HiTSeq 2014

Boston

July 11-12 2014

Presentation & Poster Abstracts

Organizing Committee

Ana Conesa, Ph.D.
Centro de Investigación Príncipe Felipe Valencia, Spain

Francisco M. De La Vega, D.Sc.
Annai Systems, Los Gatos, CA, USA.

Dirk Evers, Ph.D.
Molecular Health GmbH, Heidelberg, Germany

Kjong Lehmann, Ph.D.
Memorial Sloan-Kettering Cancer Center. New York, NY, USA

Gunnar Rätsch, Ph.D.
Memorial Sloan-Kettering Cancer Center. New York, NY, USA

Presentations

July 12, 2014 — Hynes Convention Center, Boston MA

Oral Presentations 1: 9:00am — 10:20am

Automatically Reconstructing Subclonal Composition and Evolution from Whole Genome Sequencing of Bulk Tumor Samples

Deshwar, Amit G
Vembu, Shankar
Yung, Christina
Jang, Gun Ho
Stein, Lincoln
Morris, Quaid; University of Toronto

Solid tumors often contain multiple subclonal populations of cancerous cells. These populations are defined by distinct somatic mutations that include single nucleotide variants and small indels – collectively called simple somatic mutations (SSMs) – and larger structural changes that result in copy number variations (CNVs). In some cases, the genotype and prevalence of these subclones can be reconstructed based on high-throughput, short-read sequencing of DNA in one or more bulk tumor samples. To date, CNV-based reconstructions are limited to tumors with two or fewer cancerous subclonal populations and with a small number of CNVs.

We describe a new method that incorporates CNVs and SSMs in subclonal reconstruction and demonstrate its value on subclonal reconstruction problems on WGS data from TCGA benchmarks and simulated WGS data. Our method accurately recovers subclonal composition in a TCGA benchmark for which CNV based methods fail. We also show that SSMs can supplement CNV-based subclonal reconstructions by identifying subclonal lineages missed by CNV-based methods. Through simulations using realistic parameters, we show that our method can perform subclonal reconstructions based on 30-50x coverage WGS data from single bulk tumor samples with 10's to 1000's of SSMs per subclone for as many as three cancerous subclones (i.e., four cell populations including normal) and can reliably resolve up to six populations if provided read depths of 200 or more. We also show that our method can accurately reconstruct the phylogeny of the subclonal populations and can appropriately combine overlapping SSMs with CNVs.

Our work greatly expands the range of tumor samples for which subclonal reconstruction is possible by performing accurate subclonal reconstruction of at least four subclonal populations (including normal) based on medium coverage (30-50x) WGS data from single bulk tumor samples. In particular, we show that SSM-based approaches are successful for samples for

which CNV-based approaches cannot or do not work. We also describe a principled approach to combining overlapping SSM and CNV data that is consistent with modeling assumptions implicit in subclonal reconstruction. We will release an open source Python/C++ software implementation of our method upon publication.

Quantifying Tumor Heterogeneity in Whole-Genome and Whole-Exome Sequencing Data

Oesper, Layla; Brown University, Computer Science
Satas, Gryte; Brown University, Computer Science
Raphael,Benjamin; Brown University, Computer Science

Motivation: Most tumor samples are a heterogeneous mixture of cells, including admixture by normal (non-cancerous) cells and subpopulations of cancerous cells with different complements of somatic aberrations. This intra-tumor heterogeneity complicates the analysis of somatic aberrations in DNA sequencing data from tumor samples.

Results: We describe an algorithm to infer the composition of a tumor sample – including not only tumor purity but also the number and content of tumor subpopulations – directly from both whole-genome and whole-exome high-throughput DNA sequencing data. This algorithm builds upon our earlier Tumor Heterogeneity Analysis (THetA) algorithm in several important directions. These include improved ability to analyze highly rearranged genomes using a variety of data types: both whole-genome sequencing (including low 5X coverage) and whole-exome sequencing. We apply our improved THetA algorithm to whole-genome (including low-pass) and whole-exome sequence data from 18 samples from The Cancer Genome Atlas (TCGA). We find that the improved algorithm is substantially faster and identifies numerous subclonal tumor populations in the TCGA data, including in one highly rearranged sample for which other tumor purity estimation algorithms were unable to estimate tumor purity.

Availability: An implementation of THetA is available at: http://compbio.cs.brown.edu/software

Contact: layla@cs.brown.edu, braphael@brown.edu

TIPP:Taxonomic Identification and Phylogenetic Profiling

Nguyen, Nam-phuong; University of Texas at Austin, Computer Science
Mirarab, Siavash; University of Texas at Austin, Computer Science
Liu, Bo; University of Maryland, Computer Science
Pop, Mihai; University of Maryland, Center for Bioinformatics and Computational Biology
Warnow, Tandy; University of Texas, Computer Sciences

Motivation: Abundance profiling (also called “phylogenetic profiling”) is a crucial step in understanding the diversity of a metagenomic sample, and one of the basic techniques used for this is taxonomic identification of the metagenomic reads.

Results: We present TIPP (taxon identification and phylogenetic profiling), a new marker-based taxon identification and phylogenetic profiling method. TIPP combines a highly accurate phylogenetic placement method (SEPP) with statistical techniques to control the precision and recall of the classification results. TIPP is more robust to sequencing error and has better recall than other marker-based taxon identification methods, and also yields highly accurate abundance profiles, matching or improving on many previous approaches, including NBC, PhymmBL, MetaPhyler, and MetaPhlAn.

Availability: Software and supplementary materials are available at
http://www.cs.utexas.edu/users/phylo/software/sepp/tipp-submission/.
Contact: tandy@cs.utexas.edu

KmerStream: Streaming Algorithms for k-mer Abundance Estimation

Melsted, Pall; University of Iceland, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
Halldorsson, Bjarni; deCODE genetics/Amgen, ; Reykjavík University, Biomedical Engineering

Motivation: Several applications in bioinformatics, such as genome assemblers and error corrections methods, rely on counting and keeping track of k-mers (substrings of length k). Histograms of k-mer frequencies can give valuable insight into the underlying distribution and indicate the error rate and genome size sampled in the sequencing experiment.

Results: We present KmerStream, a streaming algorithm for computing statistics for high throughput sequencing data based on the frequency of k-mers. The algorithm runs in time linear in the size of the input and the space requirement are logarithmic in the size of the input. This very low space requirement allows us to deal with much larger datasets than previously presented algorithms. We derive a simple model that allows us to estimate the error rate of the sequencing experiment, as well as the genome size, using only the aggregate statistics reported by KmerStream and validate the accuracy on sequences from a PhiX control. As an application we show how KmerStream can be used to compute the error rate of a DNA sequencing experiment. We run KmerStream on a set of 2656 whole genome sequenced individuals and compare the error rate to quality values reported by the sequencing equipment. We discover that while the quality values alone are largely reliable as a predictor of error rate, there is considerable variability in the error rates between sequencing runs, even when accounting for reported quality values.

Availability: The tool KmerStream is written in C++ and is released under a GPL license. It is freely available at https://github.com/pmelsted/KmerStream

Contact: pmelsted@hi.is

Oral Presentations 2: 10:45am — 11:45am

SplitMEM: Graphical Pan-Genome Analysis with Suffix Skips

Marcus, Shoshana; Cold Spring Harbor Larboratory, Simons Center for Quantitative Biology
Lee, Hayan; Stony Brook University, Department of Computer Science
Schatz, Michael; Cold Spring Harbor Laboratory, Simons Center for Quantitive Biology

Motivation: With the rise of improved sequencing technologies, genomics is expanding from a single reference per species paradigm into a more comprehensive pan-genome approach with multiple individuals represented and analyzed together. One of the most sophisticated data structures for representing an entire population of genomes is a compressed de Bruijn graph. The graph structure can robustly represent simple SNPs to complex structural variations far beyond what can be done from linear sequences alone. As such there
is a strong need to develop algorithms that can efficiently construct and analyze these graphs.

Results: In this paper we explore the deep topological relationships between the suffix tree and the compressed de Bruijn graph. We introduce a novel O(n log n) time and space algorithm called splitMEM, that directly constructs the compressed de Bruijn graph for a pan-genome of total length n. To achieve this time complexity, we augment the suffix tree with suffix skips, a new construct that allows us to traverse several suffix links in constant time, and use them to efficiently decompose maximal exact matches (MEMs) into the graph nodes. We demonstrate the utility of splitMEM by analyzing the pan-genomes of 9 strains of Bacillus anthracis and 9 strains of Escherichia coli to reveal the properties of their core genomes.

Availability: The source code and documentation are available open-source at
http://splitmem.sourceforge.net

Contact: mschatz@cshl.edu

Merging of Multi-String BWTs with Applications

Holt, James; University of North Carolina, Computer Science
McMillan, Leonard; University of Chapel Hill, Computer Science

Motivation: The throughput of genomic sequencing has increased to the point that is overrunning the rate of downstream analysis. This, along with the desire to revisit old data, has led to a situation where large quantities of raw, and nearly impenetrable, sequence data is rapidly filling the hard drives of modern biology labs. These datasets can be compressed via a multi-string variant of the Burrows-Wheeler Transform (BWT), which provides the side benefit of searches for arbitrary k-mers within the raw data, as well as, the ability to reconstitute arbitrary reads as needed. We propose a method for merging such datasets for both increased compression and downstream analysis.

Results: We present a novel algorithm that merges multi-string BWTs in O(LCS ∗N) time where LCS is the length of their longest common substring between any of the inputs and N is the total length of all inputs combined (number of symbols) using O(N ∗ log2 (F)) bits where F is the number of multi-string BWTs merged. This merged multi-string BWT is also shown to have a higher compressibility compared to the input multi-string BWTs separately. Additionally, we explore some uses of a merged multi-string BWT for bioinformatics applications.

Availability: The MSBWT package is available through PyPI with source code located at https://code.google.com/p/msbwt/.

Contact: holtjma@cs.unc.edu

Journaled String Tree - A Scalable Data Structure for Analyzing Thousands of Similar Genomes on your Laptop

Rahn, Rene; FU Berlin
Weese, David; FU Berlin
Reinert, Knut; FU Berlin

Motivation: Next generation sequencing (NGS) has revolutionized biomedical research in the last decade and led to a continues stream of developments in bioinformatics addressing the need for fast and space efficient solutions for analyzing NGS data. Often researchers need to analyze a set of genomic sequences which stem from closely related species or are indeed individuals of the same species. Hence the analyzed sequences are very similar. For analyses where local changes in the examined sequence induce only local changes in the results it is obviously desirable to examine identical or similar regions not repeatedly.

Results: In this work we provide a datatype which exploits data parallelism inherent in a set of similar sequences by analyzing shared regions only once. In real-world experiments we show that algorithms which otherwise would scan each reference sequentially can be speeded up by a factor of 115.

Availability: The data structure and associated tools are publicly available at http://www.seqan.de/projects/jst and are part of SeqAn, the C++ template library for sequence analysis.

Contact: rene.rahn@fu-berlin.de, knut.reinert@fu-berlin.de

Multi-Factor Data Normalization Enables the Detection of Copy Number Aberrations in Amplicon Sequencing Data

Boeva, Valentina; Curie Institute, Bioinformatics
Popova, Tatiana; INSERM, U830; Institut Curie, Genetics and Biology of Cancer
Lienard, Maxime; Institut de Pathologie et de Génétique, Bioinformatics
Toffoli, Sebastien; Institut de Pathologie et de Génétique, Bioinformatics Kamal, Maud; Institut Curie, Clinical Research Department
Le Tourneau, Christophe; Institut Curie, Department of Medical Oncology
Gentien, David; Institut Curie, Plateforme de Génomique, Département de recherche translationnelle
Servant, Nicolas; Institut Curie, U900
Gestraud, Pierre; Institut Curie, U900
Rio Frio, Thomas; Institut Curie, Next-generation sequencing platform Hupé, Philippe; Institut Curie, Bioinformatics
Barillot, Emmanuel; Institut Curie, U900 Computational Systems Biology of Cancer
Laes, Jean-François; OncoDNA, Bioinformatics

Motivation: Due to its low cost, amplicon sequencing, also known as ultra-deep targeted sequencing, is now becoming widely used in oncology for detection of actionable mutations, i.e. mutations influ- encing cell sensitivity to targeted therapies. Amplicon sequencing is based on the PCR amplification of the regions of interest, a process that considerably distorts the information on copy numbers initially present in the tumor DNA. Therefore, additional experiments such as SNP or CGH arrays often complement amplicon sequencing in clinics in order to identify copy number status of genes whose ampli- fication or deletion has direct consequences on the efficacy of a particular cancer treatment. So far there has been no proven method to extract the information on gene copy number aberrations based solely on amplicon sequencing. Results: Here we present ONCOCNV, a method that includes a multi-factor normalization and annotation technique enabling the detection of large copy number changes from amplicon sequencing data. We validated our approach on high and low amplicon density datasets and demonstrated that ONCOCNV can achieve a precision comparable to that of array CGH techniques in detecting copy num- ber aberrations. Thus, ONCOCNV applied on amplicon sequencing data would make the use of additional array CGH or SNP array experiments unnecessary.

Poster Spotlights 1: 11:45am – 12:05am

Massively Parallel Read Mapping on GPUs with PEANUT

Köster, Johannes; University Duisburg-Essen, Human Genetics, Chair of Genome Informatics
Rahmann, Sven; University Duisburg-Essen, Genome Informatics

We present PEANUT (ParallEl AligNment UTility), a highly parallel GPU-based read mapper with several distinguishing features, including a novel q-gram index (called the q-group index) with small memory footprint built on-the-fly over the reads and the possibility to output both the best hits or all hits of a read. Designing the algorithm particularly for the GPU architecture, we were able to reach maximum core occupancy for several key steps. Our benchmarks show that PEANUT outperforms other state-of-the-art mappers in terms of speed and sensitivity. The software is available at http://peanut.readthedocs.org.

