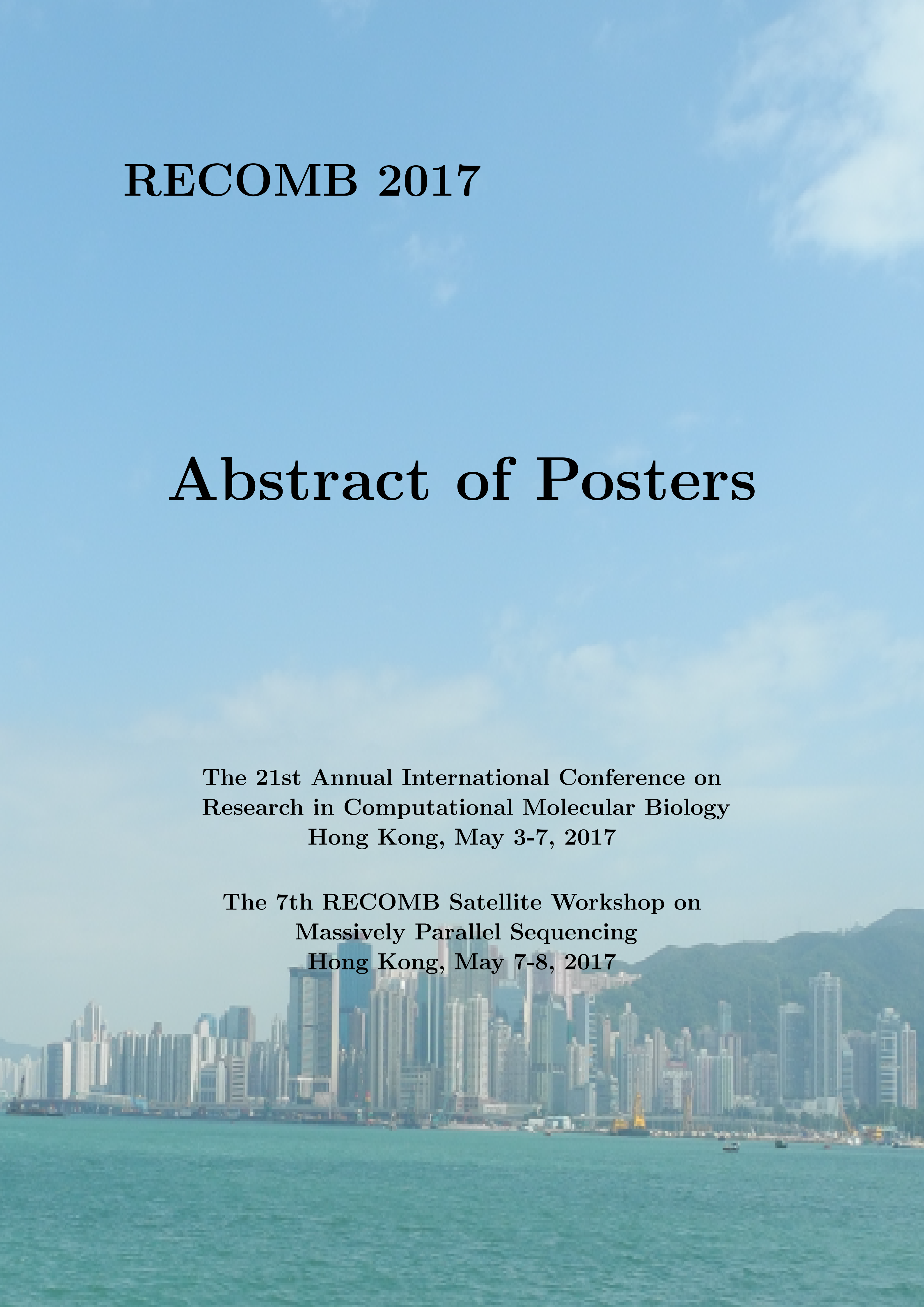

RECOMB 2017

Abstract of Posters

The 21st Annual International Conference on
Research in Computational Molecular Biology
Hong Kong, May 3-7, 2017

The 7th RECOMB Satellite Workshop on
Massively Parallel Sequencing
Hong Kong, May 7-8, 2017

P01

Is it feasible to only use tags in database search for PTM-invariant peptide identification? — A simulation-based study

Jiaan Dai[1], **Fengchao Yu**[2], **Ning Li**[2,3], **Weichuan Yu**[1,2,*]

[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
eeyu@ust.hk
[2]Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[3]Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong, China

Identifying peptides with unlimited post-translational modifications (PTMs) is a challenging task in mass spectrometry (MS) data analysis. Currently, neither *de novo* sequencing nor database search can provide satisfactory results. Restricted and unrestricted PTM identification methods work well with limited PTMs. However, their performance degrades greatly when the locations and types of PTMs keep increasing.

Recently, we proposed a method called PIPI to tackle the issue of peptide identification with unlimited PTMs. By separating "core" peptide (i.e. the peptide sequence without PTM annotations) identification from PTM characterization and by only using tags in "core" peptide candidate selection, PIPI avoids searching the huge space and allows unlimited PTMs.

However, PIPI still uses the whole spectrum-based comparison in the last step, when no more than twenty peptide candidates remain. While the performance of PIPI is encouraging, there is no clear understanding about how much sequence tags can do in peptide identification. In this work, we would like to further study the performance of tag-based methods in peptide identification through simulations. By making explicit assumptions, we generate simulated datasets to explore the performance under different settings. We obtain an upper bound as well as an estimate of the power of tag-based methods. This analysis should provide a deeper understanding of the feasibility and the limit of tag-based methods for peptide identification with unlimited PTMs.

P02

ECL 2.0: Exhaustively Identifying Cross-Linked Peptides with a Linear Computational Complexity

Fengchao Yu[1], **Ning Li**[1,2], **Weichuan Yu**[1,3,*]

[1]Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

[2]Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong, China

[3]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

eeyu@ust.hk

Chemical cross-linking coupled with mass spectrometry is a powerful tool to study protein-protein interactions and protein conformations. In order to identify cross-linked peptides from mass spectrometry data, we need to compare an experimental spectrum with theoretical spectra derived from different peptide pairs. However, most tools don't search all possible peptide pairs due to the long computational time. Consequently, significant percentages of peptide-spectrum matches (PSMs) are missed. The only remedy is to search all possible peptide pairs in the database. In our earlier work, we developed a tool named ECL to search all pairs of peptides exhaustively. However, its quadratic time complexity leads to a long running time when the database is large.

Here, we propose an advanced version of ECL, named ECL 2.0. It achieves linear time and space complexity by taking advantage of the additive property of a score function. ECL 2.0 can exhaustively search tens of thousands of spectra against a database containing thousands of proteins in a few hours. The software can be found at http://bioinformatics.ust.hk/ecl2.0.html.

P03

De novo Receptor-Based Design of new Integrin Beta 1 Inhibitors and in silico ADMET methods evaluation

João Herminio Martins Da Silva, Disraeli Vasconcelos, Lia Pinho and Beatriz Chaves

Integrins are cell adhesion receptors that transmit bidirectional signals across the plasma membrane. They are noncovalently linked heterodimeric molecules consisting of one and one subunit. Integrins play an important role in inflammation and cancer, due to their properties in cell proliferation processes. Three integrin inhibitors were selected to be docked with three receptors: 41, 51 and V1 and their ADMET properties were estimated. The aim of this study is to develop new integrin antagonists with pharmacological potential. The chosen ligands (BIO1211, BIO5192 and TCS2314) were built using Avogadro. Molecular Docking was done with Vina. Rachel was used for the design of the new compounds using ZINC database. ADMET predictions were performed using the programs ClogP/CMR, FAF-Drugs3, ProTox, and OSIRIS Property Explorer. The lowest calculated energy was -9.2 kcal/mol for 41 and BIO5192. Rachel generated more than 200 new BIO5192 derivatives. The scores ranged the first tenth ligands from 9.08 to 8.45. The results from ProTox showed low toxicity of the ligands. ClogP/CMR was used to predict the human oral bioavailability based on Lipinski Rules of 5. The molecules presented an acceptable number of violations. OSIRIS Property Explorer indicated that none of the compounds presented mutagenic or reproductive effective toxicity risks. These results show that BIO5192 is promising as a lead compound for the development of new integrin inhibitors. This is a filtering approach to improve drug discovery and development processes.

P04

Integrated somatic subtypes of localized prostate cancer with prognostic implications

Natalie Fox, Emilie Lalonde, Julie Livingstone, Julia Hopkins, Yu-Jia Shiah, Vincent Huang, Takafumi Yamaguchi, Veronica Sabelnykova, Lawrence Heisler, Michael Fraser, Theodorus van der Kwast, Robert Bristow and Paul Boutros

Treatment for localized prostate cancer is currently based on risk-stratification using Gleason Score, pre-treatment PSA and tumour size. About 40% of new diagnoses are classified as intermediate-risk by this scheme, and it is estimated that over half of these are either over- or under- treated. To help improve existing clinical classifications, genomic biomarkers have been developed, but none of these have reached routine clinical use, in part due to insufficient accuracy. We hypothesize that localized prostate cancer harbours a small set of genomic subtypes, and that distinct prognostic biomarkers will be required for each subtype. Thus the lack of clearly defined subtypes has confounded existing biomarker validations. We therefore developed a new way to create integrated subtypes, merging coding and non-coding point mutations, genomic rearrangements, copy number aberrations and methylation status. The resulting five integrated subtypes reflect major driver events in prostate cancer, including NKX3-1 deletion and TMPRSS2-ERG fusions. Further, we show accuracy of existing prognostic biomarkers are biased by these subtypes. Taken together, these data demonstrate the existence of genomic subtypes of prostate cancer whose mutational characteristics manifest across multiple genetic and epigenetic somatic alterations. New subtype-aware biomarkers are needed to improve risk stratification for patients with localized prostate cancer.

P05

Haplotype-based eQTL Mapping Increases Power to Identify eGenes

Robert Brown[1], Eleazar Eskin[1,2,3,4], Bogdan Pasaniuc[1,2,3]

[1]Bioinformatics IDP, University of California, Los Angeles, California, USA
[2]Department of Pathology and Laboratory Medicine, Geffen School of Medicine, University of California, Los Angeles, California, USA
[3]Computer Science Department, University of California, Los Angeles, California, USA
[4]Department of Human Genetics, Geffen School of Medicine, University of California, Los Angeles, California, USA

Expression quantitative trait loci (eQTLs), variations in the genome that impact gene expression, are identified through eQTL studies that test for a relationship between single nucleotide polymorphisms (SNPs) and gene expression levels. These studies typically assume an underlying additive model. However, with increasing evidence for allelic heterogeneity, more complex models will be needed to fully understand the genetic basis of many phenotypes. Here we propose using combinations of haplotypes from 10 kb regions instead of SNPs as predictors for gene expression. Simulations show that when haplotypes, rather than SNPs, are assigned non-zero effect sizes, our approach has increased power compared to the marginal SNP approach. When we apply our approach to the GEUVADIS gene expression data, we find 101 more eGenes than the marginal SNP approach. In addition to the increase in power, the sets of eGenes found by both our method and the standard SNP-based method only have a 77% overlap. This indicates our approach is able to detect signal from a large number of genes where the effect of the underlying regulatory mechanisms is not captured using marginal SNP effects.

P07

Profiling adaptive immune repertoires across multiple human tissues by RNA Sequencing

Serghei Mangul, Igor Mandric, Alex Zelikovsky and Eleazar Eskin

Assay-based approaches provide a detailed view of the adaptive immune system by profiling T and B cell receptor repertoires. However, these methods come at a high cost and lack the scale of standard RNA sequencing (RNA-seq). Here we report the development of ImReP, a novel computational method for rapid and accurate profiling of the adaptive immune repertoire from regular RNA-Seq data. We applied it to 8,555 samples across 544 individuals from 53 tissues from the Genotype-Tissue Expression (GTEx v6) project. ImReP is able to efficiently extract TCR- and BCR- derived reads from the RNA-Seq data and accurately assemble the complementarity determining regions 3 (CDR3s), the most variable regions of B- and T-cell receptors determining their antigen specificity. Using ImReP, we have created the systematic atlas of immunological sequences for B- and T-cell repertoires across a broad range of tissue types, most of which have not been studied for B and T cell receptor repertoires. We have also examined the compositional similarities of clonal populations between the GTEx tissues to track the flow of T- and B- clonotypes across immune-related tissues, including secondary lymphoid organs and organs encompassing mucosal, exocrine, and endocrine sites. The atlas of T- and B-cell receptor receptors, freely available at https://sergheimangul.wordpress.com/atlas-immune-repertoires/, is the largest collection of CDR3 sequences and tissue types. We anticipate this recourse will enhance future studies in areas such as immunology and advance development of therapies for human diseases. ImReP is freely available at https://sergheimangul.wordpress.com/imrep/.

P08

Group Sparse Optimization: An Integrative OMICs Method to Predict Master Transcription Factors for Cell Fate Conversion

Jing Qin, Yaohua Hu, Jen-Chih Yao, Yiming Qin, Ka Hou Chu and Junwen Wang

Conversion of cell fates by overexpression of defined factors is a powerful approach in regenerative medicine. However, identifying key factors for cell fate conversion needs laborious experimental efforts and many conversions are not achieved yet. Even in many published cases where the targeted cell phenotype and marker expression are shown, cell fate conversions are found to be incomplete and expressions of other important genes could not be silenced or activated properly. Thus, identify the master transcription factors required for complete cell fate conversion is crucial to moving this technology closer to clinical applications. Integrative OMICs approaches have shown to uncover the mystery of gene regulatory mechanisms in cell fate conversion processes. A systematic analysis on these OMICs data makes it possible to predict the master transcription factors for cell fate conversion. In this study, we introduce a novel computational method to predict master transcription factors based on group sparse optimization. We compared it with several state-of-the-art prediction methods and demonstrated its best performance. This method will benefit researchers in regenerative medicine by quickly identifying the key master regulators, enhancing the successful conversion rate and reducing the experimental cost.

P09

Leveraging allele-specific expression to improve fine-mapping for eQTL studies

Jennifer Zou[1], Farhad Hormozdiari[2], Jason Ernst[1,3], Jae Hoon Sul[4,*], Eleazar Eskin[1,5,*]

[1]Computer Science Department, University of California, Los Angeles, California 90095, USA

jaehoonsul@mednet.ucla.edu, eeskin@cs.ucla.edu

[2]Genetic Epidemiology and Statistical Genetics Program, Harvard University, Massachusetts 01451, USA

[3]Department of Biological Chemistry, University of California, Los Angeles, California 90095, USA

[4]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, California, USA

[5]Department of Human Genetics, University of California, Los Angeles, California 90095, USA

Many disease risk loci identified in genome-wide association studies are present in non-coding regions of the genome. It is hypothesized that these variants affect complex traits by acting as expression quantitative trait loci (eQTLs) that influence expression of a nearby gene. This indicates that many causal variants for complex traits are likely to be causal variants for gene expression. Hence, identifying causal variants for gene expression is important for elucidating the genetic basis of not only gene expression but also complex traits. However, detecting those variants is challenging due to complex genetic correlation among variants known as linkage disequilibrium (LD) and the presence of multiple causal variants within a locus. Although several fine-mapping approaches have been developed to overcome these challenges, they may not produce very accurate results when many causal variants are in high LD with non-causal variants. In eQTL studies, there is an additional source of information for fine-mapping called allele-specific expression (ASE) that measures imbalance in gene expression due to different alleles. In this work, we develop a novel statistical method that leverages both ASE and eQTL information to detect causal variants that regulate gene expression. We illustrate through simulations and application to the Genotype-Tissue Expression (GTEx) dataset that our method identifies the true causal variants more accurately than an approach that uses only eQTL information.

P10

CASC: Classification Analysis of Single Cell Sequencing Data

Luca Alessandrì and Raffaele A Calogero

Dept. of Molecular Biotechnology and Health Sciences, University of Torino, Italy

Genome-wide single-cell measurements such as transcriptome sequencing enable the characterization of cellular composition as well as functional variation in homogenic/heterogenic cell populations. An important step in the single-cell transcriptome analysis is to group cells that belong to the same sub-type based on gene expression patterns. Critical issues in cell clustering are (i) cluster stability and (ii) feature selection, i.e. the identification genes playing the major role in cluster formation. To address the above issues, we have developed CASC, a tool implemented in a docker container, that uses as core application to detect cell clusters the "kernel based similarity learning"and allows: (i) identification of the optimal number of clusters for cell partitioning using "silhouette method". (ii) The evaluation of clusters stability, measuring the permanence of a cell in a cluster upon random removal of subsets of cells. (iii) Feature selection via "nearest shrunken centroid method", applied to the gene Index Of Dispersion. CASC was tested on previously published data sets, indicating that the most critical point on cluster stability is feature selection.

