: To whom correspondence should be addressed.

Emails: AIT (tomescu@cs.helsinki.fi); AK (aekuosma@cs.helsinki.fi); RR (romeo.rizzi@univr.it); VM (vmakinen@cs.helsinki.fi);

In [1] we modeled the problem of transcript identification and quantification with RNA-Seq as that of finding the paths and their expression levels which best explain the coverage of a splicing graph, under very general fitness models. One such fitness model is least sum of squares. Without imposing a bound on the number of paths, this problem can be solved in polynomial time by reducing it to a min-cost network flow problem.

In the current work we impose a bound on the number of paths required to explain the splicing graph, and show that the problem becomes NP-hard in the strong sense. Moreover, we propose fast dynamic programming algorithms, which are given a k-tuple of expression levels and find the best k paths explaining the splicing graph with these expression levels. They work in polynomial time assuming that the number k of paths is bounded. The optimal expression levels can then be found with any local search method; we currently employ a genetic algorithm. Experimental evaluation on simulated data shows that these algorithms, at the expense of more computational resources, outperform our min-cost flow method of [1] which in turn outperforms state-of-the-art tools such as Cufflinks and IsoLasso.

References

1. Alexandru I. Tomescu, Anna Kuosmanen, Romeo Rizzi, Veli Mäkinen. A Novel Min-Cost Flow Method for Estimating Transcript Expression with RNA-Seq. *To appear in BMC Bioinformatics*, proceedings paper from *RECOMB-Seq 2013*.

Effective computational tools for next generation microbiome sequence analysis

Weizhong Li [1,*], Sitao Wu[1], Limin Fu[1]

[1]: University of California San Diego, Center for Research in Biological Systems, La Jolla California, USA

[*]: To whom correspondence should be addressed.

Emails: WL (liwz@sdsc.edu); SW (siw006@ucsd.edu); LF (l2fu@ucsd.edu)

Complex and dynamic microbial communities play a profound role in shaping the environment they inhabit. Recent advances in metagenomics approaches and high-throughput Next Generation Sequencing (NGS) technologies have enabled comprehensive study of the microbes from many diverse environments, such as the human microbiota that are thought to deeply influence human health. Vast amounts of sequences being generated impose extreme challenges in computational analyses such as sequence error correction, assembly, mapping, gene prediction, and function analysis. To address these challenges, we build a set of unique NGS-specific tools using fast clustering and alignment algorithms combined with statistical methods and visualization interface[1-11]. These tools not only allow orders of magnitude faster computational analysis but also offer inimitable way to investigate novel and complex data. We have analyzed several Terabytes of sequence for hundreds of human microbiome samples from healthy people and patients with diseases and obtained comprehensive results from our analysis.

References

1. Zhu Z, Niu B, Chen J, Wu S, Sun S, Li W: MGAviewer: A desktop visualization tool for analysis of metagenomics alignment data. *Bioinformatics* 2012.
2. Li W, Fu L, Niu B, Wu S, Wooley J: Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief Bioinform* 2012, 13(6):656-668.
3. Fu L, Niu B, Zhu Z, Wu S, Li W: CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012, 28(23):3150-3152.
4. Wu S, Zhu Z, Fu L, Niu B, Li W: WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 2011, 12:444.
5. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J *et al*: Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* 2011, 39(Database issue):D546-551.
6. Niu B, Zhu Z, Fu L, Wu S, Li W: FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* 2011, 27(12):1704-1705.
7. Niu B, Fu L, Sun S, Li W: Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* 2010, 11:187.
8. Huang Y, Niu B, Gao Y, Fu L, Li W: CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010, 26(5):680-682.
9. Huang Y, Gilna P, Li WZ: Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* 2009, 25(10):1338-1340.
10. Li W, Wooley JC, Godzik A: Probing metagenomics by rapid cluster analysis of very large datasets. *PLoS ONE* 2008, 3(10):e3375.
11. Li W: Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* 2009, 10:359.