An Infrastructure to Jointly Leverage Public and Private Genomic Data in a Co-Located Data/High-Performance Computing Environment

De La Vega, Francisco; Annai Systems Inc
Young, Stuart; Annai Systems Inc
Schlumpberger, Thomas; Annai Systems Inc
Pae, Ming; Annai Systems Inc
Hayek, Raja; Annai Systems Inc

Cancer is a disease of the genome in which an accumulation of genomic alterations leads to unregulated cell growth. Cancer remains a leading cause for disease worldwide with an expected incidence to increase to 21 million by 2030. Most cancer patients are treated with one-size-fits-all therapies based on the tumour's anatomic location, tissue of origin and stage, but because each tumour is distinct at the molecular level, response to standard therapies is highly variable. To target and truly personalize cancer therapies to the genomic alterations present in a particular patient's tumour, researchers need a comprehensive catalogue of the molecular alterations that arise during the formation of malignant tumours, and models of how these alterations interact to give rise to tumour phenotype. Researchers need access to enormous amounts of cancer data to develop such models and to truly personalize cancer therapies. Public data sets (e.g. 1000 genomes Project, TCGA, Target, ICGC, etc.) represent a vast resource with a tremendous body of cancer data. Combining public data sets with private data increases the power to develop diagnostic signatures and/or targeted therapy by joint analysis and validating and statistically refining the yield of private data with public data. The current issue to this methodology is the highly fragmented storage of public and private data and the inefficient access to public data. Researchers spend weeks to months downloading hundreds of terabytes of data from central repositories before computations can begin. Annai-ShareSeq is a data sharing resource in a collocated data/compute environment and combines access to public genomic data sets, infrastructure as a service (to store and access private data) with a compute environment and an array of tools to process and analyze genomic data. This environment leverages the technology we developed to create and manage the CGHub TCGA repository together with UCSC. ShareSeq is a hosted service that differs dramatically from the traditional cloud in two features: (i) formal mechanisms to store protected health information (PHI/HIPAA) securely and safely built into the system from the start; (ii) the system is specifically designed for scientific computing over large shared data sets supporting common bioinformatics workflow tools; (iii) Fast download and access to raw genomic information and metadata through the GeneTorrent protocol; and (iv) Provenance management to enhance analysis reproducibility. ShareSeq initially hosts normalized, processed data from the 1000 Genomes Project and ICGC data sets (whole genome sequence, transcriptomic, methylation, and other types of data), provides single tenant instances to store private data and a high performance compute environment with a large array of tools to analyze and compute a public/private data. Over time ShareSeq will host and increasing number of high value genomic public datasets.

Imseq and Imsim - A Software Toolkit for Immunogenetic Sequence Analysis

Kuchenbecker, Leon; Institute for Medical Genetics, Universitätsklinikum Charité
Nienen, Mikalai; Institute for Medical Genetics, Universitätsklinikum Charité
Hecht, Jochen; Institute for Medical Genetics, Universitätsklinikum Charité
Babel, Nina; Institute for Medical Genetics, Universitätsklinikum Charité
Neumann, Avidan; Institute for Medical Genetics, Universitätsklinikum Charité
Robinson, Peter Nick; Institute for Medical Genetics, Universitätsklinikum Charité

The identification of T- and B-cell clonotypes in mixed samples by enrichment and next-generation sequencing of the somatically recombined antigen receptor gene has recently been introduced as a powerful method of profiling the immune status. Applications include the tracking of clones specific for an antigen of interest in patients (e.g. virus or transplant specific cells) as well as single- or comparative analysis of entire immune repertoires.

We present "imseq", an analysis tool capable of identifying T- and B-cell receptor gene clonotypes from next-generation sequencing reads. The V-segment, J-segment and CDR3 sequence are identified from either single- or paired-end reads using fast filtering and alignment methods implemented in the SeqAn C++ library. Furthermore, we developed post-processing clustering methods for clonotype repertoires to correct for sequencing and PCR amplification errors inside the CDR3 region as well as clustering based on barcode sequences added during an enrichment-PCR.

Additionally, we also present "imsim", a simulator for the somatic VDJ-recombination process in T- and B-cell development. The simulator constructs the junction sites between the V, D and J gene segments according to a set of user-specified distributions for the modification operations used at the junction sites. It is also capable of simulating the PCR amplification process and can therefore be used in conjunction with an NGS read simulator such as Mason to generate simulated receptor gene sequence reads.

Evaluations with such simulated as well as real data show that "imseq" is capable of correctly identifying antigen receptor clonotypes. We also show that using paired-end sequencing should be preferred over single-end sequencing with the same overall read lengths in order to significantly reduce V-segment ambiguity and over-estimation of the repertoire size.

De novo Assembly of the North American Bullfrog Transcriptome with Trans-ABySS

Behsaz, Bahar
Raymond, Anthony
Nip, Ka Ming
Chiu, Readman
Vandervalk, Ben
Jackman, Shaun
Mohamadi, Hamid
Hammond, S Austin
Veldhoen, Nicholas
Helbing, Caren C
Birol, Inanc; BC Cancer Agency, Genome Sciences Centre; British Columbia Cancer Agency, Genome Sciences Centre

Whole transcriptome shotgun sequencing (RNA-seq) provides the ability to perform efficient and accurate transcriptome analysis and profiling. However, non-uniform coverage of transcripts in RNA-seq data due to variable expression level of transcripts, up to six orders of magnitude, has been a computational challenge for de novo assembly and analysis of RNA-seq data. Here, we report our updates on transcriptome assembly algorithm Trans-ABySS, and its application in a de novo assembly project to reconstruct the North American Bullfrog (Rana catesbeiana) transcriptome. We assessed our results with the CEGMA (Core Eukaryotic Gene Mapping Approach) tool which showed reconstruction of transcripts associated with 100% of 248 highly conserved core eukaryotic genes. We were able to map more than 95% of the original reads back to this assembled transcriptome. We used assemblies of RNA-seq data from different tissues to perform differential expression analysis. Certain genes were expected to be responding differently under different biological conditions. We observed that de novo transcriptome assemblies were effective in identifying those genes and estimating their expression levels, which correlated well with qPCR validation experiments. The results demonstrate that Trans-ABySS is a valuable tool for assembling transcriptomes of non-model organisms.

Oral Presentations 3: 1:30pm – 2:10pm

DIDA: Distributed Indexing Dispatched Alignment

Mohamadi, Hamid; BC Cancer Agency, Bioinformatics
Raymond, Anthony; BC Cancer Agency, Genome Sciences Centre
Vandervalk, Benjamin P; BC Cancer Agency, Genome Sciences Centre
Jackman, Shaun; BC Genome Sciences Centre, Bioinformatics; BC Cancer Agency, Genome Sciences Centre
Birol, Inanc; BC Cancer Agency, Genome Sciences Centre; British Columbia Cancer Agency, Genome Sciences Centre

Motivation: One essential application in bioinformatics affected by the high-throughput sequencing data deluge is the sequence alignment problem where nucleotide or amino acid sequences are queried against targets to find regions of close similarity. When queries are too many and/or targets are too large, the alignment process becomes a computationally challenging problem. This is especially true when targets are non-static, such as contigs in the intermediate steps of a de novo assembly process where a new index must be computed for each step.

Results: We present DIDA, a novel framework that distributes the indexing and alignment tasks into smaller subtasks over a cluster of compute nodes. It provides a workflow beyond the common practice of embarrassingly parallel implementations. DIDA is a cost-effective, scalable and modular framework for the sequence alignment problem in terms of memory usage and runtime. It can be employed in large-scale alignments to draft genomes and intermediate stages of de novo assembly runs.

Availability: The DIDA source code, sample files and user manual available through http://www.bcgsc.ca/platform/bioinfo/software/dida

Contact: ibirol@bcgsc.ca

Single Molecule-level Characterization of Heterogeneity in Bacterial Methylomes

Beaulaurier, John; Mount Sinai School of Medicine,
Zhu, Shijia; Mount Sinai School of Medicine,
Sebra, Robert; Mount Sinai School of Medicine,
Schadt, Eric; Mount Sinai School of Medicine,
Fang, Gang; Mount Sinai School of Medicine

In the bacterial kingdom, DNA methylation is catalyzed by three families of DNA methyltransferases (MTases), which attach methyl groups to either adenine (N6-methyladenine, 6mA) or cytosine (N4-methylcytosine, 4mC or 5-methylcytosine, 5mC). Many bacterial DNA MTases are encoded in the vicinity of restriction endonucleases (REs), together as restriction-modification (RM) systems, which aid in protecting cells from invading foreign DNA. Other MTases have also been described as orphan ones, which occur in the genome without an associated RE, and have been found to play important roles in regulating gene expression and virulence, as well as developing tolerance to antibiotics.

Genomic analyses suggest that some form of DNA methylation is present in nearly all bacteria, as putative DNA MTases were found in 94% of 3300+ sequenced bacterial genomes. Given the high frequency of MTase target sites throughout genomic sequences and the growing evidence suggesting regulatory roles of methylations, the potential scope for exploring the diversity of bacterial methylation and methylation-mediated gene regulation is vast. Although bisulfite sequencing allows the detection of m5C, there has been a lack of a high-throughput methodology for efficient genome-wide detection of the two major forms of methylation types in bacteria (6mA and 4mC).

Single-molecule real-time DNA sequencing (SMRTseq) technology developed recently represents a major advance enabling the detection of nearly twenty different types of chemical modifications in DNA, including all the three major types of DNA methylations in bacteria (6mA, 4mC, and 5mC). In SMRTseq, each DNA molecule (circular with adapters) is sequenced by a DNA polymerase multiple times as a real-time DNA synthesis process, and the time required for incorporation of each nucleotide (namely, inter-pulse duration, IPD) is monitored in addition to the specific base synthesized, and variation in the rate at which DNA polymerase acts ("IPD variation") has been shown to be highly correlated with the presence of modifications within the template.

Recently, we pioneered the first bacterial methylome study by comprehensively characterizing the methylated bases in a pathogenic strain of Escherichia coli at genome wide scale, defining simultaneously the specificity of all of the methyltransferases (MTases) and determining the functional consequences of these modifications in the context of molecular networks. The methylomes of a fast growing number of pathogenic and commensal bacteria are being

characterized. However, the heterogeneity of bacterial methylomes within a clonal population and its dynamics over different conditions has not yet been explored much yet. This is because current standard SMRTseq-based bacterial methylome analysis targets single nucleotide-resolution detection for a population of cells in an aggregated manner, which fundamentally limits the detection of complex heterogeneity and subtle dynamics.

Here, we propose the first framework for single molecule-level detection and characterization of bacterial methylations. The foundations of the framework are two mutually complementary methods that effectively use molecule-specific IPDs for inferring methylation states at single molecule resolution. First, we quantify the sensitivity and specificity of the methods over different IPD coverage, and demonstrate that highly accurate detections can be achieved. Second, we show that these single-molecule single-nucleotide methylation scores can be pooled together to accurately estimate the global methylation heterogeneity for each MTase motif, even at extremely low global coverage (0.001x). Furthermore, we present a statistically rigorous method for calling non-methylated motif sites at single molecule resolution, and demonstrate its reliability through validation by independent techniques and an enrichment analysis using transcriptional factor binding data. Finally, we present a probabilistic phasing method and show it can be used to infer MTase activity for single molecules, and differentiate between two types of global heterogeneity in a clonal bacterial population. We applied the new framework to characterize nine bacterial methylomes at single molecule resolution. The unique insights demonstrate the effectiveness and potential of the framework for discovering novel epigenetic heterogeneity and dynamics, which will enhance the analyses and interpretation of a fast growing number of bacterial methylomes, towards deeper insights into the diverse roles of bacterial methylations in bacterial virulence, host adaption and antibiotic resistance.

Poster Spotlights 2: 2:10pm – 2:30pm

Chimera: a Bioconductor Package for Secondary Analysis of Fusion Products

Beccuti, Marco; University of Torino, Department of Computer Sciences
Carrara, Matteo; University of Torino, Department of Molecular Biotechnology and Health Sciences
Cordero, Francesca; University of Torino, Department of Computer Sciences
Lazzarato, Fulvio; University of Torino, Department of Medical Sciences
Donatelli, Susanna; University of Torino, Department of Computer Sciences
Nadalin, Francesca; University of Udine, Department of Mathematics
Policriti, Alberto; University of Udine, Department of Mathematics
Calogero, Raffaele; University of Torino, Dept. Clinical and Biological Sciences

Motivation: The discovery of novel gene fusions can lead to a better comprehension of cancer progression and development. The emergence of deep sequencing of transcriptome, has opened many opportunities for the identification of this class of genomic alterations, leading to the discovery of novel chimeric transcripts in cancers. Nowadays, various computational approaches have been developed for the detection of chimeric transcripts. Since a standard format for the output of fusion detection tools it is missing then we have created chimera, which organizes the output of a set of fusion detection tools (chimeraScan, bellerophontes, deFuse, FusionFinder, FusionHunter, mapSplice, tophat-fusion, FusionMap, STAR) in a common data structure, thereby simplifying the selection of the functionally interesting fusion events.

Availability and implementation: Chimera is implemented as a Bioconductor package in R. The package and the vignette can be downloaded at bioconductor.org

Contact: raffaele.calogero@unito.it

Prioritizing Clinically Relevant Copy Number Variation from Genetic Interactions and Gene Function Data

Foong, Justin; University of Toronto, Computer Science; Sickkids, GGB
Girdea, Marta; University of Toronto, Computer Science; Sickkids, GGB
Stavropoulos, James; Sickkids, GGB
Brudno, Michael; University of Toronto, Computer Science; Sickkids, GGB

Motivation: It is becoming increasingly necessary to develop computerized methods for identifying the few disease-causing variants from hundreds discovered in each individual patient. This problem is especially relevant for Copy Number Variants (CNVs), which can be cheaply interrogated via low-cost hybridization arrays commonly used in clinical practice.

Results: We present a method to predict the disease relevance of CNVs that combines functional context and clinical phenotype to discover clinically harmful CNVs (and likely causative genes) in patients with a variety of phenotypes. We compare several feature and gene weighing systems at the gene and CNV levels. We combined the best performing methodologies and parameters on over 2,500 Agilent CGH 180k Microarray CNVs derived from 140 patients. Our method achieved an F-score of 91.59%, with 87.08% precision and 97.00% recall.