P11

SINCERITIES: Inferring gene regulatory net-works from time-stamped single cell transcrip-tional expression profiles

Nan Papili Gao, S.M Minhaz Ud-Dean and Rudiyanto Gunawan

Recent advances in single cell transcriptional profiling open up a new avenue in studying the functional role of cell-to-cell variability in physiological processes such as stem cell differentiation. The analysis of single cell expression data poses new challenges due to the distributive nature of such data and the stochastic and bursty dynamics of the gene transcription process. In particular, the reconstruction of gene regulatory networks (GRNs) using single cell transcriptional profiles is extremely challenging. Several GRN inference algorithms have been developed for single cell data, but to the best of our knowledge, none of the existing algorithms directly use the time point information of the cells in inferring the causal gene-gene relationships.

In this work, we developed a novel algorithm called SINCERITIES (SINgle CEll Regularized Inference using TIme-stamped Expression profiles), for the inference of GRNs from single cell transcriptional expression data. In particular, we focused on time-stamped cross-sectional expression data, a common type of dataset generated from transcriptional profiling of single cells collected at multiple time points after cell stimulation. SINCERITIES recovers the regulatory (causal) relationships among genes by employing regularized linear regression, particularly ridge regression, using temporal changes in the distributions of gene expressions. Meanwhile, the modes of the gene regulations (activation and repression) come from partial correlation analyses between pairs of genes. We demonstrated the efficacy of SINCERITIES in inferring GRNs using in silico time-stamped single cell expression data and single transcriptional profiles of THP-1 monocytic human leukemia cell differentiation. The case studies showed that SINCERITIES could provide accurate GRN predictions, significantly better than other GRN inference algorithms such as TSNI, GENIE3 and JUMP3. Moreover, SINCERITIES has a low computational complexity and is amenable to problems of extremely large dimensionality.

P12

The Roles of Signal Pathways mediated by AF-VHD-related microRNA combinations in Development from VHD to AF-VHD

Feng Yang[1,2], Wei Zeng[1,2], Ke Liu[3], Guangbin Wang[1,2], Zhengwen Li[1,2], Keli Huang[3], Nini Rao[1,2,4,*]

[1]Key Laboratory for NeuroInformation of Ministry of Education, University of Electronic Science and Technology of China, Chengdu, China
raonn@uestc.edu.cn
[2]Center for Information in BioMedicine, University of Electronic Science and Technology of China, Chengdu, China
[3]Subsidiary hospital, University of Electronic Science and Technology of China, Chengdu, China
[4]Guangdong Electronic Information Engineering Research Institute, University of Electronic Science and Technology of China, Dongguan, Guangdong

Valvular heart disease with atrial fibrillation (AF-VHD) is a complex comorbidity, but we know very little about the effects of various genomic alterations on its pathogenesis. This paper designed a system approach to rationalize this complex disease at the levels of microRNAs and pathways. AF-VHD-related microRNAs are screened based on the difference between the expression of AF-VHD and VHD tissues and the differences between the co-expressions within AF-VHD and VHD tissues. The qRT-PCR experiments are used to validate the partial microRNAs. The combinational patterns among AF-VHD-related microRNAs are extracted according to their co-expression network in AF-VHD group. The mediation relationships between the microRNA combinations and pathways are constructed through the target genes. The roles of the pathways mediated by AF-VHD-related microRNA combinations in development from VHD to AF-VHD are analyzed and annotated. 32 AF-VHD-related microRNAs are obtained, some of which are new findings. Three representative microRNAs all passed the verification. This illustrates the screened microRNAs have high credibility. The four types of mediation relationships are identified. The roles of a typical pathway in each class are inferred and illustrated. This paper proposes a system computational approach to study a complex disease at the molecular level, finds some new biological features of AF-VHD, and also provides some important insights into the understanding of AF-VHD molecular mechanism.

P13

Detection of long repeat expansions from PCR-free whole-genome sequence data

Egor Dolzhenko, Joke J.F.A. van Vugt, Subramanian S. Ajay, Ryan Taft, David R. Bentley, Jan H. Veldink and Michael A. Eberle

Short tandem repeats (STRs) are particularly difficult for most small variant callers, even those that perform well on SNPs and indels, and special variant callers (e.g. HipSTR) are needed to call these problematic variants. Even these STR-specific variant callers can only accurately identify repeats when they are shorter than the reads. This limitation is particularly problematic because many genetic diseases such as amyotrophic lateral sclerosis (ALS) and Fragile X syndrome are associated with repeats that are 100's to 1000's of bp in length. To fill this gap we have developed a tool called ExpansionHunter that, when applied to PCR-Free whole genome sequence (WGS) data, can accurately quantify the sizes of both short repeats and repeats that significantly exceed the read length. To demonstrate that ExpansionHunter accurately identifies large pathogenic repeat expansions, we analyzed WGS data from 3,001 ALS patients who have been tested for the presence of the repeat expansion in the C9orf72 gene with repeat-primed PCR (RP-PCR). Assessed against this dataset, ExpansionHunter correctly classified 99.5% (212/213) of the expanded samples (¿30 GGCCCC repeats) and all (2,788/2,788) of the wild type samples. Motivated by these results, we are sequencing all of the samples with repeat expansions available from Coriell to demonstrate that ExpansionHunter can also identify the repeat expansions associated with Fragile X Syndrome, Myotonic Dystrophy, Huntington's Disease and Friedreich's Ataxia. We will present the calls from ExpansionHunter on these samples and outline a process for discovering novel repeat expansions in global surveys.

P14

Computer Program for Verification of Phylogenetic Networks

Andreas D.M. Gunawan[1], **Bingxin Lu**[2], **Hon Wai Leong**[2], **Louxin Zhang**[1,*]

[1]Department of Mathematics, National University of Singapore, Singapore
matzlx@nus.edu.sg
[2] Department of Computer Science, National University of Singapore, Singapore

Genetic material is transferred from a species to another that had no descendant relationship more frequently than commonly thought, particularly in the bacteria kingdom. Extant genomes are thus considered as a fusion product of both reproductive and non-reproductive genetic transfers in comparative genomics. This has motivated researchers to adopt phylogenetic networks to model genome evolution. On the other hand, a gene's evolution is usually tree-like and has been studied for over half a century. Accordingly, the relations between phylogenetic trees and networks are the basis for the reconstruction and verification of phylogenetic networks. Therefore, one important problem in verifying a network model is determining whether existing phylogenetic trees are displayed in a phylogenetic network or not. This problem is formally called the tree containment problem (TCP). Another problem related to the TCP is the cluster containment problem (CCP) that asks whether a subset of leaves is a soft cluster in a phylogenetic network. The TCP and CCP are both NP-complete even for binary phylogenetic networks.

We developed fast algorithms for solving the two problems for arbitrary phylogenetic networks using the so-called reticulation-visible property of phylogenetic networks. They run in exponential-time in the worst case, but fast enough in practice. Additionally, the CCP algorithm is further extended into a method for computing the soft Robinson-Foulds distance between the phylogenetic networks, which is defined as the number of soft clusters in one network but not in the other. Their implementations in C are online available.

P15

Diversity of mutational loads at CTCF binding sites predicts foci of dysregulated expression

Vera Kaiser and Colin Semple

CTCF is a multi-functional DNA binding protein that plays critical roles in nuclear architecture, acting to create chromatin loops linking enhancers to target promoters, and as a boundary insulator for regulatory domains. In a previous study, we demonstrated the hyper-mutability of constitutively accessible CTCF binding sites across tumour genomes, relative to matched control sites (Kaiser et al, 2016, PLOS Genet). The mutational load accumulating at these sites varied strikingly over sites within the motif, such that CTCF binding was predicted to be compromised at mutated sites. For example, a single site within the CTCF binding motif, suffering highly elevated mutation rates, accounted for much of the excess mutation seen over the motif. Surprisingly, we found that these unusual patterns of variation were explicable by selectively neutral biases in mutation.

Using a more comprehensive collection of standardized SNV calls, we have now examined TFBS mutation loads across 18 tumour types. The characteristic site-specific mutation loads within the CTCF motif seen previously are observed again, for a number of tumour types individually. However there is substantial diversity among other cancers. Melanoma spectacularly diverges from other cancer types, showing a distinct mutational spectrum as expected, but also a significantly elevated peak of mutation at two well conserved positions within the motif. We argue that the variety of mutational profiles at CTCF active sites across cancers is, again, likely to be the product of variable mutational biases. The predicted consequences of these different CTCF site mutation loads, i.e. the effects on CTCF binding affinity, also vary widely between tumour types. From mildly deleterious effects comparable to those of common human polymorphisms, to strongly deleterious profiles predicted to abolish binding at many sites across the genome. These differences in profiles therefore predict dysregulated expression at particular foci in particular tumour types.

P16

Comparative transcriptome analysis of resistant and susceptible Korea rice genotype in response to bakanae disease

Dong-Jun Lee, Hyeon-So Gi, Dowan Kim, Jae-Hyeon Oh, Chang-Kug Kim and Tae-Ho Lee

The causal agent of bakanae is the most significant seed-born disease of rice. Molecular mechanisms regulating defense responses of rice toward this fungus are not yet fully known. To identify transcriptional mechanisms underpinning rice resistance. an RNA-seq comparative transcriptome profiling was conducted seedlings of selected Korea rice genotype(Junam and Nampeong) at post germination(five and ten days) and plant region(shoot and root)

P17

Learning hierarchical motif representations for protein analysis, search and design

Thrasyvoulos Karydis, Aditya Khosla, Manolis Kellis and Joseph M. Jacobson

Current proteomic tools, hinge on maximal sequence similarity and have segregated proteins into more than 16,000 families. As a result, a significant number of protein sequences do not fall under any of these families and their function has yet to be characterized. Here we introduce CoMET - Convolutional Motif Embeddings Tool, a deep learning framework enabling the evolutionary analysis of large protein datasets, based on a hierarchical decomposition of proteins into a set of motif embeddings. At the core of CoMET, lies a Deep Convolutional Auto-Encoder, trained to learn a non-linear combination of basis set of motifs, by minimizing the amino acid sequence reconstruction error. CoMET is successfully trained to extract all known motifs across Transcription Factors and CRISPR Associated proteins, without requiring any prior knowledge about the nature of the motifs or their distribution (unsupervised learning). We demonstrate that the motif embeddings can model efficiently known inter- and intra- protein family relationships, and allow to search for evolutionary distant protein homologs. Furthermore, we form novel protein families, by taking into account a hierarchical conserved motif phylogeny, instead of a single ultra-conserved signature profile per family. Lastly, we investigate the generative ability of CoMET, when trained to fit experimental protein function data, towards the in-silico directed evolution of proteins for novel functions. As a proof of principle, we train an accurate predictive model for the recognition code of the Type II restriction enzymes and invert the trained model to generated restriction enzymes with de-novo binding specificities.

P18

VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data

Jie Ren, Fengzhu Sun, Nathan A Ahlgren, Jed A Fuhrman and Yang Young Lu

Identifying viral sequences in mixed metagenomes containing both viral and host contigs is a critical first step in analyzing the viral component of samples. Current tools for distinguishing prokaryotic virus and host contigs primarily use gene-based homology approaches. Such approaches can significantly limit results especially for short contigs that have few predicted proteins or lack proteins with homology to previously known viruses. We have developed VirFinder, the first k-mer frequency based, machine learning method for virus contig identification that entirely avoids homology searches. VirFinder instead identifies viral sequences based on our empirical observation that viruses and hosts have discernibly different k-mer signatures. VirFinders performance in correctly identifying viral sequences was tested by training its machine learning model on sequences from host and viral genomes sequenced before 1/1/2014 and evaluating on sequences obtained after 1/1/2014. VirFinder had significantly better rates of identifying true viral contigs (true positive rates, TPRs) than VirSorter, the current state-of-the-art gene-based virus classification tool, when evaluated with either contigs subsampled from complete genomes or assembled from a simulated human gut metagenome. For example, for simulated assembled contigs VirFinder had 156-, 3.2-, and 2.6-fold higher TPRs than VirSorter for 0.5-1 kb, 1-3 kb and 3 kb contigs, respectively, at the same false positive rates as VirSorter (0%, 0.03%, and 0.53%, respectively). Thus, VirFinder works considerably better for small contigs than VirSorter. VirFinder furthermore identified several recently sequenced virus genomes (after 1/1/2014) that VirSorter could not and that have no homology to previously sequenced viruses, demonstrating VirFinder's potential advantage in identifying novel viral sequences. Application of VirFinder to a set of human gut metagenomes from healthy and liver cirrhosis patients reveals higher viral diversity in healthy individuals than cirrhosis patients. We also identified contig bins containing crAssphage-like contigs with higher abundance in healthy patients and a putative Veillonella genus prophage associated with cirrhosis patients. This innovative k-mer based tool complements gene-based approaches and will significantly improve prokaryotic viral sequence identification, especially for metagenomic-based studies of viral ecology.

P19

Applying meta-analysis to Genotype-Tissue Expression data from multiple tissues to identify eQTLs and increase the number of eGenes

Dat Duong[1,*], Lisa Gai[1], Sagi Snir[2,3], Eun Yong Kang[1], Buhm Han[4,5], Jae Hoon Sul[6], Eleazar Eskin[1,7,*]

[1]Department of Computer Science, University of California, Los Angeles, 90095, USA
datdb@cs.ucla.edu, eeskin@cs.ucla.edu
[2] Institute of Evolution, University of Haifa, Haifa, 3498838, Israel [3]Department of Evolutionary and Environmental Biology, University of Haifa, 3498838, Israel
[4]Department of Convergence Medicine, University of Ulsan College of Medicine, Ulsan, 44610, Republic of Korea
[5] Asan Institute for Life Sciences, Asan Medical Center, Seoul, 05505, Republic of Korea
[6]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, 90095, USA
[7]Department of Human Genetics, University of California, Los Angeles, 90095, USA

Gene expression data from many tissue like the Gene Tissue Expression (GTEx) data have helped interpreting GWAS results. In particular, these datasets can be used to find expression quantitative loci (eQTLs) and eGenes. eQTLs are SNPs associated to the expression of some genes, and eGenes are genes whose expressions are associated with the genotype of some SNPs in a specified tissue. One can cross-check significant SNPs and candidate genes in GWAS with the list of eQTLs and eGenes to identify the types of tissue a disease is affecting. Moreover, if a significant SNP or a candidate gene is an eQTL or eGene in some tissues then there is great evidence to further study this SNP or gene. Unfortunately, gene expression data often does not contain enough sample for some tissues. For this reason, there has been many meta-analyses designed particularly to combine gene expression data across many tissues to increase the power of finding eQTLs and eGenes. However, none of these existing methods is scalable to datasets with many tissues like the GTEx data. Second, these methods ignore a biological insight that the same SNP can be associated to the same gene in similar tissues. We introduce a meta-analysis method that addresses the problems in these existing methods. We focus on the problem of finding eGenes in the GTEx data, and show that our method is better than other types of meta-analyses.

P20

Molecular docking and dynamics simulation studies of Antitubercular compounds with Enoyl-ACP reductase enzyme of MDR-Tuberculosis

Ruban Durairaj D*, Murugesh Easwaran, Shanmughavel P

Department of Bioinformatics, Bharathiar University, Coimbatore, India
ruban.bioinfo@gmail.com, murugeshphdsch@gmail.com, shanmughavel@buc.edu.in

Tuberculosis is widely spreading disease in almost every part of the world next to HIV. The infection is getting more virulent due to its multidrug resistant factor. Hence we need to find alternate drugs and drug targets to compete with multidrug-resistant tuberculosis. The Enoyl-ACP reductase enzyme is one such a promising drug target, which plays an important role in the synthesis of mycolic acids that contributes to the formation of the Mycobacterial cell wall. The cell wall is an essential component for any bacteria in order to survive from the immune response generated by the host. Thus, by blocking Enoyl-ACP reductase enzyme it leads to the inhibition of mycolic acid synthesis, which will eventually result in the inhibition of mycobacterial cell wall construction and ultimately leads to the mycobacterial cell death. Hence 27 reported antitubercular compounds, were taken into molecular docking studies in order to identify their binding activities with Enoyl-ACP reductase enzyme. Among 27 compounds, Amikacin is having a more binding activity and the docked complex is further taken for molecular dynamics simulation studies. The molecular docking and dynamics studies revealed us that, Amikacin is a potent inhibitor of Enoyl-ACP reductase enzyme which can be used as a lead molecule in the drug designing process.

P21

Chromatin modeling reveals segregation of chromosome characteristics and fuctions in 3D space

Luming Meng[1], Wenjun Xie[1], Sirui Liu[1], Ling Zhang[1], Yiqin Gao[1,2]

[1]Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering
[2]Biodynamic Optical Imaging Center, Peking University, Beijing

Hi-C contact matrix provides plethora information for chromatin structure, which is difficult to be fully interpreted. In our work, we translate the contact information into relative distance between chromatin segments and then reconstruct 3D structures. The constructed structure is validated by reproduction of Hi-C contact matrix and consistence with distances measured by fluorescence in situ hybridization (FISH) experiments. Interestingly, our model demonstrates that the chromatin segments with different genomic features are remarkably segregated in 3D space. More information can be explored from the modeled structures.

