Fifth RECOMB Satellite Workshop on Massively Parallel Sequencing RECOMB-SEQ 2015

and

Fourth RECOMB Satellite Workshop on Computational Cancer Biology RECOMB-CCB 2015

Satellite to Conference on Research in Computational Molecular Biology

April 10-11 2015, Warsaw, Poland

Old Library of Warsaw University at the main campus

PROGRAM COMMITTEE SEQ & CCB

- **Niko Beerenwinkel** (chair SEQ), ETH Zurich, Switzerland
- **Jan Korbel** (chair SEQ), European Molecular Biology Laboratory, Germany
- **Ken Chen** (chair CCB), M.D. Anderson Cancer Center, United States
- **Ben Raphael** (chair CCB), Brown University, United States
- **Lodewyk Wessels** (chair CCB), The Netherlands Cancer Institute, Netherlands
- Alexej Abyzov, Mayo Clinic, United States
- Tatsuya Akutsu, Kyoto University, Japan
- Max Alekseyev, George Washington University, United States
- Patrick Aloy, Institute for Research in Biomedicine, Spain
- Vikas Bansal, University of California San Diego, United States
- Gurkan Bebek, Case Western Reserve University, United States
- Inanc Birol, British Columbia Genome Sciences Centre, Canada
- C.Titus Brown, Michigan State University, United States
- Dongbo Bu, Insitute of Computing Technology, Chinese Academy of Sciences, China
- Mark Chaisson, University of Washington, United States
- Kun-Mao Chao, National Taiwan University, Taiwan
- Gang Fang, Mount Sinai School of Medicine, United States
- Heng Huang, Univ. of Texas at Arlington, United States
- Rui Jiang, Tsinghua University, China
- Ekta Khurana, Weill Cornell Medical College, United States
- Gunnar W. Klau, Centrum Wiskunde & Informatica, Netherlands
- Dominique Lavenier, IRISA / INRIA, France
- Hyunju Lee, Gwangju Institute of Science and Technology, South Korea
- Ming Li, University of Waterloo, Canada
- Jinze Liu, University of Kentucky, United States
- Stefano Lonardi, University of California, Riverside, United States
- Bojan Losic, Mount Sinai Hospital, United States
- Jian Ma, University of Illinois at Urbana-Champaign, United States
- Florian Markowetz, University of Cambridge, United Kingdom
- Tobias Marschall, Max Planck Institute for Informatics, Germany
- Manja Marz, Uni Jena, Germany
- Ryan Mills, University of Michigan Medical School, United States
- Sach Mukherjee, Netherlands Cancer Institute, Netherlands
- Niranjan Nagarajan, University of Maryland, United States
- Laurent Noé, Université Lille 1, France
- Arlindo Oliveira, Instituto Superior Técnico, Portugal
- Pierre Peterlongo, Centre INRIA Rennes-Bretagne-Atlantique, France
- Julio Saez-Rodriguez, European Bioinformatics Institute, United Kingdom
- S. Cenk Sahinalp, Simon Fraser University, Canada
- Michael Schatz, Cold Spring Harbor Laboratory, United States
- Alexander Schoenhuth, Centrum Wiskunde & Informatica, Netherlands
- Sohrab Shah, University of British Columbia, Canada
- Steven Skiena, Stony Brook University, United States
- Haixu Tang, Indiana University, United States
- Helene Touzet, CNRS - Centre national de la recherche scientifique, France

- Todd Treangen, National Biodefense Analysis and Countermeasures Center, United States
- Jerome Waldispuhl, McGill University, Canada
- Lusheng Wang, City University of Hong Kong, Hong Kong
- Wenyi Wang, The University of Texas MD Anderson Cancer Center, United States
- Dong Xu, University of Missouri, United States
- Shaojie Zhang, University of Central Florida, United States
- Louxin Zhang, National University of Singapore, Singapore
- Zhongming Zhao, Vanderbilt University, United States

STEERING COMMITTEE (SEQ)

- Inanc Birol, Canada's Michael Smith Genome Sciences Centre, Canada
- Michael Brudno, University of Toronto, Canada
- Eran Halperin, Tel Aviv University, Israel
- Ben Raphael, Brown University, United States
- Cenk Sahinalp, Simon Fraser University, Canada

STEERING COMMITTEE (CCB)

- Joe Gray, Oregon Health Sciences University, United States
- Michael Hallett, McGill University, Canada
- Ben Raphael, Brown University, United States
- Sohrab Shah, BC Cancer Agency, Canada
- Zohar Yakhini, Technion and Agilent Technologies, United States

ORGANIZING COMMITTEE

- **Norbert Dojer** (chair), University of Warsaw, Poland
- **Anna Gambin** (chair), University of Warsaw, Poland
- Agata Charzyńska, Polish Academy of Sciences, Poland
- Mateusz Łącki, University of Warsaw, Poland

SPONSORS

RECOMB-SEQ and CCB 2015 are receiving support from:

Warsaw Center of Mathematics and Computer Science

UNIVERSITY OF WARSAW

Additional Information

Internet access WIFI: RECOMB-SEQ
password: Warsaw2015

Day 1: Friday, April 10

14:00 - 14:10	**Opening Remarks (Aula)**	
14:10 - 15:10	KEYNOTE: Marco Gerlinger **Structure and dynamics of cancer heterogeneity landscapes: analytical approaches and challenges**	
15:10 - 15:50	**Contributed talks (separate sessions)**	
	Session SEQ (Aula)	**Session CCB (107)**
15 min	HIGHLIGHT: Faraz Hach, Ibrahim Numanagic and S. Cenk Sahinalp **DeeZ: reference-based compression by local assembly**	SHORT TALK: Ekta Khurana **Computational identification of noncoding cancer drivers from whole-genome sequencing data**
20 min	PAPER TALK: Yuzhen Ye and Haixu Tang **Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis**	HIGHLIGHT: Hatice U. Osmanbeyoglu, Raphael Pelossof, Jacqueline F. Bromberg, Christina S. Leslie **Linking Signaling Pathways to Transcriptional Programs in Breast Cancer**
15:50 - 16:20	**Coffee break**	
16:20 - 17:30	**Contributed talks (joint session) (Aula)**	
15 min	SHORT TALK: Norbert Dojer, Abhishek Mitra, Yea-Lih Lin, Anna Kubicka, Magdalena Skrzypczak, Krzysztof Ginalski, Philippe Pasero and Maga Rowicka **Computational detection of DNA double-stranded breaks and inferring mechanisms of their formation**	
20 min	PAPER TALK: Daria Iakovishina, Isabelle Janoueix-Lerosey, Emmanuel Barillot, Mireille Regnier and Valentina Boeva **SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability**	
15 min	SHORT TALK: David Seifert, Francesca Di Giallonardo, Karin J. Metzner, Huldrych F. Günthard and Niko Beerenwinkel **A Framework for Inferring Fitness Landscapes of Patient-Derived Viruses Using Quasispecies Theory**	
20 min	PAPER TALK: Jie Ren, Kai Song, Minghua Deng, Gesine Reinert, Chuck Cannon and Fengzhu Sun **Inference of Markovian Properties of Molecular Sequences from NGS Data and Applications to Comparative Genomics**	
17:30 - 18:30	KEYNOTE: Jacek Błażewicz **DNA Sequencing - from SBH to NGS and Hybrid Algorithms**	
20:00	**Conference Dinner at Villa Foksal Restaurant** 3/5 Foksal Str.	