Herpes Beware: Eliminating False Positive Virus Detections in NGS Data Resulting from Alignment Biases

Forster, Michael; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology; Fluxus Technology Ltd,
Szymczak, Silke; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology
Ellinghaus, David; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology
Hemmrich, Georg; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology
Kraemer, Lars; Institute of Clinical Molecular Biology, Cellbiology
Mucha, Sören; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology
Wienbrandt, Lars; Christian-Albrechts-University Kiel, Department of Computer Science
Stanulla, Martin; Medical University Hannover, Department of Paedriatic Haematology and Oncology
Franke, Andre; Institute of Clinical Molecular Biology, Christian-Albrechts-University in Kiel

Motivation: Viruses are associated with several cancers, for example hepatitis B (HBV) with liver cancer or human papillomavirus (HPV) with cervical and other cancers. Recently, HBV integration into the human genome was reported at genomic rearrangement sites and tentatively associated with chromosomal instability in liver cancer. This finding fueled the search for virus/host genome integrations of known viruses in DNA or RNA sequence data of other cancer types. However, false positive virus detections can be a problem when reads align to viruses (often herpes) rather than the host.

Results: We identified highly effective filters that increase specificity without compromising sensitivity for virus/host chimera detection after paired-end sequencing and BWA-alignment. In the German Office for Radiation Protection's childhood acute lymphoblastic leukemia (ALL) study we sequenced 20 tumor and matched germline genomes from 10 patients with 80× and 40× coverage, respectively. Childhood ALL is characterised by genomic rearrangements, with radiation exposure as one suspected cause. Indeed, we found no significant evidence for virus integrations. We also applied our method to a published liver cancer transcriptome with known HBV integration. Our method eliminated 6400 false positives per 40× genome and could even detect the singleton human-phiX174-chimera caused by optical errors of the Illumina HiSeq 2000. This specificity is useful for detecting low virus integration levels using regular whole genome or whole transcriptome coverages, without the need for prior cell sorting.

Availability and Implementation: The tool Vy-PER (Virus integration detection bY Paired End Reads) is freely available on http://www.ikmb.uni-kiel.de/vy-per (or temporarily: http://www.ikmbtmp.uni-kiel.de/pibase/vy-per/index.html).

Contact: m.forster@uni-kiel.de

De novo TE Annotation with TEAM: TE Annotation from Methylation

Zynda, Gregory; Indiana University, School of Informatics and Computing

Motivation: Transposable elements (TEs) are DNA sequences that can jump and replicate throughout their host genome. The detection and classification of transposable elements is crucial since they comprise significant portions of eukaryotic genomes and may induce large-scale genome rearrangement. The number of completed genomes is growing exponentially and current de novo repeat discovery methods are insufficient. They not only misclassify many non-TE repeats such as tandem repeats, segmental duplications, and satellites, they also cant detect low copy number transposons which
are kept silent through DNA methylation.

Results: To improve the detection of low copy number transposable elements, I propose TEAM, which detects TEs in a reference genome based on its methylation signature. TEAM scans the frequencies of each methylation motif (CG, CHH, and CHG) in a sliding window across the whole genome and detects the unique methylation profiles of TEs, pseudogenes, and protein-coding genes using a hidden markov model. Not only is TEAM be more precise than existing algorithms, but it also demands less memory and processing time.

Availability: http://github.com/zyndagj/TEAM

Contact: gjzynda@indiana.edu

July 12, 2014 — Hynes Convention Center, Boston MA

Oral Presentations 4: 10:30am — 12:10pm

ClinSeK: Targeted Clinical Variant Identification from High-Throughput Sequencing Data

Zhou, Wanding
Zhao, Hao
Chong, Zechen
Eterovic, Agda
Shaw, Kenna
Meric-Bernstam, Funda
Mills, Gordon; Department of Systems Biology, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center
Chen, Ken

The current paradigm of clinical sequencing data analysis employs a lengthy discovery approach: aligning reads to the human reference assembly, discovering mutations from base to base and identifying mutations that are likely actionable. Although widely practiced, such a paradigm is not optimal for clinical applications, which demand rapid acquisition of clinically relevant, sensitive, and unambiguous molecular profiles.

We developed ClinSeK based on a knowledge-driven inversely-operating paradigm that directly tests well-characterized, clinically-relevant variants from high-throughput sequencing data, without exhaustively aligning and comparing sequencing reads to the human reference genome. We overcome challenges in analyzing repetitive regions and duplicated reads under this new paradigm. Applying ClinSeK to characterize the molecular profile of over 600 deeply sequenced cancer samples indicated that this new approach increases sensitivity in detecting low frequency variants with over 50-fold reduction in processing time than existing approaches that perform alignment and variant calling.

ClinSeK can test point mutations, indels and structural variations in single or paired samples. It supports the analysis of both DNA and RNA sequencing data. It can also serve as a quick and independent cross-validation in complement to existing variant discovery pipelines.

Multiple Group Comparisons for RNA-Seq and Stable Effect Size Estimates

Love, Michael

Comparative analysis of expression levels measured with RNA-Seq requires methods which account for the particular distribution of the data: the units of evidence of expression are discrete, positive counts of sequenced reads. One particular difficulty is that effect size estimates (fold changes) across different groups or conditions will have high variance when the counts are low, or more generally when the signal to noise ratio for the expression levels for the individual groups is low. The moderation of effect size estimates, also referred to as regularization or shrinkage estimation, can result in more stable estimates compared to the simple ratio of group averages. However, with analysis of more than two groups, arbitrary choices for the construction of the regularized model can influence the effect size estimates and the significance testing. I will present an approach for producing stable effect size estimates across groups, which is independent of the model construction and allows for the comparison of different groups, or each individual group against the common expression level.

Bayesian Transcriptome Assembly

Maretty, Lasse
Sibbesen, Jonas Andreas
Ng, Kim
Krogh, Anders

The massive throughput of second-generation RNA-sequencing methods allows for simultaneous discovery and quantification of transcript variants and has thus dramatically increased our ability to explore complex transcriptomic landscapes. However, the short sequencing fragments and noise typical of second-generation sequencing protocols complicate the assembly of transcripts from these data.

The problem of transcriptome assembly is generally divided into the subproblems of first constructing a splice-graph and subsequently estimating which combination of transcripts - or paths in the graph - and associated abundance levels best explain the data. However, finding efficient solutions to the latter problem remains a major challenge. Indeed, current algorithms depend at least partly on heuristics to maintain accuracy and more robust, probabilistic approaches are thus needed.

We introduce a new, Bayesian approach to splice-graph inference. Our main contribution is the derivation of a Bayesian model of the RNA-sequencing process, which uses a novel prior distribution over transcript abundances to model the number of expressed transcripts. Importantly, this prior does not penalise lowly abundant transcripts in contrast to existing methods. Inference of the posterior distribution over possible transcripts and abundance values is conducted using an efficient Gibbs sampling method. Samples from the posterior distribution are then used to estimate both a confidence and an abundance estimate for each possible transcript. The confidence estimates in turn determine which transcripts are included in the final assembly and thus provide a rigorous method for controlling the trade-off between recall and precision.

Using this method, we demonstrate significant improvements in both recall and precision over state-of-the-art assemblers on simulated RNA-sequencing data. More importantly, we also show marked improvements in assembly accuracy on multiple real RNA-seq datasets as determined using annotations and third-generation (PacBio) RNA sequencing data.

The inference algorithm is implemented in C++ as a complete transcriptome assembly package under the name Bayesembler and the source code and a Linux binary are freely available under the GPL license at bayesembler.binf.ku.dk.

AUTHORSHIP NOTE:
The two first authors contributed equally to this work.

Functional Alternative Splicing Analysis Using Long Read Technologies

de La Fuente, Lorena
Cristina, Martí
Rodriguez-Navarro, Susana
Moreno, Victoria
Conesa, Ana; Centro de Investigación Príncipe Felipe

RNA-seq has been claimed to be a powerful tool for the analysis of alternative splicing events. However, the analysis of the genome-wide functional implications of alternative isoforms expression remain a difficult task due to two main reasons: on one had the still existing limitations of short reads technologies and assembly algorithms to correctly predict full-length isoforms, and on the other hand, the lack of appropriate software to annotate and mine functional characteristics at the isoform resolution level. In this work we present our initial results to develop a methodology for the functional analysis of alternative isoform expression, using a mouse cell differentiation system from neural stem cells to oligodendrocytes as a test case. We have applied the smart-seq protocol to obtain full-length RNA-seq libraries and sequenced them using PacBio and Illumina technologies. We obtained around 1,2M PacBio subreads and 20 million Illumina reads per samples. The majority of the PacBio subreads contained both 3' and 5'sequencing primers and mapped at over 90% of the length of Refseq transcripts, indicating full-length transcript sequencing. PacBio reads were corrected and quantified with Illumina. We detected the expression of 6,400 transcripts of which 2,000 transcripts were differentially expressed between the two cell types. An intensive functional annotation pipeline was applied on the transcript sequences to obtain rich functional labels: GO terms, Interpro domains, miRNA target sites, functional motifs at UTR regions, repetitive sequences and post-translational modifications. Our preliminary results indicate notable functional annotation differences between the alternative isoforms of the same gene expressed in different cell types: for around 40% of the genes with alternative splicing, the expressed isoforms were annotated with different GO terms, 30% of genes expressed isoforms with differential Interpro functional domains and another 30% had differences at UTRs and miRNA target sites. Figure 2 shows some examples of annotation differences between expressed isoforms. Further characterisation of functional differences of alternative isoforms is on-going and will be presented the HiTseq meeting. These results reveal the complexity of functional differences in alternative isoforms expression and set the way for the analysis of the genome-wide functional implications of alternative splicing.

String Graph Construction using Incremental Hashing

Ben-Bassat, Ilan; Tel-Aviv University, School of Computer Science
Chor, Benny; Tel Aviv Univesity, School of Computer Science

New sequencing technologies generate larger and larger amount of sequence data at decreasing cost. Reads produced by the new technologies are relatively short. De novo sequence assembly is the problem of combining these reads back to the original genome sequence, without relying on a reference genome. This presents algorithmic and computational challenges, especially for long and repetitive genome sequences.

Most existing approaches to the assembly problem operate in the framework of de Bruijn graphs. Yet, a number of recent works employ the paradigm of string graph. We introduce a novel, hash based method for the construction of the string graph, employing a modification of the Karp–Rabin fingerprint, as well as Bloom filters. The use of probabilistic data structure might create false positive and false negative edges during the algorithm's execution, but these are all detected and corrected.

The advantages of the proposed approach over existing methods are its simplicity, and in the incorporation of established probabilistic techniques in the context of de novo genome sequencing. A preliminary implementation yields a significant improvement in time and memory with respect to the first direct string graph construction of Simpson and Durbin (2010). Further research and optimiza- tions will hopefully enable the algorithm to be incorporated, with similar performance improvement, in state of the art string graph based assembly methods.

Oral Presentations 5: 1:45pm – 3:30pm

Gustaf: Detecting and Correctly Classifying SVs in the NGS Twilight Zone

Trappe, Kathrin; Freie Universität Berlin, Department of Computer Science
Emde, Anne-Katrin; New York Genome Center,
Ehrlich, Hans-Christian; Freie Universität Berlin, Department of Computer Science
Reinert, Knut; Freie Universität Berlin, Department of Computer Science

The landscape of structural variation (SV) including com- plex duplication and translocation patterns is far from resolved. SV detection tools usually exhibit low agreement, are often geared tow- ards certain types or size ranges of variation, and struggle to correctly classify the type and exact size of SVs.

Results: We present Gustaf (Generic mUlti-SpliT Alignment Finder), a sound generic multi-split structural variation (SV) detection tool that detects and classifies deletions, inversions, dispersed duplications and translocations of 30bp or larger. Our approach is based on a generic multi-split alignment strategy that can identify SV breakpoints with base pair resolution. We show that Gustaf correctly identifies SVs especially in the range from 30 to 100 bp, which we call the NGS twilight zone of SVs, as well as larger SVs > 500 bp. Gustaf performs better than similar tools in our benchmark and is furthermore able to correctly identify size and location of dispersed duplications and translocations.

Availability: Project information and source code is available at
http://www.seqan.de/projects/gustaf/

Resolving Complex Tandem Repeats with Long Reads

Ummat, Ajay; Icahn School of Medicine at Mount Sinai, Genetics and Genomics Sciences
Bashir, Ali; Icahn School of Medicine at Mount Siani, Genetics and Genomic Sciences

Motivation: Resolving tandemly repeated genomic sequences is a necessary step in improving our understanding of the human genome. Short tandem repeats, or microsatellites, are often used as molecular markers in genetics, and clinically, variation in microsatellites can lead to genetic disorders like Huntingon's diseases. Accurately resolving repeats, and in particular tandem repeats (TRs), remains a challenging task in genome alignment, assembly, and variation calling. Though tools have been developed for detecting microsatellites in short-read sequencing data, these are limited in the size and types of events they can resolve. Single-molecule sequencing technologies may potentially resolve a broader spectrum of tandem repeats given their increased length, but require new approaches given their significantly higher raw error-profiles. However, due to inherent error profiles of the single molecule technologies, these reads presents a unique challenge in terms of accurately identifying and estimating the tandem repeats.

Results: Here we present PACMONSTR a reference-based probabilistic approach to identify the TR region and estimate the number of these TR elements in long DNA reads. We present a multi-step approach that requires as input, a reference region and the reference tandem repeat element. Initially, the TR region is identified from the long DNA reads via a 3-stage modified smith-waterman approach and then, expected number of TR elements are calculated using a pair-Hidden Markov Models (pairHMM) based method. Finally, TR based genotype selection (or clustering: homozygous/heterozygous) is performed with gaussian-mixture-models (GMMs), utilizing the Akaike information criteria (AIC), and coverage expectations.