P22

Repeat Elements Profile Across Different Tissues in GTEx Samples

Harry Taegyun Yang, Serghei Mangul, Noah Zaitlen, Sagiv Shifman and Eleazar Eskin

Next Generation RNA Sequencing technology enabled researchers to investigate transcriptomics, which is directly correlated to protein expressions, as well as other activities within cells. However, some of the transcribed RNAs are not translated into proteins but rather reincorporated into the genome through a process called retrotranscription, and those sequences that are reincorporated into the genome through this process is called retrotransposons. Retrotransposons occupy a significant portion of the human genome and are believed to have significant roles in certain diseases, such as Retts disease or Schizophrenia. We investigated the differential expression of non-long terminal repeat retrotransposons: short and long interspersed elements, denoted as SINE and LINE, respectively, as well as Alu elements. The differential expression of non-LTR retrotransposons is investigated on 3000 RNA-Seq samples across 32 tissues types in Genotype-Tissue Expression (GTEx) dataset. The repeat profile was extracted by mapping reads to the reference of repeat elements on human genome. On average, 7.47% of total RNA-Seq reads are mapped to repeat elements. Of the reads mapped to repeat elements, 32% of repeat element reads are mapped to LINE, while 41% of repeat element reads are mapped to SINE. Of the tissue types, testes showed significantly higher expression of SINE-VNTR-Alu (SVA) subclass F compared with the mean expression of SVA-F throughout the samples (p=2.46×10^{-33}), suggesting that there are important biological mechanism that up-regulates the SVA element expression. Furthermore, high co-expression of Alu elements and LINE L1 elements is detected across tissues (R^2=0.7615), signifying that Alu expression may depend on the expression of L1 elements.

P23

Structural and evolutionary analysis of sweetpotato chloroplast genomes

Ung-Han Yoon[1], JaeCheol Jeong[2], Jang-Ho Hahn[1], Jung-Wook Yang[3], Tae-Ho Kim[1], Hyeong-Un Lee[3], Young-Ju Seol[1], Sang-Sik Nam[3], Sang-Soo Kwak[4], Tae-Ho Lee[1]

[1]Genomics Division, National Institute of Agricultural Sciences, Jeonju 54875, Korea
[2]Natural Product Research Center, KRIBB, 181 Ipsin-gil, Jeongeup 56212, Korea
[3]Bioenergy Crop Research Institute, National Institute of Crop Science, Muan-gun 58545, Korea
[4]Plant System Engineering research Center, KRIBB, Daejeon 34141, Korea

Sweetpotato (*Ipomoea batatas*) is the top seven most important food crops in the world. Recently, sweetpotato is drawing interest of people as a healthy food because it has higher dietary fiber, vitamins, carotenoids and overall nutrition value. Especially, sweetpotato genome sequence has not been reported, and scientific interest on it as a valuable material for researches on *Convolvulaceae* crop evolution and investigation of useful genes is increasing. We performed whole genome sequencing of a high yield sweetpotato variety, Jeonmi, using next-generation sequencing (NGS). The genome size of hexaploid (2n=6x=90) sweetpotato was estimated to be about 3 Mb through K-mer analysis. First, we analyzed sweetpotato chloroplast genome analysis for *Convolvulaceae* crop evolution study. Among the whole genome NGS sequence data of sweetpotato, 453,000,000 bp data was assembled to one contig corresponding to chloroplast genome. The average depth of the assembled sequence was 196X, and the revealed chloroplast genome size was 161,440 bp. The chloroplast contained 113 genes which were classified into 79 coding genes, 30 tRNA genes, and 4 rRNA genes. We constructed a sweetpotato chloroplast gene map using OGDRaw program with our data. For the evolutionary analysis, phylogenetic analysis was conducted with 78 protein-coding sequences of 24 species chloroplast genomes by MEGA6.

P25

New MCMC methods for pseudo-time estimation using Gaussian processes

Magdalena Strauss, Lorenz Wernisch and John Reid

Single-cell RNA-seq provides gene expression levels of large numbers of genes for individual cells, but only a single measurement per cell. As cells progress through changes at different time scales, it is possible to obtain a form of time series data even from these cross-sectional data by means of pseudo-temporal ordering.

A number of previous approaches have provided point estimates of the order, while more recently, Gaussian process latent variable models have been applied to understand the uncertainty associated with the pseudo-temporal ordering. A good understanding of this uncertainty is crucial both because of measurement noise and the inherent stochasticity of cell development.

We present a new type of Gaussian process latent variable model for pseudo-temporal ordering, which uses MCMC methods on the possible orderings rather than on the high-dimensional space of possible pseudo-times. This allows a relatively comprehensive sampling of the pseudo-time orders, even without prior dimensionality reduction, which usually means a loss of information.

We sample not only the order, but also the parameters of the Gaussian process and present a trade-off between stability of the estimation and capturing the full extent of the uncertainty of both the parameters and the ordering of the cells. In an application to single-cell data, we obtain distributions of both pseudo-temporal orders and parameters, which allow us to understand better the time development of RNA expression levels and its uncertainty.

P26

The Potential Inhibitors in Chinese Traditional Medicine for Bcr-AblT315I Mutation of Chronic Myelogenous Leukemia

Yali Xiao[1], Ssu-Min Fang[1], Yun-Hsin Tsou[1], Cheng-Fang Tsai[1], Pei-Chun Chang[1]

[1]Department of Bioinformatics and Medical Engineering, Asia University, Taichung 41354, Taiwan

acetylcoa869@yahoo.com.tw, cck68522@gmail.com, cindy2524087@gmail.com, tsaicf@asia.edu.tw, peichun.chang@gmail.com

Recently, anticancer drug screening from the compounds of Chinese Traditional Medicine (TCM) has become a tendency in drug discovery. In this study, we focused on chronic myelogenous leukemia (CML) to mine the anticancer drug from herbs of Chinese Traditional Medicine. The gene Bcr-Abl in CML lost its regulation function of the tyrosine kinase that causes cells grow up continuously and inhibiting cells be withered. Currently, CML is treated by inhibiting the activity of Bcr-Abl tyrosine kinase that inhibits cell proliferation and induces apoptosis. Unfortunately, due to the variation of this cancer gene, the drugs such as imatinib, dasatinib, nilotinib, and bosutinib all have resistance effects or deadly side effect for Bcr-AblT315I mutation. To overcome these problems, we adopted virtual screening and QSAR modeling to discover the potential drug from TCM compounds. The results show that salvianolic acid C, baicalin, 1,4-dicaffeoylquinic acid, and dihydroisotanshinone I may have the potential for CML treatment with reducing side effects. Salvianolic acid C is present in Danshen (丹參). Baicalin was found in Huang Qin (黃芩), Chuan Huang (川黃芩), Da Che Qian (大車前), and Baihua Dan (白花丹參). 1,4-Dicaffeoylquinic acid is present in Cang Er (蒼耳). Dihydroisotanshinone I can be extracted from Bai Huad Dan (白花丹參).

P27

Exploring the landscape of regulatory elements on non-coding regions of cassava genome via motif-based screening

Chalida Rangsiwutisak[1], **Treenut Saithong**[2,3], **Saowalak Kalapanulak**[2,3,*]

[1]Bioinformatics and Systems Biology Program, School of Bioresources and Technology, and School of Information Technology, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok, 10150, Thailand

[2]Systems Biology and Bioinformatics Research Laboratory, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok, 10150, Thailand

saowalak.kal@kmutt.ac.th

[3]Bioinformatics and Systems Biology Program, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok, 10150, Thailand

Discovering the catalog of all regulatory elements on genome landscape is vital for a complete understanding of plant biology. Similar as all multicellular organisms, plants are composed of a diverse type of cells and tissue such as leaf, and root with widely divergent phenotypes and specific biological functions. Underlining the identical genomes, the various cellular phenotypes are possible because of differential gene regulation which is in turn governed by regulatory elements. Thus, regulatory element activity can be viewed as a genome-based signal driving phenotypic changes under external and internal stimuli. Herein, genome wide non-coding regions of cassava, the highlighted plant for food and energy security in the 21st century, were investigated for putative regulatory elements identification using motif-based screening. The upstream sequences were retrieved from cassava genome in the range up to 2,000 bps from translation start site without the overlapping coding sequences with previous gene. Nearly 100% of all promoters on cassava genome were scanned for the occurrences of a given motif described by a position-specific frequency matrix from PlantPAN, a database gathering transcription factor binding sites (TFBSs) and their corresponding transcription factors from 53 plant species. Finally, 1,039 TFBSs were proposed for 32,824 cassava promoters. The TFBSs frequency distribution is in the range of 1-245 TFBSs per promoter. From the top ten of the most frequently TFBSs occurring on cassava promoters are common motifs namely GATA box, E-box, TATA box and other protein-binding motifs corresponding to GATA, Dof, ZF-HD, Myb/SANT, bHLH, AT-Hook and Homeodomain transcription factor family. These results will be useful for proposing regulatory elements found in non-coding DNA, and further hypothesizing a transcriptional regulation of cassava plant.

P28

A new differential expression analysis method for single-cell RNA-Seq data

Zhun Miao[1], **Xuegong Zhang**[1,2,*]

[1]MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China
zhangxg@tsinghua.edu.cn
[2]School of Life Sciences, Tsinghua University, Beijing 100084, China

Single-cell RNA sequencing (scRNA-seq) is a promising technology emerged in recent years that enables high resolution detection of heterogeneities between individual cells. One important application of scRNA-seq is to detect differentially expressed genes (DEG) between different types or subtypes of cells or between cells of different conditions. Because of the special characteristic of scRNA-seq data, new methods are required for the detection of DEG.

We developed a new differential expression (DE) analysis method for scRNA-seq based on the zero-inflated negative binomial (ZINB) model. It has several promising properties and has better performances than existing methods. Our method will not only detect the difference of means and variances of gene expression, but will also examine the difference in the zero expression data which include genes not expressed in the sample and genes expressed but missed in the sequencing due to the so-called “drop-out”effect in scRNA-seq. With this model, we defined three different types of DE genes for scRNA-seq data, which provides a new perspective for DE analysis of single-cell data.

We evaluated our method on simulated data and got higher AUC (Area Under Curve) of ROC (Receiver Operating Characteristic) curve compared to existing methods. We also validated our method on a published scRNA-seq dataset of human preimplantation embryos. Our method found more DEG than other method and showed different patterns of the DE genes. We also found more lineage specific genes for the three lineage of preimplantation embryos, which has important biological implications.

P29

Interactive visualization of Hi-C data and epigenetic marks

Robert Schöpflin and Martin Vingron

Three-dimensional genome organization plays an important role for many biological processes, such as gene expression and DNA replication. In the last years, chromatin conformation capture became a powerful technique to investigate the chromatin structure on a large scale. High-throughput versions of this technology, such as Hi-C and Capture Hi-C, generate two-dimensional interaction maps which reveal the compartmentalization of chromatin, topological associated domains and chromatin loops. The overlay of these interaction maps with epigenetic features, such as DNA accessibility and ChIP-seq tracks of histone modifications and transcription factors indicate a tight connection between epigenetic landscape and chromatin structure and have advanced our understanding of gene regulation. Different approaches and applications have been developed to display interaction maps together with tracks of epigenetic marks interactively. However, many of them are web tools, prohibiting fluent browsing of large interaction maps or lack the possibility to display several maps at the same time for comparative analysis.

Here, we present the HiC2-viewer, a stand-alone application developed for the fast visualization of Hi-C and Capture Hi-C maps with simultaneous integration of bed and bedgraph tracks. It allows a flexible composition of different panels to visualize maps from different cell lines, tissues or even organisms, along with epigenetic information. The software offers the manual annotation of domains and chromatin loops, which can be exported and overlayed with other maps. Furthermore, multiple virtual 4C tracks can be generated enabling an in-depth analysis of the interaction landscape at selected loci. The software is implemented in a system-independent manner and available for various operating systems.

P30

Regulation of Differentiation in SH-SY5Y Neuroblastoma Cells by Translation of Upstream Open Reading Frames

Sang Chun, Caitlin Rodriguez, Ryan Mills and Peter Todd

Upstream open reading frames (uORFs) are prevalent in the human transcriptome and may alternatively regulate the abundance of canonically translated proteins through the promotion of mRNA decay, or the competitive appropriation of translational resources, among other mechanism. uORFs are conserved across species, and have been annotated to genes with diverse biological functions including oncogenes, cell cycle control, differentiation, and stress response. To that end, various algorithms have been published to profile the translational status and efficiency of canonical and non-canonical ORFs, including uORFs, through the analysis of sequence reads derived from ribosome-protected fragments (RPFs) of mRNA. We have developed a translational classification algorithm based on the magnitude of coherence between the alignment of RPFs to a transcript against an idealized reference signal based on the codon-dependent tri-nucleotide signal characteristic of active translation. Although previous studies have identified and characterized uORFs from ribosome profiling sequence data, less well studied is the context in which variable uORF translation might regulate complex biological processes, like cellular differentiation. In this study, we performed ribosome profiling and mRNA-Seq on non-differentiated SH-SY5Y neuroblastoma cells and those that were induced to undergo neuronal differentiation. We applied our spectral profiling algorithm and identified 4,542 actively translated uORFs, and then experimentally validated a subset of uORFs with potential differential regulatory potential on their downstream CDS across cells states.

P31

Quality assessment of genome assembly generated by the third-generation sequencing platform PacBio and an online editor for continuous update of genome annotation

Jian-Ying Chiu, Tse-Ching Ho, Syun-Wun Liang and Yen-Hua Huang

High-throughput sequencing data usually contains more errors than the sequences generated by Sanger sequencing method, thereby causing problems in downstream genome annotation. In particular when the third-generation seqencing platform, PacBio, is regarded, indels that cannot be completely corrected in genome assembly can lead to errors in downstream open-reading frame prediction. Additionally, unclosed gaps resulted from yet unoptimized genome assembly create difficulties in predicting gene structures and in inferring gene functions. To address the need for quality assurance of genome assembly and continuous improvement of genome annotation, we implemented a quality assessment pipeline and a genome annotation editor for genome assembly. Our pipeline is to assist the evaluation of the completeness of assembly and metrics of quality of raw data, such as seed read length and quality score distributions. According to the quality level assigned to each sample, parameters such as tentative genome size, read quality and seed read length, could hence be tuned to optimize the genome assembly. Moreover, to recover missing or wrong gene structures due to the sequence errors remained, we implemented a online genome annotation editor to allow continuous improvement of genome annotation. Our editor is a middleware system consisting of APIs to facilitate the editing of gene structures in private Ensembl database, including adjusting/merging existing gene models and creating new genes, which enable synchronization of genome annotation with new evidence for novel transcripts. We showed that the pipeline and the editor can work cooperatively to improve the confidence of the data in a genome sequencing project.

P32

Molecular signatures that can be transferred across different omics platforms

Michael Altenbuchinger, Philipp Schwarzfischer, Thorsten Rehberg, Jörg Reinders, Christian W. Kohler, Wolfram Gronwald, Julia Richter, Monika Szczepanowski, Neus Masqué-Soler, Wolfram Klapper, Peter J. Oefner and Rainer Spang

Motivation: Molecular signatures for treatment recommendations are well researched. Still it is challenging to apply them to data generated by different protocols or technical platforms.

Results: We analyzed paired data for the same tumors (Burkitt lymphoma, diffuse large B-cell lymphoma) and features that had been generated by different experimental protocols and analytical platforms including the nanoString nCounter and Affymetrix Gene Chip transcriptomics as well as the SWATH and SRM proteomics platforms. A statistical model that assumes independent sample and feature effects accounted for 69% to 94% of technical variability. We analyzed how variability is propagated through linear signatures possibly affecting predictions and treatment recommendations. Linear signatures with feature weights adding to zero were substantially more robust than unbalanced signatures. They yielded consistent predictions across data from different platforms, both for transcriptomics and proteomics data. Similarly stable were their predictions across data from fresh frozen and matching formalin-fixed paraffin-embedded human tumor tissue.

Availability: The R-package "zeroSum"can be downloaded at `https://github.com/rehbergT/zeroSum`. Complete data and R codes necessary to reproduce all our results can be received from the authors upon request.

P33

CircMarker: A Fast and Accurate Algorithm for Circular RNA Detection

Xin Li, Chong Chu, Jingwen Pei, Ion Măndoiu, and Yufeng Wu*

Computer Science & Engineering Dept., University of Connecticut, Storrs, CT, USA
{fxin.li,chong.chu,jingwen.pei,ion.mandoiu,yufeng.wug}@uconn.edu

While RNA is often created from linear splicing during transcription, circular RNA (or circRNA) is a type of RNA which forms a covalently closed contin- uous loop. It is now believed that circRNA plays an important biological role in diseases and traits. Several experimental methods, such as RNase R, have been designed to enrich circRNA while degrade linear RNA. Although several useful software tools for circRNA detection have been developed as well, these tools may miss many circular RNA. Also, existing tools are slow for large data because they often depend on reads mapping. To deal with these problems, we developed a new computational approach, named CircMarker, based on k-mers rather than reads mapping for circRNA detection. CircMarker takes advantage of transcriptome annotation files to create space-efficient k-mer table, and applies several different criteria and filters for circRNA detection. We also compared CircMarker with other three circRNA detection tools in both simulation and real data. Two different circRNA coverage cases have been applied for simulation data and CircMarker performs the best in all indicates of these two cases, including the number of called circRNA, accuracy, and running time. For real data evaluation, two different strategies have been applied. The first strategy uses RNase R treated sequence reads with public database, and the second one uses both RNase R Treated/Untreated Data from one raw specimen. Empirical results show that CircMarker can find more reliable circular RNA and obtain higher reliability ratio and consensus-based sensitivity with low bias. CircMarker is generally 5 times faster than others.

