



**NATF / INAE Conference proceedings  
Technology and Healthcare,  
15 / 16 October, Genopole d'Éry**



## Table of contents

	Page
Introduction	V
Programme, list of participants and abstracts	VII
<b>Presentations</b>	
<b>Prof. Sanghamitra Bandyopadhyay:</b> <i>MicroRNA Induced Regulatory Network in Colorectal and Breast Cancer: A Machine Learning Perspective</i>	1
<b>Prof. Florence d'Alché-Buc:</b> <i>From experimental data to biological networks and vice versa</i>	31
<b>Dr. Sharmila Shekhar Mande:</b> <i>Algorithms, Metagenomics Analysis Platform, and their Applications for Understanding the Role of Gut-Microbiome in Human Health</i>	47
<b>Dr. Jean-Jacques Codani:</b> <i>Techniques of Data Analysis from NSG (Next Generation Sequencing)/Big Data and Life Science Data</i>	61
<b>Prof. Prasum Kumar Roy:</b> <i>Towards Health Grid Initiatives in India: Approaching a Grand Challenge in Affordable Health Care – Prospects and Insights from the Brain Grid</i>	87
<b>Dr. Pierre de la Grange/Dr. Frédéric Lemoine:</b> <i>Flexible and Efficient RNA-Seq Analysis at GenoSplice</i>	105
<b>Dr. Chandrasekhar Bhaskaran Nair</b> <i>Trulab<sup>TM</sup>: Evolution of a Near Care Molecular Diagnostics Platform for Infectious Disease Diagnostics</i>	123
<b>Dr. Jean-François Deleuze:</b> <i>CNG Missions and Technology</i>	135
<b>Dr. Goutam Ghosh,</b> <i>Opportunities in Research and Manufacturing of Bio-Pharmaceuticals</i>	155
<b>Dr. William Saurin:</b> <i>Software Applications for Modelling and Simulating Biological Systems: Application to Safety and Efficacy</i> <b>PRESENTATION NOT AVAILABLE</b>	
<b>Dr. Sudhir Rachapalle Reddi:</b> <i>Enhancing Eye Care Services through Innovation, Technology and Collaboration in India</i>	177
<b>Dr. Debapriya Dutta:</b> <i>Development of Indo-French Health Care Technology Network Programme - Role of CEFIPRA"</i>	211
<b>Dr. Christophe Ambroise:</b> <i>Statistical Study of Biological Networks</i>	229
<b>Dr. Rajeev Shorey:</b> <i>Engineering Challenges in Health Care Technologies and Services</i>	257
<b>Dr. Laurent Alexandre:</b> <i>DNA, Big Data and Physicians</i> <b>PRESENTATION NOT AVAILABLE</b>	
<i>Laboratory visits</i>	285



## Introduction

Following the cooperation agreement signed in 2013 with the INAE, Indian National Academy of Engineering, a symposium was held on 15/16 October at the Évry Genopole on the following topics:

- Genomics
- Bio-informatics

Organised under the joint leadership of Prof. Rajeev Shorey and Pierre-Etienne Bost, 7 Indian and 7 French experts have, during the course of a two-day seminar, exchanged their views and experiences and visited several laboratories at the Genopole.

The seminar was an opportunity for the two countries to highlight the complementarity of their research and development activities in:

- the arena of bio-informatics, including the modelling of biological networks and systems and in particular of neural networks,
- applications verifying the safety of drugs,
- micro-analytical applications,
- research in and production of biopharmaceuticals,
- the transmission of individual health-related data by health-care computer networks,
- establishing an Indo-French health service programme

Much of the research activities in bioinformatics in India and France is related to gene regulatory networks and the mechanisms by which they are induced, with applications to cancer research and research on other diseases but also in silico research assisted by machine learning and computer aided (modelling) inference of protein-protein interactions. Another area was the role of algorithms and metagenomics in understanding natural micro-biotic ecosystems such as that of the human gut and their impact on human health. CNG, the Centre National de Génotypage has one of the largest European capacities in high throughput genotyping and sequencing dedicated to human health to ultimately support personalised health management. Genosplice provides high-level bioinformatics solutions for the analysis of genomics data. In all this, big data is a pervasive subject with opportunities for machine learning but also the need for data selectivity when accuracy is important.

Furthermore, there was much emphasis on diagnostic tools for infectious diseases, production of biopharmaceuticals and generic vaccines at a massive scale for the world with opportunities for cooperation between France and India, in particular innovators and start-ups, given the need for innovative drugs and treatments in India and the developing world at large. In addition, the Indo-French Centre for the Promotion of Advanced Research (CEFIPRA) has been created to underpin and facilitate the collaboration between the scientific communities of the two countries across the knowledge innovation chain.

While Eye-care is a strong point in the Indian Health-care system, software applications relying on modelling and simulation approaches for evaluating the safety and efficacy of new pharmaceuticals, cosmetics, and agrochemicals before testing them in an experimental setup is a strength within the French healthcare environment.

An interesting presentation on the future perspectives of healthcare pointed out that the transmission via internet or dedicated data networks of personal health-related data to medical data hubs and their computerised interpretation would play an ever increasing role in the future as all kinds of mobile applications for personal health data collection, analysis and transmission are becoming

ubiquitous. At the end of the transmission chain, there might be your physician or any medical expert prescribing you a treatment or drug to restore or enforce your health. This adds to the massive data transfer that the internet has to cope with.

The final presentation highlighted the aggressive and invasive capacity of large companies like Google and Apple for appropriating the available "massive data", including in the field of health, for possibly completely dominating the healthcare sector if we are not vigilant and do not develop our own capacities to counteract this power.

## **Conclusions**

The participants highlighted the diversity and quality of presentations and laboratory visits and concluded that there was much scope for future collaboration in various fields. Therefore, we should identify one or two topics for closer cooperation. This could for example be in the form of one or two specific diseases, looking at them from different angles. We should also look into the collaboration with SMEs as they are rather close to market, which is important for applied research. Affordable healthcare is certainly an area with large potential for newly developed applications in need of a pilot project. Ultimately, the two countries should develop a roadmap for future collaboration, eventually resulting in joint Research facilities.

Participants were asked to draw up about half a page on the areas of interest for collaboration.

In view of the wealth and the interest of the exchanges it was decided that an additional seminar would take place in India in April 2015.



## **INAE/NATF Seminar on “Technology and Health-Care” with emphasis on Bioinformatics, 15-16 October, at Évry-Genopole**

The INAE and NATF seminar presents the opportunity to discuss the increasingly technology driven progress in health-care. Ever improving affordability of DNA sequencing and genome mapping has led to a variety of applications. The seminar takes stock of and discusses these applications and the technologies behind them.

### **Programme 15<sup>th</sup> October 2014**

**09:00 – Breakfast and welcome and introduction to the programme by Bruno Revellin-Falcoz (NATF) and Rajeev Shorey (INAE).**

**09:45 – Departure for laboratory visits**

**10:00 - Visit to Genomics Institute and discussion (Jean Weissenbach, Director)**

**11:30 – Visit to IntegraGen and discussion (Mme Bérengère Génin, Bioinformatics Director)**

**13:00 – Lunch**

**14:20 – Presentations on bioinformatics and related topics**

**Introduction by Chairman Pierre-Étienne Bost, Fellow NATF, Vice-Président Généthon**

**14:30 – 15:00 Prof. Sanghamitra Bandyopadhyay, Fellow INAE, Indian Statistical Institute, Kolkata:**

*MicroRNA Induced Regulatory Network in Colorectal and Breast Cancer: A Machine Learning Perspective*

**15:00 – 15:30 Prof. Florence d’Alché-Buc, Télécom ParisTech:**

*From experimental data to biological networks and vice versa*

**15:30 – 16:00 Dr. Sharmila Shekhar Mande, Tata Consultancy Services Ltd. (TCS), Pune:**

*Algorithms, Metagenomics Analysis Platform, and their Applications for Understanding the Role of Gut-Microbiome in Human Health*

**16:00 – 16:30 Dr. Jean-Jacques Codani, CEO of Biofacet SAS:**

*Techniques of Data Analysis from NSG (Next Generation Sequencing)/Big Data and Life Science Data*

**Coffee Break (15 minutes)**

16:45 – 17:15 **Prof. Prasum Kumar Roy**, Fellow INAE, National Brain Research Centre, Gurgaon:

*Towards Health Grid Initiatives in India: Approaching a Grand Challenge in Affordable Health Care – Prospects and Insights from the Brain Grid*

17:15 – 17:45 **Dr. Pierre de la Grange**, CEO of GenoSplice /Dr. Frédéric Lemoine, GenoSplice:

*Flexible and Efficient RNA-Seq Analysis at GenoSplice*

17:45 – 18:15 **Dr. Chandrasekhar Bhaskaran Nair**, BigTec Labs, Bangalore:

*Trulab<sup>TM</sup>: Evolution of a Near Care Molecular Diagnostics Platform for Infectious Disease Diagnostics*

18:15 – 18:45 **Dr. Jean-François Deleuze**, Head of the Centre National de Génotypage (CNG):

*CNG Missions and Technology*

**19:00 – Tour around Génopôle and Migration to Hotel (by bus)**

**20:00 Dinner at Hotel, chaired by Gérard Roucairol, President of NATF**

## Programme 16<sup>th</sup> October 2014

**08:30 - Breakfast**

**08:50 – Presentations on applications in health care and related topics**

**Introduction by Chairman, Rajeev Shorey, (Fellow INAE), Program Director, ITRA, DEITY**

09:00 – 09:30 **Dr. Goutam Ghosh**, Panacea BioTech Ltd, New Delhi

*Opportunities in Research and Manufacturing of Bio-Pharmaceuticals*

09:30 – 10:00 **Dr. William Saurin**, CEO Sobios:

*Software Applications for Modelling and Simulating Biological Systems: Application to Safety and Efficacy*

10:00 – 10:30 **Dr. Sudhir Rachapalle Reddi**, Shankara Nethralaya, Chennai:

*Enhancing Eye Care Services through Innovation, Technology and Collaboration in India*

**Coffee Break (15 minutes)**

10:45 – 11:15 **Dr. Debapriya Dutta**, Director Indo-French Centre for the Promotion of Advanced Research

*Development of Indo-French Health Care Technology Network Programme - Role of CEFIPRA".*

11:15 – 11:45 **Dr. Christophe Ambroise**, Professeur Université d'Evry, Head *Statistis & genome*:

*Statistical Study of Biological Networks*

11:45 – 12:15 **Dr. Rajeev Shorey**, (Fellow INAE), Program Director, ITRA, DEITY, New Delhi:

*Engineering Challenges in Health Care Technologies and Services*

12:15 – 12:45 **Dr. Laurent Alexandre**, President DNAvision :

*DNA, Big Data and Physicians*

**13:00 - Lunch**

**14:30 – Departure for laboratory visits**

**14:15 - Visit to Généthon and discussion (Anne Galy, Head of INSERM Unit 951 for Frédéric Revah, Director)**

**15:45 – Visit to ISSB (Institute of Systems & Synthetic Biology) and discussion (Jean-Loup Faulon, Director)**

**17:30 – General discussion at Conference room**

**19:00 – Return to hotel**

## **INAE/NATF Workshop on ‘Technology and Healthcare’**

### **Participants from Indian National Academy of Engineering (INAE)**

#### **Prof. Sanghamitra BANDYOPADHYAY**

Dr. Sanghamitra Bandyopadhyay did her B Tech, M Tech and Ph. D. in Computer Science from Calcutta University, IIT Kharagpur and ISI respectively. She is currently a Professor at the Indian Statistical Institute, Kolkata, India. She has worked in various Universities and Institutes world-wide including in USA, Australia, Germany, China, Italy and Mexico and has delivered invited lectures in many more. She has authored/co-authored more than 135 journal papers and 140 articles in international conferences and book chapters, and published six authored and edited books from publishers like Springer, World Scientific and Wiley. She has also edited journals special issues in the area of soft computing, data mining, and bioinformatics. Her research interests include computational biology and bioinformatics, soft and evolutionary computation, pattern recognition and data mining. She is a Fellow of the National Academy of Sciences, Allahabad, India (NASI) and Indian National Academy of Engineering (INAE), and senior member of the IEEE. Sanghamitra is the recipient of several prestigious awards including the Dr. Shanker Dayal Sharma Gold Medal and also the Institute Silver Medal from IIT, Kharagpur, India, the Young Scientist Awards of the Indian National Science Academy (INSA), the Indian Science Congress Association (ISCA), the Young Engineer Award of the Indian National Academy of Engineering (INAE), the Swarnajayanti fellowship from the Department of Science and Technology (DST), and the Humboldt Fellowship from Germany. Recently, she has been selected as a Senior Associate of ICTP, Italy. In 2010, she was awarded the prestigious Shanti Swarup Bhatnagar Prize in Engineering Science.

**Title:** *MicroRNA Induced Regulatory Network in Colorectal and Breast Cancer: A Machine Learning Perspective*

#### **Abstract:**

It is well-known that certain proteins, called transcription factors (TFs), regulate the expression of other genes weaving a complex regulatory network in the cell. Discovery of microRNAs, small non-coding RNAs altering gene expression at a post-transcriptional level, has added a new dimension in this context. Our studies indicate that microRNAs play a crucial role in fine-tuning the balance of myriad cellular activities. It is well established that miRNAs regulate the expression of their target genes at a post-transcriptional level, i.e., after the genes have produced the corresponding messenger RNAs. Identifying the target mRNAs of an miRNA is an important task in order to determine its regulatory role on a global scale. At the same time, understanding how the miRNAs themselves are regulated is also important. This is particularly true for intergenic miRNAs that are expected to have independent transcriptional machinery. In this regard, knowing the miRNA transcription start site (TSS) becomes important. Both the tasks, viz., miRNA target prediction and TSS prediction can be formulated as classification problems, enabling the application of machine learning techniques. Here we provide an overall view of the related work going on in our group at the Indian Statistical Institute, Kolkata. We will first briefly describe our approach of miRNA target prediction and TSS prediction. This will be followed by integration of several predicted and validated results for building the network of TF-miRNA-gene. Finally, we focus on this network for two types of cancer, namely colorectal and breast. Graph theoretic analysis of the network yields an interesting three level hierarchical structure. MicroRNAs appear in majority in the topmost level of this hierarchy, indicating that these molecules may be useful for quick signal propagation and also could be potential biomarkers. Their importance in facilitating cross-talk between transcription factors is also highlighted.

## **Dr. Debapriya Dutta**

Dr. Debapriya Dutta completed his Ph.D from the Indian Agricultural Research Institute, New Delhi and joined the National Agricultural Research Service of the Indian Council of Agricultural Research (ICAR). He served as Scientist in the Central Soil and Water Conservation Research and Training Institute, Dehradun,. He joined the Natural Resources Data Management System (NRDMS) Division of the Department of Science & Technology, Government of India in 1994 as Senior Scientific Officer-I and was serving in the same Department as Scientist 'F' / Director, till May, 2008. Thereafter, he had served as the Counselor (Science and Technology) in the Embassy of India, Washington D.C till July, 2012. At this responsibility, he handled the India and USA Science and Technology Cooperation at Policy, Projects and Programme levels in the sectors of health, energy, climate and environment and education. Since, September 2012, Dr. Dutta is serving as the Director of the Indo- French Centre for Promotion of Advanced Research (IFCPAR), a bilateral organization to promote collaborative research between India and France in cutting edge Science and Technology.

Dr. Debapriya Dutta was deputed to the Blackland Soil and Water Research Institute, University of Texas, Texas, USA as UNDP visiting scientist during 1998. He completed his Post Doctoral Fellowship in Geo-information Management from the International Institute for Geo-Information Science and Earth Observation (ITC), the Netherlands in 2003-2004. He was awarded Senior Professional Research Fellowship on "Groundwater Governance" from International Water Management Institute (IWMI), Sri Lanka in 2006-2007.

His areas of expertise and research interests are application of Geo-informatics for Natural Resources Management (Focus on Land and Water), Watershed management technologies, Geo-information Management and Capacity Building for Local Level Planning and Science Diplomacy. He has number of publications in national and international journals and has authored two books.

### **Title: *Development of Indo-French Health Care Technology Network Programme - Role of CEFIPRA***

#### **Abstract:**

Indo-French Centre for the Promotion of Advanced Research (CEFIPRA) is India's first and France's only bilateral S&T organisation which has been in existence for the last 26 years, committed to promote collaboration between the scientific communities of the two countries across the knowledge innovation chain. Initially the main focus of the Centre was to support Scientific Research Programme in the cutting edge basic and applied scientific areas. Since 2002, apart from strengthening support in the cutting edge scientific areas, the Centre also started to support the Academia- industry linkages through Industrial Research Programme of CEFIPRA.

Over the journey of twenty six years, the Centre has supported 457 projects across the knowledge innovation chain out of which 224 are basic science projects, 215 are applied science projects & few have the contribution towards creation of global common goods. Under the domain Life & Health Sciences, Centre has supported 108 projects in different thematic areas including Malaria & Leishmania etc.

CEFIPRA has developed individual linkages between the individual Scientists of both the nations in different thematic areas including Genomics and computational Biology: Diagnostics, Marker, Diagnostics, markers, bioinformatics-based modelling, gene and cell therapy etc. Recognizing the importance of connecting scientists from both the countries, a High Impact Scientific Research Network programme has recently been initiated by CEFIPRA under its expanded mandate. The objective of this programme is to connect excellent research groups from India & France for fostering Interdisciplinary/ Intra disciplinary collaborative research & networking activities between identified groups.

## **Dr. Goutam GHOSH**

Dr. Goutam Ghosh is currently Senior Vice President & Head of Research & Development – Vaccines & Biologicals at Panacea Biotec Limited, New Delhi, a leading Indian Vaccine & Biopharmaceutical company. He also leads the Technology Management & IP functions of the company. Dr. Ghosh has a wide cross functional experience in Government, Public Sector Enterprise and in the Corporate world. He has been in the field Innovation & Technology Management function for over past 25 years and was instrumental in the commercialization of a number of early stage innovations and technologies with companies in India and abroad. Dr. Ghosh has his Post Graduate degree in Bio-Engineering from Indian Institute of Technology, Kharagpur and did his doctoral research in areas of Bio-chemical & Bio-process engineering at Indian Institute of Technology, New Delhi. He has also worked as a World bank fellow at a leading European university.

### **Title: *Opportunities in Research and Manufacturing of Bio-pharmaceuticals in India***

#### **Abstract :**

India has emerged as a global leader in vaccines with about one-third share of the total world market. India's bio-pharmaceutical sector is valued at \$ 26 billion and it is one of the fastest growing knowledge based market which is growing 20 per cent annually for the last few years. India is the major supplier of basic Expanded Programme on Immunisation vaccine to the United Nations Children's Fund (UNICEF). Around 75 to 80 per cent of vaccines procured by UN agencies are from the developing world and almost 80 per cent of these are from India. Similarly, India is now emerging as the global destination for the manufacture of Biologics, especially Biosimilars and cell-based therapeutics, including stem-cell research and regenerative medicine. Biosimilars have tremendous opportunity in India, particularly in monoclonal antibodies based therapeutics where innovator's patents have either expired or will be expiring soon. A cost effective Biosimilar drug must have the quality, safety and efficacy comparable to that of the innovator's product. India is globally regarded as having great potential to become a significant player in the development and commercialisation of Biosimilars due to its proven experience in generic drugs. Further, India's new regulatory policy on Biosimilar products would fast-track its development process.

Panacea Biotec is one of India's leading research-based health management companies with established capabilities in both generic formulation and vaccines. As one of the leading vaccine producers in the country it has significant presence in both institutional and private vaccines markets in India and abroad. It has a strong portfolio of vaccines against critical and life threatening diseases like Polio, Hepatitis B, Diphtheria, Tetanus, Pertussis, Haemophilus Influenza type B (Hib), pandemic flu (H1N1), and combination vaccines. Panacea Biotec has played a key role in global polio eradication program by supplying around 10 billion doses of Oral Polio Vaccines to Government of India and UN Agencies which led to polio free India since 2011. Today, it can support manufacture of up to one billion doses annually. Panacea Biotec has four distinguished research & development centers that specialize in Vaccine & Biologicals, Novel Drug Delivery Systems, Generic Formulations and Discovery Research. The Company also has state-of-the-art manufacturing facilities for Vaccines, Anti-Cancer products and other Pharmaceutical Formulations at various locations in the country. These facilities are approved by several International Regulatory Agencies such as WHO, USFDA, BfArM Germany, ANVISA Brazil, etc. and its product portfolio has expanded internationally with its products reaching out to more than 30 countries.

Therefore, Indian companies like Panacea Biotec can provide huge opportunities for French innovators and start-up companies in forming alliances for a collaborative research, manufacturing and commercialization of Biopharmaceuticals.

## **Dr. Sharmila S. MANDE**

Sharmila Mande heads the Bio-Sciences R&D activities at Innovation Labs of Tata Consultancy Services Ltd. (TCS), a leading software consultancy company. She received her PhD degree in the year 1991 in Physics from Indian Institute of Science (IISc), Bangalore, and later was trained in Protein Crystallography, through which she began to address problems of biological importance. She had her post-doctoral training at University of Groningen, The Netherlands and University of Washington, Seattle, USA. She worked in Institute of Microbial Technology and Post Graduate Institute of Medical Education and Research in Chandigarh, India, before joining Tata Consultancy Services, in 2001.

During her thirteen years at TCS, she has successfully built a group of Bioinformaticians, and trained many of them, including engineers, mathematicians, physicists and biologists in this field. She led the team from TCS in two major projects involving public-private partnership under the New Millennium Indian Technology Leadership program of Council of Scientific and Industrial Research (CSIR), Government of India.

Her main focus areas of research include algorithm development for analyzing large scale biological data and understanding human health through systems biology as well as metagenomics approaches. She has published several papers in international journals and has delivered talks at many International and National conferences. She also has a number of patented algorithms and software solutions that address challenges faced by researchers in analyzing data generated from Next Generation Sequencing technologies.

**Title:** *Algorithms, Metagenomics analysis platform, and their applications for understanding role of gut microbiome in human health*

### **Abstract:**

Majority of microbes present in diverse natural ecosystems cannot be cultured and studied in the laboratory using traditional genomic approaches. The emerging field of metagenomics facilitates the direct extraction and sequencing of the entire genomic content present in a given environment. Efficient computational methodologies are needed for analyzing the enormous volume of metagenomic data. Such metagenomic analysis tools help in profiling the microbial community as well as functional groups present in the given sample. Major challenges in analyzing metagenomics data as well as methodologies to address them will be discussed during my talk.

Malnutrition is a global health problem affecting more than 300 million pre-school children worldwide. Studies have implicated the role of gut microbiota in nutrient pre-processing, assimilation and energy harvest from food. However, our understanding of the role of gut microbiota in malnutrition is currently incomplete. Metagenomic approach was employed for analyzing the gut microbial communities sampled from rural Indian children with varying nutritional status. Results of the analyses using TCS' 'Metagenomic Analysis Platform' will also be highlighted during my presentation.

## **Dr. Chandrasekhar NAIR**

**Dr. Chandrasekhar** received his Bachelors and Masters in Chemical Engineering from BITS Pilani. He has worked in Senior Management positions with the Vittal Mallya Scientific Research Foundation. He has extensive experience in bioprocess modeling, scale up and commercial implementation of bio and chemical processes. Over the past decade, his focus has been on realization of rapid, portable, high quality and low cost diagnostics that would bring the power of a modern laboratory to near-care use. He holds a number of Indian and International patents. Chandrasekhar heads Bigtec labs and is one of the founding directors at Bigtec.

**Title:** *Truelab™: Evolution of a Near care molecular diagnostics platform for infectious disease diagnostics*

### **Abstract:**

India has a huge infectious disease load and public health systems need access to easy to use, highly sensitive and specific diagnostic tools in near care settings to tackle this problem. This talk traces the evolution of the Truelab microPCR platform and its validation across a spectrum of bacterial, viral and parasitic diseases.

## **Dr. Sudhir RACHAPALLE REDDI**

Dr R.R.Sudhir is Head of Department of Preventive Ophthalmology and Senior Consultant in the department of Cornea and Refractive Surgery. He completed his medical graduate degree from Madras Medical College, Chennai, in 1996, Diploma in Ophthalmology at the C.U.Shah Postgraduate Training Centre, Sankara Nethralaya, (1999), Diplomate of the National Board of Examinations (Ophthalmology) in s 2003. He completed one year fellowship in Public Health Ophthalmology at Dana Center for preventive ophthalmology at Wilmer Eye institute along with Master of Public Health (MPH) from the Johns Hopkins Bloomberg school of Public Health, Baltimore, USA in May 2005. He has many papers in peer reviewed journals and presentations at national and international conferences. Dr R R Sudhir is also Consultant-incharge of Electronic Medical Records. He heads the software development and implementation of Electronic Medical Records in Sankara Nethralaya and many major eye institutes all over India. He is reviewer for many Ophthalmic journal and guide for Phd students. Areas of interest: Ophthalmic epidemiology, Clinical trials, Electronic Medical Records, Operational research.

**Title: *Enhancing Eye care services through Innovation, Technology and collaboration in India***

### **Abstract:**

Medical Research Foundation (Sankara Nethralaya) is a premier tertiary eye care institute, with principle objectives of quality eye care, research and training in India. Globally, refractive error and Cataract are two commonest causes of preventable blindness. A very large number of these patients reside in rural areas where access to health care and eye care is sparse. Sankara Nethralaya (SN) first tele-ophthalmology program in India in collaboration with Indian Space Research Organisation which has benefited more than 3,50,000 patients till date. These tele-ophthalmology units are equipped to provide comprehensive eye examination to patients at their door steps and also provide spectacles within one hour through its mobile glass grinding units. Mobile operation theater units capable of performing cataract surgery in villages were developed with IIT Madras (IITM), a premier engineering and technology institute in the country. Several design techniques have been employed and appropriate sub-systems are put in place in Mobile Eye Surgical Unit (MESU) which has performed more than 1000 surgeries so far.

L & T Microbiology Research Centre of Vision Research Foundation under New Millennium Indian Technology Leadership Initiative programme on Ocular Infectious Diseases group for a Novel molecular diagnostics developed “Xcyto-screen DNA chip” is a diagnostic product of collaborative research of four major Ophthalmic centers. The DNA chip aids in the diagnosis of four ocular clinical conditions, Keratoconjunctivitis, Endophthalmitis, Retinitis and Uveitis. A Single chip can detect probable infectious etiology of a given clinical condition in a Multiplex format. Sensitivity is equivalent to individual nested PCRs by an improved detection system. The detection is done by Naked eye and can be performed even in peripheral centers.

Collaborative initiative with IIT Chennai yielded a regional Anesthesia training module with needle blocks called Ophthalmic Anesthesia Simulation System ( OASiS). This mannequin represents accurate Orbital and ocular anatomical representation using state of the art 3D printing Technology. Rugged integrated capacitive and magnetic sensor systems capable of accurately determining needle proximity to ocular muscles and optic nerve with visual alarm for the trainee. Accurate real-time measurement of rate of injection in ml/min with Hall-effect sensors along with aspiration detection and video recording.

Electronic Medical Records with Hospital management system and a detailed MIS system developed in collaboration with Tata Consultancy services, forms the backbone of the information systems at Sankara Nethralaya. This system helps in better workflows, reducing patient waiting time, providing instantaneous MIS reports that helps in better administration and rich data mine for research.

**Prof. P.K.ROY**

As a medical doctor, Prof. P. K. Roy was trained in radiology at Institute of Postgraduate Medical Education and Research, Calcutta University, and was an ICRF scholar at Royal Marsden Hospital, University of London. He has been a postdoctoral Research Scientist and Guest Professor, University of Connecticut, visiting faculty, U.C., Berkeley and Visiting Associate Professor, Medical College of Wisconsin. His areas of research are computational neuroscience, newer information theories, neuroimaging and stochastic resonance, with emphasis on neurothermodynamics, brain imaging and radiodosing.

**Title: *Towards Health Grid Initiatives in India: Approaching a Grand Challenge in Affordable Health Care – Prospects and Insights from the Brain Grid***

**Abstract :**

There looms a chronic disease pandemic both in developing and developed nations, and India has the world's highest disease burden (700 Million DALY-units), costing US\$ 17.4 Trillion during 2012-2030. A major global challenge is drawing meaningful findings from the enormous clinical, genomic and imaging data of the investigations. We develop a novel integration of Multiscale informatics approach to the issue.

An affordable country-wide Brain Grid is initiated, for effective multicentric collaboration and analysis, with possibility for high-precision applications to clinical treatment modalities. The basic infrastructure thus designed is expandable to a Health Grid by incorporating/integrating with other grids (e.g. cancer, liver or other grids), whether national and international. Such endeavours offer considerable hope for the future by ushering in proactive cooperation between informaticians, scientists, engineers and physicians.

## **Dr. Rajeev SHOREY**

Dr. Rajeev Shorey is the Program Director at IT Research Academy, Media Lab Asia, Dept of Electronics & IT, Government of India. Dr. Shorey received his Ph.D and MS (Engg) in Electrical Communication Engineering from the Indian Institute of Science (IISc), Bangalore, India in 1997 and 1991 respectively. He received his B.E degree in Computer Science and Engineering from IISc, Bangalore in 1987 and the B. Sc degree from St. Stephen's College, Delhi University in 1984.

Dr. Shorey has served as the President of NIIT University. Prior to joining NIIT University as its first President, Dr. Shorey was in General Motors India Science Laboratory (ISL), Bangalore and IBM Research Laboratory, New Delhi. He was a faculty in the Computer Science Dept at the National University of Singapore from 2003 to 2004, while on leave from IBM Research Labs in New Delhi.

Dr. Shorey's work has resulted in more than 50 publications in international journals and conferences and several 12 US patents, all in the area of wireless and wired networks. His areas of interest are Wireless Networks, Telecommunications, Telematics, Data Security, Data Analytics and Technologies for the Healthcare sector. Dr. Shorey has served on the Editorial boards of IEEE Transactions on Mobile Computing and is currently serving on the Editorial board of WINET (Wireless Networks Journal of Mobile Communication, Computation and Information) journal. He is the editor of the book titled "Mobile, Wireless and Sensor Networks: Technology, Applications and Future Directions" published by John Wiley, US in March 2006.

For his contributions in the area of Communication Networks, Dr. Shorey was elected Fellow of the Indian National Academy of Engineering in 2007. He is a Fellow of the Institution of Electronics and Telecommunication Engineers, India.

### ***Title: Engineering Challenges in Healthcare Technologies and Services***

#### **Abstract:**

The Indian healthcare industry, which comprises hospitals, medical infrastructure, medical devices, clinical trials, outsourcing, telemedicine, health insurance and medical equipment, is expected to reach US\$ 160 billion by 2017.

Technology is transforming the health care industry in ways not seen before. From pervasive, smart devices to leveraging big data and tapping into social networking, the industry is now turning its attention to how technology can be applied to keep people healthy.

Recent advances in mobile computing and networking have resulted in low cost pervasive sensing, mobile platforms with more on-device computation than ever before and ubiquitous communications networks allowing for large scale connectivity and leveraging cloud computation. While much of the sensing and mobile devices have become part of everyday life, their relevance and applicability towards developing interdisciplinary solutions for critical application domains such as Healthcare, is yet untapped, due to several limitations, both technical and contextual.

In this talk, we present technical challenges in enabling robust end-to-end healthcare systems. These systems range from tiny Body Area Networks (BANs) and wearable computing devices to wireless LANs and cellular mobile communication systems.

We argue that engineering healthcare systems requires a deep understanding of protocols, algorithms and architectures. Pervasive healthcare systems are characterized by lightweight protocols and architectures, tight networking constraints, such as, low packet loss probability and low end-to-end latencies and security related challenges. Further considering the nature of healthcare applications, engineers need to ensure that there are no false positives and false negatives. The problem is further compounded by lack of global standards.

We are seeing rapid growth in the space, in part due to the importance and criticality of healthcare systems. New paradigms in healthcare are likely to be in place in the coming decade. We are likely to see a proliferation of diverse categories of wearable devices, body area networks, sensors – both intrusive and non-intrusive, and more sophisticated medical equipment and systems. Today's large medical equipment is likely to become miniaturized, thanks to great strides in IT, VLSI and Embedded Systems. The coming decades will see a plethora of new Healthcare Services and Applications. A clear picture of what the future in Healthcare will look like in ten years from now is anyone's guess.

## **Participants from National Academy of Technologies of France (NATF)**

### **Dr. Laurent ALEXANDRE**

**Dr. Laurent Alexandre** is a Surgeon and Neurobiologist, graduated from “Sciences po”, HEC and ENA. He is the founder and developer of Doctissimo.fr and more than ten hi-tech companies. Today, he manages DNAVision that is specialised in genome interpretation. He has written several books including “La mort de la mort”, “La défaite du cancer” and “Google démocratie”. Now, he is interested in NBIC revolution.

**Title:** *DNA, Big Data and Physicians*

**Abstract:** Not available

## **Dr. Christophe AMBROISE, Statistique et Génome**

**Dr. Christophe Ambroise** (Professeur des Universités, Université d'Evry) has obtained his PhD in applied statistics in 1996 (Université de Technologie de Compiègne, Laboratoire Heudiasyc). After post-doctoral positions in Ecole des Mines and Paris 6, he was recruited at the University of Compiègne in 1998 as an assistant professor in the field of Pattern Recognition.

He has obtained his tenure in 2005 and was appointed professor in 2006. He is head of the team *Statistique et génome*, UMR CNRS 8071 since 2010 and does his research in statistics applied to biology.

**Title:** *Statistical Study of biological Networks*

### **Abstract:**

Networks are a straightforward formalism for representing interactions between objects of interest and are thus used in many scientific fields. For instance, in Biology, regulatory networks allow to describe the regulation of gene expression through transcriptional factors, while metabolic networks focus on representing pathways of biochemical reactions. Besides, the binding procedures of proteins are often described as protein-protein interaction networks.

In this presentation we discuss two related aspects of network study: the statistical inference of networks using biological high-throughput data and the use of these inferred networks for gaining biological insight via clustering.

We will discuss inference through sparse Gaussian Graphical Models (GGM), which give a sounded representation of direct relationships between biological objects (gene, protein, reaction...) and are accompanied with sparse inference strategies well suited to the high dimensional setting.

They are also versatile enough to include prior structural knowledge to drive the inference. Still, GGM are now in need for a second breath after showing some limitations: among other questionings, the state-of-the-art reconstruction strategies often suffer a lack of robustness.

One possibility to further explore networks consists in studying their group structure. This means that nodes can be spread into latent classes having similar connectivity patterns which can provide insight. We present and discuss mixture model based statistical tools dedicated to uncover latent network structures.

## **Pierre-Etienne BOST**

**Pierre-Etienne BOST** is Chemist, PhD, graduate engineer from the Massachusetts Institute of Technology, he has spent most of his career in research at RHONE-POULENC, now SANOFI-AVENTIS, where he has served as Director of research and development in the pharmaceutical field from 1983 to 1995. Responsible for the Organization of research of RHONE-POULENC health activities, the design and conduct of research and development programs, he participated very directly in the discovery and the launching of several new pharmaceutical products, drugs that today represent major therapeutic advances in public health:

-in Oncology, TAXOTERE ®, developed through collaboration with the CNRS (Institute of chemistry of natural Substances, team of Prof. P.POTIER), is an essential drug for treatment of breast and lung cancer.

-in the cardiovascular field, the development of LOVENOX ® is an important step in the prevention of venous thrombosis

-in infectious diseases, SYNERCIDE ® has proven effective against germs resistant to other antibiotics, and in the area of the central nervous system, RILUTEK ® was one of the first treatments of amyotrophic lateral sclerosis.

Many of these drugs are also industrial and commercial successes since two of them achieved turnovers of more than one billion euros. Strategic products for the company SANOFI-AVENTIS, they are also great achievements of French pharmaceutical industry research.

Pierre-Etienne BOST participated as expert in several commissions of the Ministry of National Education, higher education and research, and the Scientific Council of the Agency of medicines. Former Director of Aventis Pasteur, he was until 2006 a unit Director of research at the Institut Pasteur (URA CNRS 2128). European expert, author of numerous patents and publications, he is also one of the founding members of the Academy of Technologies.

At present he is:

- Délégué Général de l'Académie des technologies
- Membre du Conseil Supérieur de la Recherche et de la Technologie
- Vice-Président Genethon
- Président du Conseil scientifique de DNDi
- Administrateur de Texcell

**Dr. Jean-Jacques CODANI, Biofacet SAS.**

**Dr. Jean-Jacques Codani** is the CEO of Biofacet, a Computational Biology software Company.

Co-founder of Gene-IT (**aka** GenomeQuest), Dr. Codani is a Scientific Executive in Genomics/HPC, with over 10 years in Academic Research and 16 in High-Tech Software Industry. As Gene-IT's CEO, he acquired strong experience in management, fund raising and business processes. As GenomeQuest CSO, he managed R&D teams overseas, and built leading edge Software platforms for IP and NGS HPC.

Dr. Codani holds a PhD in Computer Science from INRIA. He received the IBM intensive computing award for the physical mapping of the Human Genome, and is featuring in many journals in Computer Science and Genomics, including Nature and Cell.

**Title: *Techniques of data analysis from NGS (Next Generation Sequencing) / Big data and Life-Science data***

**Abstract:**

The avalanche of data coming from Next Generation Sequencing (NGS) is a fact.

This talk will detail a couple of metrics of this arena, the dynamic of data size being quite large according to the targeted problematic. Given a single “simple” problem such as “reads mapping”, we'll show how the software techniques can be very different, for example between large-scale population SNP calling, and miRNA identification.

While the term “big data” obviously applies to NGS data management, we'll try to highlight the specificity of the field, where accuracy matters, and where scaling up with NGS results produced at a population-wide level breaks with traditional database management approach.

**Dr. Florence D'ALCHÉ-BUC, LTCI UMR CNRS 5141, IBISC, Université d'Evry and Télécom ParisTech**

**Dr. Florence d'Alché-Buc** is full professor at Telecom ParisTech and is doing research at LTCI (CNRS & Télécom Paristech) and IBISC (Université of Evry). Graduated from Telecom Paris Tech (formerly ENST), she defended her doctoral thesis at the University Paris Sud in 1993 on the learning of rules for decisions by constructive neural models. After a stint at Philips electronics laboratories, she became Lecturer at the Université Pierre et Marie Curie in 1995. She joined the University of Evry in 2004 to create and animate the team “Learning, Modelling and Data Integration: application to systems biology” and has now just joined Télécom ParisTech. Rooted in statistical learning, her research focuses essentially on structured data prediction and modelling of dynamic systems with tools on the basis of kernels and a predilection for applications in systems biology.

**Title:** *From experimental data to biological networks and vice versa*

**Abstract:**

The identification of the complex interaction and control mechanisms at work in the cell is a central goal of systems biology. At mid-term, this will open the door to a better understanding of numerous diseases and hopefully in new therapeutic targeting. Since large scale experimental techniques of measurement provide “omics” data at the scale of a whole genome, this problem of network reverse-engineering, commonly called network inference, fits well within the statistical framework of machine learning. Machine learning aims at providing a theoretical and practical framework to cope with realistic settings such as partially labeled data, replicate data, heterogeneous and structured data and finally can also provide a way to actively asks for data when they are not sufficient to provide accurate predictions.

In this talk, we present a short overview of our recent results concerning network inference. The first result concerns a new class of versatile models for network inference based on operator-valued kernel-based regression. We show how protein-protein interactions (PPI) can be inferred in the case of the PPI network around CFTR, a protein whose mutations are involved in cystic fibrosis in collaboration with Alexander Edelman. The second result is a novel general framework for experimental design based on active learning that we applied on a gene regulatory network reverse-modeling task. The classic interaction loop between biologists and modelers involve usually 5 steps : (i) some starting biological hypothesis to study (ii) a choice of relevant experiments to test this hypothesis, (iii) the development of (new) network inference methods, (iv) the application of the proposed methods to experimental data and the obtained predictions, (v) wet-lab validation of these predictions. In this work we advocate for coupling more tightly the design of experiments and the learning phase and present an original active learning algorithm that is able to ask additional experiments to get better performance. We show a first evaluation of this approach on one of the DREAM7 challenges.

## **Dr. Pierre DE LA GRANGE and Frédéric LEMOINE, GenoSplice**

**Dr. Pierre de la Grange** is the CEO of GenoSplice. Pierre started his professional career at INSERM. During his Ph.D., he played a key role in the development of bioinformatics solutions such as FAST DB, the leading alternative splicing database. From there, he was recruited as a postdoctoral fellow by EURASNET, the European Network of Excellence for Alternative Splicing. In 2008, he created GenoSplice.

**Frédéric Lemoine** is mainly responsible for the development of bioinformatic projects related to high-throughput sequencing at GenoSplice. Frederic did his PhD in computer science/bioinformatics at University Paris-XI, where he worked on integration and analysis of comparative genomics data. After a postdoc in Lausanne, Switzerland where he worked on small-RNA sequencing data, he joined GenoSplice in Feb. 2010.

### **Title: *Flexible and Efficient RNA-Seq analysis at GenoSplice***

#### **Abstract:**

GenoSplice ([www.genosplice.com](http://www.genosplice.com)) is a privately held company that provides high-level bioinformatics solutions as tailored services for analysis of genomics data (RNA, DNA and epigenetics). The company uses proprietary tools to provide its innovative services to many clients all over the world (Academics, Biotech and Big Pharma).

RNA-Seq is enabling scientists to study transcriptome in a way that was not considered before. Not only it makes possible to study gene expression and splicing without reference, but data generated by that way also allow us to study fusion transcripts, transcribed SNPs, or allele specific expression. The downside is that the analysis of this data is a challenging task in three ways:

- It generates a huge amount of data;
- Current analysis tools are computationally intensive;
- Results are very difficult to interpret.

GenoSplice developed a tool called EASANA-Seq<sup>®</sup>, which is made of two modules:

- 1) The data analysis module that perform the alignment, read counting on genes, splicing patterns, and fusion transcripts. This module is based on a flexible distributed computing platform that we can adapt to the needs of the projects.
- 2) The visualization module: EASANA<sup>®</sup>. GenoSplice designed EASANA<sup>®</sup> to facilitate access to already computed results by bringing to the user the cleanest results, in their biological context.

Each of these two modules is built on the top of FAST DB<sup>®</sup>, the GenoSplice annotation database, which integrates information about gene structure, detailed gene splicing, and public annotations. Taken together, these two modules try to resolve the 3 bottlenecks of RNA-Seq data analysis.

To illustrate the GenoSplice EASANA-Seq<sup>®</sup> solution, we will present results from analysis of uveal melanoma tumors mutated or not for *SF3BI* (a genes that participates in the splicing of pre-mRNAs). These results were published in Cancer Discovery in 2013 (Furney et al.).

## **Dr. Jean-Francois Deleuze, Ph.D, Centre National de Génotypage**

**Dr. Jean-François Deleuze** is Head of the Centre National de Génotypage (CNG) at CEA and Scientific director of Fondation Jean Dausset (CEPH). Previously he was head of the SANOFI Regenerative Medicine Platform building and running an open innovation focused “biotech like” group to implement RegMed in SANOFI (2010 to 2012), Head & Scientific Director of the SANOFI worldwide Genetics Centre and member of senior Research Management Committees (2000 to 2010) and Senior Scientist at RPR/Aventis to establish a human genetics unit (1996 to 2000).

He is a Graduate from several executive and corporate management programs (INSEAD, Wharton University) and Post-Doctorand at INSERM. He has gained his PhD in Molecular & Cellular Biology at Paris VI University, France (1988-93) and has been a Freelance Journalist for two medical journals.

His professional achievements include:

- Successful management of CNG & CEPH
- Creation and Management of the SANOFI Regenerative Medicine Platform
- Successful management of a Genomic Centre
- Identification of several disease causing genes pursued as pharmaceutical targets and identification of genetic biomarkers for clinical development
- 55 publications in peer review journals & 15 patents.

### **Title: *CNG Missions and Technology***

**Abstract:** The Centre National de Génotypage (CNG) is a strategic component of the French Genomics Institute of CEA in Evry Genopole. CNG has one of the largest European capacities in high throughput genotyping and sequencing dedicated to human health and diseases. Missions of CNG are 1) to participate to the discovery and the understanding of the « genomic » causes of human diseases but also relevant biomarkers so as to allow the development of innovative personalized therapeutic approaches in unmet medical needs by providing best in class technology platforms (wet and *in silico*) for collaborative and CNG own projects and 2) participate to the Genomic Medicine revolution of the 21st century that will make accessible and affordable the genomic information to support personalized health management. Presentation will focus on the strategic resources of CNG and a few projects will be discussed to exemplify how these technologies are used.

## **Bruno Revellin-Falcoz**

**Bruno Revellin-Falcoz** was born in 1941 in Bourgoin. He graduated in 1964 from SUPAERO.

Joined the Dassault Aviation Design Department in 1966. Focused his activities on new techniques and advanced military and civil programs, both on national and international level. Appointed Senior Vice President Research, Design and Engineering in 05/1982, he is in charge of all the technical programs (The family of business jets Falcon, Mirage 2000 and Rafale fighters).

Vice Chairman in 2000: he leads the development of the new versions of Rafale and the brand new Falcon 7X.

Leaves Dassault Aviation in 2008, after 45 years in the company. Member of the Air and Space Academy of France. Grand Prix 2008 of « Ingenieurs de France ».

Member of the scientific advisory board of the Ministry of Defense.

Member of the scientific advisory board of OPECST (Office Parlementaire d'Etude des Choix Scientifiques et Technologiques). President (2011/2012) of NATF (National Academy of Technology of France). Honorary President and Delegate for Foreign Affairs.

Member of the Executive Committee of the governing board of EIT (European Institute of Innovation and Technologies).

**Dr. William SAURIN, Sobios**

**Dr. William Saurin** is Executive Director of Sobios.

He has a long term experience in informatics and computer sciences applied to biology, in academia and industry.

In 2001, William SAURIN co-founded and managed Genomining SA, as President and CEO. Genomining was a software editor providing solutions for integrating biological databases.

William SAURIN was a member of the founding team of Genoscope (1997), one of the major contributing institution of the Human Genome Consortium. At Genoscope, he managed the informatics and bioinformatics department.

William SAURIN is a former Director of Research at CNRS. He developed his research activities at the Pasteur Institute in the field of bioinformatics. He is the author of numerous scientific articles and patents.

***Title: Software Applications for Modelling and Simulating Biological Systems: Application to Safety and Efficacy.***

**Abstract:**

Dassault Systems and Sobios develops specialized products (from molecules to biological pathways, from cells, organs, organism to clinical trials) for life sciences industries and research institutes, and in particular for pharmaceuticals, cosmetics, and agrochemicals. These applications are grounded on an integrated software environment for the discovery and development of new biological entities.

These software applications rely on modelling and simulation approaches. They allow researchers to formulate and digitally evaluate hypothesis before testing them in an experimental setup. They aim at reducing the experimental effort by a better prioritization.

The integrated software environment keeps track of models, simulation and choices performed during an R&D project, thus enhancing the collaboration between researchers from different specialties.

The software applications will be presented, and will be demonstrated on particular use cases in safety and efficacy.

## **Participants**

### ***Indian Delegation***

**Prof. Sanghamitra BANDYOPADHYAY, ISI, Kolkata, Fellow INAE**

**Dr. Debapriya Dutta, Indo-French Centre for the Promotion of Advanced Research, New Delhi**

**Dr. Goutam GHOSH, Panacea BioTech Ltd, New Delhi**

**Dr. Sharmila S. MANDE, Tata Consultancy Services Ltd (TCS), Mumbai**

**Dr. Chandrasekhar NAIR, BigTec Labs, Bangalore**

**Dr. Sudhir RACHAPALLE REDDI, Shankara Nethralaya, Chennai**

**Prof. P.K.ROY, National Brain Research Centre, Gurgaon, Fellow INAE**

**Dr. Rajeev SHOREY, Programme Director, ITRA, DEITY, New Delhi, Fellow INAE**

### ***French Delegation***

**Dr. Laurent ALEXANDRE**

**Dr. Christophe AMBROISE, Statistique et Génome**

**Dr. Jean-Jacques CODANI, Biofacet SAS.**

**Dr. Florence D'ALCHÉ-BUC, LTCI UMR CNRS 5141, IBISC, Université d'Evry and Télécom ParisTech**

**Dr. Pierre DE LA GRANGE and Frédéric LEMOINE, GenoSplice**

**Dr. Jean-Francois Deleuze, Ph.D, Centre National de Génotypage**

**Dr. William SAURIN, Sobios**

### ***Delegation NATF***

**Pierre-Etienne Bost, Vice President Généthon, Fellow NATF**

**Bruno Jarry, Fellow NATF**

**Gérard Roucairol, President NATF**

**Bruno Revellin-Falcoz, Honorary Presedent NATF**

**Wolf Gehrish, Assistant to Bruno-Revellin Falcoz**

# Exposés



# MicroRNA Induced Regulatory Network in Colorectal and Breast Cancer: A Machine Learning Perspective

Sanghamitra Bandyopadhyay

Professor, Machine Intelligence Unit

Indian Statistical Institute, Kolkata, India

Email: [sanghami@isical.ac.in](mailto:sanghami@isical.ac.in)

URL: <http://www.isical.ac.in/~sanghami>



# MicroRNA Induced Regulatory Network in Colorectal and Breast Cancer: A Machine Learning Perspective

Sanghamitra Bandyopadhyay

Professor, Machine Intelligence Unit  
Indian Statistical Institute, Kolkata, India

Email: [sanghami@isical.ac.in](mailto:sanghami@isical.ac.in)

URL: <http://www.isical.ac.in/~sanghami>

## Outline

- Introduction
- Components of the miRNA induced regulatory network
  - Regulation of miRNAs by TFs
  - Prediction of miRNA targets
- Construction of the miRNA induced regulatory network
- Network analysis
- Disease Specific Network and Analysis
- Summary

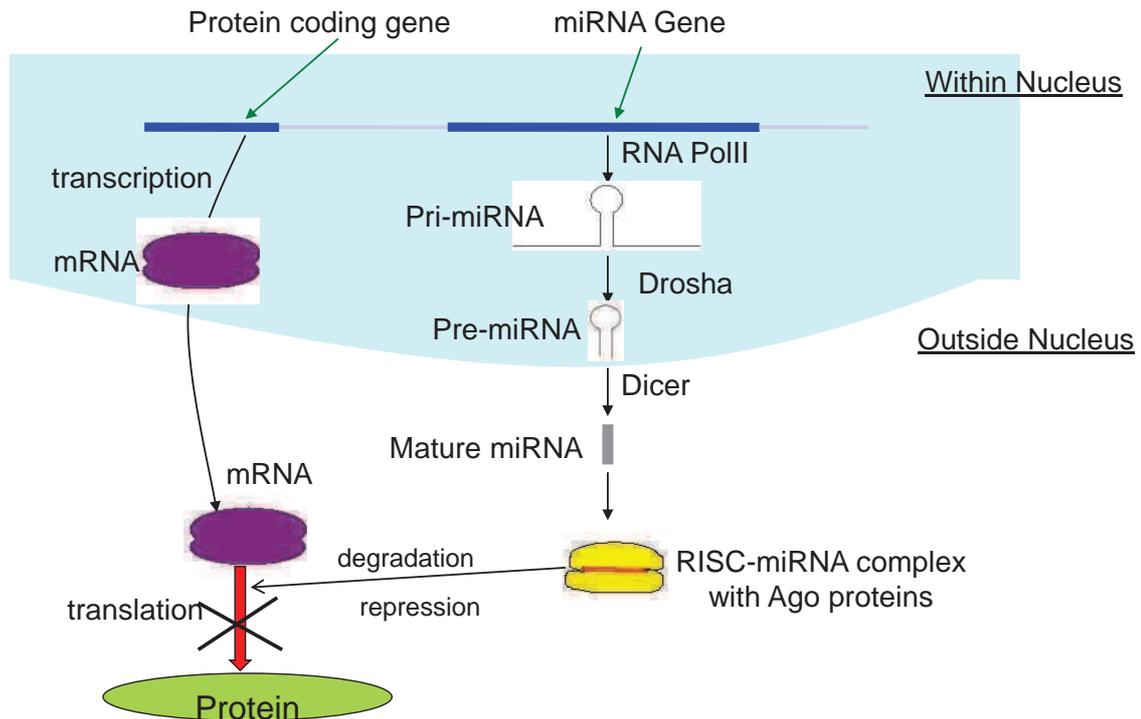
## Outline

- Introduction
- Components of the miRNA induced regulatory network
  - Regulation of miRNAs by TFs
  - Prediction of miRNA targets
- Construction of the miRNA induced regulatory network
- Network analysis
- Disease Specific Network and Analysis
- Summary

## MicroRNAs

- Small endogenous RNA molecules (Bartel 2004)
  - about 21-23 nucleotides long
- Regulates the expression of other genes post transcriptionally
  - by mRNA degradation
  - translational repression
- Take significant roles in many biological processes (Jiang *et al.*, 2008; Bartel, 2009)
- Add another level of gene regulation
- Implicated in several diseases, especially in different cancers
  - leukemia, lung, colon, breast, cervical, pancreas, thyroid, prostate

# Biogenesis of miRNA



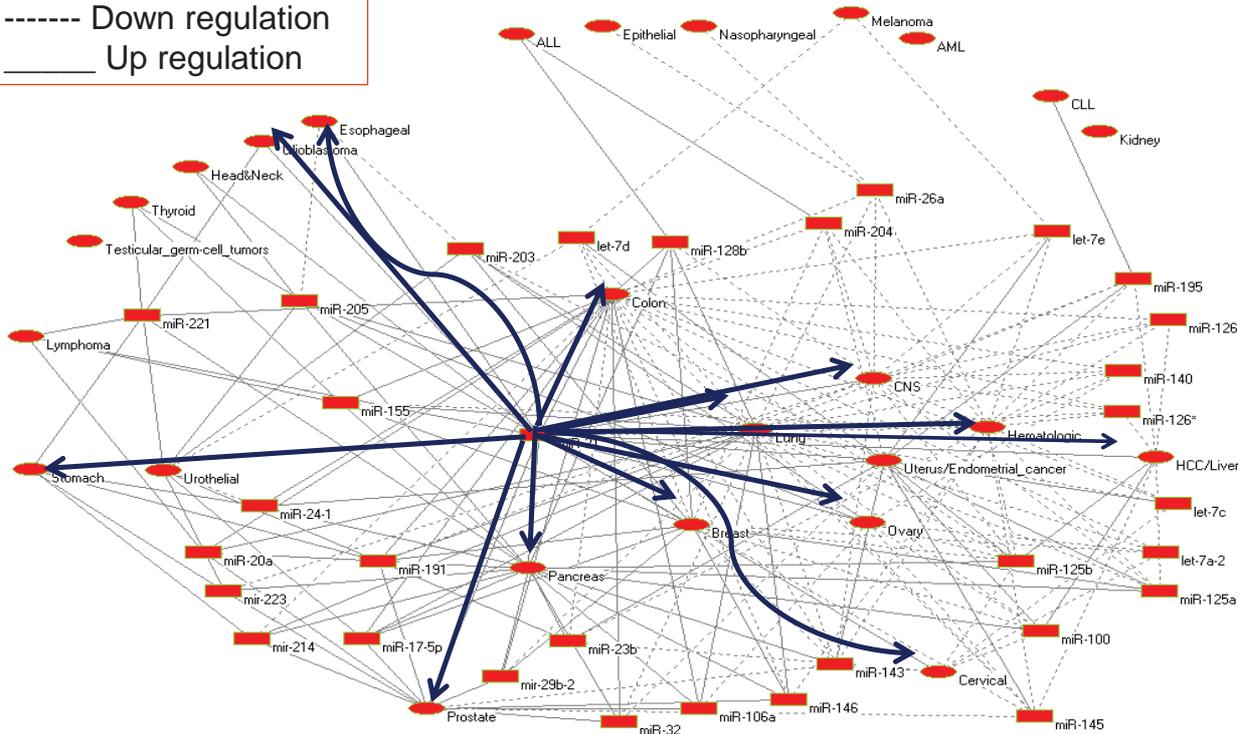
## Some miRNA Statistics

- 1,600 precursor and 2,042 mature human miRNAs in miRBase Release 19 (August 2012): <http://www.mirbase.org/cgi-bin/browse.pl?org=hsa>
- More than 600 miRNAs are known to be significantly dysregulated in various cancer malignancies
  - Expression of miRNAs are altered in most human malignancies
- More than 10,000 biologically validated miRNA-cancer relationships have been identified.
- Many miRNAs have strong oncogenic (ex: miR-21, miR-145, miR-155) or tumor suppressor (ex: let-7a-2, miR-16, miR-143) characteristics

# Cancer-miRNA network

- >200 miRNAs involved in cancer
- 1000 cancer-miRNA relations known

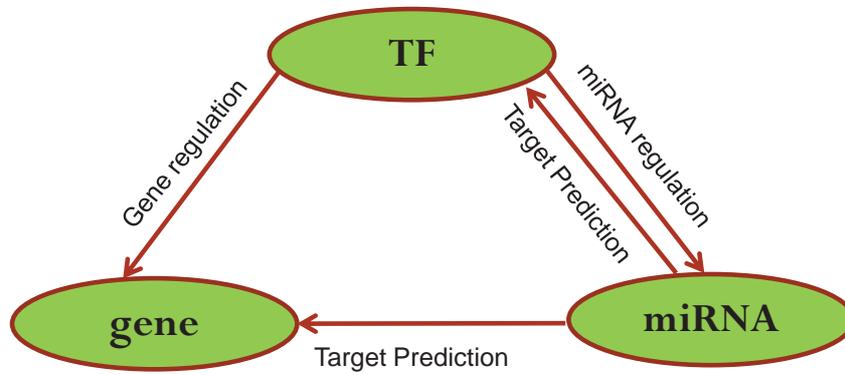
----- Down regulation  
\_\_\_\_\_ Up regulation



## Outline

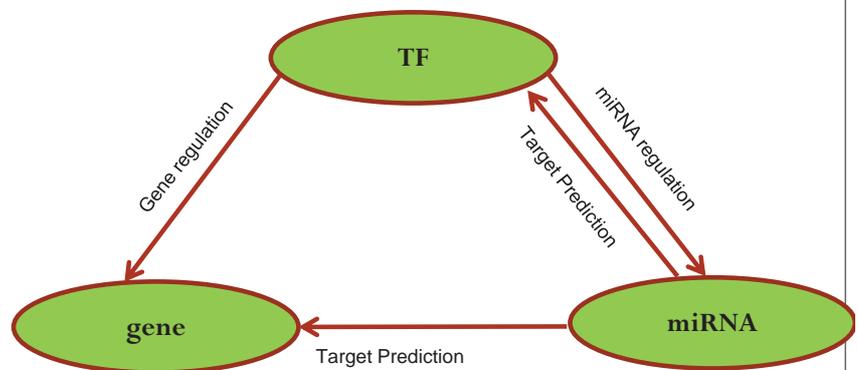
- Introduction
- Components of the miRNA induced regulatory network
  - Regulation of miRNAs by TFs
  - Prediction of miRNA targets
- Construction of the miRNA induced regulatory network
- Network analysis
- Disease Specific Network and Analysis
- Summary

# TF-miRNA-gene Interaction Network



## The Components of the Network

- TF → miRNA
  - Ongoing work on miRNA TSS Prediction
  - Recently proposed PuTMiR
- Gene (TF) → Gene (TF)
  - collected from the 2006 assembly of UCSC (original source TRANSFAC)
- miRNA → Gene (TF)
  - Recently proposed TargetMiner



## Outline

- Introduction
- Components of the miRNA induced regulatory network
  - Regulation of miRNAs by TFs
  - Prediction of miRNA targets
- Construction of the miRNA induced regulatory network
- Network analysis
- Disease Specific Network and Analysis
- Summary

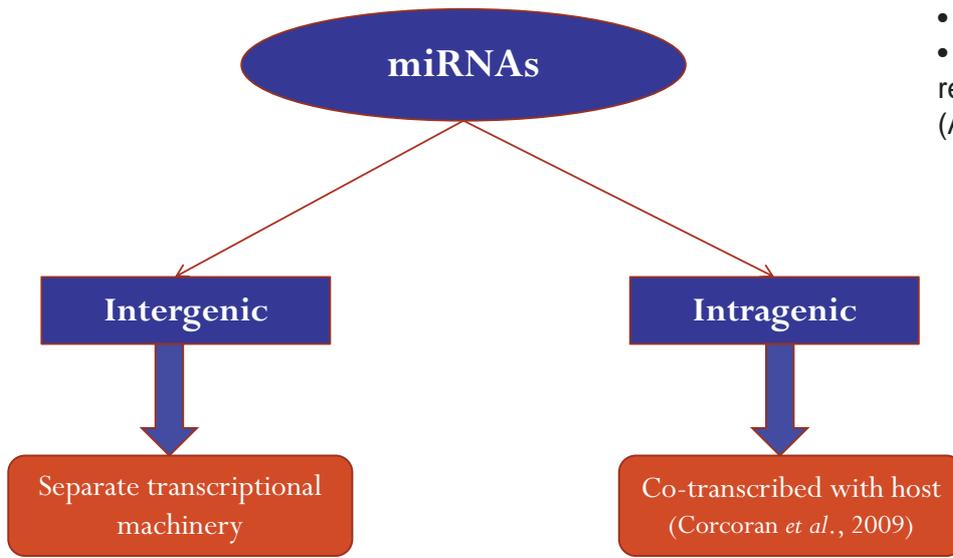
## Regulation of miRNAs by TFs

- TF → miRNA
- TransmiR: biologically validated TFs regulating miRNAs [Wang et al. (2009), *Nucleic Acids Res*]
- PuTmiR: Putative TFs that could potentially regulate miRNAs [Bandyopadhyay and Bhattacharyya (2010), *BMC Bioinformatics*]
  - TFs that have binding sites 10kb upstream of miRNAs.
- Prediction of transcription site of miRNAs [Bhattacharyya et al. (2012) *SAGMB*]

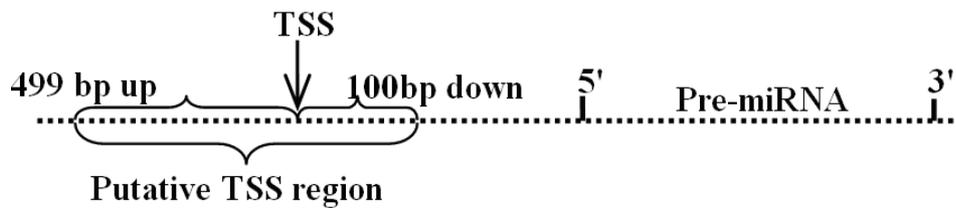
# Types of miRNAs

Observed ratio of intra and inter-miRs

- 3:2
- amongst 2042 miRNAs reported in miRBase (August 2012)



# Transcription of miRNAs



## TSS Prediction in Intergenic miRNAs

- Biologically validated data of TSSs in upstream region of miRNAs extracted.
  - 119 Positive samples (Marson et al., 2008)
- Non-TSS regions extracted that has no miRNAs with 50kb region.
  - 381 Negative samples
- Features extracted from the positive and negative samples.
  - Genetic features, CpG island based features, methylation based features
- SVM classifier with RBF kernelis trained on this data
- Trained classifier is used to predict TSSs of miRNAs.

## Experimental Results

Feature Set	# Features	Classifier Performance based on the Total Features			# Features Selected with AMOSA	Classifier Performance based on the Selected Features		
		Criteria	$\mu$	$\sigma$		Criteria	$\mu$	$\sigma$
SOUPLUSUCIUMT	478	Accuracy	0.752	0.009	225	Accuracy	0.765	0.020
		Sensitivity	0.714	0.017		Sensitivity	0.752	0.021
		Specificity	<b>0.787</b>	0.017		Specificity	0.774	0.015
SOUPLUSUCI2UMT	410	Accuracy	<b>0.765</b>	0.023	216	Accuracy	<b>0.795</b>	0.022
		Sensitivity	<b>0.779</b>	0.018		Sensitivity	<b>0.807</b>	0.037
		Specificity	<b>0.752</b>	0.034		Specificity	<b>0.784</b>	0.022
SOUPLUSUCIUMT	520	Accuracy	0.762	0.016	267	Accuracy	0.773	0.009
		Sensitivity	0.752	0.015		Sensitivity	0.741	0.018
		Specificity	0.768	0.025		Specificity	<b>0.803</b>	0.010

**Table 4.** Performance of the classifier based on various features after inclusion of methylation based features

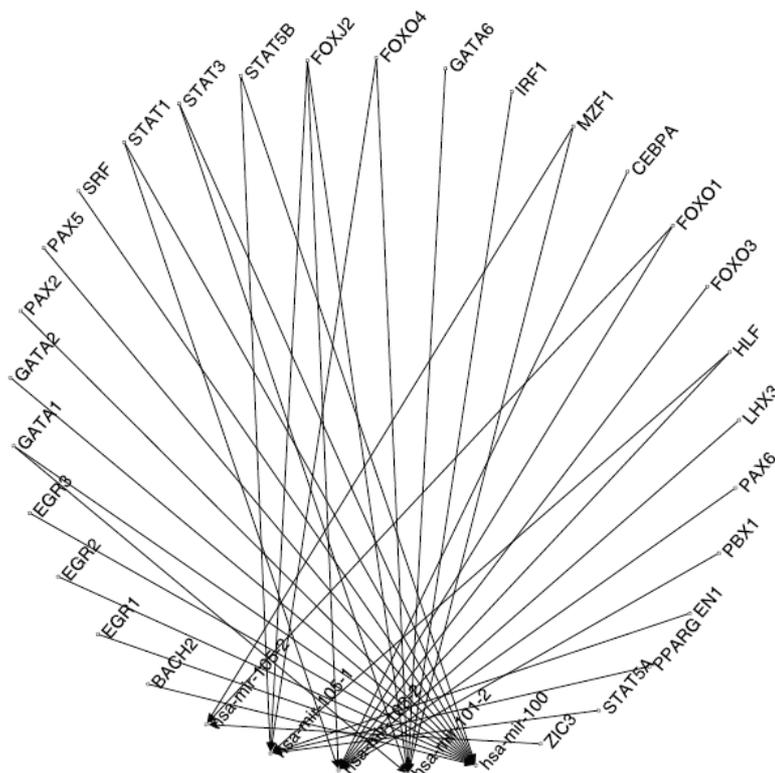
Bhattacharyya, Feuerbach, Bhadra, Lengauer, Bandyopadhyay, (2012) SAGMB

## Predicted TSS

Intergenic miRNA	Chromosomal location	Predicted site of TSS	Distance from 5' end (dis. from valid TSS)
hsa-mir-199a-1	chr19:(10789101,10789172)(-)	10789516	344 (8)
hsa-mir-181c	chr19:(13846512,13846622)(+)	13844975	1537 (21)
hsa-mir-124-2	chr8:(65454259,65454368)(+)	65445191	9068 (32)
hsa-mir-24-2	chr19:(13808100,13808173)(-)	13819222	11049 (36)
hsa-mir-596	chr8:(1752803,1752880)(+)	1752752	51 (40)
hsa-mir-10b	chr2:(176723276,176723386)(+)	176709982	13294 (52)

Bhattacharyya, Feuerbach, Bhadra, Lengauer, Bandyopadhyay, (2012) SAGMB

## Partial View of the TF→miRNA Network



## Outline

- Introduction
- Components of the miRNA induced regulatory network
  - Regulation of miRNAs by TFs
  - Prediction of miRNA targets
- Construction of the miRNA induced regulatory network
- Network analysis
- Disease Specific Network and Analysis
- Summary

## MicroRNA Target Prediction Algorithms

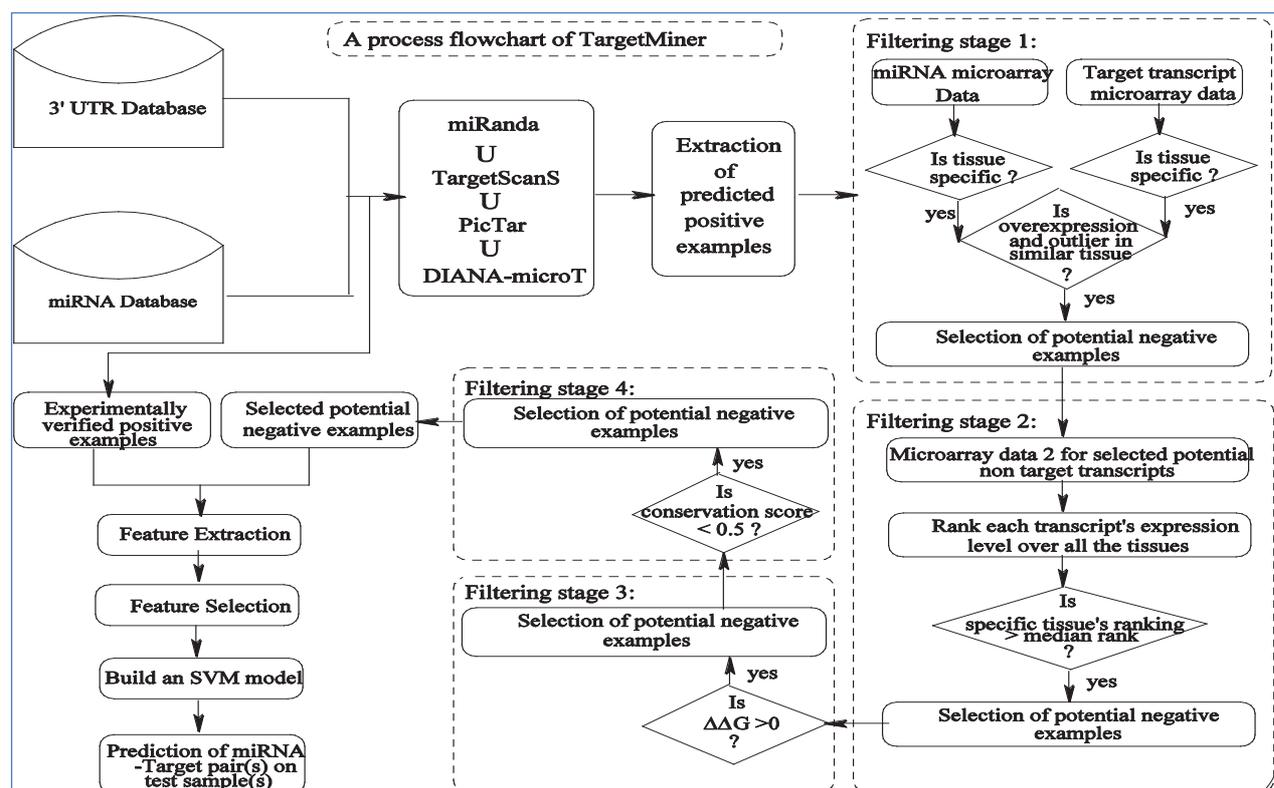
- Traditional algorithms (mainly sequence based):
  - miRanda, 2003-2008
  - PicTar, 2006
  - TargetScan, 2004-2009
- Structural interaction based algorithms:
  - PITA, 2007
  - STARmiR, 2007
- Machine learning based algorithms:
  - NBmiRTar, 2007
  - MirTarget2, 2008
  - TargetMiner, 2009

# Systematic Identification of Negative Examples

- Existing methods vary a lot in their results
  - Targets predicted by different methods have few common examples
  - Reliance on artificially generated or randomly selected negative examples
- Proposed method: TargetMiner
  - Systematic identification of tissue specific negative examples and target prediction
  - [http://www.isical.ac.in/~bioinfo\\_miu](http://www.isical.ac.in/~bioinfo_miu)
  - Genome wide predictions linked in mirbase ([www.mirbase.org](http://www.mirbase.org))

8/26/2014

## TargetMiner



# Negative Data Selection in TargetMiner

- Negative set generation
  - For an miRNA
    - Use union set of miRanda, TargetScan and PicTar to predict targets
    - Negative data extracted from this set of predicted targets
      - By filtering out those that have some potential to be true targets
    - Such negative data have target like properties
      - Training data from the overlapping region
  - Negative data extraction by filtering
    - Selection Stage 1 and Stage 2: based on microarray
      - Select those miRNA-mRNA pairs that are tissue specific and are highly over expressed in the same tissue
        - Probably are not degraded targets
    - Filter Stage 3: based on minimum free energy
      - Remove those which are stable ( $<0$  Kcal/mol)
    - Filter Stage 4: based on conservation score across 17 species
      - Remove those which are conserved ([phastCons](#) $>0.5$ )

# TargetMiner Details

contd...