Availability: https://github.com/alibashir/pacmonstr

Assessing Copy Number Alterations in Targeted Amplicon-Based Next Generation Sequencing Data

Grasso, Catherine
Butler, Timothy
Rhodes, Katherine
Quist, Micheal
Neff, Tanaya
Moore, Stephen
Tomlins, Scott
Beadling, Carol
Anderson, Mark
Corless, Christopher

Targeted sequencing using next generation technologies is effective in identifying sequence alterations in genomic DNA from cancer samples, and these alterations can inform clinical treatment decisions. Changes in gene copy number are also important in delivering precision medicine. While recent studies have established that copy number alterations (CNAs) can be detected in sequencing libraries prepared by hybridization-capture, there has been little effort on CNA assessment in amplicon-based libraries prepared by PCR. In this study we developed and validated an algorithm for detecting CNAs in amplicon-based sequencing data. CNAs from the algorithm mirrored those from a hybridization-capture library. Analysis of sequence data from 14 pairs of matched normal and breast carcinoma tissues revealed that data pooled from normal samples could be substituted for the matched normal without affecting the detection of clinically significant CNAs. Comparison of CNAs identified by array CGH and amplicon-based libraries across 10 breast carcinoma samples showed an excellent correlation. The correlation with FISH results was also very good, with agreement in 34 of 36 assessments. Factors found to influence the detection of CNAs included the number of amplicons per gene, the average read depth and, most importantly, the proportion of tumor within the sample. Our results show that CNAs can be identified in amplicon-based targeted sequencing data, and that their detection can be optimized by ensuring adequate tumor content and read coverage. Targeted sequencing is especially applicable to real treatment settings, because it can be used effectively on formalin fixed paraffin embedded (FFPE) samples, which is how most samples are prepped in clinical pathology. To explore the effectiveness of this approach, we will present targeted CNA data from an archival FFPE prostate cancer cohort that is representative of typical patient samples, including 53 PCa specimens with 8 tumor/lymph node pairs, 14 nodal/distant metastases, 26 treated specimens, and 3 high grade prostate core biopsies. Integrating this CNA data with other relevant data, such as point mutation and indel calls and gene expression, revealed interesting cases that suggest that targeted CNA data can be used to positively impact patient care. This techniques is already being used in routine patient care at OHSU, because it was shown to identify actionable mutations being missed by other techniques.

Characterization of Complex Structural Variants with Single Molecule and Hybrid Sequencing Approaches

Ritz, Anna; Virginia Tech, Computer Science
Bashir, Ali; Icahn School of Medicine at Mount Siani, Genetics and Genomic Sciences; Icahn School of Medicine at Mount Siani, Institute for Genomics and Multiscale Biology
Sindi, Suzanne; University of California, Merced, Applied Mathematics
Hsu, David; Pacific Biosciences,
Hajirasouliha, Iman; Brown University, Computer Science
Raphael, Benjamin; Brown University, Computer Science; Brown University, Center for Computational Molecular Biology

Motivation: Structural variation is common in human and cancer genomes. High-throughput DNA sequencing has enabled genome-scale surveys of structural variation. However, the short reads produced by these technologies limit the study of complex variants, particularly those involving repetitive regions. Recent “third-generation” sequencing technologies provide single-molecule templates and longer sequencing reads, but at the cost of higher per-nucleotide error rates.

Results: We present MultiBreak-SV, an algorithm to detect structural variants from single molecule sequencing data, paired read sequencing data, or a combination of sequencing data from different platforms. We demonstrate that combining low-coverage third- generation data with high-coverage paired read data is advantageous on simulated chromosomes. We apply MultiBreak-SV to third-generation sequencing data from four human fosmids and show that it detects known structural variants with high sensitivity and specificity. Finally, we perform a whole-genome analysis on third-generation sequencing data from a complete hydatidiform mole cell line and report 45 high-probability structural variants.

Availability: MultiBreak-SV is available at http://compbio.cs.brown.edu/software/

Contact: annaritz@vt.edu braphael@cs.brown.edu

Detecting Differential Peaks in ChIP-seq Signals with ODIN

Allhoff, Manuel; Helmholtz Institute for Biomedical Engineering, RWTH Aachen University
Seré, Kristin; Helmholtz Institute for Biomedical Engineering, RWTH Aachen University; Institute for Biomedical Engineering, Department of Cell Biology, RWTH Aachen University Medical School
Chauvistré, Heike; Helmholtz Institute for Biomedical Engineering, RWTH Aachen University; Institute for Biomedical Engineering, Department of Cell Biology, RWTH Aachen University Medical School
Lin, Qiong; Helmholtz Institute for Biomedical Engineering, RWTH Aachen University; Institute for Biomedical Engineering, Department of Cell Biology, RWTH Aachen University Medical School
Zenke, Martin; Helmholtz Institute for Biomedical Engineering, RWTH Aachen University; Institute for Biomedical Engineering, Department of Cell Biology, RWTH Aachen University Medical School
Costa Filho, Ivan; RWTH Aachen University Medical School, IZKF Computational Biology Research Group, Institute for Biomedical Engineering; Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University; Federal University of Pernambuco, Center of Informatics

Motivation: Detection of changes in DNA-protein interactions from ChIP-seq data is a crucial step in unraveling the regulatory networks behind biological processes. The simplest variation of this problem — differential peak calling — is defined as finding genomic regions, where the interaction of a particular protein with DNA changes between two cellular conditions. The great majority of peak calling methods can only analyse one ChIP-seq signal at a time and are unable to perform differential peak calling. Recently, a few approaches based on the combination of these peak callers with statistical tests for detecting differential digital expression have been proposed. However, these methods fail to detect detailed changes of protein-DNA interactions.

Results: We propose ODIN; a HMM-based approach to detect and analyse differential peaks in pairs of ChIP-seq data. ODIN is the first differential peak caller that performs genomic signal processing, peak calling and p-value calculation in an integrated framework. We also propose an evaluation methodology to compare ODIN with competing methods. The evaluation method is based on the association of differential peaks with expression changes in the same cellular conditions. Our empirical study based on several ChIP-seq experiments from both transcription factors and histone modifications shows that ODIN outperforms all competing methods.

Availability: http://costalab.org/wp/odin.

Contact: ivan.costa@rwth-aachen.de

Oral Presentations 6: 4:00pm – 5:00pm

LoRDEC: Accurate and Efficient Long Read Error Correction

Salmela, Leena; University of Helsinki, Department of Computer Science and Helsinki Institute for Information Technology HIIT
Rivals, Eric; LIRMM (CNRS - UM2), Computer Science

Motivation: PacBio Single Molecule, Real Time sequencing is a third generation sequencing technique producing long reads with comparatively lower throughput and higher error rate. Errors include numerous indels and complicate downstream analysis like mapping or de novo assembly. A hybrid strategy that takes advantage of the high accuracy of Second Generation short reads has been proposed for correcting long reads. Mapping of short reads on long reads provides sufficient coverage to eliminate up to 99% of errors, however at the expense of prohibitive running times and considerable amounts of disk and memory space.

Results: We present LoRDEC, a hybrid error correction method that builds a succinct de Bruijn graph representing the short reads, and seeks a corrective sequence for each erroneous region in the long reads by traversing chosen paths in the graph. In comparisons, LoRDEC is at least six times faster and requires at least 93% less memory or disk space than available tools, while achieving comparable accuracy.

Availability: LoRDEC is written in C++, is tested on Linux platforms, and is freely available at www.lirmm.fr/rivals/lordec

Contact: lordec@lirmm.fr

Using Maximum Likelihood Model to Assemble Genomes

Boža, Vladimír
Brejová, Broňa
Vinař, Tomáš

Modern genome assemblers are usually based either on an overlap–layout–consensus framework, or on deBruijn graphs (Miller et al., 2010). Neither of these frameworks is designed to systematically handle pair-end reads and additional heuristic steps are necessary to build larger scaffolds from assembled contigs. For example, Cerulean (Deshpande et al., 2013) uses PacBio long reads to provide information for scaffolding of contigs assembled from a short-read library, while ALLPATHS-LG (Gnerre et al., 2011) uses libraries with different insert lengths for the same purpose. It is clear that to obtain the best quality assemblies, it is neccesary to combine several datasets produced from different libraries or by different sequencing technologies. We propose a new framework that allows a systematic combination of diverse datasets into a single assembly, without requiring a particular type of data for specific heuristic steps. To this end, we have adapted a probabilistic model of Ghodsi et al. (2013), which was previously used to compare the quality of genome assemblers. In our work, the goal is to find the assembly that maximizes the likelihood given the available datasets. Our probabilistic model can capture characteristics of each dataset, such as sequencing error rate, length distribution of reads, length distribution and expected orientation of paired reads. We can thus transparently combine information from multiple diverse datasets into a single score. To test this framework, we have implemented a prototype genome assembler GAML (Genome Assembly by Maximum Likelihood) that can use any combination of insert sizes with Illumina or 454 reads, as well as PacBio reads. The starting point of the assembly are short contigs derived from Velvet (Zerbino and Birney, 2008) with very conservative settings in order to avoid assembly errors. We then use simulated annealing to combine these short contigs into high likelihood assemblies. Our preliminary results show that we can assemble genomes of up to 10 MB long with N50 sizes and error rates comparable to ALLPATHS-LG or Cerulean. Note that while ALLPATHS-LG and Cerulean require each very specific combination of datasets, GAML works on any combination. At present, we are working on improving data structures for likelihood computation and on designing a more effective set of proposal moves in order to increase the size of the genomes we can assemble. We also plan to apply our framework as a postprocessing step to improve the quality of already assembled genomes.

Bermuda: Bidirectional de novo Assembly of Transcripts with New Insights for Handling Uneven Coverage

Tang, Qingming; Toyota Technological Institute at Chicago, Computer Science
Wang, Sheng; Toyota Technological Institute at Chicago, Computer Science
Peng, Jian; MIT, CSAIL
Ma, Jianzhu; Toyota Technological Institute at Chicago, Computer Science
Xu, Jinbo; Toyota Technological Institute at Chicago, Computer Science

Motivation: RNA-seq has made feasible the analysis of a whole set of expressed mRNAs. Mapping-based assembly of RNA-seq reads sometimes is infeasible due to lack of high-quality references. However, de novo assembly is very challenging due to uneven expression levels among transcripts and also the read coverage variation within a single transcript. Existing methods either apply de Bruijn graphs of single-sized k-mers to assemble the full set of transcripts, or conduct multiple runs of assembly, but still apply graphs of single-sized k-mers at each run. However, a single k-mer size is not suitable for all the regions of the transcripts with varied coverage.

Contribution: This paper presents a de novo assembler Bermuda with new insights for handling uneven coverage. Opposed to existing methods that use a single k-mer size for all the transcripts in each run of assembly, Bermuda self-adaptively uses a few k-mer sizes to assemble different regions of a single transcript according to their local coverage. As such, Bermuda can deal with uneven expression levels and coverage not only among transcripts, but also within a single transcript. Extensive tests show that Bermuda outperforms popular de novo assemblers in reconstructing unevenly-expressed transcripts with longer length, better contiguity and lower redundancy. Further, Bermuda is computationally efficient with moderate memory consumption.

Availability: The software is available for review upon request and will be publicly available once the paper is accepted.

Poster Presentations

1. Adarsh Jose, Marna Yandeau-Nelson and Basil J Nikolau, *Functional annotation of de-novo assembled transcriptomes – is blast score sufficient?*
2. Alexandre Souvorov *Enhanced error correction of PacBio RNA-seq reads.*
3. Angad Pal Singh, Derek Chiang, Michael Morrissey, John Monahan, Elena Edelman and Sivakumar Gowrisankar *GC bias normalization in CNV calling for tumor-normal samples using targeted sequencing*
4. Beccuti M, Carrara M, Cordero F, Lazzarato F, Donatelli S ,Nadalin F, Policriti A and Calogero RA *Chimera: a Bioconductor package for secondary analysis of fusion products*
5. Benjamin P Vandervalk, Shaun D Jackman, Anthony Raymond, Hamid Mohamadi, Chen Yang, Dean A Attali, René L Warren and Inanç Birol *ABySS-Connector: Connecting paired-end reads using a Bloom filter de Bruijn graph*
6. Chang Sik Kim, Vipin Sachdeva, Martyn Winn, Kirk Jordan and Keywan Hassani-Pak *De Novo Assembly of the north american bullfrog transcriptome with Trans-ABySS*
7. Chang Sik Kim, Vipin Sachdeva, Martyn Winn, Kirk Jordan and Keywan Hassani-Pak *de novo transcriptome assembly using Trinity for large RNA-Seq datasets*
8. David Marron, Corbin D. Jones, Jinze Liu and Jan F. Prins *Sequence bias correction without transcript annotation*
9. Eric Talevich, A. Hunter Shain and Boris Bastian *CNVkit: Copy number variant detection and visualization from targeted resequencing using off-target reads*
10. Francisco De La Vega, Stuart Young, Thomas Schlumpberger, Ming Pae and Raja Hayek *An Infrastructure to jointly leverage public and private genomic data in a co-located data/high-performance computing enviornment*
11. Grace S. Shieh, Ping-Heng Hsieh, Yu-Chin Hsu, Chi-Li Sung, Fu-Fei Hsu and Lee-Young Chau *Interrogating the function of a cofactor in HeLa cells via ChIP-seq*
12. Gracie Zhipei Du, Matthias Hübenthal, Wolfgang Lieb, Andre Franke and Georg Hemmrich-Stanisak *Analysis of next generation sequencing derived smallRNA data: a comparison of current software tools*
13. Greg Zynda *De novo TE annotation with TEAM: TE Annotation from Methylation*
14. Harold Pimentel and Haiyan Huang *Biclustering as an application of sparse canonical correlation analysis*
15. Hongyi Xin, John Greth, Gennady Pekhimenko, Justin Meza, Carl Kingsford, Can Alkan and Onur Mutlu *Shifted Hamming Distance: A Fast and Accurate SIMD-Friendly Filter for Approximate String Matching in Read Mapping*
16. Jan Hoinka, Alexey Berezhnoy, Zuben E. Sauna, Eli Gilboa and Teresa Przytycka *AptaTools - A Toolbox to Cluster HT-SELEX Aptamer Pools and Lessons from its Application*