Day 2: Saturday, April 11

09:00 - 10:00	KEYNOTE: Francesca Ciccarelli **Proliferation history of tumors and impact on therapy**	
10:00 - 10:30	**Coffee break**	
10:30 - 12:00	**Contributed talks (separate sessions)**	
	Session SEQ (Aula)	**Session CCB (107)**
20 min	PAPER TALK: Davide Verzotto, Audrey S.M. Teo, Axel Hillmer and Niranjan Nagarajan **Index-based map-to-sequence alignment in large eukaryotic genomes**	HIGHLIGHT: Salim Akhter Chowdhury, Stanley E. Shackney, Kerstin Heselmeyer-Haddad, Alejandro A. Schäffer, Russell Schwartz **Algorithms to Model Single Gene, Single Chromosome, and Whole Genome Copy Number Changes Jointly in Tumor Phylogenetics**
15 min	SHORT TALK: Damian Wójtowicz, Fedor Kouzine, Arito Yamane, Craig J. Benham, Rafael C. Casellas, David Levens and Teresa M. Przytycka **Genome-wide Mapping and computational analysis of non-B DNA structures in vivo**	SHORT TALK: Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field and Ben Raphael **Reconstruction of clonal trees and tumor composition from multi-sample cancer sequencing data**
15 min	HIGHLIGHT: Benjamin Buchfink, Chao Xie and Daniel Huson **Fast and Sensitive Protein Alignment using DIAMOND**	SHORT TALK: Amit G. Deshwar, Levi Boyles, Jeff Wintersinger, Paul C. Boutros, Yee Whye Teh, Quaid Morris **Resolving ambiguities in tumour phylogenies**
20 min	PAPER TALK: Birte Kehr, Pall Melsted and Bjarni Halldorsson **PopIns: population-scale detection of novel sequence insertions**	PAPER TALK: Simona Constantinescu, Ewa Szczurek, Pejman Mohammadi, Jörg Rahnenführer and Niko Beerenwinkel **A Waiting Time Model for Mutually Exclusive Cancer Alterations**
15 min	HIGHLIGHT: Raluca Uricaru, Guillaume Rizk, Vincent Lacroix, Elsa Quillery, Olivier Plantard, Rayan Chikhi, Claire Lemaitre and Pierre Peterlongo **DiscoSnp: Reference-free detection of isolated SNPs**	SHORT TALK: Yuichi Shiraishi, Georg Tremmel, Satoru Miyano and Matthew Stephens **Extraction of Latent Probabilistic Mutational Signature in Cancer Genomes**
12:00 - 13:30	**Lunch**	

13:30 - 14:30	**Poster viewing**	
14:30 - 16:00	**Contributed talks (separate sessions)**	
	Session SEQ (Aula)	**Session CCB (107)**
15 min	HIGHLIGHT: Nathanael Fillmore, Bo Li, Yongsheng Bai, Mike Collins, James A. Thompson, Ron Stewart and Colin N. Dewey **Evaluation of de novo transcriptome assemblies from RNA-Seq data**	SHORT TALK: Isaac Joseph, Shannon McCurdy, Lior Pachter and Joseph F. Costello **A method for combining multiple genomic and clinical datatypes to predict recurrence grade in gliomas**
20 min	PAPER TALK: Monica Golumbeanu, Pejman Mohammadi and Niko Beerenwinkel **Probabilistic modeling of occurring substitutions in PAR-CLIP data**	SHORT TALK: Mathieu Lajoie, Sylvie Langlois, Pascal St-Onge, Patrick Beaulieu, Jasmine Healy and Daniel Sinnett **Personalized Targeted Therapy for Refractory Childhood Cancers**
20 min	PAPER TALK: Nicolas Bray and Lior Pachter **Rank Regularized RNA-seq**	SHORT TALK: Teresa Przytycka**Integrated analysis of mutual exclusivity and gene interaction in Pan-Cancer dys-regulated pathways**
15 min	SHORT TALK: Aaron Lun and Gordon Smyth **Using csaw to detect differentially bound regions in ChIP-seq data**	SHORT TALK: Ashar Ahmad and Holger Froehlich **Dirichlet Process Mixture Model with Bayesian Lasso for consistent clustering of Survival Times with Molecular Data**
15 min	SHORT TALK: Alessandro Mammana and Ho-Ryun Chung **Chromatin segmentation with a joint model for reads explains a larger portion of the epigenome**	SHORT TALK: Ewa Szczurek, Tyll Krüger, Barbara Klink and Niko Beerenwinkel **The bottleneck of metastasis formation: insights from a stochastic model**
16:00 - 16:30	**Coffee**	
16:30 - 17:30	KEYNOTE: Gerton Lunter **Analytical challenges in next-generation sequencing**	
17:30	**Closing remarks (Aula)**	

ABSTRACT INDEX

Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis

Yuzhen Ye[1*] and Haixu Tang[1]

[1] Indiana University, United States *yye@indiana.edu

Metagenomics research has accelerated the studies of microbial organisms, providing insights into the composition and potential functionality of various microbial communities. Metatranscriptomics (studies of the transcripts from a mixture of microbial species) and other meta-omics approaches hold even greater promise for providing additional insights into functional and regulatory characteristics of the microbial communities. Current metatranscriptomics projects are often carried out without matched metagenomic datasets (of the same microbial communities). For the projects that produce both metatranscriptomic and metagenomic datasets, their analyses are often not integrated. Metagenome assemblies are far from perfect, partially explaining why metagenome assemblies are not used for the analysis of metatranscriptomic datasets. Here we report a reads mapping algorithm for mapping of short reads onto a de Bruijn graph of assemblies. A hash table of junction k-mers (k-mers spanning branching structures in the de Bruijn graph) is used to facilitate fast mapping of reads to the graph. We developed an application of this mapping algorithm: a reference based approach to metatranscriptome assembly using graphs of metagenome assembly as the reference. Our results show that this new approach helps to assemble substantially more transcripts that otherwise would have been missed or truncated because of the fragmented nature of the reference metagenome.

Computational detection of DNA double-stranded breaks and inferring mechanisms of their formation

Norbert Dojer[1*], Abhishek Mitra[1], Yea-Lih Lin[2], Anna Kubicka[3], Magdalena Skrzypczak[3], Krzysztof Ginalski[3], Philippe Pasero[2] and Magda Rowicka[1†]

[1] University of Texas Medical Branch, United States *dojer@mimuw.edu.pl
[2] Institut de Genetique Humaine, CNRS, France
[3] University of Warsaw, Poland

Double-stranded DNA breaks (DSBs) are a dangerous form of DNA damage. The damage to both DNA strands precludes the straightforward use of the complementary strand as a template for repair, resulting in mutagenic lesions. Despite many studies on the mechanisms of DSB formation, our knowledge of them is very incomplete. A main reason is that, to date, DSB formation has been extensively studied only at specific loci but remains largely unexplored at the genome-wide level.

We recently developed a method to label DSBs in situ followed by deep sequencing (BLESS), and used it to map DSBs in human cells [1] with a resolution 2-3 orders of magnitude better than previously achieved. Although our protocol detects free ends of DNA with extreme single nucleotide precision, the inference of original positions of DSBs remains challenging. This problem is due to unavoidable sequencing of DNA repair intermediaries (end resection), which effectively lowers our detection resolution by several orders of magnitudes. Another challenge is that DSBs are rare events and sequencing signal originating from them is easily overpowered by background signal, such as copy number variation.

Here, we show how DSBs can be detected computationally with nucleotide resolution. First, we analyzed DSBs induced in pre-determined positions (i.e. restriction enzyme cutting sites). We compared sequencing data from two experiments: BLESS and H2AX phosphorylation (the latter is known to appear in the DSB neighborhood of a few kbp), showing that BLESS provides much stronger and accurate signal. Studying the vicinity of most evident DSBs, we developed a characteristic BLESS sequencing read pattern. Scanning the genome with this pattern, we can detect DSB locations with 2nt positional accuracy and precision of over 90%. Next, we applied our pattern to BLESS data from typical experiments, where DSB positions are not known a priori. We used other data known to correlate with DSB locations to confirm high quality of our predictions. Finally, we analyzed read patterns in larger areas surrounding DSBs and their relation to the corresponding DNA damage and repair mechanisms.

[1] Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q, et al. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat Methods. 2013;10(4):361-5.

A Framework for Inferring Fitness Landscapes of Patient-Derived Viruses Using Quasispecies Theory

David Seifert[1*], Francesca Di Giallonardo[2], Karin J. Metzner[2], Huldrych F. Günthard[2] and Niko Beerenwinkel[1†]

[1] ETH Zurich, Switzerland *david.seifert@bsse.ethz.ch
[†]niko.beerenwinkel@bsse.ethz.ch
[2] University Hospital Zurich, Switzerland

Fitness is a central quantity in evolutionary models of viruses. However, it remains difficult to determine viral fitness experimentally, and existing in vitro assays can be poor predictors of in vivo fitness of viral populations within their hosts. Next-generation sequencing can nowadays provide snapshots of evolving virus populations, and these data offer new opportunities for inferring viral fitness. Using the equilibrium distribution of the quasispecies model, an established model of intrahost viral evolution, we linked fitness parameters to the composition of the virus population, which can be estimated by next-generation sequencing. For inference, we developed a Bayesian Markov chain Monte Carlo method to sample from the posterior distribution of fitness values. The sampler can overcome situations where no maximum-likelihood estimator exists, and it can adaptively learn the posterior distribution of highly correlated fitness landscapes without prior knowledge of their shape. We tested our approach on simulated data and applied it to clinical human immunodeficiency virus 1 samples to estimate their fitness landscapes in vivo. The posterior fitness distributions allowed for differentiating viral haplotypes from each other, for determining neutral haplotype networks, in which no haplotype is more or less credibly fit than any other, and for detecting epistasis in fitness landscapes.