- Feature extraction
  - 90 features extracted
    - K-mer frequency in seed site region, AU-rich region, Watson-Crick outseed region etc.
  - Top 30 features selected
    - Based on F-score
  - SVM used for classification

## Correlation Comparisons

- Correlation between
  - Positive and negative examples (real): 0.898
  - Positive and negative examples (TargetMiner): 0.876
  - Positive and artificially generated negative examples: 0.782

8/26/2014

## Performance Comparison

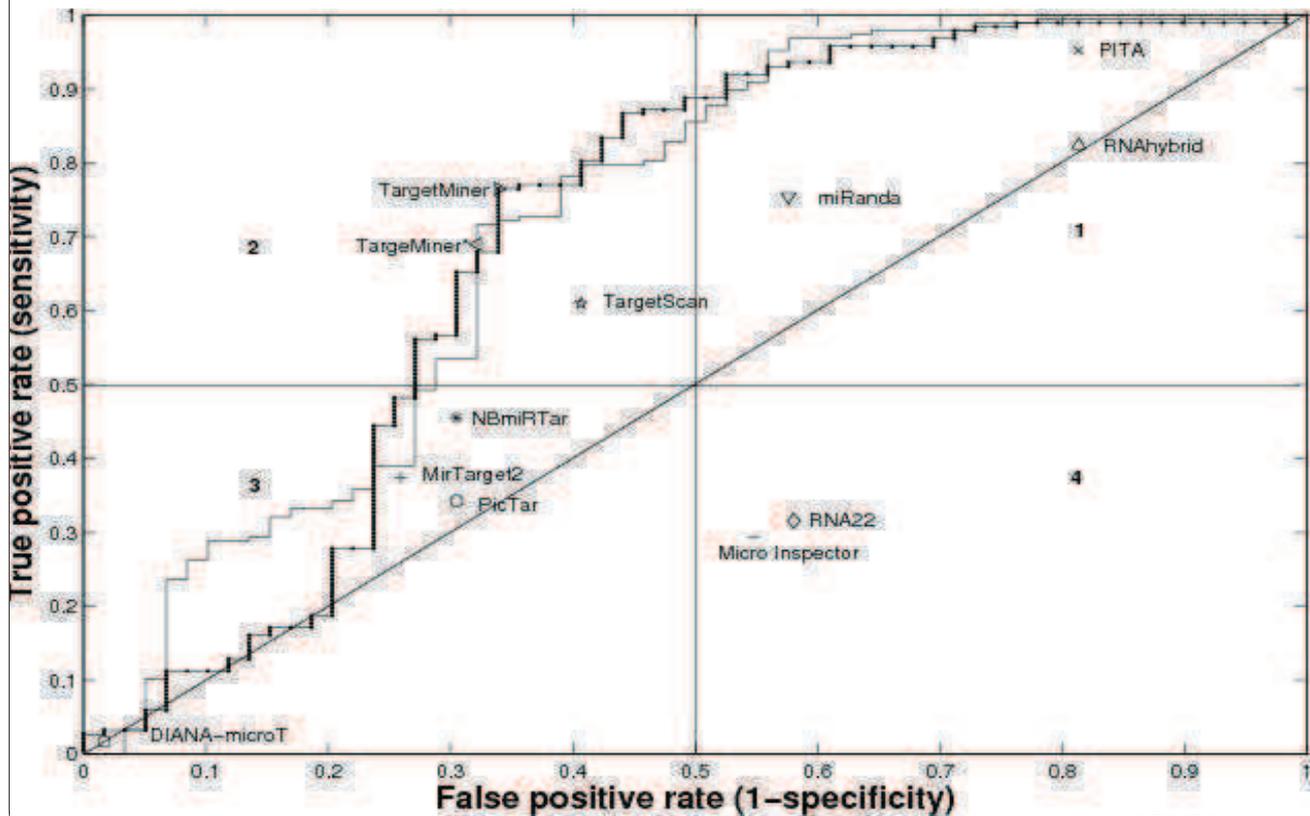
	miRanda	NBmiRTar	PicTar	TargetScan	TargetMiner	TargetMiner*
MCC	0.167	0.129	0.034	0.174	0.384	0.321
ACA	0.589	0.574	0.518	0.601	0.713	0.684

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad Specificity = \frac{TN}{TN + FP} \times 100\% \quad MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{[TP + FP][TP + FN][TN + FP][TN + FN]}}$$

$$ACA = 1/c \sum_{i=1}^c \text{accuracy of class } i, \text{ where } c = \text{number of classes}$$

8/26/2014

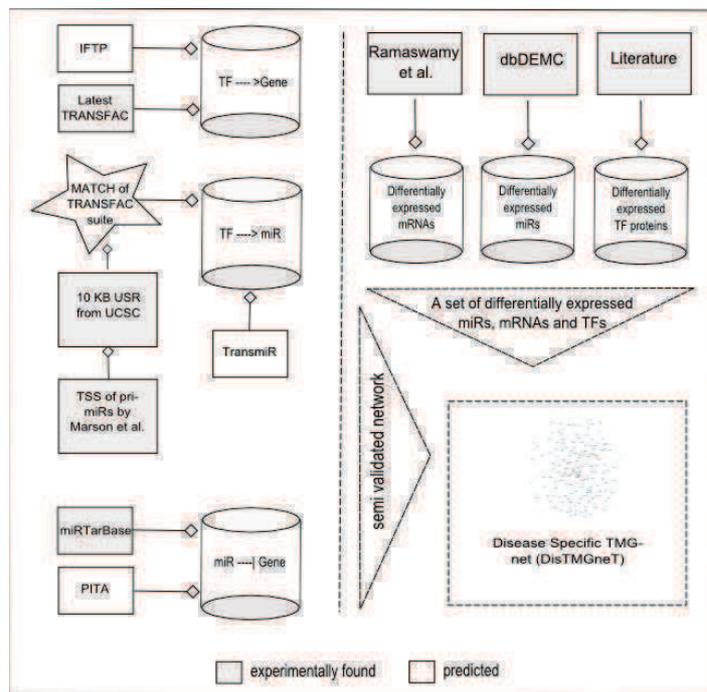
## ROC Plot



## Outline

- Introduction
- Components of the miRNA induced regulatory network
  - Regulation of miRNAs by TFs
  - Prediction of miRNA targets
- Construction of the miRNA induced regulatory network
- Network analysis
- Disease Specific Network and Analysis
- Summary

# TMG Network Generation



- Network can be built combining the three regulation types
- Interactions = Validated Interactions + Highly Recommended Predictions
- Called TMG-net (Transcription Factor-miRNA-Gene network)
- Genes are important to induce network from expression studies
- Genes were not considered in our earlier work [Sengupta and Bandyopadhyay, 2011]

## Outline

- Introduction
- Components of the miRNA induced regulatory network
  - Regulation of miRNAs by TFs
  - Prediction of miRNA targets
- Construction of the miRNA induced regulatory network
- Network analysis
- Disease Specific Network and Analysis
- Summary

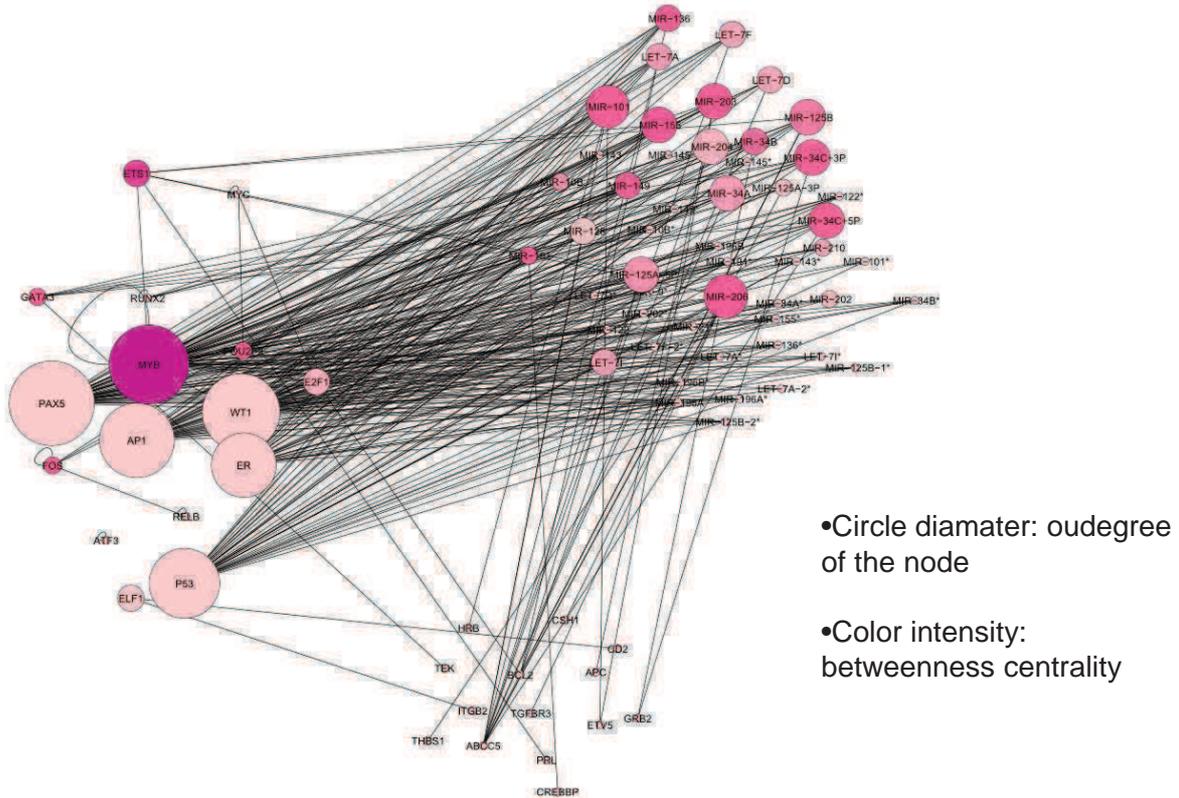
# Network Analysis - SCC

- Strongly Connected Component
  - every node has a directed path to the remaining ones
- 1009 out of 13975 molecules in TMG-Net form the large SCC
- 54% of these are TFs and rest are miRNAs
  - miRNAs are equally important
- The largest SCC plays a crucial role in inter-linking the biomolecules
- Cancer related pathways overlap significantly with the SCC
  - Pathway analysis is done separately for miRNAs and TFs in the largest SCC

## Outline

- Introduction
- Components of the miRNA induced regulatory network
  - Regulation of miRNAs by TFs
  - Prediction of miRNA targets
- Construction of the miRNA induced regulatory network
- Network analysis
- Disease Specific Network and Analysis
- Summary

## Disease Specific Network: Breast Cancer



## Network Properties

Topological attributes	Breast cancer specific DisTMGneT	Colon cancer specific DisTMGneT	Full network
No. of edges	307	441	294499
No. of vertices	81	152	13975
Density	0.05	0.02	0.002
Clustering coeff.	0.023	0.004	0.042
Largest SCC size	7	27	1009
Graph diameter	5	8	10
Characteristic path length	2.098	2.901	3.439

## Density of Disease Specific Networks

- Graph density is traditionally formulated as the ratio of existing edges to the possible number of edges
- Does not consider the distribution of density across the nodes in the graph
- InCov (Inductive Coverage): A new measure capturing the percentage of the network that can be induced by a set of nodes.
- This measure is able to capture the natural co-existence of a set of nodes

## Density of Disease Specific Networks

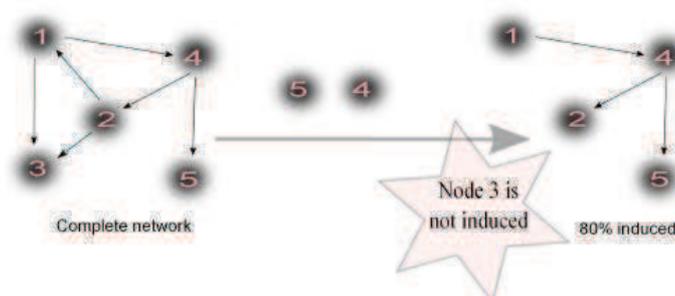
1. The complete set of vertices  $V$  is divided into  $n$  equal sized, non-overlapping subsets  $V_i$  where  $i = 1, 2, \dots, n$ .
2. For each of the vertex subset  $V_i$  compute the Inductive Coverage as follows:

$$InCov^{V_i} = \left[ \frac{|V_i \cup \bigcup_{v \in V_i} (N_v \cap V)|}{|V|} \times 100 \right] \%$$

Here,  $N_v$  indicates the set of immediate neighbors (irrespective of direction) of  $v$  irrespective of the edge directions.

3. Now compute the average InCov as follows:

$$InCov_n^V = \frac{\sum InCov^{V_i}}{n} \%$$

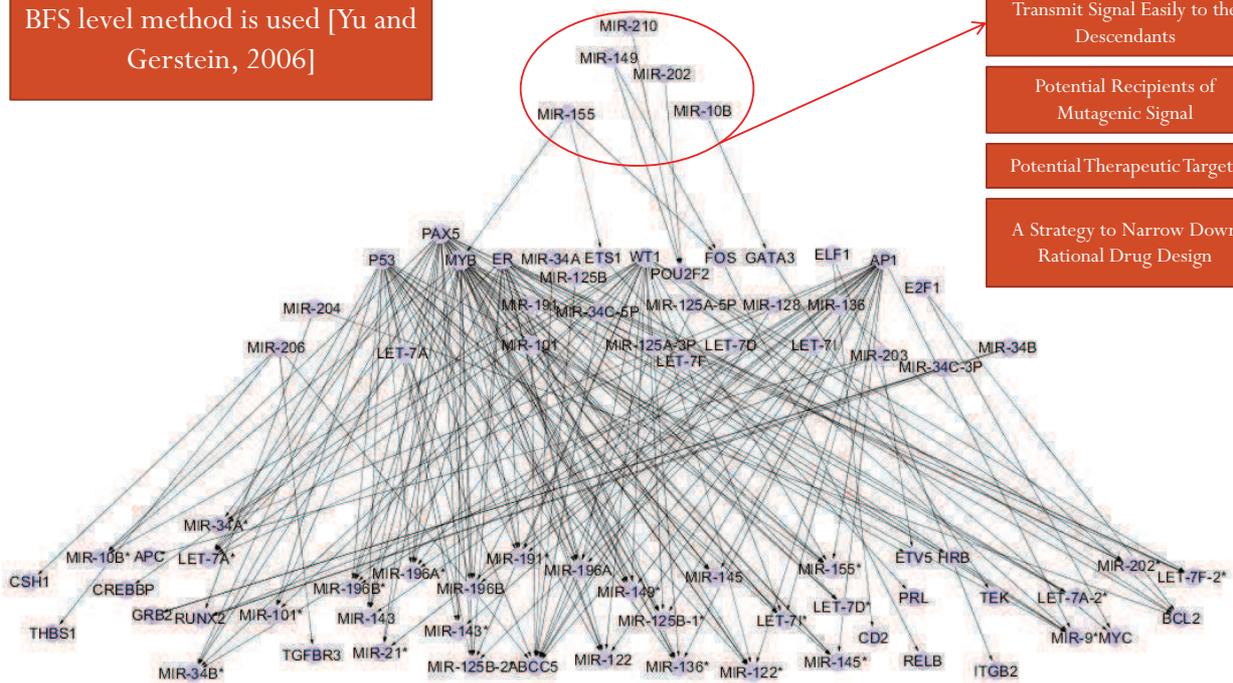


## Density of Disease Specific Networks

- InCov percentage for the breast cancer specific network = 84%
- InCov percentage for the colorectal cancer specific network = 78%
- P-value < 0.05 (degree preserving graph randomization) in both the cases

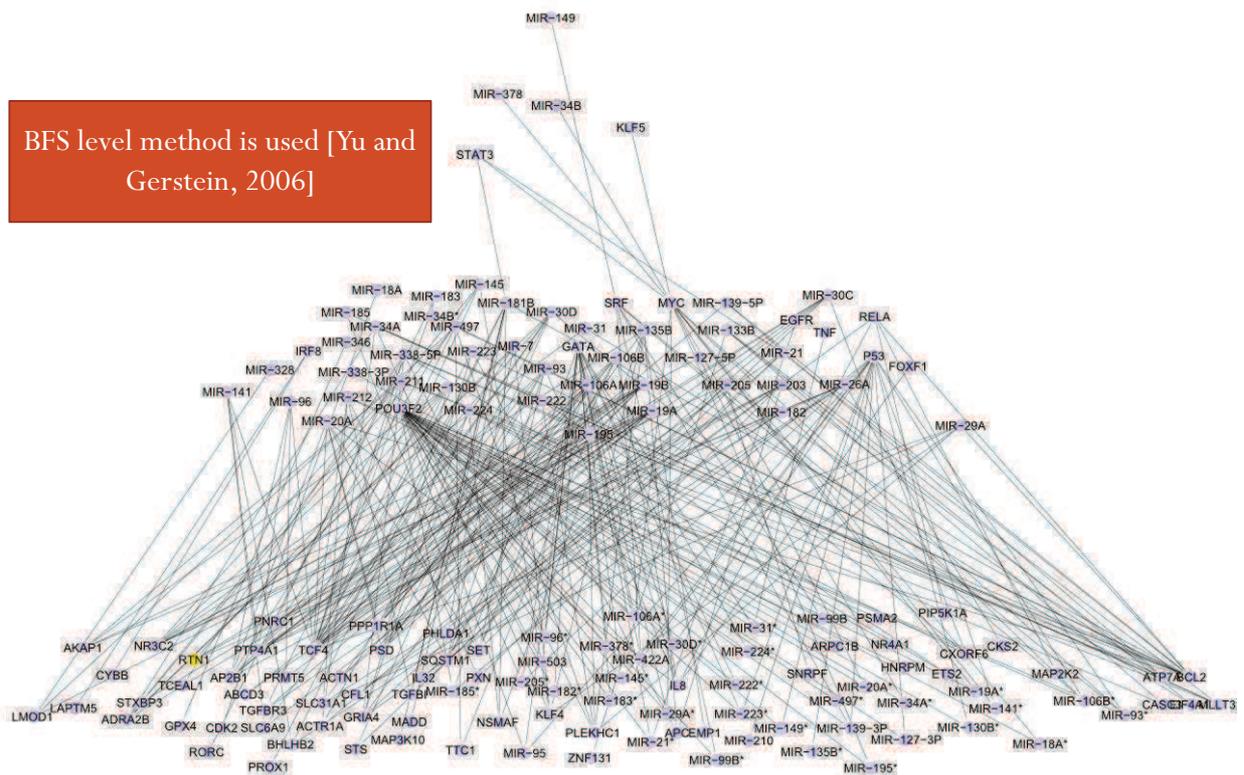
## Analyzing the Network Topology: Breast Cancer Specific Network

BFS level method is used [Yu and Gerstein, 2006]

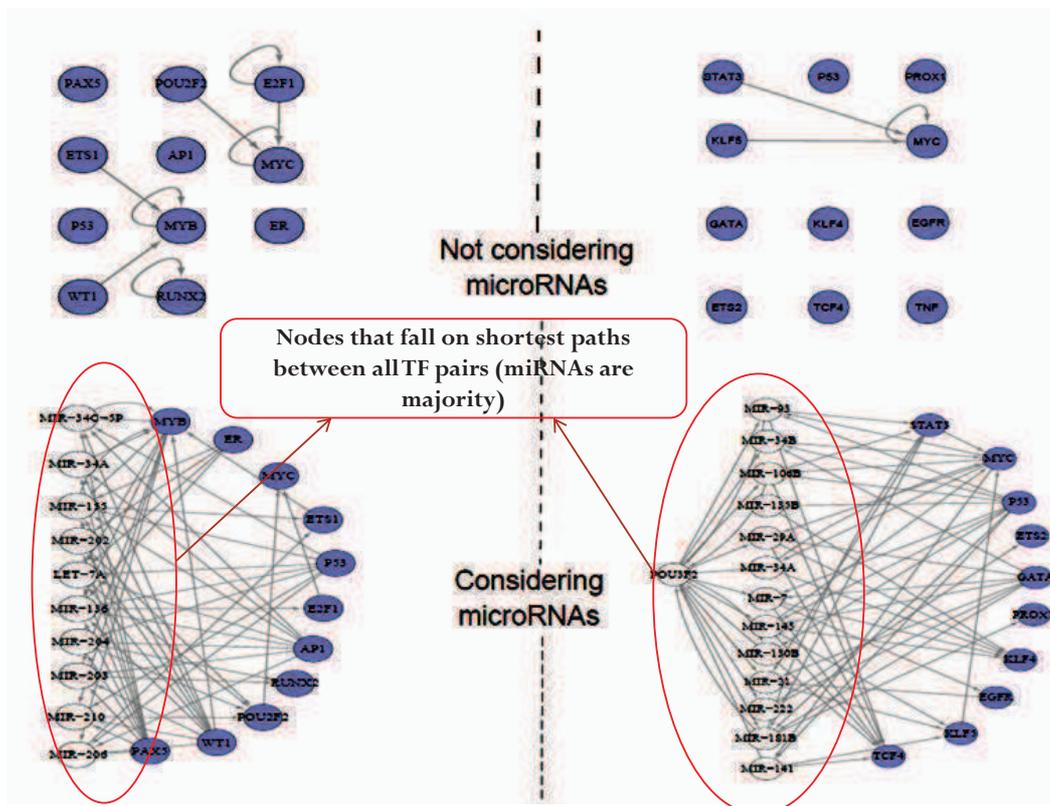


# Topology of Colorectal Cancer Specific Network

BFS level method is used [Yu and Gerstein, 2006]



## Interlinking Disease Specific TFs





## Top Level Molecules in Colorectal Cancer: Evidence from Literature

- miR-378: Oncomir specific to colorectal cancer [Feng *et al.*, 2011]
- miR-34B: Early screening marker of colorectal cancer [Kalimutho *et al.*, 2011]

**miR-149: Is this important?**

## Web Application: DisTMGneT

### DISTMGNET

DISEASE SPECIFIC TF GENE MICRORNA NETWORK

Select microRNA

hsa-miR-99b\*  
hsa-miR-99b  
hsa-miR-99a\*  
hsa-miR-99a  
hsa-miR-98  
hsa-miR-96\*  
hsa-miR-96  
hsa-miR-95  
hsa-miR-944  
hsa-miR-943  
hsa-miR-942  
hsa-miR-941  
hsa-miR-940  
hsa-miR-939  
hsa-miR-938

Select TF and Genes

A1CF  
A2BP1  
A2M  
A2ML1  
A4GALT  
AACS  
AACSL  
AADACL1  
AADAT  
AAK1  
AAMP  
AASDHPPT  
AATF  
AATK  
ABAT

Generate Network

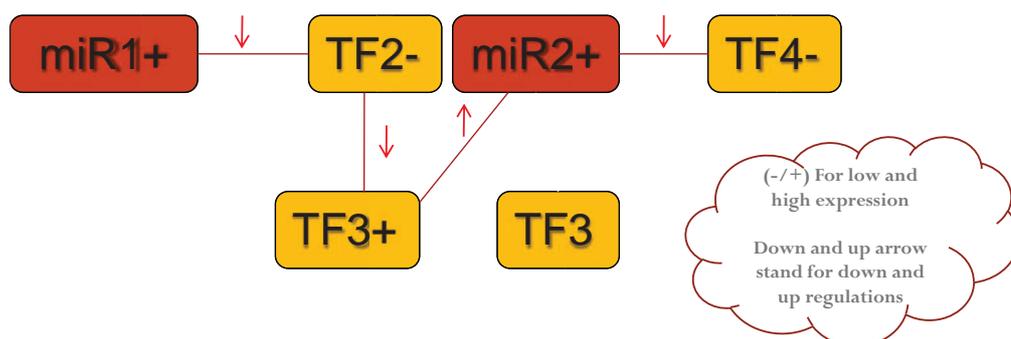


## Outline

- Introduction
- Components of the miRNA induced regulatory network
  - Regulation of miRNAs by TFs
  - Prediction of miRNA targets
- Construction of the miRNA induced regulatory network
- Network analysis
- Disease Specific Network and Analysis
- Summary

## Summary

- Importance of negative set selection for proper classifier design in target prediction
- Important to develop miRNA-specific TSS prediction systems
- MicroRNAs play important roles in regulatory network of the cell
- MicroRNAs appear to be important and quick disseminator of signals in different cancer types
- In [Sengupta 2011], we discussed about miRNA-miRNA indirect regulations



## Some Questions

- Whether oncomirs mostly belong to the largest SCC?
- Why the hierarchical topology?
- Why are all the top level miRNAs in the largest SCC?
- Why miRNAs play a major role in disease specific TF-TF regulations
  - Possibly our immune system cannot afford other TFs to be disturbed for keeping the crosstalk between TFs alive
  - The disturbance caused by miRNAs is less far-reaching than that of TFs

## Summary

- New methods of TSS prediction and target prediction of miRNAs
- Important to study the involvement of miRNAs in the transcriptional network of the cell.
  - An additional level of regulation
- Provides a holistic view of the regulatory network
- Provides information on miRNA-miRNA regulation
  - Validation based on disease analysis provides impressive results
- Likely to help in improving target prediction results.

## References

1. R. Shalgi *et al.* (2007) Global and Local Architecture of the Mammalian microRNA Transcription Factor Regulatory Network, *PLOS Computational Biology*, PMID: PMC1914371.
2. S. Bandyopadhyay and R. Mitra (2009) Targetminer: MicroRNA target prediction with systematic identification of tissue specific negative examples, *Bioinformatics*, 25: 2625-2631.
3. S. Bandyopadhyay *et al.* (2008) A simulated annealing based multi-objective optimization algorithm: AMOSA. *IEEE Transactions on Evolutionary Computation*, 12, 269–283.
4. S. Bandyopadhyay, R. Mitra, U. Maulik and M. Q. Zhang, "Development of the Human Cancer microRNA Network", *BMC Silence*, vol. 1, art no. 6, 2010
5. G. Calin *et al.* (2005) A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *New England Journal of Medicine*, 353:1793-1801.
6. J. Lu *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, 35:834-838.
7. R. Mitra and S. Bandyopadhyay, AMOSA-Targetminer: A novel multi objective optimization based miRNA-target prediction method, (under review).
8. M. Selbach *et al.* (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455:58–63.

## References

8. D. P. Bartel (2004) MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell*, 116:2, 281-297.
9. D. P. Bartel (2009) MicroRNAs: target recognition and regulatory functions, *Cell*, 136:2, 215-233.
10. D. L. Corcoran *et al.* (2009) Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data, *PLoS ONE*, 4:4, e5279.
10. T. A. Down and T. J. Hubbard (2002) Computational detection and location of transcription start sites in mammalian genomic DNA, *Genome Research*, 12, 458-461.
11. S. Fujita and H. Iba (2009) Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates, *Bioinformatics*, 24, 303-308.
12. M. Gardiner-Garden and M. Frommer (1987) CpG islands in vertebrate genomes, *Journal of Molecular Biology*, 196:2, 261-282.
13. S. Griffiths-Jones *et al.* (2008) miRBase: tools for microRNA genomics, *Nucleic Acids Research (Database issue)*, 36, D154-D158.
14. Q. Jiang *et al.* (2008) miR2Disease: a manually curated database for microRNA deregulation in human disease, *Nucleic Acids Research (Database issue)*, 37, D98-D104.
15. D. Karolchik *et al.* (2004) The UCSC Table Browser data retrieval tool, *Nucleic Acids Research (Database Issue)*, 32, D493-D496.

## References

16. R. Lister *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences, *Nature*, 462, 315-322.
17. U. Ohler *et al.* (2000) Stochastic segment models of eukaryotic promoter regions, *Proceedings of the Pacific Symposium on Biocomputing*, Honolulu, 5, 377-388.
18. H. K. Saini *et al.* (2007) Genomic analysis of human microRNA transcripts, *Proceedings of the National Academy of Sciences USA*, 104:45, 17719-17724.
19. X. Wang *et al.* (2007) Prediction of Transcription Start Sites Based on Feature Selection Using AMOSA, *Proceedings of the 6th Annual International Conference on Computational Systems Bioinformatics*, San Diego, California, 183-193.
20. S. Bandyopadhyay and M. Bhattacharyya (2009) Analyzing miRNA co-expression networks to explore TF-miRNA regulation, *BMC Bioinformatics*, 10:163.
21. S. Bandyopadhyay and M. Bhattacharyya (2010) PuTmiR: A database for extracting neighboring transcription factors of human microRNAs, *BMC Bioinformatics*, 11:190.
22. D. H. Tran *et al.* (2008) Finding microRNA regulatory modules in human genome using rule induction, *BMC Bioinformatics*, DOI: 10.1186/1471-2105-9-S12-S5.
23. S. Yoon and G. D. Micheli (2005) Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics*, 21(2):ii93-ii99.
24. S. van Drogen (2000) MCL - an algorithm for clustering graphs. (URL: <http://micans.org/mcl>).

## References

25. Zhao *et al.*, dbDEMC: a database of differentially expressed miRNAs in human cancers, *BMC Genomics*, doi:10.1186/1471-2164-11-S4-S5, 2010
26. Theis *et al.*, PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes, *Genome Biology*, 11:R6, 2010
27. Esquela-Kerscher *et al.*, Oncomirs — microRNAs with a role in cancer, *Nature Reviews Cancer* 6, 259-269, doi:10.1038/nrc1840, 2006
28. Lu M, Zhang Q, Deng M, Miao J, Guo Y, *et al.* (2008) An Analysis of Human MicroRNA and Disease Associations. *PLoS ONE* 3(10): e3420. doi:10.1371/journal.pone.0003420.

## Additional References

Camps C, Buffa FM, Colella S, Moore J, Sotiriou C, Sheldon H, Harris AL, Gleadle JM, Ragoussis J: **hsa-miR-210 Is induced by hypoxia and is an independent prognostic factor in breast cancer.** *Clin Cancer Res* 2008, **14**(5):1340–8.

Sengupta D, Bandyopadhyay S: **Participation of microRNAs in human interactome: extraction of microRNA-microRNA regulations.** *Molecular Biosystems* 2011, **7**:1966–1973.

Jiang S, Zhang HW, Lu MH, He XH, Li Y, Gu H, Liu MF, Wang ED: **MicroRNA-155 Functions as an OncomiR in Breast Cancer by Targeting the Suppressor of Cytokine Signaling 1 Gene.** *Can Res* 2010, **70**(3119).

Ma L, Teruya-Feldstein J, Weinberg RA: **Tumour invasion and metastasis initiated by microRNA-10b in breast cancer.** *Nature* 2007, **449**(7163):682–8.

Jin L, Hu WL, Jiang CC, Wang JX, Han CC, Chu P, Zhang LJ, Thorne RF, Wilmott J, Scolyer RA, Hersey P, Zhang XD, Wu M: **MicroRNA-149\*, a p53-responsive microRNA, functions as an oncogenic regulator in human melanoma.** *PNAS* 2011, **108**(38):15840–5.

Feng M, Li Z, Aau M, Wong CH, Yang X, , Yu Q: **Myc/miR-378/TOB2/cyclin D1 functional module regulates oncogenic transformation.** *Oncogene* 2011, **30**:2242–2251.

Kalimutho M, Di Cecilia S, Del Vecchio Blanco G, Roviello F, Sileri P, Cretella M, Formosa A, Corso G, Marrelli D, Pallone F, Federici G, Bernardini S: **Epigenetically silenced miR-34b/c as a novel faecal-based screening marker for colorectal cancer.** *Br. J Cancer* 2011, **104**(11):1770–8.





# From Experimental Data to Biological Networks and vice versa

Florence d'Alché-Buc

Joint work with C. Brouard, M. Szafranski<sup>1</sup>, A.  
Edelman (Paris-Descartes), A. Mezine<sup>1</sup>, A.  
Llamosi<sup>1</sup>, V. Letort (MAS, ECP) and  
M. Sebag (LRI, UPsud).





# From experimental data to biological networks and vice versa

Florence d'Alché-Buc<sup>1,2</sup>

Joint work with C. Brouard<sup>1</sup>, M. Szafranski<sup>1</sup>, A. Edelman (Paris-Descartes), A. Mezine<sup>1</sup>, A. Llamosi<sup>1</sup>, V. Letort (MAS, ECP) and M. Sebag (LRI, UPsud).

<sup>1</sup>IBISC EA 4526, Université d'Évry Val d'Essonne, Évry cedex, France

<sup>2</sup>**Current address:** LTCI, umr CNRS 5141, Télécom ParisTech, Paris, France

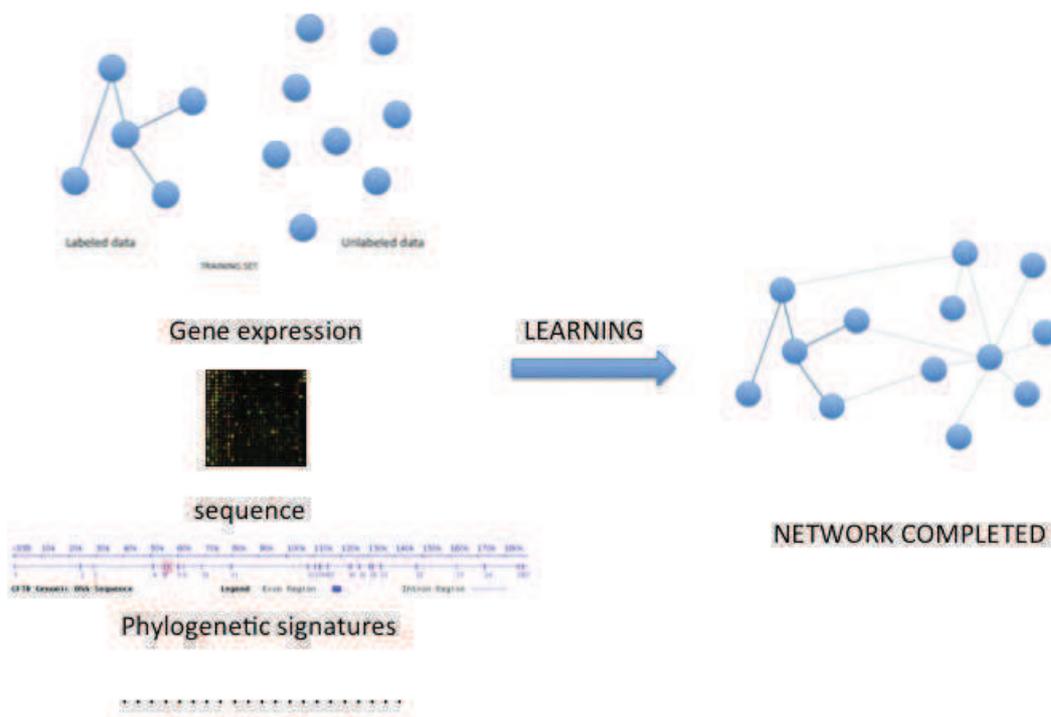
Email: [florence.dalche@ibisc.fr](mailto:florence.dalche@ibisc.fr), [florence.dalche@telecom-paristech.fr](mailto:florence.dalche@telecom-paristech.fr)



## Reverse-Modeling of biological networks

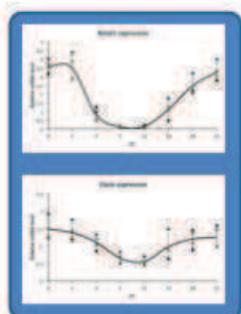
- **Task 1:** given a partially known physical protein-protein network and (experimental) data, predict unknown interactions
- **Task 2:** Infer a gene regulatory network from gene expression datasets
- **Task 3:** Given a gene regulatory network, a (dynamical) model and gene expression datasets, estimate model parameters and hidden states

# Task 1: link prediction in a protein-protein interaction



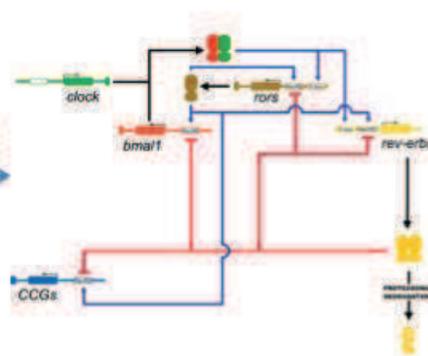
# Task 2 : gene regulatory network inference

Kinetics of gene expression  
(liver clock, mouse)

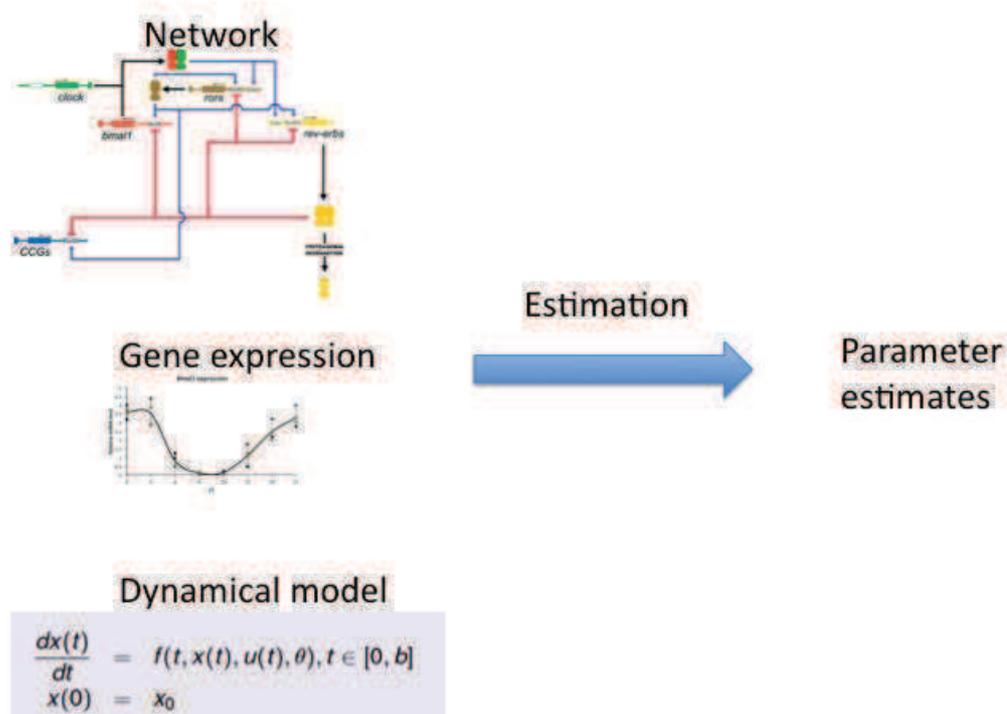


Network inference

Circadian clock



# Task 3 : parameter estimation in a gene regulatory network model



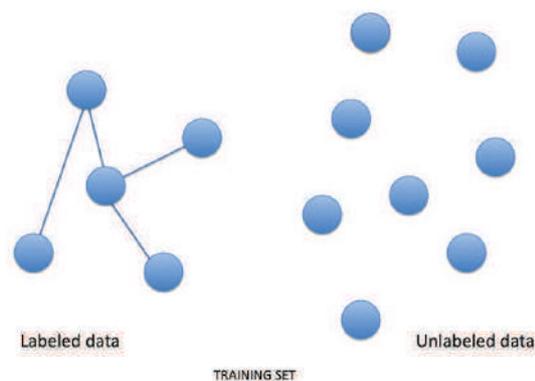
## Research activities

- Link prediction / network inference
  - ▶ A versatile theoretical framework for penalized vector-valued regression based on operator-valued kernels
  - ▶ Vector-valued functions of the following form:
$$h(x) = \sum_{i=1}^n K(x, x_i) \alpha_i$$
where  $K$  is a matrix-valued kernel and  $\alpha_i$ 's are parameter vectors
  - ▶ Brouard et al. 2011, 2014, Lim et al. 2013, 2014
- Parameter estimation in biological models
  - ▶ Active learning for experimental design
  - ▶ Meyer et al. 2014, Llamosi et al. 2014

# Outline

- 1 Overview
- 2 Link prediction in protein-protein interaction network
- 3 Active learning for parameter estimation
- 4 Conclusion

## Semi-supervised link prediction



### Goal

- Learn a function  $f : \mathcal{X} \times \mathcal{X} \rightarrow \{0, 1\}$  from:
  - ▶ descriptions of  $(\ell + u)$  proteins in  $\mathcal{X}$  (localization, sequence, gene expression ...),
  - ▶ all edges known for the  $\ell$  first proteins
- $f$  predicts 1 if there is an interaction, 0 otherwise

# Operator-valued kernel for link prediction

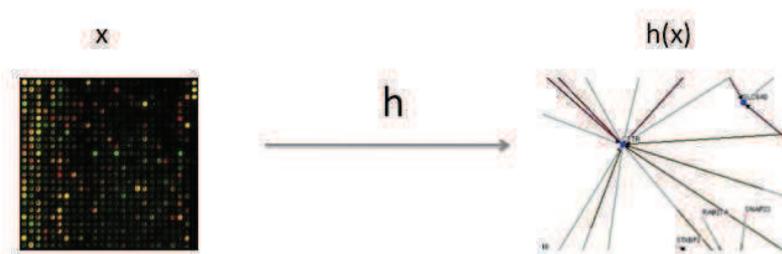
## Similarity-based model:

Given the description of two proteins  $x$  and  $x'$ ,

$$f_{\theta}(x, x') = \text{sign}(\hat{k}_y(x, x') - \theta)$$

$k_y(x, x')$  indicates the proximity between two nodes in the given graph

## Operator-valued kernel-based model



### Idea

Build a function  $h$ :  $h(x) = \sum_{i=1}^n K(x, x_i) \alpha_i$  that predicts the position of a protein in the network. Then  $k_y$  is approximated using the inner product :

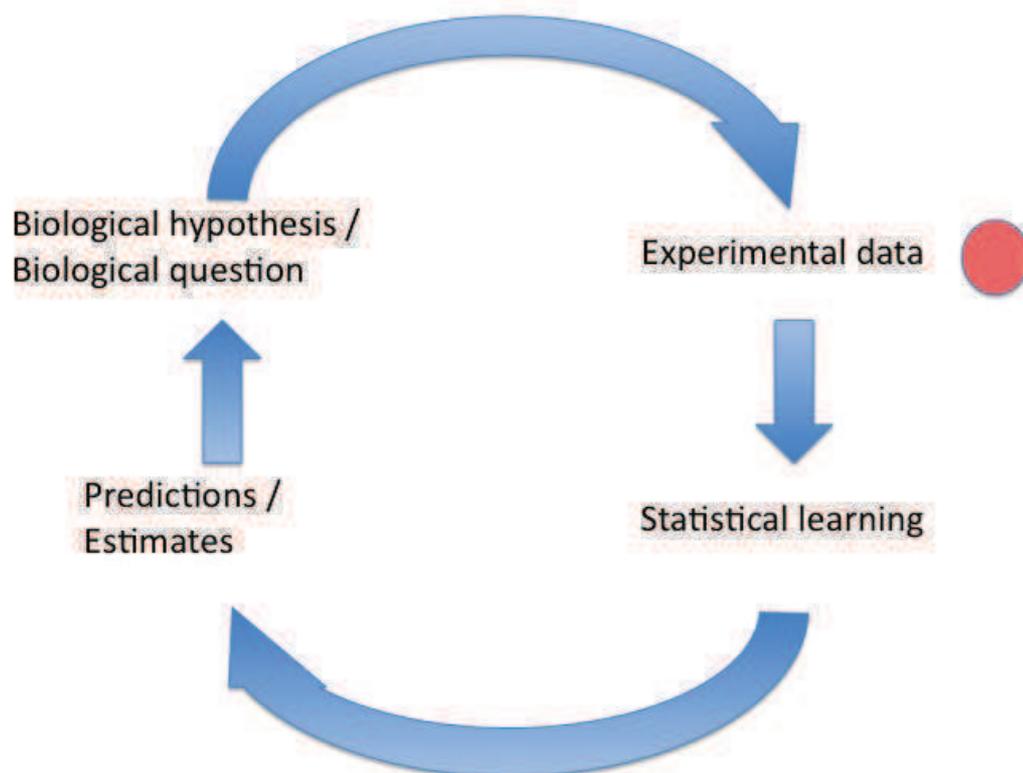
$$\hat{k}_y(x, x') = \langle h(x), h(x') \rangle,$$

where  $\langle h(x), h(x') \rangle$  is a inner product in a well chosen space.

# Results for the CFTR network

- CFTR is a protein whose mutations are involved in cystic fibrosis
- Collaboration with Alexander Edelman (Paris-Descartes), PhDthesis of Celine Brouard
- Prediction of new protein-protein interactions
- **Current work:** wet-lab validation by the group of Alexander Edelman

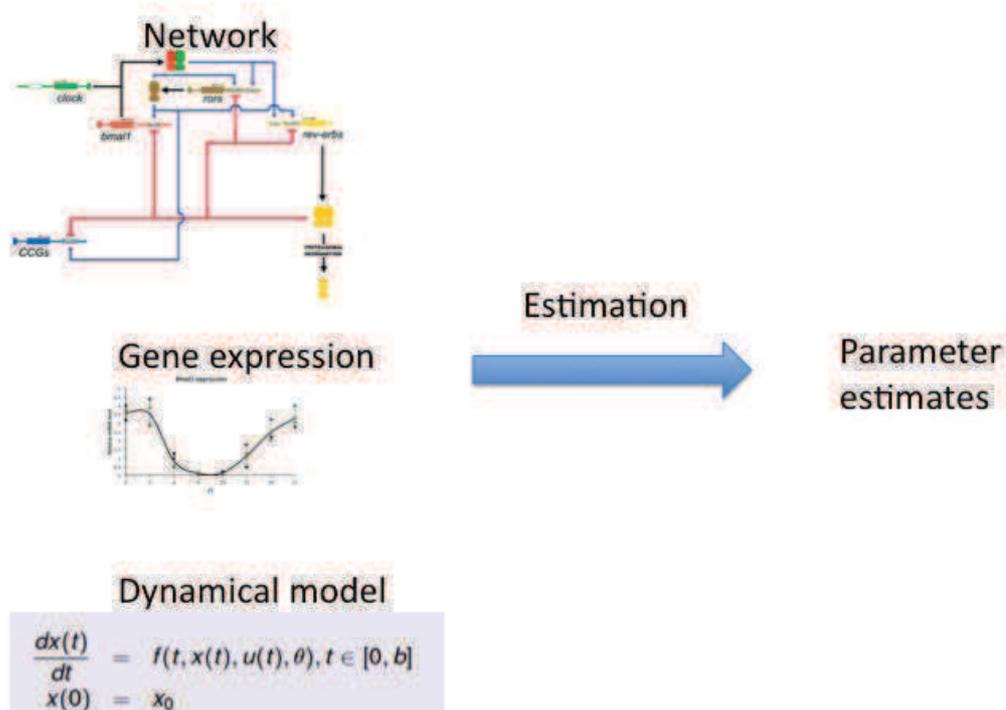
## Loop of interaction with biologists



# Outline

- 1 Overview
- 2 Link prediction in protein-protein interaction network
- 3 Active learning for parameter estimation
- 4 Conclusion

## Task 3 : parameter estimation in a gene regulatory network model



# Parameter estimation in ordinary differential equations

## Goal

- ODE model:  $f \in \mathcal{F}$ ,

$$\frac{dx(t)}{dt} = f(x(t), u(t), \theta)$$

- A partially and noisy observation model:

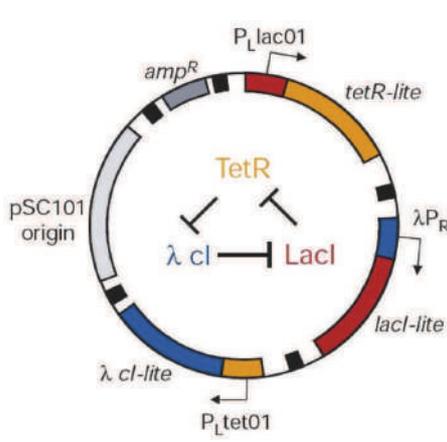
$$y(t) = h(x(t)) + \epsilon(t)$$

where  $h$  is an observation function,  $\epsilon(t)$  is a i.i.d noise

- A sequence of observed data :  $y_{0:N} = \{y_0, \dots, y_N\}$  at time  $t_0, t_2, \dots, t_N$
- **Estimate parameter  $\theta$  (and hidden state)**

## Example: the Repressilator described with Hill kinetics

[Elowitz and Leibler, Nature 2000]



$$\frac{dr_1}{dt} = v_1^{max} \frac{k_{12}^n}{k_{12}^n + p_2^n} - k_1^{mRNA} r_1$$

$$\frac{dr_2}{dt} = v_2^{max} \frac{k_{23}^n}{k_{23}^n + p_3^n} - k_2^{mRNA} r_2$$

$$\frac{dr_3}{dt} = v_3^{max} \frac{k_{31}^n}{k_{31}^n + p_1^n} - k_3^{mRNA} r_3$$

$$\frac{dp_1}{dt} = k_1 r_1 - k_1^{protein} p_1$$

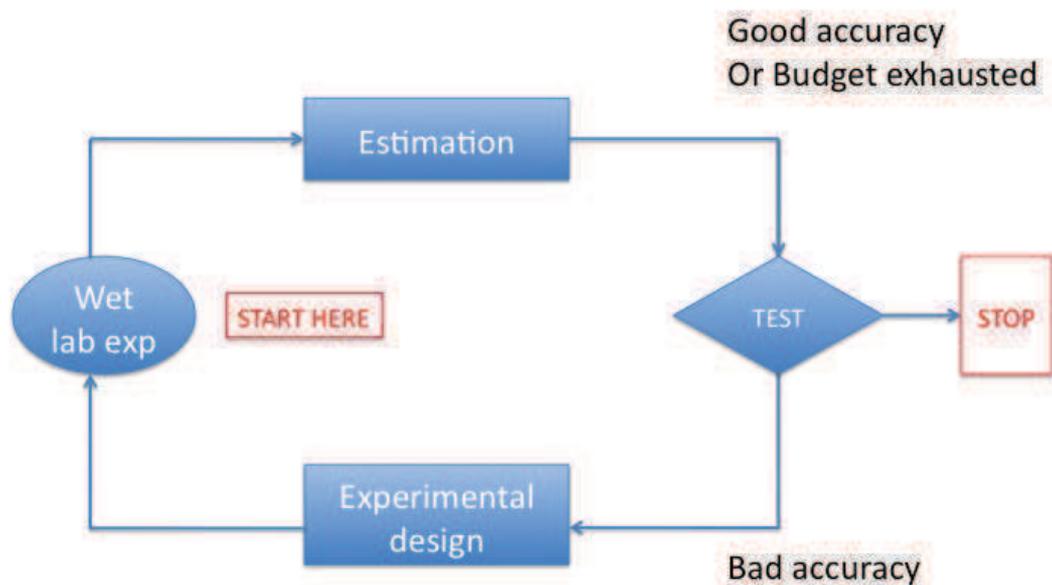
$$\frac{dp_2}{dt} = k_2 r_2 - k_2^{protein} p_2$$

$$\frac{dp_3}{dt} = k_3 r_3 - k_3^{protein} p_3$$

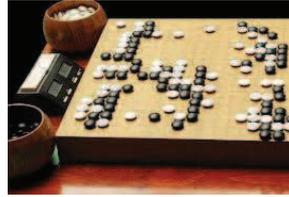
# Parameter estimation in ODE: difficulties

- Measurements of gene expression without any perturbation on the network are not sufficient to provide accurate estimate of hidden states and parameters
- A perturbation experiment (like knock-out, knock-down . . . ) allows to provide valuable information on the dynamical system and solve the **non-identifiability** issue
- **Question:** how to choose them ?

## Sequential experimental design



# Sequential experimental design as a game



- We imagine we are playing a game against nature
- We attempt to optimize our moves to win the game
- We wish to learn how to play the game when playing it

## Experimental design as a game:

### At time $t$

- **State of the game:** current set of hypotheses, current set of available experiments, current budget  $B$
- **Making the move:** we use a Monte-Carlo Tree Search (MCTS) to provide an estimation of the best move to do given it is followed by  $(B-1)$  experiments (here the best experiment to do, given the budget)
- **Nature answers in turn:** the chosen experiment is performed in the (wet) laboratory
- Then a set of new hypothesis is provided by an estimation procedure using all the available data : a posterior probability is given

# Modeling a game with a Markov Decision Process

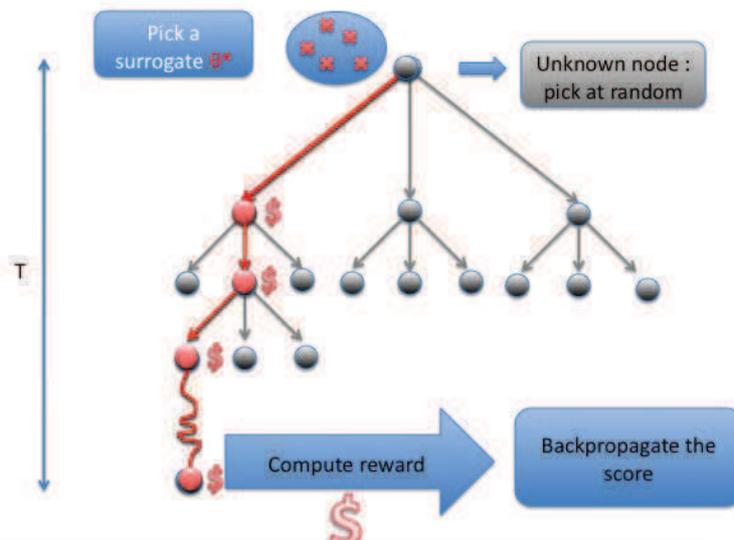
- At each time step, the process is in state  $S$
- The decision maker chooses an action  $a$
- The process responds at next step time by moving randomly to a new state  $S'$ , taking into account the action  $a$ , it also gives a reward  $R_a(S, S')$  to the decision maker
- When the rewards are unknown, it comes to *Reinforcement learning*
- Need to estimate the reward we get → Monte-Carlo Tree Search (Browne et al. 2012)

## Proposed algorithm: EDEN

- **Initialization**
- WHILE(budget not exhausted) and (estimates not accurate)
  - ▶ Design a new experiment using MCTS
  - ▶ Perform the proposed experiment and update the budget accordingly
  - ▶ Re-estimate parameters with the multiple dataset (previous dataset augmented with the new experimental data)
  - ▶ Evaluate the accuracy of estimates
- **END**

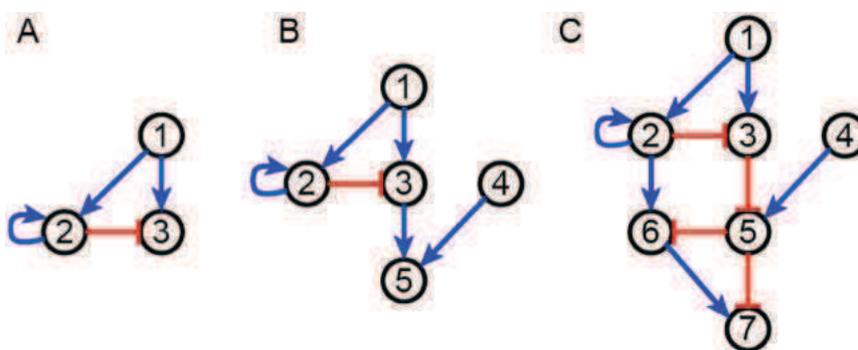
# Predict a move

- Use a set of surrogate parameters (oracles)
- Use a Monte-Carlo Tree search to estimate the average reward associated to a sequence of experiments
- Definitions of reward : variance of the estimate, distance between true parameter and estimates (bias)

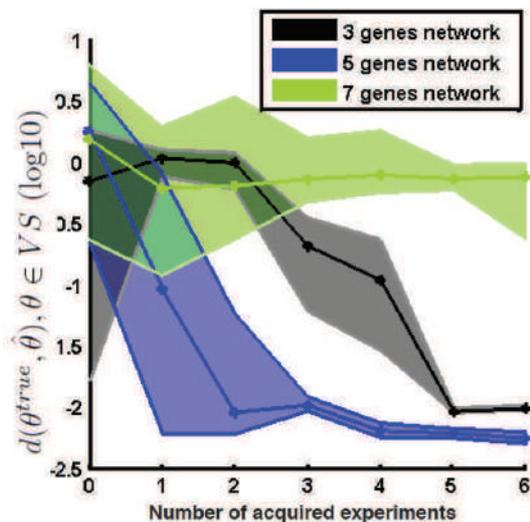


# A sample of results on DREAM7 subnetworks

- We consider KO, KD, ... experiments
- 3 networks



# A sample of results in DREAM7 subnetworks



## Outline

- 1 Overview
- 2 Link prediction in protein-protein interaction network
- 3 Active learning for parameter estimation
- 4 Conclusion

# Conclusion

## ● Part 1:

- ▶ New versatile tools for network inference with operator-valued kernels
- ▶ Classical loop of interaction with biologists
- ▶ Current progress on validation

## ● Part 2:

- ▶ Advocacy for a tighter collaboration between biologists and modelers to design experiments
- ▶ *Active learning* based on Monte-Carlo simulations allows to explore huge hypotheses sets : other target problems in systems, synthetic biology or personalized medicine can be addressed
- ▶ Scaling up these methods is still an issue !

## Related publications

### ● Protein-protein network inference

- ▶ A. Birlutiu, F. d'Alché-Buc, T. Heskes, A Bayesian Framework for Combining Protein and Network Topology Information for Predicting Protein-Protein Interactions, IEEE Trans. on Computational Biology and Bioinformatics, issue 99, Nov, 2014.
- ▶ Brouard, C., Guerrero, C., Brouillard, F., Ollero, M., Edelman, A. and d'Alché-Buc, F. (2012) Search for new CFTR-protein interactions using statistical learning. 6th European Cystic Fibrosis Young Investigator Meeting, Paris, France.
- ▶ C. Brouard, M. Szafranski, F. d'Alché-Buc, Semi-supervised penalized output kernel regression for link prediction, ICML 2011.

### ● Gene regulatory network inference

- ▶ Brouard, C., Vrain, C., Dubois, J., Castel, D., Debily, M.-A. and d'Alché-Buc, F. Learning a Markov Logic Network for supervised gene regulatory network inference. BMC Bioinformatics 14:273, 2013.
- ▶ Néhémy Lim, Yasin Senbabaoglu, George Michailidis, Florence d'Alché-Buc: OKVAR-Boost: a novel boosting algorithm to infer nonlinear dynamics and interactions in gene regulatory networks. Bioinformatics 29(11): 1416-1423 (2013)
- ▶ G. Michailidis and F. d'Alché-Buc, Autoregressive models for gene regulatory network inference: sparsity, stability and causality issues, in Special issue on : Parameter estimation in differential equations, Mathematical Biosciences, Springer, Available online 28 October (2013).

### ● Experimental design with active learning

- ▶ P. Meyer, T. Cokelaer, D. Chandran, K.H. Kim, P.-R. Loh, G. Tucker, M. Lipson, B. Berger, C. Kreutz, A. Raue, B. Steiert, J. Timmer, E. Bilal, DREAM 67 Parameter Estimation consortium (A. Mezine, A. Llamosi, V. Letort, F. d'Alché-Buc, . . .), H. M Sauro, G. Stolovitzky and J. Saez-Rodriguez, Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach, BMC Systems Biology 2014, 8:13.
- ▶ A. Llamosi, A. Mezine, F. d'Alché-Buc, V. Letort, M. Sebag, Experimental Design in Dynamical System Identification: A Bandit-Based Active Learning Approach. ECML/PKDD (2) 2014: 306-321.

# Algorithms, Metagenomics Analysis Platform, and their Applications for Understanding Role of Gut Microbiome in Human Health

Sharmila Mande, Ph.D  
Head, Bio-Sciences R&D  
TCS Innovation Labs  
Tata Consultancy Services  
Pune, India





# Algorithms, Metagenomics analysis platform, and their applications for understanding role of gut microbiome in human health

Sharmila Mande, Ph.D  
Head, Bio-Sciences R&D  
TCS Innovation Labs  
Tata Consultancy Services  
Pune, India

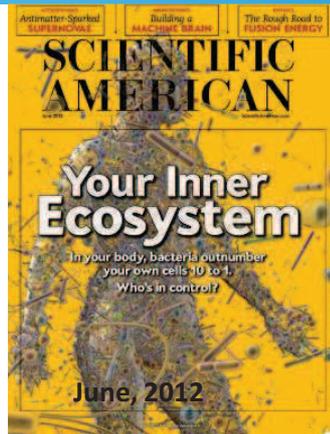
Copyright © 2011 Tata Consultancy Services Limited.

INAE/NATF Seminar on "Technology and Health-Care", Évry-Genopole, France, 15-Oct-2014

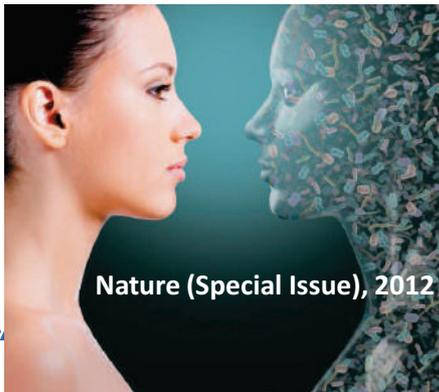
## Outline

- Metagenomics
- Analyzing Metagenomic data
- Gut microbiome and Health
  - Nutritional status
  - Antibiotic resistance profiles

# Microbes & Human health



- Human body is home to far more microbes than human cells
- Microbiome:** Totality of microbes and their genetic elements (genomes)



Nature (Special Issue), 2012

## Current Research Focus

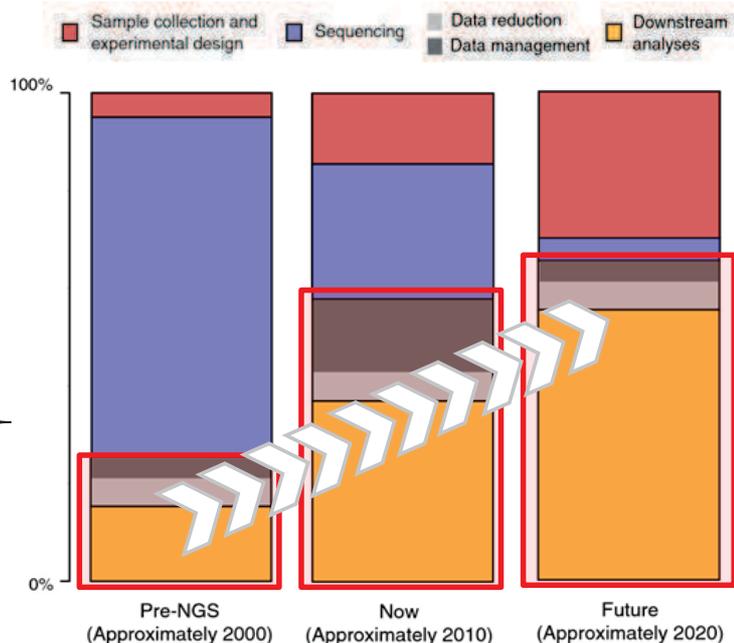
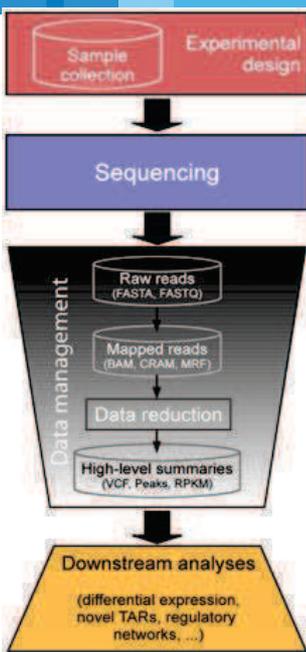
Correlating human health with resident microbial communities

## Metagenomics

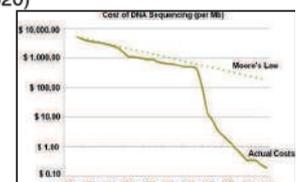
Enables characterization of microbial communities

3

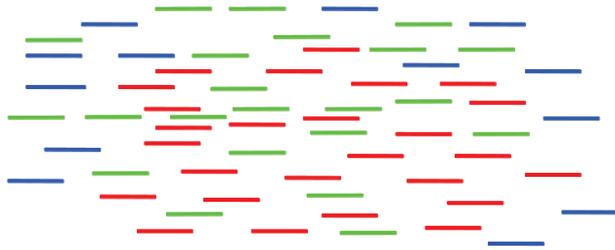
# Metagenomics: Workflow



Sboner et al. Genome Biology 2011 12:125 doi:10.1186/gb-2011-12-8-125]



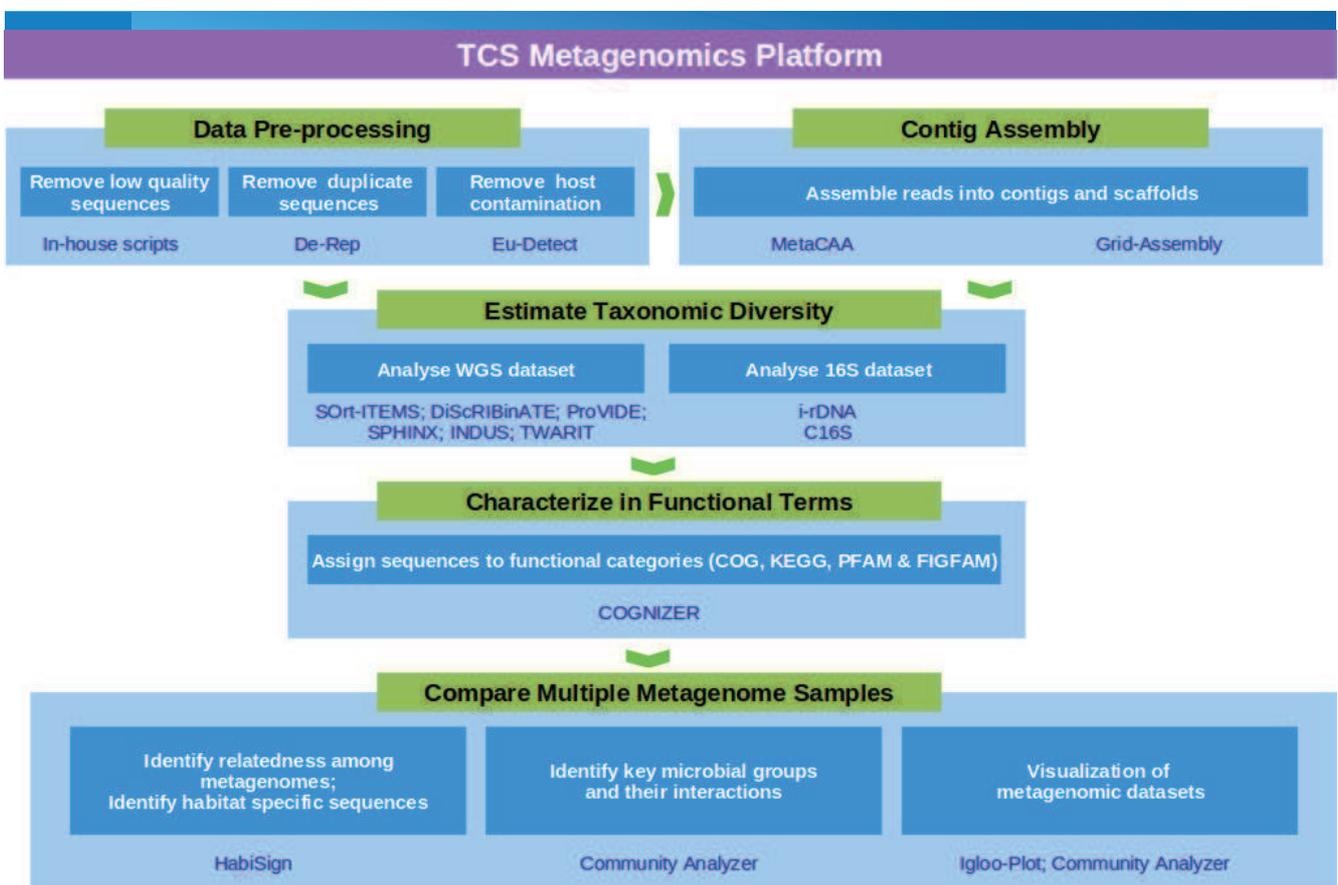
# Metagenomics : Questions to be answered



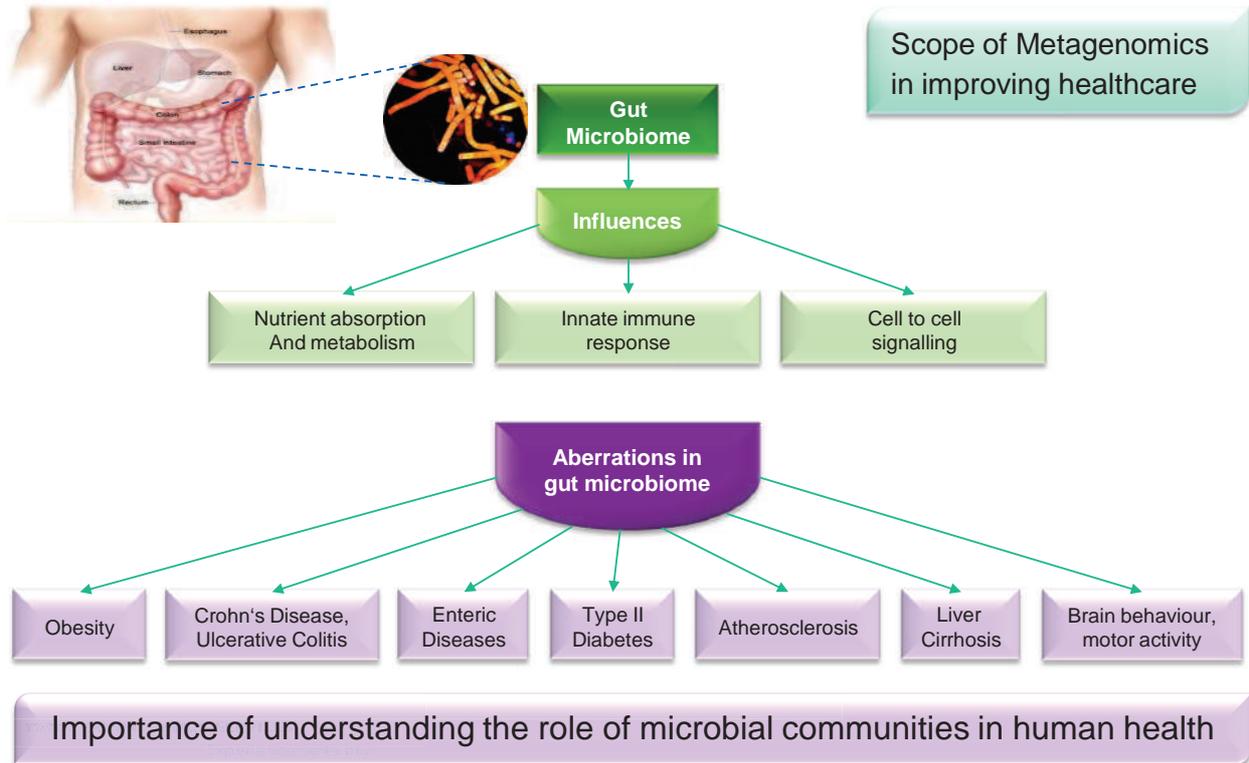
Majority of microbes are unknown!