17. Jessica Pilsworth, Readman Chiu, Ka Ming Nip, T Roderick Docking, Aly Karsan and Inanc Birol *A clinical pipeline to report actionable mutation profiles in acute myeloid leukemia*
18. Johannes Koster , Sven Rahmann *Massively parallel read mapping on GPUs with PEANUT*
19. Julia Warnke, Hesham Ali *Focus: A New Multilayer Graph Model for Short Read Analysis and Extraction of Biologically Relevant Features*
20. Justin Foong, Marta Girdea, James Stavropoulos, Michael Brudno *Prioritizing Clinically Relevant Copy Number Variation from Genetic Interactions and Gene Function Data*
21. Ken Chen, Funda Meric-Bernstam, Hao Zhao, Qingxiu Zhang, Nader Ezzeddine, Lin-Ya Tang, Ping Song, Yu-Ye Wen, Yuan Qi, Yong Mao, Chacha Horombe, Lan Zhang, Tenghui Chen, Zechen Chong, Wanding Zhou, Xiaofeng Zheng, Scott Lopetz, Michael Davies, John de Groot, Stacey Moulder, Kenneth Aldape, Mark Routbort, Raja Luthra, Kenna Shaw, John Mendelsohn, Gordon Mills and Adga Eterovic *Developing and validating a targeted exome sequencing platform for routine cancer patient care*
22. Kristoffer Sahlin, Mattias Frånberg and Lars Arvestad *Improving inference with read pair data for bioinformatic tools*
23. Leon Kuchenbecker, Mikalai Nienen, Jochen Hecht, Nina Babel, Avidan Neumann and Peter Nick Robinson *Imseq and Imsim - A Software Toolkit for Immunogenetic Sequence Analysis*
24. Michael Forster, Silke Szymczak, David Ellinghaus, Georg Hemmrich, Lars Kraemer, Sören Mucha, Lars Wienbrandt, Martin Stanulla *Herpes beware: eliminating false positive virus detections in NGS data resulting from alignment biases*
25. Michael Ryan and John Weinstein *SpliceSeq 2.0: A tool for analysis and visualization of splicing variation in RNASeq data.*
26. Milton Y. Nishiyama-Jr., Marcelo S. Reis, Daniel F. Silva, Inacio L.M. Junqueira-De-Azevedo, Julia P.C. Da Cunha, Junior Barreira, Leo K. Iwai, Solange M.T. Serrano and Hugo A. Armelin *CeTICSdb: an integrated platform for analysis of heterogeneous, high-throughput -omics and mathematical modeling of biochemical reactions*
27. Muhammad Zohaib Anwar, Vimitha Manohar and Andreas Henschel *Increased Comparability Of Environmental Sequencing Efforts Of 16S rRNA Bacterial Community Profiles*
28. Shanrong Zhao *The abundance, multiplicity and distribution of multireads in RNA-Seq data analysis*
29. Stephen Johnson, Brett Trost and Anthony Kusalik *Improved quality score generation for erroneous nucleotides in simulated (meta)genomic data*
30. Sungkyoung Choi, Sungyoung Lee, Taesung Park, and Sungho Won *FARVATX: a FAmily-based Rare Variant Association Test on X chromosome*

Poster 1

Functional Annotation of de novo Assembled Transcriptomes - Is BLAST Score Sufficient

Jose, Adarsh
Yandeau-Nelson, Marna
Nikolau, Basil J

The advent of ultra-high throughput sequencing technologies has resulted in the accumulation of several terabytes of short-read transcript sequence data. Several very efficient algorithms are available to assemble these short reads into contiguous fragments of consensus sequences (contigs). However, the power of this data is limited by our ability to associate the assembled contigs with biological functions. The standard approach in the transcriptomics community to functionally annotate the novel contigs is based on the assumption that amino acid sequences that share high homology share biological function. The contigs are typically BLASTed against the GenBank database of non-redundant protein sequences and the annotations corresponding to the most significant hits are selected as the annotation for the novel sequence. But, the fragmented nature of the transcript assemblies and miss-assemblies caused by the shortness of the reads makes the BLAST scores insufficient to assign accurate functional annotations.

Sequences that are truly orthologous are known to share functionally conserved domains with most genes in a defined gene-family, even across phylogenetically distant organisms sharing gene-families having related functions. We, propose to use this property of orthologous genes by defining the problem of identifying and ranking orthologs based on sequence similarity to a problem of finding neighbors in a cross-organism sequence similarity network. We use the concept of Random Walking with Restart to query the network using genes belonging to the same family in the source organism to identify and prioritize their homologs in the newly sequenced organism. I will demonstrate how our proposed algorithm improved the quality of functional annotations for a transcriptome we assembled de-novo from Illumina short read RNA-Seq data.

The proposed algorithm successfully uses known gene family information from functionally annotated genomes, sequence homology and topological similarity in cross organism sequence similarity networks to identify functional orthologs and assign accurate functional annotations to denovo-assembled transcriptomes.

Poster 2

Enhanced Error Correction of PacBio RNA-seq Reads.

Souvorov, Alexander; National Institutes of Health, National Center for Biotechnology Information

Motivation: New single-molecule sequencing technologies, such as PacBio, can generate long RNA-seq reads which in many cases are long enough to represent full-length transcripts. However, higher raw read error rates make it harder to use these reads directly for the gene finding. Shorter Illumina reads, which have much lower error rate and are readily available for many organisms, have been successfully used for the error correction of PacBio transcripts. In these methods the long read sequence is modified using the consensus information derived from the alignments of the short and long reads. Although the simple consensus method improves the quality of the long reads dramatically, it fails to recognize that many genes have very similar but still different regions which attract short reads originated from multiple transcripts and usually will be ‘corrected’ to represent the most expressed variant. In this paper we present a method which detects variants and uses the one which is most close to the long read.

Poster 3

GC Bias Normalization in CNV Calling for Tumor-Normal Samples using Targeted Sequencing

Singh, Angad Pal
Chiang, Derek
Morrissey, Michael
Monahan, John
Edelman, Elena
Gowrisankar, Sivakumar

Multiple tools such as ExomeCNV and Control-FREEC exist to call copy numbers in exome sequencing or targeted capture. Typically these tools identify CNVs by normalizing the coverage information from the tumor against a "control" normal sample. Factors such as capture bias, sample contamination, re-sequencing artifacts and GC bias can hamper accurate CNV calling. While capture bias which is most significant is addressed by normalizing against the control sample, Control-FREEC attempts to normalize for GC bias as well although for whole genome sequencing only.

As adoption of high coverage targeted sequencing is increasing especially for cancer studies, it becomes important to obtain accurate copy number calling for these samples. We looked at targeted sequencing data from a few dozen tumor-normal samples in primary tumor xenograft data for metastatic melanoma that developed resistance to drug treatment. We filtered the samples to retain human genome reads only. Copy number data for the samples showed a large number of fragmented CNV calls often oscillating between gains, losses and zero-alteration. Also, GC plots for some of the paired samples showed distinctly differing bias between individual pairs, thus indicating a potential for a systematic bias.

We are developing a method to normalize copy number coverage data for tumor-normal samples in targeted sequencing. Our initial results show smoother CNV calling with fewer fragmented regions or deviations from the normal. We validated our methods for a BRAF amplification found on three of the tumor samples before normalization. The BRAF amplification showed identical amplification values when compared with Amplicon sequencing data as well. We plan to test this method on other cohorts and validate the samples against SNP array data for the same.

Poster 4

Chimera: a Bioconductor Package for Secondary Analysis of Fusion Products

Beccuti, Marco; University of Torino, Department of Computer Sciences
Carrara, Matteo; University of Torino, Department of Molecular Biotechnology and Health Sciences
Cordero, Francesca; University of Torino, Department of Computer Sciences
Lazzarato, Fulvio; University of Torino, Department of Medical Sciences
Donatelli, Susanna; University of Torino, Department of Computer Sciences
Nadalin, Francesca; University of Udine, Department of Mathematics
Policriti, Alberto; University of Udine, Department of Mathematics
Calogero, Raffaele; University of Torino, Dept. Clinical and Biological Sciences

Motivation: The discovery of novel gene fusions can lead to a better comprehension of cancer progression and development. The emergence of deep sequencing of transcriptome, has opened many opportunities for the identification of this class of genomic alterations, leading to the discovery of novel chimeric transcripts in cancers. Nowadays, various computational approaches have been developed for the detection of chimeric transcripts. Since a standard format for the output of fusion detection tools it is missing then we have created chimera, which organizes the output of a set of fusion detection tools (chimeraScan, bellerophontes, deFuse, FusionFinder, FusionHunter, mapSplice, tophat-fusion, FusionMap, STAR) in a common data structure, thereby simplifying the selection of the functionally interesting fusion events.

Availability and implementation: Chimera is implemented as a Bioconductor package in R. The package and the vignette can be downloaded at bioconductor.org

Contact: raffaele.calogero@unito.it

Poster 5

ABySS-Connector: Connecting Paired-End Reads using a Bloom Filter de Bruijn Graph

Vandervalk, Benjamin P; BC Cancer Agency, Genome Sciences Centre
Jackman, Shaun; BC Cancer Agency, Genome Sciences Centre
Raymond, Anthony; BC Cancer Agency, Genome Sciences Centre
Mohamadi, Hamid; BC Cancer Agency, Genome Sciences Centre
Yang, Chen; BC Cancer Agency, Genome Sciences Centre
Attali, Dean; BC Cancer Agency, Genome Sciences Centre
Warren, Rene; BC Cancer Agency, Michael Smith Genome Sciences Centre
Birol, Inanc; British Columbia Cancer Agency, Genome Sciences Centre

Motivation: Paired-end sequencing yields a read from each end of a DNA fragment, typically leaving a gap of unsequenced nucleotides in the middle. Closing this gap using information from other reads in the same sequencing experiment offers the potential to generate longer "pseudo-reads" using short read sequencing platforms. Such long reads may benefit downstream applications such as de novo sequence assembly and variant detection.

Results: We have developed ABySS-Connector, a software tool to fill in the nucleotides of the sequence gap between read pairs by navigating a de Bruijn graph. ABySS-Connector represents the de Bruijn graph using a Bloom filter, a probabilistic and memory-efficient data structure. Our implementation is able to store the de Bruijn graph using a mean 1.5 bytes of memory per k-mer, a marked improvement over the typical hash table data structure. The memory usage per k-mer is independent of the k-mer length, enabling application of this tool to large genomes. Constructing the Bloom filter and connecting the reads are parallelized and distributed over multiple machines. We report the performance of the tool on simulated and experimental datasets, and demonstrate its utility for downstream analysis.

Availability: ABySS-Connector is open-source software, free for academic use, released under the British Columbia Cancer Agency's academic license. The source code is available at https://github.com/bcgsc/abyss/tree/connector-HiTSeq2014-submission. The name of the installed ABySS-Connector executable is "abyss-connectpairs".

Contact: ibirol@bcgsc.ca

Poster 6

De novo Assembly of the North American Bullfrog Transcriptome with Trans-ABySS

Behsaz, Bahar
Raymond, Anthony
Nip, Ka Ming
Chiu, Readman
Vandervalk, Ben
Jackman, Shaun
Mohamadi, Hamid
Hammond, S Austin
Veldhoen, Nicholas
Helbing, Caren C
Birol, Inanc; BC Cancer Agency, Genome Sciences Centre; British Columbia Cancer Agency, Genome Sciences Centre

Whole transcriptome shotgun sequencing (RNA-seq) provides the ability to perform efficient and accurate transcriptome analysis and profiling. However, non-uniform coverage of transcripts in RNA-seq data due to variable expression level of transcripts, up to six orders of magnitude, has been a computational challenge for de novo assembly and analysis of RNA-seq data. Here, we report our updates on transcriptome assembly algorithm Trans-ABySS, and its application in a de novo assembly project to reconstruct the North American Bullfrog (Rana catesbeiana) transcriptome. We assessed our results with the CEGMA (Core Eukaryotic Gene Mapping Approach) tool which showed reconstruction of transcripts associated with 100% of 248 highly conserved core eukaryotic genes. We were able to map more than 95% of the original reads back to this assembled transcriptome. We used assemblies of RNA-seq data from different tissues to perform differential expression analysis. Certain genes were expected to be responding differently under different biological conditions. We observed that de novo transcriptome assemblies were effective in identifying those genes and estimating their expression levels, which correlated well with qPCR validation experiments. The results demonstrate that Trans-ABySS is a valuable tool for assembling transcriptomes of non-model organisms.

Poster 7

De novo Transcriptome Assembly using Trinity for Large RNA-Seq Datasets

Chang Sik Kim; Hartree Centre, STFC Daresbury Laboratory, Warrington, UK
Vipin Sachdeva; IBM Thomas J. Watson Research Center, Ossining, NY
Martyn Winn; Hartree Centre, STFC Daresbury Laboratory, Warrington, UK
Kirk Jordan; IBM Thomas J. Watson Research Center, Ossining, NY
Keywan Hassani-Pak; Rothamsted Research, Harpenden, UK

Due to the recent rapid advancement of high throughput sequencing platforms, the generation of RNA-Seq data has become cheaper and more robust. The large quantities of data produced by high throughput RNA sequencing have also required the advancement of methods and programs for more efficient sequence analysis. Trinity (http://trinityrnaseq.sourceforge.net) is such a package for de novo transcriptome assembly, and consists of three independent software modules: Inchworm, Chrysalis and Butterfly. The software clusters the sequence data into many de Bruijn graphs, each of which ideally represents a set of transcripts from a gene. Trinity requires high physical memory usage and extended runtime, as confirmed with our own profiling experiments. In this study, we have addressed the challenge of de novo assembly of large RNA-seq datasets on clusters of commodity processors, such as our iDataplex compute cluster based on Intel SandyBridge processors.

