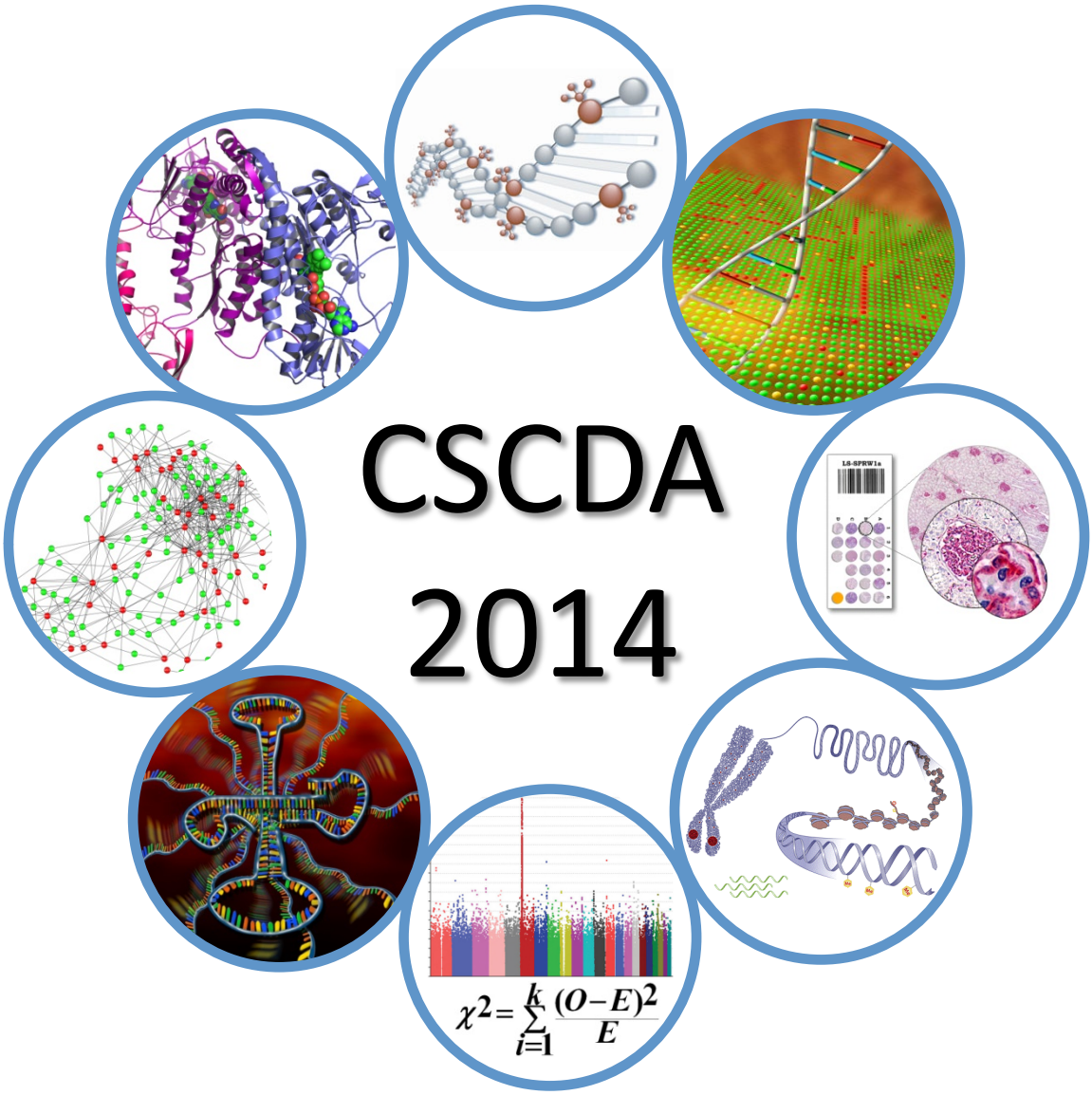
3rd meeting on

CAPITA SELECTA IN COMPLEX DISEASE ANALYSIS



Program and Abstract Book

GIGA - Liège - Belgium

24 - 26 November 2014

Sponsors


fnrs
LA LIBERTÉ DE CHERCHER


EU
PANCREAS


cost


G I G A


Université
de Liège

Welcome

CSCDA 2014

Dear Participants

It is our pleasure to welcome you to the third edition of Capita Selecta in Complex Disease Analysis (CSCDA 2014) in Liège, Belgium. CSCDA is a biennial meeting providing an interdisciplinary, international platform to discuss challenges and developments in the analysis of genetically complex diseases. The first edition was held in Leuven, Belgium in August 2010, followed by a second edition here in Liège in May – June 2012.

This year's CSCDA is held in conjunction with Annual Meeting of the EU COST Action on Pancreas Cancer, and has two overarching themes to inspire dialogue on complex disease analysis: uncovering population and patient heterogeneity in the –omics era, and data fusion and integration to improve mechanistic insights in rare diseases.

Highly evaluated in the previous editions, CSCDA 2014 kicks off with two workshops. This year's workshops address public –omics databases and bioinformatics tools for integrated -omics analysis. During the afternoon sessions of CSCDA 2014, we are very pleased to welcome invited speakers coming from a broad scientific background -and from across the globe- to share their expert insights on topics ranging from population genetics to personalized medicine. In the forenoons, the floor is for early career scientists, who will get the opportunity to share as well as discuss their work during moderated discussion sessions.

We wish you a fruitful meeting that will boost dialogue and create novel opportunities and collaborations to tackle the challenges of complex disease analysis.

Sincerely,
Kristel Van Steen and Kristel Sleegers
(Presidents of the Organizing Committee)

Welcome

CSCDA 2014

Dear Colleagues,

Pancreas cancer awareness is increasing worldwide and EUPancreas has importantly contributed to this achievement. The list of accomplishments of the Action till present is long, among them several scientific publications and events, 2 surveys, 10 short-term scientific missions, a training school, involvement of patient organizations, the webpage and dissemination actions, among others. Most importantly, the scientific initiatives each Working Group is conducting and the involvement of the EUPancreas in other international initiatives such as that of the Team Pancreatic Cancer Platform that aims to lobby in the European Parliament for increasing the awareness of this dreadful disease. This results from the effort of key members that took the responsibility to set up the framework of the Action and to whom I am deeply thankful.

It is now time to capitalize and expand on the important initial accomplishments of EUPancreas and to further enhance its impact within and beyond the Action. More activities will be pursued in cooperation with the member countries and groups for the purpose of training early-stage researchers. We further aim to reach other international initiatives on pancreas cancer research to allow our members establish fruitful collaborations and advance in the control of pancreas cancer. To increase the active involvement and participation of the members in EUPancreas activities will maximize the benefits members receive.

The Joint CSCDA & EUPancreas Conference, organized by Working Group 2, provides a unique opportunity to engage people in this common endeavour towards the above-mentioned goals. The appealing program offers the most exciting innovative methodological approaches and translational and clinical discoveries in the field of omics data integration presented by super-expert international speakers. This will serve as the basis for sharing information and exchanging ideas to make progress in the field favouring an effective and responsible translation into the clinics and public health domains.

In parallel, the EUPancreas scheduled meetings will importantly allow us to revise the already ongoing initiatives and plan future activities of the Action to meet with our main objective, namely to capitalise on emerging scientific and technological developments in the field of pancreas cancer research.

I thank Working Group 2 that have taken the lead to organize this event and the rest of the Action members that contribute to EUPancreas, and specifically to this Conference, and wish you all and enjoyable scientific and social time during the 3-day meeting in Llège.

Warm regards,

Núria Malats
Chair EUPancreas
COST Action BM1204

Committees

CSCDA 2014

Organizing committee

Nuria Malats	Professor	Genetic & Molecular Epidemiology Group, CNIO, Spain
Stuart Maudsley	Professor	VIB Department of Molecular Genetics, UA, Belgium
Anavaj Sakuntabhai	Professor	Pasteur Institute, Paris, France
Kristel Sleegers	Professor	VIB Department of Molecular Genetics, UA, Belgium
Jean-Luc Van Laethem	Professor	Erasmus Hospital, University of Brussels, Belgium
Kristel Van Steen	Professor	Montefiore Institute, University of Liège, Belgium

Scientific committee

Ilja Arts	Professor	Nutritional & Molecular Epidemiology, Maastricht University, The Netherlands
Francesco Gadaleta	PhD	Montefiore Institute, University of Liège, Belgium
Nuria Malats	Professor	Genetic & Molecular Epidemiology Group, CNIO, Spain
Stuart Maudsley	Professor	VIB Department of Molecular Genetics, UA, Belgium
Anavaj Sakuntabhai	Professor	Pasteur Institute, Paris, France
Kristel Sleegers	Professor	VIB Department of Molecular Genetics, UA, Belgium
Jean-Luc Van Laethem	Professor	Erasmus Hospital, University of Brussels, Belgium
Kristel Van Steen	Professor	Montefiore Institute, University of Liège, Belgium

Local assistants (BIO3 team)

Kyrylo Bessonov	PhD student	Montefiore Institute, University of Liège, Belgium
Kridsadakorn Chaichoompu	PhD student	Montefiore Institute, University of Liège, Belgium
Ramouna Fouladi	PhD student	Montefiore Institute, University of Liège, Belgium
Francesco Gadaleta	PhD	Montefiore Institute, University of Liège, Belgium
Elena Gusareva	PhD	Montefiore Institute, University of Liège, Belgium
Silvia Pineda	PhD student	Montefiore Institute, University of Liège, Belgium
François Van Lishout	PhD student	Montefiore Institute, University of Liège, Belgium
Benjamin Dizier	PhD student	Montefiore Institute, University of Liège, Belgium