- What microbes are there and what are their relative proportions?
- What functional genes do they have?
- How do they function?



# Gut Microbes and Diseases/Disorders



## What constitutes the healthy human gut microbiota?

**Aim 1** : Understanding the detailed role of intestinal microbiota in malnourishment

Gupta et al. *Gut Pathogens* 2011, 3:7  
<http://www.gutpathogens.com/content/3/1/7>

Highly accessed

Open access



Gut Pathogens

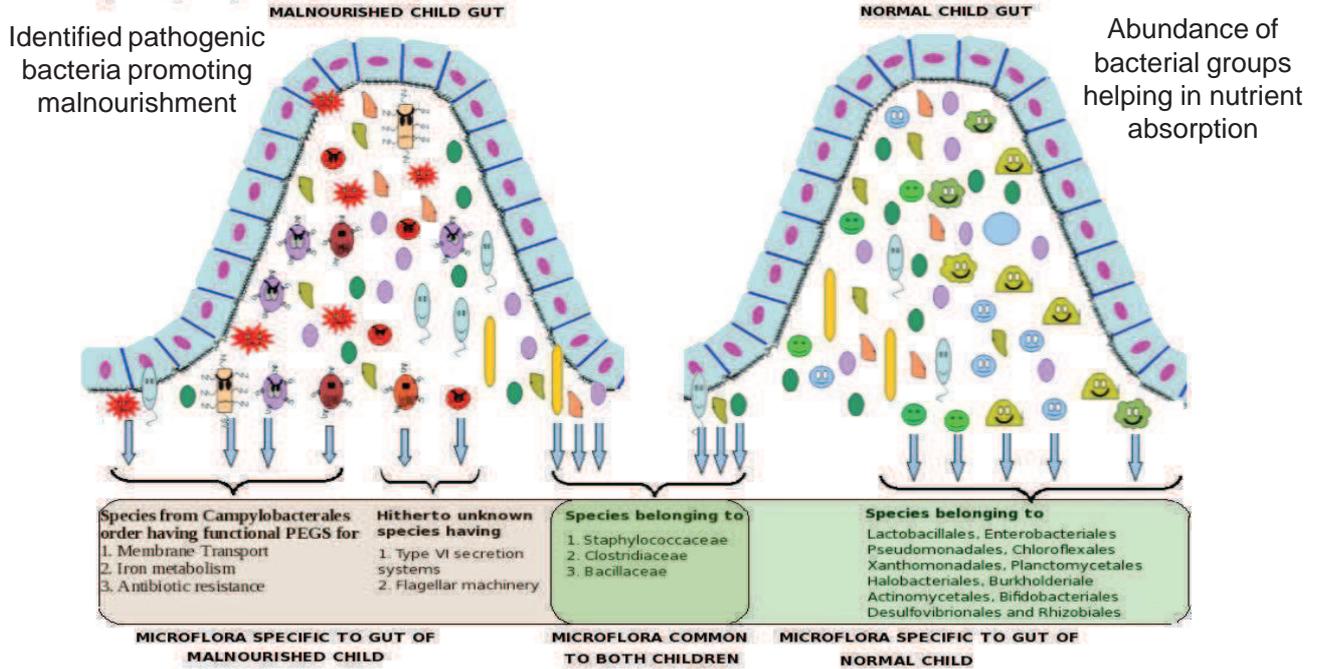
RESEARCH

Open Access

### Metagenome of the gut of a malnourished child

Sourav Sen Gupta<sup>1</sup>, Monzoorul Haque Mohammed<sup>2</sup>, Tarini Shankar Ghosh<sup>2</sup>, Suman Kanungo<sup>1</sup>, Gopinath Balakrish Nair<sup>1</sup> and Sharmila S Mande<sup>2\*</sup>

Figure-4 (Mande)



## What constitutes the healthy human gut microbiota?

**Aim 2 :** Understanding relationship between nutritional status and the microbial community in the gut.



Subject Areas

For Authors

About Us

Search



advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

1,052

VIEWS

12

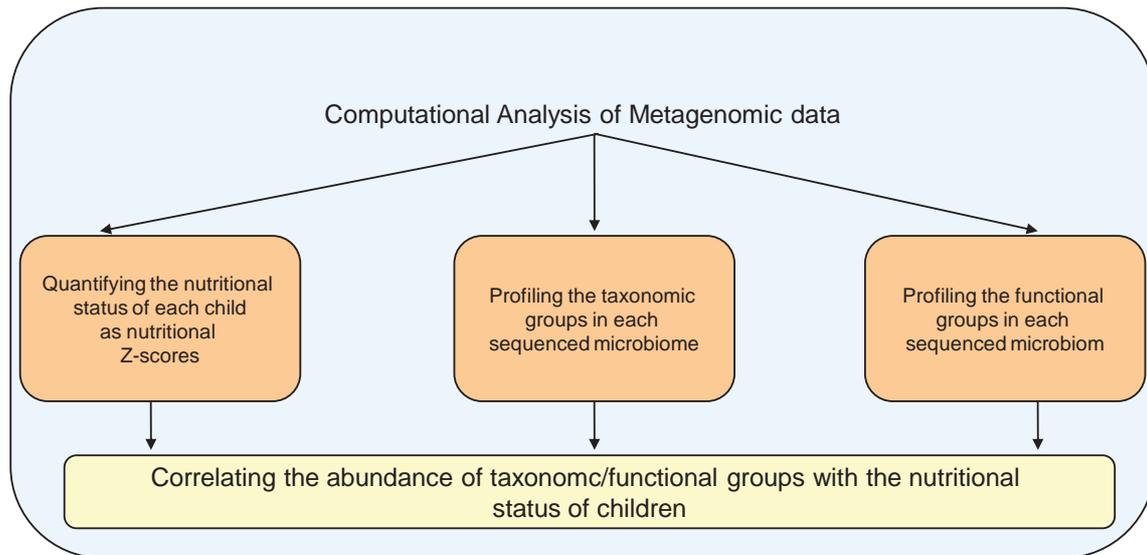
SHARES

### Gut Microbiomes of Indian Children of Varying Nutritional Status

Tarini Shankar Ghosh, Sourav Sen Gupta, Tanudeep Bhattacharya, Deepak Yadav, Anamitra Barik, Abhijit Chowdhury, Bhabatosh Das, Sharmila S. Mande, G. Balakrish Nair

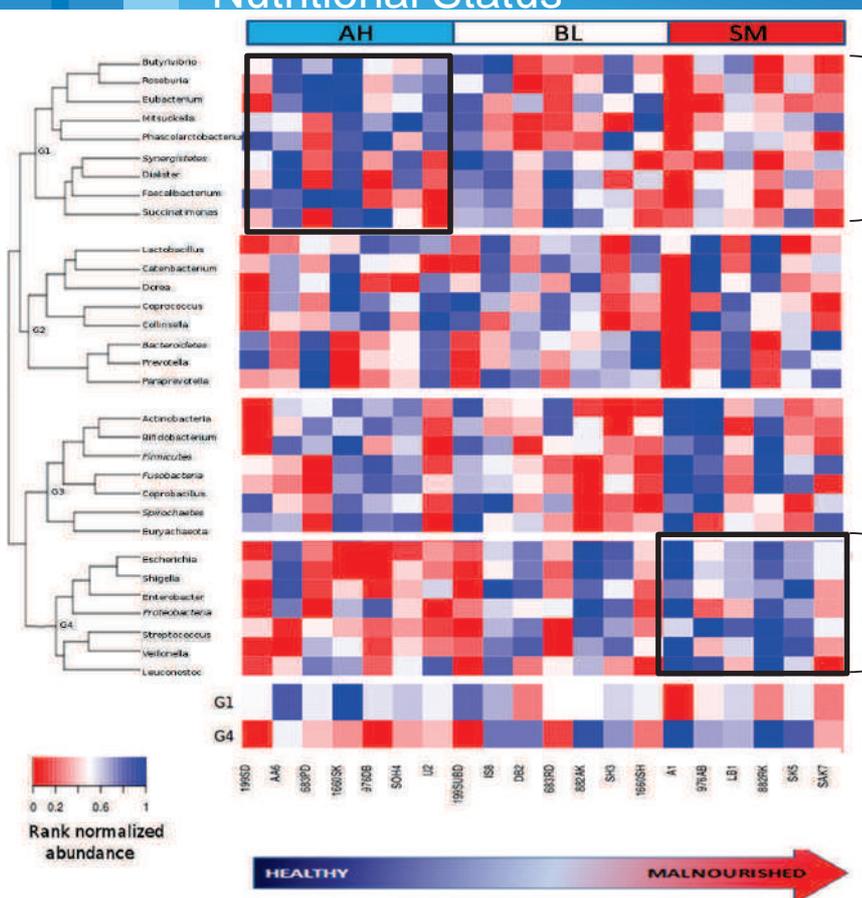
Published: April 24, 2014 • DOI: 10.1371/journal.pone.0095547

# Gut microbiome and nutritional status



Nutritional Z-scores  
 Height for age  
 Weight for age  
 Weight for height

## Correlating abundance of Taxonomic Groups with Nutritional Status



**G1** • Taxonomic markers positively or negatively influencing nutritional status

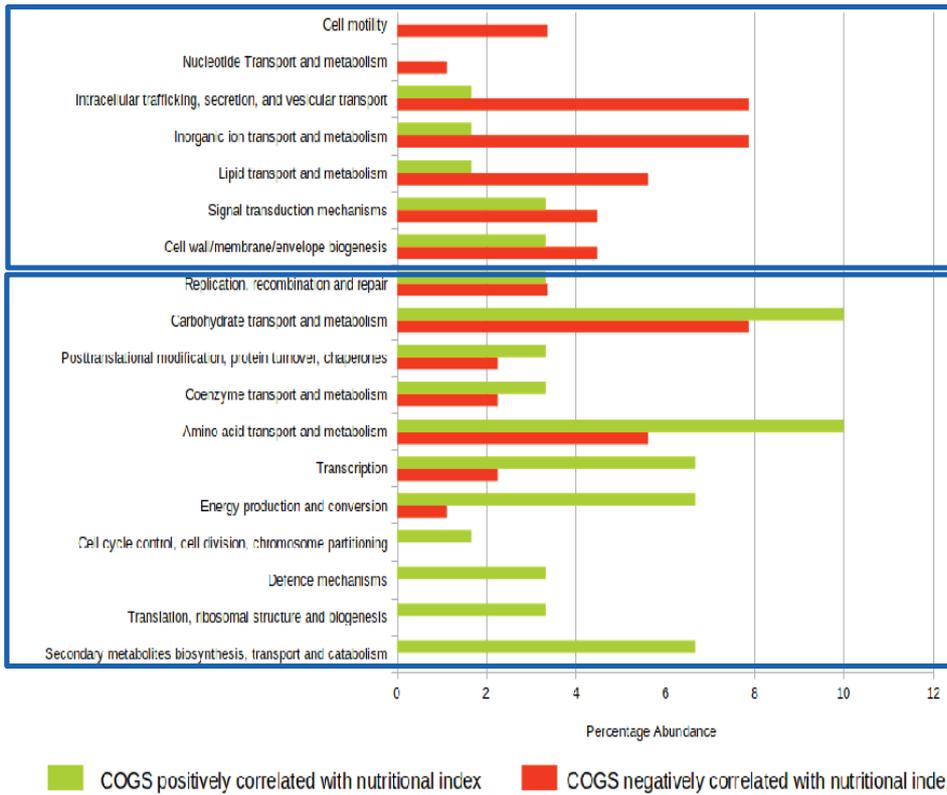
• Healthy markers :

- Probiotic and/or anti-inflammatory genera
  - Roseburia, Mitsuokella, Faecalibacterium
- Energy harvesters from dietary fiber
  - Butyrivibrio, Roseburia, Faecalibacterium
- Commensals
  - Eubacterium, Dialister

**G4** • Malnourishment markers :

- Potential Pathogenic genera
  - Escherichia, Enterobacter, Shigella, Streptococcus, Veillonella

# Comparison of Functional Categories in PC and NC COG Groups



• **Malnourished gut** enriched in COGS corresponding to

- Typically associated with pathogenesis

- Reflective of the high abundance of pathogenic genera

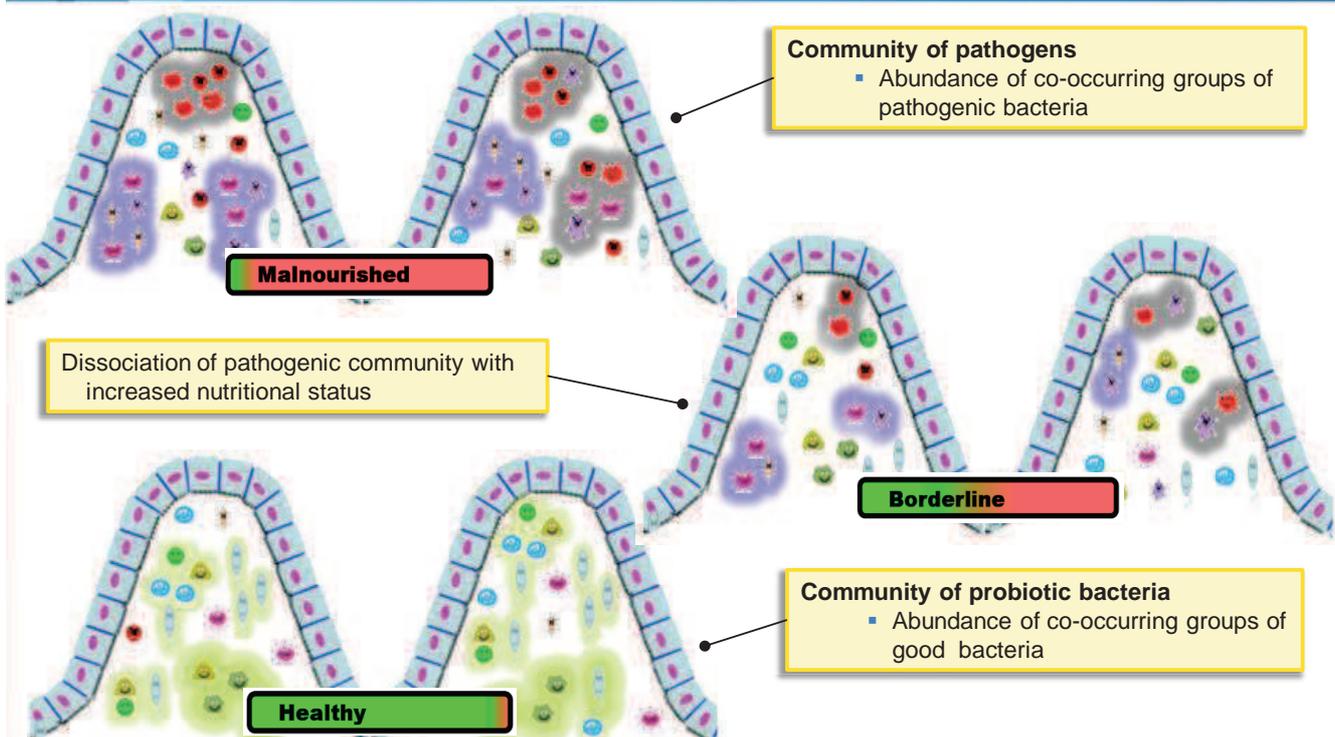
• **Healthy gut** enriched in COGS corresponding to

- 'Energy production and conversion' and 'Carbohydrate transport and metabolism'

• reflective of the high abundance energy harvesting genera

- Other functional categories associated with metabolism

# Gut microbiome and nutritional status



Prospects for development of probiotics and nutraceuticals

## Aim 3 : Antibiotic resistance genes in gut microflora

OPEN ACCESS Freely available online



### *In Silico* Analysis of Antibiotic Resistance Genes in the Gut Microflora of Individuals from Diverse Geographies and Age-Groups

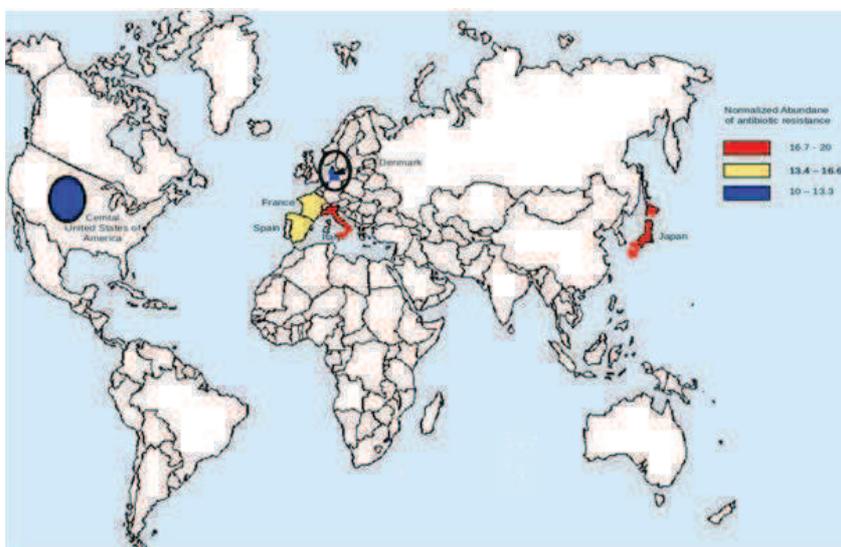
Tarini Shankar Ghosh<sup>1</sup>, Sourav Sen Gupta<sup>2</sup>, Gopinath Balakrish Nair<sup>2</sup>, Sharmila S. Mande<sup>1\*</sup>

<sup>1</sup>BioSciences R&D Division, TCS Innovation Labs, Tata Consultancy Services Ltd., Pune, Maharashtra, India, <sup>2</sup>Translational Health Sciences and Technology Institute, Gurgaon, Haryana, India

TATA CONSULTANCY SERVICES  
Experience certainty.

15

- Analyzed gut metagenomes of 275 individuals from 8 nationalities
- Checked the global trend of genes conferring resistance to around [240](#) antibiotics
- Compared country specific trends of the overall abundance of resistance genes



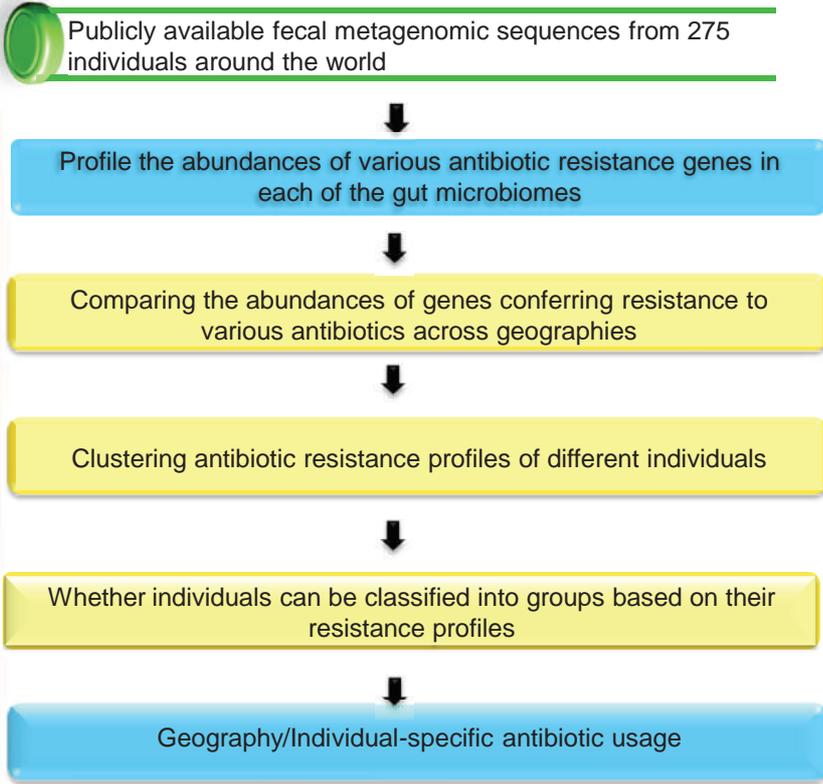
American  
Danish  
Spanish  
French  
Italian  
Japanese  
Chinese  
Indian (Children)

116 (Danish & Spanish)  
90 American (HMP)  
30 Chinese

TATA CONSULTANCY SERVICES  
Experience certainty.

16

# Methodology



TATA CONSULTANCY SERVICES  
Experience certainty.

## Abundance of antibiotic resistance genes :

Total number of antibiotic resistance genes detected per million base pairs of the corresponding metagenome.

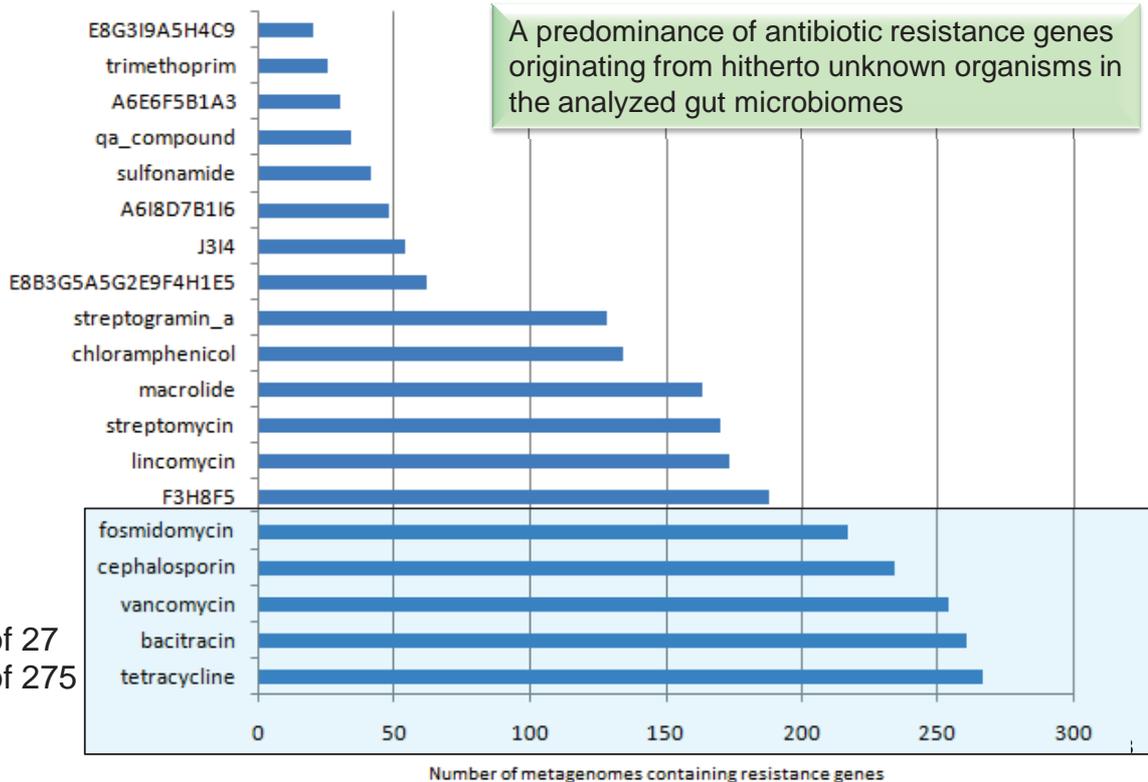
## Diversity of antibiotic resistance

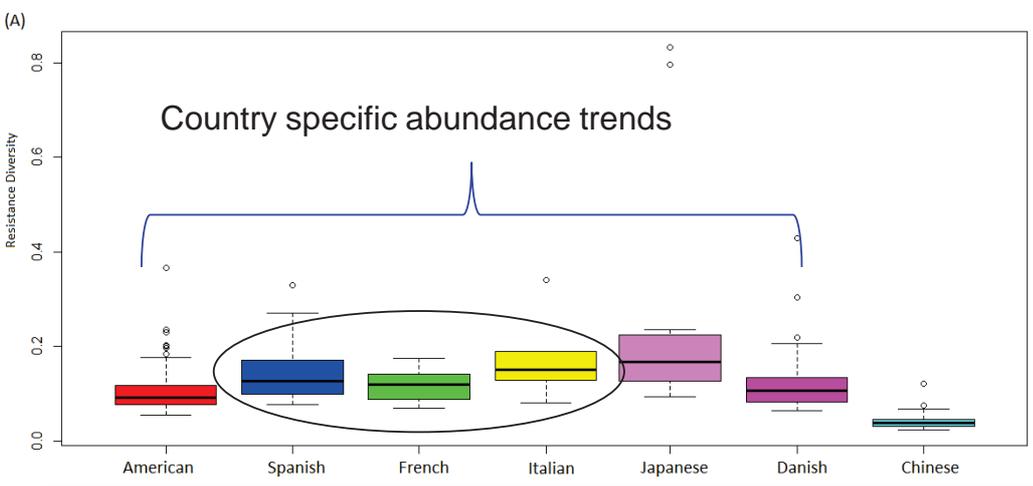
Number of antibiotics against which resistance genes were detected per million base pairs of sequence data.

17

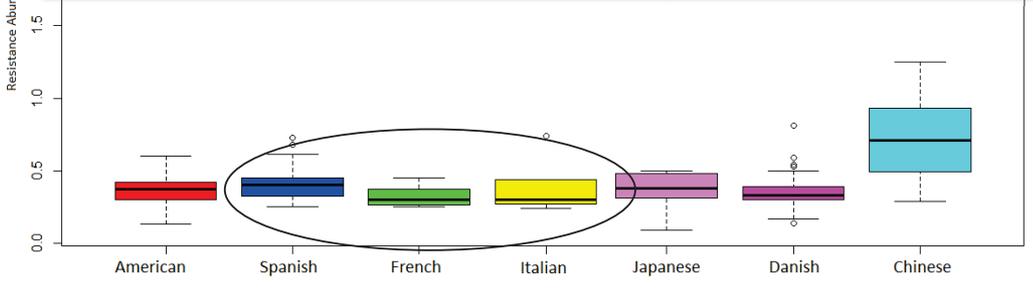
# Global trend of antibiotic resistance genes in gut microflora

Genes conferring resistance to as many as 53 antibiotics detected across the 275 gut metagenomes.

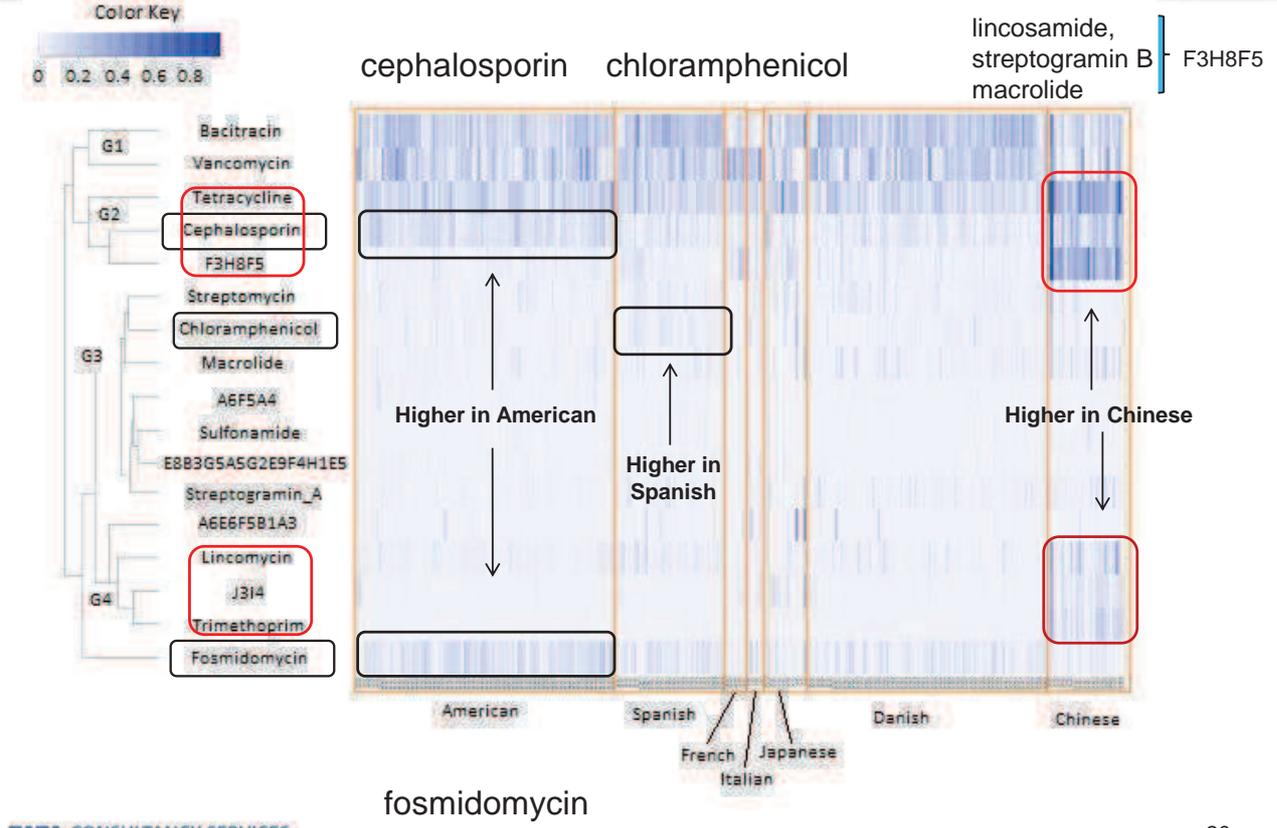




- (B)
- Gut microbiota of Southern European populations have resistance to a wider array of antibiotics as compared to the American and Danish individuals
  - Higher abundances but lower diversity of AR genes in Chinese
    - Chinese have higher number of genes conferring resistance to the same antibiotic
    - 35 AR genes (out of 157) found only in the Chinese gut metagenomes

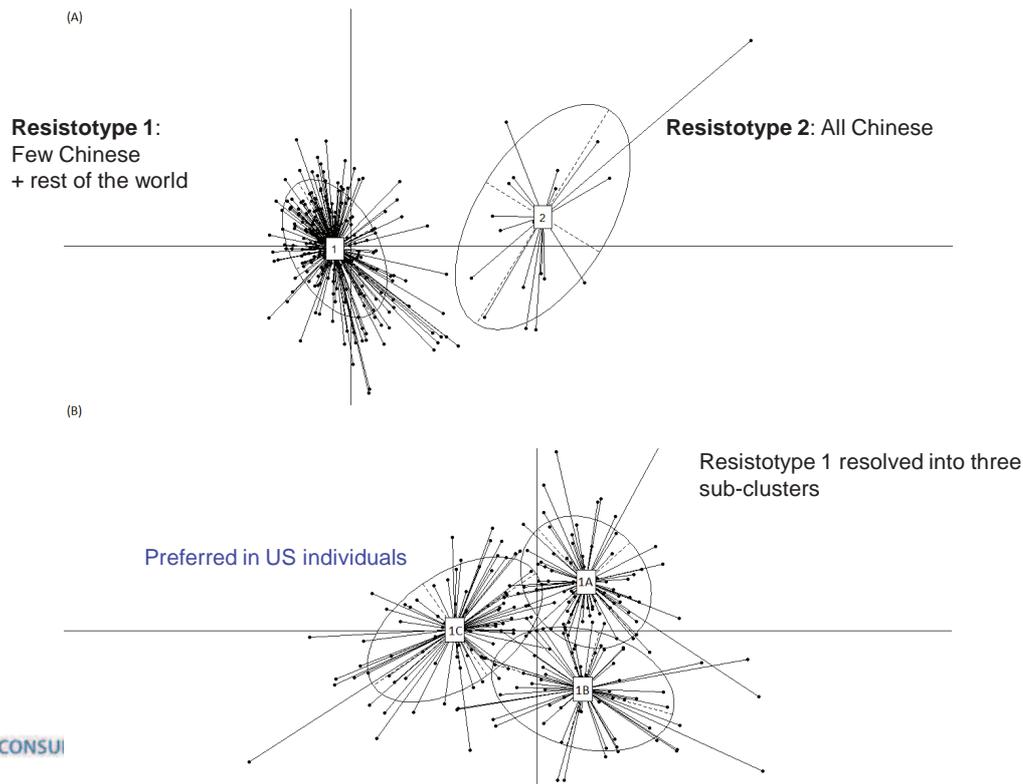


## Country specific resistance patterns



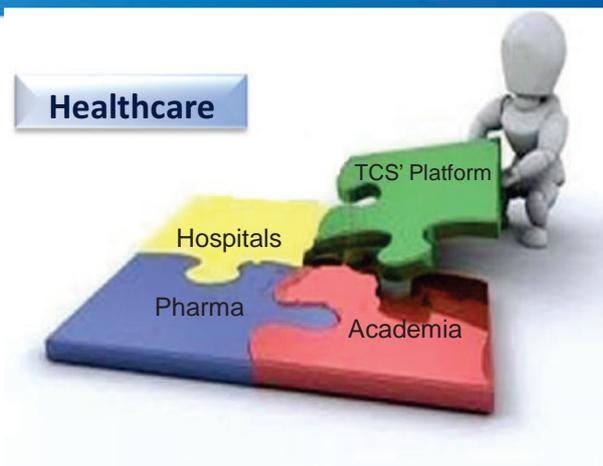
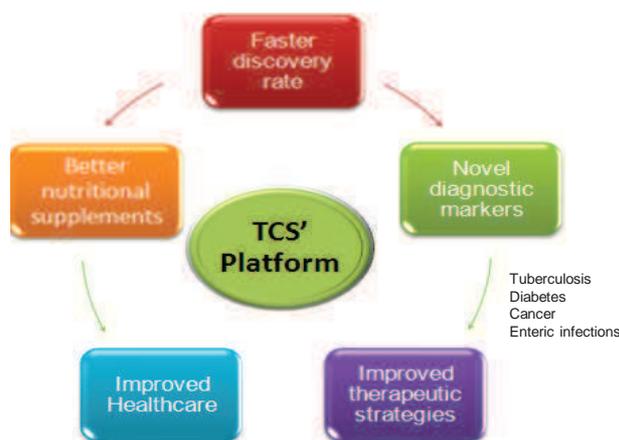
# Resistotypes: Individuals with similar resistance profiles

Individuals could be grouped into four clusters based on the **resistance profiles** in their gut microbiomes.



21

## Conclusions



- **Microbes and nutritional status**
  - Impaired nutritional status is not only due to the abundances of likely pathogenic microbial groups, but also a result of depletion of several commensal genera.
  - PC & NC functional groups
  - distinct changes in genera co-occurrence networks with nutritional status
- **'Resistotypes'**, exhibiting similarities in their antibiotic resistance profiles

# Acknowledgements

**TCS**



**Collaborator**

Dr. G.B. Nair  
(THSTI)

E-mail : [sharmila.mande@tcs.com](mailto:sharmila.mande@tcs.com)

<http://www.tcs.com/about/research/researchers/Pages/Sharmila-Mande.aspx>

## Thank you

E-mail : [sharmila.mande@tcs.com](mailto:sharmila.mande@tcs.com)

<http://www.tcs.com/about/research/researchers/Pages/Sharmila-Mande.aspx>

# NGS (Next Generation Sequencing) = Big data ?

Jean-Jacques Codani, CEO



BIOFACET SAS



# NGS = Big data ?

Jean-Jacques Codani, CEO

INAE/NATF Seminar  
GENOPOLE – Oct. 15-16 2014

## Biofacet

[Home](#) [Offering](#) [Technology](#) [Projects](#) [Management Team](#) [Contact](#)



We are the designers and developers  
of GenomeQuest Engine,  
a high-performance sequence comparison package,  
known as Biofacet® Engine.

[www.biofacet.com](http://www.biofacet.com)

## Big data

---

- ▶ Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- ▶ Big data has increased the demand of information management specialists in that [Software AG](#), [Oracle Corporation](#), [IBM](#), [FICO](#), [Microsoft](#), [SAP](#), [EMC](#), [HP](#) and [Dell](#) have spent more than \$15 billion on software firms specializing in data management and analytics. In 2010, this industry was worth more than \$100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.

Source: *wikipedia*

---

## NGS-Bioinformatics

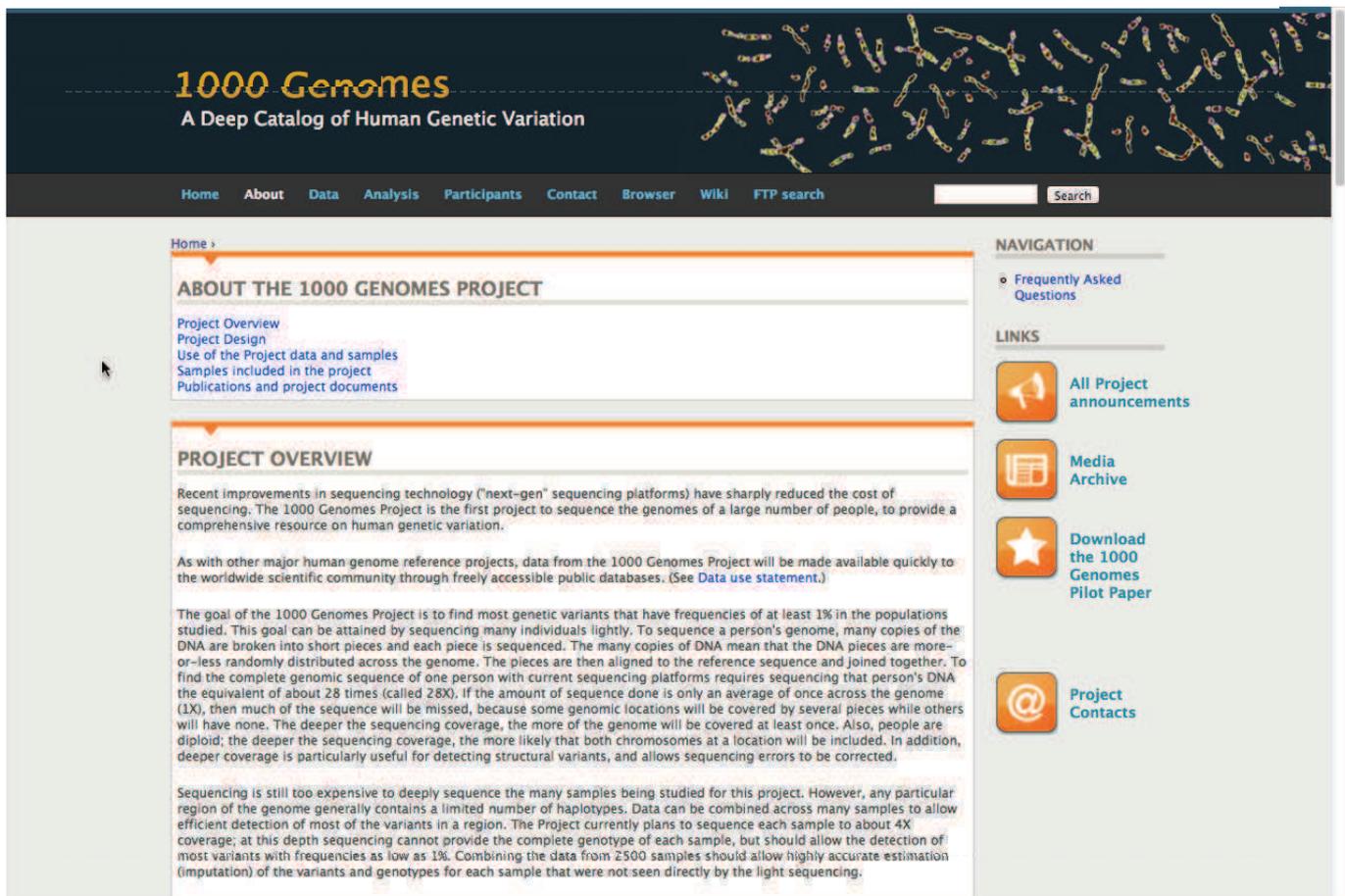
---

- ▶ Needs large to very large IT equipment:
  - ▶ Computing
  - ▶ Storage
  - ▶ Network
  - ▶ ...
- ▶ But, at the end of the day, it is NOT [Software AG](#), [Oracle Corporation](#), [IBM](#), [FICO](#), [Microsoft](#), [SAP](#), [EMC](#), [HP](#) nor [Dell](#) who will **implement** Life Sciences Bioinformatics solutions

# Plan

---

- ▶ 1000 Genomes project
    - ▶ What is-it
    - ▶ What's in there
    - ▶ For which use
  
  - ▶ Data acquisition: NGS
    - ▶ Reads, mapping, aligning, SNP-calling, annotating, etc.
  
  - ▶ Back to 1000 Genomes
    - ▶ Going further
- 



The screenshot shows the homepage of the 1000 Genomes Project website. The header features the project title "1000 Genomes" and the subtitle "A Deep Catalog of Human Genetic Variation" against a background of human chromosomes. A navigation menu includes links for Home, About, Data, Analysis, Participants, Contact, Browser, Wiki, and FTP search, along with a search bar. The main content area is divided into sections: "ABOUT THE 1000 GENOMES PROJECT" with sub-links for Project Overview, Project Design, Use of the Project data and samples, Samples included in the project, and Publications and project documents; "PROJECT OVERVIEW" with a detailed paragraph about the project's goals and methods; and "NAVIGATION" with links for Frequently Asked Questions, All Project announcements, Media Archive, Download the 1000 Genomes Pilot Paper, and Project Contacts.

1000 Genomes Samples							
Population	DNA sequenced from blood	Offspring Samples from Trios Available	Pilot Samples	Phase 1 Samples	Final Phase Discovery Sample	Final Release Sample	Total
Chinese Dai in Xishuangbanna, China (CDX)	no	yes	0	0	99	93	99
Han Chinese in Beijing, China (CHB)	no	no	91	97	103	103	106
Japanese in Tokyo, Japan (JPT)	no	no	94	89	104	104	105
Kinh in Ho Chi Minh City, Vietnam (KHV)	yes	yes	0	0	101	99	101
Southern Han Chinese, China (CHS)	no	yes	0	100	108	105	112
<b>Total East Asian Ancestry (EAS)</b>			<b>185</b>	<b>286</b>	<b>515</b>	<b>504</b>	<b>523</b>
Bengali in Bangladesh (BEB)	no	yes	0	0	86	86	86
Gujarati Indian in Houston, TX (GIH)	no	yes	0	0	106	103	106
Indian Telugu in the UK (ITU)	yes	yes	0	0	103	102	103
Punjabi in Lahore, Pakistan (PJL)	yes	yes	0	0	96	96	96
Sri Lankan Tamil in the UK (STU)	yes	yes	0	0	103	102	103
<b>Total South Asian Ancestry (SAS)</b>			<b>0</b>	<b>0</b>	<b>494</b>	<b>489</b>	<b>494</b>
African Ancestry in Southwest US (ASW)	no	yes	0	61	66	61	66
African Caribbean in Barbados (ACB)	yes	yes	0	0	96	96	96
Esan in Nigeria (ESN)	no	yes	0	0	99	99	99
Gambian in Western Division, The Gambia (GWD)	no	yes	0	0	113	113	113
Luhya in Webuye, Kenya (LWK)	no	yes	102	97	101	99	116
Mende in Sierra Leone (MSL)	no	yes	0	0	85	85	85
Yoruba in Ibadan, Nigeria (YRI)	no	yes	106	88	109	108	116
<b>Total African Ancestry (AFR)</b>			<b>208</b>	<b>246</b>	<b>669</b>	<b>661</b>	<b>691</b>
British in England and Scotland (GBR)	no	yes	0	89	92	91	94
Finnish in Finland (FIN)	no	no	0	93	99	99	100
Iberian populations in Spain (IBS)	no	yes	0	14	107	107	107
Toscani in Italy (TSI)	no	no	66	98	108	107	110
Utah residents with Northern and Western European ancestry (CEU)	no	yes	94	85	99	99	103
<b>Total European Ancestry (EUR)</b>			<b>160</b>	<b>379</b>	<b>505</b>	<b>503</b>	<b>514</b>
Colombian in Medellin, Colombia (CLM)	no	yes	0	60	94	94	95
Mexican Ancestry in Los Angeles, California (MXL)	no	yes	0	66	67	64	69
Peruvian in Lima, Peru (PEL)	yes	yes	0	0	86	85	86
Puerto Rican in Puerto Rico (PUR)	yes	yes	0	55	105	104	105
<b>Total Americas Ancestry (AMR)</b>			<b>0</b>	<b>181</b>	<b>352</b>	<b>347</b>	<b>355</b>
<b>Total</b>			<b>553</b>	<b>1092</b>	<b>2535</b>	<b>2504</b>	<b>2577</b>

1  
2  
3  
4  
5

Today:

2504 individuals divided into 5 sub-populations

Note: Some samples from this pilot or phase 1 did not progress to the final phase for reasons of quality control or changing criteria with respect to complete samples. This causes some discrepancies in the total numbers of samples compared to the numbers in the final phase.



# What's in the database ?

Samples have IDs

Samples for the 1000 Genomes Project (available and used)									
Population	Coriell sample ID	SRA individual sample accession number	Family	Sex	Relationship	Type (unrel duo trio)	Whole-genome for full project		Exomes
							Center	Platform	Center
GBR	HG00096	SRS006837		male		unrel	WUGSC		
GBR	HG00097	SRS006838		female		unrel	BCM		
GBR	HG00098	SRS006839		male		unrel	SC		
GBR	HG00099	SRS006840		female		unrel	BCM		
GBR	HG00100	SRS006841		female		unrel	SC		
GBR	HG00101	SRS006842		male		unrel	MPIMG		
GBR	HG00102	SRS006843		female		unrel	MPIMG		
GBR	HG00103	SRS006844		male		unrel	WUGSC		
GBR	HG00104	SRS006845		female		unrel	BCM		
GBR	HG00105	SRS006846		male		unrel	MPIMG		
GBR	HG00106	SRS006847		female		unrel	SC		
GBR	HG00107	SRS006848		male		unrel	MPIMG		
GBR	HG00108	SRS006849		male		unrel	BGI		
GBR	HG00109	SRS006850		male		unrel	BGI		
GBR	HG00110	SRS006851		female		unrel	BGI		

## Data: 1000G : **vcf** format : chromosome 9

Reference (0|0)

Genotype value (**GT**) per sample

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00097	HG00099	HG00100	HG00101
9	10163	.	CT	C	100	PASS	AC=15;AF=0.00299521;AN=5008;NS=2504						
0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
0 0	0 0	0 0	0 0	1 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
0/0	0/0	0/0	0/0	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
0 0	0 0	0 1	0 0	0 0	0 0	0 0	0 0	1 1	. 0	. 0	0 0	. . .	0 0

↑ Not phased     
 ↑ Heterozygotes     
 ↑ Homozygotes     
 ↑ Not called

At genomic position chr9:10163, # of a given Human Genome version and for the following 2054 samples, # HG1234, HG2345, ... NA12234, ... the variation : CT->C # a deletion of T in the reference has been observed

- on a single allele
- on both alleles
- (or not observed : sample = reference)

## Typical simple queries

- ▶ Give me variants specific to individual HG00099
- ▶ Give me all variants:
  - ▶ Present in AFR sub-pop
  - ▶ Absent in EUR sub-pop

As many genotype values as many samples x Total number of variants  
 2504 samples x 81 Millions SNPs  
 Equivalent to a RDMBS of 81M lines, 2504 columns = 200 Billions values

This starts to be Big data ... ? More Later ...

## Now ...

---

- ▶ Individuals are cohorts from disease studies
  - ▶ Comparison with known clinical mutation databases
  - ▶ Variants present in ill's, absent in healthy's
  - ▶ Homozygote variants in study A, absent in B, and many more...
- 

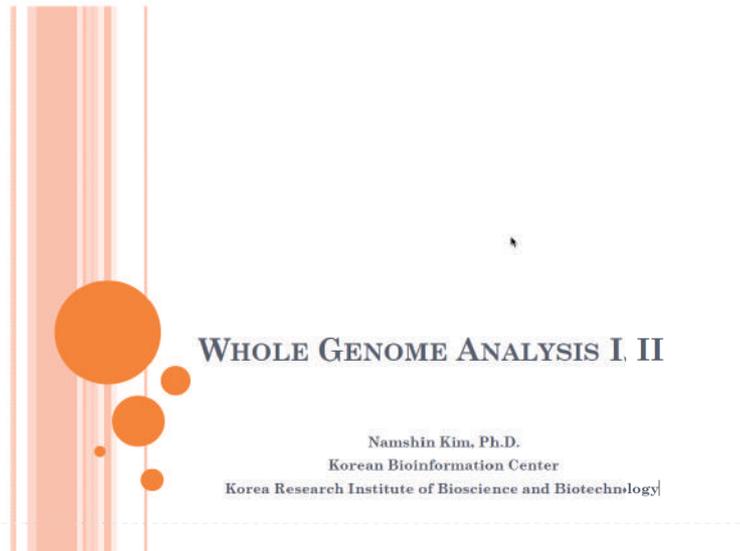
## How data is acquired?

---

- ▶ NGS
    - ▶ Reads (many, so far short)
    - ▶ Cleaning phases (e.g. extracting adaptors in small-RNAs)
    - ▶ Mapping/Aligning to genome
    - ▶ SNP-calling
    - ▶ SNP-annotating
    - ▶ ...
-

# A World ...

- ▶ Sequencing techniques: Single-End, Paired-End, ...
- ▶ Mappers : Gaps, Local/Global, ...
- ▶ Formats: FASTQ, BAM, BED, ...
- ▶ SNP-callers
- ▶ Viewers
- ▶ ...




## ABOUT THE 1000 GENOMES PROJECT

- Project Overview
- Project Design
- Use of the Project data and samples
- Samples included in the project
- Publications and project documents

### PROJECT OVERVIEW

Recent improvements in sequencing technology ("next-gen" sequencing platforms) have sharply reduced the cost of sequencing. The 1000 Genomes Project is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation.

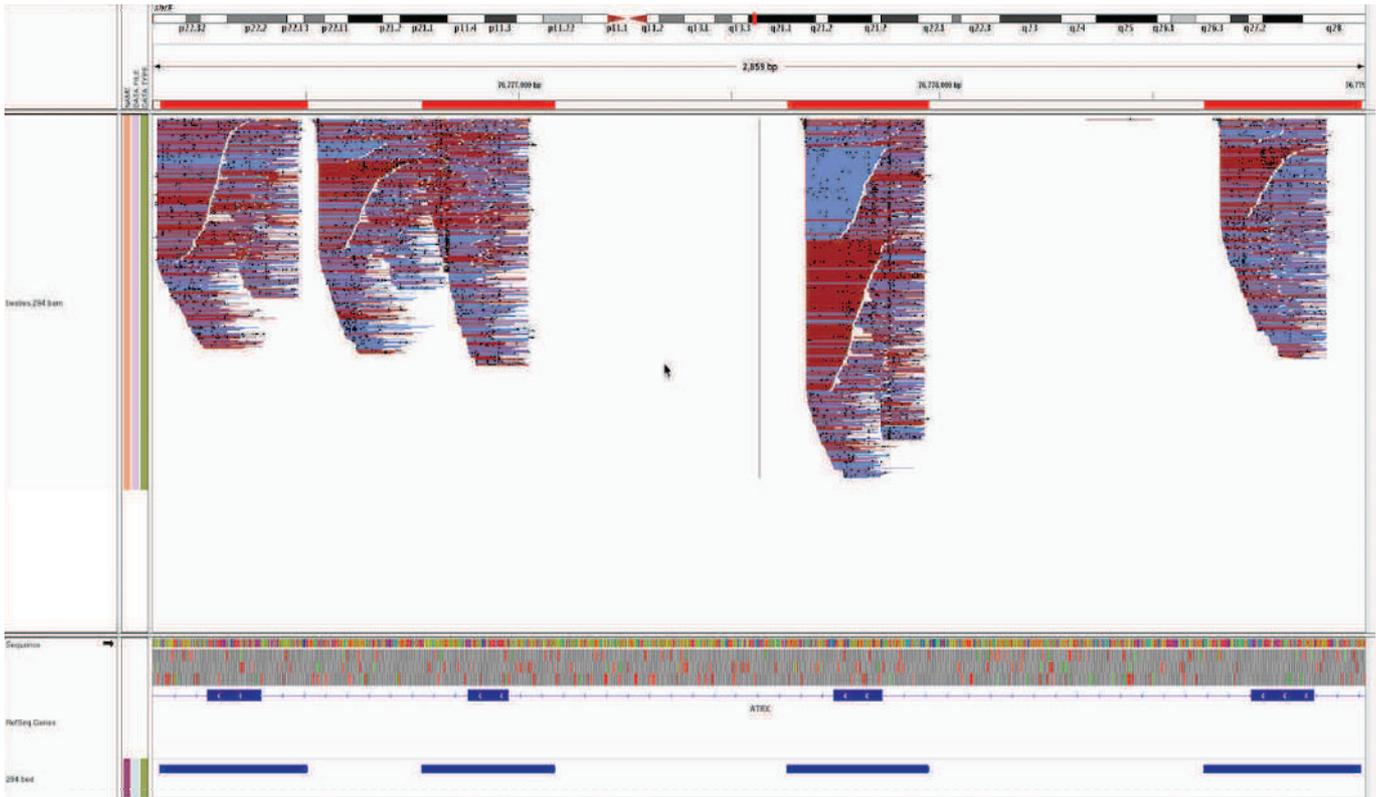
As with other major human genome reference projects, data from the 1000 Genomes Project will be made available quickly to the worldwide scientific community through freely accessible public databases. (See [Data use statement](#).)

The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied. This goal can be attained by sequencing many individuals lightly. To sequence a person's genome, many copies of the DNA are broken into short pieces and each piece is sequenced. The many copies of DNA mean that the DNA pieces are more-or-less randomly distributed across the genome. The pieces are then aligned to the reference sequence and joined together. To find the complete genomic sequence of one person with current sequencing platforms requires sequencing that person's DNA the equivalent of about 28 times (called 28X). If the amount of sequence done is only an average of once across the genome (1X), then much of the sequence will be missed, because some genomic locations will be covered by several pieces while others will have none. The deeper the sequencing coverage, the more of the genome will be covered at least once. Also, people are diploid; the deeper the sequencing coverage, the more likely that both chromosomes at a location will be included. In addition, deeper coverage is particularly useful for detecting structural variants, and allows sequencing errors to be corrected.

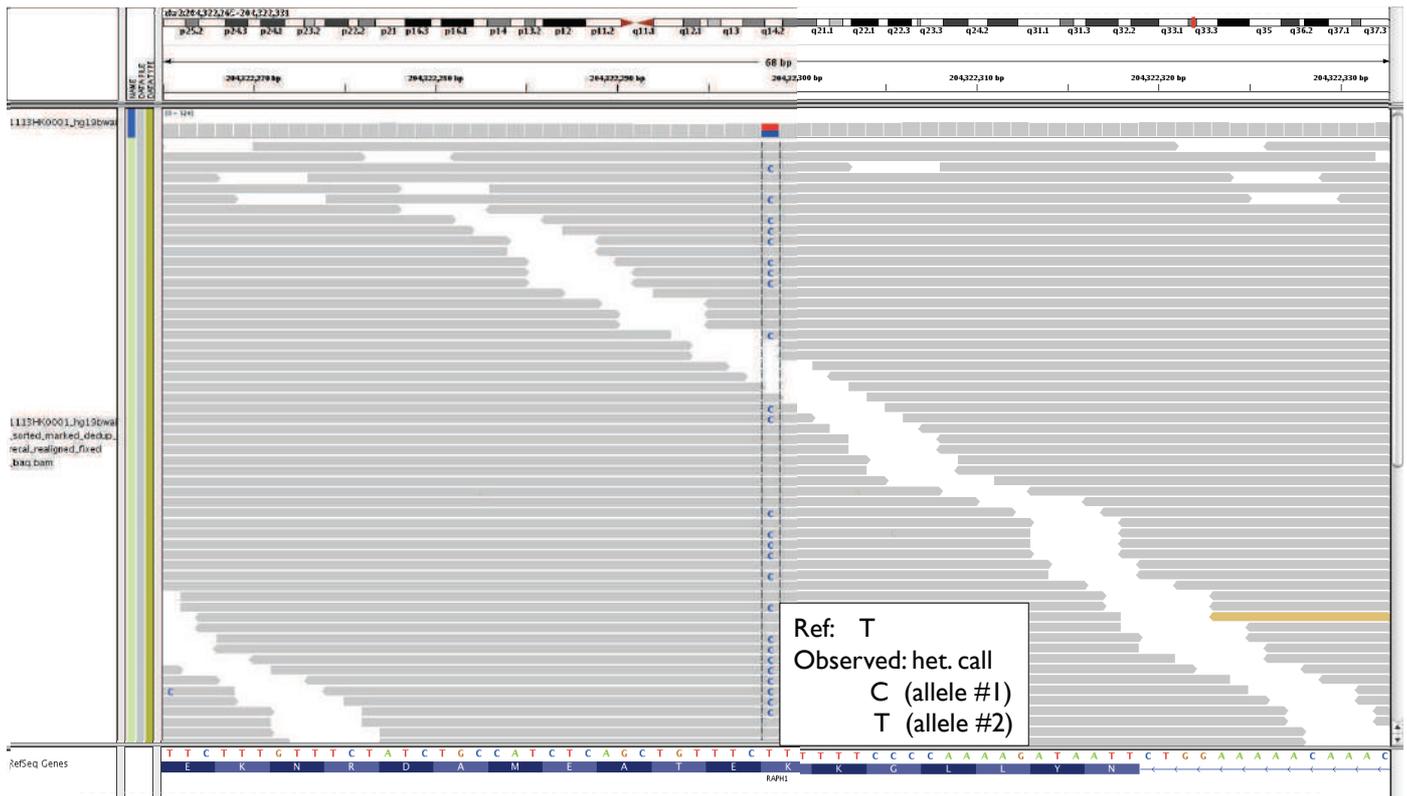
Sequencing is still too expensive to deeply sequence the many samples being studied for this project. However, any particular region of the genome generally contains a limited number of haplotypes. Data can be combined across many samples to allow efficient detection of most of the variants in a region. The Project currently plans to sequence each sample to about 4X coverage; at this depth sequencing cannot provide the complete genotype of each sample, but should allow the detection of most variants with frequencies as low as 1%. Combining the data from 2500 samples should allow highly accurate estimation (imputation) of the variants and genotypes for each sample that were not seen directly by the light sequencing.

# Exome capture @ 200x coverage

Average size of a read: 100-150nt



## A Variant mismatch



## Aligners use **filters** to **speed-up** sequence alignments

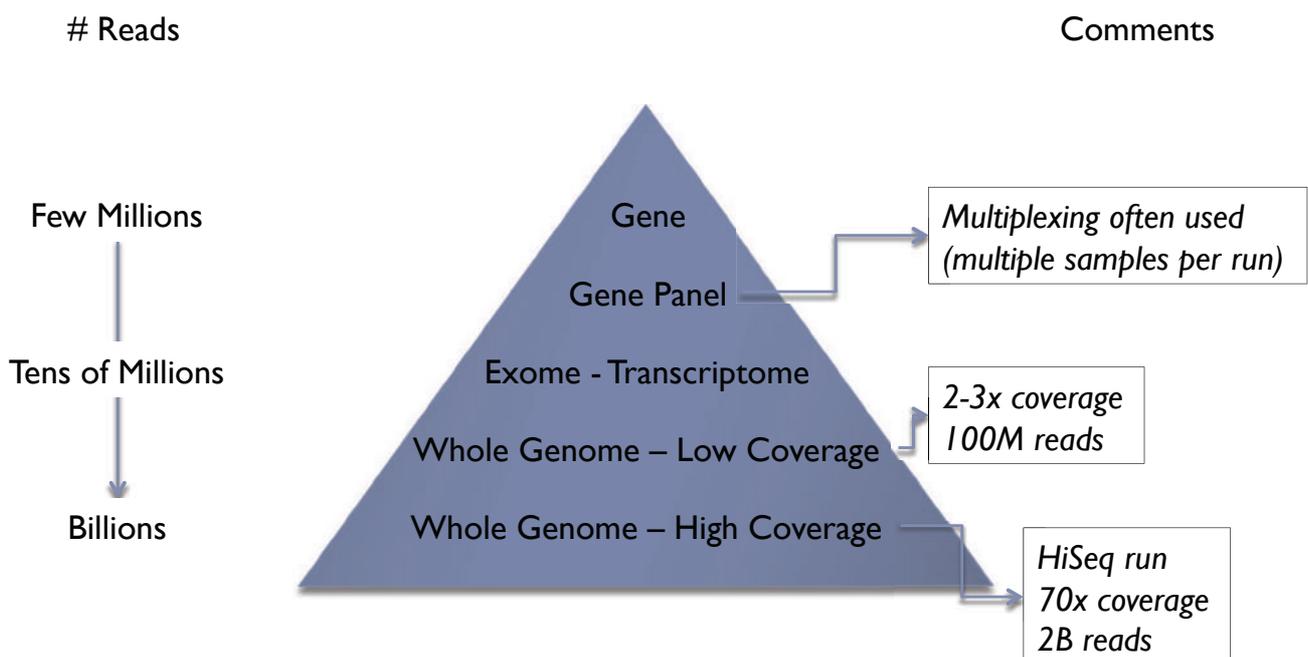
Example :

- count the # of words T, in common between a read and a region of the reference
- if  $T \geq T_0$ , align



T= 10 IS GOOD => PROCEED WITH ALIGNMENT / PILEUP / SNP-CALLER

## Big data indeed ...



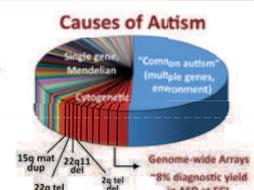
# Data & Computing

- ▶ 2B reads @ 100nt  
-> 1/2 TB of raw data (nucleotides + qualities)
- ▶ Generating the same amount of mapped data
- ▶ Compression is used, but:
  - ▶ Multiplicity of algorithms for different phases generate multiple copies.
  - ▶ For large projects, goes to PetaBytes.  
Pretty strong infrastructure is needed.
  - ▶ Becomes heavy ... such a way it is sometimes said:  
*"let's go back to the fridge and re-sequence the thing"*
- ▶ Computing: +- OK with large clusters, but ...
  - ▶ cpu-speed is NOT the issue – I/O is the issue (a lot of flat files)
  - ▶ Assuming standard (easy) mapping

Inherited Metabolic Disorders	101
Lysosomal Storage Disorders	55
Macrocephaly	11
Maturity Onset Diabetes of the Young	4
Mitochondrial Diseases	44
Multiple Epiphyseal Dysplasia	7
Neonatal and Adult Cholestasis	57
Neurological Disorders	164
Neuromuscular Disorders- Expanded	79
- Congenital Muscular Dystrophy	24
- Limb-Girdle Muscular Dystrophy	22
- Neuromuscular Disorders	46
Noonan Syndrome and Related Disorders	13
Peroxisome Biogenesis Disorders, Zellweger Syndrome Spectrum	14
Pulmonary Disease	53
- Bronchiectasis	16
- Congenital Central Hypoventilation Syndrome	7
- Cystic Lung Disease	8
- Pulmonary Fibrosis and Hermansky-Pidlak Syndrome	17
Severe Combined Immunodeficiency (SCID) B+/B-	21
- Severe Combined Immunodeficiency (SCID) B+	9
- Severe Combined Immunodeficiency (SCID) B-	7
Short Stature Panel	45
Skeletal Dysplasia	163
- Disproportionate Short Stature	77
- Limb Malformation	46
- Osteogenesis Imperfecta and Decreased Bone Density	34
- Skeletal Dysplasia With Increased Bone Density	22
Targeted Tumor Mutation	26
- Targeted Colorectal Tumor Mutation	13
- Targeted Gastric Tumor Mutation	7
- Targeted Lung Tumor Mutation	12
- Targeted Melanoma Mutation	8
- Targeted Ovarian Tumor Mutation	7
Tuberous Sclerosis	2
X-linked Intellectual Disability	91

## Autism Spectrum Disorders Panel: 60 Genes

Autism spectrum disorders (ASDs) are a group of neurodevelopmental disorders which include autism, pervasive developmental delay-not otherwise specified (PDD-NOS), and Asperger syndrome. ASDs are characterized by impairments in social relationships, variable degrees of language and communication deficits, and repetitive behaviors and/or a narrow range of interests. The age of onset is prior to age 3 with a variable clinical presentation, ranging in severity both amongst individuals as well as amongst the various subtypes of ASDs. Additional clinical features may also be observed in individuals with an ASD, such as intellectual disability (up to ~50%) and seizures (~25%).



Client  Employee  
  [Login](#)  
[Create an account](#) | [Forgot password?](#)

Your partner in genetic healthcare

[Home](#) [Test Search](#) [Test Ordering](#) [Billing](#) [Client Services](#) [About EGL](#) [Resources](#) [Contact Us](#)

**X-linked Intellectual Disability: Gene Sequencing and Deletion/Duplication Panel**

Overview
How to Order Test
Specimen Requirements
Test Description
Related Tests
Special Instructions

[Full Page View](#) | [PDF View](#)

**Name**  
X-linked Intellectual Disability: Gene Sequencing and Deletion/Duplication Panel

**Synonyms**  
XLID, X-Linked Mental Retardation, XLMR

**Test Code**  
MXLI1

**Indication**  
This test is indicated for:  

- Individuals with a clinical and family history consistent with an X-linked intellectual disability disorder after fragile X testing and genomic array testing are normal.
- Carrier testing in adult females with a family history of X-linked intellectual disability.