As an initial effort, we parallelized the most time consuming module (Chrysalis) with a hybrid implementation of MPI and OpenMP libraries to reduce the overall runtime of the Trinity workflow. The comparison of runtime for original and parallelized Chrysalis showed that the runtime was significantly reduced, by over a factor of 20 in some test cases. We also validated the hybrid parallelized Chrysalis by comparing the reconstructed transcripts to those from the original version of Chrysalis, using a range of different metrics.

Using the parallelized Chrysalis module, we are currently working on transcriptome assembly of wheat RNA-Seq dataset from Rothamsted Research. The dataset consists of ~1.5 billion reads pooled from different bread wheat cultivars. To meet the heavy physical memory requirements in the Inchworm module, ScaleMP software was used to create a virtual symmetric multiprocessing (vSMP) node for high shared memory by aggregating multiple cluster nodes. The parallelised Chrysalis module and the Butterfly processes were then run on the basic iDataPlex compute cluster. We are currently analysing two versions of the reconstructed wheat transcripts from whole and digitally normalized RNA-seq datasets, respectively. Initial result will be presented here.

Poster 8

Sequence Bias Correction Without Transcript Annotation

Marron, David; UNC Chapel Hill, Computer Science
Jones, Corbin; UNC Chapel Hill, Genetics
Liu, Jinze; University of Kentucky, Computer Science
Prins, Jan; University of North Carolina, Computer Science

Motivation: RNA-Seq is a modern sequencing method that samples and sequences both ends of mRNA transcript fragments to provide high-resolution analysis of mRNA transcription and splicing behavior. However, biases present in fragmentation and random priming result in a non-uniform sampling of reads, which impacts the accuracy of quantitative analysis. Current methods for correcting this bias rely on transcript annotations or inference. However, inference may not yield correct transcripts, and annotations may be nonexistent, incomplete, or inaccurate.

Results: We develop a method to perform sequence bias correction without knowledge of transcript annotations. The method operates on a splice graph representation constructed directly from read alignments, that may therefore incorporate novel exons and splices. We adapt two existing methods to learn the bias using likelihood models, one based on Markov chains [1] and one based on Bayesian Networks [2], and introduce an iterative method to correct the bias of individual reads on splice graphs with multiple paths (due to multiple transcript isoforms). When we correlate RNA-seq splicing data from the Microarray Quality Control samples and the experimental determination of the expression of splices using qRT-PCR, we find that our correction without annotations performs as well as method [1] with annotations, and method [2] can work at least as well in this setting with some changes. As isoform identification and quantitation substantially depend on accurate read counts across splice junctions and within exons, these results can improve downstream analyses of RNA-Seq.

Availability: This software is available at http://sourceforge.net/projects/sequencebiascorrection/

Contact: dmarron@cs.unc.edu

Poster 9

CNVkit: Copy Number Variant Detection and Visualization from Targeted Resequencing Using Off-Target Reads

Talevich, Eric
Shain, A. Hunter
Bastian, Boris

Germline copy number variants and somatic copy number alterations are found in many diseases, including cancer. Copy number can be detected using array comparative genomic hybridization (aCGH), a microarray-based assay. Next-generation sequencing is increasingly used to detect germline and somatic point mutations. Copy number can also be estimated from NGS read depth; however, this approach has limitations in the case of targeted resequencing, which leaves gaps in coverage and introduces other biases related to the efficiency of target capture and library preparation. We present a method for copy number detection, implemented in the software package CNVkit, that uses both the targeted reads and the nonspecifically captured off-target reads to infer copy number evenly across the genome. This combination achieves both exon-level resolution in targeted regions and greater overall support in the larger intronic and intergenic regions. After normalizing coverages to a pooled reference, we evaluate and correct for three biases that explain most of the extraneous variability in the sequencing read depth: GC content, target footprint, and proximity of neighboring targets. In our validation, CNVkit performed comparably to aCGH on a variety of targeted platforms, including Agilent SureSelect. In particular, we successfully inferred copy number at equivalent to 100-kilobase genome-wide resolution from a platform targeting as few as 293 genes. CNVkit is user-friendly and provides flexible visualizations, detailed reporting of significant features, and export options for compatibility with other software.

Availability: Source code is available at http://bitbucket.org/etal/cnvkit
and documentation at http://cnvkit.readthedocs.org/

Poster 10

An Infrastructure to Jointly Leverage Public and Private Genomic Data in a Co-Located Data/High-Performance Computing Environment

De La Vega, Francisco; Annai Systems Inc
Young, Stuart; Annai Systems Inc
Schlumpberger, Thomas; Annai Systems Inc
Pae, Ming; Annai Systems Inc
Hayek, Raja; Annai Systems Inc

Cancer is a disease of the genome in which an accumulation of genomic alterations leads to unregulated cell growth. Cancer remains a leading cause for disease worldwide with an expected incidence to increase to 21 million by 2030. Most cancer patients are treated with one-size-fits-all therapies based on the tumour's anatomic location, tissue of origin and stage, but because each tumour is distinct at the molecular level, response to standard therapies is highly variable. To target and truly personalize cancer therapies to the genomic alterations present in a particular patient's tumour, researchers need a comprehensive catalogue of the molecular alterations that arise during the formation of malignant tumours, and models of how these alterations interact to give rise to tumour phenotype. Researchers need access to enormous amounts of cancer data to develop such models and to truly personalize cancer therapies. Public data sets (e.g. 1000 genomes Project, TCGA, Target, ICGC, etc.) represent a vast resource with a tremendous body of cancer data. Combining public data sets with private data increases the power to develop diagnostic signatures and/or targeted therapy by joint analysis and validating and statistically refining the yield of private data with public data. The current issue to this methodology is the highly fragmented storage of public and private data and the inefficient access to public data. Researchers spend weeks to months downloading hundreds of terabytes of data from central repositories before computations can begin. Annai-ShareSeq is a data sharing resource in a collocated data/compute environment and combines access to public genomic data sets, infrastructure as a service (to store and access private data) with a compute environment and an array of tools to process and analyze genomic data. This environment leverages the technology we developed to create and manage the CGHub TCGA repository together with UCSC. ShareSeq is a hosted service that differs dramatically from the traditional cloud in two features: (i) formal mechanisms to store protected health information (PHI/HIPAA) securely and safely built into the system from the start; (ii) the system is specifically designed for scientific computing over large shared data sets supporting common bioinformatics workflow tools; (iii) Fast download and access to raw genomic information and metadata through the GeneTorrent protocol; and (iv) Provenance management to enhance analysis reproducibility. ShareSeq initially hosts normalized, processed data from the 1000 Genomes Project and ICGC data sets (whole genome sequence, transcriptomic, methylation, and other types of data), provides single tenant instances to store private data and a high performance compute environment with a large array of tools to analyze and compute a

public/private data. Over time ShareSeq will host and increasing number of high value genomic public datasets.

Poster 11

Interrogating the function of a cofactor in HeLa cells via ChIP-seq

Shieh, Grace S.; Institute of Statistical Science, Academia Sinica, Taipe, Taiwan
Hsieh, Ping-Heng
Hsu, Yu-Chin
Sung, Chi-Li
Hsu, Fu-Fei
Chau, Lee-Young

The cofactor of interest (anonymously called COF) is an essential enzyme in heme catabolism. The expression of COF is low in normal cells but induced in cancer cells. After entering nucleus, acetylation of COF promotes proliferation, migration and invasion of cancer cells. However, the biological function of COF in nucleus of cancer cells remains unclear. Here we used Illumina paired-end ChIP-seq to determine the genomewide occupancy of COF when it was induced versus not induced. Specifically, which transcription factors (TFs) COF interacts with and which target genes are regulated by the TF complexes was interrogated using Hera cells.
We mapped the sonicated reads to human genome (hg19) using Bowtie2, which resulted ~100GB in experimental and ~132GB in control groups, respectively. Next, significant peaks (p-value < 10^(-5)) were indentified via MACS. Then we mapped these peaks 3 kilo-bases (3kb) up- and downstream of transcription starting sites of RefSeq genes to determine ``COF-TF complex" target genes using CEAS; about 15 regulated genes were indentified. On the other hand, de novo motifs with sizes 6 to 7 nucleotides were searched using the significant peaks, and the detected motifs were compared with JASPAR motif database to identify potential TFs which COF interacted with.

Poster 12

Analysis of Next Generation Sequencing Derived smallRNA Data: A Comparison of Current Software Tools

Du, Gracie Zhipei ; Institute of Clinical Molecular Biology, Universität Kiel
Hübenthal, Matthias; Institute of Clinical Molecular Biology, Universität Kiel
Lieb, Wolfgang; Institute of Clinical Molecular Biology, Universität Kiel
Franke, Andre; Institute of Clinical Molecular Biology, Universität Kiel
Hemmrich-Stanisak, Georg; Institute of Clinical Molecular Biology, Universität Kiel

Next generation sequencing technology has become a popular tool for microRNA identification, transcription profiling and de novo detection. The availability of computational analysis tools for miRNA NGS-data, however, is sparse and the results must be viewed with caution. In the current study we have comprehensively evaluated the output of 6 publicly available miRNA analysis tools based on Illumina miRNA-Seq data of healthy individuals. While miRDeep2, UEA and miRanalyzer represent separately developed tools, the software packages iMir, mirTools and our in-house tool are modifications or combinations of the above. Our results generally indicate software dependent similarities but also differences for all main analytical steps: identification, transcriptional profiling as well as de novo prediction. Regarding miRNA identification, among all samples tested, miRDeep2 and miRanalyzer close, detect on average 4.6 times more miRNA species than for example the UEA toolkit. In the transcriptional profiling step, all tools show a relatively high positive correlation, regardless of the algorithms. de novo is the most challenging step during the analysis, which is also mirrored by huge differences among the tools we tested - miRDeep2 delivers moderate numbers (40-90) for novel candidates, while miRanalyzer and the UEA toolkit rarely predict more than 10. An overlap in de novo predicted miRNAs between the tools however, is missing. We also demonstrate the relevance of pre-processing and filtering smallRNA-Seq data prior to analysis to prevent false positive and false negative results. Direct comparison of miRDeep2 and our in-house pipeline shows that exhaustive adaptor trimming and filtering of non-human miRNA as well as other human ncRNAs result in increased numbers of identified miRNAs by an average of 3%, with an overlap of 98% between these two tools. Preprocessing also results in higher numbers of candidates after de novo prediction, which is
reflected by a 24% increase of predicted miRNAs compared to the standard miRDeep2 pipeline. Taken together, users should be aware of technologic and algorithmic differences when analyzing smallRNA data from NGS-experiments. Especially in the case of expression profiling and subsequent classification approaches, as conducted for many different diseases to gain predictive signatures, the choice of the software package and a proper filtering strategy might influence the results immensely.

Poster 13

De novo TE Annotation with TEAM: TE Annotation from Methylation

Zynda, Gregory; Indiana University, School of Informatics and Computing

Motivation: Transposable elements (TEs) are DNA sequences that can jump and replicate throughout their host genome. The detection and classification of transposable elements is crucial since they comprise significant portions of eukaryotic genomes and may induce large-scale genome rearrangement. The number of completed genomes is growing exponentially and current de novo repeat discovery methods are insufficient. They not only misclassify many non-TE repeats such as tandem repeats, segmental duplications, and satellites, they also cant detect low copy number transposons which
are kept silent through DNA methylation.

Results: To improve the detection of low copy number transposable elements, I propose TEAM, which detects TEs in a reference genome based on its methylation signature. TEAM scans the frequencies of each methylation motif (CG, CHH, and CHG) in a sliding window across the whole genome and detects the unique methylation profiles of TEs, pseudogenes, and protein-coding genes using a hidden markov model. Not only is TEAM be more precise than existing algorithms, but it also demands less memory and processing time.

Availability: http://github.com/zyndagj/TEAM

Contact: gjzynda@indiana.edu

Poster 14

Biclustering as an Application of Sparse Canonical Correlation Analysis

Harold, Pimentel; Electrical Engineering and Computer Science, UC Berkeley
Huang, Haiyan; Dept of Statistics, UC Berkeley

RNA expression probing (microarrays, RNA-Seq) has become sufficiently inexpensive, enabling sampling of many experimental conditions in one project. A common objective among such projects is to find genomic expression features (EF) (genes, isoforms, exon inclusion, etc.) that behave similarly in expression. Historically, this type of analysis has employed one-way clustering. With increasing sample numbers, it is practical to assume that some EF interactions diminish across some samples from diverse conditions. In such situations, one-way clustering can be insufficient, likely missing those EF as a cluster. Consequently, biclustering (or two-way clustering) methods have been introduced to find subsets of EF that behave similarly among subsets of experimental conditions.

We propose SCCA-BC, a biclustering method based on resampling, partitioning, and sparse canonical correlation analysis (SCCA), which finds numerous types of biclusters and performs well in diverse settings. By resampling and randomly separating the EF into two groups, SCCA searches for meaningful linear group relationships which, reframed in a regression setting, gives estimates proportional to partial correlations conditioned on different sets of related genes (with noisy EF eliminated through sparsity). Following that, we estimate inclusion of experimental conditions. Since SCCA-BC finds correlated biclusters, many existing models are special cases of ours and are also discoverable by SCCA-BC. Through simulation, we show that SCCA-BC performs comparably in common situations, and outperforms other methods in more difficult situations.

We applied SCCA-BC to a modENCODE data set, which consists of RNA-Seq data in developmental timepoints for 30 D. melanogaster and 14 in C. elegans. Particularly, one identified bicluster contained 760 / 3574 genes and 30 / 44 conditions. While we do not directly optimize the correlation, as doing so can often give noisy results, our resulting biclusters show higher correlation than the same set of genes across all conditions. We validated these results by performing a Gene Ontology analysis which resulted in many genes related to development, consistent with other studies.

In summary, we present a new method for biclustering based on SCCA which has the ability to find biclusters of correlated genomic expression features. Unlike many other methods, it can find numerous types of biclusters as long as the EF are linearly related in expression.

