

MEETING ABSTRACTS

Open Access

Highlights from the Seventh International Society for Computational Biology (ISCB) Student Council Symposium 2011

Vienna, Austria. 15 July 2011

Published: 21 November 2011

These abstracts are available online at <http://www.biomedcentral.com/1471-2105/12?issue=S11>

INTRODUCTION

A1

Highlights from the Student Council Symposium 2011 at the International Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology

Priscila Grynberg^{1*}, Thomas Abeel^{2,3}, Pedro Lopes⁴, Geoff Macintyre⁵, Lorena Pantano Rubiño⁶

¹Departamento de Biquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; ²VIB Department of Plant Systems Biology, Ghent University, Technologiepark 927, 9052 Zwijnaarde, Belgium; ³Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA; ⁴DETI/IEETA, University of Aveiro, Campus Universitário de Santiago, 3810 – 193 Aveiro, Portugal; ⁵NICTA Victoria Research Laboratory, The University of Melbourne, Melbourne, Victoria 3010, Australia; ⁶Department of Regulatory Genomics, Institut de Medicina Predictiva i Personalitzada del Càncer, Badalona, Catalonia, Spain

E-mail: chair@iscbsc.org

BMC Bioinformatics 2011, **12(Suppl 11):A1**

Introduction: The ISCB Student Council is a worldwide organization for students in bioinformatics and computational biology. Its main goals are the organization of events and the creation of networking opportunities for students. The main contribution of the SC is to nurture students' soft skills, such as teamwork and group dynamics, and leadership and cooperation skills, to improve their regular academic program. Since its inception, the Student Council has organized an annual student symposium for the benefit of the student community. This year, the Student Council Symposium was held in conjunction with the ISMB/ECCB conferences on July 15th. Over 85 delegates attended this edition of the Student Council Symposium. The event opened with a well-received scientific speed dating session, allowing the delegates to get to know each other before the sessions kicked off. The program featured three keynote lectures, two partner presentations, ten student oral presentations and an extensive poster session with over 50 posters.

We were honored to have three highly esteemed scientists deliver the keynote presentations. Dr. Chad Myers (University of Minnesota, Minneapolis, MN, USA) opened the scientific program with a presentation entitled: "Systems-Level Insights from Large-Scale Combinatorial Perturbation Experiments in Yeast". The afternoon session started with Dr. Ivet Bahar (University of Pittsburgh, Pittsburgh, PA, USA) presenting "Protein Dynamics: Learning from Computations and Experiments". The closing keynote lecture of the day was conducted by Dr. Curtis Huttenhower (Harvard School of Public Health, Boston, MA, USA) and was titled "Functional Metagenomics and the Human Microbiome".

This year we welcomed two talks by our industry and research partners. Dr. Izhak Haviv of Australia's Information and Communications Technology Research Center of Excellence (NICTA) and the Department of Pathology, University of Melbourne (Melbourne, Australia) gave an overview of the new efforts to reduce the processing time of algorithms used for genomic data analysis. Janna Hastings of the European Bioinformatics Institute (Cambridge, UK) provided insight into training and career opportunities for students and other early career scientists at EMBL-EBI.

During the symposium, the audience was treated to a presentation from three students taking part in the Student Council Internship Program (<http://www.iscbsc.org/internships>). This new initiative from the Student Council has resulted in four students from developing nations participating in internships of three to six months at two research laboratories in Europe, hosted by Dr. Burkhard Rost and Dr. Reinhard Schneider. The three internship students, Maina Bitar from Brazil, Mohd Rehan from India and Dedan Kinuthia Githae from Kenya, provided symposium attendees a snapshot of their experiences during their time as interns and the benefits of participating in the program.

Proceedings: This year we received almost one hundred submissions from students to present their work at the symposium. These submissions were peer-reviewed by 39 independent reviewers from whom the program committee selected 10 for oral presentation, of which 9 are included in this meeting report. An additional 50 abstracts were accepted for poster presentations and the three winners of the best poster awards also got the opportunity to have their abstract included in this report. The nine included oral presentations fall into three main fields of research: Sequence analysis, Bioinformatics tools, and Transcriptomics: gene expression analysis and data management. Each of these topics features a block of three or four presentations. Below is a short discussion of the presentations that are featured in the special proceedings issue.

1. Sequence analysis: The DNA, RNA and protein sequences generated from thousands of organisms and stored in public databases has facilitated the development of several research fields, including genome assembly and annotation, sequence alignment and polymorphism search, and protein structure prediction. Comparison of sequences provides interesting insights about the relationship between genotype and phenotype. A good example of this was presented during the symposium by Klotzbuecher *et al.* [1]. They showed a new method for determining which pairs of single nucleotide polymorphisms (SNPs) affect the flowering time of *Arabidopsis thaliana*. The method consisted of a two-step approach to determine pairs of interacting SNPs that are strongly associated with the phenotype. The first step used a branch-and-bound strategy to prune away insignificant pairs. The second step applied prior biological knowledge to reduce the search space.

Recently, the advent of new sequencing technologies has altered the bioinformatics research landscape considerably, pushing the boundaries

of what is considered high-throughput. Researchers are now able to sequence complete genomes, transcriptomes and exomes in a fast and cheap way. The large amount of data generated by such projects mean that new approaches and methods of analysis are required. As an example, Buske *et al.* [2] presented methods for analyzing sequence data from the exome of 1,000 individuals with autism in an attempt to discover associated genetic variants. The development of a specific analysis pipeline and tools for alignment, variant detection, and data visualization was demonstrated.

The advent of large scale sequencing technologies has also impacted evolutionary studies. Instead of a focus on single genes, or small groups of genes, it is now possible to sequence complete genomes from several samples or species, and study evolution from a global perspective. McElroy *et al.* [3] deep-sequenced the genome of *Pseudomonas aeruginosa* PAO1, a biofilm model of lung infection. Results showed that the bacteriophage, and not the bacterial genome, was undergoing diversification. To help with the analysis, the authors developed cutting-edge statistical techniques to reconstruct bacteriophage haplotypes.

2. Bioinformatics tools: Software tools and databases are becoming increasingly important for the analysis, visualization and comprehension of the molecular data. In this session, different topics were covered from genomic analysis to protein analysis.

Genome-wide association studies (GWAS) are a powerful approach to determine the association between variations from different individuals from the same species with specific phenotypes. GWAS analyses encounter limitations as a result of large datasets generated. Goudey *et al.* [4] presented a novel algorithm known as Optimal Pairwise Epistasis (OPE) for exhaustively examination of all pair-wise SNP interactions in GWAS. The robustness of the tool was validated through analysis of two independent datasets from Celiac patients. They also showed a 10-fold decrease in time consumed for the analysis, compared to other approaches.

A significant contribution to biological research has been the development of *in silico* prediction tools. They allow many experiments to be conducted initially *in silico*, saving time and money. Woodcroft *et al.* [5] demonstrated the development of an integrative bioinformatics predictor of protein sub-cellular localization in malaria. More than 700 publications stating the sub-cellular location of a protein were used to train and test the method. The method achieved an accuracy of approximately 60% on a seven class problem. To validate the *in silico* results, the localization of a number of proteins was verified experimentally.

Recent sequencing technologies require the development of new software suites to analyze the large amount of data generated in a quick and user-friendly way. Schultheiss *et al.* [6] showed a powerful and effective framework to analyze RNA-seq data in less than 20 minutes with a machine-learning workflow integrated in an easy-to-use Galaxy framework. The modular tool suite deals with different aspects of a typical next-generation analysis: short-read alignments, transcript identification/quantification and differential expression analysis.

3. Transcriptomics: gene expression analysis and data management:

Gene expression changes are one of the fastest and most versatile responses of an organism facing changes in their environment. The comprehension of this scenario is essential to understand an organism or cell adaptations and survival in response to stress.

Several techniques are available for measuring levels of gene expression, such as Expressed Sequence Tags (EST), Serial Analysis of Gene Expression (SAGE), microarrays, RNA-seq, and Cap Analysis of Gene Expression (CAGE). During the symposium, Franciscatto *et al.* [7] showed an application of some of these technologies. They used CAGE combined with high-throughput sequencing (deep-CAGE) to collect specific transcription start sites (TSS) and their expression levels in different areas of human aged brain. These were used to correlate expression with methylation and structural genomic variation.