**CPT Codes**  
81228 (x1), 81404 (x1), 81405 (x1), 81406 (x1)

**Turn around time**  
12 weeks

[Back](#)

for iOS and Android

© Copyright 2000-2013 Emory University
[Careers](#) | [Contact Us](#) | [Site Map](#) | [Directions](#)

**X-linked Intellectual Disability: Gene Sequencing and Deletion/Duplication Panel**

Overview
How to Order Test
Specimen Requirements
Test Description
Related Tests
Special Instructions

Intellectual disability (ID) is a nonprogressive cognitive impairment affecting 1-3% of the Western population. It is estimated that up to 50% of moderate-severe cases have genetic causes and approximately 10% are due to X-linked intellectual disability disorders (XLID). XLID can be syndromic or nonsyndromic and is observed in all ethnic groups. More than 100 XLID syndromes have been described in the literature to date. Fragile X is the most common XLID syndrome (~1 in 4000 males) while others can be quite rare with only a few patients reported in the literature. Males can have moderate to severe intellectual disability depending on a syndrome, and carrier females can also be affected, but typically have milder clinical symptoms.

A majority of individuals with XLID are non-syndromic with no other features to assist in diagnosis. Because of the number of genes involved, it is very difficult to identify which X-linked gene may be responsible for the phenotype in any given patient. Simultaneous testing of all known XLID genes in a single study provides a significant diagnostic advantage over single gene sequencing. Additional benefits for the patient and families include:

- Providing information for recurrence risk and family planning and allowing for presymptomatic support
- Helping physicians determine appropriate follow-up testing and develop a health maintenance plan
- Predicting better patient prognostic value
- Assisting researchers in the understanding of the molecular basis of disease in the hope for treatments and cures
- Assessing the possibility of therapy for some forms of XLID

The XLID next generation gene sequencing panel contains 92 genes on the X chromosome implicated in XLID.

Testing for fragile X syndrome and genomic array CGH testing are recommended as first steps for individuals who may have XLID. If those tests results are normal, XLID gene sequencing panels can be ordered.

Testing includes trinucleotide repeat analysis for the *FMR1* and *AFF2/FMR2* genes.

**Methodology**

**Next Generation Sequencing:** In solution hybridization of all coding exons contained in the genes of the X-linked Intellectual Disability Panel is performed on the patient's genomic DNA. Direct sequencing of the amplified captured regions is performed using next generation sequencing. The patient's gene sequences are then compared to a standard reference sequence. Potentially causative variants and areas of low coverage are Sanger sequenced in order to confirm variants and ensure 100% coverage of the targeted exons. Sequence variations are classified as pathogenic variants, benign variants unrelated to disease, or variants of unknown clinical significance. Variants of unknown clinical significance may require further studies of the patient and/or family members. This assay does not interrogate the promoter region, deep intronic regions, or other regulatory elements, and does not detect single or multi-exon deletions or duplications.

**Deletion/Duplication Analysis:** DNA isolated from peripheral blood is hybridized to a gene-targeted CGH array to detect deletions and duplications. The targeted CGH array has overlapping probes that cover the entire genomic region.

Please note that a "backbone" of probes across the entire genome are included on the array for analytical and quality control purposes. Rarely, off-target copy number variants causative of disease may be identified that may or may not be related to the patient's phenotype. Only known pathogenic off-target copy number variants will be reported. Off-target copy number variants of unknown clinical significance will not be reported.

**Detection**

**Next Generation Sequencing:** Clinical Sensitivity: Unknown. Mutations in the promoter region, some mutations in the introns and other regulatory element mutations cannot be detected by this analysis. Large deletions/duplications will not be detected by this analysis. Results of molecular analysis should be interpreted in the context of the patient's clinical/biochemical phenotype.

Analytical Sensitivity: ~99%

**Deletion/Duplication Analysis:** Detection is limited to duplications and deletions. The CGH array will not detect point or intronic mutations. Results of molecular analysis must be interpreted in the context of the patient's clinical and/or biochemical phenotype.

## Genome versus regions

---

- ▶ Gene panels are built by sequencing targeted regions (typically exons of the genes).

- ▶ Simple question: should I map the reads

- ▶ Against the genome?

Simpler: most mappers use complete genomes as input (because of prebuilt hash tables of words/seeds).

That case, a read belonging to the region can have a better mapping elsewhere on the genome => will this read be discarded?

- ▶ Against the regions?

Inverse effect (read can be a false positive), but guaranteed to not be missed.

Other tricks ... let's have a quick look

---

23

## Insertions/deletions

---

81bp **deletion**

```
MLL2:c2428_2508delACTGAGGAGCCGCACCTGTCTCCTGTGCCTGAGGAGCCA  
TGCTTGTCCCCCAACCTGAGGAATCACACCTGTCCCCCAG
```

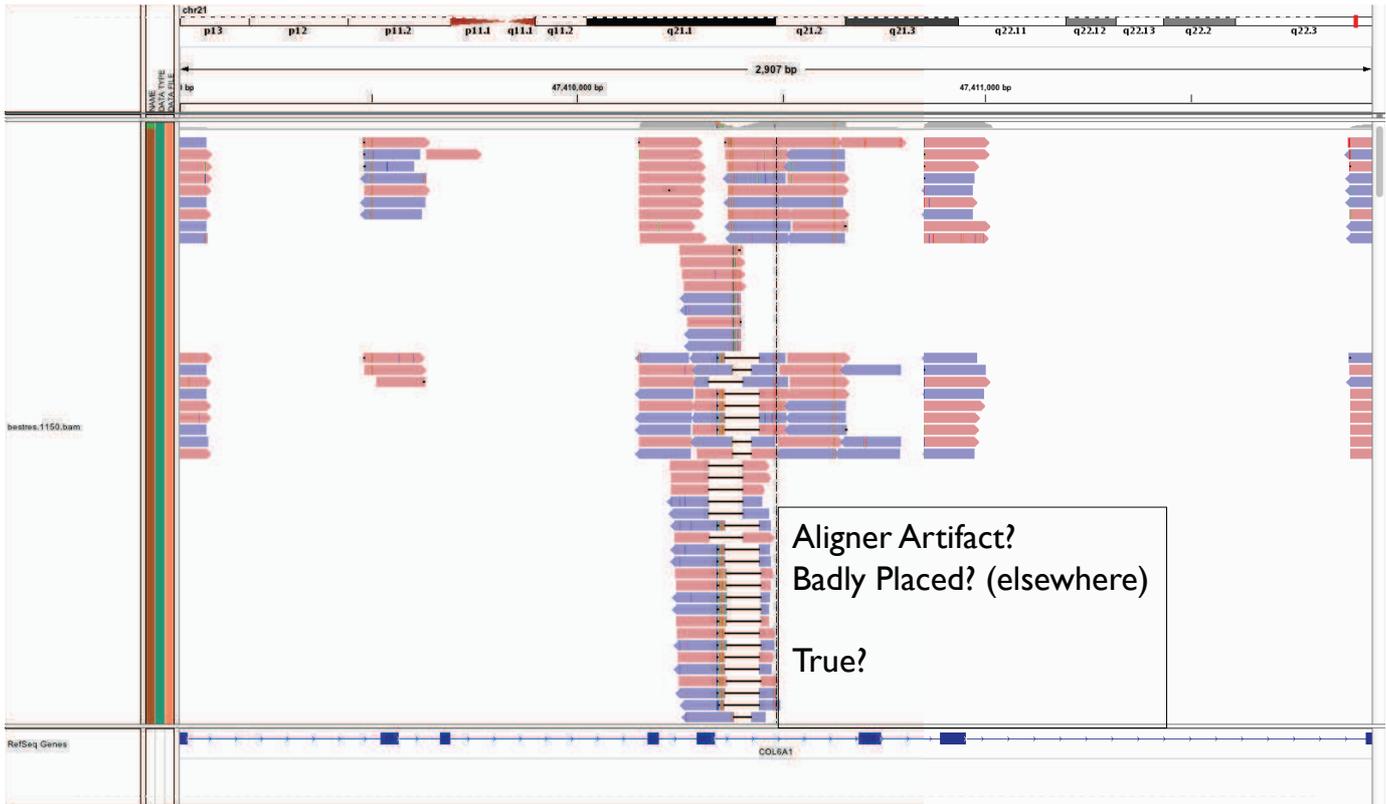
132bp **insertion**

```
NM_000203.3:c793-14_793-13insTGGACTACATCTCCCTCCACAGGAAGGTGCGC  
CCTGCCCTCCGTCCGCCCGGTGTTCTGCGCCCTCAGCCGCTGTGCCCCG  
GGCCGCGCTGACCCTGGTGGTGCTGA
```

Deletions detection can be done, but the mapper has to be guided to do so.  
Same thing (but more limited due to read sizes) for insertions.

**YOU HAVE TO KNOW  
WHAT YOU ARE LOOKING FOR**

---



Samtools manual

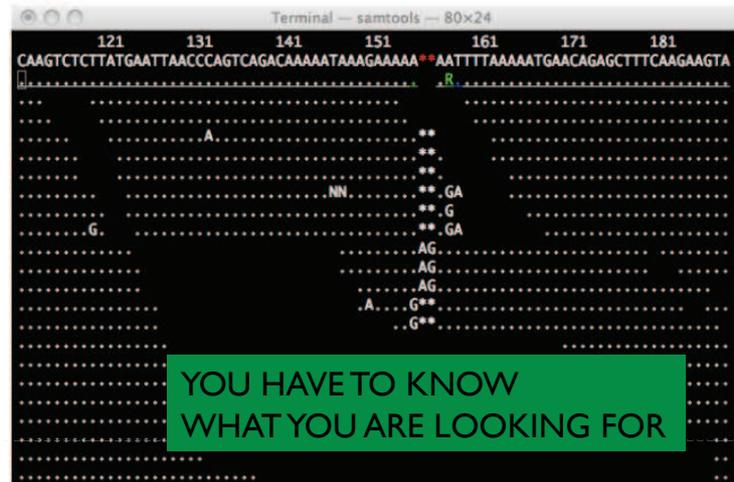
Tandem repeats  
Homopolymers

...

Short indels tend to occur around tandem repeats, but the alignment is much harder in these regions given short reads. Reads aligned without gaps may actually contain indels due to wrong alignment. The pileup command fixes this. Here is an example of a 2bp insertion to the reference:

```
seq2 151 G G 36 0 99 12 .....A :9<;7<<<<<
seq2 152 A A 63 0 99 12 ..... :9<;<;<<<<<
seq2 153 A A 63 0 99 12 ..... :7<877=<<<<<
seq2 154 A A 66 0 99 13 .$.....^~. :7<97<7<<<<<<
seq2 155 A A 63 0 99 12 .$..... 7<77<;<<<<<<
seq2 156 A A 10 0 99 11 .$.....+2AG.+2AG.+2AGGG <975;:<<<<<
seq2 156 * +AG/+AG 71 252 99 11 +AG * 3 8 0
seq2 157 A A 57 0 99 10 .$.$..... 97<<<<<<<<
seq2 158 A R 18 18 99 8 GG$G..... <;<<<<<
seq2 159 T T 8 0 99 7 A$A$..... 3:<<<<<<
```

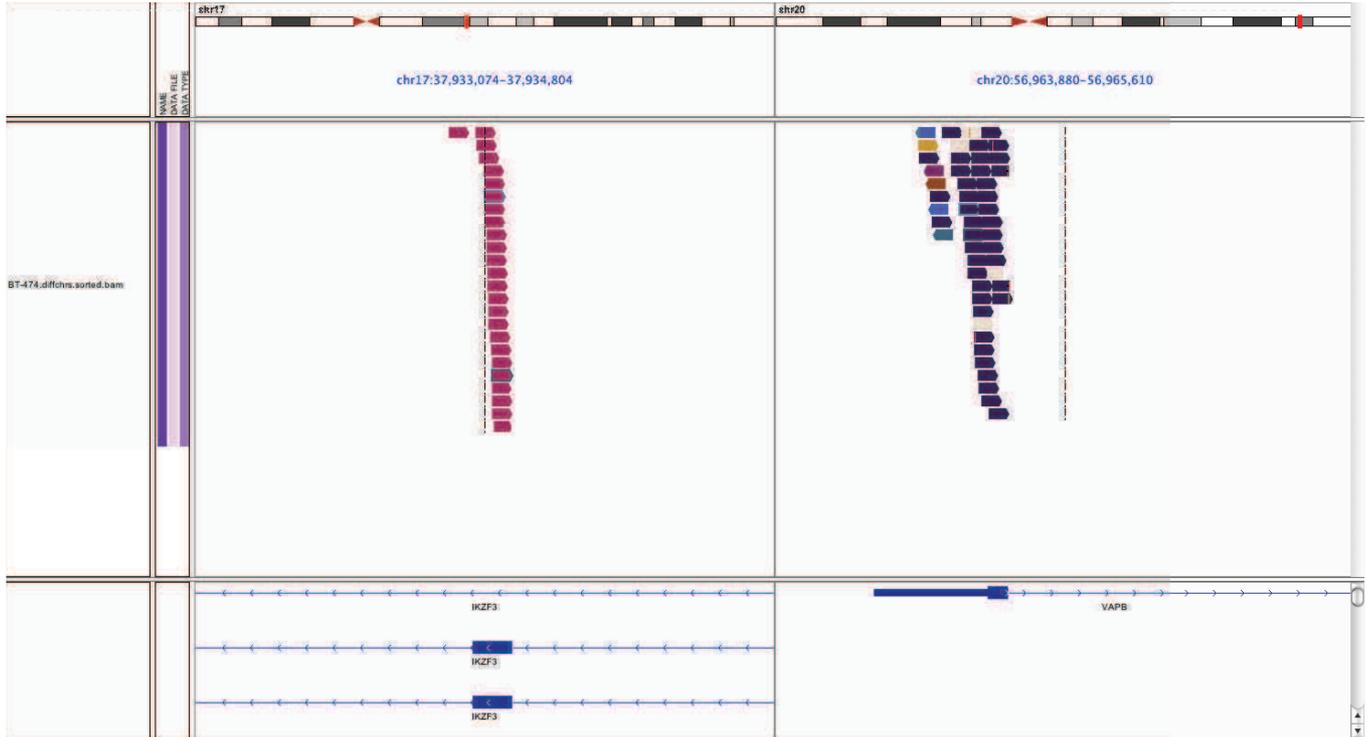
The line with the 3rd column a star indicates that the AG insertion is supported by 3 reads; 8 reads agree with the reference according to the raw alignment; no reads support a third allele. However, SAMtools infers a AC homozygous insertion with a high score 252 because when we realign the reads with the prior of an insertion, we found that the 8 reads mapped without gaps are due to a tandem repeat. This scenario is clearer from the alignment viewer:



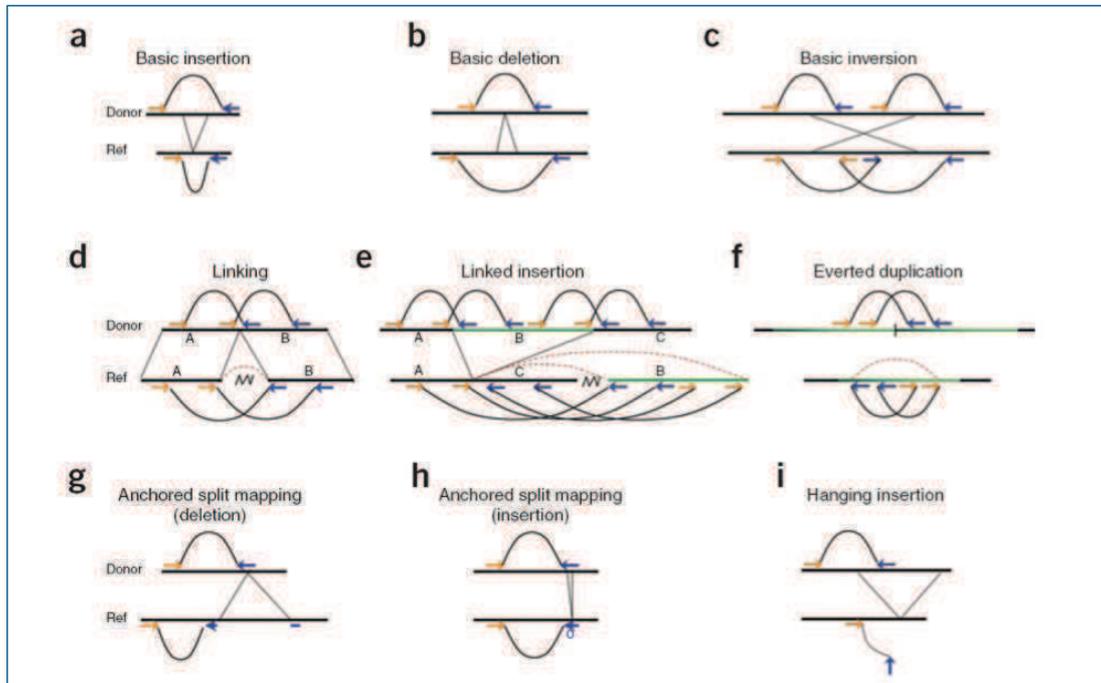
Here is another example of a 5/5 heterozygous insertion (In this case, 3 reads are aligned with gaps, but 13 reads show the evidence of the insertion):

## A Gene Fusion event

Mates from PE reads go to two different genomic locations

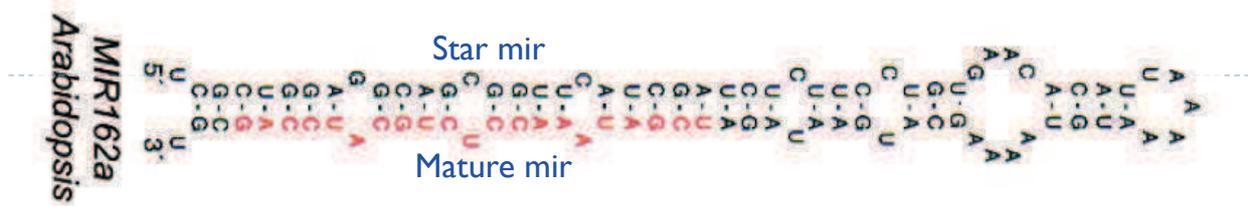


## Structural Variations



DIFFERENT MAPPING ALGORITHMS

YOU HAVE TO KNOW WHAT YOU ARE LOOKING FOR



```
== B group 70/2266 Chr01 results = 10 PIVOT=38190570
   fw_nbr=4 rv_nbr=1 zero_left=3 zero_right=2 obl=19095229 oel=19095252 obr=19095316 oer=19095340
```

```
** miRNA confirmed candidates **
1 L fw d=5 len=23      19095229      ..t..t..tc.a.C..... {118716, 118724}
1 R fw d=1 len=25      19095316      .....-g.ct..t..g.. {219648}
M & M* weak match M*/M=16.6667%
```

```
Chr01  5319      5340      6:1102:1606:43796:Y (rv)      1      21      Errs= 6
Q:      1  CCGCTAAA-AGCCTGTTTTGGT  21
      | | | | | | | | | | | | | | |
S:  5319  CCGCTAAAGATCGAGTATTAGT  19095340
```

DIFFERENT MAPPING ALGORITHMICS

YOU HAVE TO KNOW  
WHAT YOU ARE LOOKING FOR

## Biofacet Enterprise Release MAPPER

Biofacet NGS reads mapper is based on GASSST, originally published in [Bioinformatics Paper](#). Biofacet implementation is an improved version, including the following key improvements:

- Features

- ✓ Any choice of gapped/ungapped, global/local alignments (anchored, not anchored)
  - ✓ Adaptative pattern of errors, function of read lengths
  - ✓ Small insertions and large deletions
  - ✓ On the fly alignment trimming allowing clean results
  - ✓ Sensitivity/speed levels adjustment
  - ✓ Single-End and Paired-End modes
  - ✓ Read lengths up to 20K
  - ✓ SOLiD processed in native colorspace mode, DNA alignments produced
  - ✓ Integrated alignment sorting module allowing fine tuning of best results kept
  - ✓ And many more ...
- Biofacet is a proven, stable, industrial product.  
Competitive to Bowtie2, BWA, and other Open Source mappers.

## Conclusion

---

- In average, NGS sequencing and analysis “work”
  - Diagnostics of known *SNP-like* mutations: +- OK
  - Population Profiling, \$600M Clinical Trials
    - Accuracy matters
    - Choice of algorithms is critical
    - The whole technology chain MUST be controlled
    - Computing muscles comes AFTER
    - Different from Big Data
- 

## Now ... Back to 1000G

---

- ▶ Individuals are cohorts from disease studies
  - ▶ Comparison with clinical mutation databases
  - ▶ Variants present in ill's, absent in healthy's
  - ▶ Homozygote variants in study A, absent in B, and many more...
-

Mutation Details for c.1521\_1523delCTT

HGVS →

cDNA Name	c.1521_1523delCTT
Protein Name	p.Phe508del
Exon or Intron	exon 11
Legacy Exon or Intron	exon 10
Legacy Name	[delta]F508
Other Details	This is the major CF mutation; it accounts for ~70% of CF chromosomes in most Caucasian populations. For additional information please see: Rommens et al. Science 245: 1059-1065, 1989; Riordan et al. Science 245: 1066-1073, 1989; Kerem et al. Science 245: 1073-1080, 1989; The Cystic Fibrosis Genetic Analysis Consortium, Am. J. Hum. Genet. 47: 354-359, 1990; and Welsh et al. In Metabolic and Molecular Basis of Inherited Disease (7th Edition), C Scriver, AL Beaudet, WE Sly, D Valle, eds., McGraw-Hill, Chapter 127, pp. 3799-3876, 1995
Contributors	Tsui, LC Collins, FS Riordan, JR et al. 1989-08-24
Institute	Hospital for Sick Children, Toronto, Canada; University of Michigan, Ann Arbor, USA
Phenotype Information	<a href="#">CFTR2</a>
Reference	Rommens et al., Riordan et al., Kerem et al. 1989

To check if there are any papers published about this mutation/variant on PubMed, please [click here](#).

**Literature referencing this mutation. Sort by:**

Note: This reference list is not up-to-date at this stage, but may be searched for some rare variants without pubmed hits.

- Rapid screening for delta F508 deletion in cystic fibrosis. 1989 012 2;2(8675):1345-6
- Worldwide survey of the delta F508 mutation--report from the cystic fibrosis genetic analysis consortium. 1990 008;47(2):354-9
- Gradient of distribution in Europe of the major CF mutation and of its associated haplotype. European Working Group on CF Genetics (EWGCFG). 1990 009;85(4):436-45
- [Correlation between genotype and phenotype in patients with cystic fibrosis. The Cystic Fibrosis Genotype-Phenotype Consortium. 1993 010 28;329\(18\):1308-13](#)
- Abelowich D, Lavon IP, Lerer I, Cohen T, Springer C, Avital A, Cutting GR. Screening for five mutations detects 97% of cystic fibrosis (CF) chromosomes and predicts a carrier frequency of 1:29 in the Jewish Ashkenazi population. 1992 011;51(5):951-6

## Describing variants

"mutation nomenclature"

**recommendations for the description of DNA changes**



**HGVS**  
HUMAN GENOME VARIATION SOCIETY

**Johan den Dunnen**  
Human Genome Variation Society  
(HGVS)

HGVSmn@JohanDenDunnen.nl

<http://www.HGVS.org/mutnomen/>  
( HUGO-MDI initiative )



Human and Clinical Genetics

© JT den Dunnen



## Definitions

- **prevent confusion**
  - mutation*
    - change
    - disease-causing change
  - polymorphism*
    - change in >1% population
    - not disease causing change
- **better use neutral terms**
  - sequence variant*
  - allelic variant*
  - alteration*
  - CNV** (Copy Number Variant)
  - SNV** (not SNP)



Human and Clinical Genetics

© JT den Dunnen





---

*Clinical labs will tend to make the call at the 3' end.*

*This is what we do here and ... says that is the common practice at clinical labs.*

*I don't know if you can adjust the aligners and tell them to prefer gaps at the right instead of the left.*

---

## Gaps at the right: CF [delta]F508

---

THE TWO ALIGNEMENTS BELOW ARE EQUIVALENT

```
S: 117199639 ATATCATCTTTG 117199650 (159138663)
Q: 192 ATATCA---TTG 200 (254)
Q: 196 ATATCA---TTG 204 (212)
Q: 237 ATATCA---TTG 245 (259)
Q: 209 ATATCA---TTG 217 (260)
Q: 195 ATATCA---TTG 203 (264)
Q: 193 ATATCA---TTG 201 (244)
```

```
S: 117199639 ATATCATCTTTG 117199650 (159138663)
Q: 192 ATATCAT---TG 200 (254)
Q: 196 ATATCAT---TG 204 (212)
Q: 237 ATATCAT---TG 245 (259)
Q: 209 ATATCAT---TG 217 (260)
Q: 195 ATATCAT---TG 203 (264)
Q: 193 ATATCAT---TG 201 (244)
Q: 193 ATATCAT---TG 201 (208)
```

**BUT THEY LEAD TO TWO DIFFERENT HGVS**

---

## Now ... Back to 1000G

---

- ▶ Individuals are cohorts from disease studies
  - ▶ Comparison with known clinical mutation databases
    - HGVS management is key – the SNP calls must be corrected
    - Accuracy matters
    - Not really Big Data business
    - (meaning if you have 100M to compare, ok, but this problem comes next, once above is fixed)
  - ▶ Variants present in ill's, absent in healthy's
  - ▶ Homozygote variants in study A, absent in B, and many more...
- 

## Now ... Back to 1000G

---

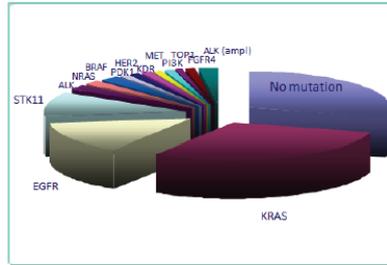
- ▶ Individuals are cohorts from disease studies
  - ▶ Comparison with known clinical mutation databases
  - ▶ Variants present in ill's, absent in healthy's
    - Equivalent to a RDMBS of 81M lines, 2504 columns = 200 Giga values
    - Actually, such a query already takes minutes on a RDBMS
  - ▶ Homozygote variants in study A, absent in B, and many more...
-

Situation will not improve ...

French Cancer Plan- INCA – Slide from Prof. Fabien Calvo



## Un nombre croissant d'altérations moléculaires



→ Le nombre d'altérations génétiques augmente régulièrement

→ Implémentation du NGS ciblé en diagnostic dans l'ensemble des 28 plateformes de génétique moléculaire

- Séquençage ciblé d'un panel de gènes
- Appel à projets pour sélection de projets pilotes

→ Objectif : 60 000 patients bénéficieront d'analyses de leur tumeur par NGS ciblé d'ici à fin 2016.

Situation will not improve ...

Technology scales up

### HiSeq X™ Ten

\$1000 human genome and extreme throughput for population-scale sequencing.

#### HiSeq X Ten Highlights

- **World's First \$1000 Human Genome:**  
First sequencing system to break the \$1000 barrier for 30x human genome sequencing, delivering unmatched cost-effectiveness for population-scale projects
- **Purpose-Built For Population-Scale Genome Sequencing**  
Highest daily throughput with the power to sequence tens of thousands of genomes, enabling whole-genome sequencing on an unrivaled scale
- **Proven Performance**  
Highly accurate Illumina sequencing by synthesis (SBS) chemistry delivers industry-leading data quality; gain confidence in your results with the most widely adopted and proven next-generation sequencing (NGS) technology

#### Introduction

Through continuous innovation, Illumina technology has broken down barriers in human genome sequencing by increasing data throughput at an astounding rate—more than doubling each year—while dramatically reducing the price to sequence a human genome. Illumina technology has forever changed the landscape of genomics research by enabling sequencing of the first genome at 30x coverage, the first cancer genome, and the first genome in a single day.<sup>1-3</sup>

Now, the HiSeq X Ten, a set of 10 HiSeq X Sequencing Systems, has reached yet another milestone, delivering the world's first \$1000 human genome and providing the throughput to sequence tens of thousands of high-quality, high-coverage genomes in a single lab. With its massive throughput and unprecedented low price per genome, HiSeq X Ten makes population-level human genome sequencing a reality (Figure 1).

#### World's First \$1000 Human Genome

The HiSeq X Ten is the world's first sequencing platform to break the \$1000 barrier for 30x coverage of a human genome. When used at scale, the HiSeq X Ten delivers a \$1000 genome, inclusive of instrument depreciation, sequencing consumables, DNA extraction, library preparation, and estimated labor for a typical high-throughput genomics laboratory.\*

Figure 1: The HiSeq X Ten



The HiSeq X Ten, a set of 10 HiSeq X Sequencing Systems, is the only high-throughput sequencing system that can produce tens of thousands of human genomes a year for under \$1000 per genome.

Figure 2: The HiSeq X Ten Enables Sequencing of Tens of Thousands of Human Genomes per Year

	HiSeq X	HiSeq X Ten
Week	32	>320
Month	>150	>1500
Year	>1800	>18000

When operating in parallel, the 10 instruments of a HiSeq X Ten generate a staggering level of throughput, capable of sequencing tens of thousands of genomes per year.

## Going further

---

- ▶ Individuals are cohorts from disease studies
  - ▶ Comparison with known clinical mutation databases
  - ▶ Variants present in ill's, absent in healthy's
  - ▶ Variants present in study A, absent in B
    - ▶ Homozygote calls
    - ▶ Inside a Coding DNA Sequence
    - ▶ Overlapping with a Transcription Factor binding site
    - ▶ Causing a AA frameshift
- 

## To do this, you need:

---

- ▶ To be able to query (index) the database(s) on:
  - ▶ Character Strings
  - ▶ Numerical values
  - ▶ **HGVS (semantical)**
  - ▶ **Intervals (positional tracks = genomic regions)**
    - ▶ **Add them dynamically (addition of a track)**
  - ▶ **All of the above simultaneously**
- ▶ To be able to perform biology-oriented programs on the fly (e.g. impact prediction if not precomputed)

No existing DBMS system does this easily.  
It is domain **specific**.

---

# Bioinformatics

---

- ▶ Needs large to very large IT equipment:
    - ▶ Computing
    - ▶ Storage
    - ▶ Network
    - ▶ ...
  
  - ▶ But, at the end of the day, it is NOT Software AG, Oracle Corporation, IBM, FICO, Microsoft, SAP, EMC, HP nor Dell who will implement Life Sciences Bioinformatics solutions
  
  - ▶ True bioinformatics (bio-IT) knowledgeable companies, or groups must do the job in collaboration with clients, and use IT big company's skills and technology, to support the big data volume.
-



# Towards Health Grid Initiatives in India: Grand Challenge in Affordable Health Care- Prospects and Insights from the Brain Grid

Prasun Kumar Roy

Tata Innovation Fellow

Senior Professor, National Brain Research Centre,  
NCR Delhi

Chief Investigator, India Brain Grid, Ministry of I.  
T., Govt. of India

Jt. Coordinator-Indian Node, International  
Neuroinformatics Coordinating Facility







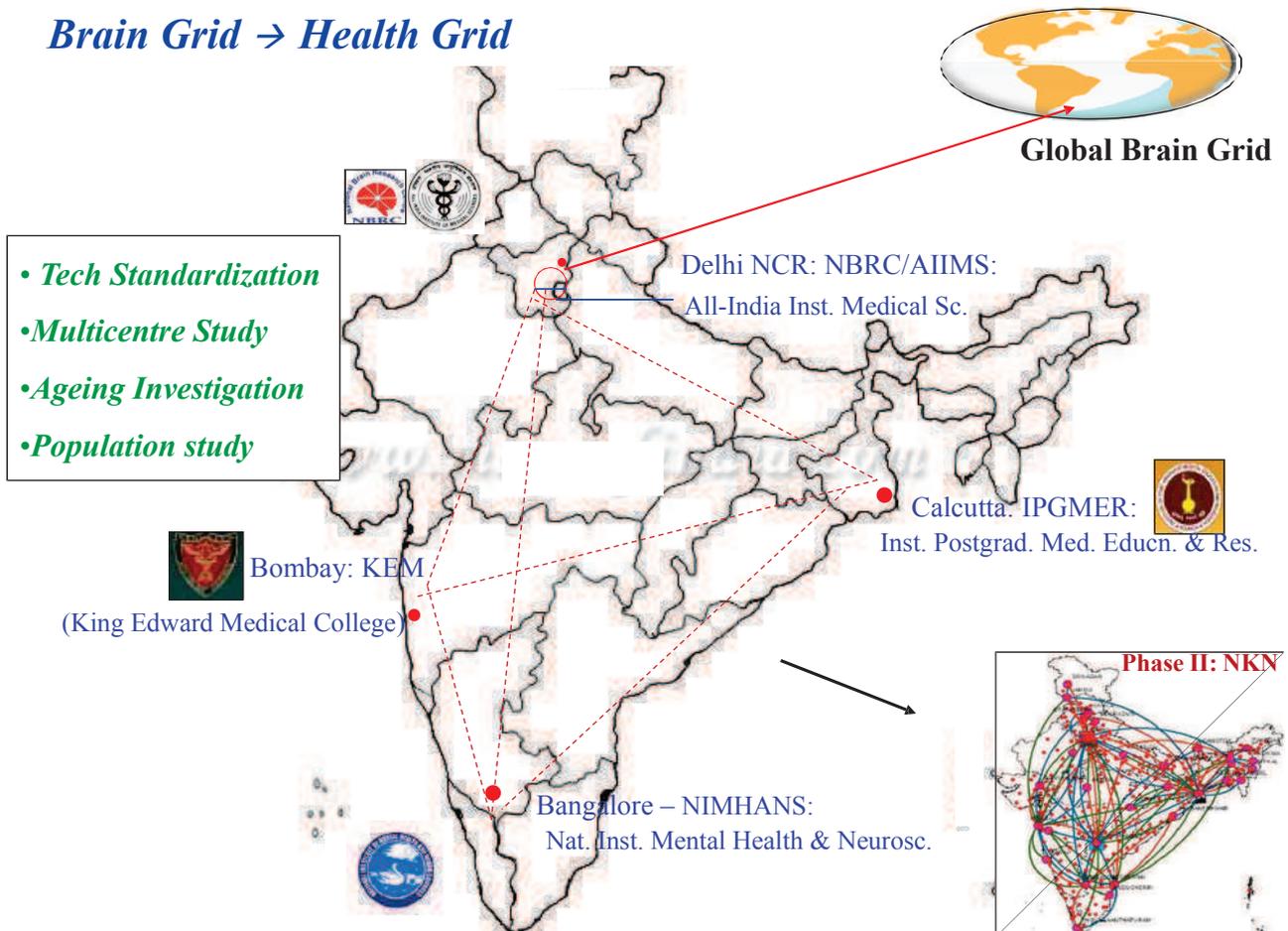
# Towards Health Grid Initiatives in India: Grand Challenge in Affordable Health Care- Prospects and Insights from the Brain Grid

**Prasun Kumar Roy**  
Tata Innovation Fellow



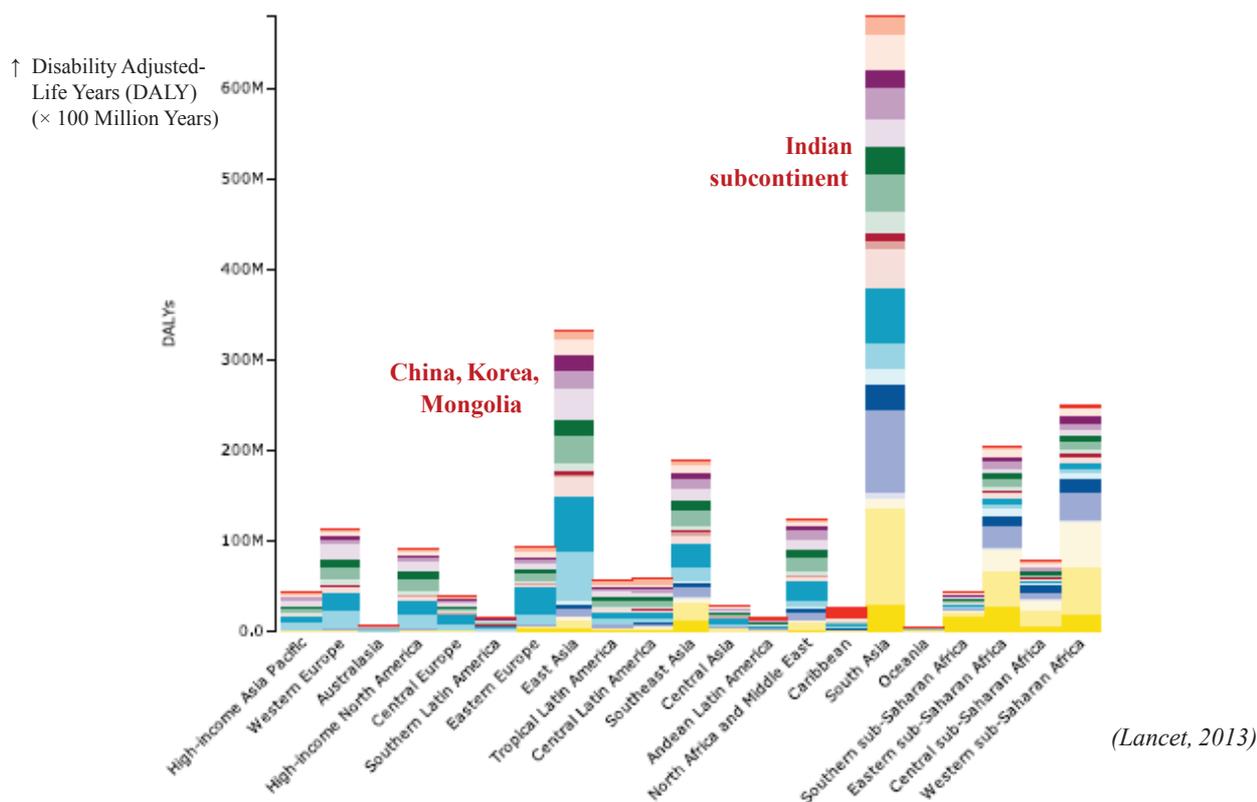
**Senior Professor, National Brain Research Centre, NCR Delhi**  
**Chief Investigator, India Brain Grid, Ministry of I. T., Govt. of India**  
**Jt. Coordinator-Indian Node, International Neuroinformatics Coordinating Facility**

## *Brain Grid → Health Grid*



## **Motivation - Pressing Need Today**

### **Indian Region – Global Highest Disease Load**



## **Rationale of Health Grid**

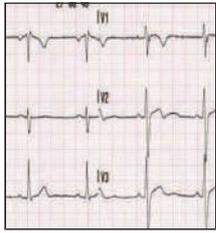
- **Optimal Utilization of Scarce Resources**
- **Global Efforts: Europe, N. America, Asia-Pacific.**
- **Organ-Specific Indian Grids initiated: Synergize them**
- **Crucial for Mass-scale Screening, Prevention, Therapy**
- **General Approach Applicable for Other Countries**
- **Scope for Cooperation with Industry & Knowledge Economy**

# Human Function: Large data vol. - Many patients:

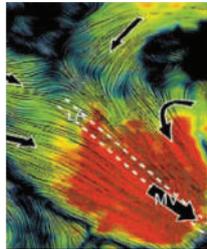
## Computed Analysis in Human Template Platform

First by Jean Talairach Template, Paris, 1960s

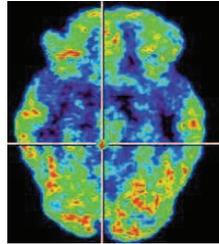
Electrophysiology



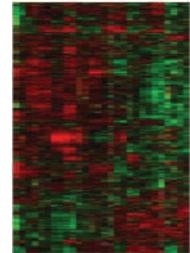
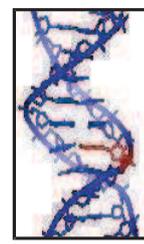
Blood flow, fMRI



Molecular PET

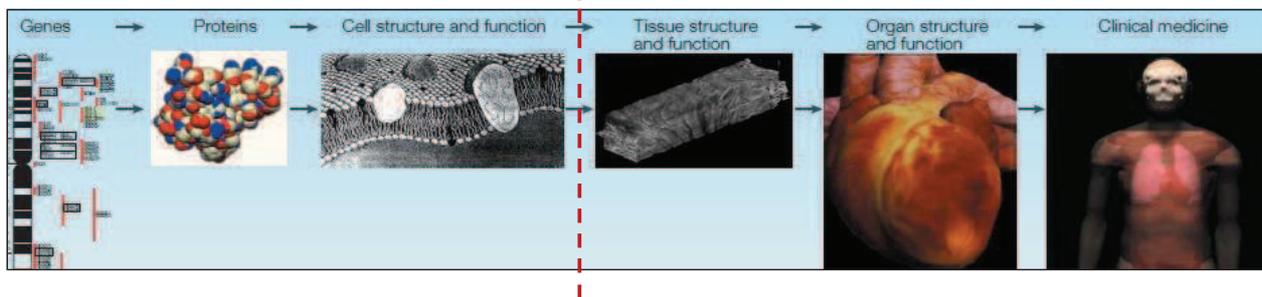


Genomics



BioInformatics

OrganoInformatics



### Some Int'l Medical Grids



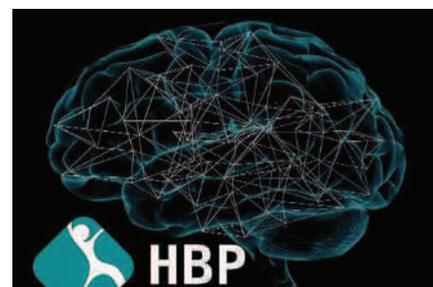
### European Grids: Neugrid, HBP

University College  
London

UK, France, Italy, Germany, Sweden, Finland, Greece



### Human Brain Project



- ♣ Child Development (UCL)
- ♣ Brain Injury (Karolinska)
- ♣ Dementia (ICRRS Venice)
- ♣ Epilepsy (CNRC Toulouse)
- ♣ Autism (MPI, Dresden)



# U.S. Grids: ICBM /ADNI / NIH

UCLA, Harvard

Int'l Consortium of Brain Mapping;

Alzheimer's Disease Neuroimaging Initiative

**NIH MRI Study of Normal Brain Development (N=500)**

Behavior/MRI for ages 0-18 yrs  
Structure-behavior relationships  
Disseminate results

What does it find (see inside the scanner?)

T1W  
T2W

1 year 2 years 10 years

SPC, CCC, DPC, DCC

- Partners:**
- Siemens
  - Philips
  - Pfizer
  - Roche
  - Allen Brain Institute

## India Brain Grid Platform Developed

Interaccession & Analysis of MRI-PET-EEG-Genomics  
Electrophysiology / Psychometry / Blood Flow Data

Navigation: Back to list

3D Surface Viewer

3D Viewer

Visit-Level controls: Visit-level feedback, QC Status (Pass, Pending, No), Save

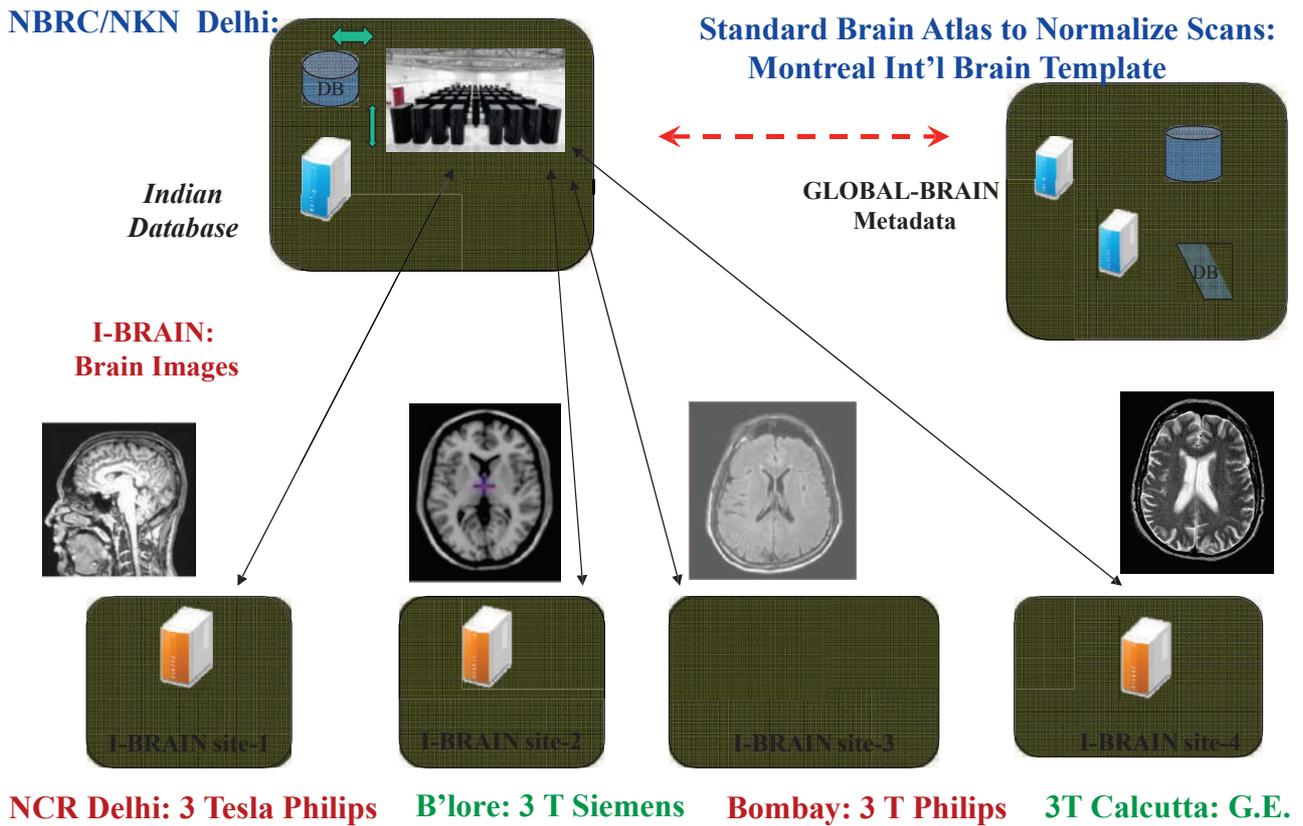
Click and drag the red dot to change slices. The red dot represents the location you are looking at.

Collab logo

Output Type: native  
Protocol: t1  
Space: native  
Acq Date: Oct 12, 2012  
Inserted: Feb 19, 2013  
Ser Desc: t1\_mpr\_1mm\_p2\_pos50  
Ser Num: 2  
Echo Time: 3.16 ms  
Rep Time: 2,400.00 ms  
Slice Thick: 1.00 mm

PLAUR  
LOX  
LIF  
STC1  
LGALS3  
VEGF

**Brain Grid Nodes initiated in the 4 zonal centres:  
Normalization of parameters of all Scanners**



**Program Initiative**

**Systems Biology & Anatomical Imaging-based  
Approach to Optimize Neuro-regenerative  
Therapy of Cerebral Stroke**

# Issues in Current Stroke Therapy & Vascular Dementia

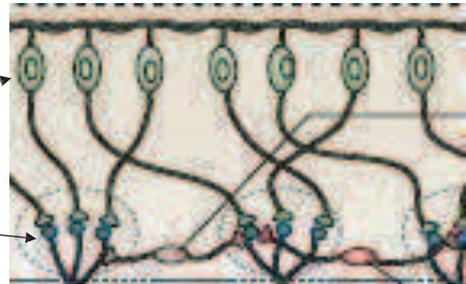
Loss of brain functions by blood supply loss, due to:

- Ischemia (lack of blood): thrombosis, embolism, hemorrhage :
- Current therapies do not target the underlying deficit, namely the loss of neurons

## Regenerative Treatment requires

Formation of Nerve Cells, Neurons (Neurogenesis)

Formation of connections, Synapses (Synaptogenesis)



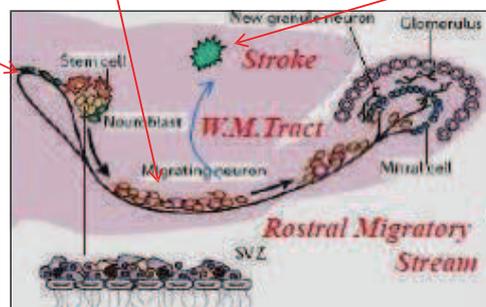
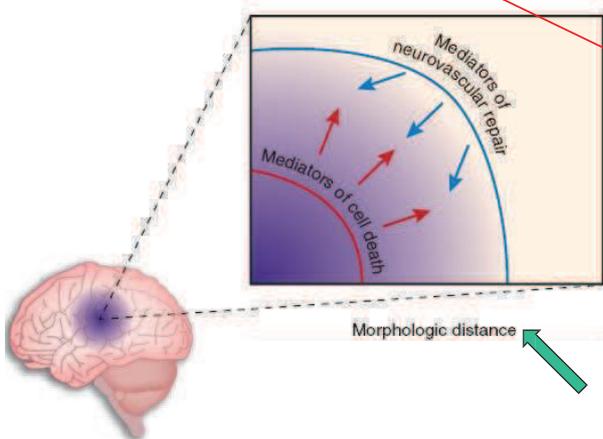
## Rationale : How to initiate Regenerative Therapy:

Activating Internal Stem Cells in Brain → Neurons → Synapses

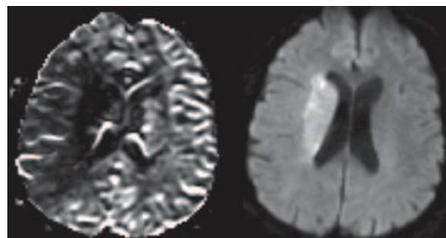
*Drugs: (i) Neurogenesis: Minocycline*

*(ii) Angiogenesis: Atorvastatin*

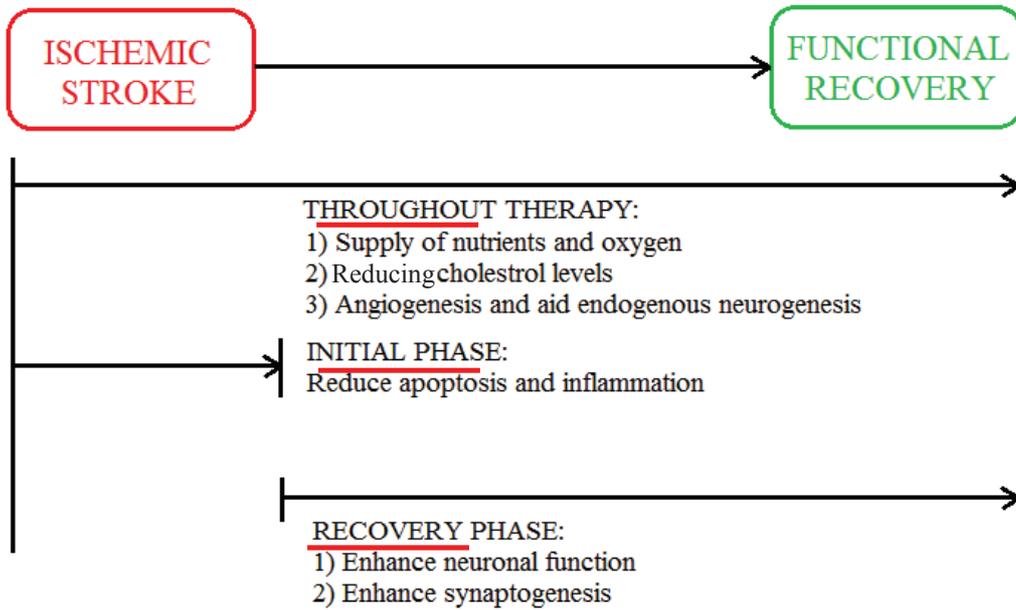
*(iii) Synaptogenesis: Fluoxetine*



**Viable Penumbra Area:**  
Perfusion-Diffusion Subtraction Imaging

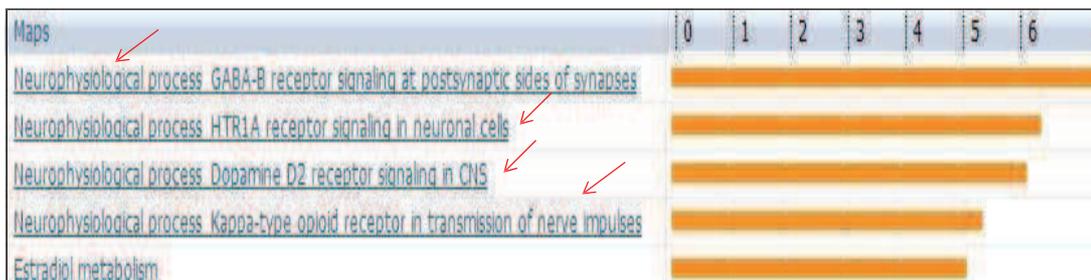


## Recapitulate Time Schedule of Proposed Therapy model



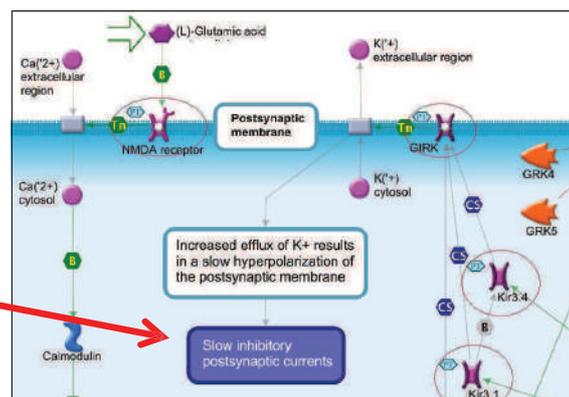
## Choosing Drugs: Grading Effectivity of Synaptogenesis Agents: Systems Biology analysis of Fluoxetine for Neuroprotection

Shannon Information →



Hence, Fluoxetine is Satisfactory

Slow Inhibitory Postsynaptic Currents



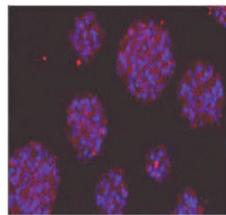
# Systems Biology Analysis of Escitalopram for Neuroprotection

Maps	0	0.5	1	1.5	2	2.5	3
Nitrogen metabolism	[Orange bar]						
Nitrogen metabolism/ Rodent version	[Orange bar]						
Acetaminophen metabolism	[Orange bar]						
Cortisol biosynthesis from Cholesterol	[Orange bar]						
Neurophysiological process Thyroliberin in cell hyperpolarization and excitability	[Orange bar]						
Estradiol metabolism	[Orange bar]						
Estradiol metabolism / Human version	[Orange bar]						

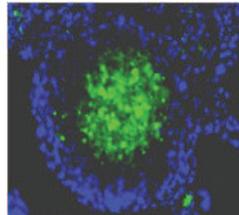
**No positive neurological functions associates with Escitalopram**

*Escitalopram is Rejected*

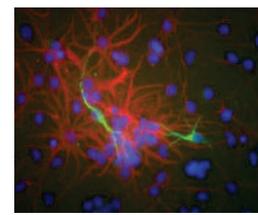
## Cell Proliferation/Migration: Computational Modelling Cell Prolif. Rates from Ventricular Zone Cultures of Neurogenesis



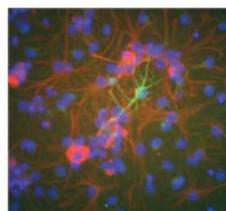
**Stage-I:**  
Neural Stem Cells:  
Neurosphere-Mushashi



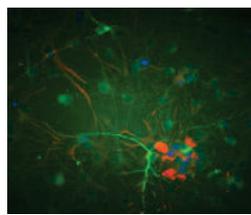
**Stage-II:**  
Prolif. Neural Precursor Cells:  
SVZ (Neurosphere-Nestin)



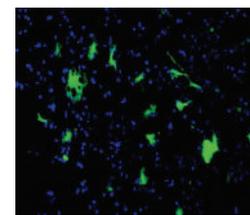
**Stage-III:**  
Early Neurons:  
Day-1 neurons (BIII, GFAP)



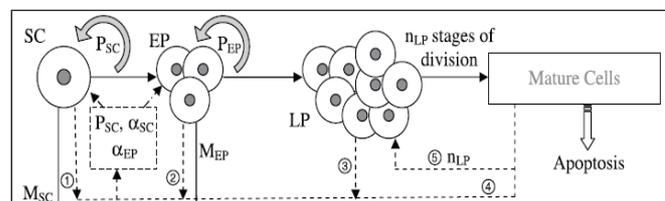
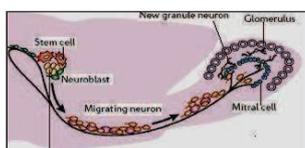
**Stage-IV:** Adult Neurons  
Day3 neurons: NeuN, GFAP



Day-3 neurons- NeuN, GFAP

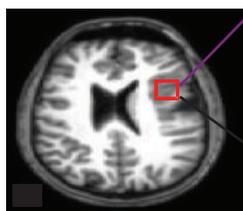
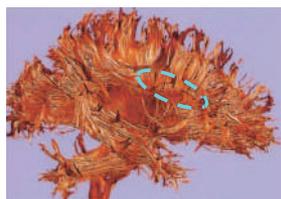


Astrocytes: GFAP



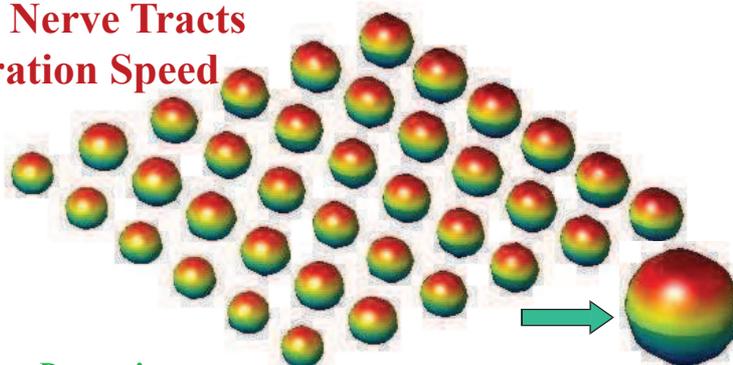
**Math Eqns.: Migration/Prolif. Stages**

# Stem Cell Migration Along Nerve Tracts (Path via MRI-DTI): Migration Speed

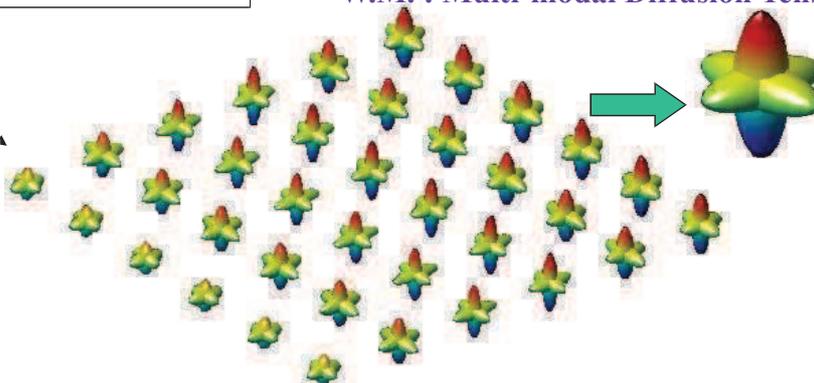


Cell Migration Dynamics

$$\frac{dV_{\perp}}{dt} = D(\nabla^2 N)$$

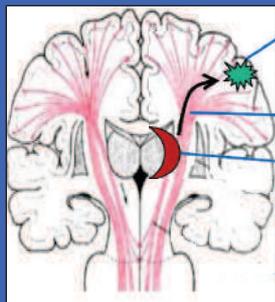


G.M. : Spheroidal Diffusion Tensors



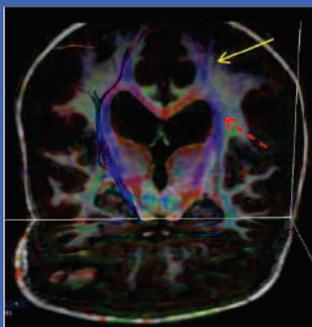
W.M. : Multi-modal Diffusion Tensors

## Temporal Scheduling: Neurogenesis & Synaptogenesis Drugs

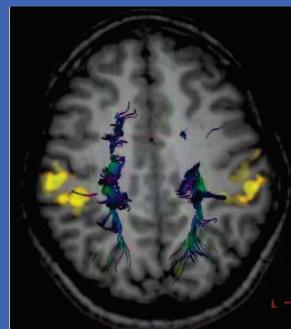


Stroke  
Fibre  
Ventricle (SVZ)

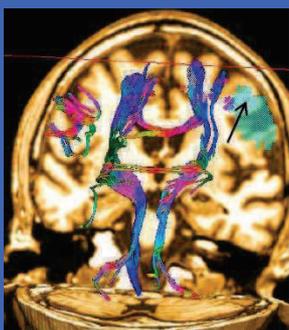
Tract Anatomy



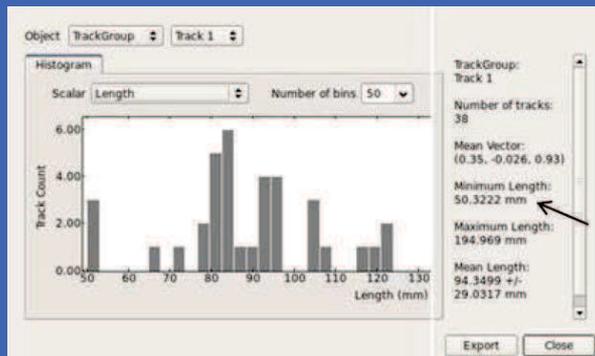
Tract Image. DTI: Stroke. NBRC



fMRI Activation + Tract: Stroke. NBRC

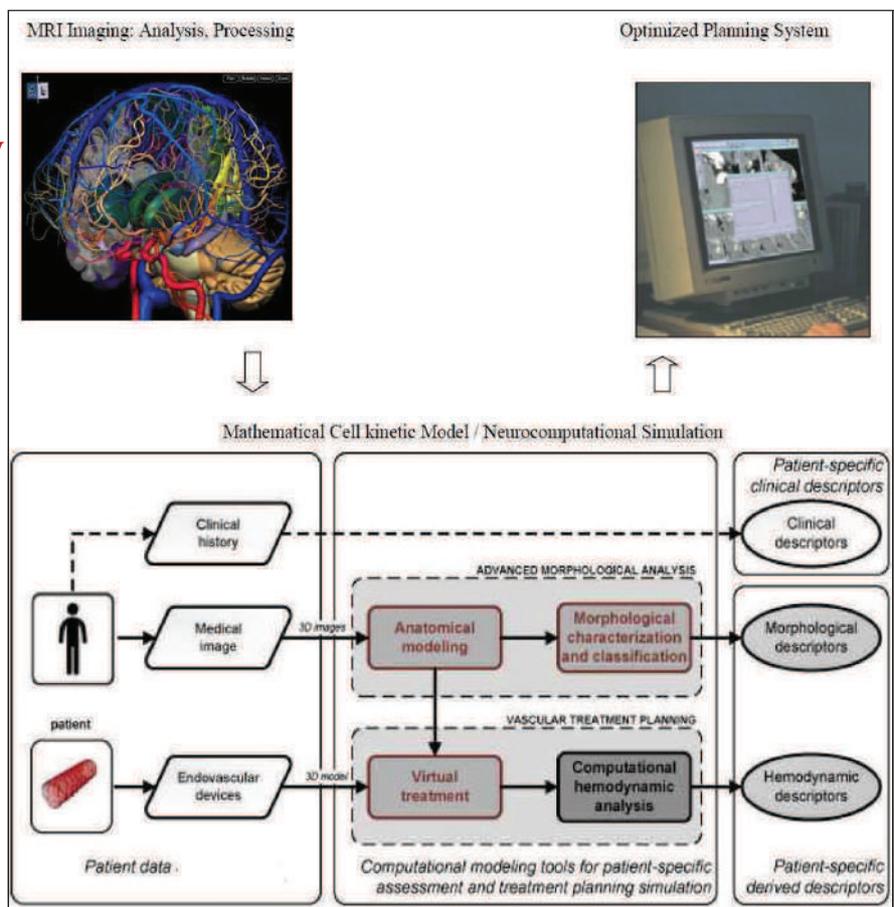


Tract Image. DTI: Stroke. AIIMS

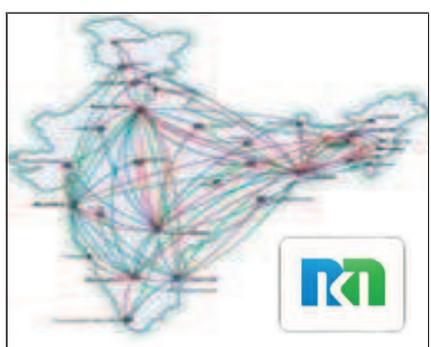


Direct Path: Tract Length Calculation: NBRC: 50.3 mm

# Neuroinformatics Expert System for Optimizing Therapy



## India Brain Grid (*iBrain*): Infrastructure Usable for Other Grids



**NBRC**  
Gurgaon  
(North)



**NIC**  
Delhi  
(North)



**AIIMS**  
Delhi  
(North)



**NIMHANS**  
Bangalore  
(South)



**KEM**  
Bombay  
(West)

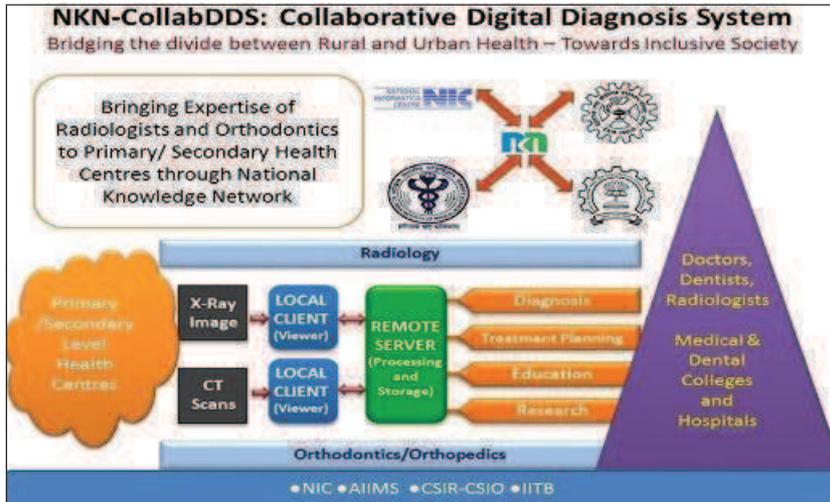


**IPGMER/Cal. Univ.**  
Calcutta  
(East)

## SkeletoMuscular Grid:

AIIMS-Delhi, IIT-Bombay, MIMS-Mangalore, CSIR-Chandigarh

(Coordinator: NIC, Delhi, Bangalore)



## Liver Grid – India

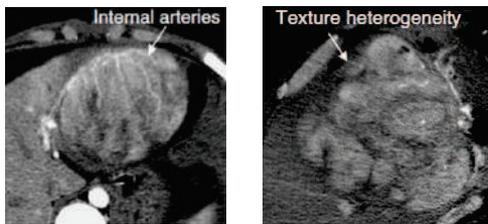
SGPGI-Lucknow,

CMC-Vellore,

LNJPI-Bombay

Coordinator: Int. of Liver & Biliary Sc.-Delhi

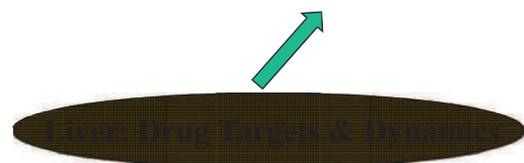
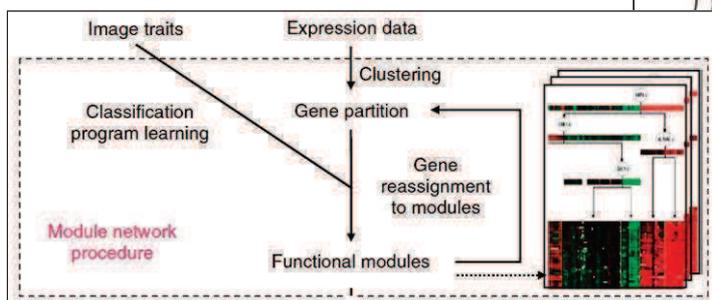
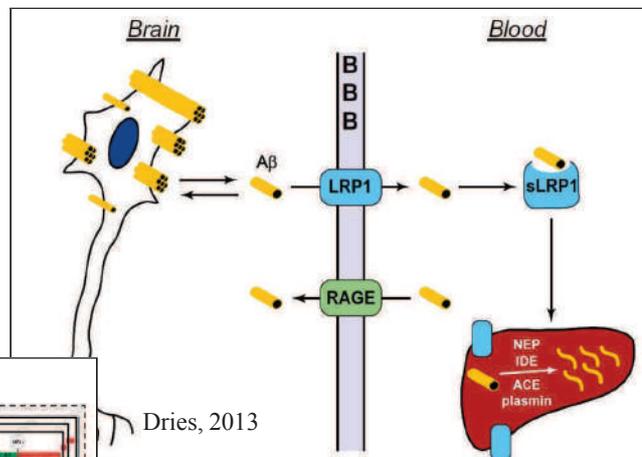
### Tissue Flow Dynamics Imaging



$$\frac{\partial A}{\partial t} = D_{\varepsilon} + \left[ \left( \frac{P_u - P_v}{8\pi u l_{eff}} \right) A_{v,eff}^2 + l_{out} \frac{\Delta P A_{in} H_{\varepsilon} + S}{E} \right] F.K'.A'.G^*L' + \left( \frac{p h^2}{12\mu} \right) \rightarrow Eq13$$

(Durga Prasad, Roy)

### Amyloid Excretion by Liver under Drug (Alzheimer's Disease: NBRC)



## Cancer Grid – India

C-NCI-Calcutta, Reg. Can. Inst.-Trivandrum, Rajiv Gandhi Can.Inst.-Delhi

(Coordinator: TMH – Bombay)

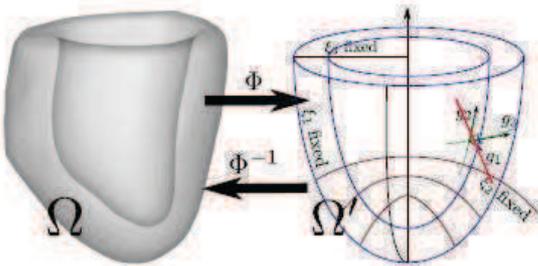


### National Cancer Grid

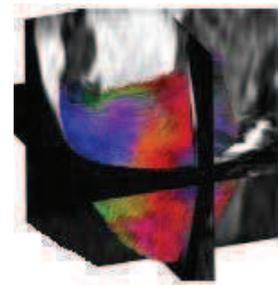
As ascertained from the Tata Memorial Hospital, Mumbai, the National Cancer Grid is envisaged to be a network of existing and future major cancer centres in the country created with the mandate of:

- i) Creating uniform standards of patient care across the length and breadth of country, bringing high quality cancer care to the doorsteps of patients.
- ii) Augmenting human resource capabilities in cancer management in the country.
- iii) Promoting collaborative research in cancer.