Our implemented approach, called QuasiFit, is available at
http://www.cbg.ethz.ch/software/quasifit

Inference of Markovian Properties of Molecular Sequences from NGS Data and Applications to Comparative Genomics

Jie Ren[1*], Kai Song[2], Minghua Deng[2], Gesine Reinert[3‡], Chuck Cannon[4] and Fengzhu Sun[1†]

[1] University of Southern California, United States [*]renj@usc.edu [†]fsun@usc.edu
[2] Peking University, China
[3] University of Oxford, United Kingdom [‡]reinert@stats.ox.ac.uk
[4] Texas Tech University, United States

Next Generation Sequencing (NGS) technologies generate large amounts of short read data for many different organisms. The fact that NGS reads are generally short makes it challenging to assemble the reads and reconstruct the original genome sequence. For clustering genomes using such NGS data, word-count based alignment-free sequence comparison is a promising approach, but for this approach, the underlying expected word counts are essential.

A plausible model for this underlying distribution of word counts is given through modelling the DNA sequence as a Markov chain. For single long sequences, efficient statistics are available to estimate the order of MCs and the transition probability matrix for the sequences. As NGS data do not provide a single long sequence, methods of inference of MC properties of sequences based on single long sequences cannot be directly used for NGS short read data.

Here we derive a normal approximation for such word counts. We also show that the traditional Chi-square statistic has an approximate gamma distribution, using the Lander-Waterman model for physical mapping. We propose several methods to estimate the order of the MC based on NGS reads and evaluate them using simulations.

We illustrate the applications of our results by clustering genomic sequences of several vertebrate and tree species based on NGS reads using alignment-free sequence comparison statistics. We find that the estimated order of the MC has a considerable effect on the clustering results, and that the clustering results that use a MC of the estimated order give a plausible clustering of the species.

Index-based map-to-sequence alignment in large eukaryotic genomes

Davide Verzotto[1*], Audrey S.M. Teo[2], Axel Hillmer[2] and Niranjan Nagarajan[1†]

[1] Computational and Systems Biology, Genome Institute of Singapore, Singapore
*verzottod@gis.a-star.edu.sg
†nagarajann@gis.a-star.edu.sg
[2] Cancer Therapeutics and Stratified Oncology, Genome Institute of Singapore

Resolution of complex repeat structures and rearrangements in the assembly and analysis of large eukaryotic genomes is often aided by a combination of high-throughput sequencing and mapping technologies (e.g. optical restriction mapping). In particular, mapping technologies can generate sparse maps of large DNA fragments (150 kbp–2 Mbp) and thus provide a unique source of information for disambiguating complex rearrangements in cancer genomes. Despite their utility, combining high-throughput sequencing and mapping technologies has been challenging due to the lack of efficient and freely available software for robustly aligning maps to sequences.

Here we introduce two new map-to-sequence alignment algorithms that efficiently and accurately align high-throughput mapping datasets to large, eukaryotic genomes while accounting for high error rates. In order to do so, these methods (OPTIMA for glocal and OPTIMA-Overlap for overlap alignment) exploit the ability to create efficient data structures that index continuous-valued mapping data while accounting for errors. We also introduce an approach for evaluating the significance of alignments that avoids expensive permutation-based tests while being agnostic to technology dependent error rates.

Our benchmarking results suggest that OPTIMA and OPTIMA-Overlap outperform state-of-the-art approaches in sensitivity (1.6–2X improvement) while simultaneously being more efficient (170–200%) and precise in their alignments (99% precision). These advantages are independent of the quality of the data, suggesting that our indexing approach and statistical evaluation are robust and provide improved sensitivity while guaranteeing high precision.

Genome-wide Mapping and computational analysis of non-B DNA structures in vivo

Damian Wójtowicz[1*], Fedor Kouzine[1], Arito Yamane[2], Craig J. Benham[3],

Rafael C. Casellas[1], David Levens[1] and Teresa M. Przytycka[1†]

[1] National Institutes of Health, United States [*]damian.wojtowicz@nih.gov
[†]przytyck@ncbi.nlm.nih.gov
[2] Gunma University, Japan
[3] University of California, Davis, United States

The canonical double helical structure of B-DNA may undergo various deformations to adopt alternative conformations, non-B DNA structures, including single-stranded DNA, Z-DNA, G-quadruplex, H-DNA, cruciform. Previous studies confirmed the existence of some non-B DNAs in a few gene promoters (e.g. c-myc, ADAM-12) and implicated their role in gene regulation, but it is not known how abundant the alternative DNA conformations might be at the genomic level. Computer-based studies uncovered a large number of sequences across the mammalian genomes that can potentially form non-B DNA structure and play functional roles in regulating DNA transactions.

We developed a new experimental technique, which combines chemical and enzymatic techniques with high-throughput sequencing, to map non-B DNA conformations at the genomic scale in vivo. The new protocol was applied to identify in vivo formation of non-B DNA structures in the genomic DNA of mouse and human cells. We performed a genome-wide analysis of occurrences of these alternative DNA conformations and compared the experimental data to genomic regions computationally predicted to have a propensity to form non-B DNA conformations. We showed a significant enrichment of ssDNA signal near computationally predicted non-B DNA motifs. Moreover, each type of predicted non-B DNA structures has a distinctive experimentally derived signature. This study provides the first look at genome-wide landscape of the in vivo formation of alternative structures. We find non-B DNA structures to be a common feature of mammalian genomes and to be associated with gene function and cell state. This study promises to be a useful resource for further studies to explore the role of non-B DNA structures in the regulation of DNA transactions.

PopIns: population-scale detection of novel sequence insertions

Birte Kehr[1*], Pall Melsted[2†] and Bjarni Halldorsson[3‡]

[1] deCODE genetics, Iceland [*]birte.kehr@decode.is
[2] University of Iceland, Iceland [†]pmelsted@gmail.com
[3] Reykjavik University, Iceland [‡]bjarnivh@ru.is

The detection of genomic structural variation (SV) has advanced tremendously in recent years due to progress in high-throughput sequencing technologies. Novel sequence insertions, insertions without similarity to a human reference genome, have received less attention than other types of SVs due to the computational challenges in their detection from short read sequencing data, which inherently involves de novo assembly. De novo assembly is not only computationally challenging, but also requires high-quality data. While the reads from a single individual may not always meet this requirement, using reads from multiple individuals can increase power to detect novel insertions.

We have developed the program PopIns, which can discover and characterize non-reference insertions of 100 bp or longer on a population scale. In this paper, we describe the approach we implemented in PopIns. It takes as input a reads-to-reference alignment, assembles unaligned reads using a standard assembly tool, merges the contigs of different individuals into high-confidence sequences, anchors the merged sequences into the reference genome, and finally genotypes all individuals for the discovered insertions. Our tests on simulated data indicate that the merging step greatly improves the quality and reliability of predicted insertions and that PopIns shows significantly better recall and precision than the recent tool MindTheGap. Preliminary results on a data set of 305 Icelanders demonstrate the practicality of the new approach.

The source code of PopIns is available from http://github.com/bkehr/popins

Probabilistic modeling of occurring substitutions in PAR-CLIP data

Monica Golumbeanu[1*], Pejman Mohammadi[1] and Niko Beerenwinkel[1†]

[1] ETH Zürich, Switzerland [*]monica.golumbeanu@bsse.ethz.ch
[†]niko.beerenwinkel@bsse.ethz.ch

Photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) is an experimental method based on next-generation sequencing for identifying the RNA interaction sites of a given protein. The method deliberately inserts T-to-C substitutions at the RNA-protein interaction sites, which provides a second layer of evidence compared to other CLIP methods. However, the experiment includes several sources of noise which cause both low-frequency errors and spurious high-frequency alterations. Therefore, rigorous statistical analysis is required in order to separate true T-to-C base changes, following cross-linking, from noise. So far, most of the existing PAR-CLIP data analysis methods focus on discarding the low-frequency errors and rely on high-frequency substitutions to report binding sites, not taking into account the possibility of high-frequency false positive substitutions. Here, we introduce BMix, a new probabilistic method which explicitly accounts for the sources of noise in PAR-CLIP data and distinguishes cross-link induced T-to-C substitutions from low and high-frequency erroneous alterations. We demonstrate the superior speed and accuracy of our method compared to existing approaches on both simulated and real, publicly available datasets.