Poster 15

Shifted Hamming Distance: A Fast and Accurate SIMD-Friendly Filter for Approximate String Matching in Read Mapping

Xin, Hongyi; Computer Science Department Carnegie Mellon University
Greth, John
Pekhimenko, Gennady; Computer Science Department Carnegie Mellon University
Meza, Justin; Electrical and Computer Engineering, Carnegie Mellon University
Kingsford, Carl; Computer Science Department Carnegie Mellon University
Alkan, Can; Department of Computer Engineering, Bilkent University
Mutlu, Onur

In local alignment, the approximate string matching problem involves evaluation of whether two strings of the same length have e-or-fewer errors, including insertions, deletions and substitutions of letters. This constitutes the primary computation performed by many seed-and-extend based DNA read mappers which compare billions of pairs of short strings.

We observe that in seed-and-extend based mappers, a majority of string pairs contains errors that significantly exceed e, and are later rejected by the mapper. Such error-abundant string pairs waste significant computational resources, and severely hinder the performance of the mapper. It is, therefore, crucial to develop a fast and accurate filter that can rapidly and efficiently detect error-abundant string pairs.

In our work, we present a simple yet efficient algorithm, Shifted Hamming Distance (SHD), which utilizes bit-parallelism and SIMD-parallelism to speed up the filtering process. SHD is a generic filtering algorithm that only filters out string pairs that have e-or-more errors. It maintains high accuracy with moderate e (up to 5% of the string length). It is also compatible with all seed-and-extend class mappers.

We implemented SHD in Intel SSE and compare it against the state-of-the-art bit-parallel and SIMD-parallel approximate string matching implementations, including seqan's implementation of Gene Myers' bit-vector algorithm and swps3. We tested them in conjunction with a popular seed-and-extend mapper, mrFAST. We observe a maximum of 3x speedup over the best previous implementation of approximate string matching algorithm at the cost of a worst-case false positive rate of 8%.

Poster 16

AptaTools - A Toolbox to Cluster HT-SELEX Aptamer Pools and Lessons from its Application

Hoinka, Jan
Berezhnoy, Alexey
Sauna, Zuben E.
Gilboa , Eli
Przytycka, Teresa; NCBI, NLM, NIH, Bethesda, MD

New sequencing technologies have recently revolutionized the SELEX protocol by allowing for deep sequencing of the selection pools after each cycle. Systematic Evolution of Ligands by EXponential Enrichment (SELEX) is a well-established experimental procedure to identify aptamers – synthetic, single-stranded (ribo)nucleic molecules that bind to a given molecular target and now the emergence of High-Throughput SELEX (HT-SELEX) has opened the field to unprecedented computational opportunities and challenges that are yet to be addressed.

To aid the analysis of the results of HT-SELEX and to advance the understanding of the selection process itself, we developed AptaCluster. This algorithm allows for an efficient clustering of whole HT-SELEX aptamer pools; a task that could not be accomplished with traditional clustering algorithms due to the enormous size of such datasets. AptaCluster also identifies the relationship between clusters of consecutive selection cycles and is capable of incorporating additional information, such as control cycles, in its computations.

In addition, our suite features a full-fledged graphical user interface, AptaGUI, which enables the user to supervise all aspects of the selection process and quickly pinpoint possible sources of noise that might have been introduced at various stages of the experiment. Furthermore, it is designed to visualize AptaCluster's results in a natural and concise manner including but not limited to sequence and cluster and enrichment analysis, cluster visualization, as well as secondary structure prediction and visualization.

AptaTools is quickly gaining popularity within the aptamer community. We performed HT-SELEX with Interleukin 10 receptor alpha chain (IL-10RA) as the target molecule and used AptaTools to analyze the resulting sequences. AptaTools allowed for the first survey of the relationships between sequences in different selection rounds and revealed previously not appreciated properties of the SELEX protocol. Using AptaTools, we were able to identify a number of functional candidates with high binding affinities (nano-molar Kd) after only 5 rounds of selection. As the first tool of this kind, AptaTools enables novel ways to analyze and to optimize the HT-SELEX procedure while being flexible enough to incorporate future technological advances as these emerge.

Poster 17

A Clinical Pipeline to Report Actionable Mutation Profiles in Acute Myeloid Leukemia

Pilsworth, Jessica; Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency
Chiu, Readman; Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency
Nip, Ka Ming; Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency
Docking, T Roderick; Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency
Karsan, Aly; Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency
Birol, Inanc; Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency

Over 80% of acute myeloid leukemias (AML) contain genomic rearrangements that involve genes related to hematopoietic lineage development. The resulting transcripts are potential drivers in tumorigenesis, and therefore represent ideal diagnostic and therapeutic targets. In a retrospective study, we performed transcriptome and genome shotgun sequencing experiments on a cohort of 20 AML patients with clinical classifications. In this presentation, we will describe the design of our clinical bioinformatics pipeline, which uses an assembly-first approach based on ABySS and Trans-ABySS.

For leukemogenesis to occur, two types of mutations are required: 1) a mutation that improves hematopoietic cells' ability to proliferate (e.g. internal tandem duplication events in the FLT3 gene); and 2) a mutation that prevents the cells from differentiating (e.g. fusion events between the genes PML and RARA, and partial tandem duplication events in the MLL gene). Our analysis pipeline consistently identifies both classes of mutations, including the above three example events, which are all clinically relevant markers for AML treatment. Conventional treatment strategies include initial induction chemotherapy to achieve complete remission followed by consolidation therapy to prevent relapse. This type of treatment is successful in approximately two thirds of patients. The reported work offers a bioinformatics pipeline that can evaluate cancer patients at the genomic level first and will enable clinicians to develop 'tailored' therapies for each individual patient.

Poster 18

Massively Parallel Read Mapping on GPUs with PEANUT

Köster, Johannes; University Duisburg-Essen, Human Genetics, Chair of Genome Informatics
Rahmann, Sven; University Duisburg-Essen, Genome Informatics

We present PEANUT (ParallEl AligNment UTility), a highly parallel GPU-based read mapper with several distinguishing features, including a novel q-gram index (called the q-group index) with small memory footprint built on-the-fly over the reads and the possibility to output both the best hits or all hits of a read. Designing the algorithm particularly for the GPU architecture, we were able to reach maximum core occupancy for several key steps. Our benchmarks show that PEANUT outperforms other state-of-the-art mappers in terms of speed and sensitivity.

The software is available at http://peanut.readthedocs.org.

Poster 19

Focus: A New Multilayer Graph Model for Short Read Analysis and Extraction of Biologically Relevant Features

Warnke, Julia; University of Nebraska Medical Center, Pathology and Microbiology; University of Nebraska Omaha, College of Information Science and Technology
Ali, Hesham; University of Nebraska Omaha, College of Information Science and Technology

Motivation: With the increasing number of applications in which a group of organisms associated with a common environment are sequenced, there is an urgent need for a new model for representing the sequenced short reads in a way that takes the nature of these organisms into consideration. In addition to facilitating the assembly process, such new models should allow for easy extraction of other useful biological information from the short reads, including conserved regions among the input genomes, sequence motifs, and other information critical to the recognition and/or classification of the organisms.

Results: We present Focus, a new multilayer graph model for short read analysis and extraction of biologically relevant features. The proposed model can be viewed as a data-mining tool that takes advantage of the multilayer graph representation of the reads to extract useful information about the associated genomes/organisms. While not primarily an assembly tool, we assessed Focus using known assemblers with excellent results. We also applied Focus in a case study on a HIV read dataset and were able to successfully extract biologically relevant graph features.

Contact: hali@unomaha.edu

Poster 20

Prioritizing Clinically Relevant Copy Number Variation from Genetic Interactions and Gene Function Data

Foong, Justin; University of Toronto, Computer Science; Sickkids, GGB
Girdea, Marta; University of Toronto, Computer Science; Sickkids, GGB
Stavropoulos, James; Sickkids, GGB
Brudno, Michael; University of Toronto, Computer Science; Sickkids, GGB

Motivation: It is becoming increasingly necessary to develop computerized methods for identifying the few disease-causing variants from hundreds discovered in each individual patient. This problem is especially relevant for Copy Number Variants (CNVs), which can be cheaply interrogated via low-cost hybridization arrays commonly used in clinical practice.

Results: We present a method to predict the disease relevance of CNVs that combines functional context and clinical phenotype to discover clinically harmful CNVs (and likely causative genes) in patients with a variety of phenotypes. We compare several feature and gene weighing systems at the gene and CNV levels. We combined the best performing methodologies and parameters on over 2,500 Agilent CGH 180k Microarray CNVs derived from 140 patients. Our method achieved an F-score of 91.59%, with 87.08% precision and 97.00% recall.

Poster 21

Developing and Validating a Targeted Exome Sequencing Platform for Routine Cancer Patient Care

Ken Chen, Funda Meric-Bernstam, Hao Zhao, Qingxiu Zhang, Nader Ezzeddine, Lin-Ya Tang, Ping Song, Yu-Ye Wen, Yuan Qi, Yong Mao, Chacha Horombe, Lan Zhang, Tenghui Chen, Zechen Chong, Wanding Zhou, Xiaofeng Zheng, Scott Lopetz, Michael Davies, John de Groot, Stacey Moulder, Kenneth Aldape, Mark Routbort, Raja Luthra, Kenna Shaw, John Mendelsohn, Gordon Mills, Adga Eterovic.

Background: Recent development in the massively parallel sequencing technology and the cancer genome atlas have revealed potential impacts of applying sequencing-based assays in routine cancer patient care. However, it is unclear what constitutes an optimal assay and whether the benefits can outweigh the costs.

Methods: We established an ultra-deep (1000 x) targeted sequencing platform of 201 genes (4874 exons) to identify relevant DNA alterations in clinical tumor samples and characterized the mutational profiles of hundreds of metastatic cancer patients. We optimized our sequencing platform for FFPE specimens and low input DNA, developed a clinical-grade variant detection pipeline, and predicted actionable alterations for clinical decision making.

Results: We assayed 515 tumor samples and their normal match (blood) from 12 disease sites in 475 patients. Unlike studies based on whole-exome or whole-genome sequencing, we were able to comprehensively detect low frequency (> 1%) mutations at a low false discovery rate (< 2.2%). Our results were highly concordant (98.2%) with those from a commercial CLIA certified hotspot sequencing panel (Ion Torrent AmpliSeq46 gene panel), abeit with twice as many (82.9%) patients found to be potentially clinically actionable. About 16.3% mutations were detected at low (<10%) allele frequency in 29.5% patients. Overall, the significantly mutated genes in our set appeared highly consistent with those identified from the cancer genome atlas (TCGA). The low frequency mutations demonstrated a landscape similar to that of high frequency mutations, which indicated their potential relevance for clinical decision making.

Conclusion: Our results demonstrate that targeted exome sequencing can potentially achieve optimal cost-benefits in a cancer care institution than other platforms based on whole-exome or whole-genome sequencing. Mutations at low frequencies can be discovered reliably from FFPE specimens and are potentially important for clinical decision making.

Poster 22

Improving Inference with Read Pair Data for Bioinformatic Tools

Sahlin, Kristoffer
Frånberg, Mattias
Arvestad, Lars

Next Generation Sequencing (NGS) data are now commonly used for answering various biological questions. In many applications, insert size distribution from paired read protocols plays an important role, for example in genome assembly and structural variation detection. However, many of the the models that are being used suffer from bias. This bias arises when assuming that all insert sizes within a distribution are equally likely to be observed, when in fact, size matters. It can be shown that these systematic errors exists in popular software even when the assumptions made about data is true.

Results: We have previously shown that bias occurs for scaffolders in genome assembly where our method was constrained to this particular application and to normally distributed insert size distributions. Here, we generalize the theory and give examples to illustrate the potential use in different settings. We also relax the assumptions about normality and account for all insert size distributions using non-parametric models with binning distributions. Furthermore, coverage is introduced as an optional parameter to the model and we show how this affects results. We provide examples on where bias occurs in state-of the-art software, explain why, and improve them using our model. The results are useful for everyone working with paired read data and insert size distributions. The theory is implemented in a tool called GetDistr. GetDistr is highly modular and easily integrated in other software.

Poster 23

Imseq and Imsim - A Software Toolkit for Immunogenetic Sequence Analysis

Kuchenbecker, Leon; Institute for Medical Genetics, Universitätsklinikum Charité
Nienen, Mikalai; Institute for Medical Genetics, Universitätsklinikum Charité
Hecht, Jochen; Institute for Medical Genetics, Universitätsklinikum Charité
Babel, Nina; Institute for Medical Genetics, Universitätsklinikum Charité
Neumann, Avidan; Institute for Medical Genetics, Universitätsklinikum Charité
Robinson, Peter Nick; Institute for Medical Genetics, Universitätsklinikum Charité

The identification of T- and B-cell clonotypes in mixed samples by enrichment and next-generation sequencing of the somatically recombined antigen receptor gene has recently been introduced as a powerful method of profiling the immune status. Applications include the tracking of clones specific for an antigen of interest in patients (e.g. virus or transplant specific cells) as well as single- or comparative analysis of entire immune repertoires.

We present "imseq", an analysis tool capable of identifying T- and B-cell receptor gene clonotypes from next-generation sequencing reads. The V-segment, J-segment and CDR3 sequence are identified from either single- or paired-end reads using fast filtering and alignment methods implemented in the SeqAn C++ library. Furthermore, we developed post-processing clustering methods for clonotype repertoires to correct for sequencing and PCR amplification errors inside the CDR3 region as well as clustering based on barcode sequences added during an enrichment-PCR.

Additionally, we also present "imsim", a simulator for the somatic VDJ-recombination process in T- and B-cell development. The simulator constructs the junction sites between the V, D and J gene segments according to a set of user-specified distributions for the modification operations used at the junction sites. It is also capable of simulating the PCR amplification process and can therefore be used in conjunction with an NGS read simulator such as Mason to generate simulated receptor gene sequence reads.

Evaluations with such simulated as well as real data show that "imseq" is capable of correctly identifying antigen receptor clonotypes. We also show that using paired-end sequencing should be preferred over single-end sequencing with the same overall read lengths in order to significantly reduce V-segment ambiguity and over-estimation of the repertoire size.