Program

CSCDA 2014

Day 1 – 24 November 2014 **Location: GIGA – floor 5**

08.00 – 08.45	**Registration**	
	CSCDA2014 workshop / EUPancreas WG2 workshop (Chair: Jörg Hoheisel)	**Parallel COST Sessions**
08.45 – 09.00	**Workshop welcome**	
09.00 – 10.45	**Silke Szymczak** Institute of Medical Informatics and Statistics, University of Kiel, Germany Title: "Public resources of omics data sets - processing omics data prior to analysis"	**WG1: 9-11 am** **WG3: 9.30 am – 4.30 pm** **(detailed agenda via WG leaders)**
10.45 – 11.15	Coffee break	
11.15 – 13.00	**Stuart Maudsley** VIB, Flanders, Belgium Title: "Getting older : it's always more complex than you think"	
13.00 – 13.45	**Walking lunch**	
	CSCDA2014 scientific session I (Chairs: Kristel Van Steen)	**Parallel COST Sessions**
13.45 – 14.00	**Welcome**	**WG3: 9.30 am – 4.30 pm** **(detailed agenda via WG leaders)** **WG2: 5.30 pm – 7pm**
14.00 – 15.00	**Luisa Pereira** Institute of Molecular Pathology and Immunology, Portugal Title: "Population genetics in infectious diseases"	
15.00 – 16.00	**Silvia Vidal** Department of Human Genetics, McGill University, Montreal, Quebec, Canada Title: "Epistasis, host-pathogen interactions and pleiotropy in the genetic control of immunity to infection: molecular insights from the mouse genome."	
16.00 – 16.30	**Coffee break**	
16.30 – 17.30	**Michael Nothnagel** Köln Center for Genomics, University of Köln, Germany Title: "Misreading epidemiological effect sizes: a note of caution."	
17.30 – 19.00		

Day 2 – 25 November 2014 **Location: GIGA – floor 5**
CSCDA2014 scientific session I (Chair: Kristel Sleegers)

08.15- 08.30	**Announcements**	
08.30 – 09.45	**Abstract based presenters (25 min each)** **Askar Obulkasim** Title: "Classification using differential network rank conservation revisited." **Kris Kridsadakorn** Title: "Iterative pruning principal component analysis to retrieve population structure." **Jeroen Huyghe** Title: "mRNA-seq of 278 diverse skeletal muscle biopsies reveals mechanistic insights about type 2 diabetes genetic risk and identifies disease state specific eQTLs."	**Optional WG / EUPancreas breakout meetings**
09.45 – 10.00	Short break	

CSCDA2014 scientific session I (Chair: Kristel Sleegers) (continued)

10.00 – 10.30	**Abstract based presenters (25 min each)** **Pablo Riesgo** Title: "Integrative visual analysis of genomic data on Spotfire." **Kyrylo Bessonov** Title: "Regulatory gene network inference from expression data via conditional inference trees."	**Optional WG / EUPancreas breakout meetings**
10.30 - 11.00	**Kim Do Kyoon** Center for System Genomics, Pennsylvania State University, U.S.A. Title: "Multi-Omics Data Integration for Predicting Cancer Clinical Outcomes."	
11.15 – 11.45	Coffee break	
11.45 – 12.30	**Ilja Arts** Molecular Epidemiology of Chronic Diseases, Maastricht University, the Netherlands Title: "Uncovering population heterogeneity in the omics era"	
12.30 – 13.15	**Anavaj Sakuntabhai** Functional Genetics of Infectious Diseases, Pasteur Institute, France Title: "Understanding complex population substructures: the dengue story."	
13.15 – 14.00	Walking lunch	

CSCDA2014 / EUPancreas session II (Chair: Núria López-Bigas)

14.00 – 15.00	**Manuel Hidalgo** Gastrointestinal (GI) Cancer Clinical Research Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain Title: "Incorporating Mouse Models into Pancreas Cancer Treatment."	
15.00 – 16.00	**Taesung Park** Department of Statistics, Seoul National University College of Natural Sciences, Seoul, South Korea Title: "Interactions to enhance functional interpretation in pancreatic research"	
16.00 – 16.30	Coffee break	
16.30 – 17.30	**Andrew Biankin** Translational Research Centre, University of Glasgow, UK Title: "The importance of Integrating, transforming and sharing data to unravel new complex disease mechanisms."	
18.30 -	Social event (registration required)	

Day 3 – 26 November 2014 — Location: GIGA – floor 5

CSCDA2014 / EUPancreas session II (Chair: Carlo La Vecchia)

08.30 – 08.45	Announcements	
08.45 – 10.00	**Abstract based presenters (25 min each)** **Pierre-Emmanuel Sugier** Title: "Integration of gene-based and text mining analyses to discover genes underlying atopy." **Silvia Pineda** Title: "Integration analysis of 'OMICS' data using penalized regression methods: An application to bladder Cancer." **Şennur Görgülü Kahyaoğlu** Title: "Novel theranostic nanocomplex for pancreatic cancer: preparation and characterization of anti-mesothelin antibody oriented and gambogic acid bioconjugated iron buried nanolactoferrin."	

CSCDA2014 / EUPancreas session II (Chair: Carlo La Vecchia) (continued)

10.00 – 10.15	Short break	
10.15 – 11.30	**Abstract based presenters (25 min each)** **Francesco Gadaleta** Title: "Are we far from correctly inferring gene interaction networks?" **Yuanlong Liu** Title: "Network-assisted investigation of signals from genome-wide association studies in childhood-onset asthma." **Ramouna Fouladi** Title: "A novel gene-based analysis method-based on MD-MBR."	
11.30 – 12.00	Coffee break	**WG4: 11.30 am - 1.30 pm (detailed agenda via WG4 leader)**
12.00 – 12.45	**Jean-Luc Van Laethem and Raphaël Maréchal** Laboratory of Experimental Gastroenterology, Erasmus Hospital, Brussels, Belgium; EORTC Gastrointestinal Tract Cancer Group (GI Group) Title: "European (EORTC) organization of research in PC. Perspectives in PC research from the clinical researcher."	
12.45 – 13.30	**Fatima Al-Shahrour** Translational Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain Title: "Bioinformatics approaches for personalized cancer therapy: From Pan-Cancer projects to Patient derived xenografts (PDX) models."	
13.30 – 13.45	**CSCDA214 closure**	

COST Session III (Chair: Núria Malats)

13.45 – 14.30	Walking lunch for COST members
14.30 – 15.30	Annual Meeting (includes joint WG meeting)
15.30 – 17.30	MC Meeting

Silke Szymczak

Silke Szymczak, Dr. is currently working as a research assistant in the Institute of Medical Informatics and Statistics in the University Medical Center Schleswig-Holstein, Kiel, Germany.
She graduated in Computer Science in the Natural Science for the University of Bielefeld, Bielefeld, Germany in 2005. She obtained her PhD in 2011 in the Institute of Medical Biometry and Statistics at the University of Lübeck (supervisor: Prof. Dr. rer. nat. Andreas Ziegler). From 2011 to 2013, she was a postdoctoral fellow at the Inherited Disease Research Branch of the National Human Genome Research Institute, NIH in Baltimore, USA (supervisor: Dr. Joan Bailey-Wilson) and from March 2013 until August 2014 she was member of Prof. Andre Franke's group at the Institute of Clinical Molecular Biology, Kiel.
Her research is focused mainly in Machine learning, especially random forest, genetic epidemiology and the analysis of omics data sets.

Public resources of omics data sets

High troughput technologies nowadays produce data of many different types of omics levels like genomics, transcriptomics and epigenomics. Analysis of these data sets often pose specific computational and statistical challenges that will require comparisons of existing methods as well as development of new approaches. Validity and power of those methods is usually evaluated based on simulations. However, little is known about correlation patterns for a specific omics level, and especially not about relationships between levels. For an evaluation under realistic conditions, it is therefore necessary to use experimental data sets.
In this talk I will present several repositories like The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO) and ArrayExpress which contain publicly available data sets from different omics levels and a wide range of experimental conditions and diseases. Furthermore, I will demonstrate how to use R to systematically download and prepare those data sets for a comparison of statistical methods.

Stuart Maudsley

Stuart Maudsley is currently the adjunct department director in the VIB Department of Molecular Genetics and professor at the Univeristy of Antwerp, faculty of Pharmaceutical, Biomedical and Veterinary Sciences. He was awarded in 2013 with the FWO Odysseus Type I Award 'Post-genomic investigation of therapeutic targets for the treatment of neurodegenerative disorders'. He graduated in Pharmacology First Class Honors at the University of Leeds in 1993 and he was awarded with the Pfizer Research Prize. He obtained her PhD in Receptor Pharmacology at the University of Leeds in 1996 as Ackroyd Brotherton & Brown Scholar. In 1997, he obtained the Howard Hughes Medical Institute Research Fellowship at Duke University supervised by Robert J. Lefkowitz (2012 Nobel Laureate in Chemistry). He worked as principal investigator in the GnRH Receptor Biology Lab at Medical Research Council (MRC), University of Edinburgh in 2000 and as a consultant in Receptor Therapeutics at Ardana Bioscience in 2002.