The reliability of gene expression studies depends largely on the method of analysis, including solid statistical approaches. For tumor gene expression analysis studies, an extra issue must be considered: tumor samples, when extracted from solid cancers, also contain healthy tissues, and this can compromise the cancer gene expression pattern. A new computational method was developed by Deshwar *et al.* [8] to purify tumor gene expression profiles using reference samples of healthy tissue to model the real contribution of cancer samples.

Collections of comprehensive studies that generate large amounts of data usually result in two or more different sources of data that are

complementary each other. Databases are important tools to store, organize and utilize such data. Bovolenta *et al.* [9] presented the Human Transcriptional Regulation Interactions database (HTRIdb), an open-access database of experimentally validated interactions among human transcription factors and their respective target genes.

Poster award winners: This year, the SC awarded three poster presentations. Awards were decided by delegate voting. The third prize was awarded to Mrinal Mishra [10]. The authors analyzed the genetic network of five pancreatic cancer candidate genes using a Cytoscape plugin. They found that two important candidate genes interact with 21 neighbor genes, revealing the importance of a change in expression level of candidate genes in causing pancreatic cancer. The second prize was given to Emre Guney [11], who presented a new framework called GUILD (Genes Underlying Inheritance Linked Disorders). The idea was based on the application of methods to prioritize the relation between genes and disease to reveal pathways associated with the disorder. First place went to Benjamin Kwan [12] who presented a new numerical representation and novel thresholding technique for classifying short exon and intron sequences using discrete Fourier transform period-3 value.

Student Council activities at ISMB: 1. Student Council career activities:

During ISMB the Student Council organized a Career Session designed to provide students with information to help them make informed decisions about their future careers. The first half of the session saw Dr. Jaap Heringa, Director of the Centre for Integrative Bioinformatics at Vrije Universiteit Amsterdam, The Netherlands, give a talk on his career choices and outline the important factors to consider when pursuing a career in bioinformatics. Following Jaap's talk, a career discussion panel made up of five early career researchers, Dr. Nils Gehlenborg, Dr. Jeroen de Ridder, Dr. Yana Bromberg, Dr. Magali Michaut and Dr. Daniela Maisel along with Dr. Jaap Heringa took questions from the audience. The panelists gave valuable insights while answering interesting and at times provocative questions from the student audience.

Besides the Career Session in the conference program, the Student Council provided an enhanced job advertising process throughout the meeting in the exhibitors hall. Job advertisers could meet with job seekers through our interactive job posting board. As a result, many impromptu interviews were carried out at the Student Council booth during the conference with positive feedback from participants, both interviewers and interviewees.

2. Social activities: As networking opportunities are a major benefit of participating in a conference, the Student Council focused on organizing activities outside conference hours to continue discussions and transference of knowledge. The daily Social HQ saw up to 50 students meet up over dinner each night, providing the ability for students to network and discuss their scientific problems in more detail. The Student Council Social Event was also a highlight with approximately 70 students participating in a walking tour of Vienna followed by dinner and a quiz. The social events provide a fun and relaxed atmosphere for students to connect and discuss all things bioinformatics and more.

Conclusions: The number of abstract submissions for the Student Council Symposium is growing every year. This year we were once again able to organize a program with three high-quality student presentation sessions, a broad poster session and three invited keynotes. Coupled with the SC activities organized during the main conference, the Student Council Symposium at ISMB/ECCB is fast becoming the premier event for students in bioinformatics and computational biology.

The future: Next year, the Student Council will feature two major international events. First, the Student Council Symposium will be held together with ISMB 2012, in Long Beach, CA, USA. Second, The European Student Council Symposium will enrich ECCB 2012, in Basel, Switzerland. Further information regarding the Student Council, its events, internships and community, please visit <http://www.iscbsc.org>.

Acknowledgements: The success of an event the size of the ISCB Student Council Symposium depends on the commitment of many. We would like to thank ISCB Executive Administrator BJ Morrison McKay, ISMB 2011 conference organizer Steven Leard and ISCB Administrative Support Suzi Smith for their logistical support and invaluable advice. Furthermore, we like to thank the ISCB Board of Directors for their continued support of the ISCB Student Council and the symposium in particular.

We are also greatly indebted to ISMB/ECCB 2011 conference chairs Burkhard Rost, Michal Linial, Peter Schuster and Kurt Zatloukal for giving us the opportunity to have the Student Council Symposium 2011 in Vienna.

The Student Council would also like to thank our keynote speakers Chad Myers, Ivet Bahar and Curtis Huttenhower who are volunteering their time to contribute to the success of the symposium and to promote the next generation of computational biologists. We would like to thank everyone involved in the organization this year for their contribution. Furthermore, we would like to thank everyone on the program committee for their time and effort. We also thank the BMC Bioinformatics Editorial Office for their help in preparing this supplement.

We are extremely grateful for the financial support that we received from our sponsors. Without their help many of the exciting opportunities that we offer to the delegates at the Student Council Symposium would not have been possible. The Student Council Symposium was made possible through generous financial support from the European Bioinformatics Institute, Genome Canada, NICTA, Iowa State University, Netherlands Bioinformatics Centre, New England BioLabs, Oxford University Press, Swiss Institute of Bioinformatics, RSG Netherlands and BMC Bioinformatics. The organizers are also grateful to Oxford University Press for sponsoring the best poster and best presentation awards.

Finally, the Student Council Symposium would not be possible without the volunteer efforts of the many Student Council members who help in organization of the symposium. A big thank you is given to all Student Council members who made the Student Council Symposium 2011 such a success.

References

1. Klotzbücher K, Kobayashi Y, Shervashidze N, Stegle O, Müller-Myhsok B, Weigel D, Borgwardt K: **Efficient branch-and-bound techniques for two-locus association mapping.** *BMC Bioinformatics* 2011, **12(Suppl 11):A3.**
2. Buske O, Dzamba M, Foong J, Lau L, Fiume M, Marshall C, Walker S, Prasad A, Brudno M: **Variant detection and the Autism sequencing project.** *BMC Bioinformatics* 2011, **12(Suppl 11):A4.**
3. McElroy K, Luciani F, Hui J, Rice S, Thomas T: **Bacteriophage evolution drives *Pseudomonas aeruginosa* PAO1 biofilm diversification.** *BMC Bioinformatics* 2011, **12(Suppl 11):A2.**
4. Goudey B, Wang Q, Rawlinson D, Zarnegar A, Kikianty E, Markham J, Macintyre G, Abraham G, Stern L, Inouye M, et al: **Replication of epistatic DNA loci in two case-control GWAS studies using OPE algorithm.** *BMC Bioinformatics* 2011, **12(Suppl 11):A5.**
5. Woodcroft B, Radloff R, Yeoh L, Scanlon K-L, Doyle M, van Dooren G, McFadden G, Tonkin C, Speed T, Ralph S: **An integrative bioinformatic predictor of protein sub-cellular localisation in malaria.** *BMC Bioinformatics* 2011, **12(Suppl 11):A6.**
6. Schultheiss SJ, Jean G, Behr J, Drewe P, Görnitz N, Kahles A, Mudrakarta P, Sreedharan VT, Zeller G, Rättsch G: **Oqtans: a galaxy-integrated workflow for quantitative transcriptome analysis from NGS data.** *BMC Bioinformatics* 2011, **12(Suppl 2):A7.**
7. Francescato M, Pardo L, Rizzu P, Vitezic M, Simón-Sánchez J, Takahashi H, Daub C, Carninci P, Heutink P: **Profiling transcription initiation in human aged brain using deep-CAGE.** *BMC Bioinformatics* 2011, **12(Suppl 11):A8.**
8. Deshwar A, Quon G, Morris Q: **Computational purification of tumor gene expression data.** *BMC Bioinformatics* 2011, **12(Suppl 11):A9.**
9. Bovolenta LA, Acencio ML, Lemke N: **The development of an open-access database for human transcriptional regulation interactions.** *BMC Bioinformatics* 2011, **12(Suppl 11):A10.**
10. Mishra M, Kumar A: **Computational analysis of genetic network involved in pancreatic cancer in human.** *BMC Bioinformatics* 2011, **12(Suppl 11):A11.**
11. Guney E, Oliva B: **Toward PWAS: discovering pathways associated with human disorders.** *BMC Bioinformatics* 2011, **12(Suppl 11):A12.**
12. Kwan BYM, Kwan JYY, Kwan HK: **Spectral classification of short numerical exon and intron sequences.** *BMC Bioinformatics* 2011, **12(Suppl 11):A13.**

MEETING ABSTRACTS

A2

Bacteriophage evolution drives *Pseudomonas aeruginosa* PAO1 biofilm diversification

Kerensa McElroy^{1*}, Fabio Luciani², Janice Hui¹, Scott Rice¹, Torsten Thomas¹

¹Centre for Marine Bioinnovation, UNSW, Sydney, NSW 2052, Australia;

²Inflammatory Diseases Research Unit, UNSW, Sydney, NSW 2052, Australia

BMC Bioinformatics 2011, **12(Suppl 11):A2**

Background: *Pseudomonas aeruginosa* infection is the leading cause of death for Cystic Fibrosis patients. Antibiotic resistance is rife, possibly due to high colonising population diversity. Our lab has replicated phenotypic diversification in a *P. aeruginosa* PAO1 biofilm model of lung infection. To reveal underlying genetic variants, we deep-sequenced PAO1 biofilm samples. Our analysis demonstrates several techniques for extracting meaningful biological information from error-prone sequencing data.