Echo: Evolutionary CHaracterization of SREBP binding mOtifs

Miaomiao Zhao [†,1], Guoqin Mai [†,1], Zhao Zhang [†,1,2], Youxi Luo [1,3], Fengfeng Zhou [*,1]

[1]: Shenzhen Institutes of Advanced Technology, and Key Laboratory of Health Informatics, Chinese Academy of Sciences, Shenzhen, Guangdong, China, 518055.

[2]: Tianjin Polytechnic University, Tianjin, China, 300387.

[3]: School of Science, Hubei University of Technology, Wuhan, Hubei, China, 430068.

[*]: Corresponding author: Fengfeng Zhou, Emails: FengfengZhou@gmail.com or ff.zhou@siat.ac.cn. Web site: http://www.HealthInformaticsLab.org/ffzhou/ .

[†]: These authors contribute equally to this work.

Transcriptional regulatory network (TRN) orchestrates biological activities in a cell by receiving regulatory signals from the signal transduction network and transcribing specific responding genes for the signals. TRN consists of directed transcription regulatory relationships from transcription factors and their substrate genes. Sterol regulatory element binding proteins (SREBPs) are transcription factors (TFs) involved in the lipid balance regulation, by controlling the expression of synthesis enzymes for endogenous cholesterol, fatty acid (FA), triacylglycerol and phospholipid. SREBPs are known to play an essential role in the development of atherosclerosis, cardiovascular diseases, obesity, diabetes mellitus, fatty liver, hypercholesterolemia, insulin resistance and lipodystrophy. The sterol regulatory elements (SREs) are the substrates regulated by the SREBPs, and their regulatory binding motifs may be detected by the ChIP-seq technology, combining the chromatin immunoprecipitation (ChIP) technology and the DNA sequencing technology. Considering the dynamic nature of SREBP binding activity and the ChIP limitation of single-TF-per-run, there is still a major requirement for the prediction of transcription factor binding sites (TFBSs). This study proposed an evolutionary algorithm (Echo) to iteratively optimize the TFBS prediction model. Based on the known TFBSs of four SREBPs obtained from TRANSFAC, Echo outperforms the widely used Position Weight Matrix (PWM) algorithm by 2-30% in sensitivity at the similar level of specificity. Echo also achieves at least 97.8% for both sensitivity and specificity for all the four SREBPs. A large-scale characterization of SREBP regulatory dynamics is underway to complement the ChIP-seq data sets.

Key words

ChIP-seq, transcription factor binding site, evolutionary algorithm.

A simple method for detecting cross-sample contamination in deep exome sequencing data

Xueya Zhou[1*], Suying Bao[2], Youqiang Song[2], and Xuegong Zhang[13*]

[1]: MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST / Department of Automation, Tsinghua University, Beijing 100084, China

[2]: Department of Biochemistry, Li Ka-Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China

[3]: School of Life Sciences, Tsinghua University, Beijing 100084, China

[*]: To whom correspondence should be addressed.

Emails: X.Y.Z (zhou-xy02@mails.tsinghua.edu.cn); S.B (sybao@hku.hk); Y.S (songy@hku.hk); X.G.Z (zhangxg@tsinghua.edu.cn)

Cross-sample contamination is a potential issue in exome sequencing which will cause genotype misclassification, create false positive variant calls, and confound the detection of somatic mutations. Whereas high level contamination or sample mixing can be easily discovered from unusual heterozygosity values; low level contamination poses a more difficult statistical problem. Motivated by our observations in ongoing exome projects, we found that the major effect of low level cross-sample contamination was to create false positive heterozygous variants with lower than expected proportion of alternative alleles in supporting reads. The resulting allele balance (AB) histogram of heterozygous SNP calls deviates from the expected normal distribution with mean 0.5 to become a mixture distribution. We propose to test the mixture proportion in AB histogram as a simple method to quantify the level of contamination. When applied to the simulated exome data, the proposed method showed similar performance to other more sophisticated methods.

High Throughput Mutation Screening of the TP53 Gene in Lung Cancer Using Single Molecule Real Time (SMRT) Sequencing

Jin Jen, JinSung Jang, Karl Oles, Ana Robles*, Jaime Davila, Bruce Eckloff, Curt Harris*, and Eric Wieben.