In 2004, he worked as Head of Receptor Pharmacology Unit, National Institutes of Health (NIH) at National Institute on Aging, Bethesda, USA. He was awarded by Co-recipient – 'UK Best Practice Award' - Association for the Study of Obesity (2010), NIH Clinical Center 'Bench-to-Bedside' Award (2011) and 'On-The-Spot' Award (2011, 2012, 2013).
In 2010, he started as senior lecturerat Johns Hopkins Bloomberg School of Public Health, Baltimore, USA and in 2011, senior lecturerat Johns Hopkins School of Medicine, Department of Endocrinology, Baltimore, USA.

Getting older : it's always more complex than you think.

Many central and peripheral diseases are strongly affected by the aging process. Aging can induce a trigger for the disease or create different outcomes based on the longitudinal nature of the process. The aging process affects every tissue in the body and represents one of the most complicated and highly integrated inevitable physiological entities. Aging can occur at a differential rate, either locally in cells or tissues, as well as globally across the whole organism. Therefore, perhaps aging should not be considered to be a chronological eventuality but a series of mechanistic systems linked to damage. The maintenance of good health during the aging process relies upon the coherent regulation of hormonal and neuronal communication between the central nervous system and the periphery. Evidence has demonstrated that the optimal regulation of energy usage in both these systems facilitates healthy aging. Therefore, understanding how complex neuronal and metabolic systems are connected at the generic molecular level will help engender a better understanding of the aging process and could also highlight the most important factors that control and potentially ameliorate the age-related damage. Identifying key factors of the global molecular aging system has the capacity to generate novel therapeutic leads for a multitude of disorders in which aging plays a role.

Luísa Pereira

Luísa Pereira has a degree in Biology, Master degree in Genetics and a PhD in the field of Human Population Genetics. She is a researcher and group leader at Ipatimup (Institute of Molecular Pathology and Immunology of the University of Porto) since 2004 and 2006 respectively, being interested in using genetics to infer the past and evolution of human populations, as well as on disentangling between neutral and pathological diversities. She is co-author of 80 peer-reviewed papers in international indexed journals.

Population genetics in infectious diseases

Genetic studies on the worldwide human population have been allowing to catalogue its genetic diversity across space and time, since its origin at around 200,000 years ago in eastern/central Africa. First migrations were restricted to the African continent, and only at 65,000 years ago did modern humans successfully migrated out-of-Africa, arriving first in Asia, then in Europe and finally in America. The out-of-Africa migration was a strong bottleneck, with a very small group of migrants giving rise to all diversity observed nowadays in non-Africans. As people moved to new environments, new genetic diversity was generated, creating population structure. Some pathogens, such as Helicobacter pylori, were companions of humans in the out-of-Africa migration, originating also heterogeneous strains specific of certain parts of the globe..

Other pathogens, such as the dengue virus, emerged in one region after the out-of-Africa migration, and are beginning to be introduced in other regions via the globalization. It has been shown that the selection pressure conferred by pathogens has been among the strongest forces acting upon the human genome, leading to an increase of the frequency (loss of diversity) of positively selected SNPs which confer resistance to the infection.

In this talk, we will discuss ancestry influence in the susceptibility/resistance to infectious diseases, and how it can be used in mapping candidate genes. In fact, if population structure can bias the evaluation of SNP association with complex diseases, leading to false positive results, new methods are being developed which take advantage of the differential susceptibility conferred by ancestry in mapping variation causing/protecting against certain diseases. Also the comparison of selected signs between cases and controls can help to identify candidate genes. I will present two examples of African resistance, when compared with other population groups, to two pathogens: dengue virus and H. pylori. Dengue virus is more recent, presents four strains that have been broadly observed in all endemic regions, and most probably African populations acquired resistance to the infection or worse manifestations of the disease. I will present our data in mapping genes conferring this African resistance to dengue in the Cuban population. The oldest H. pylori presents strains which are typical of African or Asian or European populations. Recent work performed by others in the admixed human population of Colombia showed that African H. pylori ancestry was benign in African descent people but was deleterious in Amerindian descendants. These results indicate that a more complex coevolution of H. pylori and humans modulated the risk of gastric disease.

Silvia V

Silvia Vidal, PhD (University of Geneva), is a Professor in the Departments of Human Genetics and Medicine and an Associate Member in the Department of Microbiology and Immunology, McGill University. She is Director of the Complex Traits Group at the same institution. She holds a Tier 1 Canada Research Chair in Host Responses to Virus Infections and is the recipient of the Ontario Premier's Research Excellence Award. She uses mouse genetic platforms to discover and functionally characterize the molecular interface between pathogenic viruses, inflammation and immunity. Her laboratory made inwards in characterizing mechanisms of self/non-self discrimination by NK cells during viral infection and pathways that control inflammatory responses during coxsackieviral myocarditis and influenza pneumonia. She has also developed an internationally recognized program in ENU mutagenesis and infectious diseases. Vidal's program has speared a number of projects and collaborations with academia and industry at McGill and abroad on gene discovery in human infectious and inflammatory

Epistasis, host-pathogen interactions and pleiotropy in the genetic control of immunity to infection: molecular insights from the mouse genome.

Fostered by the advent of genome-wide technologies, remarkable progress has been made in recent years in our understanding of the genetic basis of immunity-related disorders in the human population, including infectious and inflammatory autoimmune diseases. Whereas many disease variants have been identified, their additive effects explain only a portion of the trait variance; there is debate regarding the influence of non-additive genetic variance, including epistasis, in the genetic architecture of immune-related diseases. Questions also remain regarding the importance of host-pathogen interactions in shaping the genetic and phenotypic diversity of immune responses as well as in understanding genetic correlations between multiple traits, or pleiotropy. Through selected examples of mouse models of infection with viruses and other pathogens, this presentation will illustrate how the use of well-characterized mouse models of infection combined with genomic approaches, immunology and virology methods offer an excellent opportunity to examine these complex questions at the molecular level. We will provide evidence supporting genetic and molecular interactions between natural killer (NK) cell receptor and major histocompatibility complex (MHC) class I loci in the control of severity to virus infection. We will show data linking, at lease in part, the diversity of NK cell receptor genes in inbred mouse populations with their ability to respond to virus-encoded molecules during infection. Finally, we will present results from a genome-wide mouse mutagenesis screen for susceptibility or resistance to pathogen challenge, which has revealed a sub-set of genes each implicated in several infectious and inflammatory conditions. These studies complement and extend findings of human genetics by revealing unappreciated facets of immunology and providing new frameworks to understand the immune system and its pathologies.

Michael Nothnagel

Michael Nothnagel, he is professor of Statistical Genetics and Bioinformatics at the University of Cologne in Colgone, Germany. He graduated in Mathematics in 1999 at Humboldt University in Berlin, Germany. In 2005 he obatained his Ph.D. in Statistical Genetics at Charité Medical Faculty & Max Delbrück Center for Molecular Medicine (MDC) in Berlin, Germany. In 2010 he was Habilitation in Medical Statistics at Medical Faculty, Christian-Albrechts University in Kiel, Germany.

Misreading epidemiological effect sizes: a note of caution

Guided by the practice of classical epidemiology, research into the genetic basis of complex disease usually takes for granted the dictum that causative mutations are invariably over-represented among affected as compared to unaffected individuals. However, employing various models of population history and penetrance, we show that this supposition is not true and that a mutation involved in the etiology of a complex disease can under certain circumstances be depleted rather than enriched in the affected portion of the population. Such mutations are 'protective' in an epidemiological sense and would often tend to be erroneously excluded from further studies. Our apparently paradoxical finding is due to the possibility of a negative correlation between complementary causative mutations that may arise as a consequence of the specifics of the population genealogy. This phenomenon also has the potential to hamper efforts to identify rare causative mutations through whole-genome sequencing.

Kim Do Kyoon

Kim Do Kyoon, is a Postdoctoral Fellow at Center for Systems Genomics in Pennsylvania State University. He graduated in Computer Science at Korea University in Seoul, Korea, and he obtained his Ph.D. in Molecular and Genomic Medicine at College of Medicine Seoul National University in Seoul, Korea.

Multi-Omics Data Integration for Predicting Cancer Clinical Outcomes

Cancer clinical outcome prediction based on the molecular information has received increasing interest for better diagnostics, prognostics, and further therapeutics. Accurate molecular-based predictors of outcome can be used clinically to choose the best of several available therapies for a cancer patient. In the past decade, gene expression profiles have been most widely used to predict clinical outcomes in several cancers. There have been also many attempts at cancer clinical outcome prediction using a set of copy number alterations (CNA), miRNA, DNA methylation, and protein expression. However, it is still difficult to accurately predict clinical outcome since the cancer genome is neither simple nor independent but rather complicated and dysregulated by multiple levels of the biological system through genome, epigenome, transcriptome, proteome, metabolome, interactome, etc. Therefore, no single type of genomic data will be sufficient to elucidate the phenotypic end-point of events accumulated through multiple levels of biological systems involved in cancer, and hence, a consideration of incorporating the multi-layered processes in biological systems might provide much more reasonable prediction of cancer clinical outcome. Recently, emerging multi-omics data and clinical information from cancer patients such as The Cancer Genome Atlas (TCGA) have been providing unprecedented opportunities to investigate the multi-layered processes involved in cancer development and progression for improving the ability to diagnose, treat, and prevent cancer. Thus, the development of multi-scale integrative approaches is more required in order to integrate multiple types of genomic data and investigate an enhanced global view on interplays between different types of genomic data.