Materials and methods: DNA was extracted from PAO1 biofilm samples harvested after four and 11 days of growth, and Illumina sequenced to >1000x coverage. Sequencing was also simulated from computer generated PAO1 haplotypes with our program GemSIM. GemSIM uses real data (e.g., the PhiX control) to generate run-specific error models, facilitating realistic simulation. After optimisation, the program VarScan (Koboldt, 2009 #48) accurately detected SNPs in simulated data with frequencies down to 5%. VarScan was then used to identify SNPs with frequency $\geq 5\%$ in the biofilm data.

In high diversity regions, haplotypes were reconstructed using bayesian statistical techniques implemented in the program ShoRAH (Zagordi, 2010 #49), and validated through analysis of individual reads. (ShoRAH's error-correction algorithm can accurately identify variants in high diversity areas with frequencies down to 0.1%.)

Results and conclusions: Surprisingly, the PAO1 genome contained only two SNPs, with frequencies around 10%. Both were within a large hypothetical outer-membrane protein postulated to be involved in biofilm formation and antibiotic resistance. Both SNPs were silent; however converted rare codons to more common ones, potentially increasing expression. These SNPs may reflect early biofilm lifestyle adaptation.

In contrast to the negligible genetic diversity of PAO1, its associated bacteriophage Pf4 revealed ongoing diversification, characterised by an increase in Shannon's entropy between days four and 11 and an explosion in phage population size. All phage SNPs were within or upstream from the putative Repressor C gene. This gene is implicated in Pf4 superinfectivity, which results in loss of host resistance and conversion from a lysogenic form to a lethal, lytic lifecycle.

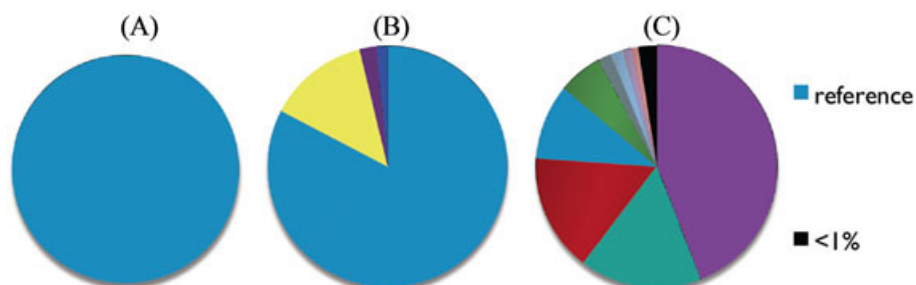


Figure 1(abstract A2) Relative phage haplotype frequencies in PAO1 samples, (A) planktonic, (B) biofilm 4 days, (C) biofilm 11 days.

In total, nine Pf4 haplotypes with frequencies > 1% emerged by day 11, while the original haplotype dropped to 9% (Fig. 1). These results suggest superinfective phage haplotype emergence drives diversification within PAO1 biofilms.

References

1. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L: **VarScan: variant detection in massively parallel sequencing of individual and pooled samples.** *Bioinformatics* 2009, **25**:2283-2285.
2. Zagordi O, Klein R, Däumler M, Beerenwinkel N: **Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies.** *Nucleic Acids Res* 2010, **38**:7400-7409.

A3

Efficient branch-and-bound techniques for two-locus association mapping

Karin Klotzbücher^{1,4*}, Yasushi Kobayashi², Nino Shervashidze¹, Oliver Stegle¹, Bertram Müller-Miyhok³, Detlef Weigel², Karsten Borgwardt¹

¹Machine Learning & Computational Biology Research Group, MPIs Tübingen, Tübingen, Germany; ²Max Planck Institute for Developmental Biology, Tübingen, Germany; ³Max Planck Institute for Psychiatry, Munich, Germany; ⁴Zentrum für Bioinformatik, Universität Tübingen, Tübingen, Germany
E-mail: karinklbr@gmx.de

BMC Bioinformatics 2011, **12**(Suppl 11):A3

Background: In this project we want to determine pairs of single nucleotide polymorphisms (SNPs) which have a statistically significant effect on the phenotypic variation of the flowering time of *Arabidopsis thaliana*.

Material and methods: For a large-scale dataset of over 200,000 SNPs from about 200 individuals together with several phenotypes, published by Atwell et al. [1], we develop efficient methods to find pairs of SNPs which are strongly associated with the phenotype. As an exhaustive search of all possible combinations of interacting SNPs is often unfeasible, even when only considering pairs of interacting SNPs, the challenge is to find methods which avoid an exhaustive search but can still guarantee to find the causal pair. We propose two distinct approaches to efficiently determine the top-scoring pairs of SNPs.

Results and conclusions: In the first approach we employ a branch-and-bound strategy to reduce the search space by pruning insignificant pairs of SNPs. Based on this branch-and-bound strategy we develop the two methods fastHSIC and COAT, which use as association measures the Hilbert-Schmidt Independence Criterion (HSIC) [2] and Pearson's correlation coefficient, respectively. The key idea is that we are able to bound the association scores of pairs of SNPs for both methods based only on the association score of one of the SNPs of the pair.

In our second approach we use prior biological knowledge to select a much smaller subset of candidate genes which, according to other findings, affect the flowering time of *Arabidopsis thaliana*. These candidate genes and interactions between them make up a network of 1,452 nodes or genes and 938 edges or gene-gene interactions, and allow us to select a subset of SNPs that lie within or in close proximity to the genes of the network.

Empirical evaluation of our own as well as traditional methods on the original and the reduced dataset shows that both our approaches can greatly reduce the runtime.

References

1. Atwell S, Huang YS, Vilhjalmsón BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, et al: **Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines.** *Nature* 2010, **465**(7298):627-631.
2. Gretton A, Bousquet O, Smola A, Schölkopf B: **Measuring statistical dependence with Hilbert-Schmidt Norms.** Proceedings of the International Conference on Algorithmic Learning Theory. *J Gen Virol* 2005, **1**:63-78.

A4

Variant detection and the Autism sequencing project

Orion Buske^{1*}, Misko Dzamba¹, Justin Foong², Lynette Lau², Marc Fiume¹, Christian Marshall², Susan Walker², Aparna Prasad², Michael Budrod^{1,2,3}

¹Department of Computer Science, University of Toronto, Toronto, Canada; ²The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada; ³Donnelly Centre, University of Toronto, Toronto, Canada

BMC Bioinformatics 2011, **12**(Suppl 11):A4

Background: Early detection of autism can improve the quality of life of affected individuals [1]. Qualitative screening methods continue to improve, but still suffer from low sensitivity despite increasing specificity [2,3]. In collaboration with the Hospital for Sick Children, we are sequencing the exomes of 1 000 individuals with autism in order to discover genetic variants associated with the disorder. Discovery of associated variants can lead to earlier diagnosis and treatment.

Materials and methods: We will present our current sequencing and analysis pipeline, from SureSelect exome capture and SOLiD sequencing through Sanger validation of predicted harmful variants, along with tools we have developed for color-space-aware alignment, variant detection, and visualization of next-generation sequencing data.

Color-space sequencing provides a tradeoff between enhanced ability to distinguish Single Nucleotide Variants (SNVs) from sequencing errors at the price of a higher sequencing error rate versus traditional letter-space sequencing. This technology has the potential to provide higher accuracy at lower cost, but opens new computational challenges that need to be addressed.