Using csaw to detect differentially bound regions in ChIP-seq data

Aaron Lun[1*] and Gordon Smyth[1]

[1] Walter and Eliza Hall Institute, Australia [*]alun@wehi.edu.au

ChIP-seq is a widely used technique for identifying the genome-wide binding sites for a protein of interest. Traditionally, ChIP-seq experiments are analysed by calling regions of absolute enrichment within each sample. However, a more direct approach can be used when differential binding (DB) is of interest. Changes in the binding profile can be identified between two or more biological conditions. Significant differences can then shed some light on the molecular mechanisms driving the differences in biology.

A DB analysis requires some definition of the genomic features of interest, in order to summarize read coverage into counts. However, a comprehensive set of known binding sites may not be available for all proteins. One alternative is to perform a de novo analysis with sliding windows. Reads are counted into sliding windows across the genome, and significant differences are identified from those counts. This avoids the need for pre-defined regions.

Here, we describe the csaw package (ChIP-seq analysis with windows) from the Bioconductor project. csaw provides a framework for the de novo detection of DB regions with sliding windows. Methods are provided for read counting, normalization of library-specific biases and statistical analysis of window-level counts. Detection of significant differences is based on the edgeR package, though some additional care is required to control the FDR. Application to real data results in the successful detection of complex DB events.

Chromatin segmentation with a joint model for reads explains a larger portion of the epigenome

Alessandro Mammana[1*] and Ho-Ryun Chung[1†]

[1] Max Planck Institute for Molecular Genetics [*]mammana@molgen.mpg.de
[†]chung@molgen.mpg.de

A central question in biology pertains to the establishment and maintenance of different phenotypes adopted by cells of a multicellular organism with a constant genotype. In part this diversification of cells is driven by the adoption of distinct epigenomes, e.g. the localization of histone modifications along the genome. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a common experimental approach to generate genome wide maps of histone modifications. There are a number of existing approaches to combine several histone modification tracks to segment and characterize cell-type-specific epigenomic landscapes. Here, we propose a novel segmentation algorithm. Owing to an accurate probabilistic model for the read counts, our method provides a useful annotation for a considerably larger portion of the genome, shows a stronger association with validation data, and yields more consistent predictions across replicate experiments compared to existing methods.

Rank Regularized RNA-seq

Nicolas Bray[1] and Lior Pachter[1*]

[1] University of California at Berkeley, United States [*]lpachter@math.berkeley.edu

Recent technological advances have enabled large RNA-Seq experiments to be performed on large numbers of related samples, with state of the art experiments involving up to tens of thousands of samples. A key tradeoff emerges in multi-sample experiments between the number of samples that are sequenced, and the depth at which each sample can be sequenced. Pooling of samples can improve the accuracy of overall abundance estimates, yet such averaging prevents analysis of differences between samples.

We describe a regularized approach to the analysis of multiple RNA-Seq experiments that allows for sharing of information between samples for the purpose of fragment assignment and transcript abundance estimation, but only to the extent that transcript abundances are homogeneous between samples. This is achieved via a rank constraint on the abundance-sample matrix. With this assumption, we show how the Lee-Seung algorithm for non-negative matrix factorization can be coupled to the standard EM algorithm for fragment assignment resulting in an EM algorithm for computational pooling.

We have implemented our method in software called R3 and demonstrate using the Geuvadis dataset that it improves on naive analysis and is practical for the scale at which multiplexed RNA-Seq experiments are currently being performed. R3 should be useful for a wide range of experiments, and can be used in conjunction with any RNA-Seq quantification tool.

DeeZ: reference-based compression by local assembly

Faraz Hach[1*], Ibrahim Numanagic[1†] and S. Cenk Sahinalp[1‡]

[1] Simon Fraser University, Canada [*]fhach@cs.sfu.ca
[†]inumanag@sfu.ca
[‡]cenk@cs.sfu.ca

Recent growth in high throughput sequencing data has necessitated novel data compression methods, especially for mapped read data. The standard approach to represent mapped reads is the SAM/BAM formats: BAM is the Lempel-Ziv compressed version of the SAM format with some additional information. Although there are alternative methods for compressing SAM files, none of them satisfy the following features of the BAM format: (i) lossless compression, i.e. being able to retrieve all the information provided in the SAM file, and (ii) random access ability, i.e. the ability to extract reads from a specific genomic window without having to decompress the entire file. As a result BAM format has been used as a standard for storing and communicating mapped sequence data.

Here, we introduce DeeZ with the aim of improving SAM/BAM format compression. The algorithmic novelty of DeeZ is that it locally assembles the reads mapped to each genomic window (of 10s of Kb in length) and encodes each read's relative locus with respect to the resulting contig. Sequence differences (SNVs, indels) between the contig and the reference are encoded separately. Through this strategy, DeeZ can almost double the compression ratio achieved by the BAM format on Illumina HiSeq data with a compression speed similar to that of the SAMtools compression package. As per BAM, DeeZ is lossless and provides random access capability. Additional features of DeeZ include support for fast flag statistics of a SAM file, and location based read sorting ability.

Among all current tools for compressing SAM files, the best compression is offered by sam_comp, which (i) omits some fields in the original SAM file and thus is lossy, and (ii) does not provide random access capability. DeeZ achieves similar or better compression ratios than sam_comp while providing random access capability. For users in need of high compression performance but not random access capability (for quality scores), DeeZ provides the option of using the quality AC model from sam_comp with partial random access capabilities.

DeeZ is available for download at http://deez-compression.sourceforge.net

Fast and Sensitive Protein Alignment using DIAMOND

Benjamin Buchfink[1*], Chao Xie[2] and Daniel Huson[1†]

[1] University of Tuebingen, Germany [*]benjamin.buchfink@uni-tuebingen.de
[†]daniel.huson@uni-tuebingen.de
[2] National University of Singapore, Singapore

The alignment of sequencing reads against a protein reference database is a major computational bottleneck in metagenomics and data-intensive evolutionary projects. Although recent tools offer improved performance over the gold standard BLASTX, they exhibit only a modest speedup or low sensitivity. We introduce DIAMOND, an open-source algorithm based on double indexing that is 20,000 times faster than BLASTX on short reads and has a similar degree of sensitivity.

For example, alignment of a whole Illumina HiSeq run of reads against NCBI-nr takes about one day on a single server using DIAMOND, while using BLASTX, this would take 30-50 years.

DiscoSnp: Reference-free detection of isolated SNPs

Raluca Uricaru[1], Guillaume Rizk[2], Vincent Lacroix[3], Elsa Quillery[4], Olivier Plantard[4], Rayan Chikhi[2], Claire Lemaitre[2] and Pierre Peterlongo[2*]

[1] Labri, France
[2] INRIA, France *pierre.peterlongo@irisa.fr
[3] Université Lyon 1, France
[4] Oniris, France

Assessing the genetic differences between individuals within a species, or between chromo- somes of an individual, is a fundamental task in many aspects of biology. Of specific interest, single nucleotide polymorphisms (SNPs) are variations of a single base, either between homolo- gous chromosomes within a single individual, or between individuals.

DiscoSnp is a method predicting, without the need of any reference genome, isolated SNPs from any number of raw read set(s). These features are of high interest while searching for markers in newly sequenced genomes from non-model species. The software runs faster than any assembly-based or mapping-based approaches, and has a tiny memory footprint enabling its usage on simple desktop machines. Moreover, results on simulated and real sequencing data show a high precision and recall rate for bacterial genomes as well as for complex eukaryotic ones, such as human and mouse. Results also show that using DiscoSnp is a better choice than assembly+mapping approaches both in terms of computational needs and result quality. As an example of application, DiscoSnp was used in a population genomics study on an arthropod species (Ixodes ricinus), in which 96% of the SNPs predicted by DiscoSnp that were tested, were experimentally validated.