P34

Finding associated variants in genome-wide associations studies on multiple traits

Lisa Gai, Dat Duong and Eleazar Eskin

Many variants identified by genome-wide association studies (GWAS) have been found to affect multiple traits, either directly or through shared pathways. There is currently a wealth of GWAS data collected in numerous phenotypes, and analyzing multiple traits at once can increase power to detect shared variant effects. However, the vast majority of studies consider one trait at a time, and studies that do analyze multiple traits are typically limited to sets of traits already believed to share a genetic basis. Although multivariate approaches increase power when combining data on related traits, they can be underpowered compared to univariate analysis when the traits are truly unrelated. However, the degree to which a pair of traits share effects is often not known, so a flexible method is necessary. Here we present a method for finding associated variants from GWAS summary statistics for multiple traits that estimates the degree of shared effects between traits from the data. Using simulations, we show that our method properly controls the false positive rate and compares favorably to trait-by-trait GWAS and multivariate regression at varying degrees of relatedness between traits. We then apply our method to real GWAS datasets in a wide variety of disease traits and medically relevant traits.

P35

Microbiome Search Engine: Enabling Rapid Microbiome Samples Search In Large-scale Database

Gongchao Jing, Jian Xu and Xiaoquan Su*

Bioinformatics Group, Single-Cell Center, Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, Shandong 266101, China
{jinggc, xujian, suxq}@qibebt.ac.cn

With the rapid growth of microbiome research across the globe, one major challenge is large-scale analysis of metagenome datasets under context of the metagenome space known to mankind to-date. Here we describe Microbiome Search Engine (MSE), a powerful database engine that enables rapid sample search against large-scale of microbiome datasets (Figure 1). This database now contains 51,032 curated microbiome samples that are of clear scientific background from 77 studies. The input query microbiome samples can be uploaded by users via web or standalone interface and then searched against the entire database for structurally or functionally similar microbiomes in a 'BLAST-like'manner. The search results, typically returned with ultrafast speed, provide visualized organismal or functional alignment patterns between queries and matches with quantitative similarity scores. The search for 'best matches'and 'top N matches'from the vast amount of microbiomes accumulated so far represents a novel way for not only annotation new microbiome datasets but also identifying scientific hypotheses that probe the complex interplay between microbiome features and ecosystem parameters. More information is available at: `http://mse.single-cell.cn`.

P37

Splicing aberration of TP53 transcripts in cancers: mechanisms and effects

Dan Huang[1], **FY Hu**[2], **Nelson Tang**[1,3]

[1]Department of Chemical Pathology, Faculty of Medicine, The Chinese University of Hong Kong, HongKong, China
[2]Department of Statistics. Faculty of Science, Wuhan University of Technology
[3]Li Ka Shing Institute of Health Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China

TP53 is a very important tumour suppressor gene. Germline mutations caused familial cancer syndromes. There are also somatic mutations causing non-sense, frame-shift and premature termination found in various cancers. However, little is known about splicing aberration in cancer.

Methods: The RNA-seq data in the TCGA dataset was screen for transcripts with novel splicing sites in the TP53 gene. Their prevalence, location, origin of the tumor and effects on disease prognosis were analyzed.

Results: There are 28 known transcript variants for TP53. In addition, we identified 7 novel transcripts. 71% of them led to a frame-shift in the protein coding sequence. The transcript causing a splicing aberration between exon 7 and exon 8 was recurrent in 30 tumour samples. These tumours were originated from breast, lung, colon and thyroid tissues. We then investigated the mechanism responsible for this splicing event. All of the samples with high expression level of this novel transcript could be explained by one of the 2 somatic mutations near the splicing sites. Namely, a C to A mutation at the splicing acceptor site and a C to A/G mutation at the 6bp downstream of the end of intron 7.

Finally, we compared the survival of patients with or without this splicing aberration. It was found that breast cancer patients without splicing aberration have a better survival.

Conclusion: Analysis of RNA-seq of cancer transcriptome provides important information about the gene expression and variation of transcripts. It also provides a means to assess the functional consequence of somatic mutations and characterize novel transcripts.

P38

Investigating residue coevolution in proteins from a structural perspective

Alexey Morgunov, Alexander Gunnarsson, Norbert Fehér and Madan Babu

In recent years, there has been remarkable progress in the development of computational methods for detecting evolutionary couplings between residues in proteins from multiple sequence alignments (MSA) of protein families. These methods have been successfully applied in predicting three-dimensional structures from amino acid sequences, as well as in identifying functionally important residues. One of the key limitations that hinder wider applicability and higher accuracy of the so-called coevolutionary methods lies in the fact that covariation between residues is not only a function of structural and functional relationships between them - two related components that are hard to disentangle in their own right - but also of potential interactions with other proteins and ligands, of phylogeny, and of stochastic noise in the imperfect MSA. In order to begin teasing apart the contributions of various sources of coevolution to the observed signal, in the first instance we looked at the structural components. We performed a large-scale analysis of protein structure datasets, focusing on statistics that describe types, expected distributions and propensities for specific structural residue contacts, as well as contacts arising within protein complexes. This quantitative information was compared directly to the output of best performing coevolutionary methods, and the expected distributions were subsequently used as prior information to improve the performance of some of the methods in detecting important residue couplings based on previously published analyses.

P39

Inferring Dynamic Network Architecture from Time Series Perturbation Data

Gregory Smith, Alan Stern and Marc Birtwistle

Constructing predictive quantitative models for networks describing cell fate decisions remains an important goal to better understand the progression of and treatment options for numerous diseases. Even for small networks, this remains challenging since a network is defined by local connections between nodes, but experimental measurements reflect overall, or global, network behavior. Current methodology struggles with inferring local connections within cycles, such as signaling cascades with feedback loops, where weak local connections can be amplified into strong global behavior. Dynamic modular response analysis (DMRA) provides a framework to calculate local connection strengths from time-series perturbation data that handles cycles in network architecture but with two drawbacks: DMRA requires two perturbation experiments per node, a large expense of time and resources, and the identification of connection strengths is sensitive to experimental noise. We address these limitations by recasting the DMRA estimating algorithm, such that the experimental approach requires only a single perturbation experiment for each node and the estimates are robust against noise in the data. We tested our method on simulated datasets representing likely network architectures and randomized network models for 2- and 3-node systems, and on a MAPK/ERK signaling cascade model with experimentally-derived rate parameters and multiple feedback conditions, applying increasing levels of noise to each simulated perturbation time course measurement. We found we can classify local connection directionality under the presence of simulated noise for each dataset. Predicting time-dependent values for more complex network dynamics and scaling to larger networks will be the next focus of our work.

P40

SeqMiner: A Weka package for mining phage display sequence data

Daniel J. Hogan[1], Bharathi Vellalore[2], Clarence R. Geyer[2], Anthony J. Kusalik[1]

[1]Dept. of Computer Science, University of Saskatchewan
djh901@mail.usask.ca
[2]Dept. of Biochemistry, University of Saskatchewan

Many standard classification and regression models used in data mining are able to handle numeric and nominal attributes, but few are able to handle string attributes like protein sequences. We have created a Weka package called SeqMiner for extracting numeric features from protein sequences that can be used to train and test classifiers in the Weka machine learning workbench. SeqMiner contains classes for extracting three sets of numeric features from protein sequences: (1) amino acid counts, (2) dipeptide counts, (3) and Kidera factors (i.e. 10 principal components from a PCA of hundreds of physicochemical properties).

A Weka KnowledgeFlow using SeqMiner classes was used to extract features from the CDR-H3 sequences of several antibody phage display (APD) experiments. The features included (1) amino acid counts, (2) dipeptide counts, and (3) Kidera factors. The features were used to train and test a random forest classifier to predict CDR-H3 sequences that become enriched during the APD experiments, which indicates binding to the target. The random forest achieved a classification accuracy of 84.4%. One potential application of this classifier is in the design of APD libraries with sequence landscapes yielding more binders.

SeqMiner can drastically reduce the effort required for data mining of short protein sequences like those generated by phage display. As part of the Weka ecosystem, SeqMiner benefits from Weka's many features, including KnowledgeFlow. The SeqMiner package is available on github (`github.com/djh901/seqminer`). Feature requests may be sent to `djh901@mail.usask.ca`.

P41

Feature reduction for practical radiation biodosimetry using combined approach of gene co-expression network and pathway knowledge

Chao Sima, Shanaz Ghandhi, Sally A. Amundson, David J. Brenner and Michael Bittner

There are numerous efforts in developing biomarkers to estimate the radiation exposure doses and injury in a large scale radiological emergency, including using gene expression profiles which were shown to predict well between radiation doses. In such emergency, it is essential to be able to get the affected population prompt treatment so efficiency in triage is crucial. There is therefore a need to speedup field tests by reducing the number of genes in the predictor signature from the typical number of close to a hundred or more, among which many are functionally related and profile-wise correlated. To achieve this, we have developed an approach that combines both the gene co-expression and pathway knowledge. Specifically, we have derived a gene co-expression network using the expression profiles for the signatures genes in the full predictor panel, and gathered pathway knowledge for these genes using multiple databases including BioCarta, KEGG, NCI, Panther and Reactome. The topological association between two genes are defined as the shortest path between them and the resulted pathway network is overlaid on the co-expression network. Functional clusters are then identified from the combined network, and the “hub”genes are used to represent each cluster. This new predictor panel, consisting of these hub genes, is of much reduced dimension. In validation, we have shown that the full-panel and reduce-panel signature genes predict nearly identically. The proposed method demonstrates an effective way of reducing the number of signatures need to be tested for biodosimetry in the field without suffering performance loss.

P42

3D Yeast Genome Prediction with Constraint Logic Programming

Kimberly Mackay, Christopher Eskiw, Anthony Kusalik

Background: In order for a cell's genetic information to fit inside its nucleus, the chromosomes must undergo extensive folding and organization. It is possible that different genomic organizations (or architectures) may relate to distinct nuclear functions. Until recently it was difficult to investigate this relationship due to the lack of high-throughput techniques for identifying genomic architectures. Fortunately, the development of Hi-C has made it possible to detect genomic regions that are in close 3D spatial proximity (or are "interacting"). Several methods have been developed that leverage Hi-C data to predict the 3D genomic architecture. None of these methods have utilized a constraint logic programming (CLP) approach despite CLP being successfully used to predict the 3D structure of other biomolecules.

Objective: Develop a CLP-based tool that utilizes Hi-C datasets to model the 3D genomic architecture.

Methodology: To generate the CLP program, Hi-C interaction frequencies were converted into a set of corresponding structural constraints. The constraints were used as a framework to determine plausible structures through constraint satisfaction with ECLiPSe. The predicted model was transformed into a graph and visualized using Cytoscape.

Conclusion: The developed program was able to predict a logical model of the 3D yeast genome within 15 minutes. A literature search verified that the predicted model recapitulated key documented features of the yeast genome. Overall, the problem formalism and program developed here demonstrates the power of CLP applications for modelling the 3D genome and are a step towards a better understanding of the relationship between genomic structure and function.

P43

Transcripsome and metabolism studies to identify alkaloid biosynthesis genes in poppies

Jae Hyeon Oh, Tae Ho Lee, Chang Kug Kim, Lee Dong-Jun and Dowan Kim

In this study, presence of the gene were identified through genetic studies about Ornamental poppy related alkaloid gene. Also we identified synthesis control of morphine and codeine on pathway through transcriptome and metabolism studies. Papaver somniferum (opium poppy) is one of the source for several pharmaceutical benzylisoquinoline alkaloids (BIAs). Many BIAs possess potent pharmacological properties such as narcotic analgesics, antimicrobials, muscle relaxants, and anticancer drug. Morphine and codeine are representatives. Samples were used Papaver rhoeas (red), Papaver nudicaule (pink), Papaver nudicaule (orange) for this study. Leaves were used 60 days and 90 days after germination for transcriptome and metabolism studies. Trinity were processed for assemble and whole genes were identified using BlastX in this Rna-seq analysis. Expression level of each genes identified using RSEM analysis. Result shows total 363,229 genes on leaf(60d). We identified related alkaloid genes 143,324 from Papaver rhoeas, 140,327 from Papaver nudicaule (pink), 135,978 from Papaver nudicaule. Also we identified total 380,950 genes on leaf(90d). it contain related alkaloid genes were 83 from Papaver rhoeas (red), 80 from Papaver nudicaule (pink), 78 from Papaver nudicaule (orange). Ultra high performance liquid chromatography coupled with quadrupole-time-of-flight mass spectrometry and metabolomics were established to characterize the chemical profiling and explore significantly potent metabolites. In total, potent 41 candidates closely related to opium poppy in benzylisoquinoline alkaloid biosynthesis were identified based on MS and MS/MS, and 18 compounds were characterized among six species. Also, Triple quadrupole mass spectrometry with UHPLC was used to investigate three targeted metabolites in different six species. Quantitative analysis was conducted. Protopine was detected in all samples. The level of protopine in Papaver rhoeas (red) was higher than other species at all different growth time. The contents of allocryptopine were various depends on species and growth time. Chelidonine was observed only in Papaver rhoeas (red). In this study, we identified locate of gene on pathway expression pattern, level of metabolites through result of transcriptome and metabolism analysis.

P44

Assessing mRNA integrity of single-cell RNA-Seq data using mRIN

Wendao Liu[1], Zhun Miao[2], Xuegong Zhang[2,1,*]

[1]School of Life Sciences and Center for Synthetic & Systems Biology, Tsinghua University, Beijing 100084, China

[2]MOE Key Laboratory of Bioinformatics; Bioinformatics Division, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China

zhangxg@tsinghua.edu.cn

Single-cell RNA sequencing (scRNA-seq) is an emerging technology recent years with broad applications, but its data quality can be low due to degraded RNA in some samples. Quality control (QC) of scRNA-seq data is a critical step before downstream analyses. Among existing QC methods for scRNA-seq data, few have focused on the assessment of the mRNA integrity.

We applied the previous tool mRIN on scRNA-seq data to assess mRNA integrity of datasets from different protocols. The method is based on quantitative modelling of 3'bias of RNA-seq read coverage. A mRNA integrity number (mRIN) was calculated from the sequencing data of each sample to measure mRNA degradation. The quality of a scRNA-seq dataset was assessed by analyzing the distribution of mRINs of the samples. A gene integrity score (GIS) derived from mRINs can indicate the gene-specific degradation.

We did a systematic assessment and comparison between datasets of different scRNA-seq protocols with mRIN, which characterized of the mRNA integrity of different protocols. Results showed that it is advisable to assess the quality of scRNA-seq data with mRIN before conducting further studies.

P45

OMView: a comprehensive visualization suite for interpreting optical mapping data

Alden King-Yung Leung, Nana Jin, Kevin Y. Yip and Ting-Fung Chan

Optical mapping is an imaging technique for capturing enzyme patterns of long DNA molecules. It has been applied on several areas of biological discovery including structural variation detection, strain typing, and assisted sequence scaffolding. Emerging high-throughput technology for optical mapping data generation raises the demand of pipelines for efficient data processing and result interpretation. Among them, data visualization is one of the essential components that helps understand and illustrate the findings. Recently, several visualization packages for optical mapping are available, but none of them offer solutions to the needs of various styles in data interpretation. Here we developed OMView, a software suite for visualizing optical mapping data. OMView provides an interactive interface and high-quality visualizations, and supports multiple visualization styles. These include visualization of (1) alignment for overall aligning patterns across regions of interest and structural variations discovery; (2) multiple alignment for genome comparison; and (3) molecules and assemblies for quality assessment. The software suite serves as a convenient and powerful visualization tool for data illustration that facilitate optical mapping analyses.

P46

RED-ML: a novel, effective RNA editing detection method based on machine learning

Heng Xiong, Dongbing Liu and Leo Lee

RNA editing is a regulated post-transcriptional modification process that can flexibly and dynamically alter the sequence of RNA transcripts during development and in a cell-type specific manner. It contributes to various diseases when misregulated, including neurological disorders and cancer. With the advancement of NGS techniques, RNA editing is now being studied under a growing number of biological conditions. However, a major barrier that prevents RNA editing from being a routine RNA-seq analysis, similar to gene expression for example, is the lack of user-friendly and effective computational tools. In this work, we developed a highly accurate, speedy and general-purpose software tool RED-ML: RNA Editing Detection based on Machine Learning. The input to RED-ML can be as simple as a single BAM file, while it can also take advantage of matched genomic variant information when available. The output not only contains detected RNA editing sites, but also a confidence score to facilitate downstream filtering. We have carefully designed high throughput validation experiments and performed extensive comparison and analysis to show the efficiency and effectiveness of RED-ML under different conditions. For example, with validation rates of 0.9 or higher, RED-ML detected 30,000-50,000 RNA editing sites in three cancer samples. It also substantially outperforms existing methods in terms of sensitivity, specificity and computing cost. With the arrival of RED-ML, which is freely available via GitHub ¡https://github.com/BGIRED/RED-ML¿, it is now possible to conveniently make RNA editing a routine analysis of RNA-seq to facilitate and accelerate our understanding of this intriguing post-transcriptional modification process.