Results and conclusions: We have sequenced over 70 individuals so far at approximately 30x mean coverage, have found and validated several interesting non-synonymous Single Nucleotide Variants (SNVs), and have identified a number of potential de novo non-synonymous mutations. After filtering, we are identifying an average of over 17 000 non-synonymous SNVs per individual, of which over 11 000 are novel to dbSNP. We also find that support from both strands is more informative than total depth of coverage for predicting SNVs from high-throughput sequencing data. This is of considerable importance in exome capture data, since only a small region at each probe captures sequence from both strands.

References

1. Lord C, McGee JP: **Committee on educational interventions for children with autism, national research council: conclusions and recommendations.** *Educating Children with Autism* Washington, DC: Press NA 2001, 211-230.
2. Baird G, Charman T, Baron-Cohen S, Cox A, Swettenham J, Wheelwright S, Drew A: **A screening instrument for autism at 18 months of age: a 6-year follow-up study.** *J Am Acad Child Adolesc Psychiatry* 2000, **39**(6):694-702.
3. Robins DL, Fein D, Barton ML, Green JA: **The modified checklist for autism in toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders.** *J Autism Dev Disord* 2001, **31**(2):131-144.

A5

Replication of epistatic DNA loci in two case-control GWAS studies using OPE algorithm

Benjamin Goudey^{1,2*}, Qiao Wang², Dave Rawlinson², Armita Zarnegar², Eder Kikianty², John Markham², Geoff Macintyre^{1,2}, Gad Abraham^{1,2}, Linda Stern¹, Michael Inouye^{3,2}, Izhak Haviv^{2,4}, Adam Kowalczyk²

¹Department of Software Engineering and Computer Science, The University of Melbourne, Parkville, Victoria 3010, Australia; ²National ICT Australia (NICTA) Victoria Research Laboratories, The University of Melbourne, Parkville, Victoria 3010, Australia; ³The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3050, Australia; ⁴Baker IDI Heart and Diabetes Institute, Melbourne, Victoria 3004, Australia; ⁵Department of Medical Biology, University of Melbourne, Parkville, Victoria 3010, Australia
E-mail: bvgoudey@csse.unimelb.edu.au

BMC Bioinformatics 2011, **12**(Suppl 11):A5

Background: One of the limiting factors of current genome-wide association studies (GWAS) is the inability of current methods to comprehensively examine SNP interactions for a reasonable sized dataset. It is hypothesised that this limitation is one of the reasons that GWAS studies have not been able to have a greater impact [1,2]. Many current methods for handling interactions are computationally expensive and do not scale to entire studies. Those methods that do scale often achieve this by pruning their datasets in some manner. This is commonly done by considering only those SNPs that show strong marginal effects, despite the fact that a strongly interacting pair may consist of SNPs with low effects individually.

Material and methods: In this presentation, we validate the robustness of a novel algorithm known as Optimal Pairwise Epistasis (OPE) for exhaustively examining all pairwise interactions in GWAS data. This

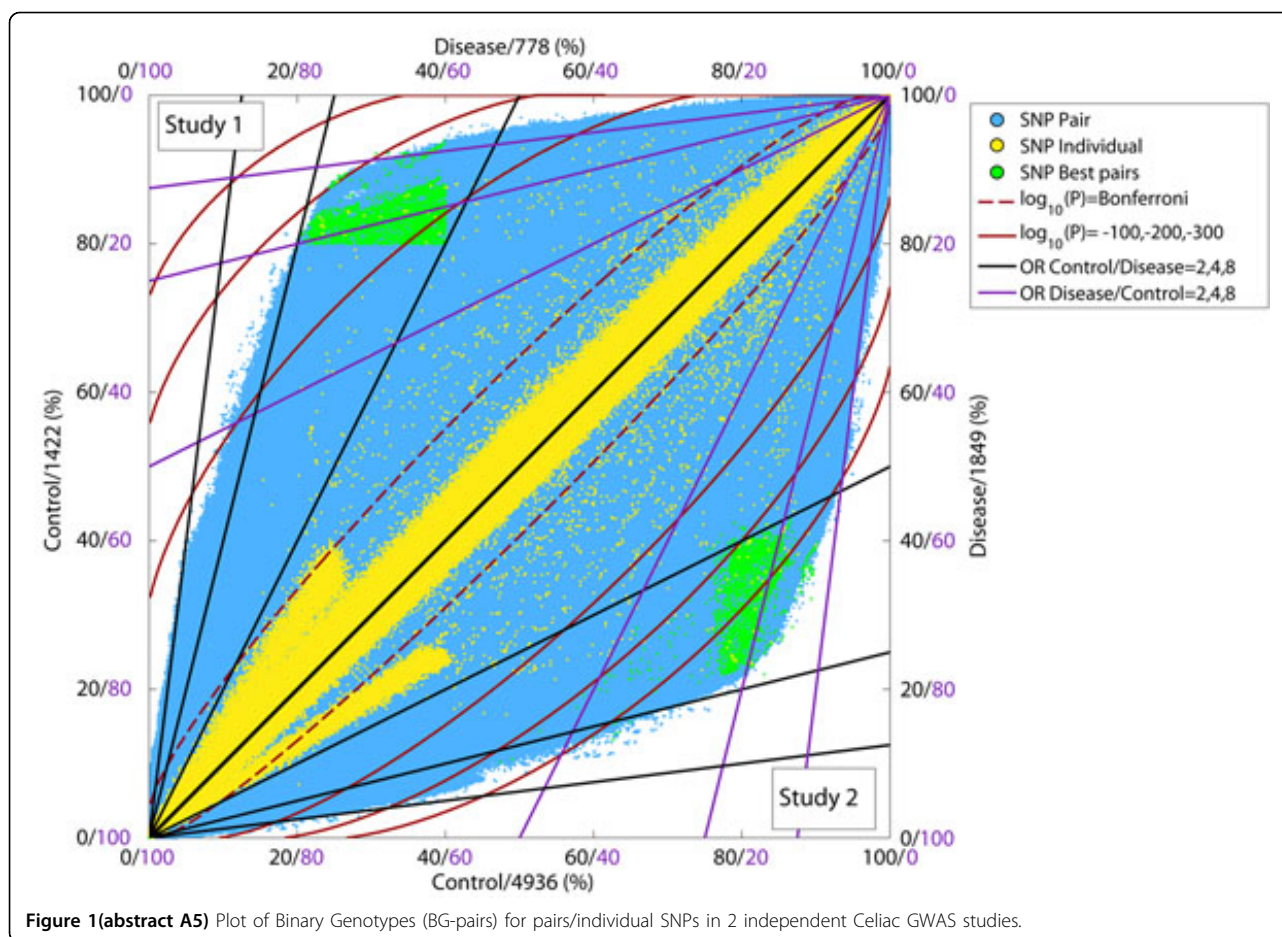


Figure 1 (abstract A5) Plot of Binary Genotypes (BG-pairs) for pairs/individual SNPs in 2 independent Celiac GWAS studies.

method is based on the systematic evaluation of “binary genotype pairs” (BG-pairs), i.e. the pairs of complementary binary classification of genotype calls for an individual SNP, or a pair of SNPs. We can quantify the discrimination potential of BG-pairs using a family of statistics based on odds ratios.

Results and conclusion: The approach is computationally efficient: the dataset reported here as Study 1 (consisting of ~310K SNPs and 2200 samples [3]) takes 12 hour to process on a single CPU (compared to 149 hours of the recent BOOST algorithm [4]). The method can be highly parallelised with a recent GPU implementation reducing this processing time to less than 15 minutes.

We have tested our approach over 2 independent GWAS studies of Celiac disease: the first (Study 1 mentioned above, [3]) with 778/1422 and the second (Study 2, [5]) with 1849/4936 of case/control samples, respectively. Each point in the figure 1 below shows the observed frequency of the BG carriers for the case and control subpopulations: in blue for a pair of SNPs or in yellow for an individual SNP. Every BG-pair can be evaluated with respect to the two sets of axes labels: purple labels for the protective BG and black labels for the risk BG. The resulting figure shows both studies related by symmetry in the main diagonal and indicates replication of results across studies. We emphasise the replicability of our approach by showing in green the same subset of SNP pairs in both studies. We also show in red contours for p-values and plot in black / purple solid diagonal lines to indicate different odds ratios.