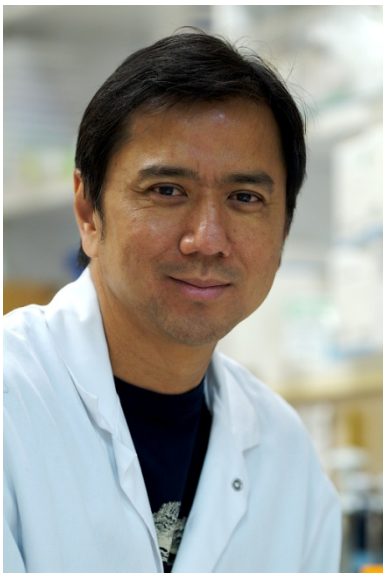
In this talk, many research schemes for multi-omics data integration will be discussed based on the experimental results on the prediction problem of cancer clinical outcomes using the TCGA data. With an abundance in of multi-omics data and clinical data from cancer patients,relevant integration frameworks will be valuable for explaining the molecular pathogenesis and underlying biology in cancer, eventually leading to more effective screening strategies and therapeutic targets in many types of cancer.

Ilja Arts

Ilja Arts, Prof.Dr. I.C.W. is Professor of Molecular Epidemiology of Chronic Diseases and affiliated with the Department of Epidemiology of Maastricht University since 2006. Her research focuses on the molecular epidemiology of chronic diseases, with a particular interest in the etiology and prognosis of diabetes and cardiovascular diseases. Molecular epidemiology entails the use of profiles of biomarkers or functional assays in epidemiological studies to improve exposure and outcome assessment, to elucidate biological mechanisms underlying epidemiological associations, and to enable better risk stratification leading to personalized health.
Prof. Arts is co-founder and co-chair of the Maastricht Molecular Epidemiology Expertise group (M2E2), and program leader in the newly established Maastricht Centre for Systems Biology (MaCSBio), where she develops methods for the integration of different types of –omics data in epidemiological studies using a systems biology approach. This interdisciplinary research is conducted in close collaboration with laboratory groups, clinical groups, and with groups specializing in complex data-analysis. She is involved in three ongoing cohort studies: the KOALA Birth Cohort Study, CODAM, and the Maastricht Study.
Prof. Arts was a research fellow at Wageningen University, and the National Institute of Public Health and the Environment in Bilthoven (The Netherlands), at the School of Public Health, University of Minnesota in Minneapolis (USA), and at the Department of Pediatrics, Uppsala University, Uppsala (Sweden). She has published more than 60 peer reviewed articles and her h-index is 32.

Uncovering population heterogeneity in the omics era

Anavaj Sakuntabhai



Anavaj SAKUNTABHAI, Dr. heads an internationally recognized research laboratory at the Institut Pasteur, is a medical doctor with 10-years experience in clinical medicine before starting his career in basic sciences research. He obtained the diploma of doctor of philosophy (D. Phil.) in human molecular genetics from University of Oxford, United Kingdom in 1999. He discovered a gene responsible for Darrier's disease, a monogenic skin disorder. He was then appointed as a senior scientist of the Institut Pasteur in 2000 to develop a program on the genetics of infectious diseases. He discovered a variant on a promoter of DC-SIGN associated with gene expression and outcome of dengue infection. He published an important finding of positive selection of G6PD (glucose 6 phosphate dehydrogenase) and its effect on Plasmodium vivax density in Science. His recent research has shown that both gene-gene and gene-environmental interactions play a significant role in susceptibility to malaria and dengue.

Dr SAKUNTABHAI has significant experience in the coordination of international projects. He successfully coordinated two important projects on genetic susceptibility to malaria and dengue involving teams from France, Thailand, and Senegal. He coordinated a global network for dengue research for the Institut Pasteur International Network. He is a principle investigator of one of the four consortial projects of the Bill and Melinda Gates financed MalariaGEN consortium.

He is now a coordinator of European FP7 project on Dengue Framework for Resisting Epidemics in Europe (DENFREE). The project aims at finding key factors determining dengue transmission and dengue epidemics in order to develop new tools and strategies for controlling dengue transmission.

Case study – Understanding Interactions in Complex trait: the malaria story

The genome wide association study (GWAS) approach, which relies on commercially available chips that represent common human variation to test for association with disease, has been successful in the effort to uncover the genetic basis for several common diseases. Success of the approach, involving testing of each marker individually, largely rests on high statistical power brought to the case-control design through large sample sizes, aided by disease-specific consortia and meta-analyses that have confirmed susceptibility loci with increasingly lower effect sizes. Nonetheless, a large proportion of estimated heritability for most common diseases remains unexplained after consideration of GWAS hits. The genetic architecture of common diseases is clearly highly complex, likely involving a combination of common and rare variants, as well as the interplay of multiple variants and the environment. New methods are required to optimally characterize this complexity.

In a case of malaria, there were several GWAS studies reported which identified two well known genetic factors, sickle cell and ABO blood group and discovered new gene with very low effect size. With the polygenic and multi-factorial aspect of malaria disease, one should allow for hypotheses that assume a cumulative and/or interactive force of several distinct genes and environmental factors, each having a weak marginal effect on the outcome of malaria infections. Here, we present a genetic study of family based, longitudinal follow up malaria cohorts. With model-free exhaustive search, we could identify several interesting gene-environmental interactions. We developed a family based method to test deviation from Mendel's law of allelic inheritance among a sample of offspring at a set of makers together. We could identify an interaction among three genes.

Manuel Hidalgo

Gastrointestinal (GI) Cancer Clinical Research Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Incorporating Mouse Models into Pancreas Cancer Treatment

Taesung Park

Department of Statistics, Seoul National University College of Natural Sciences, Seoul, South Korea

International multicenter study to characterize the individual risk of malignancy in branch duct IPMN and proposal of nomogram

Branch duct type Intraductal papillary mucinous neoplasms (BD-IPMNs) have a diverse pathologic spectrum. It is difficult to predict malignancy preoperatively using imaging or biologic diagnostic tools. Previous reports on malignancy predictors of BD-IPMN have used different definition and shown controversial results on same variables. The purpose of this study was to elucidate the malignant predictor and evaluate individual risk for malignancy and finally suggest nomogram for malignancy prediction of BD-IPMN using world largest DB of IPMN by Korea-Japan collaboration study group. We used multicenter data from Korea and Japan to find out useful variables to build nomogram for prediction of risk of malignancy. 1170 samples were from 13 Japanese hospitals and 744 samples were from 9 Korean hospitals. Malignant IPMNs were defined as those with noninvasive and invasive intraductal papillary mucinous carcinoma. 806 patients had adenoma, 374 borderline, 368 noninvasive carcinoma, and 366 invasive carcinoma. We had 9 variables: age, sex, CEA, CA19-9, pancreatitis, location (head, body/tail, and diffuse), main duct dilatation, cyst size, and mural nodule. To evaluate the performance of nomogram, we split the data into two groups in the ratio of 2 to 1 in each hospital with the consideration of proportion of malignancy. In the first group, we also divided the data evenly 200 times generating training set and validation set. In each splitting, using age and sex as covariates (also, only using age as a covariate), we built a logistic regression model using training set and calculated AUC using validation set based on the model for all combination of remaining 7 variables. Here, we used the value of natural logarithm for CEA and CA19-9.

Then, we sought for the combination that had maximum value of AUC in each divided set. Among 200 combinations, the combination for log(CEA), log(CA19-9), main duct dilatation, cyst size, and mural nodule presented the most (91 times). Also, the mean value of AUC was higher when we used age rather than both age and sex as a covariate. We built nomogram based on selected 6 variables using all the first group. We evaluated the performance of the nomogram using the second group as a test set. The value of the AUC for the nomogram is 0.7366. Also, we found out the probability cut-off value to classify malignancy and benign having maximum balanced accuracy in the first group. At the probability cut-off of 0.37, the nomogram has 0.6706, 0.7004, and 0.6408 as the value of balanced accuracy, sensitivity, and specificity. We propose malignancy predicting nomogram for BD-IPMN using meaningful variables through logistic regression model. It would be very useful to select optimal treatment.

Andrew Biankin

Translational Research Centre, University of Glasgow, UK

The importance of Integrating, transforming and sharing data to unravel new complex disease mechanisms

Jean-Luc Van Laethem

Laboratory of Experimental Gastroenterology, Erasmus Hospital, Brussels, Belgium; EORTC Gastrointestinal Tract Cancer Group (GI Group)

Raphaël Maréchal

Laboratory of Experimental Gastroenterology, Erasmus Hospital, Brussels, Belgium; EORTC Gastrointestinal Tract Cancer Group (GI Group)

European (EORTC) organization of research in PC. Perspectives in PC research from the clinical researcher

Fatima Al-Shahrour

Fatima Al-Shahrour (Madrid, 1975) obtained her PhD from Universidad Autónoma de Madrid (UAM) in 2006. During her PhD she worked at the Bioinformatics Unit at Spanish National Cancer Research Center (CNIO, Madrid, Spain) and Centro de Investigaciones Príncipe Valencia (Valencia, Spain) under Dr. Joaquín Dopazo supervision. During this period, her research work dealt with the development of new Bioinformatics tools for microarray gene expression analysis, with a particular focus on computational methods for the functional interpretation of high-throughput experiments.