P47

Reducing Runtime in CONTRAlign by Feature Reduction

Dawn Chen and Ming-Jing Hwang

This poster proposes a synthesis of feature projection method with the CONTRAlign framework. The CONTRAlign framework uses a gradient descent-based regression method to determine the proper alignment of a given pair of genetic sequences. It relies on finding Hidden Markov model representations of each alignment, and using those representations to determine features to use in regression. Since the time required for gradient descent is often dependant on the number of features, CONTRAlign having a total of 473 could be the reason it has the longest running time when compared to other machine learning methods of sequence alignment. In order to reduce CONTRAlign's running time while keeping the original accuracy, we implement feature projection on its training features after CONTRAlign calculates them. Feature projection is a method often used in machine learning to reduce the number of features by projecting the data onto a selection of its principal component vectors, while preserving the complexity of the original data. As the complexity of the original data is preserved, the accuracy is not affected too severely. A test on a dataset of 5000 examples and 1000 features implemented in MATLAB has shown that this feature projection method can maintain a level of accuracy above 90% with up to an 87.5% reduction in the number of features, with running time reduced at an approximately linear rate with the reduction in the number of features. The results of feature-reduced alignments will be discussed.

P48

drVM: detect and reconstruct known viral genomes from metagenomes

Hsin-Hung Lin, Yu-Chieh Liao*

Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan
`jade@nhri.org.tw, oliver0618@nhri.org.tw`

Background

The discovery of new or divergent viruses using high-throughput next-generation sequencing (NGS) has become more commonplace. However, although analysis of deep NGS data allows us to identity potential pathogens, the entire analytical procedure requires competency in the bioinformatics domain, which includes implementing proper software packages and preparing prerequisite databases. Simple and user-friendly bioinformatics pipelines are urgently required to obtain complete viral genome sequences from metagenomic data.

Results

This manuscript presents a pipeline, drVM (detection and reconstruction of viral genomes from metagenomes), for rapid viral read identification, genus-level read partition, read normalization, de novo assembly, sequence annotation and coverage profiling. The feasibility of drVM was validated via the analysis of over 300 sequencing runs generated by Illumina and Ion Torrent technologies to detect and reconstruct a variety of virus types including DNA viruses, RNA viruses and retroviruses. drVM is available for free download at: `https://sourceforge.net/projects/sb2nhri/files/drVM/` and is also assembled as a Docker container, an Amazon machine image and a virtual machine to facilitate seamless deployment.

Conclusions

drVM was compared with other viral detection tools to demonstrate its merits in terms of viral genome completeness and reduced computation time. This substantiates the platform's potential to produce prompt and accurate viral genome sequences from clinical samples.

P49

Predicting Allergens and Identifying Interpretable Allergenic Biological Features using Machine Learning Algorithms

Kuei-Ling Sun[1], **Onkar Singh**[1,2], **Emily Chia-Yu Su**[1,*]

[1]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan
emilysu@tmu.edu.tw
[2]Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan

Objective

Allergy is the response of the immune system to environmental stimuli and is a major health problem in human society. Although several methods have been developed to predict allergenic proteins, only few approaches can propose interpretable biological features to help biologists identify allergens.

Methods

In this study, first we collected benchmark data sets from AlgPred, AllerTop and AllerHunter, and curated the data sets. Then, we represented the protein sequences with various encoding schemes, such as amino acid composition, dipeptide composition, and pseudo amino acid composition. After that, we applied machine learning algorithms including decision trees, logistic regressions, and artificial neural networks to predict allergenic proteins. Moreover, we will propose interpretable biological features that can be used to identify allergens.

Results and Conclusion

Experiment results show that artificial neural networks achieved 0.94 in area under the receiver operation characteristics curve. However, decision tress and stepwise regression are used to identify interpretable biological findings. For example, it is observed from our study that allergenic proteins contain more positively charged amino acids than non-allergens. In addition, our study can also propose discriminative sequence patterns to distinguish allergens and non-allergens.

Conclusion

We conclude that our method can not only accurately predict allergens, but also propose valuable biological insights for experimental design.

P52

Transcriptome analysis in some varieties of Papaver rhoeas

Dowan Kim, Jae-Hyeon Oh, Dong-Jun Lee, Chang-Kug Kim and Tae-Ho Lee

Papaver rhoeas is a common plant resource around Korea. Some of these ornamental poppies have found some alkaloids, including chelidonine and cryptopine. Thus, transcript analysis was performed according to the timing of ornamental poppies in order to carry out studies related to alkaloid production. Transcript analysis was used to identify changes in the transcripts of the alkaloids and to analyze the alkaloid biosynthetic gene cluster.

P53

ABCProfiler: a program for Alignment Based Clustering and taxonomy Profiling

Hojin Gwak and Mina Rho

16S ribosomal RNA sequences are widely used to analyze the bacterial composition in the microbiome. In such studies, operational taxonomy unit (OTU) clustering is performed for taxonomy assignment. In many cases, similarity-based clustering is applied to construct OTUs. However, some bacteria show extremely high sequence similarity, thus making it challenging to distinguish individual bacterial species by using similarity-based clustering method. Although such bacterial species show extremely high similarity, many of them have polymorphic sites, and are thus distinguishable from other species. By taking advantage of this unique property, we have developed a system that performs two-step clustering. Here, similarity-based clustering is performed in the first step; further precise splitting is carried out based on polymorphic sites in the second step. Sequences from close species, such as Escherichia and Shigella, were better divided into distinctive groups. Finally, species-level profiling of each group is provided by using homology search and the Ribosomal Database Project (RDP) Classifier. Our method can successfully profile the bacterial composition with higher sensitivity than similarity-based clustering alone.

P54

Modular type I PKS cluster in sediment microbiomes

Jehyun Jeon and Mina Rho

Microbial products have significant effects on our environments and health. In particular, secondary metabolites have been studied to reveal their important roles as antibiotics and antitumor agents. A comprehensive investigation of the gene families of biosynthetic gene clusters (BGCs) under certain environment could provide valuable information. In this study, we focused on modular type I PKS cluster. We conducted extensive analysis for ketosynthase (KS) of modular type I PKS cluster using bacterial reference genomes based on sequence similarity. KS domains of the modular type I PKS cluster in intra-cluster and inter-cluster were investigated. In the network analysis of KS domains, different BGC subfamilies that share over 70% of domain similarity exist even in the same genus and species. The domain architecture of core genes of the modular type I PKS cluster affects the structure of final products. Up to date, most studies have focused on finding BGCs from cultured microbes. Novel BGCs of abundant microbes that have not been sequenced need to be studied in detail. We thus reconstructed diverse PKS proteins from twelve sediment microbiomes by applying a de novo approach to recruit and assemble the reads that are homologous to the BGC proteins of the modular type I PKS cluster. In addition, we suggested that global landscape of novel and known KS domains from environments as well as references.

P55

Learning From Sequence-Activity Data - Predicting Enantioselectivity of an Epoxide Hydrolase

Julian Zaugg, Yosephine Gumulya and Mikael Bodén

Enantioselective enzymes are highly desirable to the biochemical and pharmaceutical industries for improving economics of chemical synthesis. To improve selectivity, mutations are introduced around the catalytic active site region. In this compact environment higher-order epistatic interactions between mutations, where contributions to enzyme fitness are non-additive, play a significant role in determining the degree of selectivity. Simple linear models assuming residue independence are unlikely to accurately predict the fitness contributions of such mutations. We hypothesise that capturing epistatic interactions in predictive models will improve identification of selectivity-enhancing mutations.

In this study a number of Support Vector Machine (SVM) models have been constructed to model the relationship between mutations and the degree of enantioselectivity for (S)-glycidyl phenyl ether for a library of 136 variants of the epoxide hydrolase from the fungus Aspergillus niger (AnEH). Sequences are compared using a kernel function representing both epistatic interactions and physicochemical differences between mutations. To illustrate the effect the lack of epistatic interactions has on predictive accuracy, models are also evaluated on a set of 'interaction-minimised'sequences.

Higher-order models display improved predictive accuracy compared to linear alternatives. Comparing polynomial d = 3 and linear models, 50×5-fold cross-validation correlation r scores are 0.90 vs 0.79 respectively. Equivalent models tested on interaction-minimised sequences results in scores of r = 0.91 and r = 0.88. Testing on additional AnEH mutants, scores of r = 0.86 and r = 0.57 were observed.

P56

Using Theory to Reconcile Experiments: Understanding the Origin of Enantioselectivity of Epoxide Hydrolase

Julian Zaugg, Yosephine Gumulya, Mikael Bodén and Alpeshkumar Malde

Epoxide hydrolases (EH) are a family of enzymes exhibiting a high sequence diversity and capable of producing enantiopure epoxides and diols, which are valuable intermediates in pharmaceutical and biotech industries. The biochemical reaction involves (thermodynamic) binding of racemic epoxide substrate to EH followed by formation of an intermediate and its (kinetic) hydrolysis to release the enantiopure diol product. Despite members of the EH family largely sharing a common reaction mechanism, the origin of their enantiomeric preferences varies among species. Understanding the origin of enantioselectivity at an atomic level is crucial in guiding the design of mutants and optimising the chiral separation processes. We have used a combination of docking, molecular dynamics simulations and free energy calculations in combination with the available experimental, structural and biochemical data to understand the origin of enhanced enantioselectivity of a specific mutant LW202 (E = 193) compared to the wild type (E = 3) EH from the fungus Aspergillus niger (AnEH) for the racemic substrate glycidyl phenyl ether (GPE). The study reveals that there is no preference for a given enantiomer and both enantiomers of GPE bind to wild type and LW202 with equal affinity contrary to productive positioning of the preferred S enantiomer as reported by Reetz et al. Combining this observation with the re-interpretation of the biochemical data (Reetz et. al, 2009) reveals that improved catalysis (kinetic effect) is responsible for increasing the enantiomeric preference in LW202 mutant and catalysis of GPE by LW202 may not follow Michaelis-Menten kinetics.

P57

Identification of drug-target interactions using weighted interactome network

Ingoo Lee, Hojung Nam*

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, Republic of Korea

The identification of drug-target interaction acts as crucial role in drug discovery because most of drugs are molecular compounds which interact with target proteins. However, identifying drug-target interactions with chemical or biological experiment is very laborious, time-consuming and costly expensive. Thus, predicting drug-target interactions using computational approaches are good alternatives. To build a computational prediction model, firstly, we retrieved known drug-target interactions represented in a bipartite graph structure and converted them to feature vectors. Protein-protein interactions and drug-drug interactions were constructed as a graph structure because a drug can interact with multiple number of target proteins and vice versa. Secondly, drug-drug interactions and protein-protein interactions were reweighted by applying an affinity score that is the result of random walk with restart on each interaction graph. Finally, we re-calculated drug-target feature vectors by using newly weighted graph topology as reflection of multiple effects of a drug. The vector-transformed and re-calculated drug-target interaction features were used as a positive training set for many machine learning classification algorithms like support vector machine. To validate our model, we performed 10-fold cross-validation with the positive set and the randomly generated negative set. The prediction model showed high performance of 93% accuracy and 0.98 AUC. This result shows that our proposed method can yield new drug-target interactions from known data with the constructed prediction model.

P58

Sequence Alignment on Directed Graphs

Kavya Vaddadi, Naveen Sivadasan, Kshitij Tayal and Rajgopal Srinivasan

Genomic variations in a reference collection are naturally represented as graphs such as genome variation graphs. Such graphs encode common subsequences as vertices and the variations are captured using additional vertices and directed edges. The resulting graphs are directed graphs possibly with cycles. Existing algorithms for aligning sequences on such graphs make use of partial order alignment (POA) techniques that work on directed acyclic graphs (DAG). For this, acyclic extensions of the input graphs are first constructed through expensive loop unrolling steps (dagification). Also, such graph extensions could have considerable blow up in their size which in the worst case is proportional to the input sequence length.

We provide a novel alignment algorithm V-ALIGN that aligns the input sequence directly on the input graph while avoiding such expensive dagification steps. V-ALIGN is based on a novel dynamic programming formulation that allows gapped alignment directly on the input graph. It supports affine and linear gaps. We also propose refinements to V-ALIGN for better performance in practice. In this, the time to fill the DP table has linear dependence on the sizes of the sequence, the graph and its feedback vertex set. We perform experiments to compare against the POA based alignment. For aligning short sequences, standard approaches restrict the expensive gapped alignment to small filtered subgraphs having high 'similarity'to the input sequence. In such cases, the performance of V-ALIGN for gapped alignment on the filtered subgraph depends on the subgraph size.

P59

Kmerlight: fast and accurate k-mer abundance estimation

Naveen Sivadasan, Rajgopal Srinivasan and Kshama Goyal

Motivation:

k-mers (nucleotide sequences of length k) form the basis of several algorithms in computational genomics. In particular, k-mer abundance information in sequence data is useful in read error correction, parameter estimation for genome assembly, digital normalization etc.

Results:

We give a streaming algorithm Kmerlight for the estimation of the k-mer abundance histogram from sequence data. The algorithm uses logarithmic space and runs in linear time. Kmerlight can efficiently process genome scale and metagenome scale data using only standard desktop machines. We conduct several experiments to measure the accuracy and performance of Kmerlight. With less than 500 MB RAM (and no disk space), Kmerlight achieved 2 percent relative error. We also provide analytical bounds on the error guarantees of our algorithm. Resource frugal nature of Kmerlight allows simultaneous computation of abundance histograms with different values of k, which for instance is required in parameter estimation for genome assembly.

Few applications of abundance histograms computed by Kmerlight, such as de novo estimation of repetitiveness in the genome and estimation of k-mer error rate, are also shown. Our algorithm can be used for abundance estimation in general streams. To the best of our knowledge, our algorithm is the first streaming algorithm to solve this problem using only sublinear query time and space and with analytical guarantees. This we believe is of independent theoretical and practical interest.

Availability:

The Kmerlight tool is written in C++ and is available for download and use from `https://github.com/nsivad/kmerlight/`

P60

Species-species network analysis for microbial communities

Kang Ning, Shaojun Yu and Maozhen Han

Samples and data have been accumulating for microbiome research, which hold promise for better understanding of the microbial communities either in the environment or in our body. However, for deeper analysis of microbiome, advanced data-mining of this paramount of data should be performed. Among the various priorities for microbiome data mining, one of the most important is the examination of species-species co-occurrence patterns, which is directly related with microbial ecology, could help for better understanding of species interactions, co-modularities and key species in the community.

In this work, we have designed the Meta-Network framework for establishment and interpretation of the species-species co-occurrence relationships, based on using loose definition of network as well as mutual information. We have examined several large cohorts of microbial communities from both human gut and ocean samples, based on which we have established the species-species co-occurrence networks for respective samples. Firstly, the species-species co-occurrence network, generated by loose definition strategy, contains more rational relationships among species. Secondly, the species-species co-occurrence patterns discovered by mutual information were important while have not been identified by other approaches. Thirdly, the Meta-Network approach could also achieve best efficiencies among all methods been compared.

P61

Three-dimensional genome modeling based on ChIA-PET data

Przemyslaw Szalaj, Zhonghui Tang, Yijun Ruan and Dariusz Plewczynski

The spatial organization of the genome plays an important role in its functioning and is closely related to gene expression level, DNA replication and repair and others. Recent development of advanced chromosome conformation capture (3C) based methods such as Hi-C and ChIA-PET [1] allow to quantify the interaction frequency between distant genomic loci and to infer the 3D chromatin conformation.

Recently we developed 3D-NOME [2], a computational suite to analyze interactions data and to infer the 3D chromatin models based on the ChIA-PET data (available also through a web-based interface [3]). Our modeling is based on the main biological units of the genome organization, mainly topological domains, chromatin loops and anchors. Shortly, we use weak and non-specific interactions to guide the modeling on the low resolution, domain (megabase) scale, and then use strong, specific interactions to refine the model on the anchor (kilobase) scale. Our approach can be used to create both the population-average and ensemble structures.

To refine our modeling we made an extensive analysis of the underlying epigenetic states of domains, loops and anchors identified by ChIA-PET. We identified several distinct subpopulations of those units which may be related to different functions and activity states, demonstrating the complexity of chromatin interactions landscape.

P62

LASSIM - a network inference toolbox for genome-wide mechanistic modeling

Mika Gustafsson

Recent technological advancements has made time-resolved, quantitative, multi-omics data available for many model systems, which could be integrated for systems pharmacokinetic use. Here, we present large-scale simulation modeling (LASSIM), which are the first general mathematical tools for performing large-scale inference using mechanistically defined ordinary differential equations (ODE) for gene regulatory networks (GRNs). LASSIM integrates structural knowledge about regulatory interactions and nonlinear equations with multiple steady states and dynamic response expression datasets. The rationale behind LASSIM is that biological GRNs can be simplified using a limited subset of core genes that are assumed to regulate all other gene transcription events in the network. LASSIM models are built in two steps, where each step can integrate multiple data-types, and the method is implemented as a general-purpose toolbox using the PyGMo Python package for making the most of multi-core computers and high performance clusters, and is available at gitlab. First, LASSIM infers a non-linear ODE system of the pre-specified core genes. Second, LASSIM optimizes the parameters that models the regulation of peripheral genes by core-system genes in parallel. We show the usefulness of this method by applying LASSIM to infer a large-scale non-linear model of naive Th2 differentiation, by integrating Th2 specific bindings, time-series and six public and six novel siRNA-mediated knock-down experiments. ChIP-seq showed significant overlap for all tested TFs. Next, we performed novel time-series measurements of total T-cells during differentiation towards Th2 and verified that our LASSIM model could monitor those data significantly better than comparable models that used the same Th2 bindings. In summary, the LASSIM toolbox opens the door to a new type of model-based data analysis that combines the strengths of reliable mechanistic models with truly systems-level data. We exemplify this by inferring the first mechanistically motivated genome-wide model of the Th2 transcription regulatory system, which plays an important role in immune related diseases.