References

1. Cordell HJ: Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009, **10**:392-404.
2. Moore JH, Asselbergs FW, Williams SM: Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 2010, **26**:445-455.
3. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MC, Bethel G, Holmes GK, et al: A genome-wide

association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* 2007, **39**:827-829.

4. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W: BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* 2010, **87**:325-340.
5. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adány R, Aromaa A, et al: Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010, **42**:295-302.

A6

An integrative bioinformatic predictor of protein sub-cellular localisation in malaria

Ben J Woodcroft^{1,2}, Robert Radloff^{1,3}, Lee M Yeoh^{1,4}, Kristie-Lee Scanlon¹, Maria A Doyle^{1,5}, Giel G van Dooren⁴, Geoffrey I McFadden⁴, Christopher J Tonkin², Terence P Speed², Stuart A Ralph^{1*}

¹Department of Biochemistry & Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Australia; ²The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia; ³Institute for Biochemistry, University of Stuttgart, Stuttgart, Germany; ⁴Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Melbourne, Australia; ⁵Bioinformatics Core Facility, Peter MacCallum Cancer Centre, Melbourne, Australia
 E-mail: saralph@unimelb.edu.au

BMC Bioinformatics 2011, **12**(Suppl 11):A6

Background: The malarial parasite *Plasmodium falciparum* remains a leading international cause of mortality, with almost a million deaths each year. Determination of protein sub-cellular localisation remains a challenge in *Plasmodium* parasites due to their evolutionary distance from

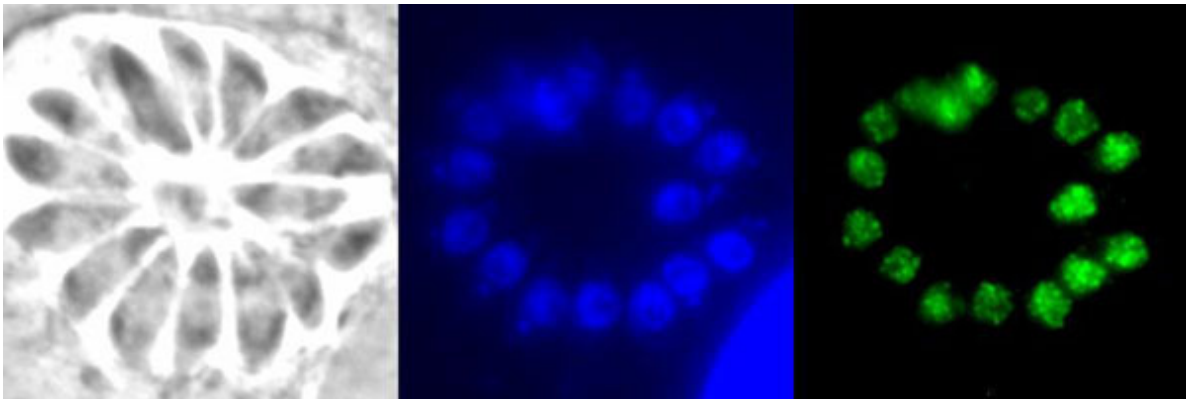


Figure 1 (abstract A6) Experimental confirmation of a Plasmarithm localisation prediction. A *P. falciparum* protein (PlasmoDB ID PFE0425w) was predicted to be nuclear, and here the HA tagged *T. gondii* orthologue (ToxoDB ID TGME49_044840) is shown exhibiting a nuclear localisation. The panels show bright field, Hoechst (nuclear) stained, and HA tagged TGME49_044840 images, respectively.

well-studied model organisms, and limited efficiency of appropriate molecular tools. However, abundant large scale systems biology information exist for several *Plasmodium* species as well as other apicomplexan parasites, including full genomic DNA sequences, plus data sets relating to the transcriptome, protein expression and interactions, polymorphisms and phyletic profiles. To date, most bioinformatic predictors of sub-cellular localisation use sequence information exclusively without consideration for other data sets.

Materials and methods: We developed the first global bioinformatic predictor of sub-cellular localisation in *Plasmodium falciparum* (called Plasmarithm) that predicts localisation for multiple cellular compartments using a variety of post-genomic information types.

Results and conclusions: We identified several non-sequence data types that are predictive of localisation, including phyletic distribution and transcript abundance at specific life stages. We performed a comprehensive literature survey of the phylum Apicomplexa to construct a database of >850 recorded protein localisations curated from >700 separate publications. The database, called ApiLoc (freely available at <http://apiloc.bio21.unimelb.edu.au>), was used to improve the accuracy of our predictor. We achieved an overall accuracy of ~60% on a seven class problem, where a

number of the classes have not previously been predicted. To further validate these in-silico analyses, we have experimentally verified localisations of a number of hypothetical proteins in the related apicomplexan *Toxoplasma gondii* (Figure 1).

A7

Oqtans: a Galaxy-integrated workflow for quantitative transcriptome analysis from NGS Data

Sebastian J Schultheiss^{1*}, Géraldine Jean¹, Jonas Behr¹, Regina Bohnert¹, Philipp Drewe¹, Nico Görnitz^{1,2}, André Kahles¹, Pramod Mudrakarta¹, Vipin T Sreedharan¹, Georg Zeller^{1,3}, Gunnar Rätsch¹

¹Machine Learning in Biology Group, Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany; ²Department of Software Engineering and Theoretical Computer Science, Technical University Berlin, 10578 Berlin, Germany; ³Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
 E-mail: sebastian@bioweb.me

BMC Bioinformatics 2011, **12**(Suppl 11):A7

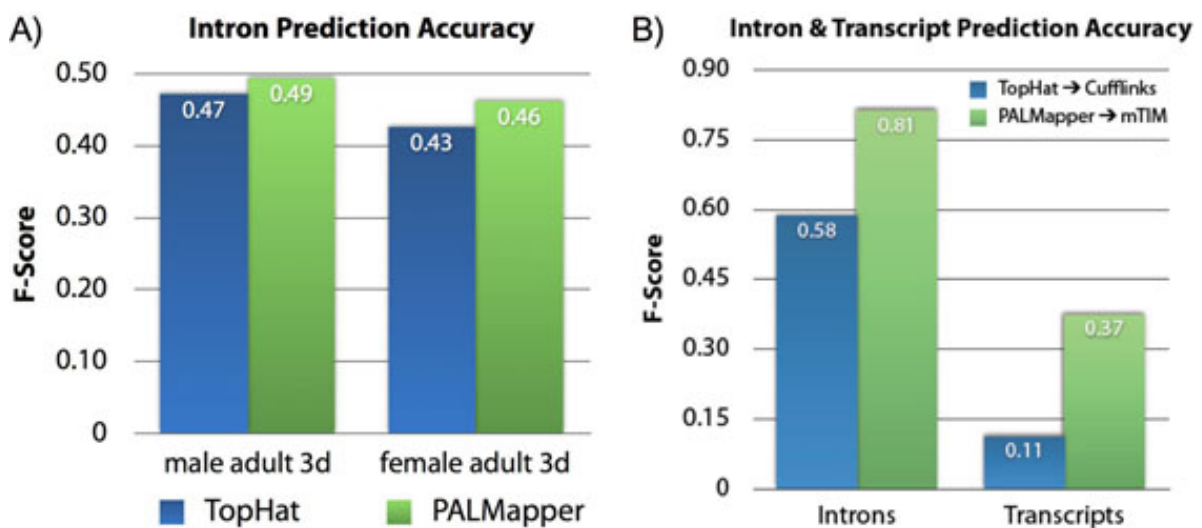


Figure 1 (abstract A7) A) Accuracy (F-score) of intron predictions in 3-day-old adults of *D. melanogaster* with aligners PALMapper (green) and TopHat (blue). B) Accuracy of intron predictions with the same aligners and transcript predictions with mTIM (green) and Cufflinks (blue) on *C. elegans* RNA-seq transcriptome data.

Background: The current revolution in sequencing technologies allows us to obtain a much more detailed picture of transcriptomes via RNA-Sequencing. We have developed the first integrative online platform, oqtans, for quantitatively analyzing RNA-Seq experiments. Our approach of providing a self-contained machine image with the accessible, transparent Galaxy framework [1] minimizes the risk of using a third-party web service for data analysis. These services often disappear a few years after publication and render results irreproducible [2]. With oqtans, bioinformatics becomes reproducible by providing analysis building blocks for a customized workflow of read mapping, transcript reconstruction and quantitation as well as differential expression analysis.