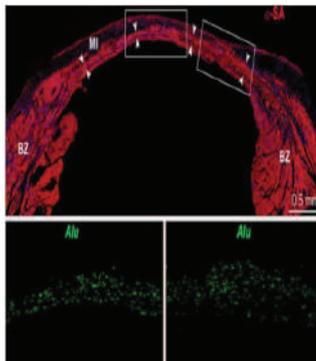
## The Heart Grid: Cardiac Mapping & Regeneration



Computational Fluid Dynamics



Cardiac Muscle Connectomics  
(MRI-DTI)



Ischaemic Regeneration by  
Cardiac Endogenous Stem Cell

## Health Grid: Possible directions

### Database & Tracking for Indians: Towards Country's Productivity

- Imaging      ♠ Electrophysiology
- Biochemistry      ♠ Genetics      ♠ Psychometry & Cognition

**Time Changes: Child's Biological/Psychological Dev.; Ageing Effects**  
In Normality and in Diseases

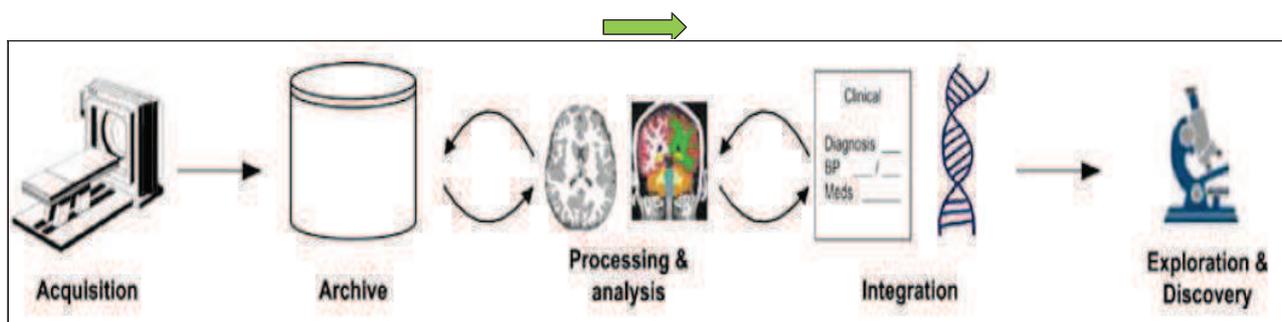
**Clinical Biomarkers for Screening Abnormality/High Risk cases**  
For Early Disease Diagnosis & Grading

**Cross-Sectional Study:**  
Mapping Variation across Gender and Different Ethnic Group in India

**Monitoring of Treatment Response:**  
At early treatment stage, to reject ineffective drugs

### Health Grid Summary:

*Digital Organ Initiative → Digital Human Initiative*



Computing Platforms available;

To Use/Interpret: Signals/Images/Biochemical Data

From Data Analysis & Causality Analysis :-

- ◆ Obtain Scientific Models,
- ◆ Delineation of Risks, Screening, Choosing Drugs
- ◆ Quantitative Relationships & Principles

## **Collaborative Initiative for International Partnership**



**Human Organism: Unique Dynamics Platform**

**Big Data :**  
**Making Sense from Large-scale**  
**National-International Data Streams**

**National/Global Problem:**  
**Growth, Ageing, Life-span**

**Screening, Diagnostics, Therapeutics**

**Decisions for Policy-Makers**

**Economic Productivity**

**Blue Sky Products & Services**

### **Some Things To Do Now (?)** **Any Others (?) :**

- **What to Do Now....?**
- **Compose A White Paper**
- **Include 2-3 Specimen sub-projects**
- **Approach the Office of the Agencies/Govts. of Partnering Countries**
- **Funding Models of the Endeavour**
- **Long-Term Financial Viability/Financial Paradigms of the Health Grid**
- **Towards integrated country-level Health Care Initiative**
- **.....(What Else)..... ?**

## Cooperation Members of Indian Institutions

**National Informatics Centre**

Dr Savita Dawar

**Indian Instt. Of Technology-Delhi**

Dr Rajesh Khanna

**National Knowledge Network**

D K Gupta

**Office of Principal Scientific Advisor, Govt. of India**

Neeraj Sinha

**KEM Medical Instt.**

**Bombay**

Dr S Sankhe

**IPGMER**

**Calcutta**

Dr S P Basu

**AIIMS**

**Delhi**

Dr S Kumaran

**NIMHANS**

**Bangalore**

Dr John P John

## Team Members



*Subramanyam*



*Subhadip*

Computational Analysis



*Jyothi*  
*(now in Italy)*



*Rajeev*

Neuroimaging



*Budha*  
*(now in Canada)*



*Suhela*

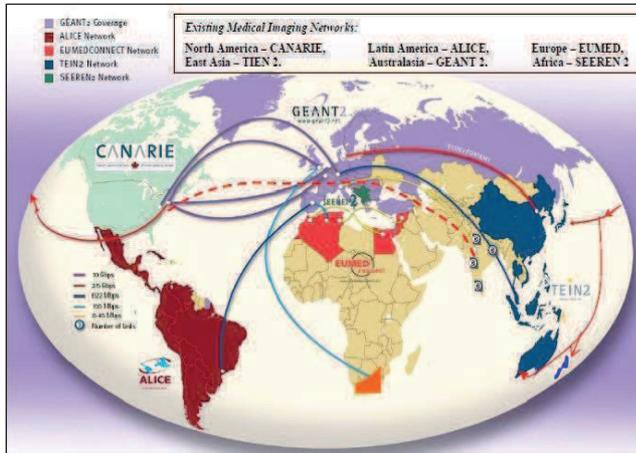
Biomedical Engg.

Molecular Biology

*Digital Human Initiative*  
*Towards Well-being of Nations*

Indeed, Health Grid can be a Model for other regions globally

**International Facilitating Agencies:**



**International ' Telecommunication Union.**



**Office of Principal Science Advisor, GoI**



**International Neuroinformatics Coordinating Facility**



**National Knowledge Network**  
*Connecting Knowledge Institutions*

# Flexible and Efficient RNA-Seq analysis at GenoSplice

Frédéric Lemoine, PhD







# Flexible and Efficient RNA-Seq analysis at GenoSplice

Frédéric Lemoine, PhD

2014/10/15

## Summary

1. **GenoSplice: A few words about the company**
2. **RNA-Seq: Several levels of analyses (and complexity)**
3. **Focus on RNA-Seq splicing analysis: The GenoSplice strategy**
4. **Example from a GenoSplice project**
5. **Conclusion**

1. **GenoSplice: A few words about the company**
2. RNA-Seq: Several levels of analyses (and complexity)
3. Focus on RNA-Seq splicing analysis: The GenoSplice strategy
4. Example from a GenoSplice project
5. Conclusion

## GenoSplice is a bioinformatics service provider

### ► ***Omics data analyses:***

- RNA (expression, splicing, small RNA, fusion transcript...);
- DNA (SNP/indel, translocation...);
- Protein (RPPA...);
- Epigenetics (methylation...).

### ► ***Other services:***

- Functional analyses (Pathway analysis, Gene Ontology...);
- Biomarker discovery/signature definition;
- Data and content mining;
- Bioinformatics advisory/custom development.

# Main customers

## Private companies



## Academics Labs

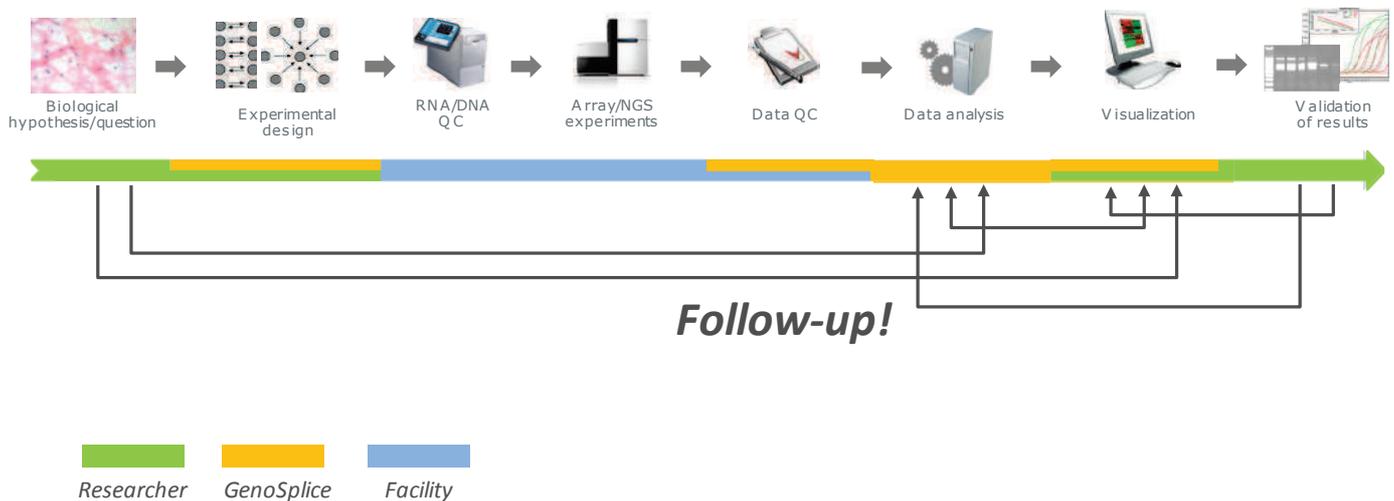


5

2014/10/15

# GenoSplice philosophy: BaaS (Bioinformatics As A Service)

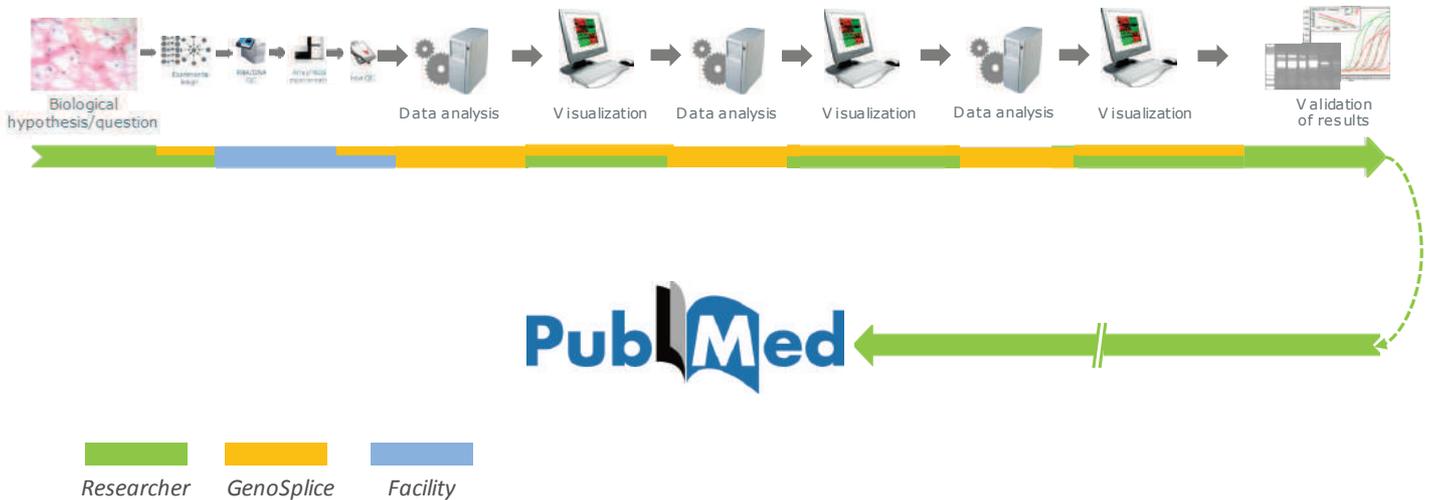
*Steps of Omics project:  
from hypothesis to validation*



169

2014/10/15

## Steps of Omics project - the "real view": from hypothesis to validation



## Most relevant scientific publications involving GenoSplice services

### ► 2014

- Cottu et al., *Clin Cancer Res*;
- Bechara et al., *Molecular Cell*;
- Pesson et al., *PLoS One*;
- Pesson et al., *Molecular Cancer*;
- Vogt et al., *Genome Research* (under review);
- Almeida et al., *PLoS One* (submitted);
- Vallerand et al., *Molecular Oncology* (submitted).

### ► 2013

- Furney et al., *Cancer Discovery*;
- Convertini et al., *Biochimica et Biophysica Acta*;
- Astori et al., *Oncotarget*;
- Beurlet et al., *Blood*;
- Shen et al., *Nucleic Acids Research*;
- Gandoura et al., *Journal of Hepatology*;
- Sousa et al., *EMBO Molecular Medicine*;
- Chapat et al., *EMBO Journal*;
- Lustremant et al., *Cel. Reprog*;
- Falaleeva et al., *Clinical Medicine Insights: Case Reports*;
- Zhang et al., *PLoS One*.



### ► 2012

- Jia et al., *Cell*;
- Ameyar-Zazoua et al., *Nat Struct Mol Biol*;
- Ballarino et al., *Oncogene*;
- Wang et al., *PLoS One*;
- Gaudineau et al., *J Cell Sci*;
- Hadj-Hamou et al., *Carcinogenesis*;
- Montjean et al., *J Assist Reprod Genet*

### ► 2011:

- Fugier et al., *Nature Med*;
- Saint-André et al., *Nat Struct Mol Biol*.
- Rajan et al., *PLoS One*;
- Lemonnier et al., *Human Mol Genet*;
- Ramgolam et al., *PLoS One*;

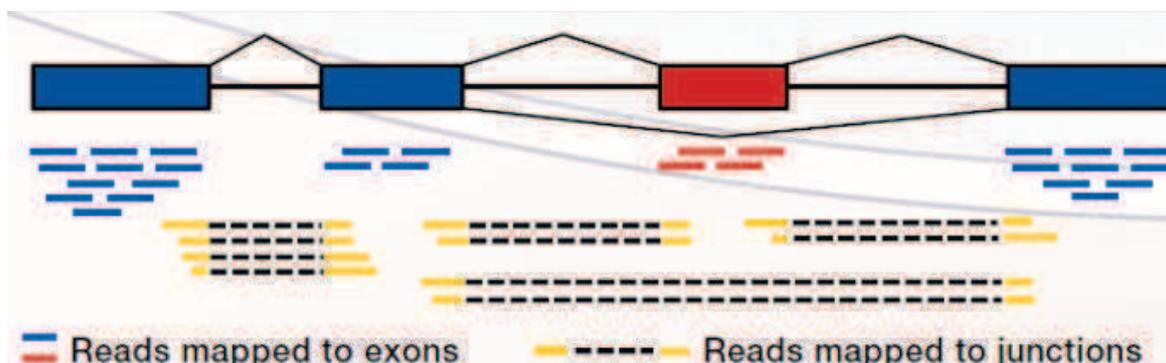
### ► 2010:

- Llorian et al., *Nat Struct Mol Biol*;
- Dutertre et al.2, *Cancer Research*;
- de la Grange et al., *Nucleic Acids Research*;
- Dutertre et al.2, *Cancer Research*.

1. GenoSplice: A few words about the company
2. RNA-Seq: Several levels of analyses (and complexity)
3. Focus on RNA-Seq splicing analysis: The GenoSplice strategy
4. Example from a GenoSplice project
5. Conclusion

## RNA-Seq: Several levels of analyses (and complexity)

RNA-Seq allows to study a snapshot of RNA presence and quantity from a genome at a given moment in time



### ► *Different levels of analyses:*

- Gene Expression;
- Alternative Splicing;
- Fusion transcripts;
- SNP/INDEL/Editing in transcribed regions.

1. GenoSplice: A few words about the company
2. RNA-Seq: Several levels of analyses (and complexity)
- 3. Focus on RNA-Seq splicing analysis: The GenoSplice strategy**
4. Example from a GenoSplice project
5. Conclusion

## Main steps of GenoSplice analysis for splicing

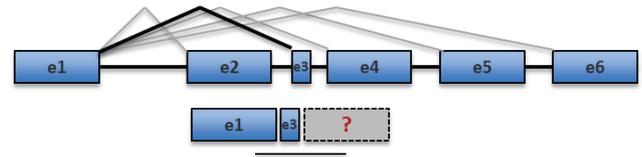
- 1. Read Mapping;**
- 2. Gene expression estimation;**
- 3. Exon level analysis;**
- 4. Pattern level analysis.**

## Mapping: GenoSplice's mixed approach

### ► *Genomic reference;*

### ► *Exon-exon junctions :*

- Using exon-exon boundaries (known);
- Taking into account very short exons;
- Depending on read length.



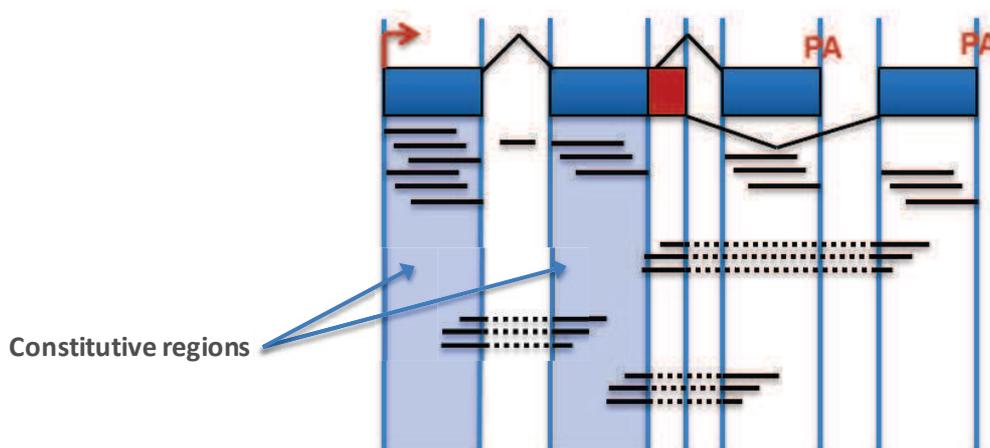
### ► *FAST DB : Annotation Database for Exons and junctions*

- Up to date database on genes, exons and splicing patterns;
- Enriched with RNA-Seq/microarray experiment results;
- Only accessible for clients/collaborators (diff from publically accessible FAST DB);
- FAST DB annotations used for:
  - RNA-Seq/microarray projects;
  - Integrative projects (e.g., CHIP-Seq/RNA-Seq/DNA-Seq);
  - Target annotations/gene documentation (in particular, for pharma clients)



## Gene expression estimation from GenoSplice

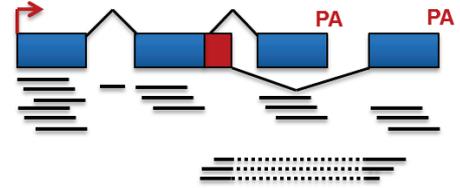
**GenoSplice's analysis: We count reads that map on constitutive regions of genes and use DESeq:**



## Exon level analysis

### ► At the exon level

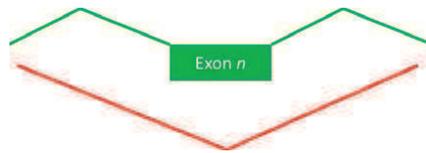
- We can use the same tools than expression to assess exon level differential expression, using **exon counts**



- But it depends on the expression of the **containing Gene**

### ► We use **splicing index** to take into account gene expression

- The gene expression is estimated using **constitutive regions**
- A **confidence value** is added to the exons using **junction expression**



## Pattern level analysis: Generalities

### ► Different kind of splicing events

- Alternative First Exons;



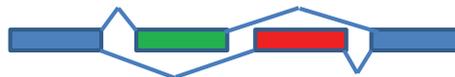
- Alternative Last Exons;



- Exon skipping;



- Mutually exclusive exons;



- Intron retention;



- Alternative Acceptor sites;



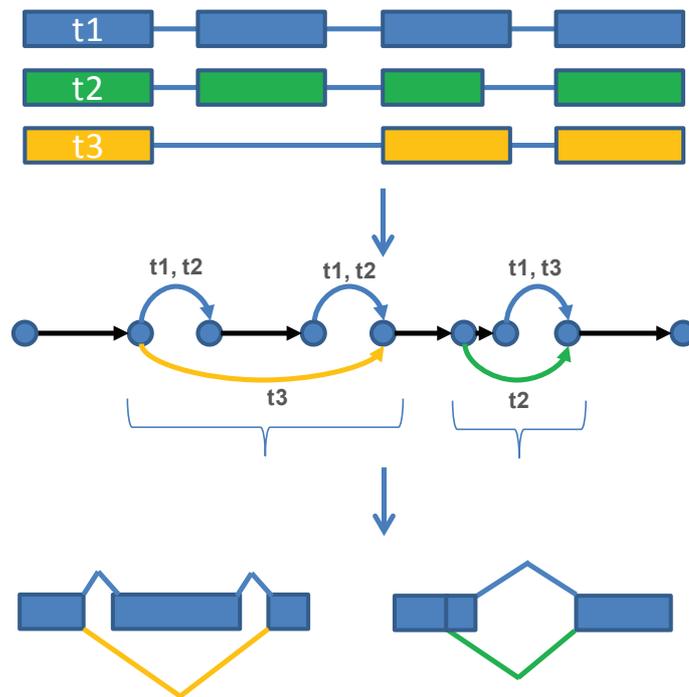
- Alternative Donor sites.



## Pattern level analysis: Definition of patterns 1/2

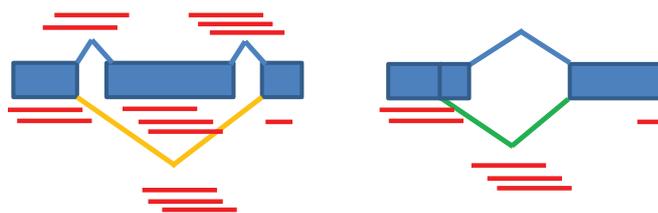
Creation of a Splicing pattern database using splicing graphs

For each gene:

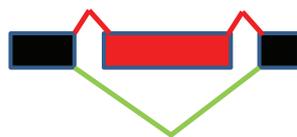


## Pattern level analysis: Definition of patterns 2/2

- ▶ Assigning mapped reads to splicing pattern regions:



- ▶ Comparing counts on different samples using Splicing Index:



- ▶ Sorting and Listing regulated splicing patterns

- ▶ Visualizing results with EASANA

1. GenoSplice: A few words about the company
2. RNA-Seq: Several levels of analyses (and complexity)
3. Focus on RNA-Seq splicing analysis: The GenoSplice strategy
4. Example from a GenoSplice project
5. Conclusion

## Example of a RNA-Seq GenoSplice project: Introduction

- ▶ SF3B1 was found recurrently mutated in:
  - Myelodysplastic Syndromes<sup>1</sup>;
  - Chronic lymphocytic leukemia<sup>2</sup>;
  - ER+ breast cancers<sup>3</sup>;
  - Uveal Melanoma<sup>4</sup>.
  
- ▶ Very few (or no) splicing defects were found (and no validation on possible aberrant splicing were presented);

1: Papaemmanuil et al., N Engl J Med. 2011

2: Wang et al., N Engl J Med. 2011

3: Ellis et al., Nature 2012

4: Harbour et al., Nature Genetics 2013

# Example of a RNA-Seq GenoSplice project: Results



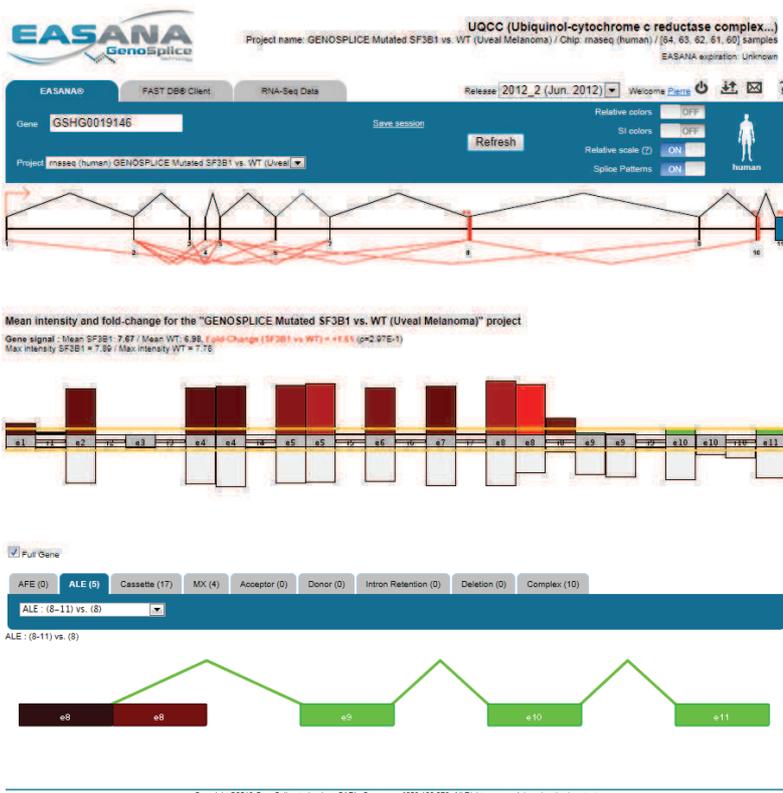
Analysis of RNA-Seq data by Harbour *et al.*: **NO RESULT!**

Reanalysis of this publicly available RNA-Seq data by GenoSplice and comparison with HTA2.0 results:

Gene Symbol	Possible Alternative Event	Harbour et al. (RNA-Seq)	RT-qPCR validation
ABCC5	Retention of intron 8	✓	✓
CRNDE	Alternative acceptor site (exon 5)	✓	✓
UQCC	Alternative terminal exons (exon 8 vs. exon 11)	✓	✓
GUSBP11	Cassette exon 7	✓	✓
ANKHD1	Alternative acceptors site (exon 3)	✓	✓
ADAM12	Alternative terminal exons (exon 18 vs. exon 23)	✓	✓

**Same RNA-Seq dataset was analyzed: 2 different results**

## RNA-Seq data: Visualization using GenoSplice's EASANA® 1/2

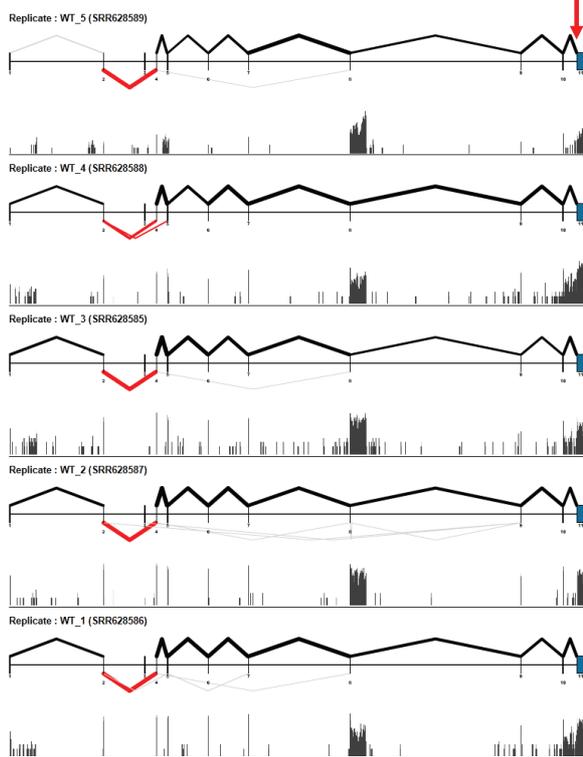


- Information regarding the project and different options for this interface (e.g., change color mode, scale...)
- Exon/intron gene structure with alternative events in red (according to GenoSplice annotations)
- Expression and expression regulation of the different gene parts (i.e., exon, part of exon, retained intron). Red means up-regulated in SF3B1 mutated tumors; Green means up-regulated in WT SF3B1 tumors
- Splicing pattern view: All potential splicing patterns of the gene can be displayed. Colors correspond to splicing-index regulation.

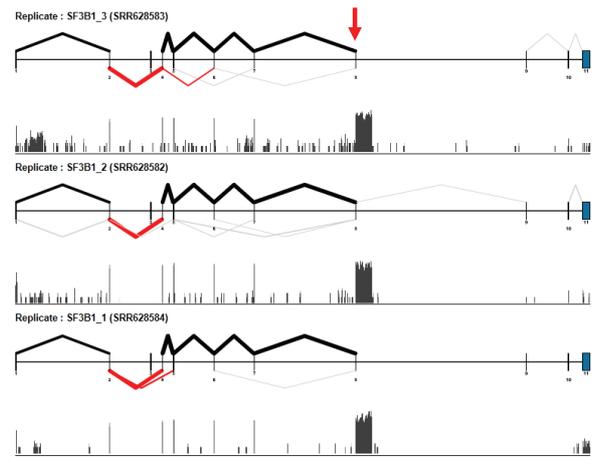
# RNA-Seq data: Visualization using GenoSplice's EASANA® 2/2

Another EASANA® option: Read coverage and main used exon-exon junctions:

## WT SF3B1 tumors: transcripts end in exon 11

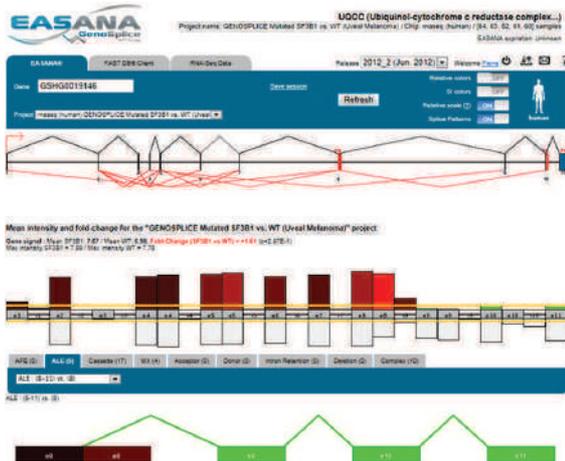


## SF3B1 mutated tumors: transcripts end in exon 8



In SF3B1 mutated tumors, transcripts end in exon 8 and subsequent protein product loss the Ubiquinol-cytochrome C chaperone domain

# GenoSplice's expertise for primer design



Targeted splicing variant	Primer ID (forward)	Primer ID (reverse)	Expected size (bp)
Exon 11 as terminal exon	GSHG0019146_F_e8	GSHG0019146_R_e9	170
Exon 8 as terminal exon	GSHG0019146_F_e8	GSHG0019146_R_i8	132

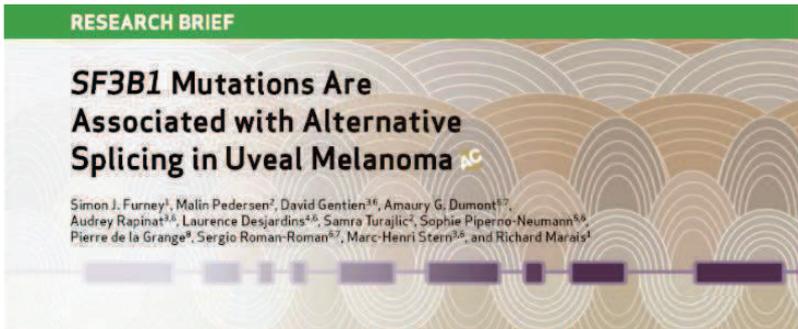
Up-regulated in SF3B1 mutated tumors

Down-regulated in SF3B1 mutated tumors

### List of the 25 distinct primers for RT-qPCR validations

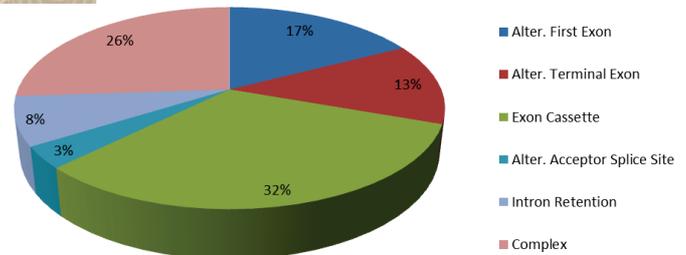
Primer ID	Primer Sens	Gene Symbol	Localization	Sequence	Length (bp)	Tm (°C)	GC (%)
GSHG0022278_F_e8	Forward	ABCC5	e8	CGAAGGGTIGTIGGATCTT	20	60,0	50,0
GSHG0022278_R_i8	Reverse	ABCC5	i8	GAGACTGTCGGAAGGATGG	20	59,7	55,0
GSHG0022278_R_e8-e9	Reverse	ABCC5	e8-e9	ATCCTGAAAATTTGGTCCACTG	22	60,2	40,9
GSHG0022278_R_e8-ae9	Reverse	ABCC5	e8-ae9	CACCAAGCAAGTGGTCCAC	19	60,1	57,9
GSHG0011855_F_e4-e5	Forward	CRNDE	e4-e5	CTGAAGATAAGGAGGTAAACCTG	24	57,7	41,7
GSHG0011855_F_e4-ae5	Forward	CRNDE	e4-ae5	TGAAGATAAGGAGGTGCCACTG	23	62,7	47,8
GSHG0011855_R_e6	Reverse	CRNDE	e6	CATATTTAAACCCTCGAGCACT	23	57,6	39,1
GSHG0011481_F_e11	Forward	GAS8	e11	GTCTGGCTGCCITTAACCT	20	60,8	60,0
GSHG0011481_R_e13	Reverse	GAS8	e13	TGATGGTGTCTTCTTCGAC	20	59,8	50,0
GSHG0011481_R_e12	Reverse	GAS8	e12	CAIACCACAGGGTITGTCAG	20	60,0	55,0
GSHG0020243_F_e7	Forward	GUSBP11	e7	TGCGTGTGTCAGTITCTTTA	20	60,7	45,0
GSHG0020243_F_e6-e8	Forward	GUSBP11	e6-e7	ATAAGCAGTTCAGGCCAAG	19	59,0	52,6
GSHG0020243_R_e8	Reverse	GUSBP11	e8	AGGTGGGACTTCCTTCCTTC	20	59,5	55,0
GSHG0019146_F_e8	Forward	UQC	e8	AGTCCGAATGAAGCAGGAGG	20	59,4	50,0
GSHG0019146_R_e9	Reverse	UQC	e9	ATCCCAAGATCGCTGCATAG	20	60,2	50,0
GSHG0019146_R_i8	Reverse	UQC	i8	ITCCAGGCACAGCATCTGACA	20	60,4	50,0

# Example of a RNA-Seq GenoSplice project: Validations

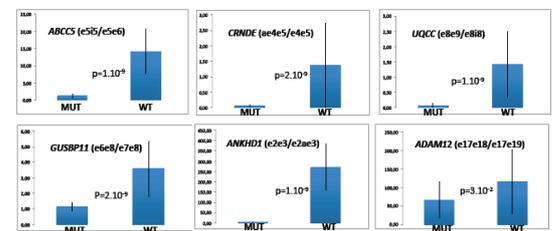


## ► Validation: 100% of true positive at the splicing level

- All events tested by RT-qPCR on an independent cohort of 74 patients from Curie Institute were validated;
- Different types of alternative patterns.



Gene symbol	Gene name	Possible alternative event (HTA2/Geno Splice EASANA)
ABCC5	ATP-binding cassette, sub-family C (CFTR/MRP), member 5	Retention of intron 5
CRNDE	Colorectal neoplasia differentially expressed (non-protein coding)	Alternative acceptor site (exon 4)
UQCC	Ubiquinol-cytochrome c reductase complex chaperone	Alternative terminal exons
GUSBP11	Glucuronidase, beta pseudogene 11	Cassette exon 7
ANKHD1	Ankyrin repeat and KH domain containing 1	Alternative acceptors site (exon 3)
ADAM12	ADAM metallopeptidase domain 12	Alternative terminal exons (exon 18 vs. exon 19)



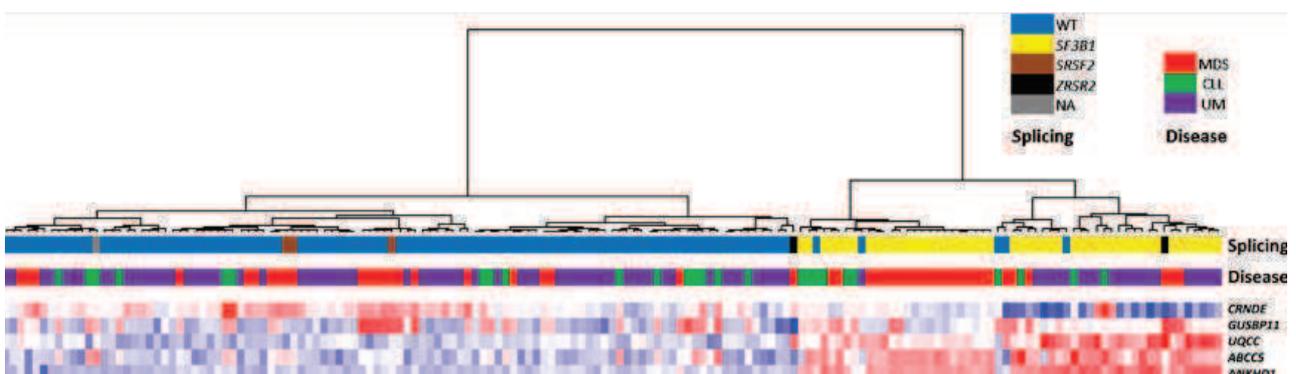
## Application of splicing signature in other pathologies

### ► Three kinds of patients

- 26 patients with chronic lymphocytic leukemia (CLL);
- 48 patients with myelodysplastic syndromes (MDS);
- 87 patients with uveal melanoma (UM).

### ► Signature based on five splicing forms;

### ► Better patient classification according to SF3B1 status rather than pathology



1. GenoSplice: A few words about the company
2. RNA-Seq: Several levels of analyses (and complexity)
3. Focus on RNA-Seq splicing analysis: The GenoSplice strategy
4. Example from a GenoSplice project
5. **Conclusion**

## Conclusion

- ▶ **GenoSplice is able to analyze/integrate all kinds of omics data;**
- ▶ **GenoSplice is specialized in expression data, and in particular in RNA-Seq data;**
- ▶ **GenoSplice provides dedicated bioinformatics services with strong track records (respect of delay, detailed methods and clear presentation of results).**

## Some comments from GenoSplice clients to conclude

"I fully recommend GenoSplice for bioinformatics analysis of array experiments [...] the service by GenoSplice is highly cost-effective, as I do not have to hire and train a bioinformatician for the analysis. GenoSplice allows me to concentrate on the biological question, rather than the technique of data analysis"

**Stefan Stamm, University of Kentucky** 🇺🇸

"Congratulations to Genosplice for developing this easy-to-use platform that is far superior to other web-based methods of splicing analysis"

**David Elliott, University of Newcastle** 🇬🇧

"We were very satisfied with GenoSplice services: [...] a very good interface to evaluate results and very helpful assistance for analysis from experts in the splicing field"

**Juan Valcarcel, CRG** 🇪🇸

"With GenoSplice I found more than a company but a strong collaboration that allowed me to use state of the art approaches to analyze gene expression"

**Sandrine Humbert, Institut Curie** 🇫🇷

"The follow up of the project ensured by GenoSplice is also excellent on both professional [...] and scientific side"

**Françoise Bachelier, Institut Pasteur** 🇫🇷

"FAST DB® from GenoSplice is a very useful resource concerning alternative splicing in human and mouse. Having such a high-quality database greatly helps researchers that want to have access to a comprehensive and well organized compendium of information on such an important facet of the transcriptome"

**Amos Bairoch, SIB** 🇨🇭



29

2014/10/15

## Contacts

**[www.genosplice.com](http://www.genosplice.com)**

**Email: [contact@genosplice.com](mailto:contact@genosplice.com)**

**Phone: +33 1 572 747 52/53**



IUH - Hôpital Saint-Louis  
75010 Paris, France  
(activity address)



ICM - Hôpital de la Pitié-Salpêtrière  
75013 Paris, France  
(activity address)



Genopole® Entrepise,  
Evry, France  
(legal address)



1301

2014/10/15



# Truelab™: Point of Care Molecular Diagnostics

Chandrasekhar Nair

[bc@bigtec.co.in](mailto:bc@bigtec.co.in)

**bigtec Labs**  
*Enabling better Medicine*



# Truelab™: Point of Care Molecular Diagnostics

Chandrasekhar Nair  
[bc@bigtec.co.in](mailto:bc@bigtec.co.in)

INAE-NATF Seminar on Technology and Health-Care,  
Oct 2014

## Acknowledgements

- INAE – NATF for inviting me to share my thoughts
- My fantastic cross-disciplinary team of highly motivated scientists
- Government of India's support (NMITLI-CSIR, BIPP-DBT, ICMR, NPMASS) & Prof. Venkataraman (IISc)
- My very tolerant colleagues, investors and well wishers

# About bigtec Labs

- Founded in 2000
- HQ – Bangalore, India
- State of the Art, ~20,000 sqft R&D facility in Rajaji Nagar, Bangalore
- Skills
  - Engineering
  - Biology
  - Chemistry
  - Product Development



Located at Golden Heights,  
Rajaji Nagar

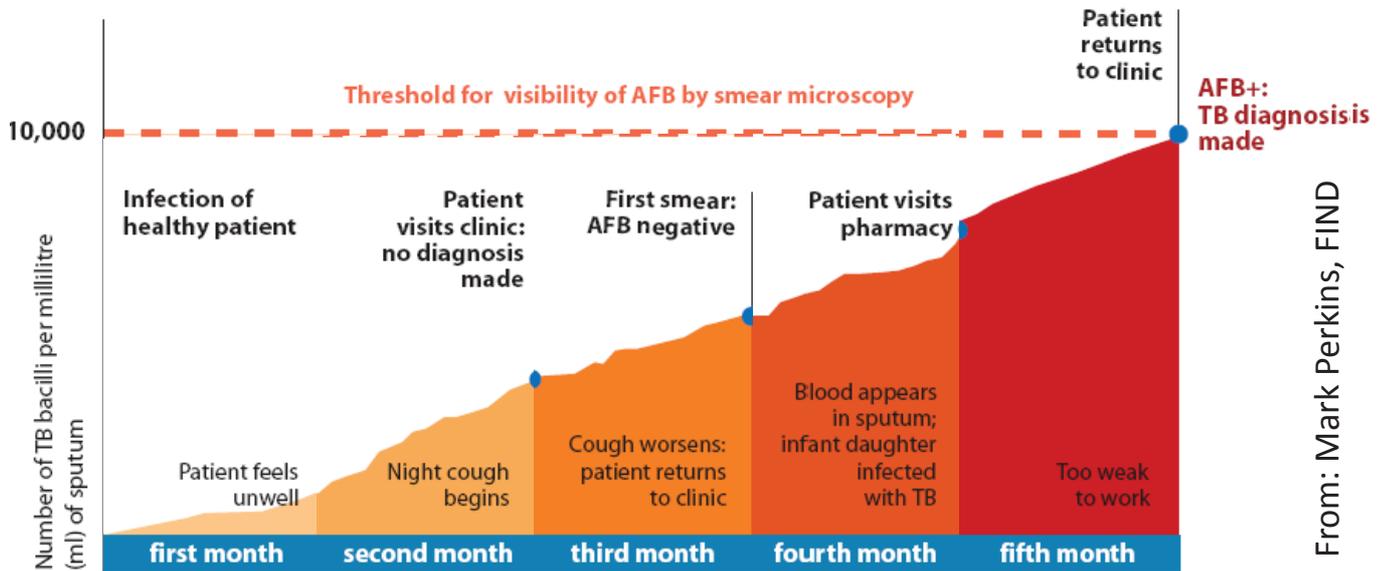
bigtec Confidential

## microPCR -What did we set out to do

- Make the PCR platform for pathogen detection available at near-care settings by making this
  - Fast: Sample to result in < 1 hr
  - Portable by Miniaturizing the PCR
  - Battery operated
  - Easy to use by a person with minimal or no training

# Why?

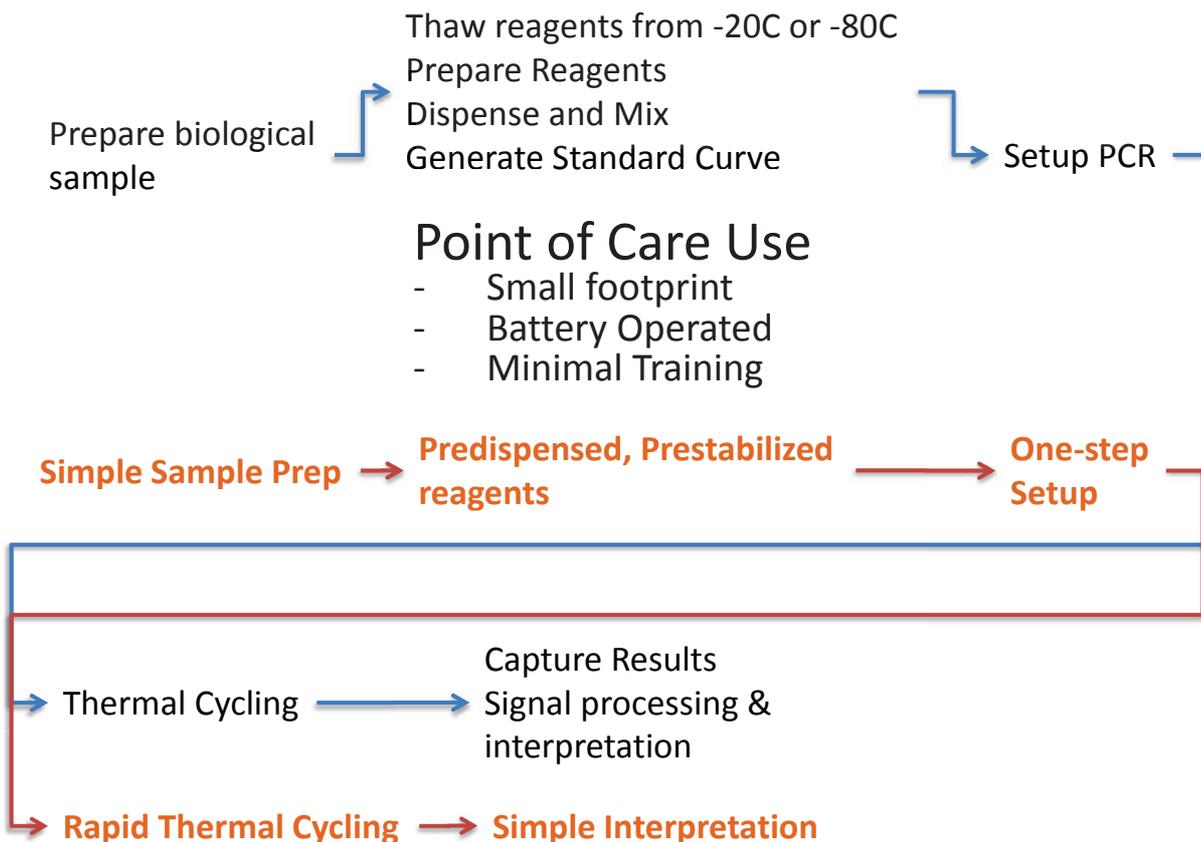
**Smear**  
**Culture**  
**Immunological tests**



Tuberculosis, an example of an infectious disease that can really use a rapid, sensitive, specific diagnosis

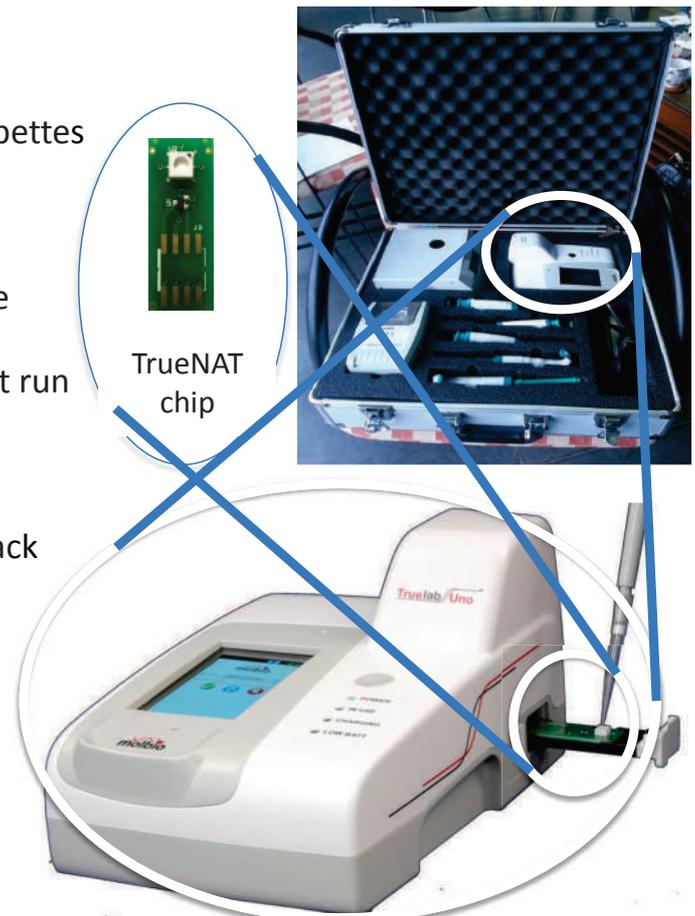
bigtec Confidential

# What did we need to do?



# Truelab™ Workstation

- Complete -Truelab, Trueprep, Printer, Pipettes
- Portable – Light weight, rugged
- Carry case - Sample to result - any where
- No setup/calibration- Power on and start run
- Minimal training, easy to use
- Runs on AC / Battery power ~ 8 hours back
- Real-time detection and Data transfer
- Global IP



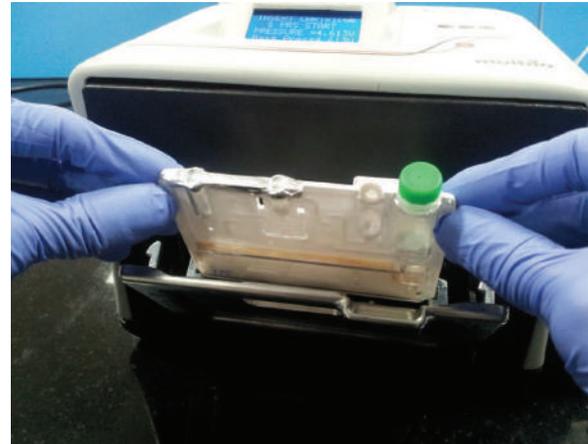
# Trueprep-MAG

- Samples are prepared through a simple menu-driven process
- User adds buffers and aspirates waste as per screen instructions



# Trueprep Auto

- Fully automated sample handling done on disposable cartridge
- All waste contained in dump chamber of cartridge
- Bio-safe design minimizes sample handling and user training
- User just needs to add sample to cartridge and insert cartridge into device
- Full process internal control validates each run
- Totally hands-free, usable in minimally equipped lab



# Truenat™ MTB

- Single copy genomic MTB specific region
- Result as CFU/ml in sputum
- Sample: 500 uL raw sputum
- Positive control DNA is co-extracted to validate the test.

Truenat™ MTB			
Center	molbio		
Date	Friday 24 August 2012 08:40:16		
Operator	Satheesh		
Profile	MTB		
Lot	12345	Expiry Date	0912
Sample	Sputum		
Patient Details			
Name	Gopal Mohan		
ID	PN5063		
Age	45	Sex	Male
Referred By	Divakar Kohli		
Result			
Control C <sub>t</sub>	27.2	Test C <sub>t</sub>	24.0
Run Status	Valid		
MTB	DETECTED 1.1x10 <sup>6</sup> CFU/ml		
Print	SMS	Share	Back

Result screen: Representation only

# MTB: Clinical evaluations

- Study 1:
  - Sensitivity: S+C+: 99.12 % & S-C+: 75.86 % ( Archived panel, n=226)
  - *Published in PLOS one*
- Study II: Truelab vs GeneXpert

		GeneXpert		Total
		POS	NEG	
Truelab	POS	87	3	90
	NEG	1	27	28
Total		88	30	118

- PD Hinduja Hospital and Research Centre, Mumbai, India
- Published in IJMYCO

bigtec Confidential

## Use of Truelab at Point of Care

- Indian Government Tuberculosis Control Program
  - 13,000 Designated Microscopy Centres (DMCs) :7.5 M suspects screened with Smear, 630,165 cases
  - Smear sensitivity ~ 50%
  - 37 accredited Culture and DST labs : 366,381 cases
- Truelab
  - ~ 255,000 cases of S-C+ cases could be identified ***at the DMC at first contact***
- Private Lab Testing ~ 11 M pulmonary TB suspects

# Potential Impact

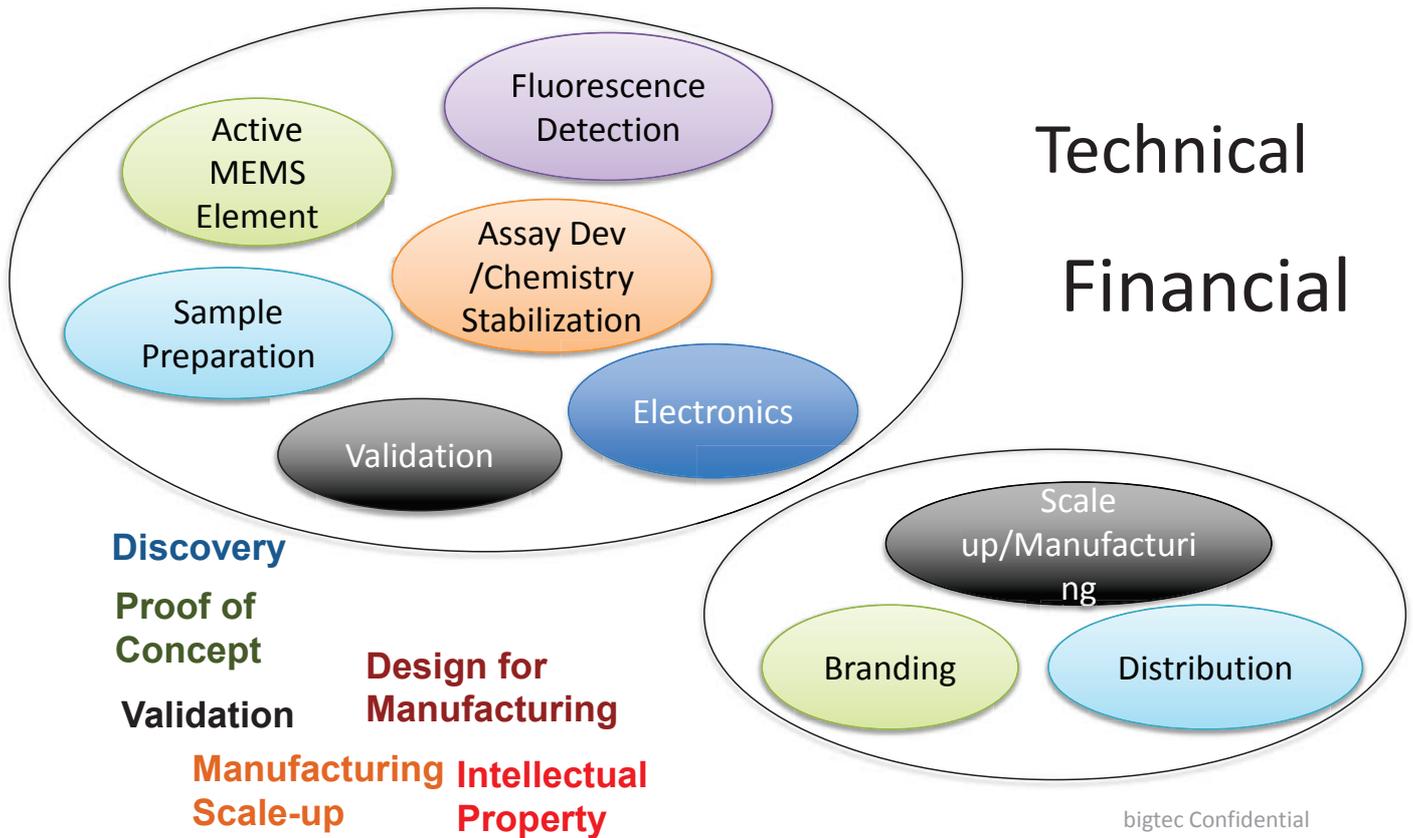
- Time to positive (TTP) could be drastically reduced
- Platform can be used for active case finding
- A battery operated, field usable molecular test like Truelab is ***the only way*** to bring down the very slowly declining TB incidence in India and other resource limited countries



## Validations: Commercial Assays

Disease	Validation centre(s)	Reference test	No. of samples	Result
Hepatitis B	AIIMS, St. Johns, CMC	PCR Viral load assays that are FDA/CE marked	400	>95% sensitivity
TB	Hinduja, CMC, NIRT	Smear, Culture, CRS	800	99% sensitivity in S+C+ samples
Malaria	NIMR	Smear, RDT	300	99% sensitivity for P. fal and P. vivax
H1N1 (Swine Flu)	KMC, NCDC, NIMHANS, NIV	Reference PCR, RDT	500	~95% sensitivity
Dengue/ Chikungunya	NIMHANS	Reference PCR	150	99% sensitivity

# Ecosystem



## Truelab™ @ point of care: True impact in health care



Truelab workstation being used for Malaria (*P. fal*) study at Gobei Health Centre, Kenya  
- Organized by UMN, USA & KEMRI, Kenya



[www.molbiodiagnostics.com](http://www.molbiodiagnostics.com)

A joint venture company between bigtec Labs, Bangalore & Tulip Group of companies, Goa.

- ▶ IVD manufacturing experience of 22 years
- ▶ ISO 9001 and ISO 13485 certified facility at Goa
- ▶ Marketing footprint in 70 countries

Thank You



Centre National de  
Génotypage  
Missions,  
Technologies, Projects

Deleuze, Ph.D  
Head of  
Centre National de Génotypage





DE LA RECHERCHE À L'INDUSTRIE



Jean-François Deleuze, Ph.D  
Head of Centre National de Génotypage

www.cea.fr

## Centre National de Génotypage Missions, technologies, projects

INAE/NATF Seminar on "Technology and Health-Care"  
September 16th, 2014  
Evry Genopole



## Plan

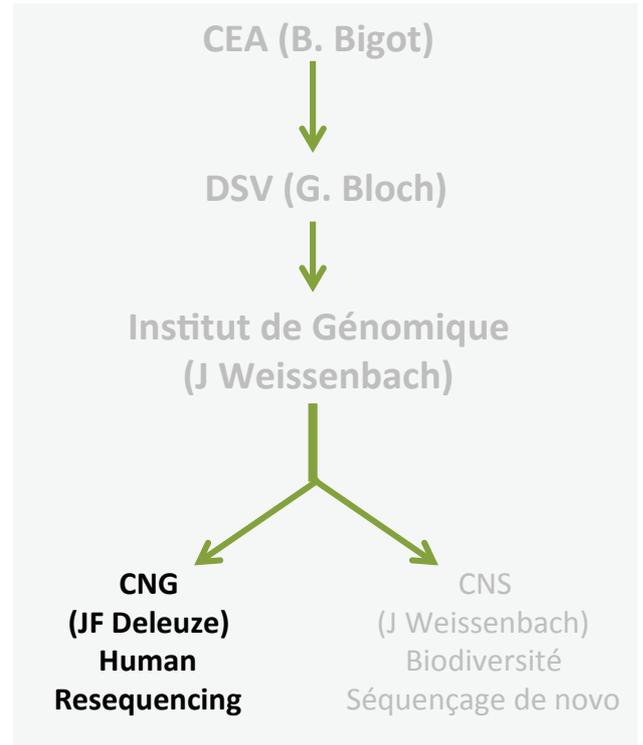
- ❑ Missions
- ❑ Technologies
- ❑ Projects with focus on potential collaboration with India

# Already together on the road!?



## Plan

- Missions
- Technologies
- Projects with focus on potential collaboration with India

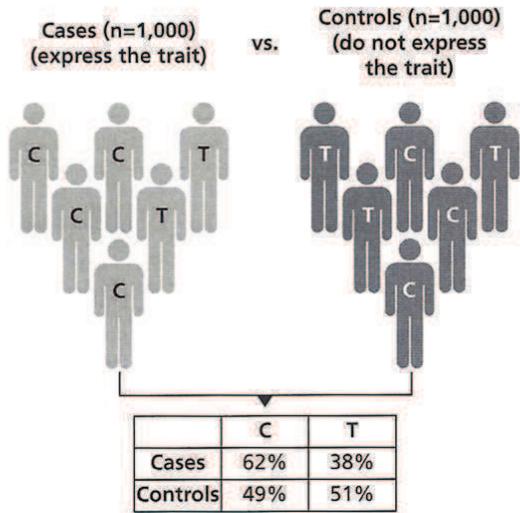


## CNG Missions

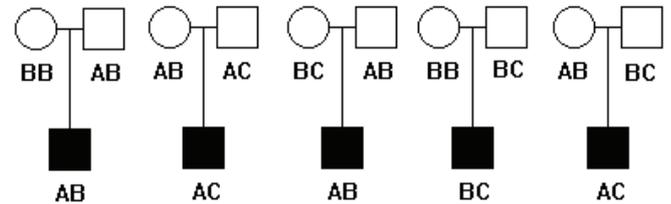
- ❑ Participate to the discovery and understand 1) the « genomic » causes of **human** diseases and 2) relevant biomarkers, to allow the development of innovatives personalized therapeutic approaches in unmet medical needs by providing **best in class tech platforms (wet and *in silico*)** for **collaborative** and CNG own projects.
- ❑ Participate to the Genomic Medicine revolution of the 21st ce
  - ✓ Genomic information available to all (diagnostic and beyond)
  - ✓ Personalized medicine, personalized health management



## Case-control study for genetic association



## Family based association studies



Qualitative and quantitative information!

$$(P) = (G) + (E)$$

Phenotype (P) = Genotype (G) + Environment (E)



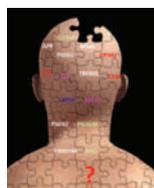
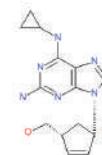
CFTR $\Delta$ F508

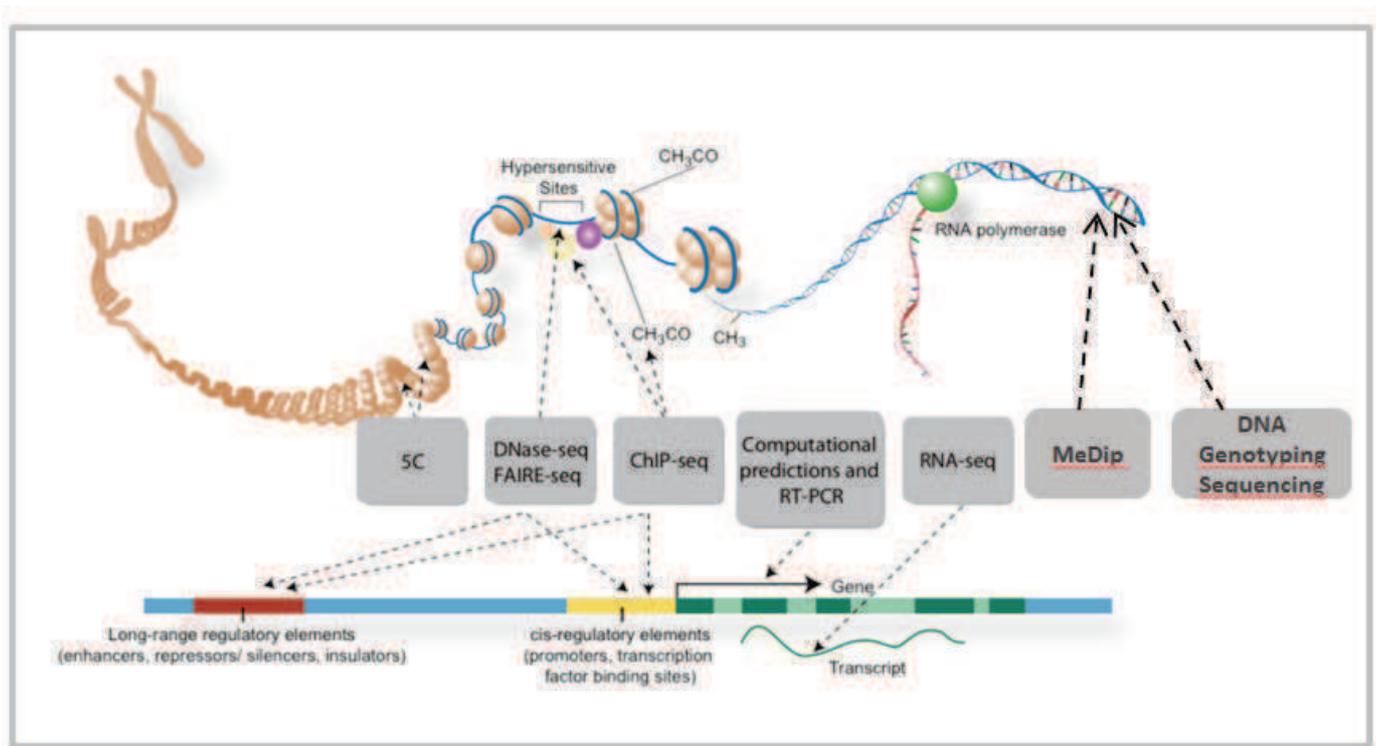


Chrna5



HLA B\*5701

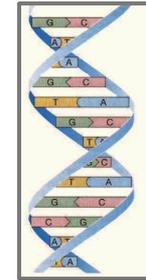
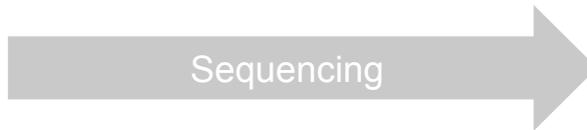
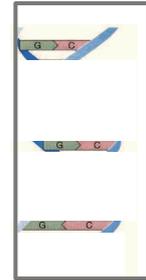
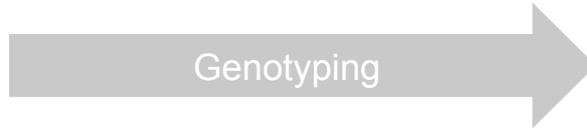
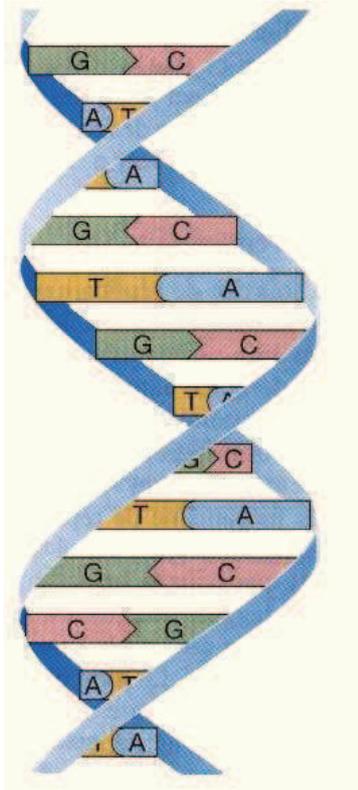




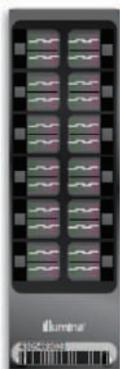
## Plan

- Missions
- Technologies
- Projects with focus on potential collaboration with India

# Tools for genomic analysis



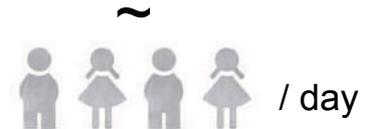
# European Leader in High Throughput Genotyping



OMNI5	4.5.10 <sup>6</sup> pan genomic SNPs (~85%)
CoreExome	265K TagSNPs + 245K exonic SNPs (~50%)
ExomeCHIP	240K Exonic SNPs (~10%)
Omni Express	715K tag SNPs (60%)
Human CH3	450 000 methylation sites



6600Gb (raw data) in 2 weeks  
 440 Gb per day  
 30x Whole Genome = 100 Gb



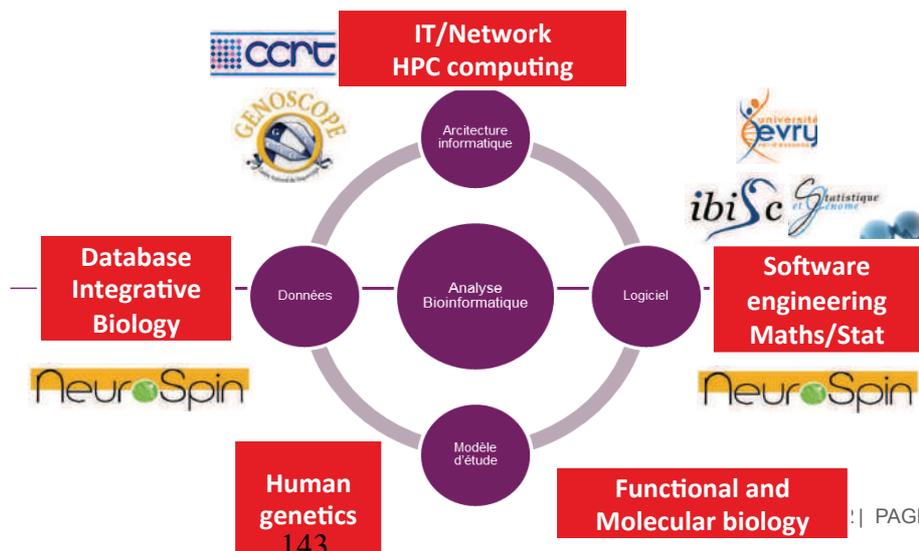
NGS@ CNG: a generic readout to illuminate the variety of genomics variations:  
 Exome, WG, RNA, Met.....