P63

Combinatorial and Probabilistic Aspects of the Multiple RNA Interaction Problem

Saad Mneimneh[1] **and Syed Ali Ahmedh**[2]

[1]Hunter College City University of New York, New York, USA
saad@hunter.cuny.edu
[2]The Graduate Center City University of New York, New York, USA
sahmed3@gradcenter.cuny.edu

The interaction of multiple RNAs poses a relatively new computational problem when compared to its pairwise counterpart. Our recent formulation of this problem was based on a combinatorial optimization called *Pegs and Rubber Bands*. Whether pairwise or multiple, the "optimal"solution is typically driven by an energy-minimization-like approach, which may not entirely capture the interaction structure. Moreover, the actual structure may not be unique. Therefore, alternative suboptimal solutions are needed.

We extend our formulation for the Multiple RNA Interaction problem on a path, where RNAs interact sequentially, to handle more elaborate interaction patterns using bipartite graphs, including cycle and star interactions. We also integrate the combinatorial optimization techniques with an approach based on Gibbs sampling and MCMC, to efficiently generate a reasonable number of suboptimal solutions. In addition, when viable structures are far from optimal, we show how exploring dependence among different parts of their interactions can boost their candidacy for the sampling algorithm. Such dependence can arise, for instance, when the breaking of a stem in the folded structure of a given RNA can cause the two regions of the stem, which are now seemingly separate, to simultaneously favor to interact with other RNAs. Finally, by clustering the solutions, we identify few representative clusters that are distinct enough to suggest possible alternative structures.

We successfully apply the developed techniques to known RNA interactions, such as the CopA-CopT pair and ribozyme complexes of three and four RNAs, where the correct structure is not computationally optimal (for instance, due to reversible kissing loops in CopA-CopT), and other examples of multiple RNAs; for instance, a four RNA complex in the spliceosome of yeast, where structural variation has been reported.

P64

GI-Cluster: detecting genomic islands in newly sequenced microbial genomes via consensus clustering on multiple features

Bingxin Lu, Hon Wai Leong

Department of Computer Science, National University of Singapore, Singapore

Lateral gene transfer (LGT), the transfer of genetic materials between two reproductively isolated organisms, is an important process in evolution. A large continuous genomic region acquired by LGT is often called a genomic island (GI). GIs can promote microbial genome evolution and adaptation of microbes to environments. They may also contain genes involved in pathogenesis and antibiotic resistance. Thus, the accurate inference of GIs is important for both evolutionary study and medical research.

Many computational methods have been developed to predict GIs in microbial genomes. However, their precision and recall are still not high enough. To get a more accurate prediction, it is necessary to combine multiple GI-associated evidences. Given that there are still no reliable golden datasets for GIs, it is tempting to use unsupervised machine learning methods to detect GIs from newly sequenced genomes.

To address this need, we developed a method called GI-Cluster, which provides a novel and effective way to integrate multiple GI-related features via database search and consensus clustering. GI-Cluster also provides tools to visualize the distribution of predicted GIs along a microbial genome and the features of each GI.

GI-Cluster does not require training datasets, but it can still achieve comparable or better performance than supervised machine learning methods in the evaluation. The preliminary results suggest that it had higher recall and precision than programs with similar input on some datasets. GI-Cluster is applicable to not only an unannotated genome sequence but also initial predictions from programs with high recall and low precision.

P65

Investigation of canonical metabolic network of plants through topology-based analysis

Wanatsanan Siriwat[1], **Saowalak Kalapanulak**[1,2], **and Treenut Saithong**[1,2,*]

[1]Department of Computer Science, National University of Singapore, 1Systems Biology and Bioinformatics Research Laboratory, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok, 10150, Thailand

treenut.sai@kmutt.ac.th

[2]Bioinformatics and Systems Biology Program, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok, 10150, Thailand

Cell factory is a rational strategy to enhance yield capacity of a desired or high-value compound production in living organisms. Plant cells have advantage over many simple eukaryotic hosts in terms of being an autotropic organism, but engineering of plant metabolism often faces failure due to the complexity of the system. The characteristic of plant metabolism is, thus, the crucial knowledge for increasing the success rate of transferring metabolic engineering technology from unicellular to more sophisticated multi-cellular organisms. Here, the metabolism of plants was investigated based on topological analysis, in which global topology of the metabolic network and hub metabolites were determined. The metabolite graphs of cassava, rice, Arabidopsis, and maize showed that the metabolic networks of plants follow the scale-free properties. The average path length (9.57) and diameters (45-49) of the networks are in comparable range with the other eukaryotes. The top 15 hub metabolites are mostly related to the primary metabolism and up to 66 percent of which (12/18) are particular identified in plant metabolism, and not reported in board 80 organisms. The results suggested that the global structure of plant metabolism may be not significantly evolved from simple eukaryotes, but the interaction of the constituent pathways and reactions that underlines the metabolic regulation is distinct and may cause different metabolic behavior of plants.

P66

Mining the Proteome of Fusobacterium nucleatum subsp. nucleatum ATCC 25586 for Potential Therapeutics Discovery: An In Silico Approach

Abdul Musaweer Habib, Md. Saiful Islam, Md. Sohel, Md. Habibul Hasan Mazumder, Mohd. Omar Faruk Sikder, Shah Md. Shahik*

The plethora of genome sequence information of bacteria in recent times has ushered in many novel strategies for antibacterial drug discovery and facilitated medical science to take up the challenge of the increasing resistance of pathogenic bacteria to current antibiotics. In this study, we adopted subtractive genomics approach to analyze the whole genome sequence of the Fusobacterium nucleatum, a human oral pathogen having association with colorectal cancer. Our study divulged 1,499 proteins of F. nucleatum, which have no homolog's in human genome. These proteins were subjected to screening further by using the Database of Essential Genes (DEG) that resulted in the identification of 32 vitally important proteins for the bacterium. Subsequent analysis of the identified pivotal proteins, using the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automated Annotation Server (KAAS) resulted in sorting 3 key enzymes of F. nucleatum that may be good candidates as potential drug targets, since they are unique for the bacterium and absent in humans. In addition, we have demonstrated the three dimensional structure of these three proteins. Finally, determination of ligand binding sites of the 2 key proteins as well as screening for functional inhibitors that best fitted with the ligands sites were conducted to discover effective novel therapeutic compounds against F. nucleatum.

P67

Classification of promoter and enhancer pairs based on expression patterns for building a predictive model

Abhishek Das, Subhadeep Das, Samrat Ghosh, Sucheta Tripathy*

Structural Biology and Bioinformatics Division, CSIR - Indian Institute of Chemical Biology, Kolkata, India 700032
`tsucheta@gmail.com, tsucheta@iicb.res.in`

Enhancers, “the non-coding region”of DNA are responsible for regulating transcription of its interacting genes. As different region acts as enhancer at different cell lineage, so is the difference in expressions and cellular functions. To study this difference in the regulations and number of interactions of enhancers with genes in different cell lines, three major cell-lines Gm12878, K562 and H1-hesc were chosen. Using CAGE data (downloaded from ENCODE database) of all three cell-lines, TPM count of TSS were calculated using cageR package. Box plot suggested that chromosome number 2, 6, 9, 13, 18, 20 and X of K562 cell lines showed significant difference in the distribution as compared to Gm12878 and H1hesc. And chromosome number 8, 12, 14, 16, 21 of K562 and H1hesc shows relatively same distribution when compared with Gm12878. The above results suggested that difference in expression leads to different cell lineage and functions. It also suggest that specific chromosomes and genes involved in disease manifestations. Further, edgeR package was used to calculate the differential expressed TSS between Gm12878 and K562 which was then mapped with promoter and enhancer cis-Interactions data taken from DENdb and 4D genome database. After mapping 4 major classes were obtained; (a) Upregulated promoters and enhancers, (b) Downregulated promoters and enhancers, (c) Upregulated enhancers, downregulated promoters and (d) Downregulated enhancers, upregulated promoters. For better understanding we have also done the network analysis taking the interacting promoters and enhancers into account. Top 25 clusters were taken based on number of nodes. 13 out of 25 clusters showed correlations in their expressions. In future we want to build a model that can predict the expression of TSS and its effect on disease manifestations based on enhancer interaction.

P68

Identification of novel peptidic antibiotics by searching large-scale mass spectra against natural products databases

Alexander Shlemov, Alexey Gurevich, Alla Mikheenko, Anastasiia Abramova, Anton Korobeynikov, Hosein Mohimani and Pavel Pevzner

Peptidic natural products (PNPs) are important biomedical compounds that include many antibiotics and a variety of other bioactive peptides. Most PNPs contain non-proteinogenic amino acids and are cyclic or branch-cyclic. Although recent breakthroughs in PNP discovery raised the challenge of developing new algorithms for their analysis, identification of PNPs via database search of tandem mass spectra remains an open problem. To address this problem, natural product researchers use search strategies that identify novel PNPs, even in cases when the reference spectra are not present in spectral libraries.

Modern mass spectrometers produce millions of spectra per microbial samples. Metagenome mining approaches, on the other hand, predict millions of candidate PNPs per microbial sample. To investigate which of the candidate PNPs are expressed in the microbial sample, there is a need for computational tools capable of searching millions of mass spectra against millions of candidate PNPs allowing for small variations.

Here we propose a fast strategy to search millions of spectra against millions of PNPs and compute statistical significance (P-value) of a high score matches. Our method speeds up the search by selecting a small set of PNPs from the database that is guaranteed (with high probability) to contain a PNP that gave rise to the variant PNP that produced the spectrum. Moreover, P-values of peptide spectrum matches are approximated using a fast heuristic. Our method discovers an order of magnitude more known PNPs and PNP variants than previous efforts.

P69

Identifying molecular targets of drugs from gene transcriptional profiles

Heeju Noh, Ziyi Hua and Rudiyanto Gunawan

Identifying the molecular targets of a compound has great importance in drug discovery, especially for understanding the mechanism of drug actions as well as for drug repurposing. A number of computational algorithms, e.g. Detecting Mechanism of Action by Network Dysregulation (DeMAND), have been developed for this purpose using gene expression data from drug treatments of human cell lines. The-state-of-the-art methods including DeMAND, typically ignore the dynamical changes in the gene expression (e.g. time-series profiles) and are often incapable of predicting the type of drug actions (e.g., inhibition or induction). In this work, we developed Systems Analysis and Learning for inferring Modifiers of Networks (SALMON), a novel strategy for predicting drug targets using a combination of gene network perturbation and upstream analysis. SALMON uses a prior information on protein-gene network, which is constructed using a combination of transcription factor binding sites, protein-protein interactions and our recent network inference algorithm called DeltaNeTS. To determine the targets of each drug, the proteins are scored based on the dysregulation of the protein-gene network in the drug treatment sample, more specifically using the deviation of the (log-fold changes of) gene expression and the regulatory modes of the protein-gene interactions. We tested the performance of SALMON using microarray expression data from DREAM/NCI compound synergy challenge, which comprise time-series gene expression of human lymphoma cells from 14 compound treatment experiments. The results demonstrated that SALMON could provide significantly more accurate target predictions than DeMAND.

P70

Modeling genome coverage and estimating genome length in metagenomics

Kui Hua and Xuegong Zhang

Microbes play important roles in human health and other environments. Combined with some computational analysis, shotgun metagenomics has made unprecedented achievements in uncovering the structure of microbial communities and their relationship with environmental factors. Due to the high complexity of microbial communities and inadequate sequencing depth, however, the fraction of the whole genomes represented in a metagenomic sequencing sample can vary a lot across communities and experiments. Here we concatenate genomes of different microbes in a community and raise an interesting question, "can we possibly estimate the total length of the concatenated genome using a metagenomic sample?"Similar problems have been well studied with powerful statistical models proposed to address these problems in single genomic data. However, things become quite difficult when it comes to metagenomics considering the complexity of metagenomic data. Apparently, it is an intractable problem without further constraints. We examine carefully what these constraints should be on both simulation data and HMP mock community data. The high unevenness of different genome coverage turns out to be a vital obstacle for the accurate estimation. Therefore we build a mixture model to describe the coverage across the concatenated genome. To guarantee the accuracy of the estimation, we remove components with average coverage lower than a threshold and only estimate the total length of the rest parts. We apply our model to real data set and some interesting observations are provided.

P71

Constructing prediction of drug-target interaction model using deep neural network approach

Jongsoo Keum, Hojung Nam*

School of Electrical Engineering and Computer Science (EECS), Gwangju Institute of Science and Technology (GIST), 261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, Republic of Korea

Identification of compound-target interactions helps not only develop novel drugs and repositioning of drugs but also understand mechanism of action of drug. However, time consuming and cost problem cannot be disregarded for identification of compound-target protein interactions. The computational screening methods can predict the interactions well in a reasonable time. In many in silico screening methods, similarity based methods which use similarity score of each compound and target protein as features and construct the prediction model using machine learning such as Support vector machine (SVM) show promising effect. However, the methods have some limitations especially when unseen data sets are predicted.

In this study, we constructed a deep-learning prediction model of compound-target interactions by using a deep learning approach. Deep learning approaches show the best performance of prediction studies such as image recognition and natural language processing. Although drug- target prediction methods which use general machine learning methods have quite good performance, deep neural networks concept gives an advantage to the prediction of compound- target protein interactions. In this work, we used compound structural properties and amino acid sequence as features, and constructed a classification model. Our model demonstrated high performance in the external validation with unseen data sets, the predicted result achieved more than 70% accuracy.

P72

RENT+: An Improved Method for Inferring Local Genealogical Trees from Haplotypes with Recombination

Sajad Mirzaei, Yufeng Wu

Dept. of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA
sajad@engr.uconn.edu, yufeng.wu@uconn.edu

Haplotypes from one or multiple related populations share a common genealogical history. If this shared genealogy can be inferred from haplotypes, it can be very useful for many population genetics problems. However, with the presence of recombination, the genealogical history of haplotypes is complex and cannot be represented by a single genealogical tree. Therefore, inference of genealogical history with recombination is much more challenging than the case of no recombination.

In this work, we present a new approach called RENT+ for the inference of local genealogical trees from haplotypes with the presence of recombination. RENT+ builds on a previous genealogy inference approach called RENT, which infers a set of related genealogical trees at different genomic positions. RENT+ represents a significant improvement over RENT in the sense that it is more effective in extracting information contained in the haplotype data about the underlying genealogy than RENT. The key components of RENT+ are several greatly enhanced genealogy inference rules. Through simulation, we show that RENT+ is more efficient and accurate than several existing genealogy inference methods. As an application, we apply RENT+ in the inference of population demographic history from haplotypes, which outperforms several existing methods.

RENT+ is implemented in Java, and is freely available for download from: https://github.com/SajadMirzaei/RentPlus.

P73

de novo assembly and haplotyping of Sweet Potato (Ipomoea Batatas [L.] Lam) genome

Mohammadhossein Moeinzadeh, Jun Yang and Martin Vingron

Although the sweet potato, Ipomoea batatas, is the seventh most important crop in the world and the fourth most significant in China, its genome has not yet been sequenced. The reason, at least in part, is that the genome has proven very difficult to assemble, being hexaploid and highly polymorphic; it has a presumptive composition of two B1 and four B2 component genomes (B1B1B2B2B2B2). In this work, we proposed a novel haplotyping method and applied it on the sweet potato genome. We have produced a half haplotype-resolved genome from 267Gb of paired-end sequence reads amounting to roughly 60-fold coverage. Then, we recruited the assembled haplotypes in order to improve the the connectivity, ordering and orientation of the scaffolds in the assembly.

P74

STELLS2: Fast and Accurate Coalescent-based Maximum Likelihood Inference of Species Trees from Gene Tree Topologies

Jingwen Pei, Yufeng Wu

Dept. of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA

jingwen.pei@uconn.edu, yufeng.wu@uconn.edu

It is well known that gene trees and species trees may have different topologies. One explanation is incomplete lineage sorting, which is commonly modeled by the coalescent process. In multispecies coalescent, a gene tree topology is observed with some probability (called the gene tree probability) for a given species tree. Gene tree probability is the main tool for the program STELLS, which finds maximum likelihood estimate of the species tree from the given gene tree topologies. One main drawback of the existing probabilistic species tree inference approaches such as STELLS is that they usually become slow when data size increases. Recently, several fast species tree inference methods have been developed, which can handle large data. However, these methods are often heuristic. In this paper, we present an algorithm (called STELLS2) for computing the gene tree probability much more efficiently than the original STELLS. The key idea of STELLS2 is taking some "shortcuts"during the computation and computing the gene tree probability approximately. We apply the STELLS2 algorithm in the species tree inference approach implemented in the original STELLS, which leads to a new maximum likelihood species tree inference method (also called STELLS2). Through simulation we demonstrate that STELLS2 is almost as accurate in species tree inference as the original STELLS. Also STELLS2 is usually more accurate than several existing methods when there is one allele per species, although STELLS2 is still slower than these methods. STELLS2 outperforms these existing methods significantly when there are multiple alleles per species.