Method: Oqtans includes a comprehensive machine-learning-powered tool suite developed by the authors for NGS data analysis. PALMapper is a short-read mapper which efficiently computes both unspliced and spliced alignments at high accuracy by taking advantage of base quality information and computational splice site predictions [3]. mTIM is a transcript reconstruction method, which exploits features derived from RNA-seq read alignments and from computational splice site predictions to infer the exon-intron structure of the corresponding transcripts. rQuant is based on quadratic programming. It simultaneously estimates biases inherent in library preparation, sequencing, and read mapping, and accurately determines the abundances of given transcripts [4]. rDiff is a set of statistical test techniques that determine significant differences between two RNA-seq experiments to find differentially expressed regions with or without knowledge of transcripts.

Results: We compare predictions to the published annotation at the intron and transcript levels. The performance of read aligners is shown in Fig. 1A from D. melanogaster data, and transcript segmentation tools in Fig. 1B, on C. elegans. Our tools, PALMapper and mTIM, outperform TopHat [5] and Cufflinks [6]. Oqtans is available free and open-source, from <http://oqtans.org> as a virtual machine for cloud computing environments, and ready to use on our public compute cluster at <http://bioweb.me/mlb-galaxy>.

References

1. Goecks J, Nekrutenko A, Taylor J: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* 2010, **11**(8):R86.
2. Schultheiss SJ, Münch MC, Andreeva GD, Ratsch G: Persistence and Availability of Web Services in Computational Biology. *PLoS computational biology* 2011, **6**(9):e24914.
3. Jean G, Kahles A, Sreedharan VT, De Bona F, Ratsch G: RNA-Seq read alignments with PALMapper. *Current protocols in bioinformatics* Andreas D Baxevasis [et al] 2010, Chapter 11:Unit 11 16.
4. Bohnert R, Ratsch G: rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic acids research* 2010, **38**(Web Server):W348-351.
5. Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009, **25**(9):1105-1111.
6. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology* 2011, **12**(3):R22.

A8

Profiling transcription initiation in human aged brain using deep-CAGE

Margherita Francescato^{1,2*}, Luba Pardo¹, Patrizia Rizzu¹, Morana Vitezic^{3,4}, Javier Simón-Sánchez¹, Hazuki Takahashi³, Carsten Daub³, Piero Carninci³, Peter Heutink¹

¹Department of Clinical Genetics, Section Medical Genomics, VU Medical Center, Amsterdam, The Netherlands; ²GABBA Program, Instituto de Ciências Biomédicas Abel Salazar, UP, Porto, Portugal; ³RIKEN Omics Science Center, RIKEN Yokohama Institute, Yokohama, Japan; ⁴Department of Cell and Molecular Biology (CMB), Karolinska Institute, Stockholm, Sweden

E-mail: m.francescato@vumc.nl

BMC Bioinformatics 2011, **12**(Suppl 11):A8

Introduction: The genome sequencing projects completed in recent years revealed that the number of protein-coding genes does not change appreciably with increasing complexity of the organisms, and it is now generally accepted that this divergence is largely due to variation at the regulatory level. Mechanisms such as alternative splicing, alternative promoters and antisense transcription allow to both obtain a high number of transcripts from a relatively small number of genes and to fine tune isoforms expression in a cell-specific or developmentally-restricted manner.

It is likely that the extensive use of such mechanisms plays a pivotal role in development, adult function and ageing of complex tissues like brain.

The aim of this study was to characterize transcription start sites (TSSs) in different areas of human aged brain and correlate expression with methylation and structural genomic variation. Since its ability to profile TSSs at high resolution and at a genome wide level, we used Cap Analysis of Gene Expression (CAGE) combined with high-throughput sequencing (deep-CAGE) to collect exact TSSs and their expression levels. We present here our findings on alternative promoters and antisense transcription. Post-mortem tissue from 5 different brain regions was collected from 5 human donors and used to prepare 25 libraries.

Results: On average 2 million CAGE tags for each sample were sequenced. Mapping, expression normalization and clustering of the tags were carried out using automated pipelines. Core promoters were defined by merging tags within 300 base pairs of each other on the same strand. We found 22023 promoters, 50% of which mapped to either the promoter region or the 5' UTR of RefSeq transcripts. Ca. 32% of the genes expressed use alternative promoters. Ca. 15% of the promoters found were either part of a bi-directionally transcribed pair or antisense to an annotated RefSeq gene. A promoter was considered preferentially expressed (PEP) in one of the regions if at least 50% of its expression was derived from that region. Around 30% of the alternative promoters were PEP in one of the regions. In 8% of the bi-directional pairs identified, at least one of the members was PEP while 35% of the antisense promoters was.

Conclusions: This study confirms deep-CAGE as a suitable approach to characterize mechanisms involved in the regulation of gene expression, such as alternative promoter usage and antisense transcription, even in the challenging context given by the use of post-mortem tissue from aged human brain.

A9

Computational purification of tumor gene expression data

Amit Deshwar^{1*}, Gerald Quon², Qaid Morris³

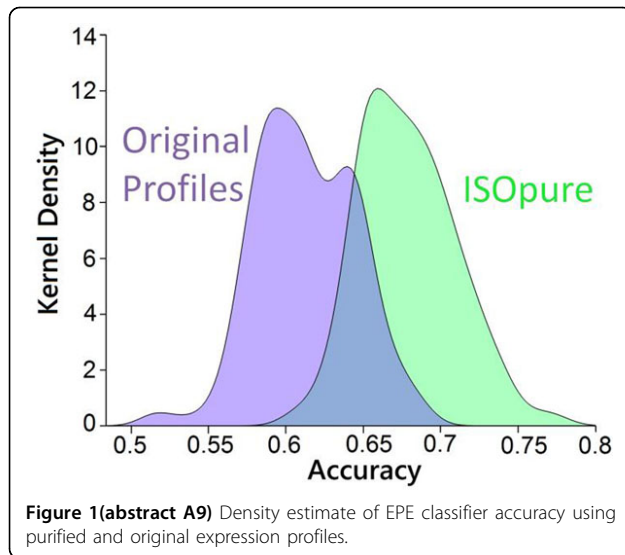
¹Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada; ²Department of Computer Science, University of Toronto, Toronto, Canada; ³Banting and Best Department of Medical Research, University of Toronto, Toronto, Canada
BMC Bioinformatics 2011, **12**(Suppl 11):A9

Background: Cancer gene expression profiling is an indispensable tool for identifying drivers of tumor progression, identifying subtypes, and predicting clinical outcome. An outstanding challenge faced by cancer gene expression studies is the limited concordance between studies [1], driven in part by lack of statistical power [2]. Part of this lack of statistical power is due to the fact that tumor samples from some solid cancers contain between 30%-70% healthy tissue [3]. This healthy tissue contaminates tumor expression profiles and variable amounts of healthy tissue leads to increased variability between tumor expression profiles. Physical purification of these tumor samples before profiling is often not feasible.

Materials and methods: We have developed ISOpure [4], a computational method to purify tumor gene expression profiles using reference samples of healthy tissue to model the contribution of healthy tissue. For every tumor expression profile in the input, ISOpure estimates the percentage of cancerous tissue and outputs a purified cancer expression profile from which the impact of healthy tissue has been removed. We verified our purification procedure by measuring the performance of expression-based predictive models of patient outcome in cancer, using either the original or ISOpure-purified expression profiles. We predicted extraprostatic extension (EPE) in 89 prostate tumor samples and patient survival for a set of 443 lung cancer patients.

Results and conclusions: Purified expression profiles showed significant improvements in prognostic model performance. 93% of the EPE classifiers constructed using the purified profiles had higher accuracy on held-out data in cross-validation than the matching classifier trained using the original expression data ($p = 1.58 \times 10^{-77}$), with an average improvement of 11% in performance (Fig. 1). For lung cancer, the prognostic model based on the purified profiles improved hazard modeling by 39% over the model based on the unpurified profiles ($p = 0.016$).

We have demonstrated that ISOpure improves our ability to predict patient phenotype based on gene expression, and expect to see similar impro-



vements for other cancer gene expression analyses such as subtype identification and classification. We are currently generating a compendium of purified gene expression profiles from 1600 tumor samples representing 15 different types of solid cancer using archival data from GEO. We are excited to work with the community at large to generate a resource of computationally purified cancer datasets, in order to facilitate more accurate analysis of cancer gene expression.