The Medical Genome Facility, Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota, and * the Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, United States.

TP53 is one of the most commonly mutated genes in human cancer. In wild type form, TP53 functions as a tumor suppressor gene by negatively regulateing cell cycle and inducing apoptosis when needed to preserve genomic stability. In non-small cell lung cancer (NSCLC), mutation of TP53 occurs in nearly 90% of squamous cell carcinomas and roughly 50% of adenocarcinomas. Clinical studies suggest that NSCLCs with TP53 alterations have less favorable prognosis and may confer tumor resistant to chemotherapy and radiation. Therefore, it is highly desirable to develop robust and efficient ways to identify the genetic status for TP53 or other highly targetable genes, utilizing the Next Generation Sequencing (NGS) based high throughput technologies. Here we explore the use of the PacBio Single Molecule SMRT sequencer for rapid and high-throughput, targeted gene sequencing for the entire coding region of the TP53 gene. Using multiplex PCR, the entire 11 coding exons of the TP53 gene were amplified in a single tube and indexed for each sample. SMRTbell libraries were then generated after pooling up to 12 samples and bead purification of the PCR products. Single molecule sequencing was then carried out to generate up to 100,000 circular consensuses reads (CCR) per SMRT cell. Our results show that the CCRs generated high quality sequences that were accurate and detected mutations present at less than 10% of the alleles in the sample. Our experience using paired tumor and normal lung cancer samples with and without TP53 gene mutations demonstrate that PacBio single molecule sequencing can provide a highly robust, reliable and cost effective method for rapid identification of tumor associated mutations in the TP53 gene. This approach is also easily applicable for the rapid analysis of any other candidate genes of biological and clinical interest.

Statistical methods for comparative metagenomic analysis

Hongmei Jiang [1,*], Lingling An [2], Zhenyu Zhao [1], Issac Jenkins [2], Naruekamol Pookhao[2],

[1]: Department of Statistics, Northwestern University, Evanston IL, 60208, USA.

[2]: Department of Agricultural and Biosystems Engineering, The University of Arizona, Tucson AZ, 85004, USA.

[*]: To whom correspondence should be addressed.

Emails: HJ (Hongmei@northwestern.edu); LA (anling@email.arizona.edu)

Recent advent of high-throughput sequencing technologies has greatly promoted metagenomics which studies the entire microbial communities without the need of culturing the individual member organisms in the laboratory. In the past few years, the number of metagenomic projects in natural (such as soil and water) or host-associated (such as human) environments has grown exponentially. Comparative analysis of two or more metagenomic samples is necessary to understand the similarities and dissimilarities of the complex microbial communities between different environments or different hosts. Comparisons of metagenomes can be performed at the levels of average genome size, taxonomic composition, and functional roles. In this paper we give an overview of comparative metagenomics, a new branch of metagenomics, with particular emphasis on the current development in this field. We will compare several statistical methods and computational algorithms by using comprehensive simulation studies and real data analyses.

PacBio RS long Read Applications in Plant Genomics

Jifeng Tang[1*], Erwin Datema[1], Rui Peng Wang[1], Alexander Wittenberg[1], Rolf Mank[1], Rudie Antonise[1], Rik op den Camp[1], Peter van Dijk[1], Antoine Janssen[1*]

[1]: Keygene N.V., Agro Business Park 90, P.O. Box 216, 6700 AE Wageningen, The Netherlands

[*]: To whom correspondence should be addressed.

E-mail: jifeng.tang@keygene.com

Many plant species with high economic value have large and repetitive genomes, which hamper construction of high quality assemblies and its subsequent exploitation. The currently available next-generation sequencing technologies, such as Illumina, are high-throughput, low cost, but produce relatively short reads (<~250 bases). As a result, most sequenced plant genomes consist of tens of thousands of scaffolds containing many gaps. The PacBio RS sequencing technology can generate very long reads (≥4 kb on average), which can be used to generate high quality assemblies and improve the existing draft genomes. Here we present several examples of high quality assemblies and improved existing draft genomes in plants generated using PacBio RS reads.