Evaluation of de novo transcriptome assemblies from RNA-Seq data

Nathanael Fillmore[1], Bo Li[2], Yongsheng Bai[3], Mike Collins[3], James A. Thompson[3], Ron Stewart[3] and Colin N. Dewey[1]

[1] University of Wisconsin, Madison, United States *easychair@nate-fillmore.com
[2] University of California, Berkeley, United States
[3] Morgridge Institute for Research, Madison, WI, United States

De novo RNA-Seq assembly facilitates the study of transcriptomes for species without sequenced genomes. With the recent development of several de novo transcriptome assemblers, each of which has its own set of user-tunable parameters, there are many options for constructing an assembly. However, selecting the most accurate assembler and parameter settings for a given RNA-Seq data set has remained challenging, especially when the ground truth is unknown. To address this challenge, we have developed DETONATE, a collection of methods for evaluating de novo transcriptome assemblies with or without a ground truth transcript set. DETONATE consists of two components: RSEM-EVAL, a model-based score for evaluating assemblies when the ground truth is unknown, and REF-EVAL, a refined set of ground-truth-based scores. Our experiments show that RSEM-EVAL correctly reflects assembly accuracy, as measured by REF-EVAL. The RSEM-EVAL score has a broad range of applications, including selecting the best assembler for a particular data set, optimizing parameter settings for an assembler, and guiding new assembler design. With the guidance of RSEM-EVAL, we assembled the transcriptome of the regenerating axolotl limb; this assembly compares favorably to a previously published axolotl assembly, identifying more expressed and differentially expressed genes, many of which are known to play a role in limb regeneration.

Computational identification of noncoding cancer drivers from whole-genome sequencing data

Ekta Khurana[1*]

[1]Weill Cornell Medical College, United States [*]ekk2003@med.cornell.edu

Plummeting sequencing costs have led to a great increase in the number of personal genomes. Interpreting the large number of variants in them, particularly in non-coding regions, is a central challenge for genomics. We investigated patterns of selection in DNA elements from the ENCODE project using the full spectrum of sequence variants from 1,092 individuals in the 1000 Genomes Project, including single-nucleotide variants (SNVs), short insertions and deletions (indels) and structural variants (SVs). We analyzed both coding and non-coding regions, with the former corroborating the latter. We identified a specific sub-group of non-coding categories that exhibit very strong selection constraint, comparable to coding genes: “ultra-sensitive” regions. We also find variants that are disruptive due to mechanistic effects on transcription-factor binding (i.e. "motif-breakers"). Using connectivity information between elements from protein-protein interaction and regulatory networks, we find that variants in regions with higher network centrality tend to be deleterious. Indels and SVs follow a similar pattern as SNVs, with some notable exceptions (e.g. certain deletions and enhancers). Using these results, we developed a scheme and a practical tool to prioritize non-coding variants based on their potential deleterious impact. In particular, identification of noncoding cancer "drivers" from thousands of somatic alterations is a difficult and unsolved problem. We developed the computational framework to annotate and prioritize cancer regulatory mutations. The framework combines an adjustable data context summarizing large-scale genomics and cancer-relevant datasets with an efficient variant prioritization pipeline. To prioritize high impact variants, we developed a weighted scoring scheme to score each mutation’s impact through analyzing conservation, loss-of and gain-of function events, gene associations, network topology and across-sample recurrence. Cancer specific information is used to further highlight potential oncogenic relevant candidates. Using this scheme, we identified candidate noncoding drivers in 570 samples from 10 different cancer types. This approach can be readily used in precision medicine to prioritize variants.

SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability

Daria Iakovishina[1 2], Isabelle Janoueix-Lerosey[3], Emmanuel Barillot[3],

Mireille Regnier[1 2] and Valentina Boeva[2*]

[1] INRIA projet AMIB, France
[2] Ecole Polytechnique, France
[3] Institut Curie, France *valentina.boeva@curie.fr

Motivation: Whole genome sequencing of paired-end reads can be applied to characterize the landscape of large somatic rearrangements of cancer genomes. Several methods for detecting structural variants with whole genome sequencing data have been developed. So far, none of these methods has combined information about abnormally mapped read pairs connecting rearranged regions and associated copy number changes. Our aim was create a computational method that could use both types of information, i.e., normal and abnormal reads, and demonstrate that by doing so we can highly improve both sensitivity and specificity rates of structural variant prediction.

Results: We developed a computational method, SV-Bay, to detect structural variants from whole genome sequencing mate-pair or paired-end data using a probabilistic Bayesian approach. This approach takes into account depth of coverage by normal reads and abnormalities in read pair mappings. To estimate the model likelihood, SV-Bay considers GC-content and read mappability of the genome, thus making important corrections to the expected read count. For the detection of somatic variants, SV-Bay makes use of a matched normal sample when it is available. We validated SV-Bay on simulated datasets and an experimental mate-pair dataset for the CLB-GA neuroblastoma cell line. The comparison of SV-Bay with several other methods for structural variant detection demonstrated that SV-Bay has better prediction accuracy both in terms of sensitivity and false positive detection rate.

Availability: https://github.com/InstitutCurie/SV-Bay
Contact: SV.Bay@curie.fr

Reconstruction of clonal trees and tumor composition from multi-sample cancer sequencing data

Mohammed El-Kebir[1*], Layla Oesper[1†], Hannah Acheson-Field[1] and Ben Raphael[1‡]

[1] Brown University, United States [*]mohammed_el-kebir@brown.edu
[†]layla@cs.brown.edu
[‡]braphael@brown.edu

Cancer is a disease resulting from somatic mutations that accumulate during an individual's lifetime. The clonal theory of cancer (Nowell, 1976) posits that a tumor evolves over time as di_erent descendants of the original founding cell acquire new somatic mutations. A clone is a set of cells having the same complement of somatic mutations. Since somatic mutations are typically measured in human solid tumors only at a single time point, when the patient undergoes surgery, the clonal evolution is not directly observed. Thus, one is faced with the problem of inferring the ancestral relationships between cells in a tumor from measurements at one time point. Recent studies that sequence multiple samples of a tumor from the same time point provide additional data to analyze the process of clonal evolution in the population of cells that give rise to a tumor.

We apply the resulting AncesTree algorithm to 22 tumors from three different studies (Schuh et al., 2012; Zhang et al., 2014; Gerlinger et al., 2014). We find that AncesTree is better able to identify ancestral relationships between individual mutations than existing approaches, particularly in ultra-deep sequencing data when high read counts for mutations yield high con_dence variant allele frequencies. For instance, Figure 1 shows the clonal tree inferred by AncesTree for chronic lymphocitic leukemia (CLL) patient 077 (Schuh et al., 2012). The structure of this tree closely resembles that found by other algorithms for inferring tumor evolutionary history and composition (Jiao et al., 2014; Malikic et al., 2015). However, our method is able to exploit the high coverage of the input data and con_dently orders mutations as successive clonal expansion whereas previous methods are unable to do so.

Resolving ambiguities in tumour phylogenies

Amit G. Deshwar[1 4], Levi Boyles[5], Jeff Wintersinger[2 4], Paul C. Boutros[6],

Yee Whye Teh[5] and Quaid Morris[1 2 3 4*]

*Presenting author.
[1]Edward S. Rogers Sr. Department of Electrical and Computer Engineering,
[2]Department of Computer Science,
[3]Department of Molecular Genetics
[4]Donnelly Center for Cellular and Biomolecular Research, University of Toronto.
[5]Department of Statistics, University of Oxford.
[6]Ontario Institute for Cancer Research, Toronto Ontario

Tumours are composed of genetically heterogeneous subpopulations of cancerous cells that can differ in their metastatic potential and response to treatment. Somatic mutations present in some (or all) of the tumour subpopulations can be identified using high-throughput, short-read sequencing. Also, by clustering the variant allele frequencies (VAFs) of these somatic mutations (estimated based on sequencing depth), it is possible to not only estimate the number of major subpopulations under selection during the evolution of the tumour, but also determine a set of mutations unique to each subpopulation when it first expanded. However, further analysis (and possibly different data) is needed to determine how the subpopulations relate to one another and whether they share any mutations. Answering these questions can be important for designing drug cocktails to target all tumour cells; and for understanding the acquisition of multiple cancer hallmarks as well as the biogenesis of metastatic, drug-resistant tumours.

A widely made, and relatively safe, assumption is that each somatic single nucleotide variant (SNV) (i.e. mutation) only occurred once during the evolution of the tumour. This 'infinite sites assumption' can sometimes permit full reconstruction of subpopulation genotypes based solely on cellular prevalence. For example, if a set of mutations, A, is present in 80% of the tumour cells and another, B, is in 60% then some cells must contain both A and B, and therefore, by the infinite sites assumption, the mutation B must have occurred in cells already containing A. So, in this case there are three tumour subpopulations: 20% of cells have neither A nor B, 20% of cells contain only A; and 60% of cells have both A and B. However, if the cellular prevalences were only 40% (A) and 30% (B) then it is unclear whether both A and B occur in the same cell.