In 2007, she was awarded with the José Castillejo mobility research grant from the Spanish Ministry of Science and Innovation to join the Computational Biology and Bioinformatics group at Cancer Program under supervision of Prf. Dr. Jill P. Mesirov at Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard (Cambridge, USA). Her project was focusing mainly in the development of a computational methodology for identifying molecular signatures of oncogenes and tumor suppressor genes based on gene expression data.

Bioinformatics approaches for personalized cancer therapy: From Pan-Cancer projects to Patient derived xenografts (PDX) models.

The success of personalized treatment of cancer patients depends on matching the most effective therapeutic regimen with the characteristics of the individual patient, balancing benefit against risk of adverse events. The primary challenge in achieving this goal is the heterogeneity of the disease, recognizing that the majority of cancers are not single diseases but rather an array of disorders with distinct molecular mechanisms.

High-throughput technologies such as DNA next generation sequencing are being used to dissect the genetic heterogeneity of tumors, and in parallel, Bioinformatics has emerged as a critical discipline to transform the huge amount of genomic data in comprehensive models. However, these analyses have resulted in the identification of hundreds (or thousands) of mutations and other alterations in the same tumor; therefore, we need new approaches to establish the relevance of these changes, and more importantly, to prioritize those that could be clinically useful for cancer therapy.

Here, we present an integrative bioinformatics approach to demonstrate the value in integrating genomic and clinical data with available drugs, to refine and to improve therapeutic strategies for cancer patients. Publicly available data from Pan-cancer project and PDX models generated at our institution are used to validate this new approach.

CLASSIFICATION USING DIFFERENTIAL NETWORK RANK CONSERVATION REVISITED

Askar Obulkasim[1], Maarten Fornerod[1], Michel C Zwaan[1,2], Marry M van den Heuvel-Eibrink[1,2]

[1] Department of Pediatric Oncology/Hematology, Erasmus-MC Sophia Childrens Hospital, The Netherlands
[2] Dutch Children's Oncology Group, Erasmus-MC Sophia Children's Hospital

Many characteristics of transcriptomic data, such as redundant features and technical artifacts, make over-fitting commonplace. Promising classification results often fail to generalize across datasets with different sources, platforms, or preprocessing. In their pioneering work Eddy et al. (2010) [1] proposed a novel differential network rank conservation (DIRAC) algorithm to characterize cancer phenotypes using transcriptomic data. DIRAC is a member of a family of algorithms that have shown useful for disease classification based on the relative expression of genes. Combining the robustness of this family's simple decision rules with known biological relationships, this systems approach identifies interpretable, yet highly discriminate networks. While DIRAC has been briefly employed for several classification problems in the original paper, the potentials of DIRAC in cancer phenotype classification, and especially robustness against artifacts in transcriptomic data have not been fully characterized yet.

We thoroughly investigate the potentials of DIRAC by applying it to multiple real-world datasets and examine the variation in classification performances when datasets are treated and untreated for batch effect. Performances of the DIRAC-based classifier when the dataset is preprocessed with different techniques are also investigated. We also propose the first DIRAC-based classifier to integrate multiple networks as features using standard machine learning methods. we show that the DIRAC-based classifier is very robust in the examined scenarios. Therefore, we propose that DIRAC is a promising solution to the lack of generalizability in classification efforts that uses transcriptomic data.

We recommend the DIRAC-based classifier when the choices of preprocessing and batch effect correction methods are not obvious, either by limited statistical resources or due to limited data availability. it is, for example, often the case that raw data are not publicly available. Thus, albeit numerous databases exist to store tremendous amounts of genomic data, it is surprisingly difficult to find a dataset that has been preprocessed in exactly the same way as the one from which new findings were discovered and are in need to be validated. As we demonstrated in this study, these issues, to a large extent, are ameliorated when the DIRAC-based classifier is used. To our surprise, the DIRAC-based classifier even translated well to a dataset with different biological characteristics in the presence of substantial batch effects that, as shown here, plagued the standard expression value based classifier. In addition, the DIRAC-based classifier also suggests pathways to target in specific subtypes, which may enhance the establishment of personalized therapy in diseases such as cancer.

[1] Eddy, J. A.; Hood, L.; Price, N. D. et al. (2010), PLoS Comput Biol. 6: 5923--5928.
[2]Walpert, D. H. (1996), Neural Computation. 8: 1341-1390.

COMBINING GENOTYPE WITH LD-BASED HAPLOTYPE INFORMATION AS INPUT FOR ITERATIVE PRUNING PRINCIPAL COMPONENT ANALYSIS (IPPCA) TO IMPROVE POPULATION CLUSTERING

Kridsadakorn Chaichoompu[1,2], Ramouna Fouladi1,2, Pongsakorn Wangkumhang3, Alisa Wilantho3, Wanwisa Chareanchim3, Philip James Shaw3, Sissades Tongsima3, Anavaj Sakuntabhai4, Kristel Van Steen1,2

[1]Montefiore Institute, University of Liege, Belgium
[2]GIGA-R, University of Liege, Belgium
[3]Genome Institute, National Center for Genetic Engineering and Biotechnology, Thailand
[4]Functional Genetics of Infectious Diseases Unit, Institut Pasteur, France

Correspondence: kridsadakorn.chaichoompu@ulg.ac.be

Key words: ipPCA, population clustering, LD-based haplotype

Single Nucleotide Polymorphisms (SNPs) are commonly used to capture variations between populations and often genome-wide SNP data are pruned based on linkage disequilibrium (LD) patterns. To identify and differentiate between subpopulations using a rich set of genetic markers, as using reduced sets of genetic markers for these purposes, can become challenging especially when similar geographic regions are involved or when spurious patterns are likely to exist. Notably, haplotype composition and the pattern of LD between markers may vary between larger populations but may also play a role within more confined geographic regions. Indeed, the structure of haplotypes in unrelated individuals can reveal useful information about genetic ancestry.

Here, we use iterative pruning principal component analysis (ipPCA) [1] to identify and characterize subpopulations in an unsupervised way. Furthermore, we purpose to combine an LD-based haplotype encoding scheme with the ipPCA machinery to retrieve fine population substructures. Despite the complexities that are associated with haplotype inference, added value can be obtained when the LD structure between SNPs is exploited in the search for relevant population strata. As input data, either pruned genome-wide SNP data are used or multilocus haplotype information derived from the genome-wide SNP panel. Preliminary results indicate that ipPCA applied to pruned SNP data or ipPCA that explicitly uses multilocus information (haplotypes) give complementary information about population substructure for geographically confined populations. In fact, both methods address different aspects of population structure.

[1] Intarapanich, A. et al. (2009), BMC Bioinformatics. 10: p. 382.

mRNA-seq of 278 Diverse Skeletal Muscle Biopsies Reveals Mechanistic Insights about Type 2 Diabetes Genetic Risk and Identifies Disease State Specific eQTLs

Jeroen R. Huyghe[1], Stephen C.J. Parker[2], Michael R. Erdos[2], Heikki Koistinen[3], Peter S. Chines[2], Ryan Welch[1], Xiaoquan Wen[1], Hui Jiang[1], Narisu Narisu[2], Leland Taylor[2], Brooke Wolford[2], Laura J. Scott[1], Heather Stringham[1], Leena Kinnunen[3], Tom Blackwell[1], Anne U. Jackson[1], Yeji Lee[1], Amy J. Swift[2], Lorry Bonnycastle[2],
Michael L. Stitzel[4], Richard M. Watanabe[5,6], Karen Mohlke[7], Timmo Lakka[8], Markku Laakso[8], Jaakko Tuomilehto[3], Francis S. Collins[2], Michael Boehnke[1]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, USA
[2]National Human Genome Research Institute, National Institutes of Health, USA
[3]National Institute for Health and Welfare, Finland
[4]The Jackson Laboratory for Genomic Medicine, USA
[5]Department of Preventive Medicine, University of Southern California (USC) Keck School of Medicine, USA
[6]Department of Physiology and Biophysics, Keck School of Medicine of USC, USA
[7]Department of Genetics, University of North Carolina at Chapel Hill, USA
[8]University of Eastern Finland, Finland

Correspondence: jhuyghe@umich.edu

Key words: genome-wide association studies, eQTL, type 2 diabetes, transcriptome sequencing, mRNA-seq

Type 2 diabetes (T2D) is a complex disease caused by an interplay between genes, environment, and behavioral factors, acting over time and across multiple tissues. Genome-wide association studies (GWAS) have identified >80 loci associated with T2D risk. For most identified loci, the causal gene and functional variant(s) remain elusive because the associated region resides in noncoding DNA, suggesting a major contribution of transcriptional regulatory elements to disease risk. Regulatory element usage is often tissue-specific. Therefore, a crucial next step to guide the functional follow-up of GWAS is to determine the relationship between single nucleotide polymorphisms (SNPs) associated with T2D or related traits, and gene expression in disease-relevant tissues and across disease progression. As part of the Finland-United States Investigation of NIDDM Genetics (FUSION) study, we obtained vastus lateralis skeletal muscle biopsies from 278 clinically well-characterized Finns with normal and impaired glucose tolerance, and with newly diagnosed T2D without antihyperglycemic medication. Skeletal muscle is a major insulin target tissue and accounts for ~25-30% of postprandial glucose uptake. We constructed and sequenced strand-specific mRNA-seq libraries (mean depth 55 million 101 base read pairs) and performed dense genotyping and imputation. We identified >8000 genes with expression and/or splicing quantitative trait loci (e/sQTL) (5% FDR). Some of these eQTL (e.g., for the genes TTNT3 and SDCCAG8) appear disease state specific. Multiple eQTL are in high linkage disequilibrium with GWAS SNPs for T2D or related traits, highlighting genes at these loci as probable candidates for a role in T2D risk. Interestingly, for a subset of these GWAS SNP overlapping eQTL, gene expression is also significantly associated with T2D or a glycemic trait. E.g., T2D GWAS index SNP rs516946 is the most significant eQTL SNP for the ANK1 gene, which is differentially expressed between normal glucose tolerant vs. T2D individuals. Similarly, our eQTL analysis points out CCHCR1, associated to glycemic traits and BMI, as a candidate gene for the T2D GWAS index SNP rs3130501. This rich data resource enables identification of diverse molecular processes involved in skeletal-muscle-based insulin resistance and changes in gene transcription with progression towards T2D, and reveals mechanistic insights about T2D risk.