We have sequenced a mitochondrial genome using the PacBio RS and assembled the sequences into two contigs, in contrast with 835 scaffolds with N50 size of 2Kb generated from Illumina reads only. Also, we have successfully sequenced and assembled individual BACs and BAC pools using the PacBio RS. Compared to several hundreds of contigs per BAC generated from 454 sequencing alone, the PacBio assemblies consist of only a single contig per BAC and have almost identical sequence content. We have further checked the quality of the assembled BACs using Whole Genome Profiling (WGP™), a sequence-based physical mapping technology. We confirmed that the order of the sequence tags of these BACs on the WGP map is collinear with the assembled BACs. In addition, we have developed an approach to select BACs from a WGP map for gap closure within scaffolds and between scaffolds. The selected BACs were sequenced using the PacBio RS, and most of the gaps that we expected were (partially) closed by the PacBio sub-reads.

Inferring Intra-Tumor Heterogeneity from Copy Number Aberrations

Layla Oesper [1,*], Ahmad Mahmoody [1], Benjamin J. Raphael [1,2]

[1]: Department of Computer Science, Brown University, Providence, RI 02912.

[2]: Center for Computation Molecular Biology, Brown University, Providence, RI 02912.

[*]: To whom correspondence should be addressed.

Emails: LO (layla@cs.brown.edu); AM (ahmad@cs.brown.edu); BJR (braphael@cs.brown.edu)

Motivation: Cancer is a heterogeneous disease with individual cells in a tumor typically having different complements of somatic mutations. Most cancer sequencing projects sequence a mixture of cells from a tumor sample including admixture by normal (non-cancerous) cells and different subpopulations of cancerous cells. Sequenced reads are typically aligned to a reference genome yielding information on read depth. Regions of differing read depth suggest regions with different copy number in the tumor sample. Factors such as normal cellular contamination, and multiple subpopulations are known to cause changes in the observed read depth. However, most solid tumors exhibit extensive aneuploidy and copy number aberrations. Highly rearranged genomes containing many copy number aberrations may also cause shifts in read depth when the length of the cancer genome is far from the length of the reference genome. This further complicates the analysis of somatic mutations in sequenced tumor samples.

Methods: We describe an algorithm - Tumor Heterogeneity Analysis (THetA) - to infer tumor purity (and more generally, the fraction of each subpopulation in the sample) and clonal/subclonal tumor subpopulations directly from read depth information for whole-genome DNA sequencing data. THetA uses large copy number aberrations to estimate tumor purity and distinct subpopulations, and does so while considering the length of all component genomes in the mixture.

Results: We demonstrate the advantages of THetA on both simulated and real data. In simulation we show how changes in the length of genomes in the mixture can alter read depth - leading to highly inaccurate tumor purity estimates for other methods that do not fully account for the altered genome length. We also demonstrate the advantage of using THetA on whole-genome sequencing data from breast carcinoma and glioblastoma multiforme samples. In each case we identify tumor purity and subpopulations of tumor cells, some of which have genomes whose length is significantly different from the reference.

Bringing next generation sequencing to the clinic: Analytical validation and initial deployment of a comprehensive cancer genomic profiling test

Kai Wang[1], Garrett M. Frampton[1], Alex Fichtenholtz[1], Sean Downing[1], Jie He[1], Frank Juhn[1], Tina Brennan[1], Geoff Otto[1], Alex Parker[1], Vincent A. Miller[1], Jeffrey S Ross[1,2], John Curran[1], Philip J. Stephens[1], Doron Lipson[1], Roman Yelensky[1,*]

[1]Foundation Medicine, Cambridge, Massachusetts,

[2]Albany Medical College, Albany, NY

Emails: K Wang(kwang@foundationmedicine.com);R Yelensky(ryelensky@foundationmedicine.com)

Background: As the number of clinically relevant cancer genes increases, next-generation sequencing (NGS) is becoming an attractive diagnostic tool, as it can detect most genomic alterations in a single assay on limited tissue. However, rigorous analytical validation and comparison to current tests is required before clinical use.