We will present recent work in our labs aimed at resolving these structural ambiguities in tumour phylogenies. In some cases, it is possible to determine the whether or not two SNVs are both present in the same cell. We have developed a new method, PhyloSpan, to incorporate data from multi-SNV loci such as these. Although it is rare that two SNVs are close enough to be covered by a single read pair; only a handful of multi-SNV loci can resolve substantial structural ambiguity. Also, long (>10k) read technologies, such as PacBio, can be used to supplement short read sequence. Our approach also generalizes to permit the integration of single cell sequencing with bulk tumour sequencing and we can use our framework to identify informative nearby SNV pairs. We will present data on how often whole genome sequenced (WGS) tumour samples have useful multi-SNV loci and we can show, in simulation, that these loci can be used to resolve structural ambiguities even when the SNVs are not phased. We are currently applying our method to real tumour WGS data. Also, we have developed a method, PhyloWGS, that combines SNVs with copy number variants. In some cases, it is possible to uniquely infer branching phylogenies from a single sample using PhyloWGS.

A Waiting Time Model for Mutually Exclusive Cancer Alterations

Simona Constantinescu[1*], Ewa Szczurek[1], Pejman Mohammadi[2], Jörg Rahnenführer[3] and Niko Beerenwinkel[1†]

[1] ETH Zurich, Switzerland [*]simona.constantinescu@bsse.ethz.ch
[†]niko.beerenwinkel@bsse.ethz.ch
[2] NYGC, United States
[3]Technische Universität Dortmund, Germany

Despite recent technological advances in genomic sciences, our understanding of cancer progression and its driving genetic alterations is still incomplete. Here, we introduce TiMEx, a generative probabilistic model for detecting patterns of various degrees of mutual exclusivity across genetic alterations, which can indicate pathways involved in cancer progression. TiMEx explicitly accounts for the temporal interplay between the waiting times to alterations and the observation time. In simulation studies, we show that our model outperforms previous methods for detecting mutual exclusivity. On large-scale biological datasets, we show that TiMEx identifies gene groups with stronger functional biological relevance than other methods, while also proposing many new candidates for biological validation. TiMEx possesses several advantages over previous methods, including a novel generative probabilistic model of tumorigenesis, direct estimation of the probability of mutual exclusivity interaction, computational eciency, as well as high sensitivity in detecting gene groups involving low-frequency alterations.

The R code implementing our procedure is available at
www.cbg.bsse.ethz.ch/software/TiMEx

Extraction of Latent Probabilistic Mutational Signature in Cancer Genomes

Yuichi Shiraishi[1*], Georg Tremmel[1], Satoru Miyano[1] and Matthew Stephens[2]

[1] University of Tokyo, Japan [*]yshira@hgc.jp
[2] University of Chicago, United States

Thanks to the advances in recent high throughput sequencing technologies, massive amounts of somatic mutations from cancer genome sequencing data have become available. Accordingly, it is now possible to detect characteristic patterns of somatic mutations or "mutation signatures," which reflect driving forces causing somatic mutations, at an unprecedented resolution. And revealing novel mutation signatures can lead to identification of novel mutagens and prevention of cancer.

For extracting prominent mutation signatures from vast amounts of somatic mutation data, some statistical latent variable models are necessary. Currently, only a few statistical approaches for extracting characteristic mutation signatures have been proposed, and majority of the researchers are using the framework using nonnegative matrix factorization proposed by Alexandrov et al. 2013. However, in this approach, since the number of parameters increases exponentially as we increase the number of contextual factors to take into account, we could not treat many contextual factors at a time because of instability of estimation results. Furthermore, interpretation of mutations signatures of huge dimensional vectors is often troublesome (see Figure 1).

In this paper, we propose a novel approach based on hierarchical probabilistic modeling. The proposed approaches adopt the independence assumption on each factor of mutation signatures so that we can obtain more robust and interpretable estimates (see Figure 2) because of reduced number of parameters. In addition, we clarify the relationships between the proposed approach and the "mixed-membership models," that have been actively studied in statistical machine learning and statistical genetics community. Recognizing these relationships will help us to develop a grape of the formulation of mutation signature extraction problems, and will allow us to utilize a lot of techniques accumulated on those fields to further improve the statistical methods. Using synthetic and real data, we demonstrate that the proposed approach can not only gives us highly robust estimates, but also capture novel characteristics of mutation signature such as base frequency at the two base 5' to the mutated sites.

The R package of the proposed approach (probabilistic mutation signature, pmsignature) is available at https://github.com/friend1ws/pmsignature

A method for combining multiple genomic and clinical datatypes to predict recurrence grade in gliomas

Isaac Joseph[1*], Shannon McCurdy[1], Lior Pachter[1] and Joseph F. Costello[1]

[1] University of California, United States [*]ijoseph@berkeley.edu

A central goal of oncology is making treatment decisions that result in the best patient outcome, which must be consequently predicted. In gliomas, a common type of brain tumor with a nearly 100% recurrence rate following initial surgery, recurrence grade is an impactful, yet difficult to predict outcome. Molecular evidence suggests that recurrence grade may be increased by application of chemotherapy agent Temozolomide, modulated by various known molecular mutation-related and adduct-removal pathways. Consequently, the usage of high-throughput sequencing (HTS) genomic data of multiple types may improve our ability to predict recurrence grade by measuring related molecular mutations and epigenetic alterations. Framed as a machine-learning problem, the high-dimensionality of genomic data poses a prediction accuracy challenge due to overfitting, and consequently limits transferability of findings to further patients. To reduce overfitting, we developed and implemented a novel statistical dimensionality reduction method that relies on the following assumptions: (1) correlation of genomic modalities being informative for outcome, (2) linear relationship of predictors to response, and (3) collinearity of genomic/ clinical predictors. Initial tests of an approximation of the method shows an improved generalizability and interpretability of the relationship between high-dimensional genomics data and clinical outcomes within data from The Cancer Genome Atlas and UCSF Department of Neurooncology. Successful application of the method results in identification of multi-HTS-assay signatures of recurrence grade, useful for understanding mechanisms and treatment decisions in gliomas.

Personalized Targeted Therapy for Refractory Childhood Cancers

Mathieu Lajoie[1*], Sylvie Langlois[1], Pascal St-Onge[1], Patrick Beaulieu[1], Jasmine Healy[1] and Daniel Sinnett[1]

[1] Research Center, Sainte-Justine Hospital, Canada [*]mathieu.lajoie@gmail.com

In Canada, about 1500 pediatric cancers are diagnosed each year. Despite improvements in risk-based treatment protocols, ~20% of childhood cancer patients do not respond to current therapies and ultimately succumb to their disease, urging the need for new and more effective therapeutic approaches. Since individual tumours of the same clinical type harbour diverse sets of genomic alterations that drive oncogenesis and modulate drug response, personalized targeted therapy based on next generation sequencing may be a key to increase cure rates and decrease treatment-related morbidity and mortality.

In this project, we implemented an automated pipeline to identify single nucleotide variants, indels, gene fusions, and copy number variations from DNA and RNA sequencing data. We use Perl wrappers to encapsulate well-established softwares (e.g. Bowtie2, STAR, Picard Tools) and connect them using a common interface. Our goal is to detect prognostic markers and drug-actionable targets in a rapid and standardized way, going from biopsy to detailed tumour analysis within a clinically-relevant timeframe. Actionable variations are validated by re-sequencing or qPCR, detailed in a report with supportive information and communicated to the physician. Preliminary results on 10 cases indicate the feasibility and great potential of our approach; indeed, the final report can be in the treating oncologist's hands in less than 8 weeks. This represents a first step towards the implementation of a targeted therapy program for children with relapse or refractory cancer.

Integrated analysis of mutual exclusivity and gene interaction in Pan-Cancer dys-regulated pathways

Teresa Przytycka[1*]

[1] National Center for Biotechnology Information, Canada [*]przytyck@ncbi.nlm.nih.gov

The data gathered by the Pan-Cancer initiative has created an unprecedented opportunity for illuminating common features across different cancer types. However separating tissue specific features from across cancer signatures has proven to be challenging. One of the often-observed properties of the mutational landscape of cancer is the mutual exclusivity of cancer driving mutations. Even though studies based on individual cancer types suggested that mutually exclusive pairs often share the same functional pathway, the relationship between across cancer mutual exclusivity and functional connectivity has not been previously investigated. Here we introduce a classification of mutual exclusivity into three basic classes: within tissue type exclusivity, across tissue type exclusivity, and between tissue type exclusivity. We then combined across-cancer mutual exclusivity with interactions data to uncover pan-cancer dysregulated pathways. Our method, Mutual Exclusivity Module Cover (MEMCover) not only identified previously known Pan-Cancer dysregulated subnetworks but also novel subnetworks whose across cancer role has not been appreciated well before. In addition, we demonstrate several interesting properties of mutual exclusivity classes.