## • Missions

- Analytical expertise to support a variety of NGS based technologies
- Innovation for polymorphism studies (exome variant server, WGS...)
- Support translational medicine objectives of the Center

## • Competences

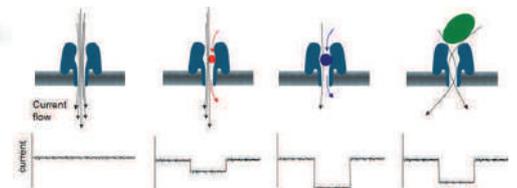
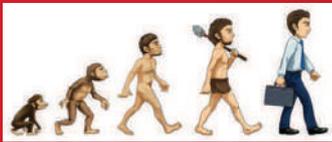
- Bioinformatic network
- Statistical genetics
- Data integration (provides technology and data)
- Bioinformatics development (conception/implementation)



- 20th in Nov'13 Top 500 ranking
- Total capacity >120 000 computing cores
- ~ 2-3 Petaflops
- Simulation

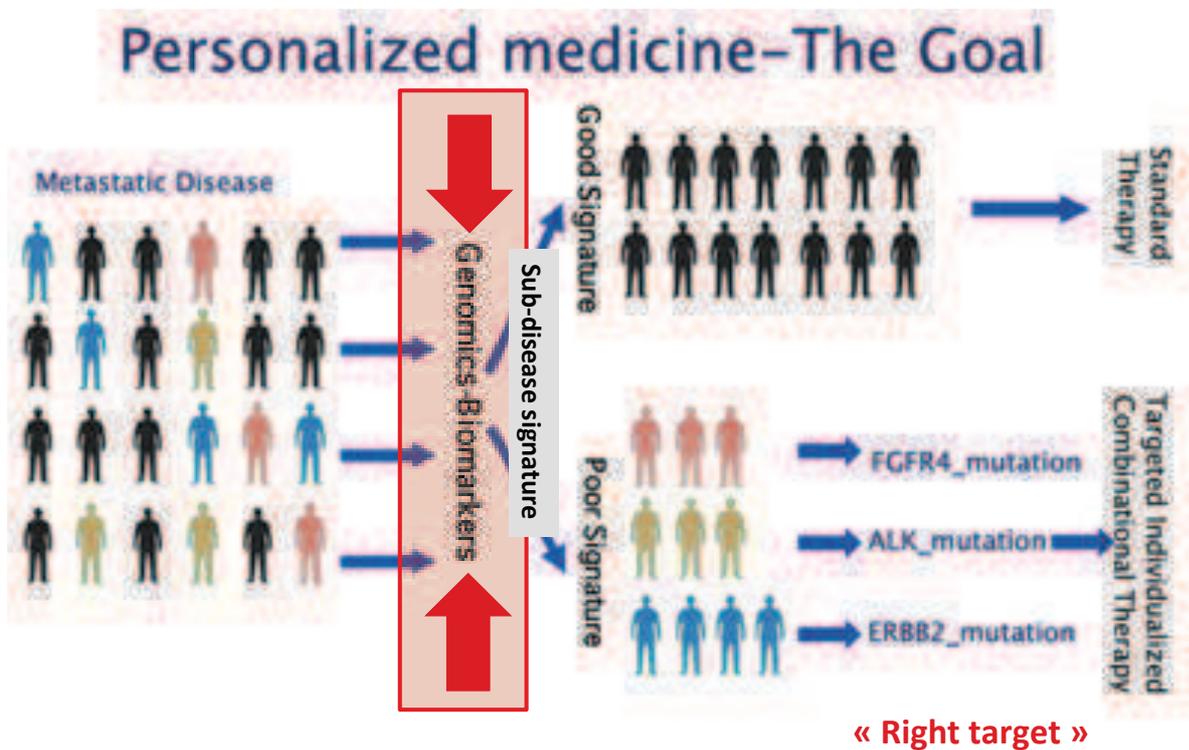
### Dedicated to Genomics

- 3000 cores
- 5 petabytes
- Exome and WGS analysis
- Initial productivity gain: x10

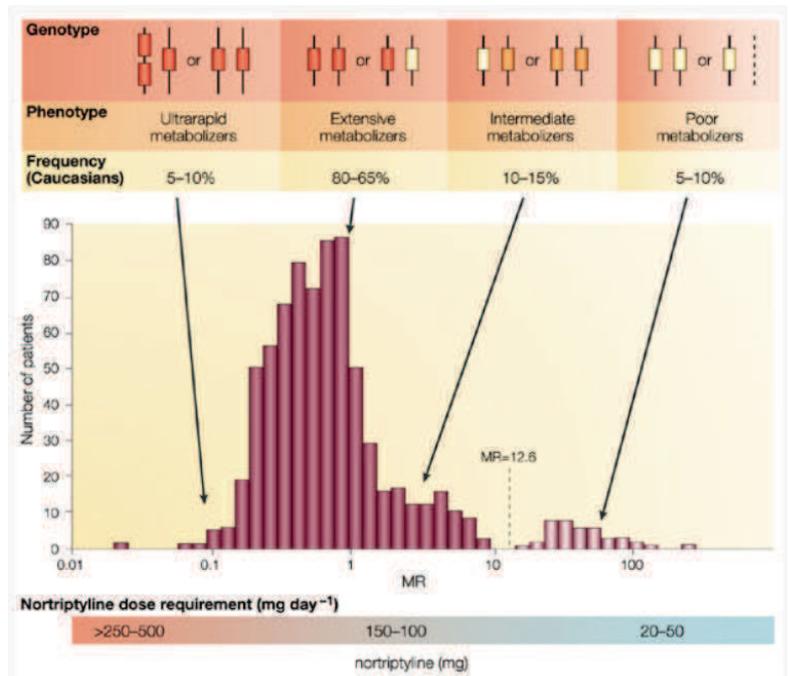
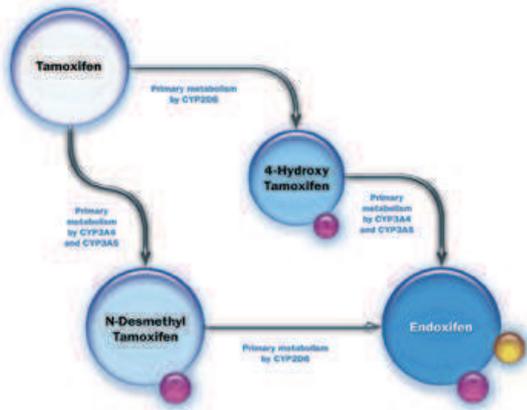


- Missions
- Technologies
- Projects with focus on potential collaboration with India

« The right drug at the right dose to the right patient »



# Drug metabolizing enzymes are polymorphic



# Genotypic-based drug dosage

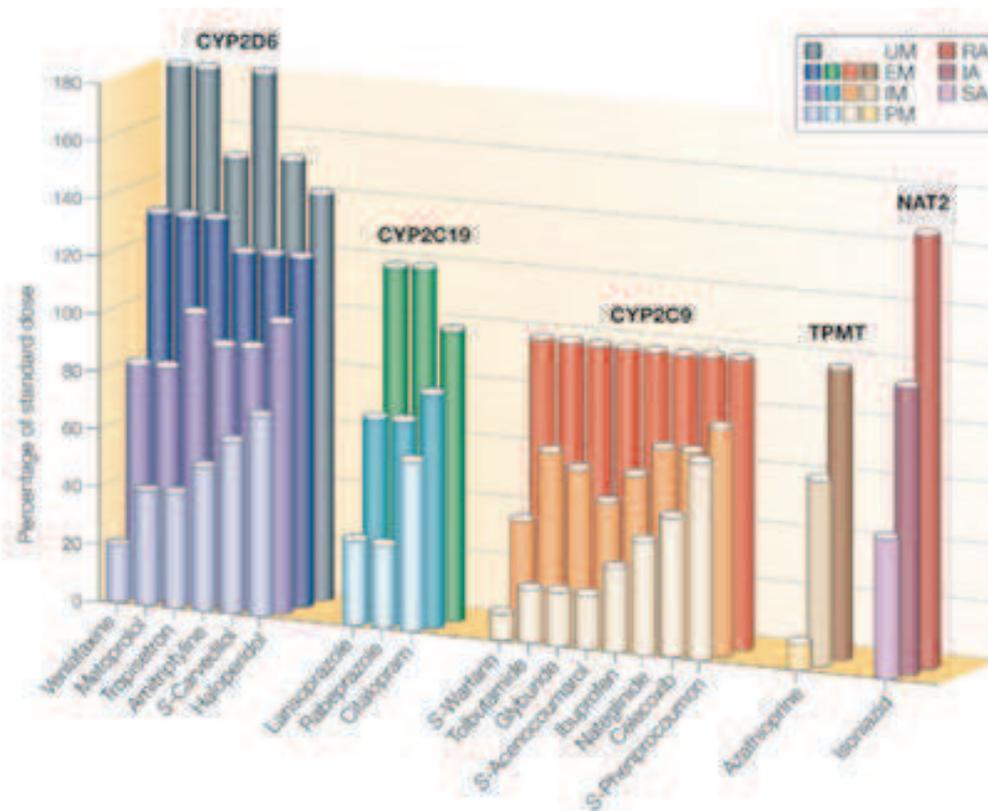


Table 1 Continued

	Substrates	Variant alleles	Alteration in function	Allele frequencies (%)				
				White	Black	Asian	Chinese	Japanese
ABC81 (P-gp) <sup>g</sup> (ref. 26)	Digoxin Cyclosporin Loperamide Verapamil Quinidine	*1	Unknown	15	15	15	—	—
		*13	Unknown	32	5	27	—	—
		*26	Unknown	10	9	5	—	—
		*21	Unknown	3	8	—	—	—
		*11	Unknown	1	2	23	—	—
ABC22 (BCRP) (ref. 27)	Topotecan Irinotecan Mitoxantrone Doxorubicin Rosuvastatin Methotrexate	34G>A	Reduced	2	4	45 <sup>g</sup>	20	15
		421C>A	Reduced	14	0	15 <sup>g</sup>	35	35
SLCO1B1 (OATP1B1) (refs. 28–32)	Pravastatin Rosuvastatin	*1b (388A>G)	Increased (possibly protein expression)	30, 30–51, 38	74, 75, 77	57–88	59.9	46.9; 53.7, 63–67
		*2	Unknown	2	0	—	—	0
		*4	Unknown	16	2	—	—	0
		*5 (521T>C)	Decreased	14; 2, 4 <sup>g</sup>	2	—	—	0.7
		*6	Unknown	2	0	—	—	—
		*7	Unknown	1	0	—	—	—
		*8	Unknown	1	0	—	—	—
		*9	Unknown	0	9	—	—	—
		*10	Unknown	2	0	—	—	—
		*11	Unknown	2	34	—	—	—
		*15 (both 388A>G and 521T>C)	Decreased	2, 7 <sup>g</sup>	—	—	14	3.7; 10.3
UGT1A1 (refs. 33–35)	Irinotecan (SN38)	*6 (211G>A)	Homozygous Reduced	0	0	—	—	4
			Heterozygous	3.3	0	—	—	23
		*27 (229C>A)	Reduced	0	0	<1–3	—	—
		*28 (TA <sub>7/7</sub> )	Reduced	12; 13	23	5	8	2
		*28 (TA <sub>6/7</sub> )	Reduced	39	—	20	14	—
UGT2B7 (ref. 36)	Morphine Zidovudine	*1*1	None	25	—	—	—	43
		*1*2	None	43	—	—	—	37
		*2*2	Reduced transcription	23	—	—	—	4
VKORC1 <sup>h</sup> (refs. 13,37)	Warfarin	-1639G>A	AA	14.2	—	—	82.1	—
			AG	46.7	—	—	17.9	—
			GG	39.1	—	—	0	—
		1173C>T	CC	37.5 <sup>g</sup> ; 34	80.4	—	<1	—
			CT	50 <sup>g</sup> ; 49	18.7	—	14	—
	TT	12.5 <sup>g</sup> ; 17	0.9	—	85	—		

[Login](#) [Blog](#) [Contact Us](#) [f](#) [t](#) [in](#) [s+](#)

[Company](#)

[Genetic Assessment](#)

[Health Signals](#)

[Knowledge Center](#)

## CLOPIDOGREL

**Introduction**

Why you need this product:

Information for Patients:

Information for Doctors:

Benefits to Customer:

How this assessment works?

Revalidation

FAQs

Blood clot formation due to excessive 'bad cholesterol' results in narrowing or blockage of the artery. This leads to permanent organ damage as the organ is now deprived of much needed oxygen. In order to prevent organ damage it is vital to inhibit platelet clumping. Clopidogrel, an anti-platelet drug prevents platelets sticking to the plaque and their resultant adverse after effects.

Clopidogrel is used for the treatment of:

- Coronary Artery Disease (Heart Disease)
- Before and after Coronary Artery Stenting
- Cerebrovascular Disease (Stroke) and
- Peripheral Artery Disease

Clopidogrel has to be converted into its active form for its anti-platelet activity. This conversion is done by the enzyme Cytochrome P450. The activity of this enzyme is controlled by the gene CYP2C19. If an individual has genetic variations in CYP2C19 gene, Clopidogrel is not converted into its active form. Hence such a person, despite taking Clopidogrel in the dosage prescribed will continue be at a risk of platelet induced organ damage such as heart attack or stroke and even death. Did you know that 30-40% of the Indian populations are at a high risk of such adverse cardiac events because of these genetic variations?

# How a statin might destroy a drug company?

## THE LANCET

Volume 361, Number 9360

### How a statin might destroy a drug company

Bayer, the German drugs giant, is backed against the ropes because of compensation claims for cerivastatin. In the face of massive potential losses, the company's share price fell last week to €8.53 from a high in the past year of £29.07. Bayer believes the market is over-reacting. The company placed advertisements in several German national newspapers at the end of February. "Fakten statt Stimmungsmache" headlined the advert, which translates as "Facts rather than spin". The advert continues: "Because we can again comment, we would like to immediately bring facts into the forefront. We very much hope that we—especially in the interests of our shareholders—can thus counteract the current upheaval in the share market."

Bayer's woes arose because cerivastatin, launched as the sixth statin in the marketplace at the end of 1997, became linked with serious myopathy, severe enough to lead to rhabdomyolysis and death, especially when given with a fibrate such as gemfibrozil to lower triglycerides. Cerivastatin was launched at lower doses, and Bayer subsequently sought and gained approval for doses of 0.4 and 0.8 mg. The company voluntarily withdrew the drug in the USA on Aug 8, 2001. By then, the drug had been linked to over 100 deaths from rhabdomyolysis. In a letter in February, 2002, that reported a study of prescription monitoring and adverse-reaction reporting (*N Engl J Med* 2002; 346: 539-40), scientists at the US Food and Drug Administration (FDA) said that the rate of fatal rhabdomyolysis with cerivastatin was 16-60 times higher than for any other statin (31 deaths with cerivastatin vs 42 with all five other statins). When cases of concomitant use of cerivastatin with gemfibrozil or lovastatin were excluded, the rate was still 10-50 times higher. With cerivastatin alone, six investigated deaths occurred after use of 0.4 mg and 12 with 0.8 mg.

There are about 7800 claims for compensation in the USA and about 900 in Germany. About 450 of the US cases have been settled by Bayer without coming to trial and without the company admitting liability. The first case came to trial in Corpus Christi, Texas, last week, and Bayer was immediately in deep

water in court. The company had sent a letter to over 2000 local residents, reminding them that it employs nearly 2000 people in Texas and contributes about US\$185 million to the state's economy in payroll, taxes, and support of local groups. In its letter, Bayer lumped all statins together when discussing the potential for serious side-effects in a "small segment" of the population, but failed to mention the FDA comparative data for rhabdomyolysis and statins, alone or with gemfibrozil—which, of course, would show cerivastatin outlying the other drugs in the class. The judge said the sending of the letter was "outlandish", and has deferred a decision on how to deal with Bayer until the case on trial is over. One of Bayer's lawyers apologised in court and said the letter was a mistake; it was meant to go only to the town's chamber of commerce.

Bayer intends to fight the compensation claims vigorously. It said the action in Texas and a class action that is forming in Minnesota quoted internal company documents out of context, and it will put in the context later in the hearings. Bayer insists that it acted properly and in a timely manner when informing regulatory authorities about myopathy with cerivastatin. Bayer also seeks to blame prescribers for ignoring labelling changes—i.e. giving gemfibrozil with cerivastatin when this became contraindicated and not starting patients on a low dose.

If Bayer is found guilty of negligence in the marketing and timing of its eventual withdrawal of cerivastatin, the company may founder. But a strong general signal will then be sent to the drug industry, which is especially pertinent for the statins. These drugs are highly effective in the prevention of heart attacks—indeed, some experts advocate much wider use of statins. But the message following the cerivastatin disaster is of the law of unintended consequences. In the enthusiasm for wider use of a class of drugs, all must remember that rarer side-effects are unlikely to be seen in clinical trials before a drug is approved. Post-marketing surveillance can teach salutary lessons about the need for caution when new drugs are promoted to physicians.

The Lancet

THE LANCET • Vol 361 • March 6, 2003 • www.thelancet.com

793

For personal use. Only reproduce with permission from The Lancet Publishing Group.

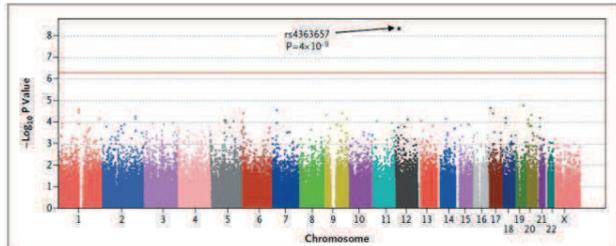
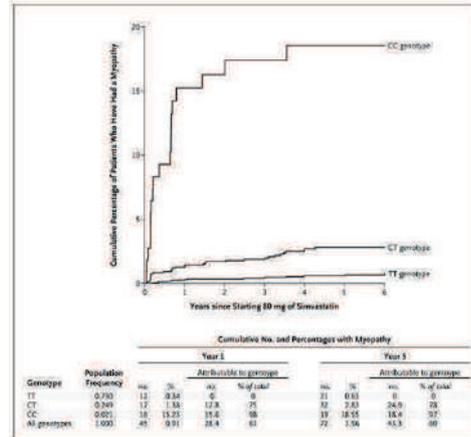
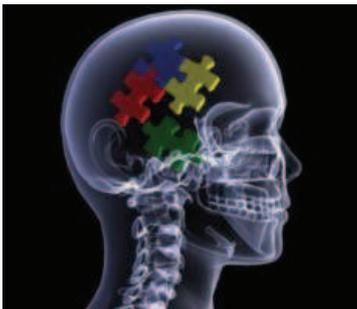


Figure 1. Results of Tests for a Trend in the Association between Myopathy and Each SNP Measured in the Genome-wide Association Study.

P values are shown for each SNP measured among 85 participants with myopathy and 90 matched controls who were taking 80 mg of simvastatin daily. Analyses are based on 316,184 of the 318,237 SNPs (99.4%) on the Sentrix HumanHap300-Duo BeadChip (Illumina). A result above the horizontal red line indicates strong evidence of an association ( $P < 5 \times 10^{-8}$ ).



# Ongoing pharmacogenetic projects



## Optimise (Schizophrenia)

- ✓ Identification of genes involved in response to antipsychotic



## Canto (Breast cancer, LABEX GENMED)

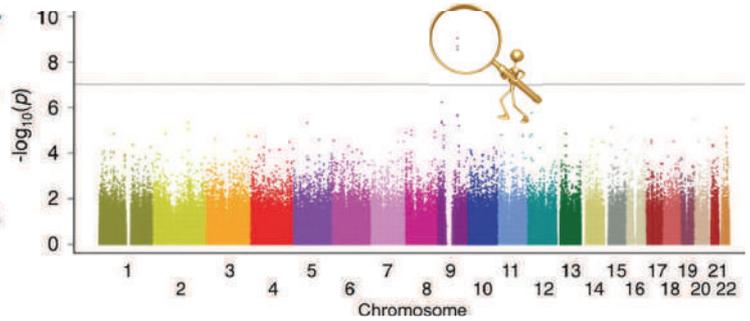
- ✓ Identification of genes involved in response to treatment

*Human Molecular Genetics*, 2010, Vol. 19, No. 12 2516–2523  
doi:10.1093/hmg/ddq123  
Advance Access published on March 29, 2010

## The *FOXE1* locus is a major genetic determinant for radiation-related thyroid carcinoma in Chernobyl

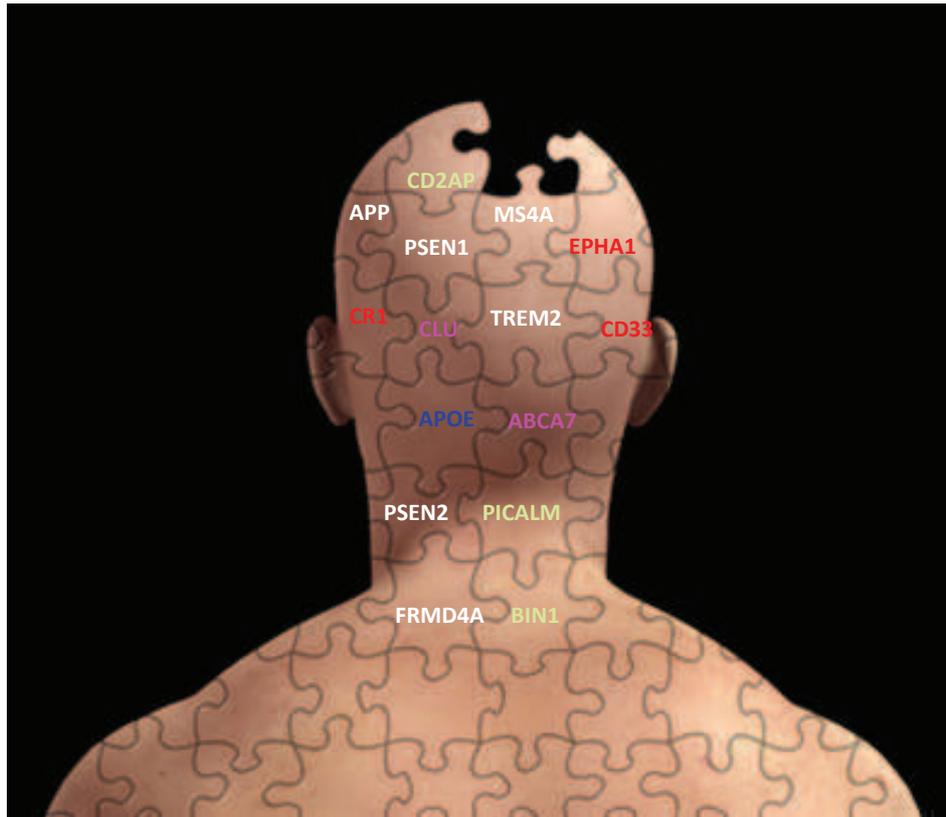
Meiko Takahashi<sup>1,2,1</sup>, Vladimir A. Saenko<sup>3,1</sup>, Tatiana I. Rogounovitch<sup>4</sup>, Takahisa Kawaguchi<sup>1,2</sup>,  
Valentina M. Drozd<sup>5</sup>, Hisako Takigawa-Imamura<sup>1</sup>, Natalia M. Akulevich<sup>4</sup>,  
Chanavee Ratanajaraya<sup>1</sup>, Norisato Mitsutake<sup>6</sup>, Noboru Takamura<sup>4</sup>, Larisa I. Danilova<sup>6</sup>,  
Maxim L. Lushchik<sup>5</sup>, Yuri E. Demidchik<sup>7</sup>, Simon Heath<sup>8</sup>, Ryo Yamada<sup>1</sup>, Mark Lathrop<sup>8,9</sup>,  
Fumihiko Matsuda<sup>1,2,\*</sup> and Shunichi Yamashita<sup>3,4</sup>

<sup>1</sup>Center for Genomic Medicine and <sup>2</sup>Institut National de la Santé et de la Recherche Médicale (INSERM) Unit U852, Kyoto University Graduate School of Medicine, Kyoto 606-8501, Japan, <sup>3</sup>Department of International Health and Radiation Research and <sup>4</sup>Department of Molecular Medicine, Atomic Bomb Disease Institute, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki 852-8523, Japan, <sup>5</sup>Department of Thyroid Disease Research, <sup>6</sup>Department of Endocrinology and <sup>7</sup>Belarusian Medical Academy for Postgraduate Education, Minsk 220013, Republic of Belarus, <sup>8</sup>Centre National de Genotypage, Institut Génomique, Commissariat à l’Energie Atomique, Evry 91000, France and <sup>9</sup>Fondation Jean Dausset-CEPH, Paris 75010, France



- FOXE1 is a a thyroid-specific transcription factor with pivotal roles in thyroid morphogenesis





- Immune system
- Lipid processing
- Endocytosis

## CNG Portfolio

### □ National

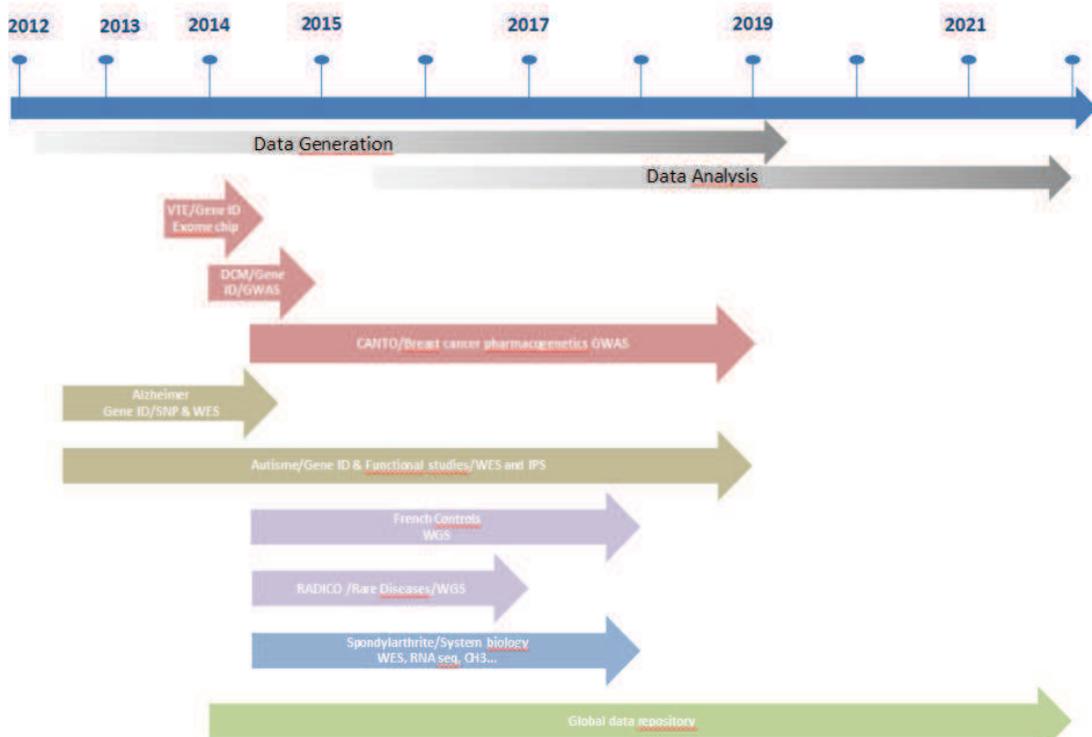
- ✓ France Génomique (13 projects among which the « French Exome Project »)
- ✓ Labex GENMED (8 projects among which the « French Whole Genome Project »)
- ✓ Other numerous collaborations (~ 50)
- ✓ Own projets (mitochondria, metagenomics (Microbiome+Gwas...))...

### □ European



- Single cell sequencing in neurone (neurodegeneration)
- Healthy Aging (centenarians and above)

# LaBex GENMED to build the foundation of medical genomics in France



## GENMED at a Glance

# Recent Collaborative Results

nature genetics

## Seven new loci associated with age-related macular degeneration

**Age-related macular degeneration (AMD)** is a common cause of blindness in older individuals. To accelerate the understanding of AMD biology and help design new therapies, we conducted a collaborative genome-wide association study, including 177,788 advanced AMD cases and 240,000 controls of European and Asian ancestry. We identified 19 loci associated with  $P < 5 \times 10^{-8}$ . These loci show enrichment for genes involved in the regulation of complement activity, lipid metabolism, retinal cellular matrix remodeling and angiogenesis. Our results indicate novel loci with associations reaching  $P < 5 \times 10^{-8}$  for the first time, near the genes *CELF4*, *CEP350*, *MEIS1*, *UCP2*, *UCP3*, *PCSK9*, *PCSK2*, *ADAMTS1* and *ESRRA2*. A genetic risk score combining 19p genotypes from all loci showed similar ability to distinguish cases and controls in all samples sequenced. Our findings provide new information for biological, genetic and therapeutic studies of AMD.

AMD is a highly heritable, progressive neurodegenerative disease that leads to loss of central vision through death of photoreceptors<sup>1</sup>. In developed countries, AMD is the leading cause of blindness in those over 50 years of age<sup>2</sup>. Despite the complexity of the disease, a large number of genes have been implicated in the pathogenesis of AMD, with the first gene identified in 2005<sup>3</sup>. In addition, we have found several additional loci<sup>4-14</sup>, each providing an entry point for AMD biology and potential therapeutic targets.

To accelerate the pace of discovery in macular degeneration, we led a research group from across the world to form the AMD Gene Consortium in early 2010, with support from the National Eye Institute of the US National Institutes of Health (Table 1). Supplementary Tables 1 and Supplementary Note 1, in addition to the study of common associated variants, we first conducted a meta-analysis of genome-wide association studies (GWAS) combining data from 177,788 cases with advanced AMD (with geographic ancestry non-Asian) and 240,000 controls. Each study was first subjected to a GWAS quality control (QC) pipeline to reduce any potential study-specific biases<sup>15</sup>, as detailed in Supplementary Table 2, and then entered into the large-scale meta-analysis using standard procedures<sup>16</sup>. Results were combined through meta-analysis<sup>17</sup>, and variants associated with both remaining variants of advanced AMD were genotyped in an additional 14,500 cases and 14,500 controls.

10 of 16 authors and affiliations appear in the end of the paper  
Received 16 May 2012; accepted 7 January 2013; published online 2 March 2013; doi:10.1038/ng.1208

Supplementary Information is available at [www.nature.com/ng](http://www.nature.com/ng).

LETTERS

LETTERS

## Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease

Alzheimer's disease is a devastating neurodegenerative disorder affecting the elderly. The disease manifests with progressive dementia, loss of cognitive functions, leading to loss of autonomy. The APOE ε4 allele is strongly associated with a major genetic risk factor for Alzheimer's disease<sup>1</sup>. Previous GWAS in individuals of European ancestry identified several genetic regions associated with AD<sup>2-10</sup>. Recently, a meta-analysis of 29,839 individuals<sup>11</sup> identified 11 new susceptibility loci for AD. The search for additional genetic risk factors requires large-scale meta-analysis of GWAS in additional genetic populations. The largest GWAS meta-analysis of AD to date was conducted by the International Genetics of Alzheimer's Disease Consortium (IGAD), the Collaborative Cross and Genetically Diverse European (CCGDE) and the Genetic and Environmental Risk in Alzheimer's Disease (GERAD) Consortium (Table 1). Other methods, Supplementary Table 1 and Supplementary Note 1, describe the study design and QC pipeline. We used a genome-wide association study (GWAS) approach to identify new susceptibility loci for AD. We conducted a meta-analysis of 74,046 individuals of European ancestry, including 11,884 cases and 22,162 controls (Supplementary Table 1). We identified 11 new susceptibility loci for AD, each reaching genome-wide significance ( $P < 5 \times 10^{-8}$ ) in the combined study. In addition, we identified 11 new susceptibility loci for AD, each reaching genome-wide significance ( $P < 5 \times 10^{-8}$ ) in the combined study. In addition, we identified 11 new susceptibility loci for AD, each reaching genome-wide significance ( $P < 5 \times 10^{-8}$ ) in the combined study.

10 of 16 authors and affiliations appear in the end of the paper  
Received 10 March 2012; accepted 17 September 2012; published online 27 October 2012; doi:10.1038/ng.1207

Supplementary Information is available at [www.nature.com/ng](http://www.nature.com/ng).

nature genetics

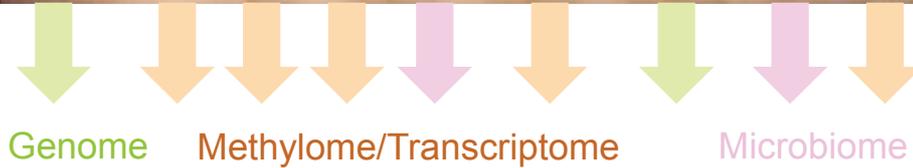
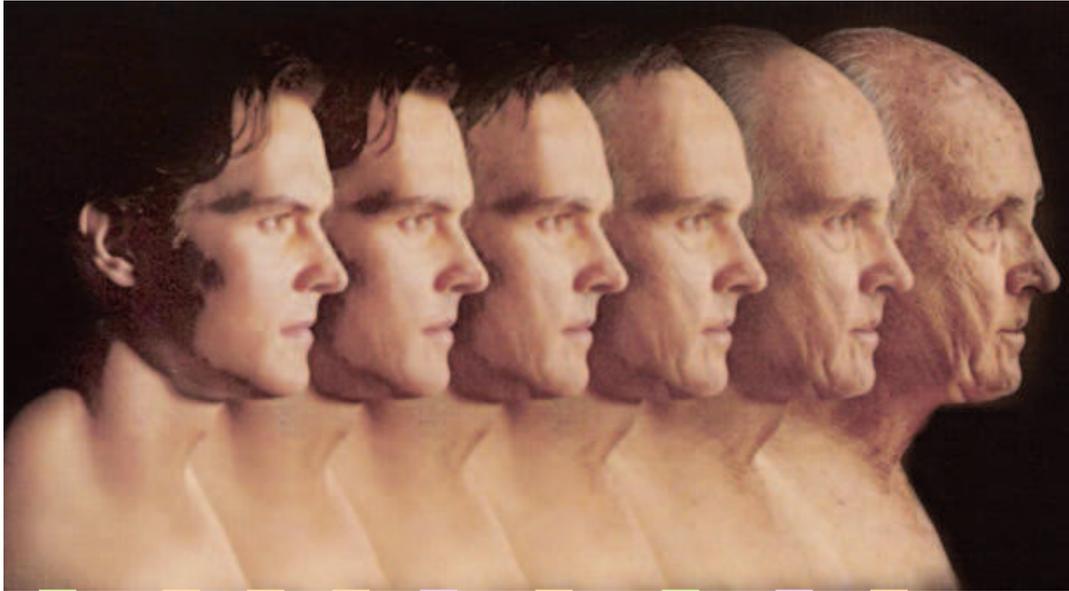
## Mutations in *TUBG1*, *DYNCH1*, *KIF5C* and *KIF2A* cause malformations of cortical development and microcephaly

The genetic causes of malformations of cortical development (MCD) are largely unknown. Here we report the discovery of multiple pathogenic nonsense mutations in *TUBG1*, *DYNCH1* and *KIF2A*, as well as a single genomic region mutation in *KIF5C*, in subjects with MCD. We found a recurrent rearrangement of mutations in *KIF5C*, implying that this gene is a major locus for MCD. We further show that mutations in *KIF5C*, *KIF2A* and *DYNCH1* affect *KIF5C* protein, produce protein folding and microcephaly, respectively. In addition, we show that aggregates of mutant *Kif5c* proteins in mice interfere with proper neurogenesis, whereas expression of altered tubulin proteins in *Saccharomyces cerevisiae* disrupts normal microtubule behavior. Our data reinforce the importance of microtubule-related proteins in cortical development and strongly suggest that microtubule-dependent events and postnatal processes are major contributors to the pathogenesis of MCD.

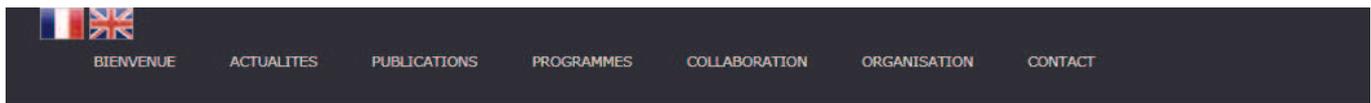
The formation of the complex architecture of the mammalian brain requires coordinated timing of proliferation, migration and layering, as well as differentiation of distinct neuronal populations that migrate long distances from their development sites to their final destinations. The molecular mechanisms underlying these processes are largely unknown. Here we report the discovery of multiple pathogenic nonsense mutations in *TUBG1*, *DYNCH1* and *KIF2A*, as well as a single genomic region mutation in *KIF5C*, in subjects with MCD. We further show that mutations in *KIF5C*, *KIF2A* and *DYNCH1* affect *Kif5c* protein, produce protein folding and microcephaly, respectively. In addition, we show that aggregates of mutant *Kif5c* proteins in mice interfere with proper neurogenesis, whereas expression of altered tubulin proteins in *Saccharomyces cerevisiae* disrupts normal microtubule behavior. Our data reinforce the importance of microtubule-related proteins in cortical development and strongly suggest that microtubule-dependent events and postnatal processes are major contributors to the pathogenesis of MCD.

10 of 16 authors and affiliations appear in the end of the paper  
Received 16 November 2012; accepted 22 March 2013; published online 27 April 2013; doi:10.1038/ng.1213

Supplementary Information is available at [www.nature.com/ng](http://www.nature.com/ng).



Thank you and come visit us at CNG or @ <http://www.cng.fr/>



**L'appel à propositions  
« Grands Projets de Séquençage »**



**Actualités**

**Fête de la science**  
400 visiteurs en provenance de différents départements de la région parisienne ont été accueillis durant les portes ouvertes de l'Institut de Génétique - CNG à Evry. [\[plus...\]](#)

**Collaboration**

L'appel à propositions est ouvert en continu [\[plus...\]](#)





# Opportunities in Bio-Pharmaceuticals Research and Manufacturing in India

Dr. Goutam Ghosh  
Panacea Biotec Ltd., India







**Panacea Biotec**



# OPPORTUNITIES IN BIO-PHARMACEUTICALS RESEARCH AND MANUFACTURING IN INDIA

- Dr. Goutam Ghosh  
Panacea Biotec Ltd., India

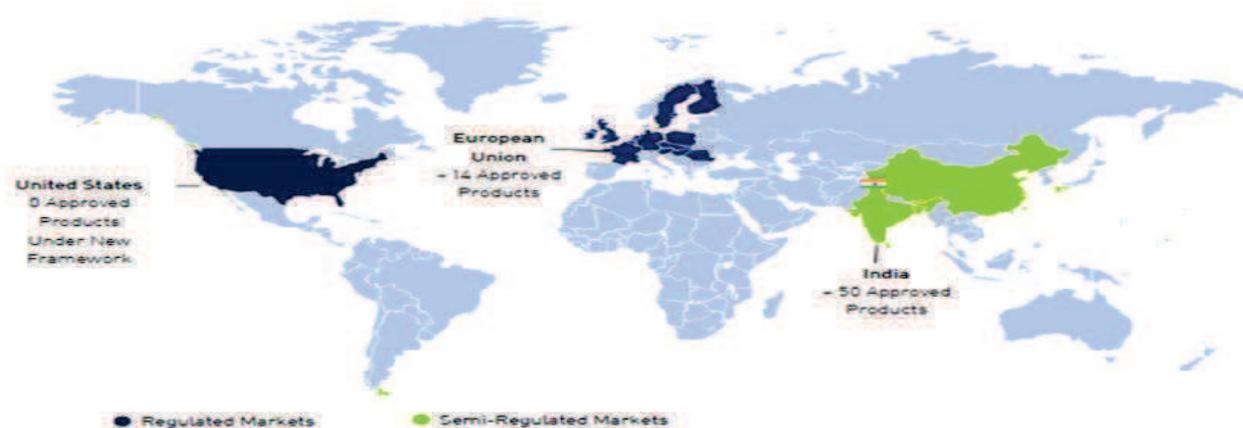


Biopharmaceuticals include all recombinant proteins, monoclonal antibodies, vaccines, hormones, blood/plasma-derived products, non recombinant culture-derived proteins, and cultured cells and tissues

- Biopharmaceuticals broadly comprise the following:
  - Biologics
  - Vaccines
  - Diagnostics
  - Stem Cells
- Biologics may include the following categories:
  - Novel Biologics
  - Biosimilar/ Similar Biologics
  - Bio Equivalents
  - Bio Betters

## A COMPLEX GLOBAL INDUSTRY

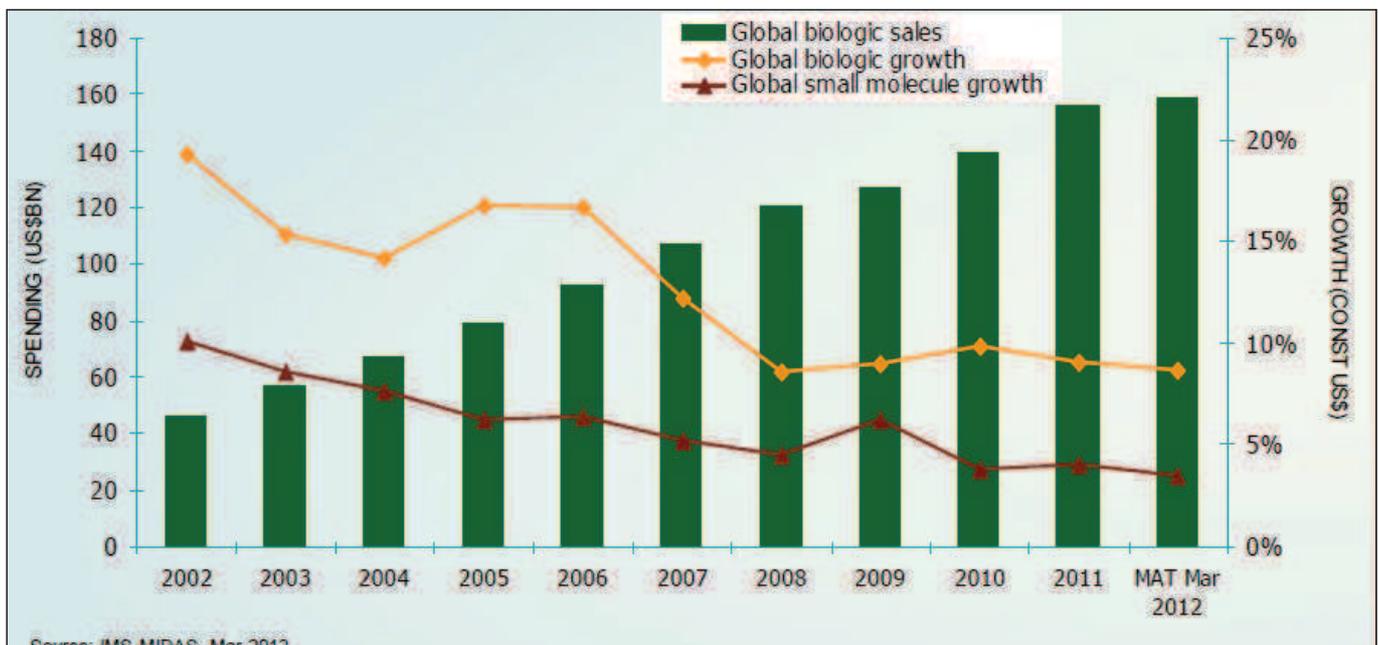
- Market and competitive pressures for Biopharmaceuticals vary from country to country, but may be broadly segmented according to countries that are:
  - Regulated markets, and
  - Semi-regulated markets



- Biologics product grew worldwide and is currently >\$150 billion market
- *The global biosimilars market set to grow to about USD \$ 20 billion by 2015*
- The emerging pharmaceutical markets of Asia, Latin America and Eastern Europe offer especially attractive locations for biosimilars research and commercialization
- They are typically generics-driven pharmaceutical markets; this provides a positive medical and commercial environment for biosimilars

## GLOBAL BIOLOGICS SPENDING & GROWTH

*Biologics products continue to grow worldwide and are >\$150 billion market*

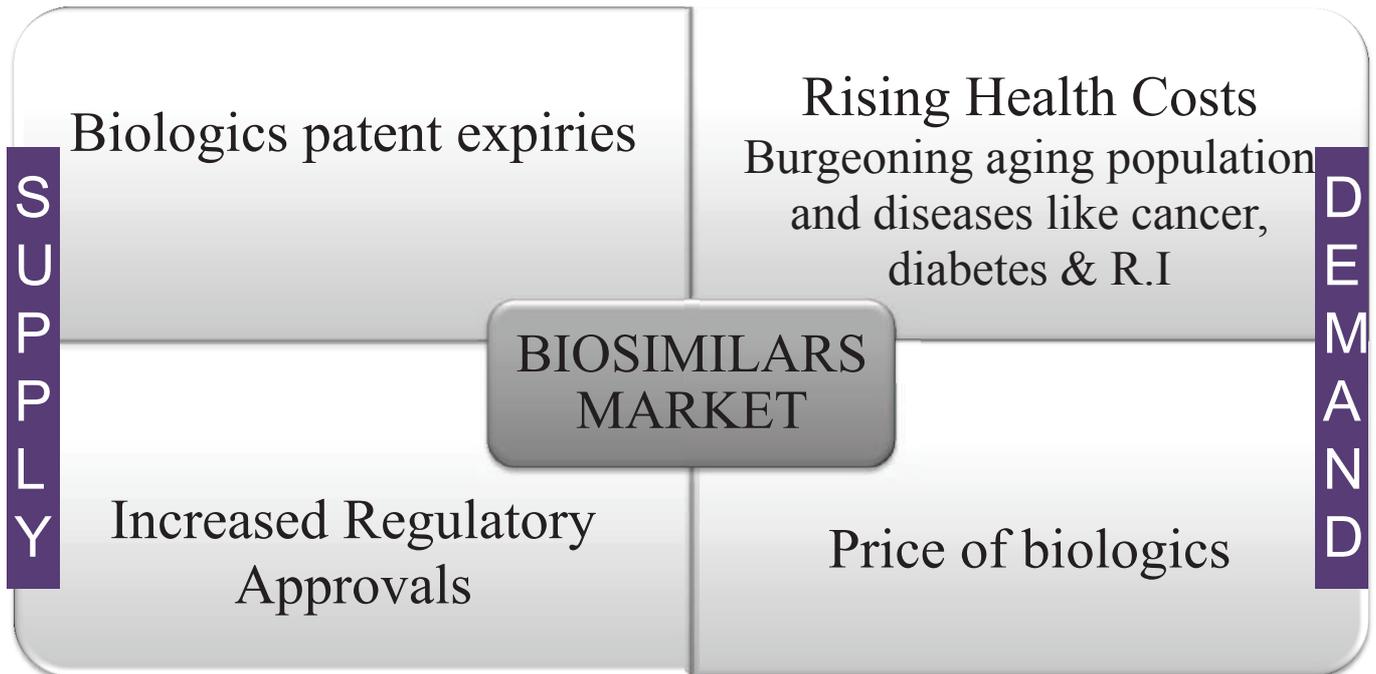


Insulin	15.9
Anti-TNF	15.8
Oncology	12.5
EPO	7.6
Multiple sclerosis	7.3
CSF-G	5.0
Blood coagulation	3.1
Ocular antineovascular	2.0
Anti viral (no-HIV)	1.5
Others	16.5

## OPPORTUNITIES & BARRIERS IN BIOSIMILARS MARKET

- The global Biosimilars market is expected to grow to US\$ 10 billion by 2015, with growth largely driven by US\$79 billion worth of biologics going off-patent by 2015.
- Biosimilars are typically marketed at prices 25 to 40 percent below innovators' products, which is the primary appeal to customers
- Highest regulatory hurdles to market biosimilar products in US & Europe

## GROWTH DRIVERS FOR THE BIOSIMILAR MARKET



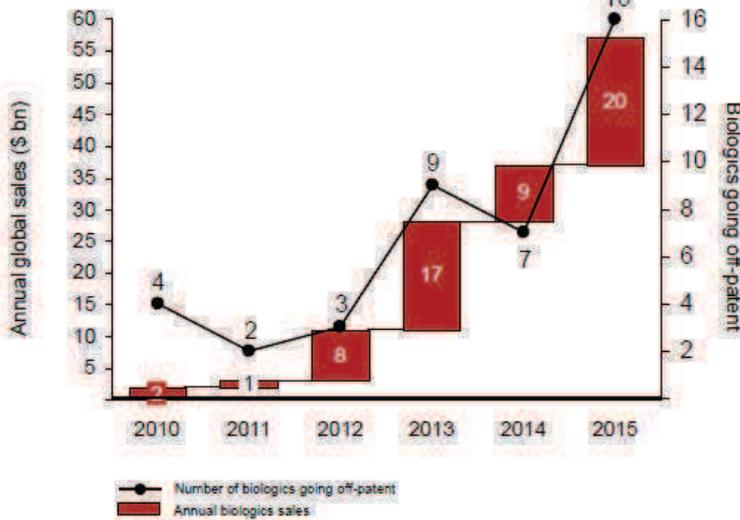
cost-effective alternatives to biologics and patent expiry of the blockbuster biologics are the key drivers

## ANNUAL SALE OF TOP BIOSIMILARS

Brand Name	Therapeutic Indications	Innovator Company	2013 Sales (in billion US\$)	US Patent Expiry
Humira (Adalimumab)	Rheumatoid Arthritis	Abbott	10.7	2016
Remicade (Infliximab)	Rheumatoid Arthritis	Janssen	8.9	Expired
Rituxan (Rituximab)	Non-Hodgkins lymphoma	Roche/ Biogen Idec	8.9	2015
Enbrel (Etanercept)	Rheumatoid Arthritis	Amgen/ Wyeth	8.3	Expired
Lantus (Insulin glargine)	Diabetes	Sanofi	7.8	2015
Avastin (Bevacizumab)	Metastatic colorectal cancer	Roche	7.0	2019
Herceptin (Trastuzumab)	HER2-positive metastatic breast cancer	Roche	6.8	2019
Neulasta (Pegfilgrastim)	Neutropenia	Amgen	4.4	2015

**Through 2015, 45 biologic drugs worth more than \$60 billion in global sales will lose patent protection presenting a major opportunity**

**Number and value of biologic drugs set to lose patent protection per year through 2015**



**Blockbuster biologic drugs set to lose patent protection per year through 2015**

Biologic	Global sales 2008 (bn \$)	US/EU patent expiry
Enbrel	6,5	2012
Remicade	5,3	2013
Rituxan	5,5	2015

Source: Biosimilar series: Forecast analysis: Datamonitor, June 2009; Biosimilars 101, Credit Suisse, August 2009

**BIO-PHARMACEUTICAL INDUSTRY IN INDIA**

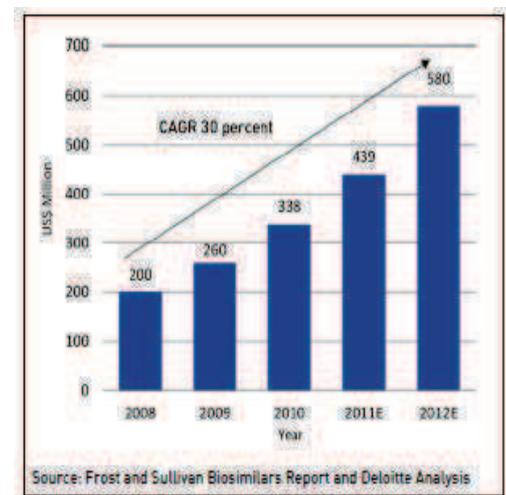
- India has emerged as a global leader in vaccines
- India's bio-pharmaceutical sector is valued at \$ 26 billion
- India is the major supplier of basic Expanded Programme on Immunisation vaccine to the United Nations Children's Fund (UNICEF)

- India is now emerging as the global destination for the manufacture of Biologics, especially Biosimilars and cell-based therapeutics
- India is globally regarded as potential hub for the development and commercialisation of Biosimilars due to its proven experience in generic drugs
- India's new regulatory policy on Biosimilar products would fast-track its development

13

## INDIA – EMERGING BIOSIMILAR DESTINATION

- The Indian Biosimilars industry is estimated to be a US\$ 338 million industry that has been growing at a Compounded Annual Growth Rate (CAGR) of 30%
- There are around 25 Indian companies operating in the Biosimilars space, marketing close to 50 products in the Indian market
- Few Indian companies have already received marketing approval to sell their product in regulated market (Europe)



- ✓ Indian generics companies have gross margins closer to 50%. Straight 50-60% discount e.g. bio-generic of Roche's Rituxan was launched at 50% lesser price in India
- ✓ Development cost for biosimilar requires US\$ 10-20 million as compared to \$50 to \$100 million in developed countries
- ✓ India is one of the major contributors in the world generic market having key skills to replicate fast for bio-generic as well
- ✓ Out of 50 biotech drugs ,13 are available in India and 7 drugs are indigenously developed and produced by Indian companies
- ✓ Of the 40 biologicals marketed in India, 30 are biosimilars
- Highest number of US FDA approved plants outside the US
- Compliance with GCP guidelines is one rise with Indian companies
- Highly qualified human resource availability
- Low capital and operational cost
- Favorable IP scenario

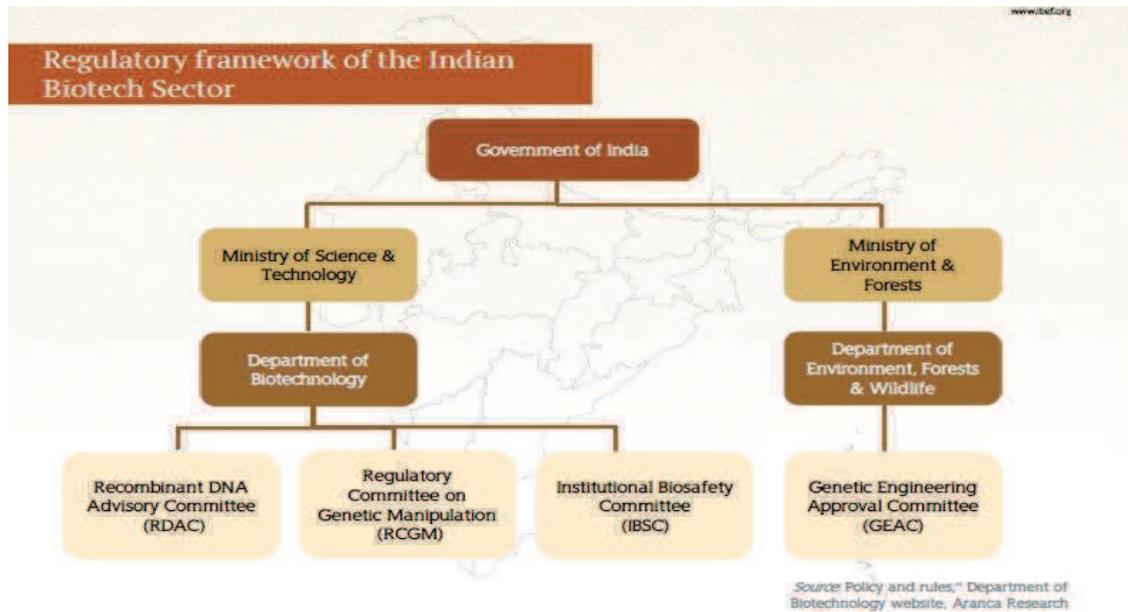
### Notable trends in the Indian biotech sector

Remarkable global positioning

- India is amongst the top 12 biotech destinations in the world
- India ranks second in Asia, after China
- India is the largest producer of recombinant Hepatitis B vaccine in the world

Pharma companies are focusing on biotech

- Ranbaxy, Cadila Healthcare, Lupin, Wockhardt and Dr Reddy's are among the major Indian pharmaceutical companies that operate in the bio-pharma segment



Notwithstanding the above, Government is mandated to set up the National Biotechnology Regulatory Authority (NBRA), an independent, autonomous and professionally led body to provide a single window mechanism for biosafety clearance of genetically modified products and processes

About 20 Indian companies entered into market

Presently, there are about 15 epoetin, 8 G-CSF and 4 insulin biosimilars available

While almost all major Indian drug makers have outlined plans, identified products and set aside investment budgets to develop a robust product pipeline, some have even started rolling them into the market

## Multinational Company

- ✓ To gain access to pipeline of generic products developed by Indian companies to leverage cost effective R&D and low cost manufacturing having similar value
- ✓ In order to increase number of blockbuster drugs going off patent & focus on generic adoption globally
- ✓ To manufacture the product at least 30% cheaper in india that in US
- ✓ Cost of developing generic drug in India about half of that in US

## Indian Company

- ✓ Indian Pharma Co's lack marketing their setup in local geographies . Tie ups with MNC's help Indian companies to develop expertise outside India
- ✓ Foreign companies help in in creating base for infrastructure
- ✓ Help commercializing Indian company products due to well established distributors and retail chain



## PANACEA BIOTEC: CORPORATE OVERVIEW



### Business Lines

- Pharmaceuticals
- Vaccines
- Biotherapeutics (Biosimilars, mAbs & Peptides)

### Ranking

- 2<sup>nd</sup> Largest Vaccine Producer in India
- Ranked 48<sup>th</sup> amongst Pharmaceutical Companies in India (ORG-IMS, 2012)

### Infrastructure

- Four Research & Development Centers
- Established Sales & Distribution Network in India, 50 branded products
- Direct presence in Germany for specialty hospital products segment
- Presence in 55 ROW and Emerging Markets
- cGMP Manufacturing Facilities

### Manpower

- 3,800 Human Resource
- 275 in R&D
- 1,200 in Sales and Marketing

21

## COLLABORATIONS & PARTNERSHIPS OF PANACEA BIOTEC



In-licensing technology for development of hormone based formulation for hair growth.



Panacea Biotec to manufacture and market IPV based combination vaccines for Global Markets.



Panacea Biotec to manufacture and market Measles vaccine for Global Markets.



10 year exclusive in-licensing agreement for Japanese Encephalitis vaccine.

50:50 Joint Venture with Chiron Corporation (now Novartis Vaccines & Diagnostics) for marketing of vaccines in India.

Partnership for development of monovalent - Polio Vaccine (Type 1).

Marketing Partnership in US

Manufacturing collaboration

**Every collaboration - An enduring success**

- Established presence of over 25 years vaccines
- Reliable partner to WHO, UNICEF: has been largest supplier of vaccines to UNICEF from India
- First Indian Company to launch innovative branded combination vaccine – Easyfive (DTP/Hep B/Hib) & other combination vaccines (Easyfour, Ecovac)
- One of the 3 Companies chosen by Govt. of India to develop Flu vaccine (Pandemic Flu)



## INNOVATION INFRASTRUCTURE



**OneStream, New Delhi**

### Drug Discovery: Novel Biologicals

- Target identification to development of pre-clinical candidate
- Focus areas for Novel Peptides: Metabolic Disorders
- Vaccines & Biosimilars



**Laksh, Mohali**

### Drug Discovery: Small Molecules

- Target identification to development of pre-clinical candidate
- Focus areas : Metabolic disorders, Anti-infectives, CNS



**Sampann, Lalru**

### Difficult to Develop Products

- High barrier to entry generics
- NDDS technologies: Depot Injections, Oral modified release, SMEDDS in Softgels, MD tablets, Critical dose drugs
- Vaccine Formulation & Bio-therapeutics Development



**GRAND, Navi Mumbai**

### NDDS Product Development

- Platform NDDS technologies : Nanoparticles, Liposomes, Micro-particles, Depot Injections, SPORT, Oral films etc.
- High barrier to entry generics

Location: Delhi	Branches	Projects
	<b>Bacterial Vaccine Development (Recombinant and native)</b>	- 10 and 15 Valent Pneumococcal Vaccine, - Tetravalent Meningococcal Vaccine
	<b>Viral Vaccine Development</b>	<ul style="list-style-type: none"> <li>• Cell based: Dengue vaccine</li> <li>• JEV vaccine</li> <li>• Egg based vaccine: Pandemic flu (H1N1)</li> </ul>
	<b>Biotherapeutics / Recombinant Proteins</b>	Darbepoetin and Mabs Like; Trastuzumab (herceptin), Bevacizumab (avastin), Adalimumab (humira) & others
	<b>Peptide Biology</b>	Hair Growth Peptide, Anti diabetic molecules

VACCINE MANUFACTURING CAPABILITIES

Location	Bulk Antigen Facilities	Built Up Area
<b>Lalru</b>	Recombinant Vaccines	~ 40,000 sq. ft
	Bacterial Vaccines	~ 18,000 sq. ft
	Tetanus Vaccine	~ 20,000 sq. ft
	Cell Culture Vaccines	~ 30,000 sq. ft
<b>Baddi</b>	<b>Formulation Capacity doses p.a. (Includes Vials &amp; PFS) of 1.6 billion</b>	

- Facility Bulk Antigens manufacture & formulation –
  - ✓ Diphtheria Toxoid • Tetanus Toxoid • Whole cell / acellular Pertussis
  - ✓ Haemophilus influenzae type b conjugate • Recombinant Hepatitis B
  - ✓ Inactivated H1N1 split viron influenza vaccine bulk (egg based technology)
  - ✓ Vaccines : Dengue Vaccine, Japanese Encephalitis, Sabin IPV, egg based seasonal flu vaccine, Yellow Fever vaccine
  - ✓ Bio therapeutics : Viral proteins, non-viral recombinant bio molecules on cell culture in both conventional & disposable formats



## Mammalian Cell Culture Products

- MABs
- Biosimilars



### Fermentation Capacity at Production Level

- ✓ Pre inoculum preparation - 1 x 5 L
- ✓ Inoculum Preparation - 1 x 50 L
- ✓ Production fermentation - 2 x 500 L
- ✓ Harvesting - Microfiltration (TFF)



- Low pressure Chromatography
  - ✓ Gel Filtration Chromatography
  - ✓ Ion exchange chromatography
- Tangential flow filtration (TFF)
  - ✓ Concentration / Diafiltration
- HPLC
- Centrifugation
- Sterile filtration
- Purification
  - ✓ High Pressure Chromatography
  - ✓ Medium Pressure Chromatography System





# PROCESS AREA- CIP/SIP Fermenter



31

# PROCESS AREA- Zonal centrifuge



172

32



**Sartorius C 30**  
42 L / 30 L fermenter



**Production Fermenters**  
(B.Braun) D 600 fermenter



# PROCESS AREA

**Bioengineering  
750 L / 500 L fermenter**



**Disc stack centrifuge  
(Westfalia)**



35

# PROCESS AREA

**Automated  
Chromatography System**



**Tubular centrifuges**



174

36

- **State of the art GMP complied facilities for R&D and manufacturing**
- **Trained manpower**
- **Collaborative research and end to end process development**
- **Process development and scale up from bench to 500 L scale for various biologicals**
- **Fast-track the development, Pre-clinical & clinical trials up to the commercialization of product through co-development with partner companies in areas of mAb, vaccines, small molecules, generic & novel (NDDS) formulations**

**THANK YOU**



# Enhancing Eye Care Services through Innovation, Technology and Collaboration in India

Dr R R Sudhir

Senior Consultant – Cornea Services

Head Dept of Preventive Ophtalmology

Consultant-incharge Electronic Medical Records

Sankara Nethralaya, Chennai





# Enhancing Eye Care Services through Innovation, Technology and Collaboration in India

## Dr R R Sudhir

MBBS, DO, DNB, MPH (JOHNS HOPKINS USA)  
Senior Consultant - Cornea Services  
Head Dept of Preventive Ophthalmology  
Consultant-incharge Electronic Medical Records  
Sankara Nethralaya  
Chennai



## Eye Health burden in India



# Cataract Burden in India

	2001	2005	2010	2015	2020	Avg output / yr
Ophthalmologists		13,000	17000	21000	25000	1200
Optometrists		12000	25000	40000	60000	4500
No-of cataract blind (Million)	7.749	7.390	7.536	7.823	8.252	
Prevalence+ Incidence	9.229	8.868	9.043	<b>9.394</b>	9.903	
Population ( Million)	149.863	163.165	192.598	229.340	274.272	

- Currently no-of cataract surgeries performed per annum > 6 Million
- One need to perform 500 -600 cataract surgeries per annum per head
- 1/3<sup>rd</sup> of Ophthalmologist do only medical practice

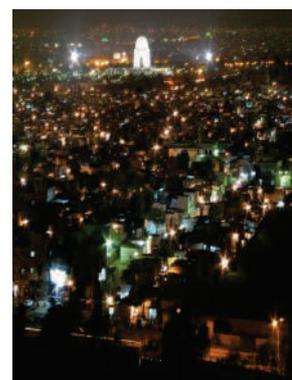
*Murthy G, Gupta SK, John N, Vashist P. Current status of cataract blindness and Vision 2020: the right to sight initiative in India. Indian J Ophthalmol. 2008 Nov-Dec;56(6):489-94*

2-02-2012

70th AIOS, Cochin

# Health care availability in India

- 80 of population resides in rural area
- 70 of health care resources are in urban area
- 70 of them practice in Urban areas.
- 1 Ophthalmologist / 100,000 population.



# Tele-Ophthalmology



Comprehensive  
eye examination & spectacle dispensing at  
door step

2/2/2012

## Rural-Urban Divide

- A large number of (26.1 %) of persons who did not attend free eye camps cited 'inability to leave family or work responsibilities'.
- Cataract surgery was recommended for nearly 36 % of persons identified by their community as having eye problem, but who *did not attend* the eye camps.

Fletcher AE, Donoghue M, Devavaram J, Thulasiraj RD, Scott S, Abdalla M, Shanmugham AK, Murugan B. Low Uptake Of Eye Services in Rural India – A Challenge for Programmes of Blindness Prevention. Arch Ophthalmol. 1999;117:1393-1399.

# MOBILE EYE SURGICAL UNITS



SANKARA NETHRALAYA – IIT MADRAS MOBILE EYE CARE UNITS IS DESIGNED TO SERVE THESE PATIENTS IN INACCESSIBLE AREAS

- MESU Video

## Challenges faced in designing a Mobile Eye Surgical Unit

### Water sterility

Sterile water is made available from a reverse osmosis plant installed in the vehicles. This water is used both for sterilization and scrubbing.

### Air sterility

An Air Handling Unit provided in the sterile vehicle functions on a three-stage air filtration/purification system and ensures elimination of air pollution.

### Air sterility

The walls of the Operation theater will be made up of SS 316 - a food grade steel.

Stability during surgery

Both the sterile and the preparation vehicles are provided with hydraulic jacks to keep them steady and level even in uneven grounds to facilitate safe surgery.

Consistent power supply

A 20KVA generator set will be mounted in the vehicle to provide steady electric supply, even in places where there is no local electric supply.



- Patient waiting area outside Mobile Units



- Local Anesthesia under cardiac monitoring in a preparatory vehicle



- Ophthalmic nurse scrubbing



- Patient monitor and autoclave machine



- Nursing assistant preparing trolley

- Surgeon getting ready



- Operation in progress



- Surgery in progress



- Patient being helped out after surgery



- Post Operative instructions before discharge



- Post Operative review at camp site

## The mobile comprehensive eye care unit will

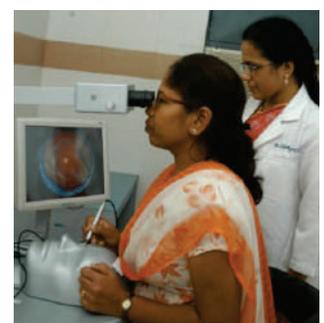
- Screen over 25000 patients every year
- Carry out about 3,000 cataract surgeries
- Dispense spectacles to patients having refractive errors.
- Conduct awareness programmes to disseminate information on eye care and various eye diseases.