Methods: We have developed an NGS-based test to characterize all classes of genomic alteration across 4,604 exons of 287 cancer-related genes from routine FFPE clinical specimens, including needle biopsies. To validate the test, we created reference samples reflecting key drivers of detection accuracy for somatic alterations across the targeted regions: For base substitutions, we mixed 2 pools of 10 normal cell-lines, testing 1,650 variants at allele frequencies (AF) 5%-100%. For indels, 28 tumor cell lines with 47 alterations 1-40bp long were used to generate 41 pools, testing 161 events at AF≥10%. For copy number alterations (CNAs), 7 tumor cell-lines with 19 gene amplifications and 9 homozygous gene deletions were pooled with their matched normal in 5 ratios with tumor content 20-75%. We confirmed accuracy in 308 FFPE tumors characterized for 95 alterations in 12 genes (e.g., EGFR, BRAF, HER2) by other assays, including PCR, mass-spec, FISH, and IHC. Precision was established on two control FFPE specimens processed a total of 150 times. We then assessed the potential clinical impact of comprehensive NGS by examining the nature and prevalence of genomic alterations revealed by the validated test in >2,000 consecutive patient specimens.

Results: On reference samples, sensitivity reached >99% (1,649/1,650) for base substitutions (AF≥5%), 98% (157/161) for indels (AF≥10%), >99% (56/56) for gene amplifications at CN≥8 and 97% (35/36) for homozygous deletions (tumor purity ≥30%), all with high specificity (PPV>99%). Robust performance translated to FFPE: concordance averaged 97% across subs/indels (109/113) and CNAs (41/42) relative to prior calls. All known alterations were called in all replicates of precision control specimens, including a base substitution present at only 4%. 2112/2221 (95%) of clinical specimens were successfully profiled (mean coverage 1134X), with 76% containing at least one alteration directly linked to a clinically available targeted treatment option or a mechanism-driven clinical trial. 963 unique such "actionable" alterations were reported, less than 1/3rd of which would have been identified by standard clinical tests, highlighting the advantages of a comprehensive NGS-based approach.

Conclusions: We present rigorous validation of a comprehensive NGS-based diagnostic test optimized for use in oncologic care and advocate that such genomic profiling be used to drive the appropriate use of targeted therapy and expand treatment choices for cancer patients.

Phylogenetic analyses unravel the evolutionary history of NAC proteins in plants

Tingting Zhu [1, *], Eviatar Nevo [2], Dongfa Sun [3], Junhua Peng [4, 5], Xuan Li [1, *]

[1]: Institute of Plant Physiology and Ecology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China.

[2]: Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel.

[3]: College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, China.

[4]: Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei 430074, China.

[5]: Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO 80526-1170, USA.

[*]: To whom correspondence should be addressed.

Emails: TZ (yooheez@gmail.com); EN (nevo@research.haifa.ac.il);
DS (sundongfa@mail.hzau.edu.cn); JP (junhuapeng@yahoo.com); XL (lixuan@sippes.ac.cn)