Dirichlet Process Mixture Model with Bayesian Lasso for consistent clustering of Survival Times with Molecular Data

Ashar Ahmad[1*] and Holger Froehlich[1]

[1] Bonn Aachen Institute of Information Technology, Germany [*]ashar799@gmail.com

We develop a fully Bayesian Model for the identification of relevant molecular signatures associated with survival times. Our probabilistic graphical model combines a Dirichlet Process Mixture Model for Clustering based on molecular data with a Bayesian Lasso for the Accelerated Failure Time Model. Our approach does not need pre-specified number of cluster, instead using a model based

criterion to discover the number of clusters. A Conditionally Conjugate Hierarchical Multivariate Gaussian Mixture Model has been used to make inference more robust.

Gibb's Sampling has been employed to make model inference. Our approach handles censored survival times as hidden variables in the graphical model and makes inference over them. Pathway-based Principal Components Analysis is used for reducing the dimensionality of molecular data.

The principal goal of this model is to identify key prognostic signature or bio-markers for each of the cancer sub-types. This approach is developed with the specific problem of Glioblastoma in mind. It has been relatively well known that Glioblastoma has four different sub-types which show significant differences in their survival curves, this association with survival times is, however, post-hoc and is not built into the model of clustering itself. Our model alleviates the shortcomings of the past methodology and results in coherent clustering patterns between survival times and molecular data.

The Graphical model implemented makes it relatively straightforward for the integration of multiple sources of molecular data.

The bottleneck of metastasis formation: insights from a stochastic model

Ewa Szczurek[1*], Tyll Krüger[2], Barbara Klink[3] and Niko Beerenwinkel[1]

[1] ETH Zurich, Switzerland [*]ewa.szczurek@bsse.ethz.ch
[†]niko.beerenwinkel@bsse.ethz.ch
[2] Wroclaw University of Technology, Poland
[3] Dresden University of Technology, Germany

Metastasis formation is a complex process in which cancer cells spread from their primary tumor of origin to distant organs where they initiate new tumors. However, only a tiny fraction of disseminating tumor cells succeeds in establishing a stable colony. There is evidence that while the early steps of this process, including release to the vascular system and infiltration of the secondary organ are efficient, the bottleneck is the initial expansion of the metastatic colony in the new environment. Here, we study this rate-limiting step of metastasis initiation in a quantitative fashion using a size-dependent branching process model. Our model adds to a systematic understanding of metastasis formation and may help defining medical intervention strategies.

We compute the probability of metastatic colony survival and derive critical colony sizes under different plausible initial growth assumptions. One refers to self-stimulation, where tumor cells benefit from each-others company, and another to surface hostility, where the tumor cells on the boundary are exposed to detrimental forces of the alien environment. Using established models of primary tumor growth together with our metastasis initiation model, we further obtain the probability of metastatic invasion and expected patient survival given the tumor size. These models fit well to epidemiological data collected for eleven cancers, were validated with independent datasets, and used to predict the impact of treatment delay on metastasis incidence and survival.

Linking Signaling Pathways to Transcriptional Programs in Breast Cancer

Hatice U. Osmanbeyoglu[1*], Raphael Pelossof[1†], Jacqueline F. Bromberg[1‡] and Christina S. Leslie[2◇]

[1] MSKCC, United States [*]uosmanbey@gmail.com
[†]pelossof@cbio.mskcc.org
[‡]bromberj@mskcc.org
[2] Memorial Sloan-Kettering Cancer Center, United States [◇]cleslie@cbio.mskcc.org

Cancer cells acquire genetic and epigenetic alterations that often lead to dysregulation of oncogenic signal transduction pathways, which in turn alters downstream transcriptional programs. Numerous methods attempt to deduce aberrant signaling pathways in tumors from mRNA data alone, but these pathway analysis approaches remain qualitative and imprecise. In this study, we present a statistical method to link upstream signaling to downstream transcriptional response by exploiting reverse phase protein array (RPPA) and mRNA expression data in The Cancer Genome Atlas (TCGA) breast cancer project. Formally, we use an algorithm called affinity regression to learn an interaction matrix between upstream signal transduction proteins and downstream transcription factors (TFs) that explains target gene expression. The trained model can then predict the TF activity, given a tumor sample's protein expression profile, or infer the signaling protein activity, given a tumor sample's gene expression profile. Breast cancers are comprised of molecularly distinct subtypes that respond differently to pathway-targeted therapies. We trained our model on the TCGA breast cancer data set and identified subtype-specific and common TF regulators of gene expression. We then used the trained tumor model to predict signaling protein activity in a panel of breast cancer cell lines for which gene expression and drug response data was available. Correlations between inferred protein activities and drug responses in breast cancer cell lines grouped several drugs that are clinically used in combination. Finally, inferred protein activity predicted the clinical outcome within the METABRIC Luminal A cohort, identifying high- and low-risk patient groups within this heterogeneous subtype.

Algorithms to Model Single Gene, Single Chromosome, and Whole Genome Copy Number Changes Jointly in Tumor Phylogenetics

Salim Akhter Chowdhury[1], Stanley E. Shackney[2], Kerstin Heselmeyer-Haddad[3], Alejandro A. Schäffer[3] and Russell Schwartz[1*]

[1] Carnegie Mellon University, United States [*]russells@andrew.cmu.edu
[2] Intelligent Oncotherapeutics, United States
[3] National Institutes of Health, United States

We present methods to construct phylogenetic models of tumor progression at the cellular level that include copy number changes at the scale of single genes, entire chromosomes, and the whole genome. The methods are designed for data collected by fluorescence in situ hybridization (FISH), an experimental technique especially well suited to characterizing intratumor heterogeneity using counts of probes to genetic regions frequently gained or lost in tumor development. Here, we develop new provably optimal methods for computing an edit distance between the copy number states of two cells given evolution by copy number changes of single probes, all probes on a chromosome, or all probes in the genome. We then apply this theory to develop a practical heuristic algorithm, implemented in publicly available software, for inferring tumor phylogenies on data from potentially hundreds of single cells by this evolutionary model. We demonstrate and validate the methods on simulated data and published FISH data from cervical cancers and breast cancers. Our computational experiments show that the new model and algorithm lead to more parsimonious trees than prior methods for single-tumor phylogenetics and to improved performance on various classification tasks, such as distinguishing primary tumors from metastases obtained from the same patient population.

Deep sequencing characterization of Sus scrofa piRNA fraction shared between female and male gonads

Aleksandra Swiercz[1 2*], Dorota Kowalczywiewicz[2], Luiza Handschuh[2],

Katarzyna Lesniak[2], Marek;Figlerowicz[2] and Jan;Wrzesinski[2]

[1] Institute of Computing Science, Poznan University of Technology, Poland
[*]aswiercz@cs.put.poznan.pl
[2] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland

Small non-coding RNAs (snc RNA) are essential for proper germ cell development and analysis of mechanisms involving these RNAs in germ cell regulation is a big challenge of molecular genetics. To address this challenge we characterized three families of small RNAs (piRNA, miRNA and tRF) present in Sus scrofa gonads using Illumina sequencer. Particular attention was paid for RNA fraction shared between male and female gonads. In the case of piRNA, we demonstrated that despite of similar number of reads for the gonad's piRNAs, the number of unique piRNA sequences in ovaries was almost several times lower. In addition 2.5% of piRNA and 10% of miRNA occurring in testis were also presented in ovaries.

Notably, the majority of the shared piRNAs mapped to ribosomal RNAs and were derived from clustered loci. In addition, the most abundant miRNAs present in the ovaries and testes are conserved and are involved in many biological processes such as the regulation of homeobox genes, the control of cell proliferation, and carcinogenesis. Unexpectedly, we detected a novel sncRNA type, the tRFs, which are 30–36-nt RNA fragments derived from tRNA molecules, in gonads. The tRF family has been suggested to be involved in the stress response and in tumor suppression.

Analysis of S. scrofa piRNAs show that testes specific piRNAs are biased for 5' uracil but both testes and ovaries specific piRNAs are not biased for adenine at the 10th nucleotide position. These observations indicate that adult porcine piRNAs are predominantly produced by a primary processing pathway or other mechanisms and secondary piRNAs generated by ping-pong mechanism are absent.