References

1. Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao MS, Penn LZ, Jurisica I: Prognostic gene signatures for non-small-cell lung cancer. *Proc Natl Acad Sci USA* 2009, **106**(8):2824-2828.
2. Ein-Dor L, Zuk O, Domany E: Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci USA* 2006, **103**(15):5923-5928.
3. Wang Y, Xia XQ, Jia Z, Sawyers A, Yao H, Wang-Rodriguez J, Mercola D, McClelland M: In silico estimates of tissue components in surgical samples based on expression profiling data. *Cancer Res* 2010, **70**(16):6448-6455.
4. Quon C, Haider S, Deshwar AG, Cui A, Boutros PC, Morris QD: Patient-specific computational purification of gene expression profiles. *Nature Biotechnology* 2011, in review.

A10

The development of an open-access database for human transcriptional regulation interactions

Luiz Augusto Bovolenta*, Marcio Luis Acencio, Ney Lemke
Department of Physics and Biophysics, Instituto de Biociências de Botucatu, Unesp -Univ Estadual Paulista, Botucatu -SP, Brazil
E-mail: labovolenta@gmail.com
BMC Bioinformatics 2011, **12**(Suppl 11):A10

Background: The modeling of interactions among transcription factors (TFs) and their respective target genes (TGs) into transcriptional regulatory networks is important for the complete understanding of regulation of biological processes. In the case of human TF-TG interactions, there is no database at present that explicitly provides such information even though many databases containing human TF-TG interaction data have been available, such as TRANSFAC [1] and TRED [2]. In an effort to provide researchers with a repository of TF-TG interactions from which such interactions can be directly extracted, we present here the human transcriptional regulation interactions database (HTRldb), an open-access database of experimentally validated interactions among human TFs and their respective TGs.

Materials and methods: The HTRldb is implemented as a relational database PostgreSQL that is connected to a web interface via the JBOSS AS

application server that dynamically generates user-friendly HTML front-end queries using the Apache Tomcat web server. For the visualization of TF-TG interactions, we embedded in the HTRldb the Cytoscape Web.

Results: The HTRldb offers several mechanisms of data query and extraction, such as download in spreadsheet or text format and the visualization of TF-TG interactions. There is an update mechanism that allows scientists to send new data. The HTRldb currently holds a collection of 2,114 unique transcriptional regulation interactions among 163 TFs and 1,034 TGs.

Conclusion: HTRldb is a powerful user-friendly tool from which human TF-TG interactions can be easily extracted.

Acknowledgements: We would like to thank The State of São Paulo Research Foundation (FAPESP) (grants 2009/10382-2 and 2010/20684-3) and CNPQ for the financial support.

References

1. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al: TRANSFAC and its module TRANSCCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006, **34**(Database):D108-110.
2. Jiang C, Xuan Z, Zhao F, Zhang MQ: TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* 2007, **35**(Database):D137-140.

A11

Computational analysis of genetic network involved in pancreatic cancer in human

Mrinal Mishra*, Ambuj Kumar
School of Bio Sciences and Technology, VIT University, Vellore, India
BMC Bioinformatics 2011, **12**(Suppl 11):A11

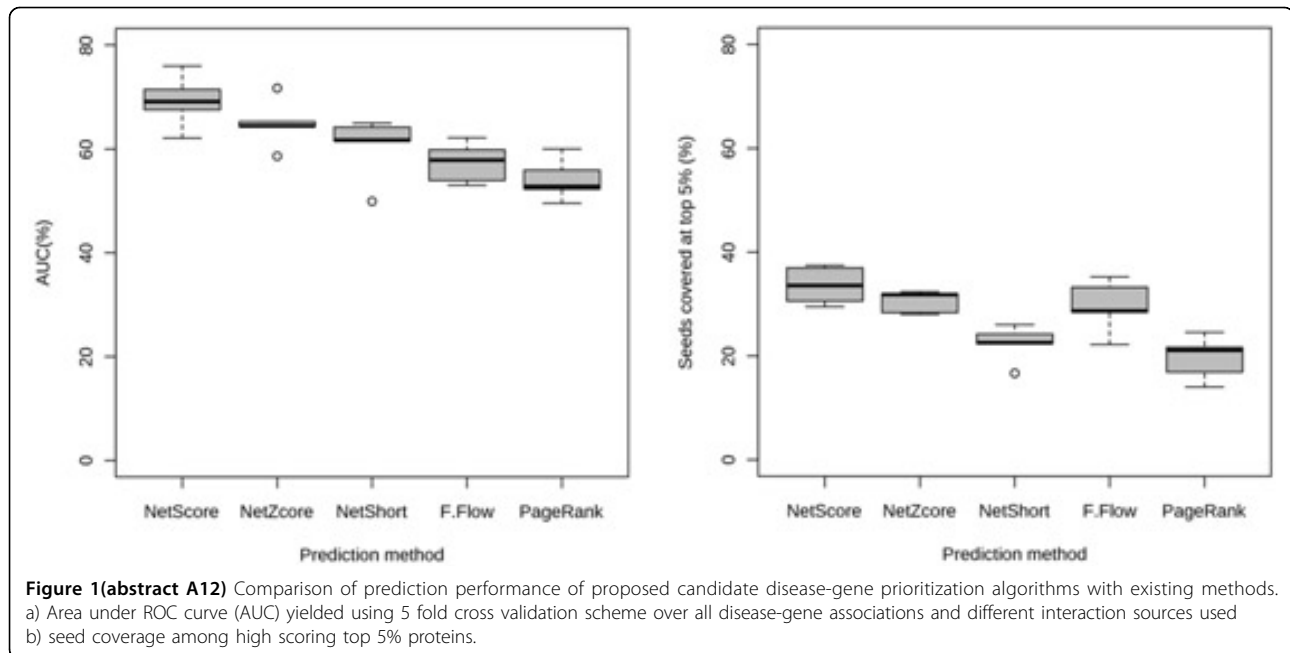
The poster is based on the in silico identification and analysis of the variation in the gene network related to the pancreatic cancer. Pancreatic cancer has been a major cause of death in Asia and European countries. This cancer is not easily detectable in its initial stages and at the later stages it becomes very hard to cure this disease. So the in-depth understanding of the variation in the genetic pathway of this disease is very important. We selected 5 candidate genes from various published journals which are involved in the pancreatic cancer pathway (KRAS, CDKN2A, MADH4, TP53 and ARMET) in generating their interaction network using Agilent literature search plugin in cytoscape. The organic layout of interaction revealed the cross interaction between these genes and the other neighbour genes. Merging the expression profile data of the pancreatic cancer to the parent network helped us in understanding the variation of the network in the diseased state. Using the merged profile network we found out the importance of the KRAS & CDKN2A gene interaction with other 21 neighbour genes among which PIK3CA and TP53 interactions were showing major variation on their expression pattern. This study reveals the importance of change in expression level of candidate genes (KRAS, CDKN2A and TP53) in causing pancreatic cancer. The results obtained in our study will be very much useful in detecting the disease in its initial stages and in finding the cure.

A12

Toward PWAS: discovering pathways associated with human disorders

Emre Guney, Baldo Oliva*
Structural Bioinformatics Group (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Research Park of Biomedicine (PRBB), 08003, Barcelona, Catalonia, Spain
BMC Bioinformatics 2011, **12**(Suppl 11):A12

Introduction: The past decade has witnessed dramatic advances in genome sequencing and a substantial shift in the number of genome wide association studies (GWAS). These efforts have expanded considerably our knowledge on the sequential variations in Human DNA and their consequences on the human biology. Nevertheless, complex genetic disorders often involve products of multiple genes acting cooperatively and pinpointing the decisive elements of such disease pathways remains a challenge. Network biology recently proved its use in identifying candidate disease genes based on the simple observation that proteins translated by



phenotypically related genes tend to interact, the so called guilt-by-association principle.

Methods: Here, we present GUILD (Genes Underlying Inheritance Linked Disorders), a network-based candidate disease-gene prioritization framework which reveals the pathways associated with the disorder (pathway wide association study, PWAS). We exploit several distinct plausible communication mechanisms of known genes associated with the phenotype emerging from the topology of the interaction network. We used three sources of gene-phenotypic association to specify nodes involved in a disorder (seeds for the methods proposed): Online Mendelian Inheritance in Man (OMIM) database [1] and two published data sets (by Goh et al. [2], and Chen et al. [3]).