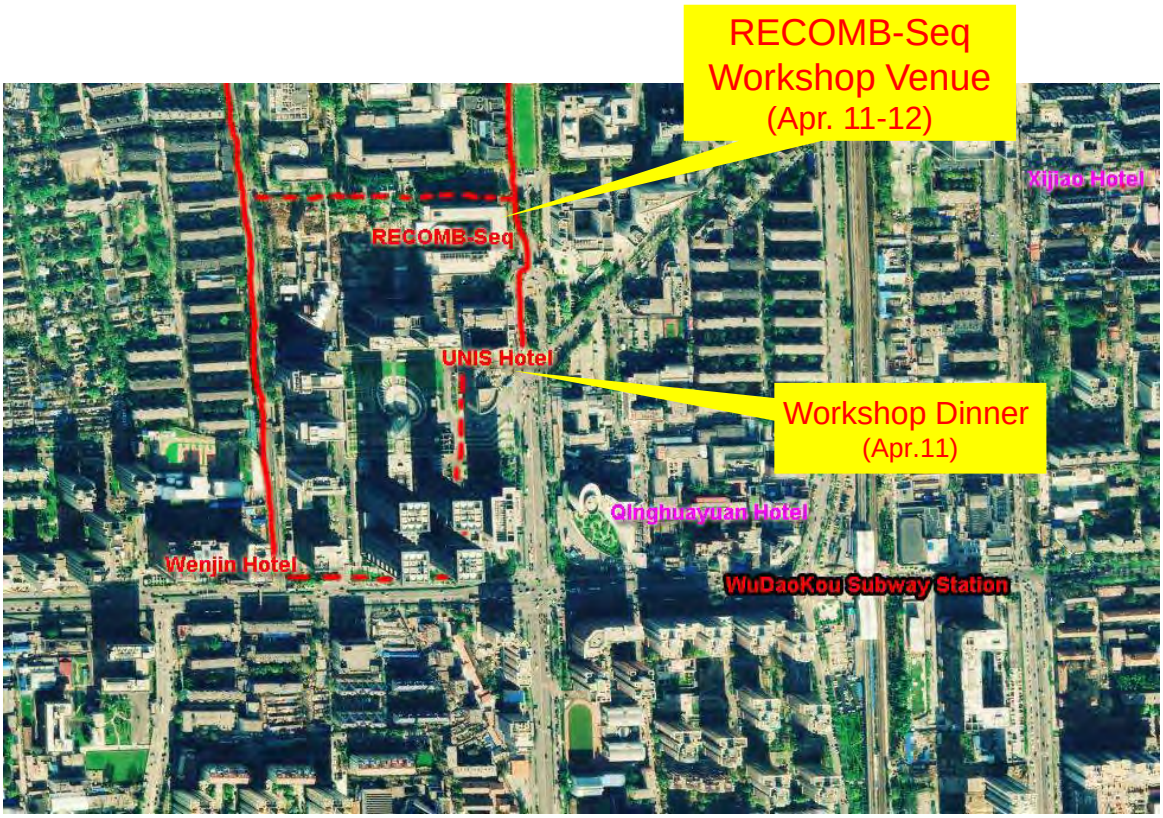
NAC (NAM/ATAF/CUC) proteins are one of the largest groups of transcription factors in plants. Although many NAC proteins based on Arabidopsis and rice genomes have been reported in a number of species, a complete survey and classification of all NAC genes in plant species from disparate evolutionary groups is lacking. In this study, we analyzed whole-genome sequences from nine major lineages of land plants to unveil the relationships between these proteins. Our results show that there are fewer than 30 NAC proteins present in both mosses and lycophytes, whereas more than 100 were found in most of the angiosperms. Phylogenetic analyses suggest that NAC proteins consist of 21 subfamilies, most of which have highly conserved non-NAC domain motifs. Six of these subfamilies existed in early-diverged land plants, whereas the remainder diverged only within the angiosperms. We hypothesize that NAC proteins probably originated sometime more than 400 million years ago and expanded together with the differentiation of plants into organisms of increasing complexity possibly after the divergence of lycophytes from the other vascular plants.