## Surgical Training for Cataract

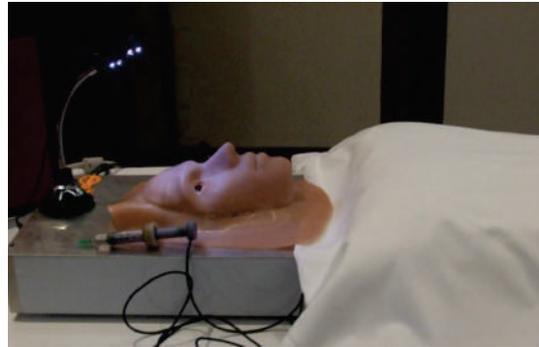
- Observation
- Wet Lab Training
- One to One supervision
- Live Recording of surgeries
- Virtual Reality Training



2/2/2012



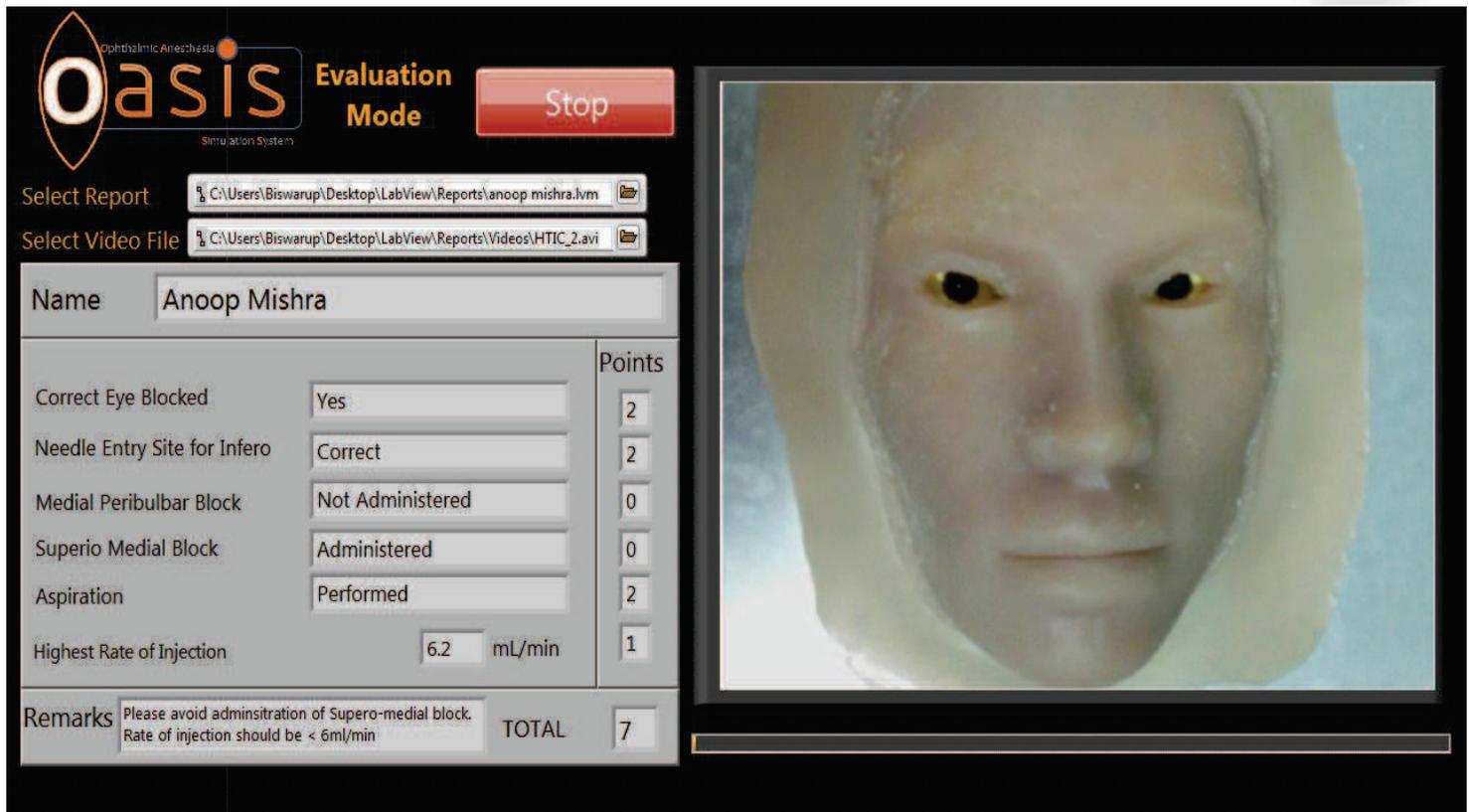
# Ophthalmic Anaesthesia Simulation System - OASiS



Developed in collaboration with IIT Madras

## What it can do?

- 1 Designated eye has been blocked.
- 2 Needle touch of the muscles/globe/Optic Nerve.
- 3 Needle entry site for Infero-lateral block.
- 4 Administration of Supero-medial and medial peribulbar block.
- 5 Aspiration performed or not?
- 6 Measures rate of injection.
- 7 Video recording and playback features for future reference
- 8 Special instructor console to monitor and add comments



**OASiS** Evaluation Mode  
Simulation System

Select Report: C:\Users\Biswarup\Desktop\LabView\Reports\anoop mishra.lvm

Select Video File: C:\Users\Biswarup\Desktop\LabView\Reports\Videos\HTIC\_2.avi

Name: Anoop Mishra

Task	Value	Points
Correct Eye Blocked	Yes	2
Needle Entry Site for Infero	Correct	2
Medial Peribulbar Block	Not Administered	0
Superio Medial Block	Administered	0
Aspiration	Performed	2
Highest Rate of Injection	6.2 mL/min	1
<b>TOTAL</b>		<b>7</b>

Remarks: Please avoid administration of Supero-medial block. Rate of injection should be < 6ml/min

## Benefits of OASiS

- It enables to learn and practice a safe orbital regional anaesthesia
- Helps to teach the post-graduates/trainees
- Helps to conduct workshops
- Hands-on practice sessions
- Also enables to train the paramedics for giving blocks
- To learn and teach anatomy



## Critical Infections



- **High Mortality**
- Most of the deaths in first 72 - 96 hours of admission in to the hospital
- Transplant recipients and febrile neutropenics
- **Loss of function**
- Blindness or visual impairment in 96 hours after the start of first symptom.
- Residual damage in the form of paralysis or sensory loss, epilepsy in 96 hours
- Septicaemia leaves behind residual Renal & hepatic damage and Arthritis

# Diagnostic Challenges of Critical infections



- Diagnosis needed in the first 24 hours of admission
- Infection is localized. No evidence of infection in Blood / serum / plasma
- Clinical specimen available for eye infections
  - Corneal scrapings – a few cells
  - Conjunctival swab – a few cells
  - Aqueous humor – 50  $\mu$ L
  - Vitreous fluid - 50  $\mu$ L

## Current Diagnostic solutions for Critical infections



### **Bacterial & Fungal Cultures:**

- Take 72- 96 hours Not useful.
- Sample is too small for Culture
  - Ten to fifteen bacteria or fungal particles in eye samples and about 50 particles in CSF
- Anaerobes cannot be cultured
- Mycobacteria take too long to grow 12-14 days in BacT Alert)
- Over all success rate is less than 15 (7000 blood cultures in CMC 1100 positive and 840 are pathogens)

# Diagnostic Challenges of Critical infections

## Viral Identification

- Culture takes 7 days
- Requires additional techniques such as immunofluorescence for diagnostic validity
- Immunodiagnosis: serology: Antibody detection; Useful only after 5 days of infection
- Immunodiagnosis: antigen Detection: Very Low sensitivity useful in 15% of Pyogenic Meningitis or Cryptococcal meningitis.

## Diagnostic Challenges of Critical infections

- PCR and / or Other Nucleic Acid Based Detection tests:
  - Very High Sensitivity
  - Low volume of clinical specimen to Perform multiple PCR Reactions
  - Conventional PCR read out is Gel Electrophoresis – identification basis is size of the gene fragment amplified – not sequence – leading to unacceptable level of false alarms

# Diagnostic Challenges of Critical infections

## XCyton's Solution- A Paradigm Shift

### Disease Based Diagnosis

- Sequentially looking for one organism after the other
- A
- B
- C
- D
- .....N

### Syndrome Evaluation System

- Simultaneously looking for all pathogens that could probably cause a disease
- A, B, C, D.....N all tested

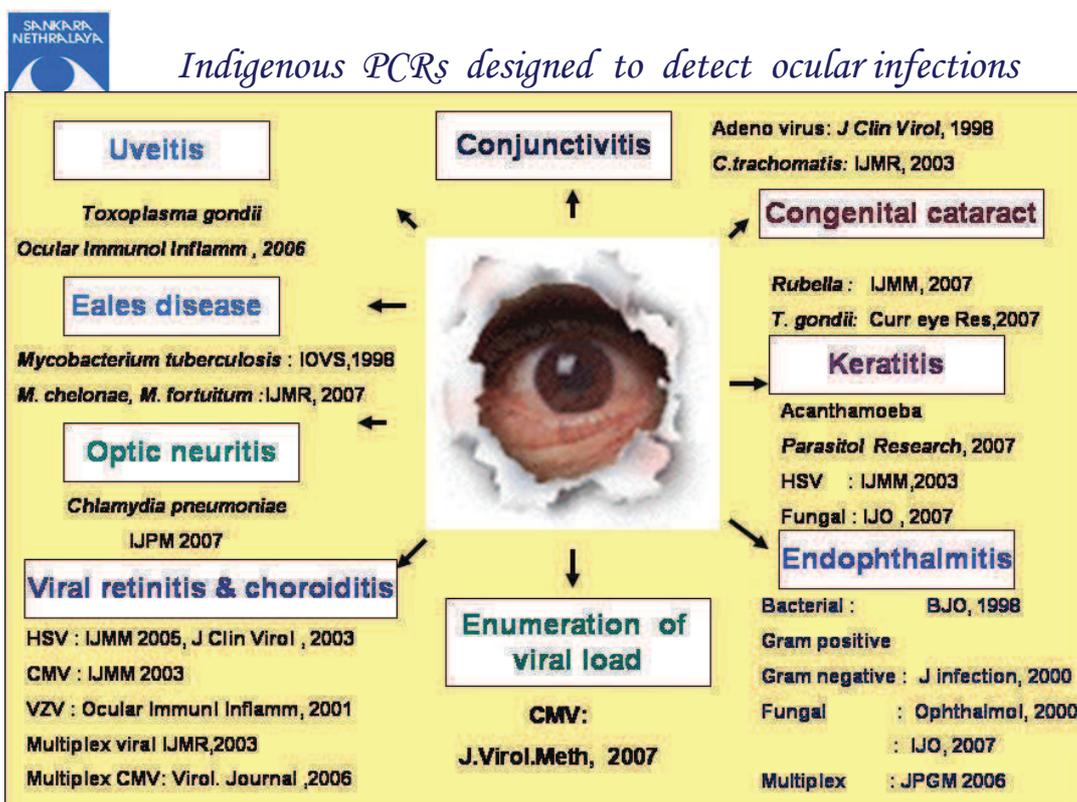
## What is available?



- DNA Micro Chips with about 60,000 features defining pathogen sequences
- End read out is not a definite “yes” or “No”
- Enormous data on signal to noise ratios and probabilities
- Not useful in clinical setting

# What is available?

- Multiplex Real Time PCR for all Herpes Viruses using a single set of primers and three probes
- Multiplex to distinguish Mycobacterium tuberculosis and avium intracellulare
- **No multiplex System that can simultaneously detect more than one class of pathogens such as viruses or bacteria**



# DNA CHIP TECHNOLOGY

## ONE CHIP FOR ONE TYPE OF CLINICAL OCULAR INFECTIOUS DISEASE

WORLD'S FIRST DNA CHIP FOR EYE INFECTIONS



### CHIP DESIGN & DEVELOPMENT

Infectious agents detectable by DNA chip

S.No.	Organism	S.No.	Organism
1	HSV ( 3 Genes)	8	Gram-negative
2	CMV ( 3 Genes)	9	<i>P. acnes</i>
3	VZV ( 2 Genes)	10	<i>M. tuberculosis</i>
4	Adenovirus	11	<i>M. fortuitum</i>
5	<i>C. trachomatis</i>	12	<i>M. chelonae</i>
6	Eubacteria	13	Panfungus
7	Gram-positive	14	<i>T. Gondii</i>

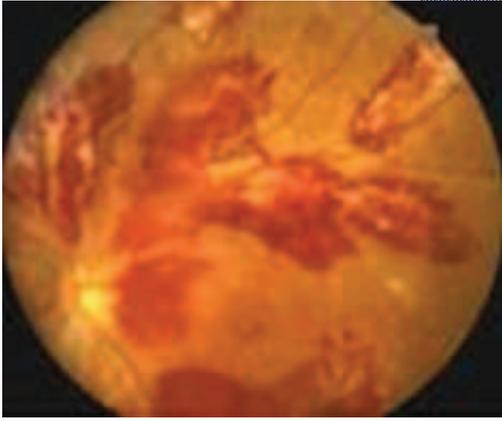
(Total - 21 regions)

14 infectious agents + 1  
Human β globin gene

In-house PCR  
available  
for the detection  
of all the 14 infectious  
agents

# DNA CHIPS FOR **FOUR** DIFFERENT OCULAR INFECTIOUS CLINICAL CONDITIONS

## VISION CHIP



I. VIRAL RETINITIS



II. KERATO CONJUNCTIVITIS



III. UVEITIS & CLINICALLY SUSPECTED MYCOBACTERIAL INFECTIONS



IV. INFECTIOUS ENDOPTHALMITIS

## STEPS INVOLVED



**DNA extract**  
From ocular  
clinical  
specimen  
Eg: Actap

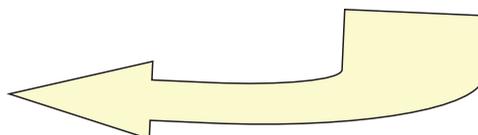


**PCR reaction mix**  
(all 21 primer sets)

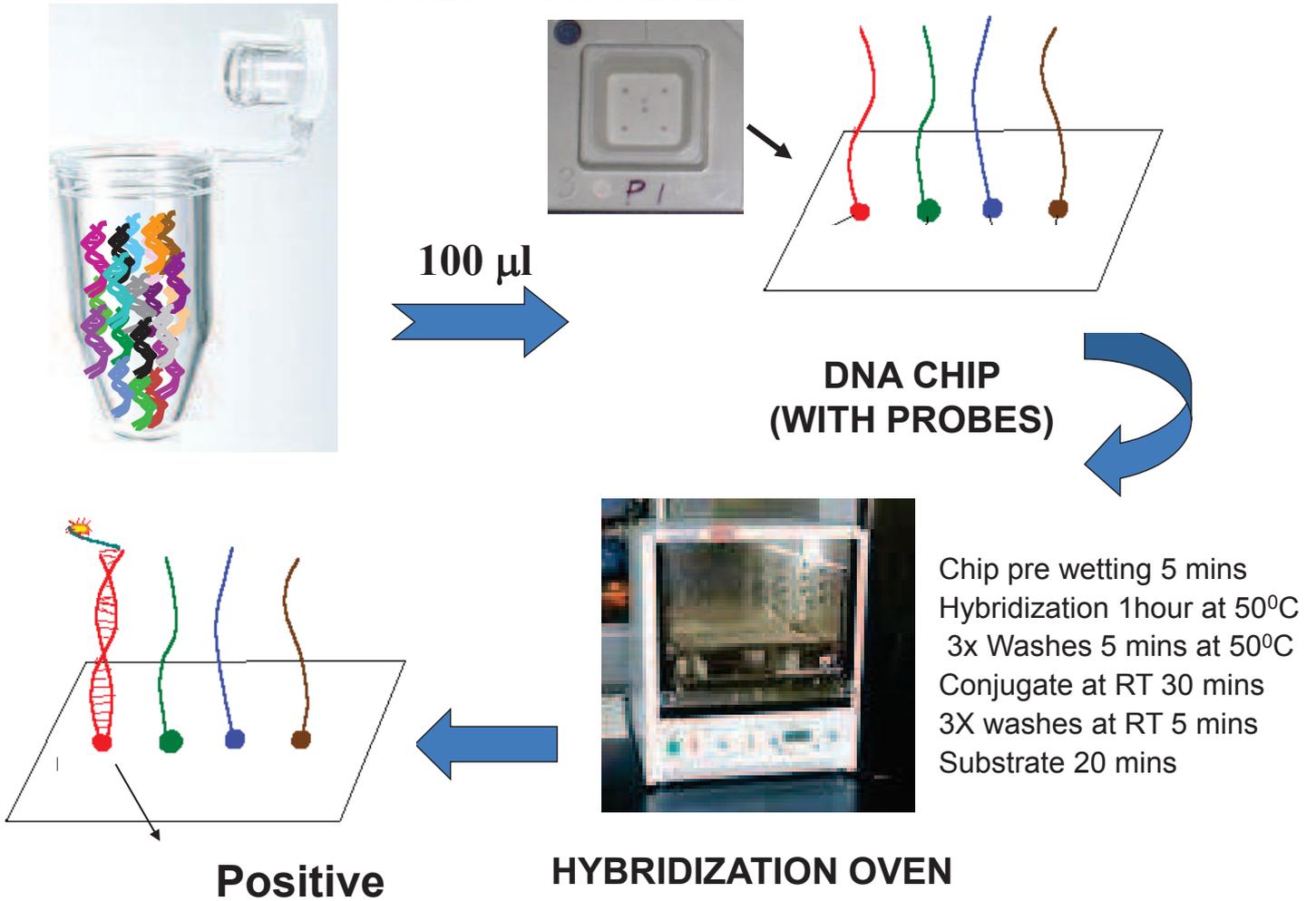


**Amplification**

95°C- 10mins  
95°C 45 sec  
60°C 45 sec  
72°C 45 sec } 35 cycles



## STEPS INVOLVED



## CLINICAL SPECIMENS – VIRAL RETINITIS

Presence of UL-44 helpful in serotyping HSV ← UL 44 gene

HSV1

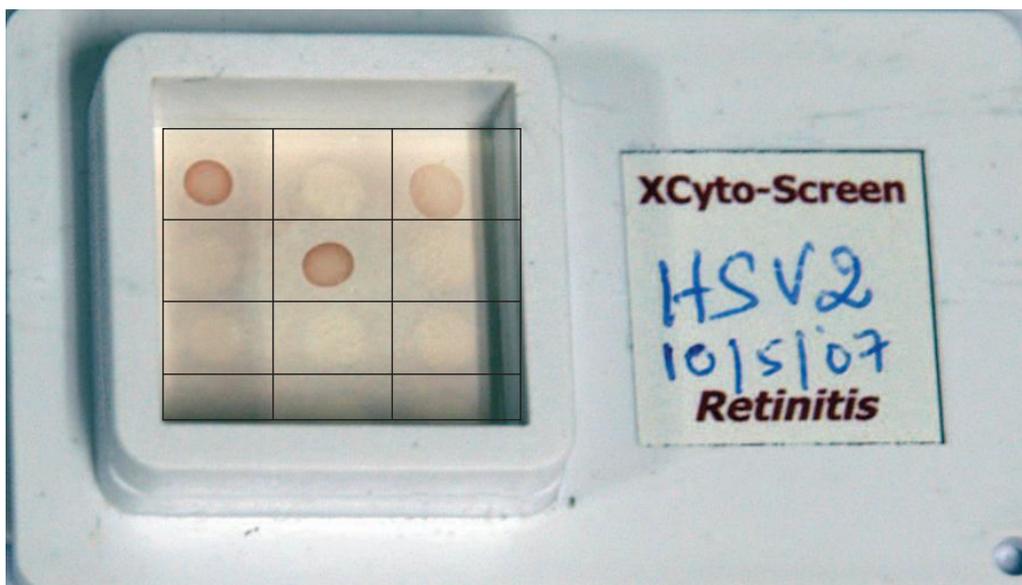
HSV2

CMV

HSV DP	HSV UL44	HSV gD
VZV DP	HSV gD (c)	VZV ORF 29
CMV mtrI	CMV UL83	CMV gO
	CMV mtrI (c)	β globin

VZV

← β globin gene



# Product Specifications

- Total process time (DNA Extraction, amplification and Hybridization) less than 7 hours allowing same day reporting
- “Yes” or “No” answers
- No quantification of the pathogen
- End read out to be by naked eye without loss of clinical sensitivity
- No use of fluorescence signals and fluorescent scanners
- Usable at even district level hospitals

## How many genes per pathogen?

- Some pathogens require multiple genes to be amplified to get adequate clinical sensitivity
  - HSV Required three Genes
  - CMV Required three Genes
  - VZV Required only two genes

## Sensitivity & Specificity

- **Sensitivity** is conferred by
  - Nucleic acid amplification
  - Amplification of signal by the enzyme at the level of hybridization
- **Specificity:**
  - Sequence specific Hybridization

Determination of the **clinical sensitivity** of DNA microchip using DNA of prospective clinical specimens 

Total no. of specimens collected : 40

Intraocular fluids : 26

External ocular specimens: 14

**All the specimens stored in liquid nitrogen;  
DNA of the clinical specimens stored at – 80°C**

**Necessary laboratory PCRs performed on the clinical specimens and results recorded**

## ACURACY STUDY OF EYE SAMPLES CONDUCTED AT SHANKAR NETRALAYA, CHENNAI

Specimen Category	No. Tested	No. Positive
HSV Culture Positive	05	18
HSV FAT Positive	13	
CMV FAT Positive	13	13
Adenovirus Culture Positive	02	07
Adenovirus Culture negative in conjunctivitis	05	
Varicella zoster virus	03	03
Aspergillus( KOH Positive)	04	16
Aspergillus (KOH Negative)	12	
Propionibacterium(PCR positive)	22	22
Bacterial Endophthalmitis ( culture/smear positive)	25	25
Mycobacterium tuberculosis (PCR Positive)	10	10
Toxoplasma ( PCR positive)	8	8
Mycobacterium chelonae (PCR Positive)	7	7
Mycobacterium fortuitum (PCR positive)	3	3
Aqueous Humour obtained at cataract surgeries (Non-infectious controls)	30	0

### THE XCYTON SYNDROME EVALUATION SYSTEM – DNA CHIP

Not only used for Eye infections the same technology is used for diagnosis of

**ENCEPHALITIS**  
**SEPTICAEMIA**  
**FEBRILE NEUTROPENIA**  
**TRANSPLANT INFECTIONS**  
**PNEUMONIA**

# APPLICATION OF PCR BASED dHPLC IN CULTURE NEGATIVE PCR POSITIVE INTRAOCULAR SPECIMENS

Journal of Microbiological Methods 85 (2011) 47–52



Contents lists available at ScienceDirect

Journal of Microbiological Methods

2011; 85 (1): 47-52

journal homepage: [www.elsevier.com/locate/jmicmeth](http://www.elsevier.com/locate/jmicmeth)



## Identification of bacteria in culture negative and polymerase chain reaction (PCR) positive intraocular specimen from patients with infectious endophthalmitis

Pasupathi Aarthi, Rajagopal Harini, Murali Sowmiya, Jambulingam Malathi, K. Lily Therese, Hajib N. Madhavan\*

*J. E. T. Microbiology Research Center, Sankara Nethralaya, 18, College Road, Chennai 600 006, India*

### ARTICLE INFO

#### Article history:

Received 15 October 2010

Received in revised form 11 January 2011

Accepted 12 January 2011

Available online 22 January 2011

#### Keywords:

dHPLC

DNA sequencing

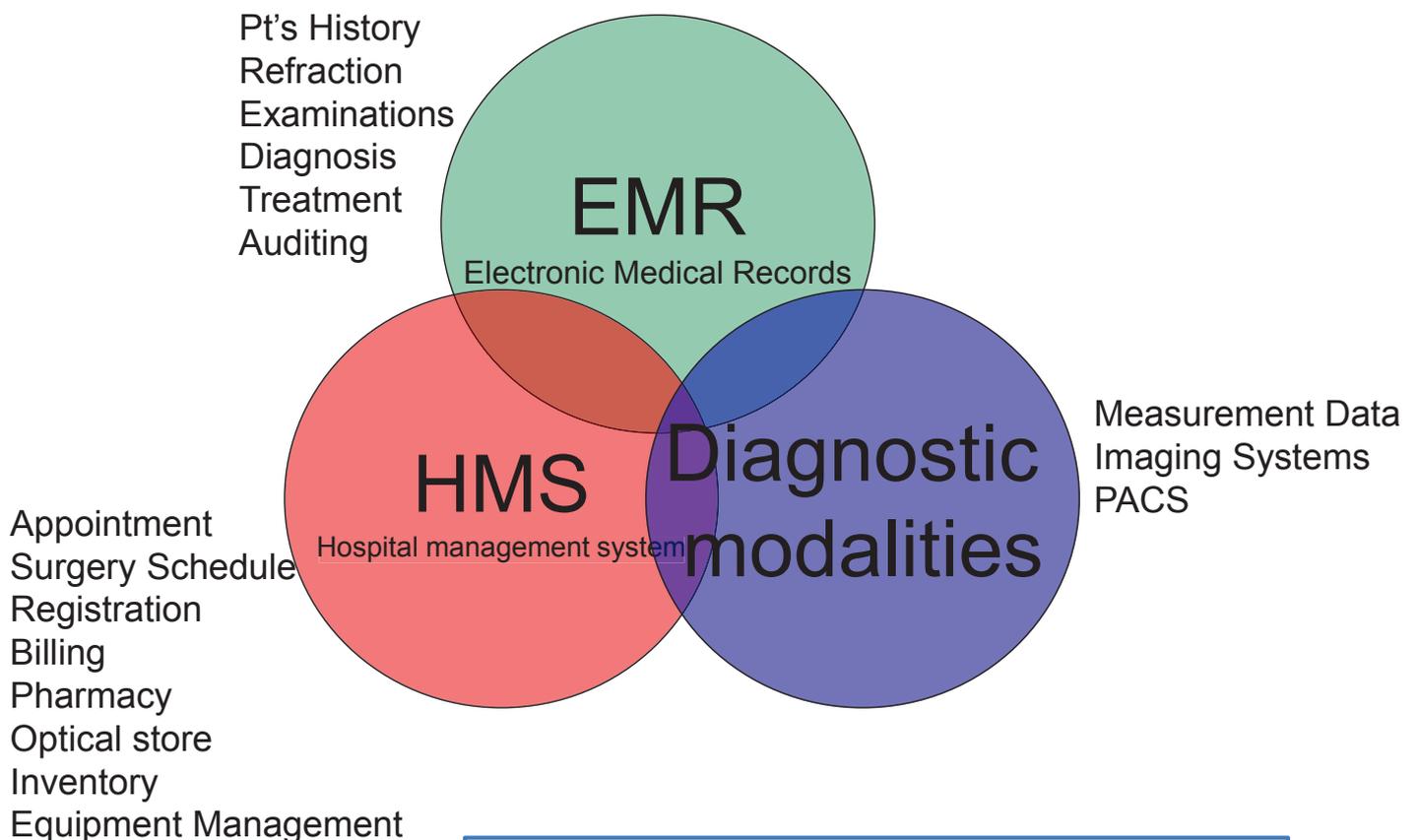
Infectious endophthalmitis

### ABSTRACT

A novel Denaturing High-Performance Liquid Chromatography (dHPLC)-based technique allows rapid high-resolution analysis of PCR products. We show the application of this PCR/dHPLC approach for direct detection and identification of bacterium from the Eubacterial PCR amplified products of aqueous and vitreous aspirates from patients with endophthalmitis and to differentially identify the culture negative cases and initiate appropriate therapy. The aim of this study is to identify culture negative PCR positive cases by the application of PCR based DNA sequencing. A total of 116 intraocular specimens were subjected for the study. Sixty-nine different bacteria were identified using dHPLC based DNA sequencing of which predominant ones were Gram-positive bacteria and cannot be cultured by conventional methods. Forty eight different bacteria detected in this study is being reported for the first time in infectious endophthalmitis.

© 2011 Elsevier B.V. All rights reserved.

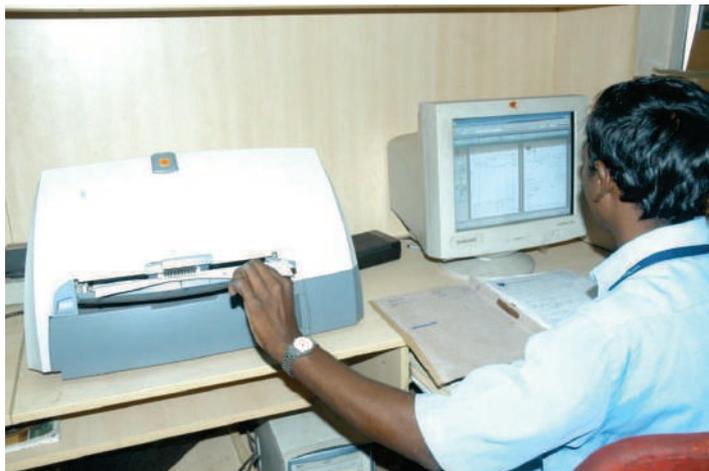
## PAPER LESS HOSPITAL - ELECTRONIC MEDICAL RECORDS



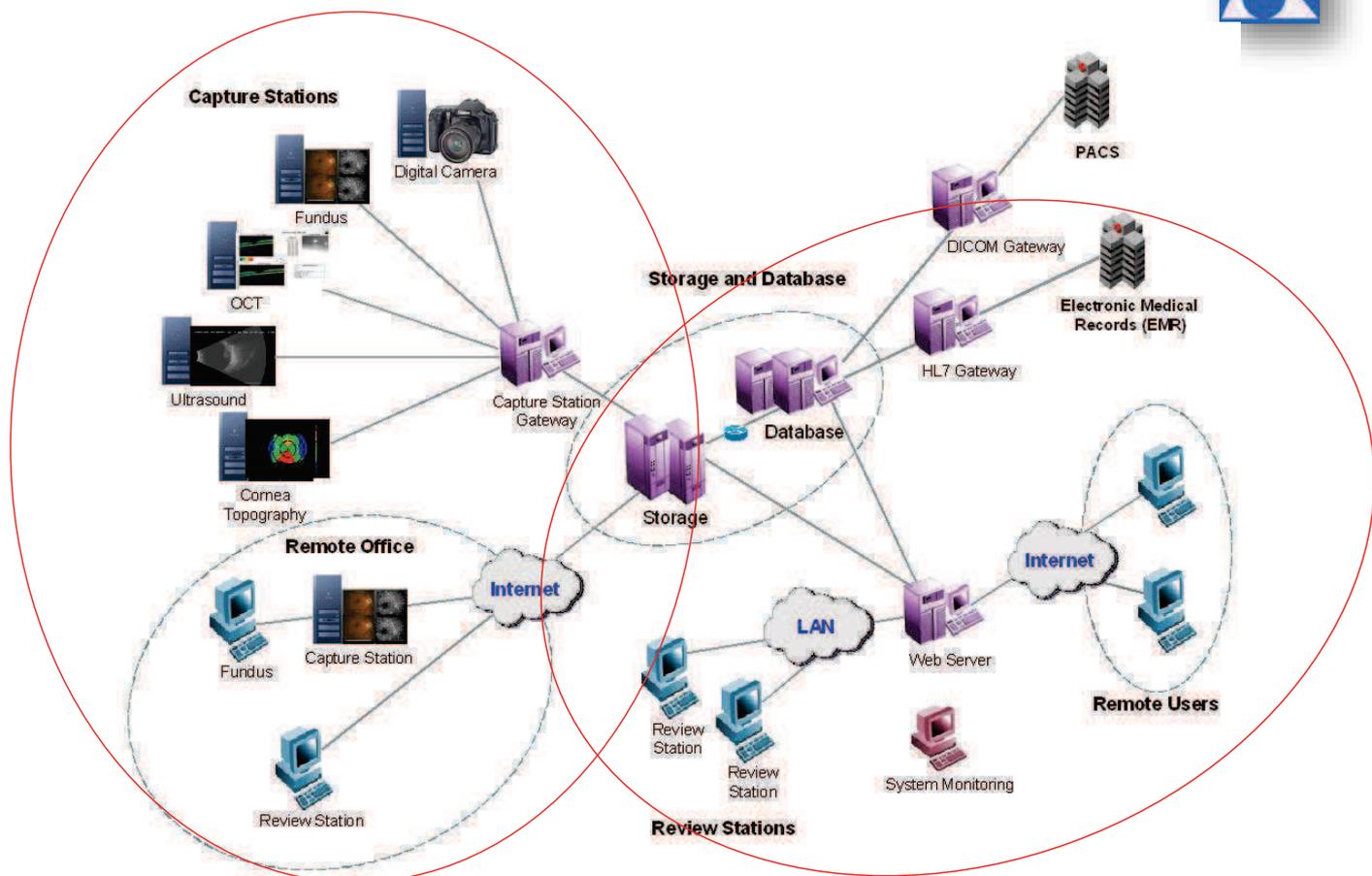
In Collaboration with TCS – Tata Consultancy Services

# Why do we need EMR

- Case Study : Sankara Nethralaya - MRD
- > 2 Million files (1978-2013)
- >1200 patients / day (500:New)
- >100 Surgeries /day
  - *Huge Storage space and man power*



## INTEGRATION WITH EQUIPMENTS



**SANKARA NETHRALAYA**  
Electronic Medical Records

Search  
this website on the web Google

welcome Doctor SUDHAKAR  
Operating in: JKCIN  
Logout | Change Password

Schedule Case Summary Examination Reports

784214 Go File Status - ELRC IOL Request Form Patient Case Summary Patient Case Summary Tracking

Case Summary Files Attached IOP Visual Acuity Customized Reports

RAUT, PRIYANKA ID#: 784214 DOB: 7/25/1991 Sex: F MEDICAL RESEARCH FOUNDATION  
Date: 4/27/2010 2:06:14 PM Exam 4

RAUT, PRIYANKA ID#: 784214 DOB: 7/25/1991 Sex: F MEDICAL RESEARCH FOUNDATION  
Date: 4/27/2010 2:06:50 PM Exam 6

OD OS

Ka: 65.60 @ 95° Kt: 66.48 @ 182° AwaK: 69.28  
MiniK: 66.36 @ 172° Est: 0.58 / Err: 0.60 Cyl: 7.61  
SRL: 1.52 PVA: 20/40-20/60 SAI: 0.45

Ka: 65.57 @ 88° Kt: 66.14 @ 6° AwaK: 69.68  
MiniK: 66.13 @ 7° Est: 0.73 / Err: 0.85 Cyl: 7.42  
SRL: 1.25 PVA: 20/30-20/40 SAI: 1.92

Encounter 1  
12-29-2009 / REG  
12-29-2009 / PAC  
12-29-2009 / CTG  
12-29-2009 / CTG  
12-29-2009 / PEN  
12-29-2009 / PEN  
12-29-2009 / CC

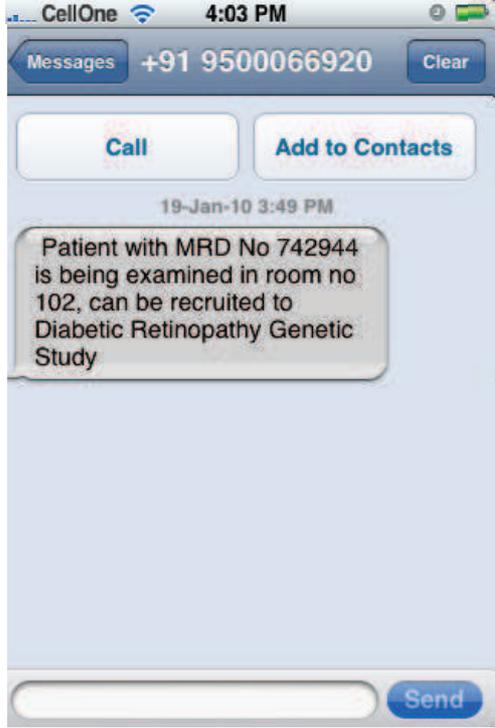
Encounter 2  
04-27-2010 / REG  
04-27-2010 / PEN  
04-27-2010 / PEN  
04-27-2010 / PAC  
04-27-2010 / CTG  
04-27-2010 / CTG  
04-30-2010 / CNCD  
05-01-2010 / PO  
05-03-2010 / CC

Encounter 3  
09-28-2010 / REG

# Recruitment of subjects for Clinical Studies



Alert message is mapped to the IP address of the desktop machine and room no and linked to automatic SMS programme installed in the application



# Send an Automatic SMS

- To Improve Compliance to treatment protocols
  - Reminders for taking medications
  - Reminders for Scheduling and attending appointments
  - Reminder for patching – Amblyopia treatment



Welcome Doctor SUDHIR R.R.  
Operating in: SN MAIN  
[Logout](#) | [Change Password](#)

## INBUILT DATA MINING

Schedule ▾ Case Summary ▾ Examination ▾ Reports ▾

Dynamic Report

Categories: **Diagnosis** Screens: **Diagnosis** Fields: **Diagnosis**  
 Contains: **And** **<select>** **And** **<select>** **And** **<select>**  
**KERATOCONUS UNSPECI**

Category	Screen	Field	Assign	Value	Condition	Delete
1	Diagnosis	Diagnosis	Contains	KERATOCONUS UNSPECIFIED	And	Delete

Report Format:  DataGrid  Excel  Chart  
 Period From: **01/04/2010** Period To: **08/10/2010**

Display Parameters

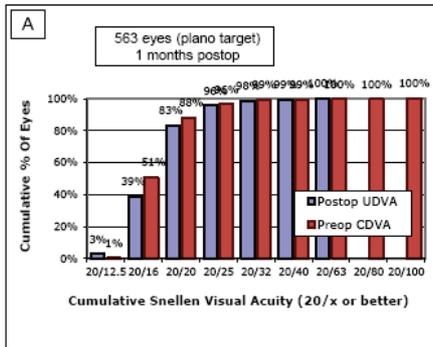
Categories: **Visual Acuity / Refraction** Screens: **Visual Acuity/Refraction** Fields: **<select>**

Category	Screen	Field	Delete

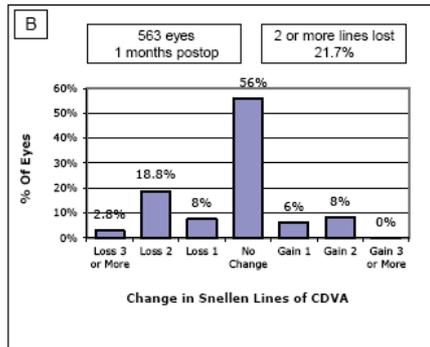
<select>  
 AC OS ADD SPHERICAL OI  
 AC OS ADD BCVA OPT  
 AC OS ADD DISTANCE OP  
 AC OD PREFERENCE SE OPT  
 AC OS PREFERENCE OPT  
 AC OD DV SPHERICAL  
 AC OD DV CYLINDRICAL  
 AC OD DV AXIS  
 AC OD DV BCVA  
 AC OS DV SPHERICAL  
 AC OS DV CYLINDRICAL

207

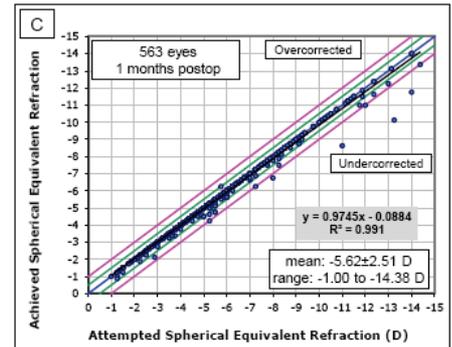
Internet



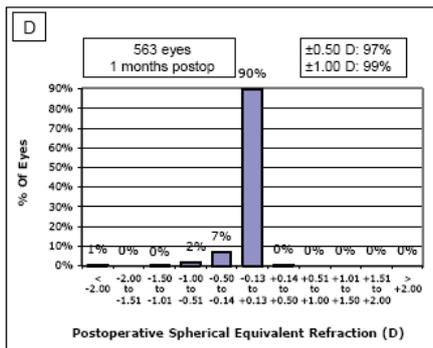
Uncorrected Distance Visual Acuity



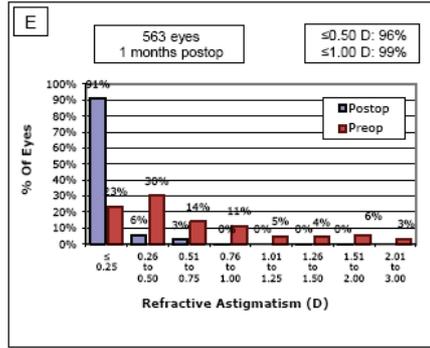
Change in Corrected Distance Visual Acuity



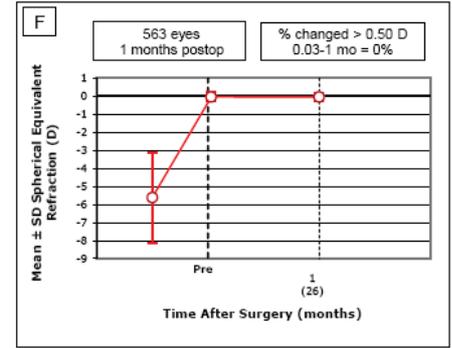
Spherical Equivalent Attempted vs Achieved



Spherical Equivalent Refractive Accuracy



Refractive Astigmatism



Stability of Spherical Equivalent Refraction

## Conclusion

- Innovation, adapting Technology advances helps in solving major eye health problems
- Collaboration with leaders in different fields will enhance the speed of translational research.

# Acknowledgements

- Dr H N Madhavan  
( Director Microbiology – Sankara Nethralaya)
- Dr Mohan Shankar Sivaprakasam  
( IIT – Madras)
- TCS

Thank You





Development of Indo-French  
Health Care Technology Network  
Programme-Role of  
CEFIPRA

Dr. Debapriya Dutta  
Director

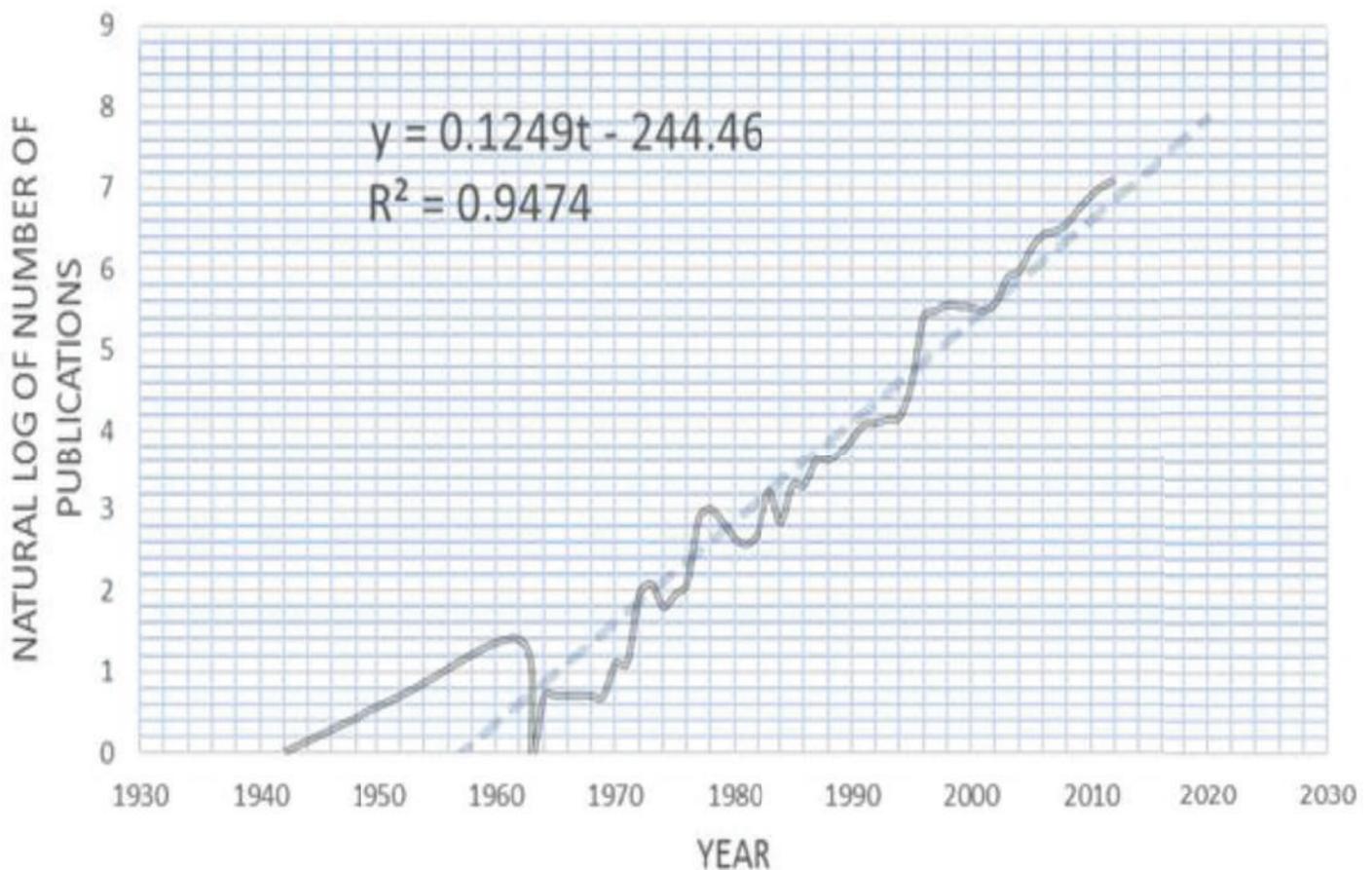


*Bonjour ! Good Morning!*

**Development of Indo-French Health care  
technology network programme- role of  
CEFIPRA**

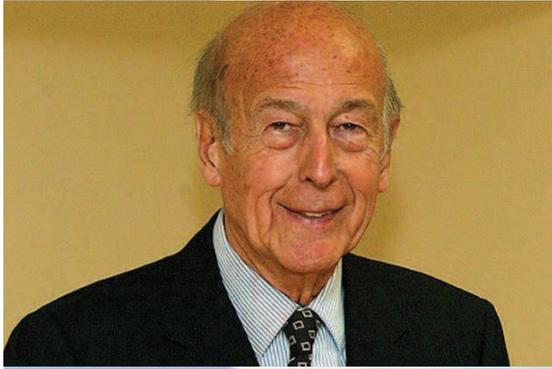
**Dr. Debapriya Dutta  
Director**

**INDO-FRENCH PUBLICATIONS TREND**

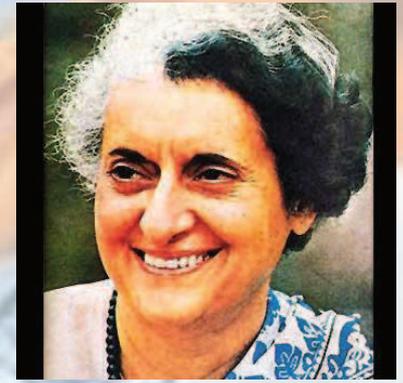


Source: Data from Scopus

# CEFIPRA: A Unique Model of Indo French S&T Cooperation



Mr. Valery Giscard d'Estaing



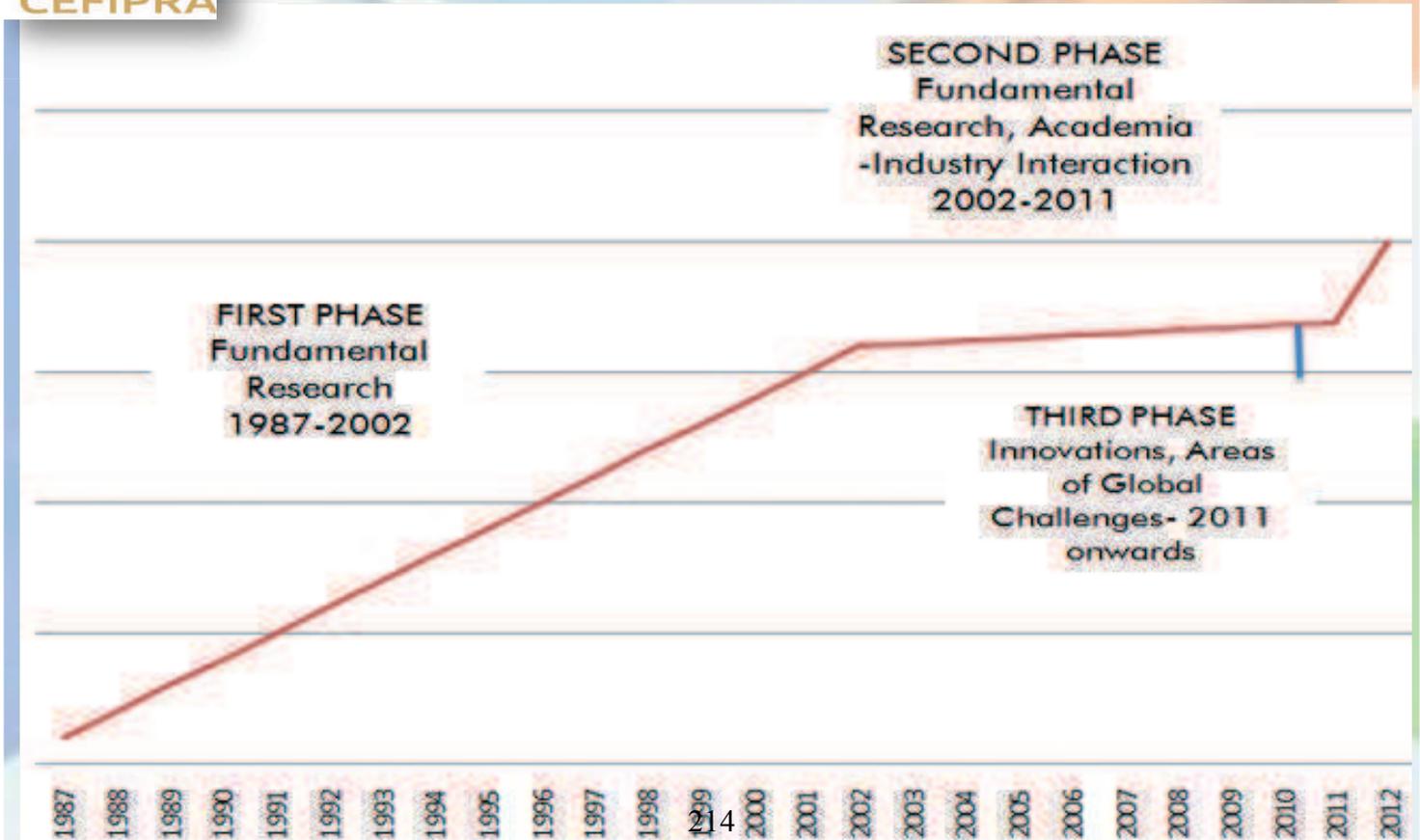
Mrs. Indira Gandhi

## The Mandate

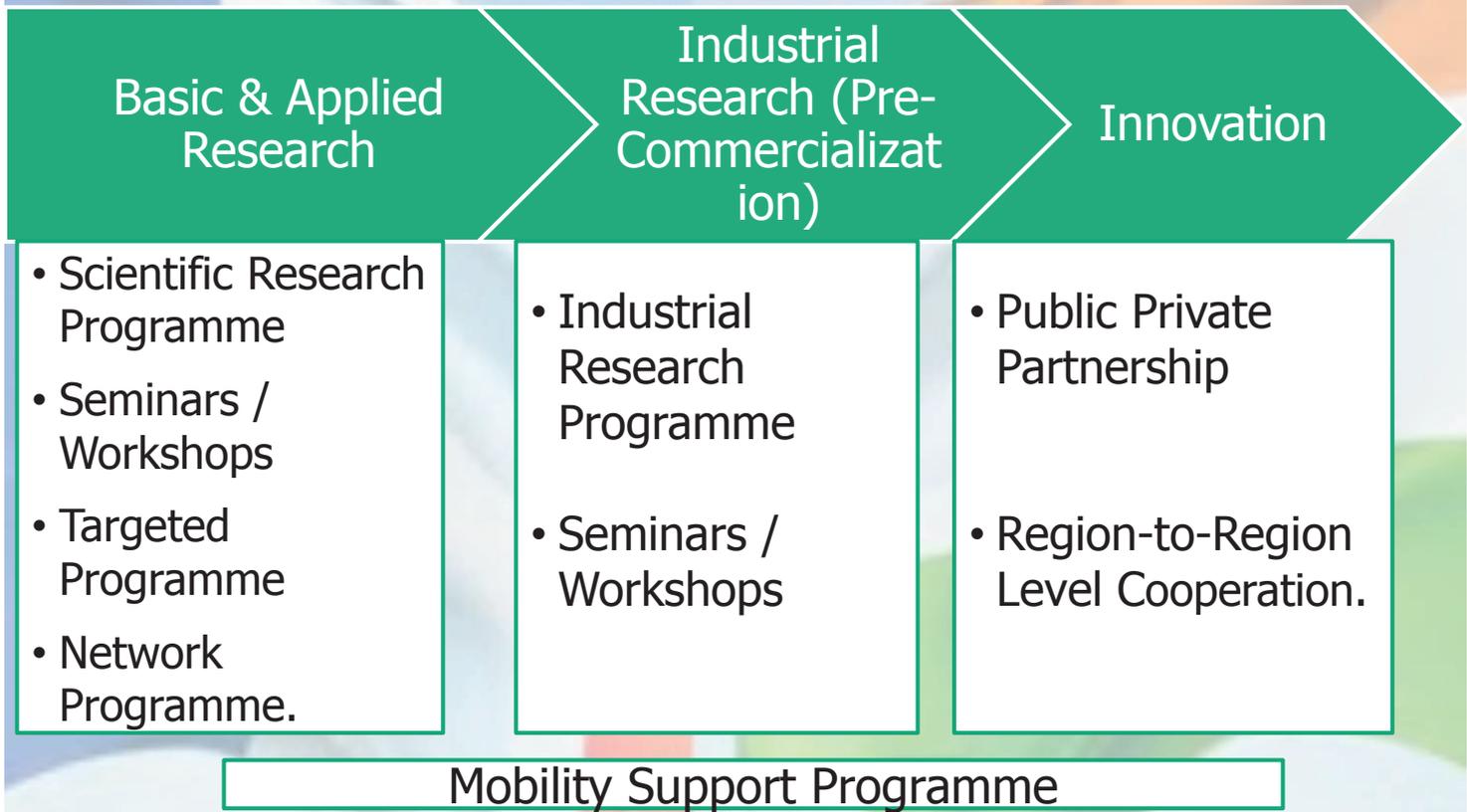
To promote collaborative research in advanced areas of Science & Technology between scientists and technologists of the two countries.



## Phases of Evolution



# Programme Profile across the Knowledge Innovation Chain



## Scientific Research Programme

### **Goal:-**

- Support High Quality Research Groups to **nurture scientific competency** in cutting-edge areas.

### **Areas :-**

- Mathematics, Computer and Information Sciences, Life and Health Sciences, Physics, Chemistry, Instrumentation, Earth and Planetary Sciences, Material Sciences, Environmental Sciences, Others (Water, Biotechnology, ICT)

### **Eligibility :-**

- Permanent Position in a University / Institution

### **Submission Deadlines:-**

- April 1 & October 1 of each year

# Industrial Research Programme

## Goal:-

### Improving Industrial Competitiveness

Leveraging research for industrial competitiveness

Industry Centric program

Targeted at SME/MSME

Developing/Improving product/process

## Eligibility Criteria:-

- At least one industry from one of the two countries
- At least one academic institution from the other country

## Domains:-

- All areas of Technology

## Submission Deadlines:-

- Accepts Proposals Throughout the Year

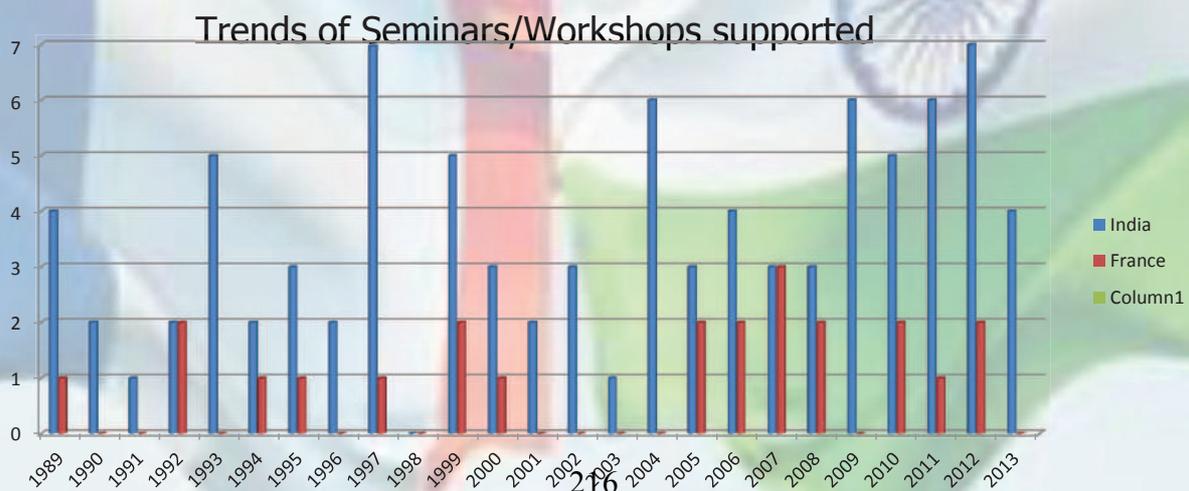
# Seminars / Workshops

## Objectives:-

- Platform to Share Knowledge / Expertise in an Advanced Area
- To Result in Collaborative Projects between Indian and French Scientists

## Areas of Interest:-

- Topics of Indian and French Interests
- Topics of Current Relevance



# Mobility Support

## **Raman Charpak Fellowship**

### ***Goal:-***

- To Increase Indian and French Laboratories Collaboration by Supporting the Highly Qualified PhD Student Mobility between India and France

### ***Eligibility Criteria:-***

- Applicants must be Indian or French Citizens
- Master's Degree in Science, Pursuing PhD in a Recognized University

### ***Research Domains:-***

- Life Science, Mathematical & Computational Sciences, Material Science, Physical Science, Chemical Science, Atmospheric and Earth Sciences, Engineering Sciences

# Mobility Support (contd..)

## **ESONN Training Programmes**

### ***Objectives:-***

- To Emphasize the Role of Laboratory Courses and
- To Highlight the Fundamental and Technological Advances in Specific Domains.

### ***Eligibility Criteria:-***

- Exclusively for Indian Students
- Pursuing PhD in a Recognized University

### ***Domains:-***

- Quantum NanoElectronics
- Interface between Physics and Biology

# Mobility support through Public Private partnership mode

- MOU has been signed with ANRT
- Joint call for proposal for placing Indian doctoral students in French Academia & industries will be launched shortly



## Targeted Programmes-CEFIPRA as facilitator

### ***Goal:-***

- Scientific Collaboration in Specific Areas of Mutual Interest

### ***Eligibility Criteria:-***

- Permanent Positions in University / Research Institutions for India
- French Participants Should be Recognized by the Concerned French Agencies

### ***Research Domains:-***

- Infectious Diseases, Engineering Science (Materials, Energy, Chemistry, Intelligent Transport System), NeuroSciences
- Big Data, Cyber Physical Systems, High Performance Computing
- Global Climate Change

# Targeted Programme under CEFIPRA

<u>FRENCH</u>	<u>INDIAN</u>	<u>NATURE OF COOPERATION</u>
ANR	DST	Call for proposals to support projects on “Infectious Diseases” and “Engineering Sciences”. 2 <sup>nd</sup> call focused on “Neurosciences” and “Engineering Sciences”.
INRA	DST	A project on “Adaptation of Irrigated Agriculture to Climate Change”
INRIA-CNRS	DST	1 <sup>st</sup> call focused on Big Data, Cyber Physical Systems, High Performance Computing 2 <sup>nd</sup> call focused on Big Data, High performance Computing & Computational Biology

Role of CEFIPRA: CEFIPRA provide a platform to the national funding agencies of both the countries to come together for supporting collaborative research activities in specific thematic areas of Science & Technology

## Linkages between Industry & Academia in 2+ 2 model

### BIRAC-CEFIPRA

- Indo-French Challenge-oriented call for proposal was launched in the following areas:
  - Molecular diagnostics for prediction of cardiac stroke.
  - Rapid diagnostics for Alzheimers and /or dementia in elderly or molecular diagnostics for detection of neurological disorders in neonates especially related to cerebral palsy.
  - Generation of new assistive technologies for mobility of physically challenged including elderly.

2<sup>nd</sup> call in the area of green & red Biotechnology will be launched soon

## **CEFIPRA's Public Private partnership programme**

### **EADS- CEFIPRA Aerospace Programme**

- Expression of Interest signed
- MoU would be signed soon
- Joint call for proposal would be launch thereafter

### **Astrium- CEFIPRA Aerospace Programme**

- **Letter of Intent has been signed**

## **Public Private Programme**

### **Indo-French Innovation Programme of Bpi-France & TDB**

- MOU has been signed
- Joint call for proposal will be launched shortly

### **ALSTOM-NEB Incubator Programme**

- The Programme would be supported both by Alstom & NEB, DST in partnership mode
- 4-5 projects will be supported every year
- One Incubator Centre will be establish

# Public Private Programme- CEFIPRA as Partner



## Funding Share

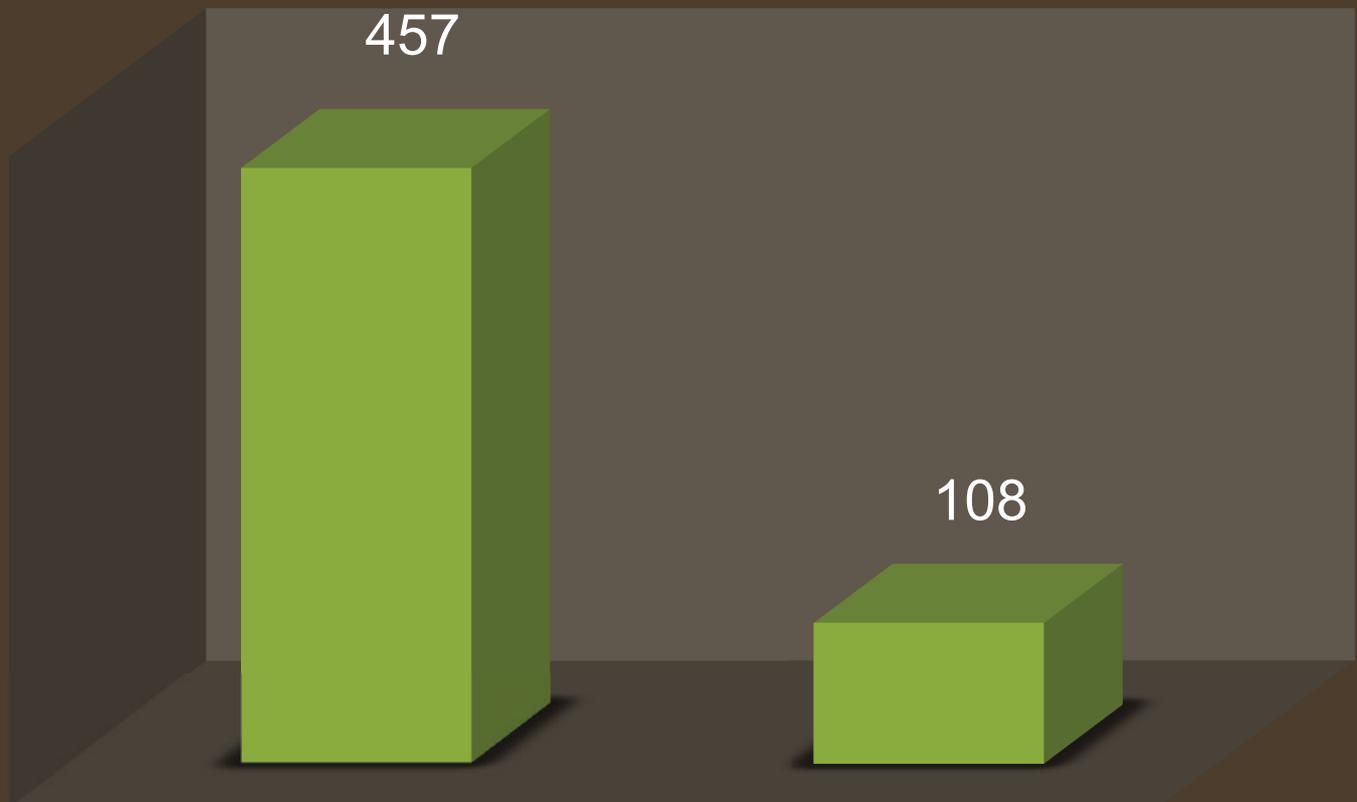
Equivalent amount of CEFIPRA support for a maximum period of 3 years as per agreement time to time



## Funding Share

Financial support for a maximum period of 3 years as per agreement time to time

Joint call for proposals On “sustainable habitat for hot and/ or humid climates”, launched on 15<sup>th</sup> November 2013.



Total No. of Projects Supported by CEFIPRA

No. of Projects supported in Biotechnology and Life & Health Sciences

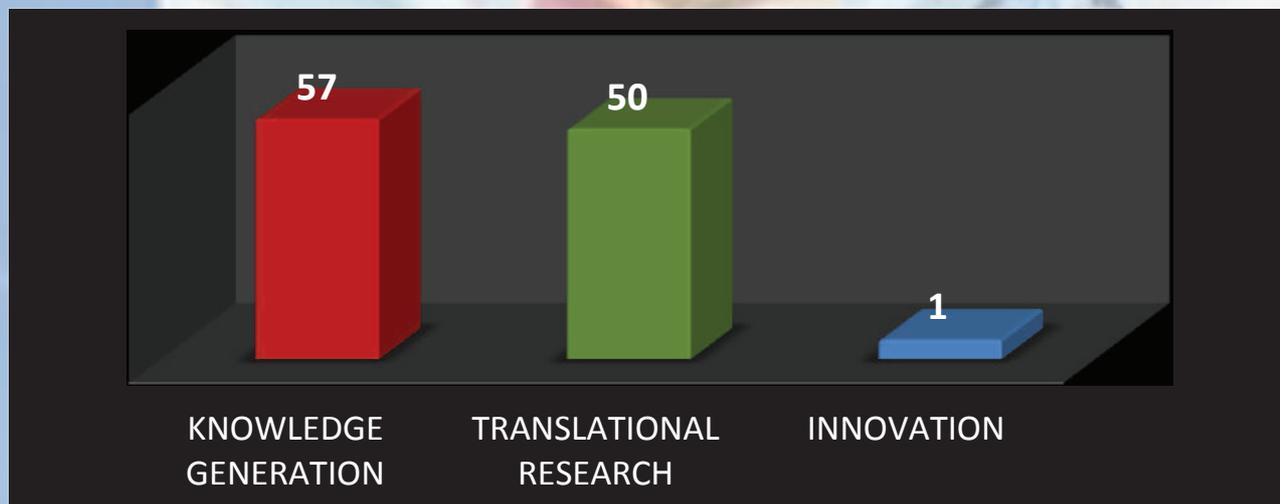
# CEFIPRA SUPPORTED HEALTH & LIFE SCIENCE PROJECTS ACROSS THE KNOWLEDGE INNOVATION CHAIN

Basic Research

Applied Research

Industrial Research

Product Development



Risk/ Uncertainty

## BASIC SCIENCE PROJECT

### Analysis of Protein Flexibility in Biological Recognition

#### Objectives

1) To study structural and dynamic aspects of molecular recognition in proteins, which underlies much of biology. This work involved creating datasets containing structures of protein-protein complexes and their individual components. The collaborators analyzed each structure to delineate flexible regions and determine what concerted movements they undergo upon biological complex formation.

2) The interface region was compared between the free, or unbound (U), and the bound (B) forms of proteins to identify changes in secondary structure, side-chain conformation, crystallographic temperature factors, accessible surface areas, etc. Project also targeted novel strategies to assist in computer docking of proteins and small molecules to proteins.

#### Potential for Knowledge Forward Chain

- More extensive mining of structural data
- Better data (crystallization processes)
- Understanding particular biological path ways
- Flexible protein docking: structure prediction
- Drug design: predicted structures as targets

#### Collaborators



**Pinak Chakrabarti,**  
Bose Institute Kolkata



**Charles Robert**  
Institut de Biologie Physico Chimique  
Paris

#### Knowledge Products Developed

- Homologous NDP kinases and their complexes with NDPs (71 complexes)
- The Structure Affinity Benchmark includes KD values (148)
- Permanent and transient homodimeric protein complexes along with KD's (315)
- Database of human hemoglobin tetraners (165)
- Flexbase: Systematic analysis of paired component structures

# Anti-factor H Autoantibody Associated Hemolytic Uremic Syndrome

## Objectives

- Validate anti-factor H antibody assay in India and establish a normal threshold
- Obtain genetic insights into the mechanisms of immunization against factor H by: (i) performing MHC haplotype determination of all patients; (ii) examine deficiency of CFHR1 in patients, relatives and controls
- Screen for mutations in genes implicated in susceptibility to HUS
- Determine microbial triggers associated with the disease, by (i) collection of clinical and biological data, (ii) parasitological and bacterial examination of stools, and screening for fecal shigatoxin, (iii) serological identity of infections
- • Study anti-factor H cellular immune response, through constitution of a peripheral blood mononuclear cell and plasma samples bank of patients with HUS and anti-FH IgG antibodies

## Collaborators



**Arvind Bagga**  
All India Institute of Medical Sciences, New Delhi



**Marie-Agnès Dragon-Durey**  
INSERM UMRS 82, Cordeliers Research Center, Paris



# Catalytic Antibodies in Immune-Mediated Disorders

## Objectives

### I. FVIII-hydrolyzing IgG

- Prevalence FVIII-hydrolyzing IgG in patients with acquired hemophilia and follow-up during progression of the disease
- Follow-up of FVIII-hydrolyzing IgG in patients with congenital hemophilia A who have developed FVIII inhibitors and are under protocols of 'immune tolerance induction'
- Generate monoclonal anti-FVIII antibodies with proteolytic activity to FVIII
- Screen for specific inhibitors of FVIII-hydrolyzing antibodies

### II. DNA-hydrolyzing IgG in patients with systemic lupus erythematosus

- Longitudinal follow-up of DNA-hydrolyzing IgG in patients with SLE
- To generate murine monoclonal anti-DNA antibodies with hydrolytic activity to DNA
- Topoisomerase I-hydrolyzing IgG in patients with scleroderma
- To investigate the presence of topoisomerase I-hydrolyzing IgG in patients with scleroderma

## Collaborators



**Valakunja Nagaraja**  
Indian Institute of Science, Bangalore



**Srinivas V. Kaveri**  
Immunopathologie et Immunointervention Thérapeutique, Paris

## Knowledge Products Developed

- Discovery of factor IX-hydrolyzing IgG in the plasma of patients with acquired hemophilia A
- Demonstration of presence and prevalence of topoisomerase-hydrolyzing IgG and DNA-hydrolyzing IgG in patients with scleroderma

## Mechanism based lead generation in cancer chemotherapy from natural products

### Objectives

The core objectives pursued under the project are listed as under:

- To conduct highly interdisciplinary project between chemistry and biology in India and France and develop novel anticancer compounds.
- Identify new selective apoptosis inducers for cancer cells, with special attention paid to NCEs.
- Synthesis and lead optimization of natural products with apoptosis restoring capacity for cancer cells.
- Develop new methodologies for screening compounds using HTS/MTS.
- Identify at least three lead compounds for further development as anticancer drugs.
- Conduct in vivo studies for selected compounds.