P75

Single cell RNA-seq immunodynamics in peanut allergy

Xintong Chen, David Chiang, M. Cecilia Berlin and Bojan Losic

To understand food-specific T cell responses in the context of tolerance or allergy, we identified peanut-responsive CD4+ T cells by CD154 assay in peripheral blood of subjects with or without peanut allergy (PA) and performed single cell sequencing. Single cell RNA-seq data were available for 212 peanut-responsive CD154+ cells from 5 PA subjects and 122 antiCD3/CD28-activated CD154+ cells from 3 PA subjects. As an additional reference population, we collected 97 resting Tregs from freshly isolated PBMCs from PA or HC subjects, for a total of 431 cells. We observed substantial heterogeneity in the phenotype of peanut-responsive cells as identified by CD154 expression after exposure to peanut extract. Using an unsupervised approach we were able to cluster peanut-responsive CD154+ cells into three states independent of subjects and cell cycle. We identified a clearly pro-inflammatory Th2-associated state that's preferentially associated with peanut-activated T cells. When projecting the state associated genes to single cell network constructed from our data, we found key cytokines interact in highly differentiated Th2 cells in the allergic immune response to peanut. We reconstructed the full complementarity-determining region 3 (CDR3) directly from the single cell RNAseq data to characterize the TCR sequence of alpha and beta chains and observed clonal expansion only took place in peanut-responsive CD4+ T cell population. We were able to identify 4 clonal expansions each comprised of a pair of cells with identical alpha and beta TCR sequences from a single PA individual and that a key gene driving clonal expansion was PLA2G15, which results in a lymphoproliferative and autoimmune disorder when deleted.

P76

Breaking the Scalability Barrier for Core-genome Identity Computation

Chirag Jain, Luis M. Rodriguez, Alexander Dilthey, Adam Phillippy, Kostas Konstantinidis and Srinivas Aluru

With the increasing role of DNA sequencing and data-driven research in microbiology, core-genome identity (a.k.a. ANI) estimation is now a widely used computational technique used by microbiologists to study the evolutionary relationship of an unknown assembled microbial genome against the known taxonomy. This measure also serves as a robust measure for species demarcation in prokaryotes. However, due to the ongoing expansion of microbial reference genome databases and the known diversity since last 10 years, the alignment-based computational technique faces a scalability barrier.

We propose a new method called fast-ANI to compute core-genome identity of a query genome against a large reference database. This method utilizes our approximate sequence mapping algorithm (Jain et. al. RECOMB 2017) that models the sequence mutations using a Poisson distribution and estimates the sequence identity using the Jaccard similarity index. Thus, we bypass the need of an expensive alignment based method in this application at the cost of a practically insignificant estimation error. Core-genome identity is computed using the best bidirectional mappings between fragments of the query genome against the reference genomes. Using the underlying probabilistic model, we also estimate the error in our identity prediction.

Our experiments with real data show two major empirical advances over the original method. First, we demonstrate the capability to utilize large reference databases with thousands of bacterial genomes. Second, the computation time we achieve is about two orders of magnitude faster than the original method proposed by Goris et. al. (2007). This is achieved while maintaining the required accuracy in the identity estimates using both complete and incomplete draft genome assemblies.

P77

An improved approach for reconstructing consensus repeats from short sequence reads

Chong Chu, Jingwen Pei, Yufeng Wu

Computer Science & Engineering Dept., University of Connecticut, Storrs,CT, USA
{chong.chuJtngwen.pet,yufeng.wu}@uconn.edu

Repeat elements are important components of most eukaryotic genomes. Most existing tools for repeat analysis rely either on high quality reference genomes or existing repeat libraries. Repeat analysis is still difficult especially for species with highly repetitive or complex genomes which often do not have good reference genomes or annotated repeat libraries. Recently we develop a computational method called REPdenovo that construct consensus repeat sequences directly from short sequence reads, which outperforms an existing tool called RepARK. One major issue with REPdenovo is that it doesn't perform well for repeats with relatively high divergence rates or low copy numbers. In this paper, we present an improved approach for constructing consensus repeats directly from short reads. Comparing with the original REPdenovo, this improved approach uses more repeat-related k-mers. In addition, our new approach improves repeat assembly quality using a consensus-based k-mer processing method. We compare the performance of the new method with REPdenovo and RepARK on hiDUan and Arabidopsis thalia.na. short sequence data. The results show that our new method can assemble more complete repeats than REPdenovo (and also RepARK). We apply our new method on hummingbird which has no known repeat library, and construct many repeat elements that are validated using PacBio long reads. Many of these repeats are likely to be true repeats that are not in public repeat libraries. Our new approach has been implemented as part of the REPdenovo software package, which is available for download at `https://github.com/Reedwarbler/REPdenovo`.

P78

GAPPadder: A Sensitive Approach for Closing Gaps on Draft Genomes with Short Sequence Reads

Chong Chu*, Xin Li, Yufeng Wu

Dept. of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

Closing gaps in draft genomes is an important post processing step in genome assembly. It leads to more complete genomes, which benefits downstream genome analysis such as annotation and genotyping. Several tools have been developed for gap closing. However, these tools don't fully utilize the information contained in the sequence data. For example, while it is known that many gaps are caused by genomic repeats, existing tools often ignore many sequence reads that originate from a repeat-caused gap. In this paper, we propose a new approach called GAPPadder for gap closing. The main advantage of GAPPadder is that it uses more information in sequence data for gap closing. In particular, GAPPadder finds and uses reads that originate from repeate-caused gaps. We show that these repeat-associated reads are useful for gap closing, even though they are ignored by all existing tools. Other main features of GAPPadder include utilizing the information in sequence reads with different insert sizes and performing two-stage local assembly of gap sequences. We compare GAPPadder with GapCloser, GapFiller and Sealer on one bacterial genome, human chromosome 14 and the human whole genome with sequence reads of both short and long insert size. Empirical results show that GAPPadder can close more gaps than these existing tools. We also show GAPPadder is efficient in both time and memory usage.

P79

Differential Expression Analysis of Alternatively Polyadenylated 3'UTR using RNA-Seq

Chen Yang, Readman Chiu, Daniel MacMillan, Zhuyi Xue, Rene Warren and Inanc Birol

Alternative polyadenylation (APA) is a widespread RNA-processing mechanism that generates distinct 3' untranslated regions (UTRs) with different lengths and contributes to the complexity of transcriptome. 3' UTR-APA is well known to play a fundamental role in gene regulation, and have been observed in tumours. However, studies aimed at profiling APA in large cohort sequencing data have had limited scope. Here, we present an analysis pipeline designed to profile APA events using RNA-Seq reads. We have developed KLEAT, a post-processing tool for low FDR identification of cleavage sites from 3' UTRs assembled with Trans-ABySS. We applied KLEAT on 674 matched normal/tumour paired samples across 22 cancer types obtained from The Cancer Genome Atlas (TCGA). Based on reported cleavage sites for each tumour type, 3' UTRs are recovered using ensembl gene annotation and quantified using Salmon. Raw counts are filtered and normalized using 16 housekeeping genes. For genes with multiple 3' UTR regions, we perform differential expression analysis to identify regions of significance. Furthermore, genes showing clear preference for long or short 3' UTR isoforms are investigated using pathway analysis, demonstrating the potential clinical utility of KLEAT.

P80

The Cancer Genome Collaboratory

Christina K Yung[6], George L Mihaiescu[1], Bob Tiernay[1], Junjun Zhang[1], Francois Gerthoffert[1], Andy Yang[1], Jared Baker[1], Guillaume Bourque[2], Paul C Boutros[1,3], Bartha M Knoppers[2], B.F. Francis Ouellette[1,3], S Cenk Sahinalp[4,7], Sohrab P Shah[5], Michelle D Brazas[1], Vincent Ferretti[1], Lincoln D Stein[1,3]

[1]Ontario Institute for Cancer Research, Toronto ON, Canada
[2]McGill University, Montreal QC, Canada
[3]University of Toronto, Toronto ON, Canada
[4]Simon Fraser University, Vancouver BC, Canada
[5]BC Cancer Agency, Vancouver BC, Canada
[6]University of Chicago, Chicago IL, USA
[7]Indiana University, Bloomington, IN, USA

The Cancer Genome Collaboratory is an academic compute cloud designed to enable computational research on the world's largest and most comprehensive cancer genome dataset, the International Cancer Genome Consortium (ICGC). A subproject of ICGC, the Pan-Cancer Analysis of Whole Genomes (PCAWG) alone has generated over 800TB of harmonized sequence alignments, variants and interpreted data from over 2,800 cancer patients. A dataset of this size requires months to download and significant resources to store and process. By making the ICGC data available in cloud compute form in the Collaboratory, researchers can bring their analysis methods to the cloud, yielding benefits from the high availability, scalability and economy.

To facilitate the computational analysis on the ICGC data, the Collaboratory has developed software solutions that are optimized for typical cancer genomics workloads, including well tested and accurate genome aligners and somatic variant calling pipelines. We have developed a simple to use, fast and secure, data transfer tool that imports genomic data from cloud object storage into the user's compute instances. Because a growing number of cancer datasets have restrictions on their storage locations, it is important to have software solutions that are interoperable across multiple cloud environments. We have successfully demonstrated interoperability across The Cancer Genome Atlas (TCGA) dataset hosted at University of Chicago's Bionimbus Protected Data Cloud, the ICGC dataset hosted at the Collaboratory, and ICGC datasets stored in the Amazon Web Services (AWS) S3 storage. Lastly, we have developed a non-intrusive user authorization system that allows the Collaboratory to authenticate against the ICGC Data Access Compliance Office (DACO) when researchers require access to controlled tier data. We anticipate that our software solutions will be implemented on additional commercial and academic clouds.

P81

Antibody repertoire construction for organisms with unknown germline V/J genes

Alexander Shlemov, Sergey Bankevich, Andrey Bzikadze and Yana Safonova

Construction of an antibody repertoire from Rep-seq reads is an important preliminary step in various immunological applications. There are at least three tools: MiXCR, pRESTO and IgReC (developed by our group) designed for this problem. Thereby, generally the problem seems to be solved. However, all these tools use input reads alignment against germline V/J genes for contamination/chimera filtering and read trimming. This step is trivial in case of known germline. At the same time, for many popular model organisms (hamster, camel, some mice strains, walrus, shark, etc) germline V/J genes have not been assembled or have been assembled with insufficient quality.

Thus, the problem of germline-less antibody repertoire construction is naturally arisen. We generalized our algorithm and implemented germline-less construction as an IgReC option. The poster presents the results of this research.

P82

Antibody repertoire construction from short HiSeq Rep-Seq reads

Sergey Bankevich, Alexander Shlemov, Andrey Bzikadze and Yana Safonova

During the years of cancer clinical studies a huge bank of tumor-specific RNA-seq datasets was accumulated. This data naturally includes immune-related sequences among the others. Such sequences could be of interest for researchers as they are probably related to cancer-specific B- and T-cells. Unfortunately, the vast majority of these datasets were sequenced by old Illumina HiSeq technology. Relatively short reads (typically, 100 nt) provided by this technology do not cover the whole variable region of an antibody. Hence, currently available tools for repertoire construction designed for Illumina MiSeq fail to produce decent results. Existing solutions addressing immune repertoire construction from short reads were designed for TCRs only and are not able to deal with clonally expanded antibodies. In this poster we present our results on construction of full-length repertoire construction from cancer RNA samples.

P83

Antibody repertoire construction from Ion Torrent Rep-seq reads

Sergey Bankevich, Alexander Shlemov, Andrey Bzikadze and Yana Safonova

Construction of an antibody repertoire is an important preliminary step in various immunological applications. Reconstruction of antibody repertoires from noisy immunosequencing reads (Rep-seq) removes sequencing and PCR errors and thus opens a possibility to analyze natural diversity and clonal features.

Most modern studies employ Illumina MiSeq technology for antibody sequencing. In 2015 our group developed IgRepertoireConstructor, a tool for antibody repertoire reconstruction from Illumina MiSeq reads and immunoproteogenomics validation. In 2016 we also released a tool for high-throughput data and another one for barcoded datasets.

However, Illumina MiSeq is not an exclusive option. In particular, it is possible to obtain Rep-seq datasets using Ion Torrent technology. Additionally, Ion Torrent sequencing is relatively cheap making this technology attractive to many laboratories and researchers. However, there is also a downside. In contrast to Illumina MiSeq, Ion Torrent sequencing introduces a solid number of indel errors. At the same time, existing repertoire construction solutions cannot handle reads with significant number of indels. Therefore, a novel algorithmic approach is needed. Using molecular barcoding a statistical model for sequencing errors can be constructed. Further it can be applied to correct errors in non-barcoded Rep-seq Ion Torrent data. The poster presents the results of this research.

P84

AntEvolo: a novel approach for clonal analysis of antibody repertoires

Andrey Bzikadze, Sergey Bankevich, Alexander Shlemov and Yana Safonova

An antibody repertoire is the result of a fast evolution that is achieved by various processes of the secondary diversification. As a result of multiple cycles of the secondary diversification, antibody repertoire represents a set of clonal lineages with various abundances. Each such lineage can be viewed as a clonal tree. Constructing clonal trees for antibody repertoire and analysis of disease-specific response are widely used in drug and vaccine design.

Despite the fact that the problem of evolutionary analysis is widely addressed, existing phylogenetics tools can not be directly applied to antibody repertoires. The main difference between standard and immune settings is in presence of intermediate species. Phylogenetic algorithms expect that all species present leafs of an evolutionary tree that is not correct for antibody repertoires. We developed AntEvolo algorithm for construction of clonal trees for antibody repertoires.

To compute a clonal tree for an antibody repertoire, one needs, in particular, to evaluate the direction of its edges. This problem can be approached using a statistical model for the somatic hypermutagenesis. Several studies were conducted to construct a SHM model (Yaari et al. (2013), Elhanati et al. (2015)). We propose a novel approach based on Yaari et al. that overcomes the flaws of the original model and additionally considers the existence of parameters fluctuations between individuals by design. We additionally enhanced the model by considering separately the FRs and the CDRs.

P85

Improving imputation by maximizing power

Yue Wu, Eleazar Eskin and Sriram Sankararaman

GWAS estimates the correlations between disease status and collected genetic variants. After estimating the correlations, we perform a statistical test to indicate if each of the estimated correlation is statistically significant. If the absolute value of the association statistics is smaller than the threshold we set, we reject the null hypothesis. In GWAS, imputation is used to aid the interpretation of a GWAS by predicting the association statistics at untyped variants. People perform imputation in two sets of ways. One way is that to impute the genotypes directly at the untyped variants, and then perform a statistical test. The other sets of ways is to utilize summary statistics and impute the association statistics directly. Having the predicted statistics of the untyped variants, people indicate significance if two ways. One way is to adjust the rejection threshold for the untyped SNPs, basing on the correlation between the untyped and typed variants, to control the family-wise error rate. The other way is to set the same reject threshold for all association statistics of the typed and untyped SNPs. In this paper, we compared different previous methods on imputation, show that the previous ways lose power. And we propose a new method, that can set a set of rejection threshold for both typed and untyped variants to maximize the power, and at the same time control the family-wise error rate to a desired level.

P86

High-quality, fast, and memory-efficient assembly of metagenomes and large genomes using Minia-pipeline

R. Chikhi, C. Deltel, G. Rizk, C. Lemaitre, P. Peterlongo, K. Sahlin, L. Arvestad, P. Medvedev, D. Lavenier

Gigabase-scale genome projects and large metagenomics studies have flourished thank to high-throughput sequencing technologies. However, performing de novo assembly of such data remains challenging. In the landscape of assembly software, the tools that produce high-quality assemblies typically require significant computational resources, while the fast and memory-efficient ones yield relatively inferior results. We present Minia-pipeline: an assembler that combines efficiency and high-quality results. Minia-pipeline is geared towards large datasets of metagenomes and eukaryotic genomes, and recently provided high-ranking assemblies in the Critical Assessment of Metagenomic Interpretation challenge. This poster describes the overall architecture of the pipeline, key algorithmic improvements, and demonstrate its effectiveness on both large genome and metagenome samples. The pipeline is modular and integrates several components: an error-correction module, a unitig assembly tool (BCALM 2, (Chikhi et al, 2016)), a multi-k contigs assembly module (Minia 3), and a scaffolder (BESST, (Sahlin et al, 2016)). Software is available at `https://github.com/GATB/gatb-minia-pipeline`

P87

Decentralized indexes for public genomic data

Luiz Carlos Irber Junior, C. Titus Brown and Tim Head

MinHashes can be used to estimate the similarity of two or more datasets. Expanding on the work pioneered by mash and extended in our library Sourmash, we calculated signatures for 412 thousand microbial reads datasets on the Sequence Read Archive. To be able to efficiently search for matches of these signatures in the RefSeq microbial genomes database we developed a new data structure based on Sequence Bloom Trees adapted for searching MinHash signatures (named SBTMH) and made it available to whoever wants to use it.