List of Paper Authors

List of Paper Authors

Akalin, Altuna	Weill Cornell Medical College	USA
Allhoff, Manuel	RWTH Aachen	Germany
Ander, Christina	Bielefeld University	Germany
Ashoor, Haitham	King Abdullah University of Science and Technology	Saudi Arabia
Bafna, Vineet	University of California, San Diego	USA
Bajic, Vladimir B.	King Abdullah University of Science and Technology	Saudi Arabia
Bareke, Eric	Universite de Montreal	Canada
Barillot, Emmanuel	INSERM U900, Institut Curie, Mines ParisTech	France
Benson, Gary	Boston University	USA
Biesinger, Jacob	University of California, Irvine	USA
Boeva, Valentina	INSERM U900, Institut Curie, Mines ParisTech	France
Bresler, Guy	University of California at Berkeley	USA
Bresler, Ma'Ayan	University of California at Berkeley	USA
Cairns, Murray J.	University of Newcastle	Australia
Chen, Kun-Tze	National Tsing Hua University	Taiwan
Chen, Liang	University of Southern California	USA
Costa, Ivan G.	RWTH Aachen	Germany
Cox, Anthony	Illumina UK	United Kingdom
Csuros, Miklos	Universite de Montreal	Canada
Florea, Liliana	Johns Hopkins University	USA
Garrett--bakelman, Francine	Weill Cornell Medical College	USA
Gelfand, Yevgeniy	Boston University	USA
Guo, Jiangtao	Peking University	China
Hayes, Matthew	Case Western Reserve University	USA
Healy, Jasmine	Universite de Montreal	Canada
Hérault, Aurélie	UMR 144 CNRS, Institut Curie	France
Hernandez, Yozen	Boston University	USA
Hu, Gangqing	ational Institutes of Health	USA
Kambadur, Prabhanjan	IBM TJ Watson Research Center	USA
Kapun, Evgeny	St. Petersburg National Research University of IT, Mechanics and Optics	Russia
Kim, Sangwoo	University of California, San Diego	USA
Kuo, C.-C. Jay	University of Southern California	USA
Kuosmanen, Anna	University of Helsinki	Finland
Lederman, Roy	Yale	USA
Lee, Seunghak	Carnegie Mellon University	USA
Levine, Ross	Memorial Sloan-Kettering Cancer Center	USA
Li, Chi-Long	National Tsing Hua University	Taiwan
Li, Jing	Case Western Reserve University	USA
Li, Sheng	Weill Cornell Medical College	USA
Li, Yi	University of California, Irvine	USA
Liu, Taigang	Shanghai Ocean University	China
Liu, Yongchu	Peking University	China
Lo, Christine	University of California, San Diego	USA
Loving, Joshua	Boston University	USA
Lozano, Aurelie	IBM TJ Watson Research Center	USA

Lu, Chin Lung	National Tsing Hua University	Taiwan
Mäkinen, Veli	University of Helsinki	Finland
Marschall, Tobias	CWI Amsterdam	Netherlands
Martin, Marcel	TU Dortmund	Germany
Mason, Christopher	Weill Cornell Medical College	USA
Melnick, Ari	Weill Cornell Medical College	USA
Qin, Yufang	Shanghai Ocean University	China
Radvanyi, François	UMR 144 CNRS, Institut Curie	France
Rahmann, Sven	University of Duisburg-Essen	Germany
Rizzi, Romeo	University of Verona	Italy
Schoenhuth, Alexander	Centrum Wiskunde & Informatica	Netherlands
Schulz-Trieglaff, Ole	Illumina UK	United Kingdom
Sinnett, Daniel	Universite de Montreal	Canada
Song, Li	Johns Hopkins University	USA
Sonmez, Kemal	Oregon Health & Science University	USA
Spinella, Jean-Francois	Ste-Justine Hospital	Canada
Stoye, Jens	Bielefeld University	Germany
Tomescu, Alexandru I.	University of Helsinki	Finland
Tsarev, Fedor	St. Petersburg National Research University of IT, Mechanics and Optics	Russia
Tse, David	University of California at Berkeley	USA
Vidal, Ramon	Ste-Justine Hospital	Canada
Wang, Jun	Shanghai Normal University	China
Wang, Xi	University of Newcastle, Australia	Australia
Wang, Yuanfeng	University of California, Irvine	USA
Whelan, Christopher	Oregon Health & Science University	USA
Xie, Xiaohui	University of California, Irvine	USA
Xing, Eric	Carnegie Mellon University	USA
Zakov, Shay	University of California, San Diego	USA
Zhang, Jing	University of Southern California	USA
Zhang, Yu	Penn State University	USA
Zheng, Xiaoqi	Shanghai Normal University	China
Zhu, Huaiqiu	Peking University	China
Zhu, Jie	Shanghai Normal University	China
Zumbo, Paul	Weill Cornell Medical College	USA



Sponsors:

清華大學

TNList

国家自然科学基金委员会 NSFC

illumina®

MONSANTO imagine®

EMC® ISILON

中科曙光 Sugon

intel