Integrative visual analysis of genomic data on Spotfire

Pablo Riesgo

Visual analytics is a key step in a data analysis pipeline, and in particular for the integrative analysis of results from different sources. This is specially true in the post-genomic era as the amount of omics technologies increase. Spotfire is a business intelligence (BI) platform that allows to integrate diverse sets of data and visualise them graphically in highly interactive dashboards to facilitate the visual exploration. We have developed solutions to integrate genomic data in Spotfire. In the oncology context, we present a proof of concept that integrates differential expression data with gene driver selection approaches and simple variant association, combining different tertiary analysis from DNAseq and RNAseq.

Regulatory gene network inference from expression data via conditional inference trees

Kyrylo Bessonov[1], Kristel Van Steen[2]

[1] Systems and Modeling Unit, Montefiore Institute, University of Liege, B-4000 Liege, Belgium
[2] Bioinformatics and Modeling, GIGA-R, University of Liege, B-4000 Liege, Belgium

Correspondence: kbessonov@ulg.ac.be

Key words: network inference; gene regulation, transcriptome; microarrays, microarray expression data, conditional inference trees and forests, biological interactions;

Trees are classical data structures allowing to effectively classify and predic responses. Due to their versatility and high performance in classification and prediction, plenty of tree-based methods exist, including popular Conditional Inference Tree (*CIT*) and Forests (*CIF*) [1,2], Random Forests (*RF*), Randomized Trees (*RT*), randomized C4.5, etc. In this work we assessed the performance of tree-based methodologies (*CIT*s, *CIF*s, and Random Forests *RF*s [4], *GENIE3* [5]) with each other and with Weighted Gene Coexpression Network Analysis (*WGCNA*) [3] to infer both directed and undirected gene-gene expression networks. Our results show that CIF_{mean} methodology based on *CIF* [1,2], with mean variable importance averaging across trees, is the best performer under directed gene regulatory network (GRN) contexts. The second best performer is random forests (*RF*) for synthetic microarray datasets with either 100 or >1000 genes (respectively, *DREAM4* and *DREAM5* data). Analysis of real-life *E.coli K12* microarray expression data (1637 genes x 24 samples) [6] showed top performance and practical applicability of the CIF_{mean} methodology, especially under directed network inference contexts. We believe that our work is of interest to those experts aiming to infer biological interactions from transcriptome data via machine learning approaches.

[1] Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinformatics 9: 307.
[2] Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8: 25.
[3] Horvath S, Dong J (2008) Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4: e1000117.
[4] Breiman L (2001) Random Forests. Machine Learning 45: 5-32.
[5] Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods. PLoS One 5
[6] Chen T, Wang J, Zeng L, Li R, Li J, et al. (2012) Significant rewiring of the transcriptome and proteome of an Escherichia coli Strain Harboring a Tailored Exogenous Global Regulator IrrE

Integration of gene-based and text mining analyses to discover genes underlying atopy

P-E Sugier[1,2,3], A. Vaysse[1,2], C. Sarnowski[1,2], C. Loucoubar[1,2], P. Margaritte-Jeannin[1,2], M-H Dizier[1,2], M. Lathrop[4], F. Demenais[1,2], E. Bouzigon[1,2]

[1]INSERM, UMR-946, Paris, France
[2]Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, Paris, France
[3]UPMC, Paris, France
[4]McGill University and Génome Québec Innovation Centre, Montréal, Canada

Correspondence: pierre-emmanuel.sugier@inserm.fr

Key words: Skin prick test, atopy, genome-wide association study, gene-based tests, data mining

The prevalence of allergic diseases such as asthma has reached pandemic proportions in industrialized countries. The genetic component of allergy is substantial but has been rarely investigated at the genome-wide level. Two genome-wide association studies (GWAS) of atopy defined as increased specific IgE levels or skin prick test (SPT) reactivity to allergens have been conducted in population-based cohorts with inconsistent results [1,2].
We aimed to identify genetic determinants of atopy using a GWAS approach combining a single SNP analysis followed by 1) gene-based association tests and 2) data-mining analysis.
We conducted GWAS of atopy phenotypes in 1,660 subjects (925 atopics and 735 non-atopics) from the French Epidemiological study on the Genetics and Environment of Asthma (EGEA) which includes families ascertained through asthmatics. These subjects were genotyped with Illumina 610K Array. Four atopy traits were investigated: 1) atopy as a whole and defined by a positive SPT response to at least one of 11 aeroallergens, and then three groups of aeroallergens: 2) indoors, 3) outdoors, and 4) molds.
For each phenotype, we first performed a genome-wide single SNP analysis. Then, we conducted gene-based association tests using VEGAS (Versatile Gene-based Association Study) [2]. VEGAS method combines single SNPs results within a gene and accounts for linkage disequilibrium between markers by using simulations from the multivariate normal distribution to produce a gene-based test statistic. Finally, a data-mining method using GRAIL (Gene Relationships Among Implicated Loci) [3] was applied to the top SNPs detected at P<10-4 by single SNP analysis. GRAIL takes a list of disease regions and assesses a degree of relatedness of implicated genes using text-based metric build from words included in articles of the scientific literature in which these genes are referenced (PubMed abstracts until December 2006).
Six atopy loci were detected at P-value≤5x10-7 by single SNP analyses: 3q24, 5q13, 12q24, 4q24, 15q13 and 19q13. Our best signal was located in GPR98 (5q13, P=2.9 x10-7) and was closed to the genome wide significant level (P=1.25x10-7). Gene-based tests both strengthened the evidence for association of 5q13 and 19q13 loci, and increased evidence for two other loci on chromosome 19. Using GRAIL approach, we identified a relationship between GPR98 and TYRP1 genes (PGRAIL=1.6x10-3), this latter gene was previously detected at P=5x10-5 by single SNP analysis.
Further investigations are needed to confirm the hypothesis of potential interactions between GPR98 and TYRP1 genetic variants. SNP-SNP interaction tests are currently conducted across these two genes and will be presented. This study highlights that combining at the genome-wide level single SNP analyses and multi-marker analyses may facilitate the identification of new susceptibility genes and could have a best power to detect epistasis by a previous genes selection to reduce both the computational burden and the statistical burden of an exhaustive search.

[1] Andiappan, A. K.; Wang de, Y.; Anantharaman, R. *et al.* (2011), PLoS One. 6(5):e19719
[2] Ramasamy, A.; Curjuric, I.; Coin, L. J. *et al.* (2011). J Allergy Clin Immunol. 128(5):996-1005.
[3] Liu, J. Z.; McRae, A. F.; Nyholt, D. R. *et al.* (2010), Am J Hum Genet. 87(1): 139-45
[4] Raychaudhuri, S.; Plenge, R. M.; Rossin, E. J. *et al.* (2009) PLoS Genet. 5(6):e1000534

Integration analysis of 'OMICS' data using penalized regression methods: An application to bladder cancer

Silvia Pineda[1,2], Núria Malats[1] , Kristel Van Steen[2]

[1] *Spanish National Cancer Research Center (CNIO), Madrid, Spain.*
[2] *University of Liege (ULg), Liege, Belgium.*

Key words: LASSO, ENET, permutation-based test, omics

Combining different 'omics' data such as common genetic variation, DNA methylation, and gene expression may allow discovering new biological mechanisms of complex diseases. In the cancer field, the development and progression of a tumor is the consequence of multiple processes and alterations including gene aberrations, epigenetic changes, modifications in gene regulation, environmental influences, etc. To integrate all this information, advanced statistical techniques are being developed and novel techniques are continuously emerging. Moreover, interpretation and validation of new biological data becomes an important challenge.

In this work, where large-sample statistics can no longer be used, we rely on variable selection methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) approach and the Elastic Net method to obtain sparse models with better precision, accuracy and statistical power. These methods can control also for multicollinearity that may arise from high correlated 'omics' features. Although promising in the context of high-throughput data, one of their drawbacks is that they do not provide p-values to assess statistical significance of relationships, nor give a formal assessment of the overall goodness-of-fit. Therefore, we adopted a permutation-based strategy to assess significance of discovered relationships combined with a FWER multiple testing correction approach (maxT algorithm) building upon the statistical concept of "deviance". Our strategy was illustrated on the pilot Spanish Bladder Cancer/EPICURO study (27 bladder cancer cases recruited in 2 hospitals in Spain in 1997-1998). The aim was to assess how much the variability in gene expression was explained by DNA methylation and genome-wide SNP data measured in tumor samples. We detected significant genes when using SNP data and DNA methylation data individually to explain gene expression levels. Additional results were highlighted when combining the three data sets, suggesting the importance of integrating 'omics' data. Moreover, ENET selected different significant genes than LASSO, suggesting the difference in the correlation structure between and within DNA methylation and SNP data. In conclusion, applying advanced statistical methods and adopting novel strategies to integrate high-throughput data gives us the opportunity to gain new insights in the development and progression of complex diseases.