We explore how to encode the SBTMH structure as objects in a MerkleDAG and store it in IPFS (InterPlanetary File System), a decentralized system for data sharing, and how to load and spread the SBTMH indexes as well as the signatures calculated. The SBTMH behaves like a persistent data structure, where updates and new nodes can share parts of the structure of previous versions of the SBTMH. While this property can be used to avoid duplicating data, in this case it is important because it allows common nodes in trees to be shared, leading to increased availability and facilitating sharing and remixing of indexes and signatures.

This design can be extended to change how databases and archives (like the SRA) are offered and implemented, since users can collaborate by choosing to share subsets of the archive and spread the network bandwidth. More important, it avoids the central point of failure, while still allowing for curation and quality assurance of the data. We present a prototype showing how this can be achieved.

P88

Deciphering general characteristics of residues constituting allosteric communication paths

Andrzej Kloczkowski, Girik Malik and Anirban Banerji

There has been great interest and excitement in studying proteins' dynamics and structural changes involved in function to gain insights into the machinery of proteins. Due to difficulty in retaining atomic details in mode decomposition of large system dynamics, there have been significant computational challenges, which make the study of large system dynamics very complex.

Considering all the PDB annotated allosteric proteins (from ASD, the allosteric database) belonging to four different classes, this work has attempted to decipher certain consistent patterns present in the residues constituting the allosteric communication sub-system(ACSS). While the thermal fluctuations of hydrophobic residues in ACSSs were found to be significantly higher than those present in the non-ACSS part of the same proteins, thermal fluctuations recorded for the polar residues showed the opposite trend.

While the basic residues and hydroxyl residues were found to be slightly more predominant than the acidic residues and amide residues in ACSSs, hydrophobic residues were found extremely frequently in kinase ACSSs. Despite having different sequences and different lengths of ACSS, they were found to be structurally quite similar to each other - suggesting a preferred structural template for communication. ACSS structures recorded low RMSD and high Akaike Information Criterion(AIC) scores among themselves. While the ACSS networks for all the groups of allosteric proteins recorded low degree centrality and closeness centrality, the betweenness centrality magnitudes revealed nonuniform behavior. Though cliques and communities could be identified within the ACSS, maximal-common-subgraph considering all the ACSS could not be generated, primarily due to the diversity in the dataset. Barring one particular case, the entire ACSS for any class of allosteric proteins did not demonstrate "small world"behavior, though the sub-graphs of the ACSSs, in certain cases, were found to form small-world networks.

P89

Computational approaches to identify alternative splice variants as biomarkers of disease progression and drug resistance in Acute Myeloid Leukemia

Hayati Sh., Mitrofanova A.*

Department of Health Informatics, Rutgers School of Health Professions, Rutgers Biomedical and Health Sciences, Newark, New Jersey 07107, USA

The rate of mortality for Acute Myeloid Leukemia (AML) remains high despite recent advances in therapeutic targeting of the disease. Patients with AML often fail a conventional line of treatments, including induction therapy followed by bone marrow or stem cell transplantation, mainly due to chemotherapy resistance after achieved remission. Response to chemotherapy differs from patient to patient, possibly depending on genomic mechanisms uniquely altered in each patient. We have examined differential splicing patterns in blood samples from AML patients enrolled in TCGA project. To uncover splicing patterns implicated in therapy resistance, we compared chemotherapy-treated patients with poor and favorable response to therapy. Our analysis identified 112 differentially spliced skipped-exon events (FDR corrected p-value¡0.05), which span 37 genes, including CD44 that in addition to been involved in cell proliferation, migration and apoptosis, has been implicated in chemotherapy resistance in AML cell lines; RTN3, shown to have pro-apoptotic properties; a tumor suppressor HINT1; NONO, which is a known regulator of transcription and splicing; among others. Furthermore, we have tested the ability of differentially spliced genes to identify patients that are at risk to develop resistance and have demonstrated that a signature of the top 35% of significantly spliced genes was able to differentiate patients with poor and favorable chemotherapy response in an independent patient cohort (log-rank p=0.002). Our studies have identified alternatively spliced genes that can serve as biomarkers of chemotherapy resistance and potentially shed light on treatment choices for patients with AML.

P90

Molecular Insights of Peptide Folding Propensities for Cancer Drug Target Improvisation and Anti Microbial Peptide Library

Praveen Kumar[1], Murugesh Easwaran[2], Dr.C.Nilavamuthan[2], Dr.P.Shanmughavel[*]

[1]Computational Biology Lab, Department of Bioinformatics, Bharathiar University, Coimbatore, India

shanmughavel@buc.edu.in

Peptide plays characteristic role in Drug Discovery, Development and Drug Improvisation purpose. Experimentally, the peptide has different monomeric domains to occupy the structure which is inevitably making a peptide functional. Molecular cause of a peptide secondary structure profile has more than one function when it interacts with target protein or act as a linker or anchored one. Solvent property influence the parameter of a peptide structure eg: Temperature, pH, ratios of solvent volume, multimeric composition in the solvent. Before to design druggable molecule and to study drug likeliness disease, the properties of peptide and functions should be studied. But there is no proper computational annotations for peptide nature and behavior in different solvent. So this study targets to annotate functional monomeric peptides that could target ”dis-ease” and ”dis-order” profile. To probable short sequence of residues, Microbial Surface proteins were retrieved from Meta-Protein sequence Database. Library of Protein Sequence were subjected to variate differentially in length and functional features. There are 1809 peptide sequence were retrieved and analyzed features of is same. Totally six features were adopted from obtained peptide sequence such as length, residual classification, secondary structure properties, conformational features (co-ordinates), volumes of peptide structure and functional features. Output of this study concentrates on making a peptides confined to make use for Cancer Drug Discovery purpose and Library Construction.

P91

scTree: reconstructing complex cellular lineage trees from single-cell RNA-seq data

Parra RG*, Papadopoulos N, Ahumada-Arranz L, Soeding J

Quantitative and Computational Biology Group, Max Planck Institute for Biophysical Chemistry, Goettingen, Germany.
gonzalo.parra@mpibpc.mpg.de

Introduction:

Recent advances on single-cell sequencing techniques have facilitated the obtention of detailed expression profiles for thousands of cells at different stages on time course cellular processes such as embryonic development. If correctly treated, these expression profiles can help to dissect the intricate gene interaction networks that regulate differentiation from one progenitor cell type into different ones. Although different techniques like Linear Local Embedding, Independent Component Analysis and Diffusion Maps have shown to be successful in finding a low dimensional manifold, where cells are embedded, identification of a meaningful lineage tree topology structure, is still, not solved for complex topologies.

Results:

We have developed scTree, a tool to reconstruct lineage tree topologies from single cell RNA-seq data. Given a distribution of cells in a given manifold, like Diffusion Maps, we apply different heuristics to find the best tree structure that explains the dispersion of cells on the low dimensional space. Endpoints, branching points and their connectivity are detected and cells are assigned to the different branches. Additionally, functions for pseudotime assignment and detection of differentially expressed genes among the different branches in the tree are provided.

Discussion:

scTree is able to reconstruct complex tree topologies with multiple branching points in a robust and accurate way. With more and more complex datasets being periodically published, scTree constitutes a valuable tool for reconstructing the hierarchical structure of the cells being studied as well as to characterize their specific transcriptional signatures.

P92

IDICAP2: An Improved Tool for Integrating Drug Intervention Based on Cancer Panel

Eric Ho

By the time the sun ascends from the sea line tomorrow, cancer has claimed 1,632 lives, and upended the future of another 4,617 people in the United States on average, indiscriminately of age and gender. Cancer persistently maintains as the 2nd killer in the U.S., trailing only after heart disease. Although the number of cancer survivors has been rising steadily since the establishment of the National Cancer Act in 1971, the upward trajectory of cancer mortality remains untamed, in direct opposite to the declining mortality of heart disease. Combating cancer occupies the center stage of basic and translational biomedical research, leading to significant advancement of cancer diagnosis and treatment such as targeted cancer therapies, which specifically pinpoint cancer-driving proteins. Accurate diagnosis of cancer subtypes accompanied with targeted cancer therapies, generating sheer volume of therapeutic options, challenges physicians in formulating the best treatment that aims at patients' unique genetic mutations. Developing a web service that allows clinicians as well as patients to identify effective targeted therapies, both approved and experimental, would be helpful for both parties. We have developed an innovative web service called IDICAP (`http://idicap.lafayette.edu:8000/`), which stands for Integrated Drug Intervention for CAncer Panel. It uses genes that have been linked to a patient's cancer to search for targeted therapies in the market or under late phase of clinical trials in the proximity of the patient. Although IDICAP has proven to be successful in achieving the goal, selecting the true clinical trials of targeted therapy remains a challenge as the narrative exhibits high variability, e.g. a gene mentioned in the trial may not necessarily mean it is the molecular target. Our goal is to improve the sensitivity and specificity of IDICAP in retrieving clinical trials. Natural language processing (NLP) is the foremost technique we have used to boost the accuracy of searches by identifying latent connections among clinical trials through named entities i.e. gene symbols and drug names. Thus, we used trials from approved targeted therapies as positive samples and trials of non-targeted cancer therapies as negative samples to train our model followed by latent semantic analysis and its competitive method random indexing. Preliminary results were promising but there is room for further improvement. We hope IDICAP2 can unleash the power of biomedical information, aiding patients and clinicians in obtaining informative therapeutic options, that may ultimately hamper the rising trajectory of cancer death toll.

P93

Comparative extreme microbiomes exploration using bioinformatics methodologies

Pravin Dudhagara, Anjana Ghelani and Rajesh Patel

Next generation technologies allow quick and inexpensive analysis of metagenomes and revolutionizing microbial ecology studies of extreme environments. However, the effective analysis of the huge biological data is a key challenge of computational biology. Comparative analysis of the hot spring in the present study has been conducted. We have analyzed 4 metagenomes of Indian hot springs and compared with 5 publically available hot spring metagenomes from the Indian subcontinent using MG-RAST, EBI- EMBL and MetaGenAssist. The phylogenetic and functional profiling of the hot springs' microbial communities were performed to decode the hidden microbial ecosystem. All hot springs were found to dominate by bacteria and viruses with a significant presence of unassigned sequences. The dominating Firmicutes phylum in most of the hot spring was reported due to their thermo-tolerance nature. However, Taptapani, Arti, and Unkeshwar hot springs dominated by Bacteriodetes, Cyanobacteria, and Actinobacteria respectively indicated the geographic variation play the significant role in a distribution of microorganisms. Deinococcus-Thermus group in 5 host spring can be significantly used as a metagenomic biomarker to identify the radioactive substances. The co-occurrence associations between complex microbial communities at a taxonomic and functional level were also significant. Bacillus and Clostridium are the most abundant genus in most of the hot spring. The detection of stress response genes in hot spring metagenomes reveals the secrets hold by thermophiles for survival at elevated temperature. Uncharacterized genes detection in all metagenomes is the key indication towards the hidden unculturable microbes. The diversity of unculturable bacteria in all hot springs demonstrates a vast gene pool for biotechnological exploration and creates a major face for microbiologists to know the phylogenetic correlation and ecological implication of habitats. The result also enlightens the abundance, diversity, distribution and coexistence of microorganisms.

P94

PARP inhibition cytotoxicity or cytoprotection - inferring molecular pathways heterogeneity from transcriptional data

Weronika Wronowska, Krzysztof Gogolewski, Bogdan Lesyng and Anna Gambin

RNA microarrays and RNA-Seq are nowadays standard technologies to study the transcriptional activity of the cell. The obtained information about the activation of specific genes expression is inferred from the behaviour of the relatively large population of cells. Even assuming perfect sample homogeneity, different sub-populations of cells can exhibit diverse transcriptomic profiles. This is due to the random molecular diversity of the population, which leads to the activation of various pathways of signals transduction in particular cells.

Here, we propose a novel computational method based on matrix decomposition techniques to infer the proportion between cells that entered the cell death pathway (either apoptotic and necrotic) and those that active pathways related to proliferation as a reaction to experimental conditions. The method was applied to investigate the influence of C2 ceramide and poly(ADP-ribose) polymerase-1 inhibitor (PJ34) on the viability of neuroblastoma cells. Our results shown neurotoxic effect of ceremide which was slightly increased by PJ34. Our finding regarding toxicity of PJ34 comes out to be surprising however consistent with the results of tests carried out using the Ingenuity Pathway Analysis method. Currently we conduct a series of biological assays for further validation of our computational method.

Proposed novel method to dissect the effects of different cell populations in the sample can be easily adapted to investigate various molecular processes. The presented methodology complement standard approaches for inferring the regulatory network from transcriptomic data.

P95

BioCarian: A Search Engine for Performing Exploratory Searches of Biological Databases

Nazar Zaki and Chandana Tennakoon

There are a large number of web-based biological databases publicly available for scientists. In addition, many private databases are generated in the course of research projects. These databases are in a wide variety of formats. Generally, exploratory searches of these databases require customized solutions, especially when multiple databases are involved. This process is cumbersome for scientists who do not have a sufficient background in computer science. Web standards have evolved in recent times, and semantic web technologies are now available to interconnect diverse and heterogeneous sources of data. Therefore, the integration and querying of biological databases can be facilitated by semantic web techniques. The key idea is to describe data in the Resource Description Format (RDF) and to query them using the SPARQL language. In this poster, we present an efficient and user-friendly search engine that can be used to perform exploratory searches of biological databases. The search engine acts as an interface for running SPARQL queries in RDF databases. Databases in tabular format are first converted into RDF format so that SPARQL queries can be run directly on these databases. In addition, we also provide a graphical interface based on facets that can be used to create advanced SPARQL queries. The facet interface is more advanced and novel than conventional facets. It allows complex queries to be constructed and includes additional features such as allowing facet values to be ranked based on several criteria, visually indicating the relevance of a facet value, and presenting the most important facet values when a large number of choices are available. We have constructed a prototype database that combines public databases that have gene-level, protein-level, and disease-level information.

P96

A comprehensive variation analysis based on whole-genome of 62 Koreans and constructing Korean variome database called as HYKVB

Sunhye Park, Young Chan Park, Kiejung Park and In-Song Koh

Since the Human Genome Project had been done, many international projects have been conducted to construct genome-wide reference information and many human genetic variants were deciphered in various populations. We analyzed whole genomes of 62 individuals on Korean population to identify Korean-specific variation and construct a pilot reference database of Korean variation, where a comprehensive variant information is included such as SNVs, INDELs, Large deletions and CNVs.

We identified a total of 9,096,853 SNV, 1,000,193 INDELs, 4,112 Large deletions and 4,712 CNVs in 62 Koreans. We also compared our detected sequence variants with dbSNP (build 142) and found Korean-specific variants of 1,021,582 SNVs and 182,957 INDELs. Particularly, using genes with 5,415 nonsynonymous SNVs of all Korean-specific SNVs, we performed the gene-set enrichment analysis and discovered the relatively high associations with some metabolic pathways and cellular signal pathways in Korean population. Based on our results, we constructed Korean variome database, called as the HanYang Korean Variome Browser (HYKVB).

P97

Characterizing RNA-seq assembly graphs: when is enough, enough?

Camille Scott and C. Titus Brown

With sequencing experiments regularly reaching into the billions of fragments, assembly graphs have become a core feature of most extant assemblers. Traversals of these graphs yield images of the underlying sequence, and the assembled sequences that result are studied in downstream analyses. However, less studied are features of the assembly graph itself. Motivated by the observation that assembly graphs succinctly encode a universe of possible assemblies, we explore some fundamental features of assembly graphs from RNA-seq. Our work is guided by the question: how much sequence is needed to build a reliable and useful image of the underlying transcripts? Using a large body of RNA-seq experiments available through the Marine Microbial Eukaryotic Sequencing Project (MMETSP), we describe transcriptome assembly graphs though their component size and coverage distributions, and harness this information to present a model for streaming transcriptome assembly graph partitioning. This work moves us toward our goal of producing a fully streaming transcriptome assembler.

P99

model-based multiple variants test considering causal status

Jong Wha J. Joo, Farhad Hormozdiari, Jeahoon Sul, Eleazar Eskin*

Over the past decade, GWAS have successfully identified many variants associated with diseases and complex traits. The standard GWAS examine one variant at a time to identify causal variants. Recently, several studies have demonstrated that multiple causal variants may exist in a region. For those regions, the standard GWAS may be inappropriate due to its low statistical power. Alternatively, an approach considering multiple causal variants simultaneously may increase statistical power by aggregating the effects of causal variants in a region. Unfortunately, we do not know in advance which variants are causal and there are too many possible causal status. Recently, there are many progresses in fine mapping approaches that try to identify causal variants in a region. We extend the likelihood model of one of the fine mapping approaches, CAVIAR, and propose a new model-based multiple variants test considering causal status, referred to as MARS(Model-based Association test Reflecting causal Status).