GPU-accelerated whole genome assembly

Michał Kierzynka[1*], Wojciech Frohmberg[1], Jacek Błażewicz[1], Piotr Żurkowski[1],

Marta Kasprzak[1] and Paweł Wojciechowski[1]

[1] Poznan University of Technology, Poland [*]michal.kierzynka@cs.put.poznan.pl

Sequencing has recently become a primary method used by life scientists to investigate biologically relevant problems related to genomics. As modern sequencers can only read very short fragments of the DNA strands, an algorithm is needed to assemble them into the original sequence. When this sequence is not known beforehand, this process is called DNA de novo assembly. There are a couple of types of methods available that address this problem. However, one of them, based on the overlap-layout-consensus approach, despite its high accuracy, has recently been nearly supplanted from the market due to its time consumption, especially in the context of constantly increasing number of sequences. In response to this, we propose a new algorithm based on this classical approach, but being able to accurately handle large data sets coming from next generation sequencing machines.

We proposed a unique way to construct the DNA overlap graph model by employing the power of alignment-free sequence comparison. The novelty of our solution lies in a special sorting technique that puts similar sequences close to each other without performing the sequence alignment. This phase is very fast and serves as preselection of similar pairs of sequences. Then, an ultra fast exact sequence comparison implemented on graphics cards (GPUs) verifies previously selected candidates, resulting in very accurate results. As a consequence, both sensitivity and precision parameters of the algorithm are very high: 99% and 97%, respectively. The high performance computations employing both multiple CPUs and GPUs make the method very efficient even for large data sets.

Having the DNA graph, the algorithm goes on to traverse it in a parallel way to obtain so called contigs and scaffolds, i.e. long fragments of reconstructed genome. Again the approach is novel, as resulting contigs are precisely cut in places where the repetitive fragments are detected. Therefore, the results are more accurate compared to other state-of-the-art algorithms. The information about paired-end reads is used to further increase the accuracy of the method. Moreover, the user may visualize the dependencies and possible connections between resulting contigs and scaffolds in a form of a graph which should greatly facilitate any further genome analysis.

The software was tested on a variety of real data coming from modern Illumina sequencing machines, and was proved to deal with them particularly well. Tests show that the accuracy of the algorithm is higher compared to many well-established assemblers like WGS Celera Assembler, SOAPdenovo, Velvet or AS-ASM. As a result, we think that our method for the DNA de novo assembly may revolutionize the world of DNA assemblers.

The software is an academic and non-commercial tool and will be available publicly soon. The poster is meant to present the main algorithm and its high accuracy measured on real data.

Scaling ABySS to longer reads using spaced k-mers and Bloom filters

Shaun Jackman[1*], Karthika Raghavan[1], Benjamin Vandervalk[1], Daniel Paulino[1], Justin Chu[1], Hamid Mohamadi[1], Anthony Raymond[1], Rene Warren[1], Inanc Birol[1†]

[1] BC Cancer Agency Genome Sciences Centre, Canada [*]sjackman@bcgsc.ca [†]ibirol@bcgsc.ca

Adapting to the continually changing landscape of sequencing technology is a particular challenge when maintaining an assembly software package such as ABySS that spans years of development. It also offers opportunities for better assemblies if new algorithms capitalize on the technology improvements.

Illumina read lengths were shoter than 50 nucleotides at the initial release of ABySS, and overlapping MiSeq reads now exceed 500 nucleotides. ABySS and other de Bruijn graph (dBG) assemblers use a hash table to store k-mers, sequences of k nucleotides. A standard hash table requires memory that scales with the value of k. To make better use of longer read lengths without a commensurate increase in memory requires space-efficient data structures. In a new release of ABySS, we use spaced seeds to represent large k-mers while storing a fraction of their nucleotides. For example, two 32-mer separated by a space of 300 nucleotides represents a dBG comparable to a 364-mer dBG, while using the memory of a 64-mer dBG.

We also introduce an assembly finishing tool to close scaffolding gaps in draft assemblies. The Sealer algorithm fills these gaps by navigating a dBG represented probabilistically by a Bloom filter. Because a Bloom filter is space-efficient, we can employ multiple such filters, using smaller k to span regions of low coverage and larger k to resolve repeats.

We assemble Escherichia coli overlapping MiSeq reads with ABySS producing an assembly with a contig NGA50 of 176 kbp and no misassembled contigs, shown in Figure 1. We assemble Caenorhabditis elegans Illumina TruSeq Synthetic Long Reads and Illumina mate pair reads with ABySS producing an assembly with a scaffold NGA50 of 200 kbp. ABySS is a flexible assembly pipeline that may be used to assemble a variety of sequencing data types and read lengths, from 500 bp overlapping MiSeq reads to 10 kbp pseudo-long reads. The assembly algorithms of ABySS will scale to exploit the length of the long reads from PacBio and Oxford Nanpore, though correcting the sequencing errors from these technologies remains a challenge.

We present in this work the performance of ABySS, with a detailed look at the data structures used, and the utility of automated finishing. We demonstrate the scalability of these efficient tools to long reads and large genomes.

Conditional Entropy in Variation-Adjusted Windows Detects Positive Selection Signatures Relevant to Next Generation Sequencing

Samuel K. Handelman[1], Michal Seweryn, Ryan M. Smith, Katherine Hartmann, Danxin Wang, Maciej Pietrzak, Andrew D. Johnson, Andrzej Kloczkowski and Wolfgang Sadee

[1] Columbia University, United States

Background

The onset of the "Great Leap Forward", some 50,000 years ago, marks a dramatic shift in natural selection among humans, driving population-specific differences in the frequencies and linkage patterns of human genetic variants. In particular, shifts in the ways humans interacted with their environment and one another following the "Great Leap Forward" have likely shaped genetic differences between human populations, and genetic differences within human populations arising from frequent variants under positive selection. Such variation plays a major role in the genetics of modern health, disease, and treatment outcome. Thus, identifying regions of the genome that have recently undergone positive selection will improve our understanding of health and disease. Here, we use conditional entropy to explore the relationship between recent human adaptations and variations altering the expression level of nearby genes, known as cis-expression quantitative trait loci (eQTLs), as well as SNPs showing allelic expresion imbalance (AEI). In a modern context, both of these categories of markers are identified using next generation sequencing experiments.

Results & Conclusions

Drawn from six independent high-throughput studies, eQTLs associated with genes in pathways related to anatomical development, lipids and G-protein/GTPase signaling show especially strong positive selection compared to other markers in the same genes, when using conditional entropy. Conditional entropy gives a far stronger result than previous methods of assessing positive selection. However, a second group of eQTLs show either no selection, purifying, or balancing selection, including eQTLs in genes associated with innate immunity, and with transcription and RNA processing. SNPs showing AEI also show strong positive selection signals. All of these signals are adjusted for the background level of positive selection in the corresponding gene. Finally, positive selection signatures identified using our new method can be used to prioritize cis-acting but non-expressed driver mutations for AEI marker SNPs within the transcribed portion of a gene.

Onctopus: Combinatorial Optimization For Lineage-Based Subclonal Composition Reconstruction

Linda K. Sundermann[1*], Amit G. Deshwar[2], Quaid Morris[2] and Gunnar Rätsch[3]

[1] Bielefeld University, Germany [*]lsunderm@cebitec.uni-bielefeld.de
[2] University of Toronto, Canada
[3] Memorial Sloan Kettering Cancer Center

Cancer samples are often genetically heterogeneous, harboring subclonal populations with different mutations. Information about copy number variations (CNVs) or simple somatic mutations (SSMs; i.e., single nucleotide variants and small indels) in the sub- populations can help to identify driver mutations or to choose targeted therapies. As methods that analyze individual cells with the help of fluorescent markers or single cell sequencing still have various drawbacks (for instance, limited number of features or high cost), bulk tumor sequencing is often used.

Recently, several methods that attempt to infer the genotype of subclonal populations using either CNVs, SSMs, or both have been published. Here, we present Onctopus, a new method to reconstruct the subclonal composition of bulk tumor samples in terms of SSMs and CNVs, using information about read depth and variant count data of heterozygous germline SNPs and SSMs, as well as information about segments affected by CNVs. We define a subclonal lineage as the set of subclonal populations that contains a subclonal population and all its descendants. We model the tumor as consisting of a mixture of these lineages where each comprises a characteristic set of CNVs and SSMs. Onctopus is designed to infer a partial order on these lineages while also simultaneously phasing SSMs and SNPs whose copy number is altered by CNVs. Currently, we are refining our model and implementation and compare it to similar tools on simulated data, plan to test it on real data sets and validate it with single cell sequencing data.