NOVEL THERANOSTIC NANOCOMPLEX FOR PANCREATIC CANCER: PREPARATION AND CHARACTERIZATION OF ANTI-MESOTHELIN ANTIBODY ORIENTED AND GAMBOGIC ACID BIOCONJUGATED IRON BURIED NANOLACTOFERRIN

Lütfi GENÇ[1,2*], Şennur GÖRGÜLÜ KAHYAOĞLU[2,3], Betül MÜJDECİ[2], Emel ERGENE[4], Arzu Ersöz[5], Rıdvan SAY[5]

[1]Anadolu University, Faculty of Pharmacy, Department of Pharmaceutical Technology, Eskişehir. TÜRKİYE
[2]Anadolu University, Plant, Drug and Scientific Researches Center, Eskişehir, TÜRKİYE
[3]Anadolu University, Department of Pharmacology, Eskişehir, TÜRKİYE
[4]Anadolu University, Department of Biology, Eskişehir, TÜRKİYE
[5]Anadolu University, Department of Chemistry, Eskişehir, TÜRKİYE
**Correspondence: lgenc@anadolu.edu.tr*

Key words: Lactoferrin, Anti-mesothelin antibody, Theranostic, Pancreatic cancer, ANADOLUCA method

Pancreatic cancer is one of the worst mortality cancer species which has poor diagnose at an early stage, difficult detection and surgery. The underlying problem in the use of anticancer drug is their toxicity and poor bioavailability. In spite of, there are a lot of anti-cancer drug in the market, they are not selective and they affect both cancer cells and normal body cells at the same time. This, development of novel, specific, tumor targeted drug delivery systems is urgently needed for this terrible disease. Antibody based drug delivery systems can not show long term and effective availability according to their brittle structure [1].

Nanotechnology is able to provide nanoparticulate drug delivery systems (NPDDSs) for treatment, diagnosis and imaging of diseases. This is an attempt to cover the recent trends and emerging technologies in the area of NPDDSs namely nanosuspensions, polymer-based nanoparticles, nanocrystals, nanospheres, nanocapsules, lipid-based nanoparticles, nanocomplex, quantum dots and etc. [2]. The current progress on targeted drug delivery in the diagnosis and theraphy of pancreatic cancer are included the therapeutic agents (Gene theraphy, suicide gen theraphy, small molecule inhibitor, antibody theraphy, etc.), drug delivery systems and potential targets for specific delivery. Multi functional magnetic nanoparticles have been used for early diagnosis, drug delivery and biomedical purposes [3,4].

Lactoferrin (Lf) is a multi functional bioactive glycoprotein (composed of a protein and a carbohydrate) found in breast milk and in small quantities in most body fluids [5,6]. Lf prevents the production of tumor necrosis factor (TNF) and interleukin 1 and 6 from monocytes [7,8]. These functions become by conjugating of Lf in suitable receptor in the cell membrane. Additionally, these receptors are found in the T and B cells, NK cells, Pancreatic cells, blood cells, intestinal cells and brain cell membrane [5]. This is important starting point to develop Lf based targeted drug and imaging systems. In this study, nanolactoferrin drug platform (NanoLf) are prepared as a new generation polymeric material by using crosslinking polymerization technique, called ANADOLUCA method [9].

In this manner, new generation theranostic nanocomplex can be improved to show both flourescense and magnetic properties. And then, anti-mesothelin antibody are oriented and cross-linked, and iron atoms buried and gambogic acid conjugated to the NanoLf drug platform. These nanocomplexes have too many advantages such as narrow particle size, light sensitivity, reusability, durability, therapy and imaging functions. Being multifunctional of these nanostructures show that they can be targeted to the tumor cells for imaging, diagnosis and therapy [10,11].

[1] Wang, C., Zhang, H., Chen, Y. et al. (2012), Int. J. Nanomedicine. 2012(7), 781-787.
[2] Sinha, R., Kim, G.J., Nie, S. et al. (2006), Mol. Cancer Ther.5 (8), 1909.
[3] McCarthy, J.R. and Weissleder, R. (2008), Adv. Drug Deliv. Rev. 60(11), 1241-51.
[4] Mahapatro, A. and Singh, D.K. (2011), J. Nanobiotechnology. 9, 55.
[5] Suzuki, Y. A., Lopez V., Lönnerdal, B. (2005), Cell. Mol. Life Sci. 62, 2560–2575.
[6] Ward, P. P., Paz E., Conneely O. M. (2005), Cell. Mol. Life Sci. 62, 2540–2548.
[7] Puddu, P., Latorre, D., Valenti, V. et al. (2010), Biometals. 23:387–397.
[8] Legrand, D., Elass, E., Carpentier, M. et al. (2005), Cell. Mol. Life Sci. 62 2549–2559.
[9] Say, R. Patent (2009), WO2011070402 A1.
[10] Say, R, Aydoğan Kılıç, G., Atilir Ozcan, A. et al. (2011), Histochem Cell Biol. 135, 523-530.
[11]Say, R, Uzun, L., Yazar, S. et al. (2014), Artificial Cells, Nanomedicine, and Biotechnology, 42, 138-145.

Are we far from correctly inferring gene interaction networks?"

Francesco Gadaleta[1,2]

[1] Systems and Modeling Unit, Montefiore Institute, University of Liege, B-4000 Liege, Belgium
[2] Bioinformatics and Modeling, GIGA-R, University of Liege, B-4000 Liege, Belgium

Genome-wide association studies can potentially unravel the mechanisms behind complex traits and common genetic diseases. Despite the valuable results produced thus far, many questions remain unanswered. For instance, which specific common variants are linked to the risk of the disease under investigation, what biological mechanism do they act through or how do they interact with environmental and other external factors? The driving force of computational biology is the constantly growing amount of big data generated by high-throughput technologies. The amount of available data and its heterogeneity seem to play a beneficial role rather than a detrimental one in discovering new genetic insights. Each type of data directly contributes to complete the overall puzzling picture of genetic disorders, with its own unique local knowledge. In such a scenario, data integration, as the practice
of combining evidence from different data sources, represents the most challenging activity, due to the unattainable task of merging large and heterogeneous data sets. A practical framework that fulfils the needs of integration is provided by means of networks. Variable selection is a fundamental step to mitigate the curse of dimensionality, a common issue in computational biology. Experiments on synthetic data show that networks are helpful in detecting genetic compounds that are potentially implicated in the disease under study.

Network-Assisted Investigation of Signals from Genome-Wide Association Studies in Childhood-onset Asthma

Y. Liu[1,2], M. Brossard[1,2,3], P. Margaritte-Jeannin[1,2], F. Llinares[4], C. Sarnowski[1,2,3], L. Al-Shikhley[1,2], N. Lavielle[1,2], A. Vaysse[1,2], M.H. Dizier[1,2], E. Bouzigon[1,2], F. Demenais[1,2]

[1]U946, INSERM, Paris, France
[2]Université Paris Diderot, Paris, France
[3]Université Paris Sud, Paris, France
[4]ETH, Basel, Switzerland

Correspondence: yuanlong.liu@inserm.fr

Key words: GWAS, multi-marker analysis, human protein interaction network, network-assisted analysis

Genome-Wide Association Studies (GWASs) have consisted of testing association of disease with single SNPs and of highlighting those SNPs reaching a stringent genome-wide significant level. In comparison, the joint analysis of multiple SNPs (multi-marker analysis) allows detecting sets of SNPs with small effect and can provide more insight into the biological mechanisms underlying disease. In this study, we focused on multi-marker analysis integrating biological knowledge based on the Human Protein Interaction

Network (HPIN). Two datasets were used for analysis in a discovery-evaluation scheme. These datasets consisted of the outcomes (single SNP association statistics) of two meta-analyses of 9 childhood-onset asthma GWASs each (total of 3,031 cases / 2,893 controls for meta-analysis 1 and 2,679 cases / 3,364 controls for meta-analysis 2), that were part of the large GABRIEL-European Asthma Consortium [1]. GWAS signals were first overlaid to HPIN, by assigning intragenic SNPs to genes and using as gene-wise P-value either the best SNP P-value (as classically used) or an empirical P-value computed from Circular Genomic Permutation (CGP)[2] to adjust for gene length while preserving linkage disequilibrium (LD) between SNPs. A Dense Module Search algorithm (DMS)[3] was applied to generate gene modules enriched in GWAS signals. We modified the original DMS implementation by adding a hierarchical merging step to reduce module overlap and redundancy within each dataset. Further, to select consistent modules, we computed pairwise module similarity between datasets. Module pairs with high similarity were then evaluated for significance of 1) gene modules and 2) association of modules with disease in both discovery and evaluation datasets.
As a result, 20 significant gene modules (P≤0.05 after Bonferroni correction for multiple hypothesis testing) were identified with each method and consisting of 239 or 184 unique genes when the initial gene-wise P-value was that of the best SNP within the gene or was derived from CGP respectively. These two sets of unique genes shared 34 genes in common, which shows the sensitivity of DMS method to the gene-wise P-value used. Use of a gene-based statistic that is adjusted for gene length and LD may be of importance to reduce bias and potentially false discoveries. Further evaluation of our modified DMS approach is needed. Moreover, the set of genes characterized by the present study require further assessment of their association and potential interaction underlying childhood-onset asthma.