Results and discussion: Analysis on multiple human disease phenotypes demonstrate that the methods proposed in GUILD surpass state-of-the-art prioritization methods such as PageRank with priors [3] and Functional-Flow [4] (Fig. 1). We tested the robustness of the approaches proving the effect of the network properties and the independence with the number of original genes/proteins associated with the function or phenotype. We applied the prioritization methods to study the implication of pathways in various diseases and highlight the relationship between AD and aging. Our findings confirm that most prioritization methods introduced in this study are able to distinguish between groups of connected genes with functions identical to those of the known disease-associated genes (disease pathways). In addition, using prioritization methods, we increased the coverage of genes known to play important roles in the interplay of AD and aging, most of which would not be otherwise identified by just inspecting the direct neighborhood in the network.

References

1. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33(Database):** D514-517.
2. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: **The human disease network.** *Proc Natl Acad Sci USA* 2007, **104(21):**8685-8690.
3. Chen J, Xu H, Aronow BJ, Jegga AG: **Improved human disease candidate gene prioritization using mouse phenotype.** *BMC Bioinformatics* 2007, **8:**392.
4. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M: **Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps.** *Bioinformatics* 2005, **21(Suppl 1):**i302-310.

A13

Spectral classification of short numerical exon and intron sequences

Benjamin YM Kwan^{1*}, Jennifer YY Kwan², Hon Keung Kwan³

¹Faculty of Medicine, University of Ottawa, Ottawa, Ontario K1H 8M5,

Canada; ²School of Medicine, Queen's University, Kingston, Ontario K7L 3N6,

Canada; ³Department of Electrical & Computer Engineering, University of

Windsor, Windsor, Ontario N9B 3P4, Canada

E-mail: bkwan066@uottawa.ca

BMC Bioinformatics 2011, **12(Suppl 1):**A13

Background: Current methods for genome annotation focus on sequence similarity or motif matching to known genes and there is a need for a complementary or more effective approach. It is known that protein coding (exonic or C-G rich) regions exhibit a period-3 property which is less prominent in noncoding (intronic or A-T rich) regions. The boundary between these 2 regions becomes less apparent as sequence length becomes shorter. The period-3 property is likely due to the 3-base-length of codons. C-G rich content in coding regions is due to nonuniform codon usage. For spectral analysis of period-3 value, a nucleotide sequence has to be converted to a numerical sequence. The choice of numerical representation affects how well its biological properties can be preserved and reflected.

Methodology: Based on exon and intron sequences downloaded from UCSC Genome Browser on Human (GRCh37/hg19) (<http://genome.ucsc.edu/cgi-bin/hgText>) using [1-3], the classification performance in precisions (%) were computed by applying the spectral analysis and thresholding of [4] to the following twelve numerical representation methods: 1. Integer Number; 2. Single Galois Indicator; 3. Paired Nucleotide Atomic Number; 4. Atomic Number; 5. Molecular Mass; 6. EIIP; 7. Paired Numeric; 8. Real Number; 9. Complex Number; 10. Complex Twin-Pair (C, G = -1; A, T = j); 11. Complex Bipolar-Pair Code I (C = -1; G = 1; A = j; T = -j); 12. Complex Bipolar-Pair Code II (C = -1; G = 1; A = -j; T = j). Methods 1-9 are specified in [4] and Methods 10-12 are new numerical representations. In simulations, two adjacent windows are overlapped by 3 bases.

Results and conclusions: The results summarized in Figure 1 indicate that the approach is capable for effective classification of untrained short exon and intron sequences. Among the 3 new numerical representations, the Complex Twin-Pair (Method 10) achieves a precision of about 79% to 92% for a sequence length of 150 bases to 600 bases and a window

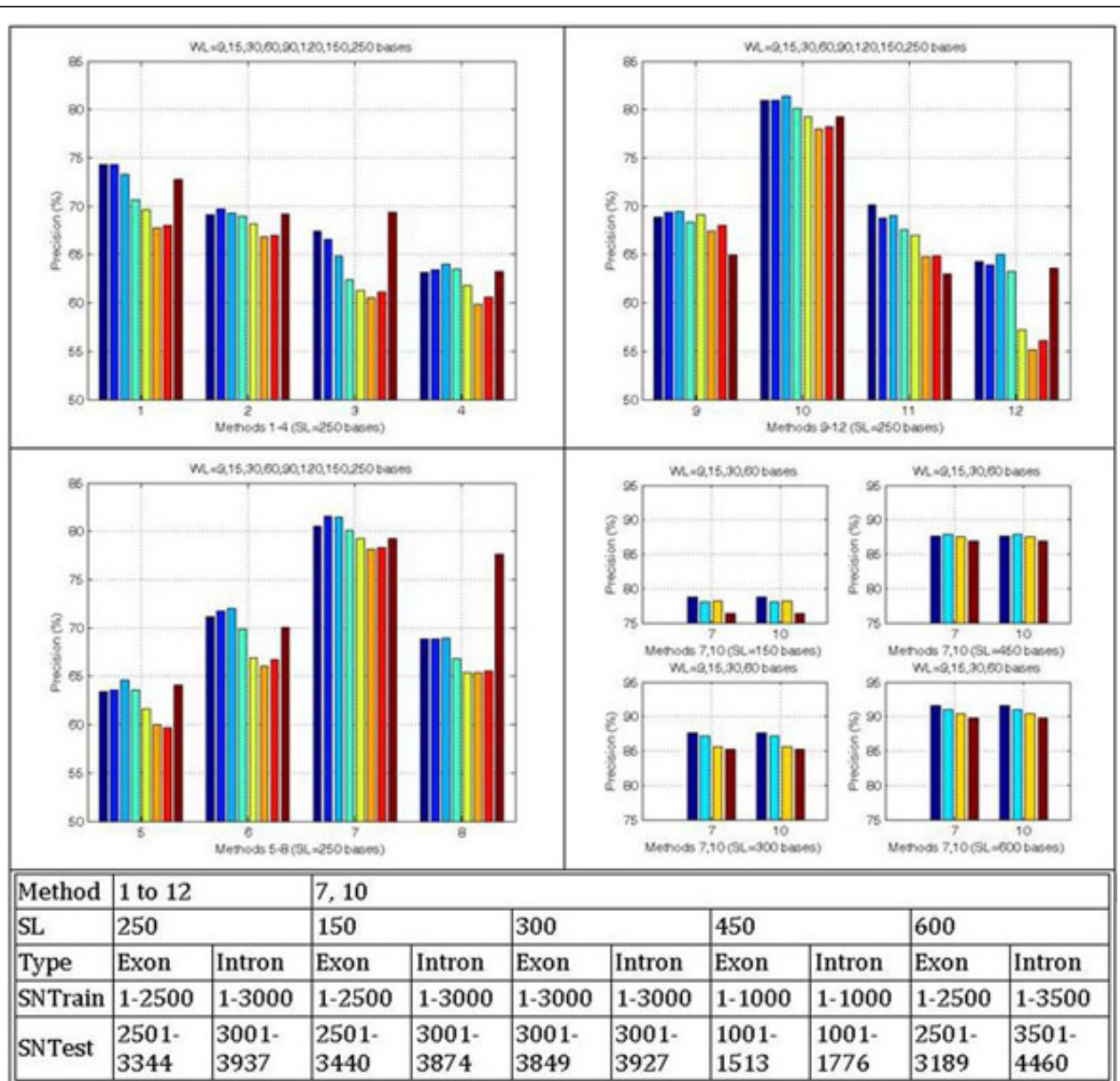


Figure 1 (abstract A13) Precisions of methods obtained by different WL set and SL (SL: Sequence length in bases; WL: Window length in bases; SNTrain: Sequence numbers for training; SNTTest: Sequence numbers for testing).

length of 9 bases which is comparable with those of the Paired Numeric (Method 7).

References

1. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool.** *Nucleic Acids Res* 2004, **32(Database issue)**:D493-496.
2. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11(8)**:R86.
3. Blackenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: **Galaxy: a web-based genome analysis tool for experimentalists.** *Curr Protoc Mol Biol* 2010, Chapter 19(Unit 19.10):1-21.

4. Kwan JYY, Kwan BYM, Kwan HK: **Spectral analysis of numerical exon and intron sequences.** *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)* Hong Kong 2010, 876-877.

Cite abstracts in this supplement using the relevant abstract number, e.g.: Kwan et al.: Spectral classification of short numerical exon and intron sequences. *BMC Bioinformatics* 2011, 12(Suppl 11):A13