Poster 24

Herpes Beware: Eliminating False Positive Virus Detections in NGS Data Resulting from Alignment Biases

Forster, Michael; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology; Fluxus Technology Ltd,
Szymczak, Silke; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology
Ellinghaus, David; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology
Hemmrich, Georg; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology
Kraemer, Lars; Institute of Clinical Molecular Biology, Cellbiology
Mucha, Sören; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology
Wienbrandt, Lars; Christian-Albrechts-University Kiel, Department of Computer Science
Stanulla, Martin; Medical University Hannover, Department of Paedriatic Haematology and Oncology
Franke, Andre; Institute of Clinical Molecular Biology, Christian-Albrechts-University in Kiel

Motivation: Viruses are associated with several cancers, for example hepatitis B (HBV) with liver cancer or human papillomavirus (HPV) with cervical and other cancers. Recently, HBV integration into the human genome was reported at genomic rearrangement sites and tentatively associated with chromosomal instability in liver cancer. This finding fueled the search for virus/host genome integrations of known viruses in DNA or RNA sequence data of other cancer types. However, false positive virus detections can be a problem when reads align to viruses (often herpes) rather than the host.

Results: We identified highly effective filters that increase specificity without compromising sensitivity for virus/host chimera detection after paired-end sequencing and BWA-alignment. In the German Office for Radiation Protection's childhood acute lymphoblastic leukemia (ALL) study we sequenced 20 tumor and matched germline genomes from 10 patients with 80× and 40× coverage, respectively. Childhood ALL is characterised by genomic rearrangements, with radiation exposure as one suspected cause. Indeed, we found no significant evidence for virus integrations. We also applied our method to a published liver cancer transcriptome with known HBV integration. Our method eliminated 6400 false positives per 40× genome and could even detect the singleton human-phiX174-chimera caused by optical errors of the Illumina HiSeq 2000. This specificity is useful for detecting low virus integration levels using regular whole genome or whole transcriptome coverages, without the need for prior cell sorting.

Availability and Implementation: The tool Vy-PER (Virus integration detection bY Paired End Reads) is freely available on http://www.ikmb.uni-kiel.de/vy-per (or temporarily: http://www.ikmbtmp.uni-kiel.de/pibase/vy-per/index.html).
Contact: m.forster@uni-kiel.de

Poster 25

SpliceSeq 2.0: A Tool for Analysis and Visualization of Splicing Variation in RNASeq Data.

Ryan, Michael; In Silico Solutions LLC, Falls Church, VA
Weinstein, John; MD Anderson Cancer Center, University of Texas, Houston, TX

SpliceSeq is a biologist friendly application for identifying changes in mRNA splicing patterns and exploring the related potential functional impact. A splice graph model for each gene is used to align reads to a single, unique transcriptome location and to serve as a common reference for comparative analysis. Weighted graph traversals are used to predict transcript isoforms and translation of these sequences are aligned with annotated UniProt sequences to provide functional context to splicing events. The SpliceSeq interface allows for user-driven thresholds, sorting, and filtering to identify significant, interesting alternative splicing events. It then provides intuitive, drill-down visualization and summary statistics for splicing events including details of exon and splice expression for each sample. SpliceSeq has been successfully utilized by a variety of studies both within and external to MD Anderson. We will present representative findings, and a companion website resource developed through the application of SpliceSeq to 650+ RNASeq samples of 24 different TCGA tumor types.

SpliceSeq is freely available for academic, government, or commercial use at http://bioinformatics.mdanderson.org/main/SpliceSeq:Overview

Poster 26

CeTICSdb: An Integrated Platform for Analysis of Heterogeneous, High-Throughput -Omics and Mathematical Modeling of Biochemical Reactions

Milton Y. Nishiyama-Jr., Marcelo S. Reis, Daniel F. Silva, Inacio L.M. Junqueira-De-Azevedo, Julia P.C. Da Cunha, Junior Barreira, Leo K. Iwai, Solange M.T. Serrano, Hugo A. Armelin

The Center of Toxins, Immune-response and Cell Signaling (CeTICS) studies biochemical, molecular, and cellular mechanisms of toxins that have therapeutic potential, aiming to understanding the short-term and long-term behavior of biological systems based on analyses of signaling networks. Once those analyses involve collaborations across disciplines such as Biology, Mathematics and Computer Science, the research developed in CeTICS is intrinsically interdisciplinary; this fact, coupled to the heterogeneous, high-throughput data produced by modern high-throughput techniques in genomics and proteomics, implies the necessity of data organization and integration to carry out scientific investigations. To this end, we are developing CeTICSdb, an integrated platform for interdisciplinary research that allows quantitative and qualitative –omics analysis and mathematical modeling of biochemical reactions (e.g., metabolic pathways, molecular signaling networks), based on a relational database to minimize the issues of heterogeneous data representation based on suitable semantics. Some of characteristics of CeTICSdb are: a) submission system that receives biological data (e.g., sequenced DNA) attached to its semantics (i.e., the context in which data were produced); b) automatic preprocessing of submitted data; c) creation of “projects” to analyze data and to execute computational simulations; d) multivariate methods to the integration, comparison and visualization of multiple datasets on same and different –omics technologies. To validate the platform, we integrated transcripts or protein expression profile and Metabolic Pathways to: i) estimate the metabolic activity between different conditions or treatments; ii) define and compare the functional activity for the metabolic pathways in each condition. Finally, our mid-term objective is to make the CeTICSdb platform available as a dry lab to the scientific community.

Poster 27

Increased Comparability Of Environmental Sequencing Efforts Of 16S rRNA Bacterial Community Profiles

Anwar, Muhammad Zohaib
Manohar, Vimitha
Henschel, Andreas

Advances and democratization of DNA sequencing technology have revolutionized the approach of studying microbial community profiles in all environments around the globe. In order to elucidate the environmental factors that drive bacterial community composition, it is desirable to collect large, comprehensive sets of samples from independent studies in various conditions. One attempt to coordinate this data deluge is the Earth Microbiome Project (Gilbert et al., 2010). Phylogeny-based beta-diversity comparison (such as UniFrac) can then be used to detect clusters of similar communities and possibly relate them to environmental attributes such as salinity. However, large scale community comparisons must overcome innate sample differences such as uneven sampling size, different OTU calling methods and inconsistent or lacking environmental annotation. We here propose a novel data collection and integration procedure that tackles these challenges and thus enables meta-analyses amongst and across environment categories to detect community-level differences.
Methodology

Sequence analysis. We composed a large meta-dataset from heterogeneous sources such as the SRA (Wheeler et al., 2008), QIIME-DB (Caporaso et al., 2010), a data collection provided by Chaffron et al., 2010 (mainly small samples but many independent studies) and some locally sampled data. In total, we collected 20,472 distinct 16S rRNA from 2,462 different studies and stored them in a relational database (MySQL) for fast retrieval. The sequence analysis includes: merging of paired end sequences (if required and possible), quality filtering, demultiplexing samples, consistent closed-reference OTU calling against GreenGenes (v 13.5), assignment of taxonomy for representative OTU sequences using RDP. The QIIME toolkit was used extensively.

Environment Ontology (EnvO) Annotation. The aim of Ontology annotation is to use controlled vocabulary for different types of environments to hierarchically categorize samples so to be able to relate clusters of communities to environmental determinants. We used text-mining (weighted Jaccard phrase similarity) to automatically annotate samples lacking MIMARKS annotation (SRA, Chaffron's and our own) with EnvO-terms based on sample description texts like isolation-source.

Adaptive Rarefaction. In order to compare a large set of samples with strongly varying sizes, we device a method that for (almost) each pairwise sample comparison subsample only to a size

necessary for that individual pair, rather than subsampling all samples to the size of the smallest sample (as is commonly done with tools that perform rarefaction on BIOM tables, such as QIIME/UniFrac). By keeping the subsample size as large as possible, we increase the pairwise UniFrac distance precision. As an example, we calculated the UniFrac based beta-diversity distance matrix for 4158 soil samples (min sample size: 85, average: 41000) taken from 68 different studies using adaptive rarefaction. Figure 1 shows that the diversity of the Arctic soil samples, as postulated in (Chu et al., 2010), but evidenced by a much larger collection of diverse, independent samples and studies.

Conclusion

The presented methodology works as a single platform for studying Microbial communities using advanced downstream analysis such as phylogenetic Beta-Diversity (UniFrac).Our data-collection, EnvO annotation and adaptive rarefaction facilitates the large scale meta-analysis of microbial communities, which in turn elucidates the environmental attributes that drive community composition.

Poster 28

The Abundance, Multiplicity and Distribution of Multireads in RNA-Seq Data Analysis

Zhao, Shanrong

Motivation: RNA-Seq is rapidly becoming the method of choice for transcriptional profiling experiments. Owing to the short read length and the presence of paralogs and homologous regions with a gene, RNA-Seq mappers typically report a substantial portion of multireads, i.e. reads that map equally well to multiple genomic locations. Although some earlier studies have estimated the number of multireads for a mammalian genome between 10 and 40%, little is done to quantify the abundance, multiplicity and distribution of multireads in detail. In this paper, we filled this gap and investigated the influential factors that determine the abundance and distribution of multireads by analyzing the RNA-Seq data from the Human Body Map 2.0 Project at Illumina.

Results: We demonstrated a gene model is the most influential factor that impacts the abundance, multiplicity and distribution of multireads. Ensembl reports a much higher percentage of multireads that RefGene and UCSC gene model for the same RNA-Seq dataset, while UCSC reports the highest average multiplicity for multireads in all samples. To ensure the majority of multireads to be reported when mapping reads, UCSC requires a higher multi-read cut-off compared to RefGene and Ensembl. The second important factor is read length. The shorter the read length, the more reads become multireads, and the more genomic locations a multiread tends to be mapped to. It has also been shown that the abundance, multiplicity and distribution of multireads are tissue dependent as well. The cumulative distribution of multireads can help a user to make an informed decision, and accordingly to choose a more appropriate cut-off to report multireads in alignment step
when analyzing RNA-Seq data. We also found there is poor linear relationship between the multiplicity and abundance of multireads. Our study complements nicely to those researches focusing on how to accurately allocate a multiread to its true genomic origin.

Poster 29

Improved Quality Score Generation for Erroneous Nucleotides in Simulated (Meta)Genomic Data

Johnson, Stephen
Trost, Brett
Kusalik, Anthony

Motivation and Objectives: In previous work, we presented BEAR (Better Emulation of Artificial Reads, available at https://github.com/sej917/BEAR), a collection of Perl and Python scripts that used machine learning techniques to emulate and/or automate characteristics (read length distributions, quality score profiles, error rates, abundance profiles) of a given sample of WGS reads from any sequencing platform. We demonstrated that BEAR was able to emulate those characteristics better than similar programs (MetaSim, Grinder, 454sim, SimSeq, GemSIM), but did not have the opportunity to demonstrate BEAR's ability to generate realistic quality scores for erroneous nucleotides. Here, we describe in detail how BEAR creates error-quality models for simulated data and examine how well the models fit to real sequencing data from different platforms.

Methods: The first stage of error-quality model generation involves the use of a modified version of DRISEE to cluster duplicate reads in a given FASTQ file. DRISEE classifies reads with identical prefixes of a certain length (default: 50bp, lower for short reads) as duplicates and clusters them. BEAR cross-references the clusters with the FASTQ file to determine the average quality scores of all substitution and indel errors for all nucleotides at all positions. BEAR then performs second-degree polynomial regression on the average quality scores as a function of their position within the reads to derive two error-quality models: one for indel errors, and another for substitution errors.

A methodology for evaluating BEAR has been described previously (Johnson et al: A better sequence-read simulator program for metagenomics. RECOMB-seq 2014). Here, a similar methodology is used to evaluate the success of BEAR's error-quality model. Three sets of sequencing data from different platforms are used: 689,365 reads from the E. coli DH10B genome (Personal Genome Machine with Ion 318 chip); 122,737 reads from from the L. rhamnosus ATCC 8530 genome (Roche 454 Genome Sequence FLX platform); and 400,000 reads from the P. claussenii ATCC BAA-344 genome (Illumina Genome Analyzer IIx). Simulated data is generated from each dataset and compared to the original. An R2 value is used to measure the accuracy of each error-quality model.

Results and Discussion: BEAR's error-quality models allow for realistic quality scores to be generated for all indel and substitution errors. The error-quality models are least accurate for the short Illumina reads (67bp) with the lowest R2 values of 0.32 and 0.59 for the substitution and

indel models, respectively, while the models for the 454 reads (1bp to 1138bp) are most accurate with R2 values of 0.82 and 0.93. Quality scores behave differently for indel and substitution errors for all datasets, demonstrating the need for quality scores to be modeled separately for different types of errors. We also observe that quality scores for erroneous bases can even vary between nucleotides at a given position within a read, especially in Illumina data, justifying the need for nucleotide-specific error models.

In summary, we demonstrate that BEAR, a run-specific, platform-agnostic sequencing simulator programs is able to accurately model the quality scores of erroneous nucleotides for a variety of sequencing platforms.

Poster 30

FARVATX: a Family-Based Rare Variant Association Test on X Chromosome

Choi, Sungkyoung; Seoul National University,
Park, Taesung; Seoul National University, Statistics
Won, Sungho; Chung-Ang University, Dept of Applied Statistics

Motivation: Genes on X chromosome contains useful information about human evolutionary history and it has been repeatedly addressed that they are often functionally related with human phenotypes. However genetic analyses of variants on X chromosome should consider the unique biological properties of X chromosome, which makes genetic analyses of variants on X chromosome have lagged behind those on autosomes - from the genome-wide association studies to sequence-based association studies.

Results: In this report, we propose a family-based association analysis with rare variants on X chromosome which takes account of dosage compensation. Both burden and variance component tests were proposed, and they were extended to the robust method. Extensive simulation studies were conducted to show the efficiency of the proposed method, and we performed genetic association analysis of rare variants on X chromosome with smoking dependency. The significant results indicate the importance of analysis with rare variants on X chromosome and the practical value of the proposed method.

Availability: The software for the proposed method is freely downloadable from http://bibs.snu.ac.kr/software/farvatx/.

Contact: won1@snu.ac.kr, tspark@stats.snu.ac.kr