[1] Moffatt, M. F.; Gut, I.G.; Demenais, F. et al. (2010), N. Engl. J. Med.,363:1211-1221.
[2] Cabrera, C. P.; Navarro, P.; Huffman, J. E. et al. (2012), G3 Bethesda Md. 2: 1067-1075.
[3] Jia P.; Zheng S.; Long J. et al. (2011), Bioinformatics, 27: 95-102.

A novel gene-based analysis method based on MB-MDR

Ramouna Fouladi [1 2] , Kyrylo Bessonov [1 2] , Francois Van Lishout [1 2] , Kristel Van Steen [1 2]

[1] Systems and Modeling Unit, Montefiore Institute, University of Liege, Liege, Belgium
2 Bioinformatics and Modeling, GIGA-R, University of Liege, Liege, Belgium

A novel omics integrated association analysis framework is proposed that builds upon the Model-Based Multifactor Dimensionality Reduction (MB-MDR) method. This method primarily aimed at identifying higher-order interactions using large collections of SNP panels. Difficulties in interpreting results from SNP x SNP interaction studies, as well as concerns about replication, may be overcome in part by adopting a gene-based approach. At the basis of the so-called genomic MB-MDR method lies a data organization step that involves clustering of individuals according to features mapped to a region of interest (ROI). ROIs can be selected in different way, for instance based on functionality or entire genes. Features may include common and rare variants, as well as epigenetic markers. Any feature can be analyzed, irrespective of whether it is measured on a continuous scale or categorical scale, and features that are mapped to pre-selected ROIs are submitted to a clustering algorithm to find ROI-based groups of alike individuals. Subsequently, MB-MDR is applied to individuals that are organized into multi-locus ROI-based (e.g., gene-based) classes, in the same way that MB-MDR was used on individuals organized according to multi-locus SNP-based (genotype) classes.
Restricting attention to a genome-wide association setting and exome sequence data, we propose to first cluster individuals according to their similarities based on rare and common variants in preselected genes. In particular, per gene we transform the genotype data using kernel PCA and assess the number of homogeneous clusters of individuals using Dynamic Tree Cut on hierarchical clustering. Second, we apply MB-MDR in one-dimension. Genomic MB-MDR for rare variants appears to be a robust scalable method with acceptable overall performance in terms of power and type I error, under various realistic scenarios. In conclusion, several statistical methods have been proposed to uncover the association of rare variants with complex diseases, but none of them is the clear winner in all possible scenarios of study design and assumed underlying disease model. Application of MB-MDR to genes as unit of analysis, rather than SNPs as units of analysis, to discover new associations based on common and rare variants showed that it is a promising new rare variants analysis approach, with adequate power while maintaining low false positive rates. The assertions are based on synthetic data from GAW17 and comparison of our method with other state-of-the art rare variants analysis methods, including SKAT, SKAT-O, CMC, and VT.

Probabilistic Dimensionality Reduction for Heterogenous Data Integration

Max Zwiessele[1]

[1]*Department of Computer Science, Uiversity of Sheffield, UK*

Keyword: Gaussian Processes, Dimensionality, Reduction, Data Integration, Missing Data

In the time of next generation sequencing, we are able to extract immense amounts of data from biological systems in order to analyse the effects of microbiological phenotypes to clinical phenotypes, most importantly diseases. This data will inevitable enable us to increase curation efficiancy and reduce the amout of drugs necessary tailored to the individual. I am working on a data integration technique using Gaussian Process Latent Variable Models [1], using Bayesian integration to jointly model microbiological and clinical data from heterogeneous sources.

Handling heterogeneous data sources gives great advantage in data collection, but introduces massive amounts of missing data. In clinical data aquisition for example, it is not possible to administer eye tests, when a kidney failure is at hands. This means we need a model which can cope with and is consistent under missing data. Gaussian Processes are a great tool for handling missing data, as they are closed under marginalization, that means, that the marginals of measured experiments to not depend on missing experiments. The further addition of a variational approximation gives rise to an efficient algorithm, which scales with the amount of data we can aquire these days. We then proceed to variationally integrate out the latent inputs to the algorithm, which allows for a learning of the non-linear latent embedding of the observed data.

Extensions of the so called Bayesian GPLVM [2] allow for mutliple dataset integration (Manifold Relevance Determination [3]). This enables the model to learn shared and private latent space features automatically, which can be used to dissect datasets, or account for confounding variation, such as cell-cycle activity.

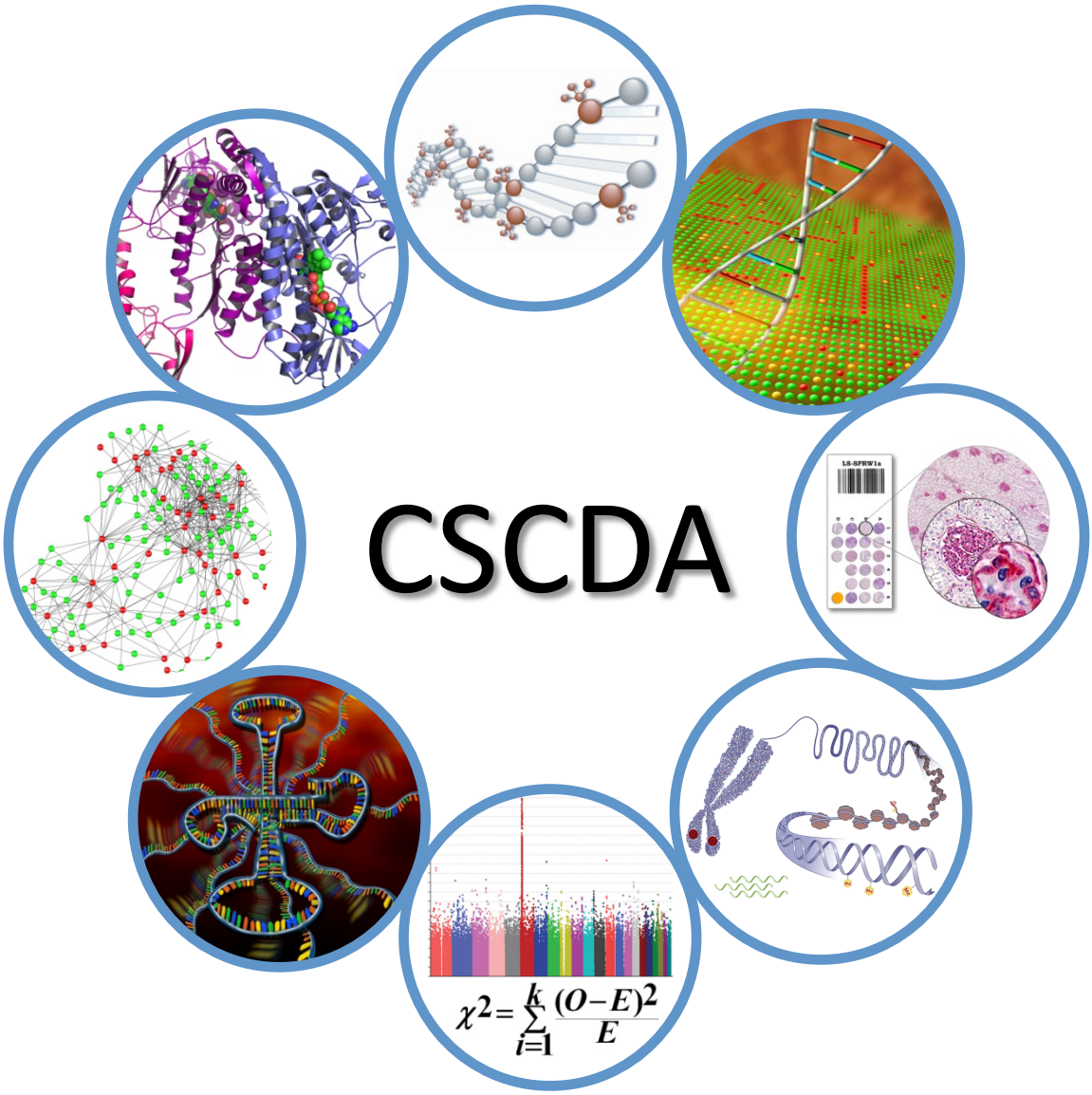
Future directions will be to include handling of missing data in a more principled way, allowing to learn expected distributions over missing entries.

Modeling heterogeneous data with massively missing values is a hard problem and can be tackles by using principled approaches such as probabilistic modeling and variational approximations.

[1] Lawrence, N. (2004), Advances in Neural Information Processing Systems, 16, 329-336.
[2] Titsias, M. K. and Lawrence, N. D. Bayesian Gaussian Process Latent Variable Model Artificial Intelligence and Statistics, 2010.
[3] Damianou, A. C.; Ek, C. H.; Titsias, M. K. and Lawrence, N. D. Manifold Relevance Determination ICML, 2012.

List of authors

CSCDA 2014



Thank you and see you at **CSCDA 2016** !!!

http://www.statgen.ulg.ac.be/