### Potential for Knowledge Forward Chain

The discovery of new chemical entities able to restore apoptosis selectively on cancer cells in a very promising new approach to anti-cancer drugs.

### Collaborators



**J. S. Yadav**

Indian Institute of Chemical Technology, Hyderabad



**René Grée**

Université de Rennes 1  
Rennes

### Knowledge Products Developed

- Efficient strategized to access the target molecules in different series.
- Small chemical libraries of designed compounds.
- Preliminary biological screening validated some of our working hypotheses, affording first series of active analogues.

## Evaluation of Cellular and Immune Response in Mice and Patients with Acute Promyelocytic Leukemia Treated with Arsenic Trioxide

### Objectives

The overall goal of this collaboration is to study the effects of novel agents used in the treatment of acute promyelocytic leukemia on the immune response through preclinical studies in mice models and ongoing clinical trials in patients. In specific:

- Study antibody responses to acute promyelocytic leukemia (APL) in mouse model of acute promyelocytic leukemia and in APL patients with newly diagnosed and relapsed acute promyelocytic leukemia treated with an arsenic trioxide based regimen.
- Study immune reconstitution and cellular response to APL in patients with newly diagnosed and relapsed acute promyelocytic leukemia treated with an arsenic trioxide based regimen.
- Study safety and efficacy of PML $\downarrow$ RAR $\beta$  targeted DNA vaccine as an adjunct to arsenic trioxide in the treatment of a mouse model of acute promyelocytic leukemia mice with and without all-trans retinoic acid.

### Collaborators



**Vikram Mathews**

Christian Medical College  
Vellore



**Christine Chomienne**

Institute of Universitaire Hématologie  
Paris

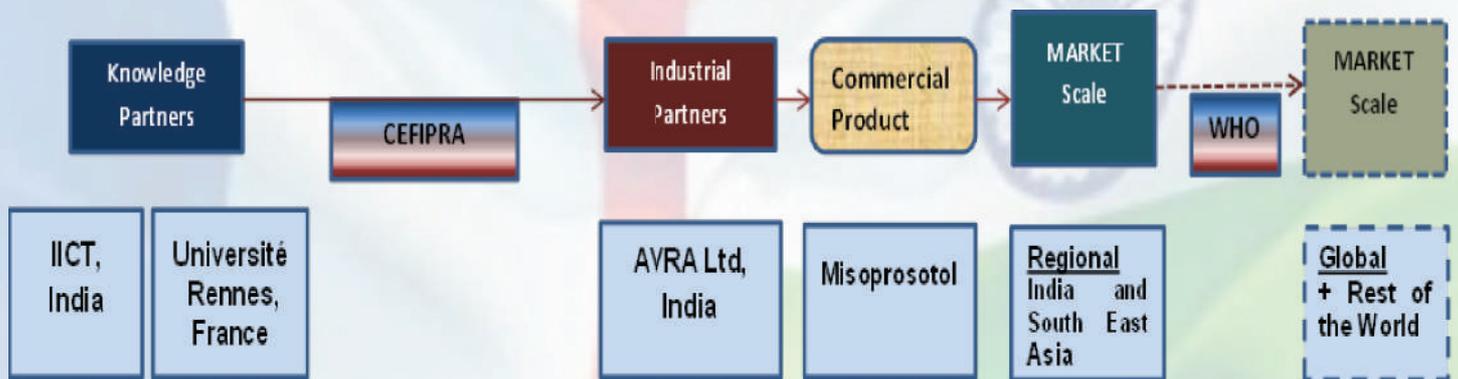
### Knowledge Products Developed

- Transfer of animal model of acute promyelocytic leukemia and APL transplantable model to India and establishment of the preclinical animal model of APL treated by arsenic trioxide
- Induction of the APL in FVBN mice
- Transfer of DNA plasmid vaccine and its evaluation in a mouse model
- Technology for in-house evaluation of antibody response to APL

# Value of CEFIPRA in Knowledge-Product Transformation

## Misoprosotol

- A Drug, Developed by Avra Labs India Under CEFIPRA Project
- Used in Conjunction with Non-Steroidal Anti-Inflammatory Drugs to Prevent Ulcers



## Connecting Individual Scientist for Indo-French network programme

As is

Bridging of individual scientist



To be

Research networking





## *The Goal of the Proposed Network Programme Of CEFIPRA*

# **Knowledge Linkages & Knowledge Forward chaining**



## **The Objectives**

To connect excellent research groups from India & France supported by the respective National agencies and CEFIPRA

To foster Interdisciplinary/ Intra disciplinary collaborative research & networking activities between identified groups

To generate world class collaborative research projects by establishing linkages between groups for solving specific problem

To develop mechanism for knowledge linkages & Knowledge forward chain by facilitating exchange of domain expertise, innovative ideas and technological knowhow between selected groups



**Mapping of Indian National S&T agencies against French alliance-  
Aviesan to develop Indo-French targeted programme for Health Care  
Technology**

National Institute of Health and Medical Research (INSERM)	Indian Institute of Medical Research (ICMR)
The National Centre for Scientific Research (CNRS)	Council of Scientific and industrial research ( CSIR), New Delhi
The Commissioner for Atomic Energy (CEA)	Department of Atomic Energy (DAE)
The National Institute of Agronomic Research (INRA)	Indian Council of Agricultural Research (ICAR)
Institute National Computer Science and Control (INRIA)	Council of Scientific and industrial research ( CSIR), New Delhi: Department of Bio-Technology ( DBT) under Ministry of Science & Technology: Department of Electronics & information Technology
The Institute of Research for Development (IRD)	Council of Scientific and industrial research ( CSIR), New Delhi Ministry of Earth Science
Pasteur Institute	Indian Institute of Medical Research (ICMR)
Conference of University Presidents (CPU)	University Grand Commission (UGC).



**THANK YOU!**

**MERCI !**

**CEFIPRA**

# Statistical Study of biological Networks

Christophe Ambroise

Statistique et Génome, CNRS & Université d'Évry  
Val d'Essonne

<http://stat.genopole.cnrs.fr/~cambroise>





# Statistical Study of biological Networks

C. Ambroise

Statistique et Génome, CNRS & Université d'Évry Val d'Essonne

octobre 2014, INAE/NATF Seminar

<http://stat.genopole.cnrs.fr/~cambroise>



1

## Outline

### Introduction

- Motivations
- Data

### Inference

- Problem
- Statistical models
- An example of penalty: multitask learning
- Example

### Concluding Remarks

## Introduction

Motivations

Data

## Inference

Problem

Statistical models

An example of penalty: multitask learning

Example

## Concluding Remarks

3

# Biological Networks

## Six chemical elements

Carbon, Hydrogen, Oxygen, Nitrogen, Phosphorus, Sulfur

## Four types of molecules in living systems

- ▶ carbohydrates: structural tissue, carry energy
- ▶ lipids: cell membranes, energy storage
- ▶ nucleic acids: information storage
- ▶ proteins: chemical workhorses of the cell

## Many interactions

Intra-cellular Biological network = set of interactions between various components from the cell

- ▶ Genes : Gene interaction networks
- ▶ Protein : Physical interaction between proteins, PPI networks
- ▶ Metabolites : Metabolic networks

# Biological Networks

## Six chemical elements

Carbon, Hydrogen, Oxygen, Nitrogen, Phosphorus, Sulfur

## Four types of molecules in living systems

- ▶ carbohydrates: structural tissue, carry energy
- ▶ lipids: cell membranes, energy storage
- ▶ nucleic acids: information storage
- ▶ proteins: chemical workhorses of the cell

## Many interactions

Intra-cellular Biological network = set of interactions between various components from the cell

- ▶ Genes : Gene interaction networks
- ▶ Protein : Physical interaction between proteins, PPI networks
- ▶ Metabolites : Metabolic networks

4

# Biological Networks

## Six chemical elements

Carbon, Hydrogen, Oxygen, Nitrogen, Phosphorus, Sulfur

## Four types of molecules in living systems

- ▶ carbohydrates: structural tissue, carry energy
- ▶ lipids: cell membranes, energy storage
- ▶ nucleic acids: information storage
- ▶ proteins: chemical workhorses of the cell

## Many interactions

Intra-cellular Biological network = set of interactions between various components from the cell

- ▶ Genes : Gene interaction networks
- ▶ Protein : Physical interaction between proteins, PPI networks
- ▶ Metabolites : Metabolic networks

# Biological Networks

## Six chemical elements

Carbon, Hydrogen, Oxygen, Nitrogen, Phosphorus, Sulfur

## Four types of molecules in living systems

- ▶ carbohydrates: structural tissue, carry energy
- ▶ lipids: cell membranes, energy storage
- ▶ nucleic acids: information storage
- ▶ proteins: chemical workhorses of the cell

## Many interactions

Intra-cellular Biological network = set of interactions between various components from the cell

- ▶ Genes : Gene interaction networks
- ▶ Protein : Physical interaction between proteins, PPI networks
- ▶ Metabolites: : Metabolic networks

4

# Biological networks

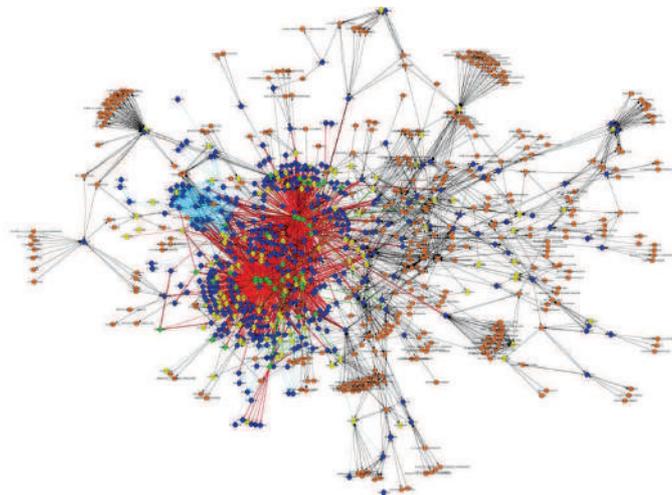


Figure : Regulatory network identified in mammalian cells: highly structured

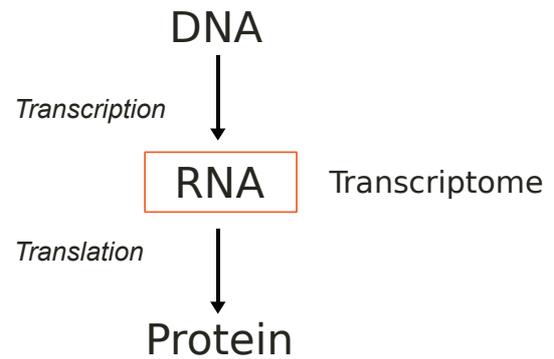
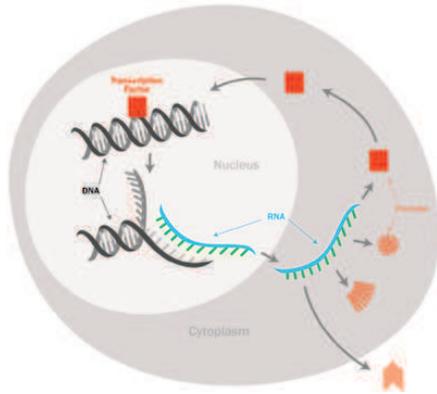
Mathematical representation = graph composed of nodes and edge

1. Inference: Inferring the structure of a graph based on observations at its nodes
2. Structure: Understanding the global structure of an observed graph.
3. ...

# What can we observe ?

## Gene expression

Process of information transmission from DNA to proteins



## Differential regulation of the gene expression

- ▶ among tissues,
- ▶ among developmental stages,
- ▶ or in response to environmental signals.

6

# Outline

## Introduction

Motivations

Data

## Inference

Problem

Statistical models

An example of penalty: multitask learning

Example

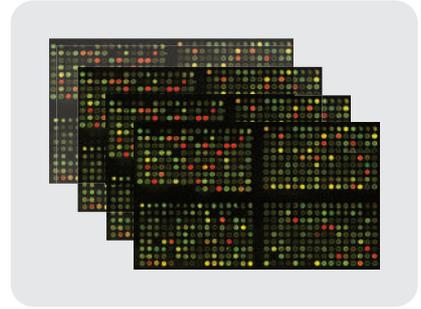
## Concluding Remarks

# How is this measured? (1)

Microarray technology: parallel measurement of many biological features



signal processing



Matrix of features  $n \ll p$

Expression levels of  $p$  probes are simultaneously monitored for  $n$  individuals

pretreatment

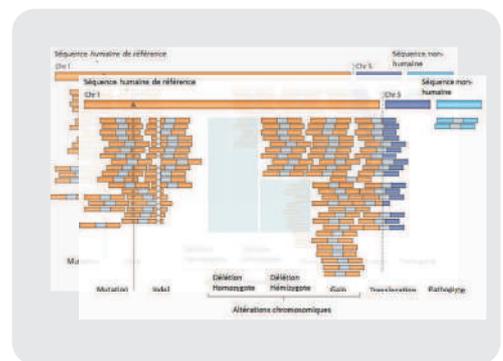
$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & x_1^3 & \dots & x_1^p \\ \vdots & & & & \\ x_n^1 & x_n^2 & x_n^3 & \dots & x_n^p \end{pmatrix}$$

# How is this measured? (2)

Next Generation Sequencing: parallel measurement of **even** many **more** biological features



assembling



Matrix of features  $n \lll p$

Expression counts are extracted from small repeated sequences and monitored for  $n$  individuals

pretreatment

$$\mathbf{X} = \begin{pmatrix} k_1^1 & k_1^2 & k_1^3 & \dots & k_1^p \\ \vdots & & & & \\ k_n^1 & k_n^2 & k_n^3 & \dots & k_n^p \end{pmatrix}$$

## Introduction

- Motivations
- Data

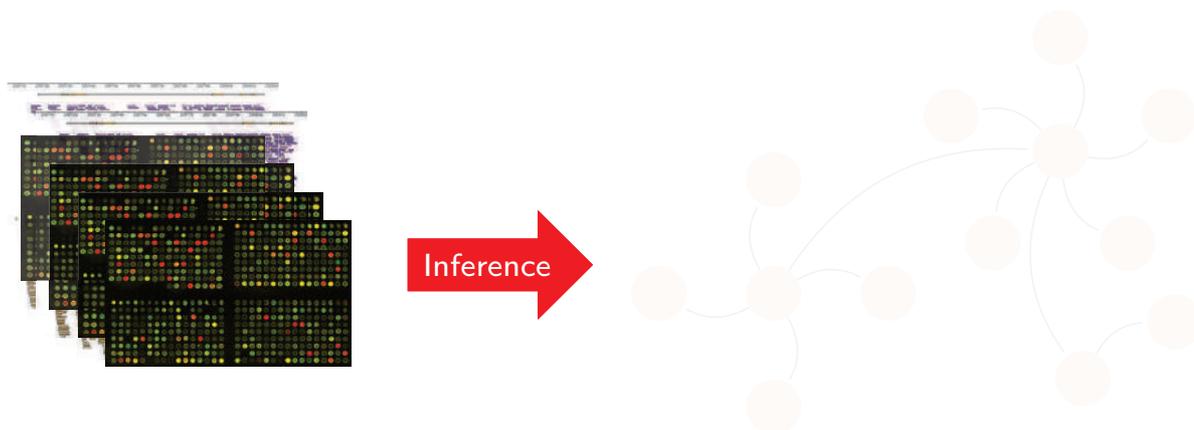
## Inference

- Problem
- Statistical models
- An example of penalty: multitask learning
- Example

## Concluding Remarks

10

# The problem at hand



≈ 10s/100s microarray/sequencing experiments

≈ 1000s probes (“genes”)

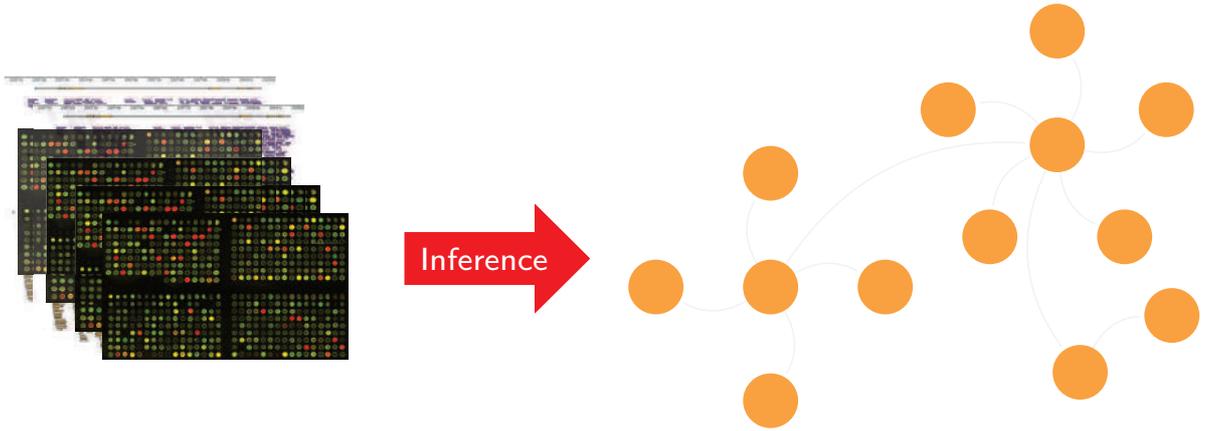
## Question ?

How to infer gene regulation networks from gene expression data ?

## Statistician’s answer

Estimating the dependence between gene expression profiles inform us about regulation mechanisms between genes

# The problem at hand



≈ 10s/100s microarray/sequencing experiments  
≈ 1000s probes (“genes”)

## Question ?

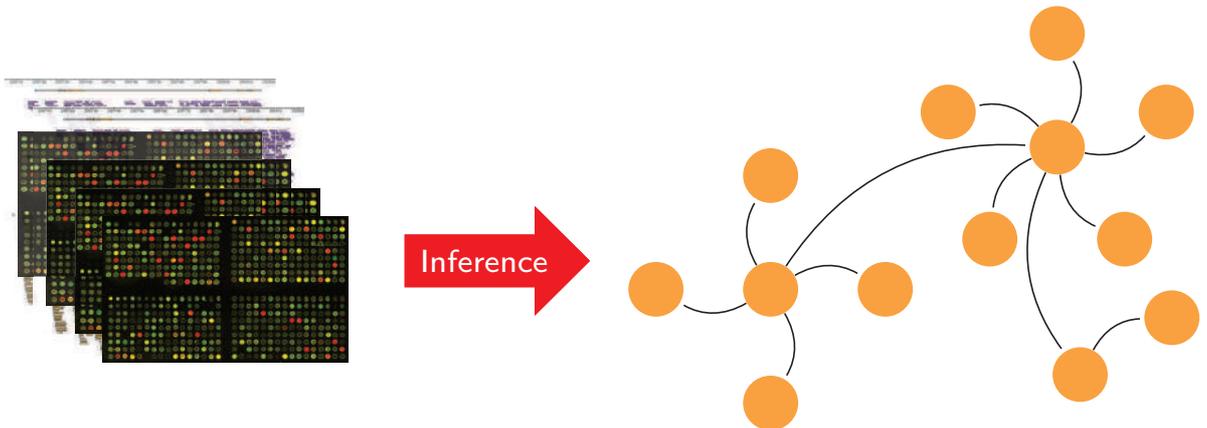
How to infer gene regulation networks from gene expression data ?

## Statistician’s answer

Estimating the dependence between gene expression profiles inform us about regulation mechanisms between genes

11

# The problem at hand



≈ 10s/100s microarray/sequencing experiments  
≈ 1000s probes (“genes”)

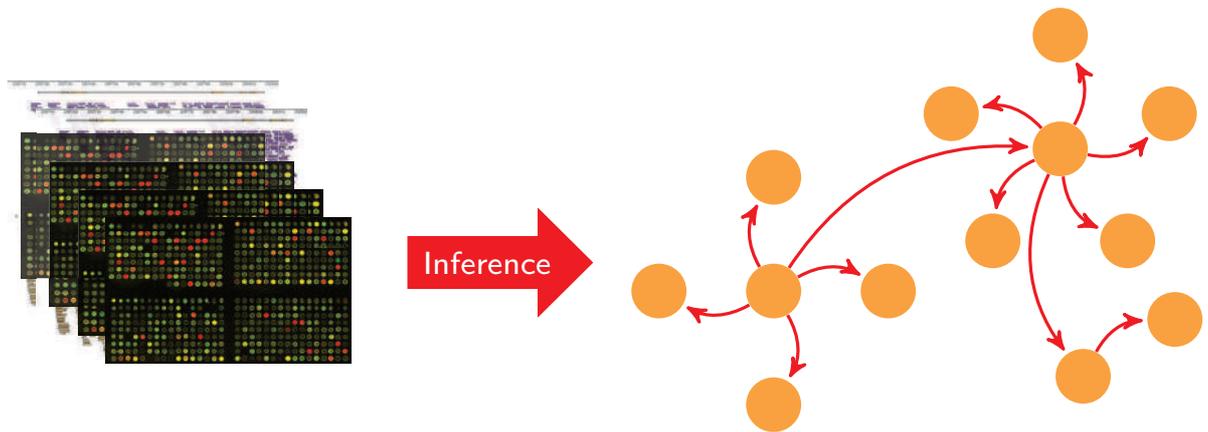
## Question ?

How to infer gene regulation networks from gene expression data ?

## Statistician’s answer

Estimating the dependence between gene expression profiles inform us about regulation mechanisms between genes

# The problem at hand



≈ 10s/100s microarray/sequencing experiments  
≈ 1000s probes ("genes")

## Question ?

How to infer gene regulation networks from gene expression data ?

## Statistician's answer

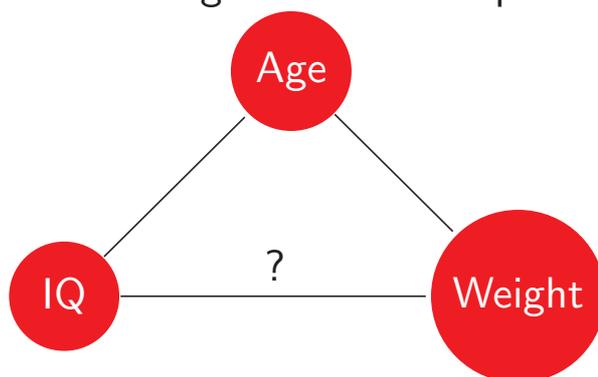
Estimating the dependence between gene expression profiles inform us about regulation mechanisms between genes

11

# Which dependence?

## Spurious links

Considering conditional dependence may avoid some spurious links



## Gaussian Model

- ▶ Correlation is a direct measure of dependence
- ▶ Partial Correlation is a direct measure of conditional dependence

The **Gaussian model is really convenient** for estimating (conditional) dependence

## Introduction

Motivations  
Data

## Inference

Problem  
Statistical models  
An example of penalty: multitask learning  
Example

## Concluding Remarks

13

# The graphical models: general settings

## Assumption

A microarray can be represented as a **multivariate Gaussian** vector  
 $X = (X(1), \dots, X(p)) \in \mathbb{R}^p$ .

## Graphical interpretation

$i$

$j$



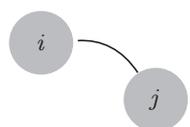
conditional dependency between  $X(i)$  and  $X(j)$

# The graphical models: general settings

## Assumption

A microarray can be represented as a **multivariate Gaussian** vector  $X = (X(1), \dots, X(p)) \in \mathbb{R}^p$ .

## Graphical interpretation



conditional dependency between  $X(i)$  and  $X(j)$

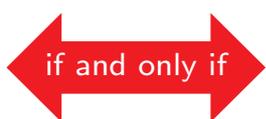
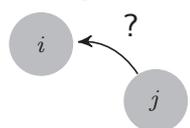
14

# The graphical models: general settings

## Assumption

A microarray can be represented as a **multivariate Gaussian** vector  $X = (X(1), \dots, X(p)) \in \mathbb{R}^p$ .

## Graphical interpretation



conditional dependency between  $X_t(i)$  and  $X_{t-1}(j)$

# The Gaussian model for an i.i.d. sample

Let

- ▶  $X \sim \mathcal{N}(\mathbf{0}_p, \Sigma)$  with  $X_1, \dots, X_n$  **i.i.d.** copies of  $X$ ,
- ▶  $\mathbf{X}$  be the  $n \times p$  matrix whose  $k$ th row is  $X_k$ ,
- ▶  $\Theta = (\theta_{ij})_{i,j \in \mathcal{P}} \triangleq \Sigma^{-1}$  be the **concentration matrix**.

## Graphical interpretation

Since  $\text{cor}_{ij|\mathcal{P}\setminus\{i,j\}} = -\theta_{ij}/\sqrt{\theta_{ii}\theta_{jj}}$  for  $i \neq j$ ,

$$X(i) \perp\!\!\!\perp X(j) | X(\mathcal{P} \setminus \{i, j\}) \Leftrightarrow \begin{cases} \theta_{ij} = 0 \\ \text{or} \\ \text{edge } (i, j) \notin \text{network.} \end{cases}$$

$\rightsquigarrow \Theta$  describes the undirected graph of **conditional dependencies**.

15

# The general statistical approach

Let  $\Theta$  be the parameters to infer (the edges).

## A penalized likelihood approach

$$\hat{\Theta}_\lambda = \arg \max_{\Theta} \mathcal{L}(\Theta; \text{data}) - \lambda \text{pen}(\Theta, \mathbf{Z}),$$

- ▶  $\mathcal{L}$  is the model log-likelihood,
- ▶  $\mathbf{Z}$  is a latent or additional variable,
- ▶  $\text{pen}()$  is a penalty function tuned by  $\lambda > 0$ .

It performs

1. *regularization* (needed when  $n \ll p$ ),
2. *selection* (sparsity induced when using  $\ell_1$ -norm),
3. *model-driven inference* (penalty adapted according to  $\mathbf{Z}$ ).

# The general statistical approach

Let  $\Theta$  be the parameters to infer (the edges).

## A penalized likelihood approach

$$\hat{\Theta}_\lambda = \arg \max_{\Theta} \mathcal{L}(\Theta; \text{data}) - \lambda \text{pen}(\Theta, \mathbf{Z}),$$

- ▶  $\mathcal{L}$  is the model log-likelihood,
- ▶  $\mathbf{Z}$  is a latent or additional variable,
- ▶  $\text{pen}()$  is a **penalty function** tuned by  $\lambda > 0$ .

It performs

1. *regularization* (needed when  $n \ll p$ ),
2. *selection* (sparsity induced when using  $\ell_1$ -norm),
3. *model-driven inference* (penalty adapted according to  $\mathbf{Z}$ ).

16

## A first approach: Neighborhood selection

Let

- ▶  $\mathbf{X}_i$  be the  $i$ th column of  $\mathbf{X}$ ,
- ▶  $\mathbf{X}_{\setminus i}$  be  $\mathbf{X}$  deprived of  $\mathbf{X}_i$ .

$$\mathbf{X}_i = \mathbf{X}_{\setminus i} \boldsymbol{\beta} + \varepsilon, \quad \text{where } \beta_j = -\frac{\theta_{ij}}{\theta_{ii}}.$$

### Meinshausen and Bühlman, 2006

Since  $\text{sign}(\text{cor}_{ij|\mathcal{P}\setminus\{i,j\}}) = \text{sign}(\beta_j)$ , select the neighbors of  $i$  with

$$\text{Argmin}_{\boldsymbol{\beta}} \frac{1}{n} \|\mathbf{X}_i - \mathbf{X}_{\setminus i} \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_1}.$$

- ↪ The sign pattern of  $\Theta_\lambda$  is inferred after a **symmetrization** step.
- ↪ Can be formulated as the maximisation of a **pseudo-likelihood**

## Second approach: Graphical Lasso

Banerjee *et al.*, JMLR 2008

$$\hat{\Theta}_\lambda = \underset{\Theta}{\operatorname{Argmax}} \mathcal{L}_{\text{iid}}(\Theta; \mathbf{S}) - \lambda \|\Theta\|_{\ell_1},$$

efficiently solved by the graphical Lasso of Friedman *et al.*, 2008.

Ambroise, Chiquet, Matias, *EJS* 2009

Use **adaptive** penalty parameters for different coefficients

$$\mathcal{L}_{\text{iid}}(\Theta; \mathbf{S}) - \lambda \|\mathbf{P}_Z \star \Theta\|_{\ell_1},$$

where  $\mathbf{P}_Z$  is a matrix of weights depending on the underlying clustering  $\mathbf{Z}$ .

18

## Second approach: Graphical Lasso

Banerjee *et al.*, JMLR 2008

$$\hat{\Theta}_\lambda = \underset{\Theta}{\operatorname{Argmax}} \mathcal{L}_{\text{iid}}(\Theta; \mathbf{S}) - \lambda \|\Theta\|_{\ell_1},$$

efficiently solved by the graphical Lasso of Friedman *et al.*, 2008.

Ambroise, Chiquet, Matias, *EJS* 2009

Use **adaptive** penalty parameters for different coefficients

$$\tilde{\mathcal{L}}_{\text{iid}}(\Theta; \mathbf{S}) - \lambda \|\mathbf{P}_Z \star \Theta\|_{\ell_1},$$

where  $\mathbf{P}_Z$  is a matrix of weights depending on the underlying clustering  $\mathbf{Z}$ .

## Introduction

- Motivations
- Data

## Inference

- Problem
- Statistical models
- An example of penalty: multitask learning
- Example

## Concluding Remarks

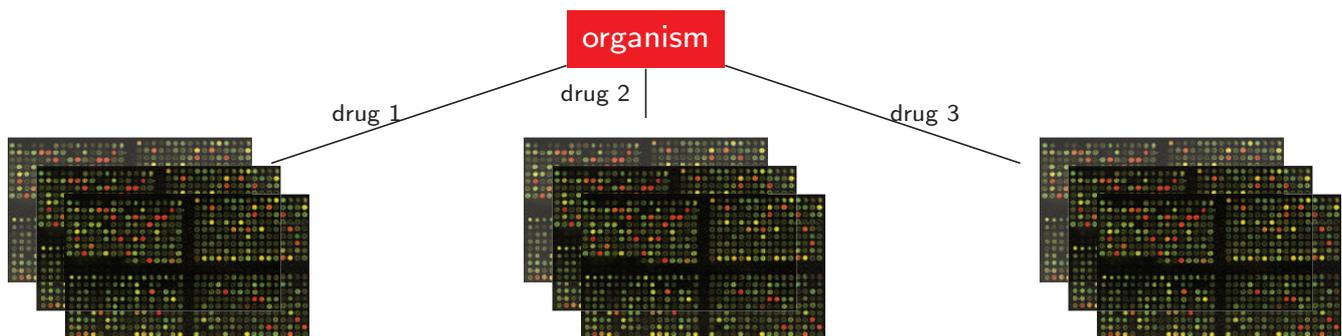
19

# Handling the scarcity of data

By collecting as many observations as possible

## Multitask learning

How should we merge the data?

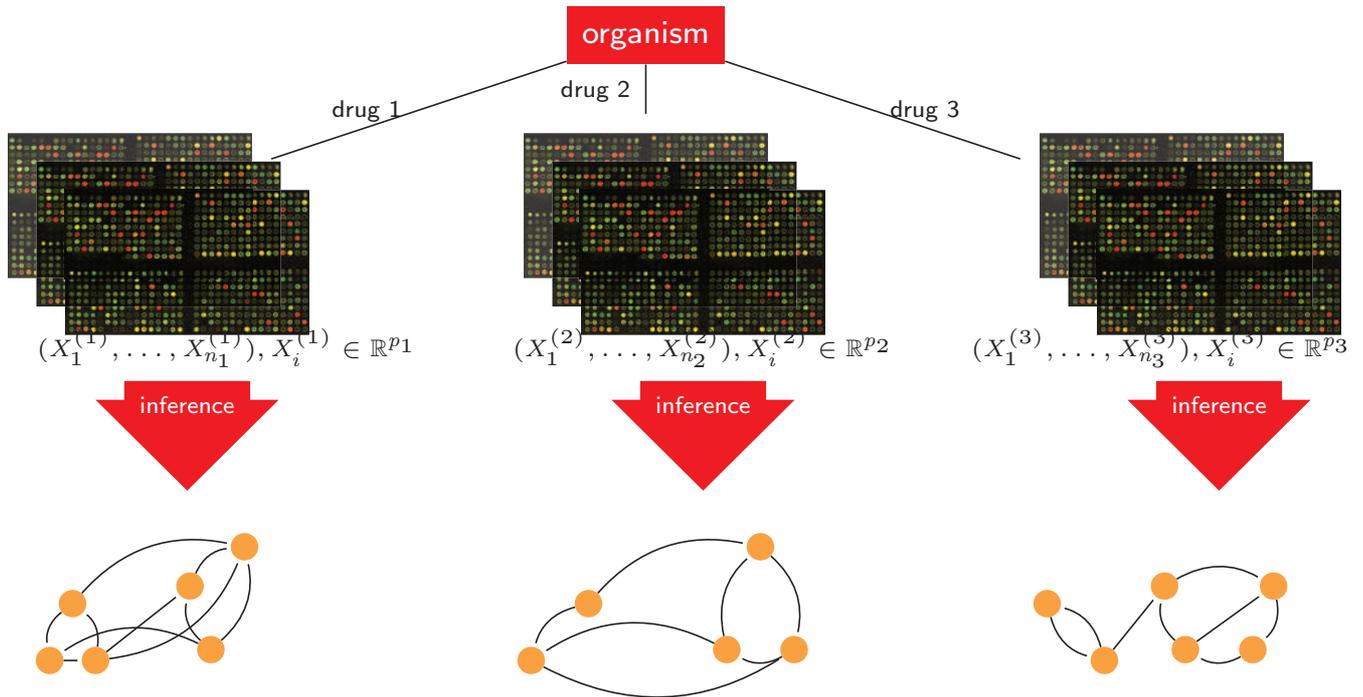


# Handling the scarcity of data

By collecting as many observations as possible

## Multitask learning

by inferring each network **independently**



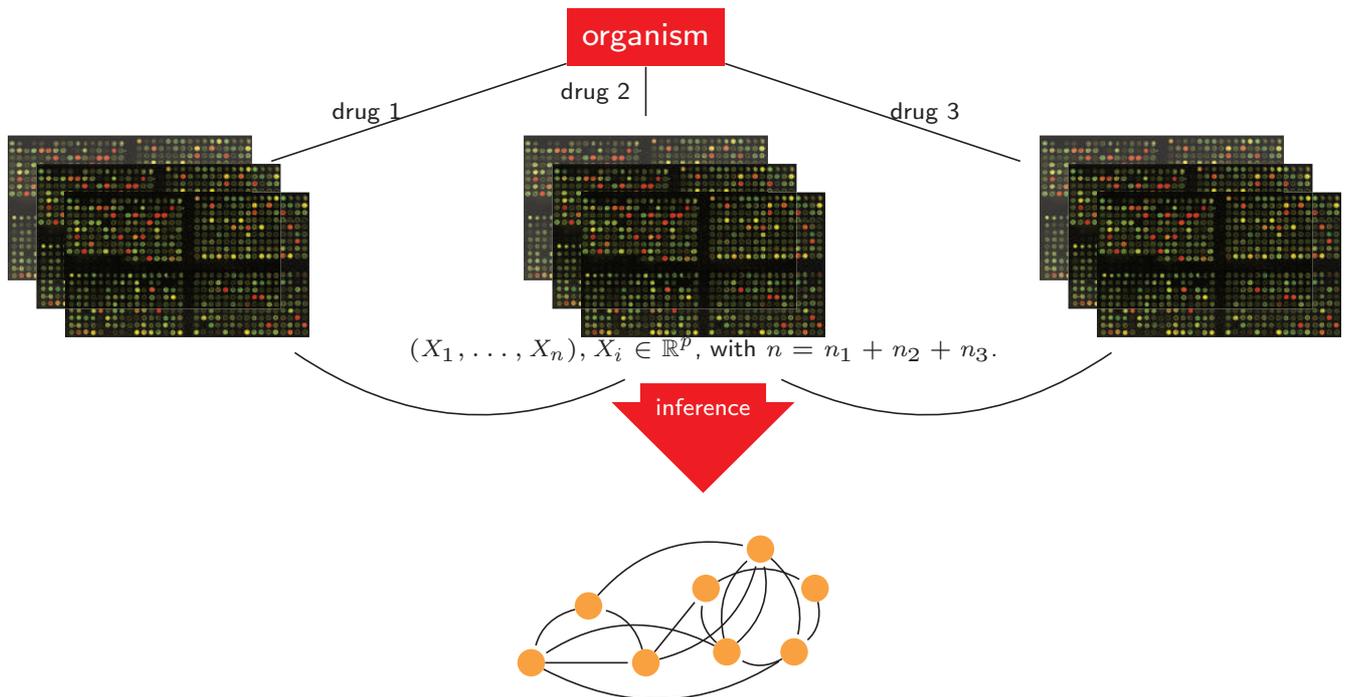
20

# Handling the scarcity of data

By collecting as many observations as possible

## Multitask learning

by **pooling** all the available data



246

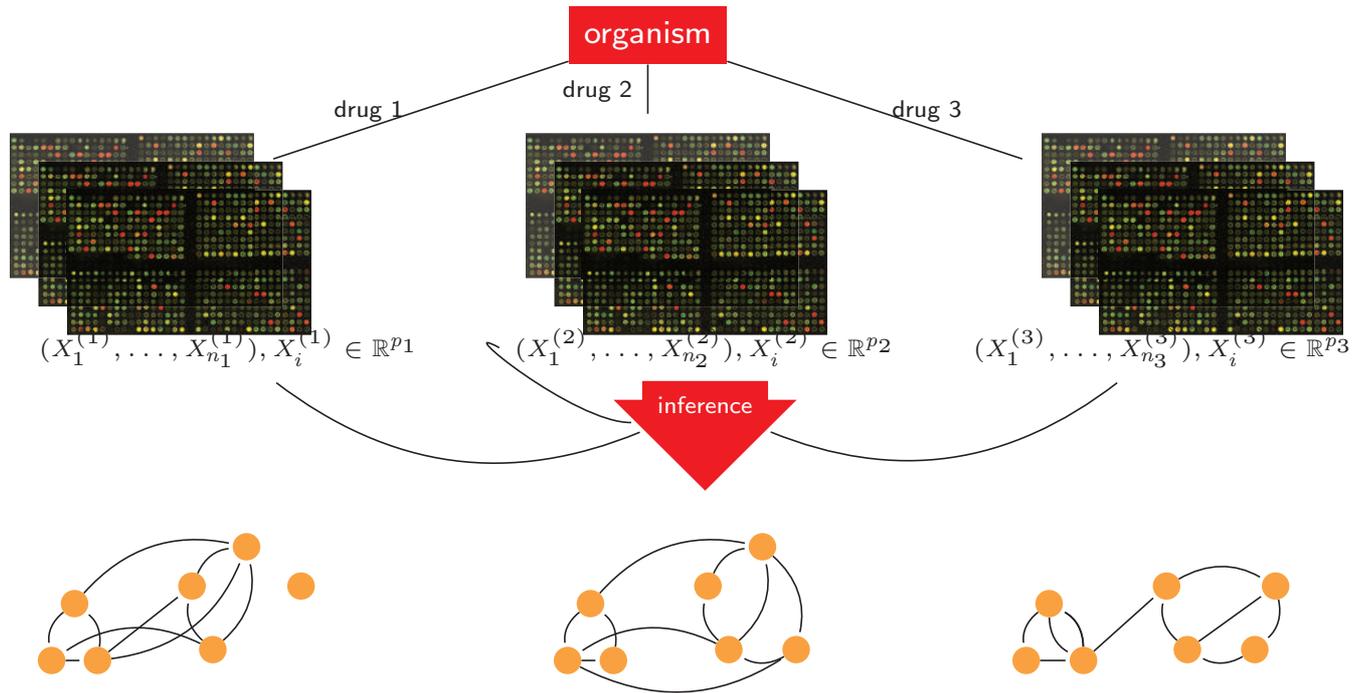
20

# Handling the scarcity of data

By collecting as many observations as possible

## Multitask learning

by **breaking** the separability



20

## Coupling related problems

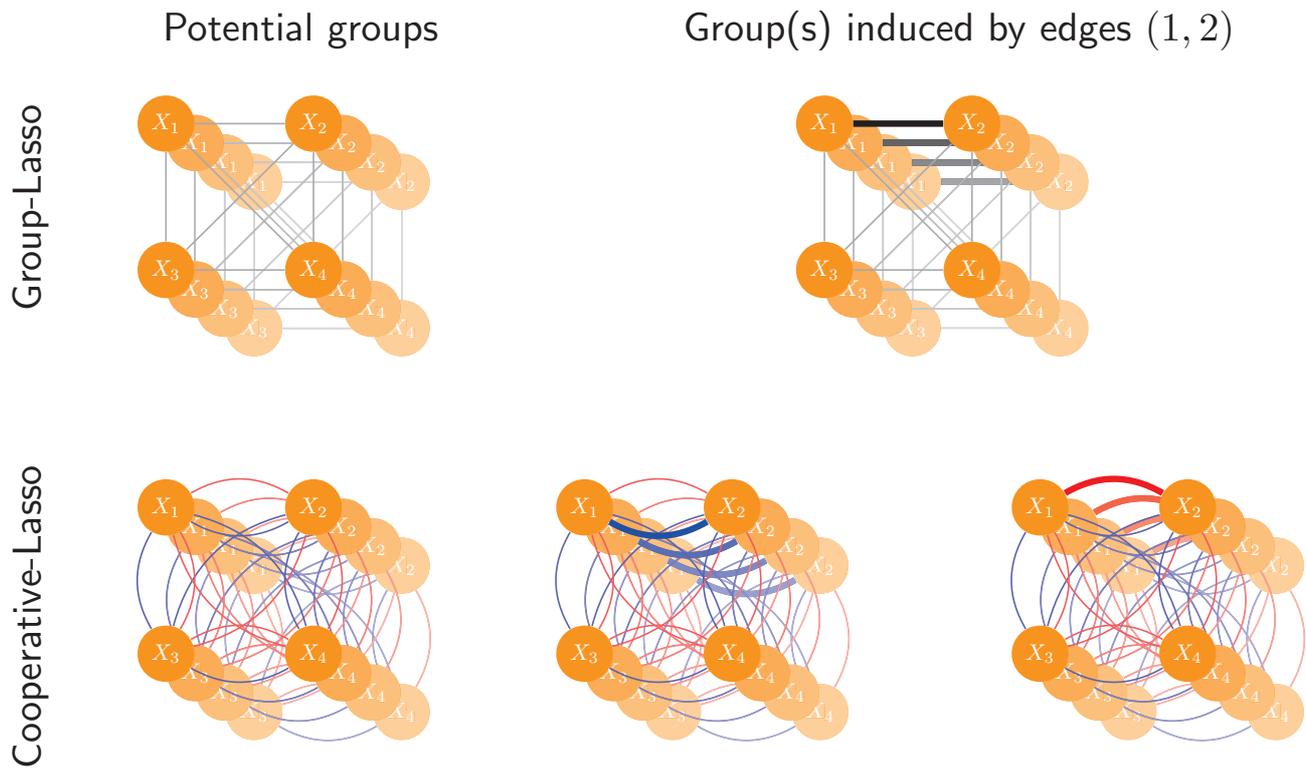
Consider

- ▶  $T$  samples concerning the expressions of the same  $p$  genes,
- ▶  $X_1^{(t)}, \dots, X_{n_t}^{(t)}$  is the  $t^{\text{th}}$  sample drawn from  $\mathcal{N}(\mathbf{0}_p, \Sigma^{(t)})$ , with covariance matrix  $\mathbf{S}^{(t)}$ .

### Multiple samples setup

Ignoring the relationships between the tasks leads to

$$\underset{\Theta^{(t)}, t=1, \dots, T}{\text{Argmax}} \sum_{t=1}^T \mathcal{L}(\Theta^{(t)}; \mathbf{S}^{(t)}) - \lambda \text{pen}(\Theta^{(t)})$$



22

## Coupling problems through penalty

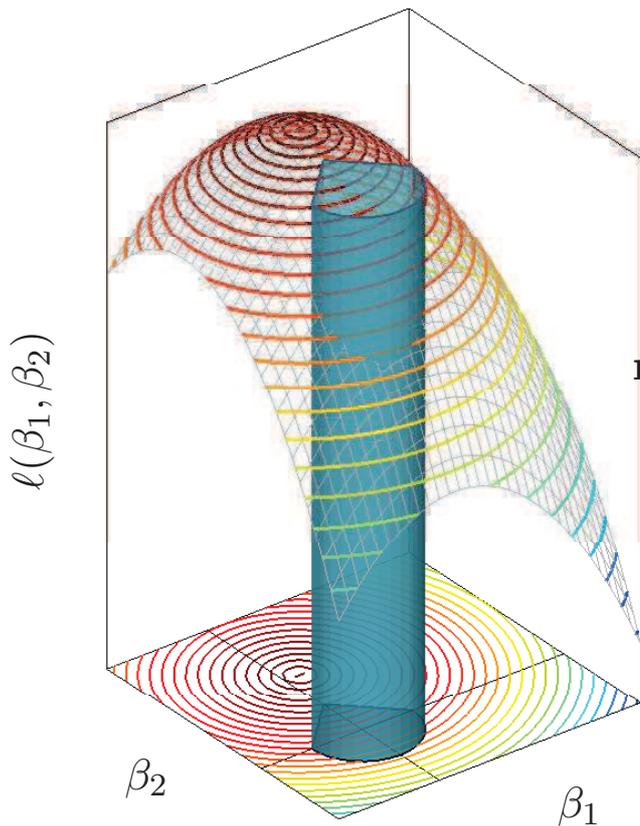
### Groups definition

- ▶ Groups are the  $T$ -tuple composed by the  $(i, j)$  entries of each  $\hat{\theta}^{(t)}, t = 1, \dots, T$ .
- ▶ Most relationships between the genes are kept or removed across all tasks simultaneously.

### The graphical group-Lasso

$$\max_{\hat{\theta}^{(t)}, t=1, \dots, T} \sum_{t=1}^T \tilde{\mathcal{L}} \left( \Theta^{(t)}; \mathbf{S}^{(t)} \right) - \lambda \sum_{\substack{i, j \in \mathcal{P} \\ i \neq j}} \left( \sum_{t=1}^T \left( \theta_{ij}^{(t)} \right)^2 \right)^{1/2}.$$

# A geometric view of sparsity



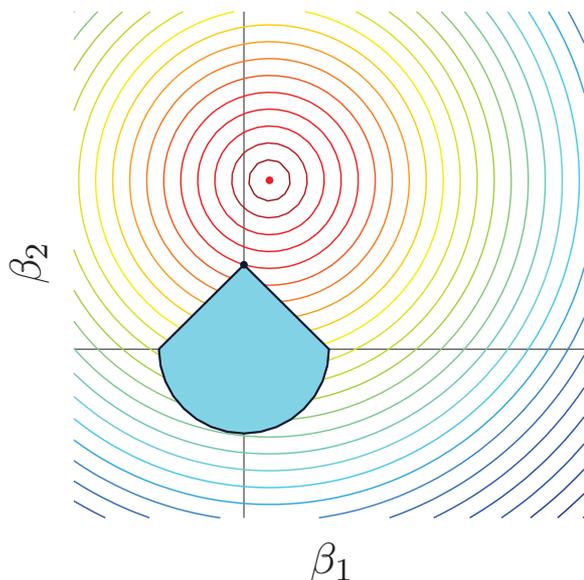
$$\text{minimize}_{\beta_1, \beta_2} -\ell(\beta_1, \beta_2) + \lambda\Omega(\beta_1, \beta_2)$$

$$\Updownarrow$$

$$\begin{cases} \text{maximize}_{\beta_1, \beta_2} & \ell(\beta_1, \beta_2) \\ \text{s.t.} & \Omega(\beta_1, \beta_2) \leq c \end{cases}$$

24

# A geometric view of sparsity



$$\text{minimize}_{\beta_1, \beta_2} -\ell(\beta_1, \beta_2) + \lambda\Omega(\beta_1, \beta_2)$$

$$\Updownarrow$$

$$\begin{cases} \text{maximize}_{\beta_1, \beta_2} & \ell(\beta_1, \beta_2) \\ \text{s.t.} & \Omega(\beta_1, \beta_2) \leq c \end{cases}$$

249

24

## Theorem

Under classical assumptions, the coop-Lasso is asymptotically unbiased and has the property of exact support recovery

$$\hat{\beta}_n^{\text{coop}} \xrightarrow{P} \beta^* \quad \text{and} \quad \mathbb{P} \left( \mathcal{S}(\hat{\beta}_n^{\text{coop}}) = \mathcal{S} \right) \rightarrow 1,$$

for every sequence  $\lambda_n$  such that  $\lambda_n = \lambda_0 n^{-\gamma}$ ,  $\gamma \in (0, 1/2)$ .



Chiquet, J., Grandvalet, Y. and Charbonnier, C. *The Annals of Applied Statistics* (2012).  
Sparsity in sign-coherent groups of variables via the cooperative-Lasso.

## Warning

Detection limits in Gaussian regression models vary with the problem

1. Prediction
2. Estimation
3. Testing (related to selection)
4. Detection

Verzelen (2012) EJS shows that a transition phase in terms of minimax risk is observed between  $k \log \frac{p}{k} \leq n$  and  $k \log \frac{p}{k} > n \log n$  where  $k$  is the size of the support.

## Warning

Detection limits in Gaussian regression models vary with the problem

1. Prediction
2. Estimation
3. Testing (related to selection)
4. Detection

Verzelen (2012) EJS shows that a transition phase in terms of minimax risk is observed between  $k \log \frac{p}{k} \leq n$  and  $k \log \frac{p}{k} > n \log n$  where  $k$  is the size of the support.

26

## Warning

Detection limits in Gaussian regression models vary with the problem

1. Prediction
2. Estimation
3. Testing (related to selection)
4. Detection

Verzelen (2012) EJS shows that a transition phase in terms of minimax risk is observed between  $k \log \frac{p}{k} \leq n$  and  $k \log \frac{p}{k} > n \log n$  where  $k$  is the size of the support.

## Warning

Detection limits in Gaussian regression models vary with the problem

1. Prediction
2. Estimation
3. Testing (related to selection)
4. Detection

Verzelen (2012) EJS shows that a transition phase in terms of minimax risk is observed between  $k \log \frac{p}{k} \leq n$  and  $k \log \frac{p}{k} > n \log n$  where  $k$  is the size of the support.

26

## Warning

Detection limits in Gaussian regression models vary with the problem

1. Prediction
2. Estimation
3. Testing (related to selection)
4. Detection

Verzelen (2012) EJS shows that a transition phase in terms of minimax risk is observed between  $k \log \frac{p}{k} \leq n$  and  $k \log \frac{p}{k} > n \log n$  where  $k$  is the size of the support.

## Introduction

Motivations

Data

## Inference

Problem

Statistical models

An example of penalty: multitask learning

Example

## Concluding Remarks

27

# Application to breast cancer

## Dataset : Metastatic relapse of breast cancer



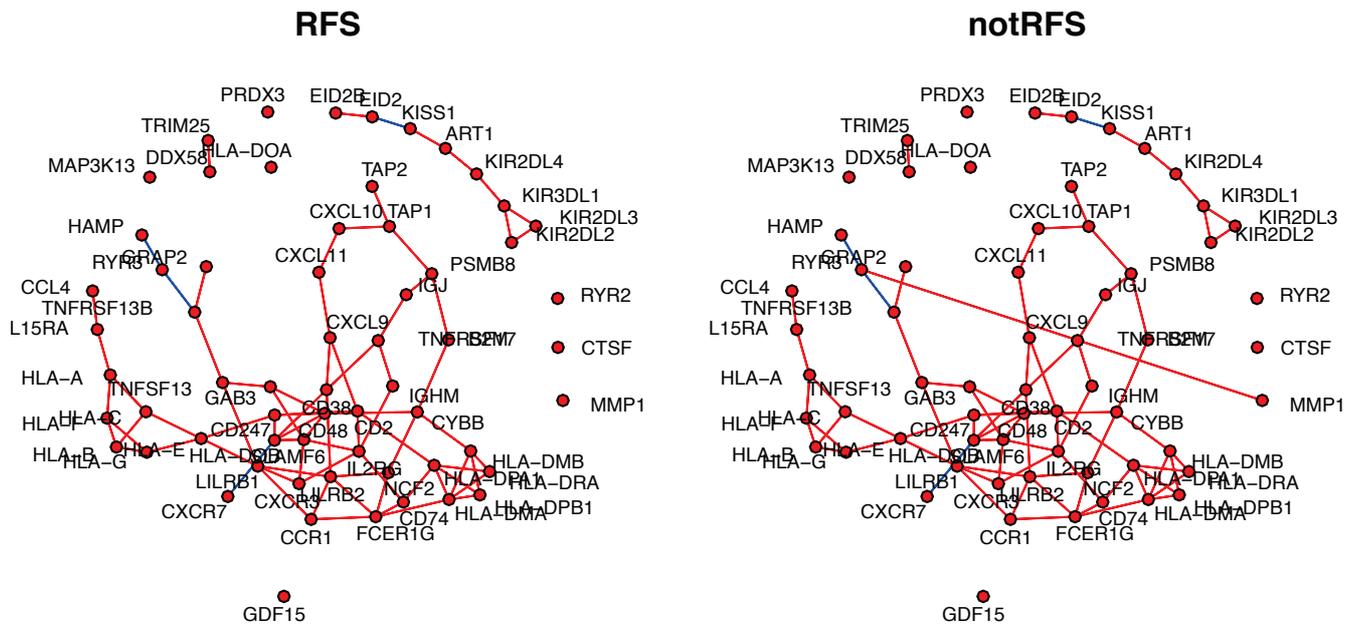
Guedj *et al*, *Oncogene* (2012).



Jeanmougin, *JOBIM* (2012).

- ▶ A total of 82 ER- tumors;
- ▶ 31 women had a relapse (RFS);
- ▶ 51 had no relapse (notRFS).
- ▶ Signature of 62 genes, which were selected by M. Jeanmougin (DIAM's):
  - ▶ Selection of gene modules
  - ▶ With dense interaction in PPI network
  - ▶ With significant differential expression.

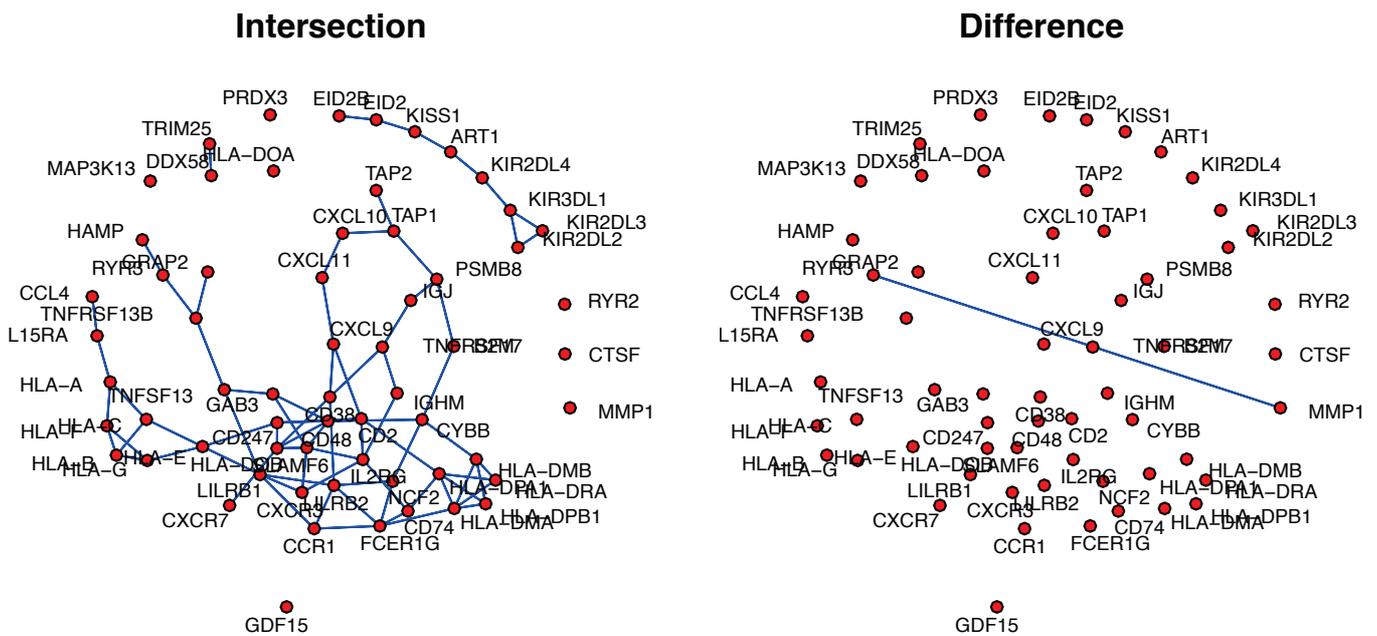
# Application to breast cancer



► Are these differences significant ?

29

# Application to breast cancer

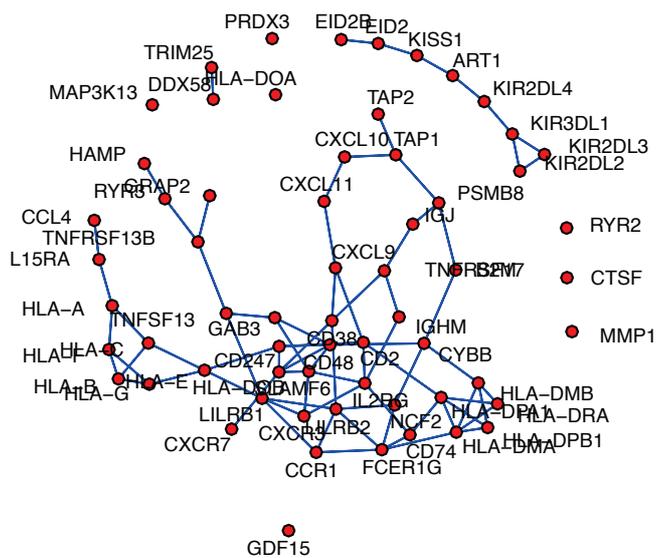


► Are these differences significant ?

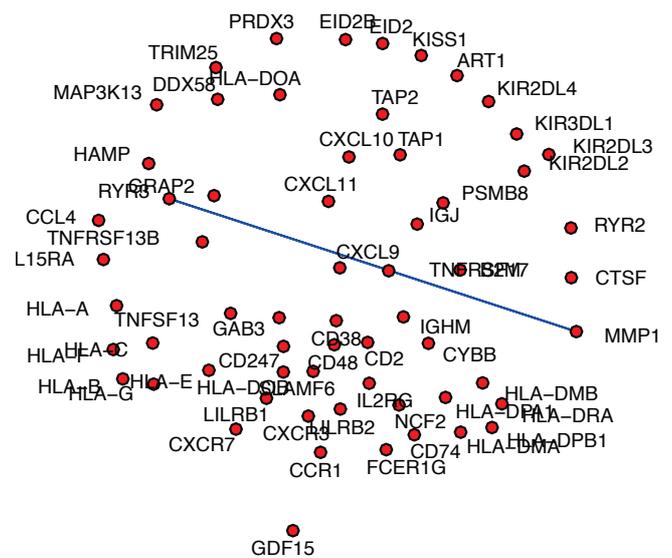
254

29

## Intersection



## Difference



- ▶ Are these differences significant ?

29

## Conclusion

### Versatile Statistical Tools with limitations

- ▶ Gaussian Graphical Model
- ▶ Sparse Regression
- ▶ Model Based Graph Clustering

### Unaddressed questions

- ▶ Structure : are there specific local structures (motifs) ?
- ▶ Exploitation of the graph as prior information
- ▶ ....

### Available Package on CRAN

- ▶ **Simone** to infer graphs from various kind of data
- ▶ **quadrupen** to solve efficiently various penalized regressions problems
- ▶ **osbm** and **mixer** to search for structure in graphs (to be updated)



# Engineering Challenges & Technologies for Healthcare

Rajeev Shorey

(Ph.D, Fellow Indian National Academy of  
Engineering)

IT Research Academy

Department of Electronics & IT (DeitY)

Government of India

(Formerly IBM Research & GM Research)



# Engineering Challenges & Technologies for Healthcare



**Rajeev Shorey**  
(Ph.D, Fellow Indian National Academy of Engineering)

**IT Research Academy**  
**Department of Electronics & IT (DeitY)**  
**Government of India**  
(Formerly IBM Research & GM Research)



**Evry-Genopole, France**  
**15-16 October, 2014**



## Structure of the Talk

- Introduction
- Statistics
- Trends in Healthcare Technologies
- Mobile Health
- Engineering challenges
  - Standardization
- A Peek at the Future of Healthcare
- Concluding Remarks

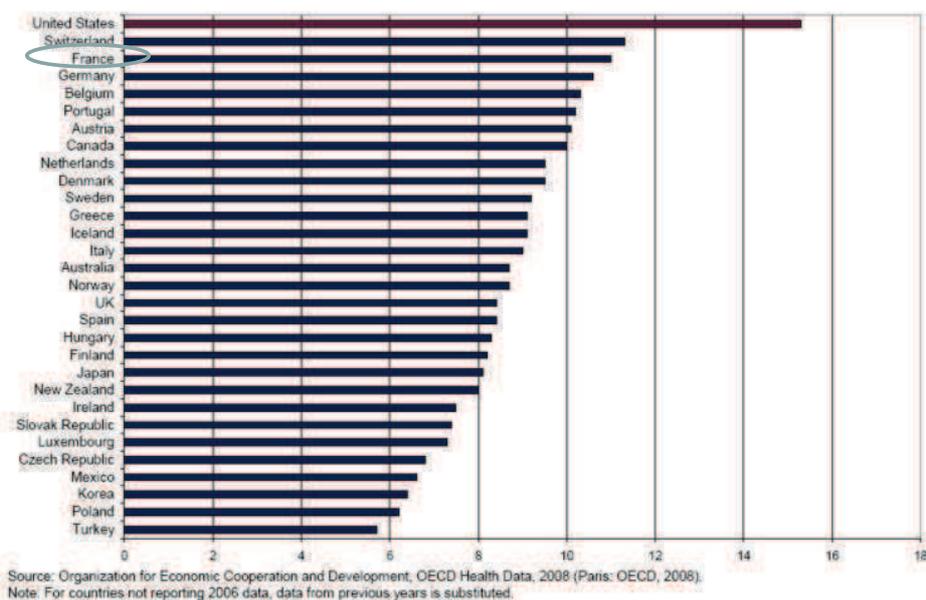
# Facts

- Populations across the world are getting older and there will be fewer young people to look after them !
  - Wireless Health Institute

3

## Healthcare Spending as % GDP

Healthcare Spending as % GDP



4

# India's Healthcare Sector

- India's healthcare sector has been growing rapidly
- Revenues from the healthcare sector account for 5.2% of the GDP
  - Third largest growth segment in India
- Indian healthcare industry is expected to become
  - US \$ 158 billion in 2017
  - US \$ 280 billion industry by 2022 !



Source: [www.ibef.org](http://www.ibef.org)

5

# India's Healthcare Sector

- Key drivers that are responsible for the phenomenal growth of the healthcare sector in India (and other Asian Countries)
  - *Growing Population and Economy*
  - *Rising Middle Class*
  - *Rise of Disease*
  - *Deteriorating Infrastructure*
  - *Lack of Insurance*

6

# Trends in Healthcare Technologies

7

## Next Generation of Real Time Control, Communication and Computation (C<sup>3</sup>) for Wireless Systems

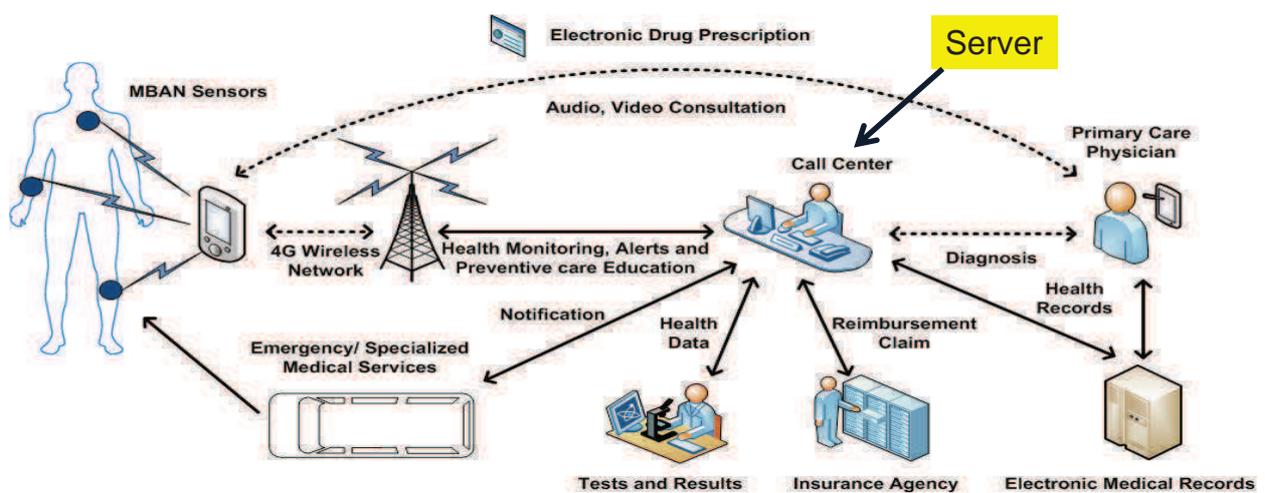




# Pervasive Healthcare

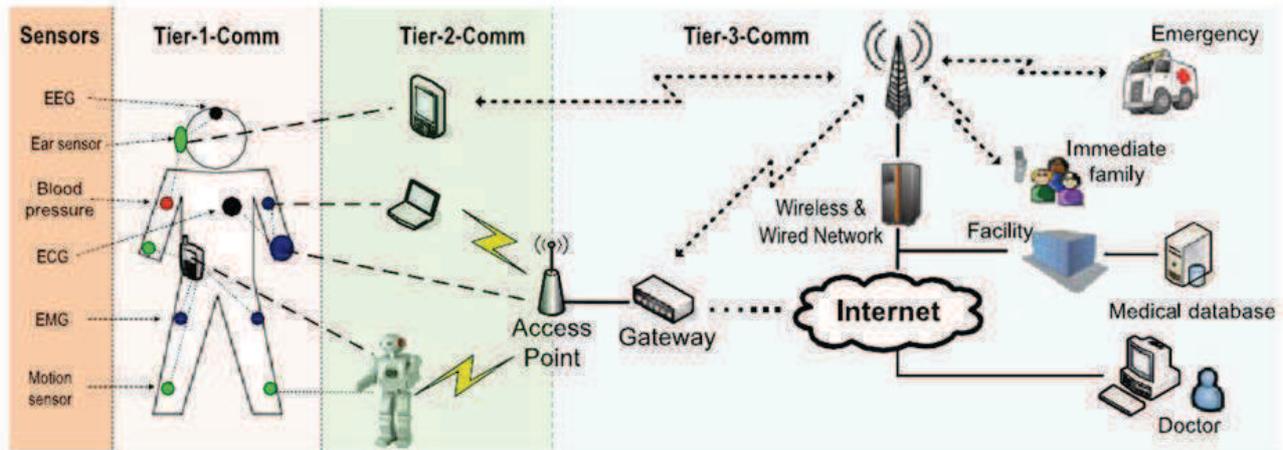


# nG Health: The Future



Mobile broadband technologies have the potential to revolutionize healthcare management

# Wireless Body Sensor Networks: A Hierarchical Network



Complexity Increasing with Increasing Miniaturization

13

## Mobile Health

Refers to the use of Mobile Communication Networks in the Healthcare Sector

14

# Socio-Economic Reasons for mHealthcare

- In the year 2009, data from Center for Disease Control (CDC, 2009) show the observation as given in the table (CDC 2009).
- Researches predict a shortage of 35,000 to 44,000 primary care physicians in the USA by 2025 (Glied et. al, 2008) and approximately 0.8 million shortage of nurses by 2020 (HHS, 2002).
- A mobile healthcare management (MHM) system is expected to almost eliminate visits to a physician for general medical examination as the system can monitor physiological metrics regularly, issue alerts if necessary and recommend appropriate action

Number of visits to a physician in USA	902 million
Number of visits to a physician per 100 persons	306.6
Percent of visits made to primary care physicians	49 per cent
Most frequent reason for visit	General medical examination
Average cost of hospital visit	\$100.00
Approximate total savings to US healthcare using MHM	<b>\$ 44.19 billion</b>

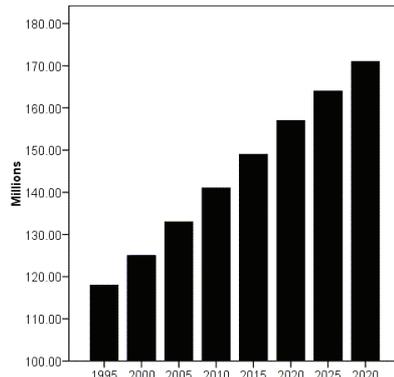


Figure 1. US population (in millions) with one or more chronic diseases.

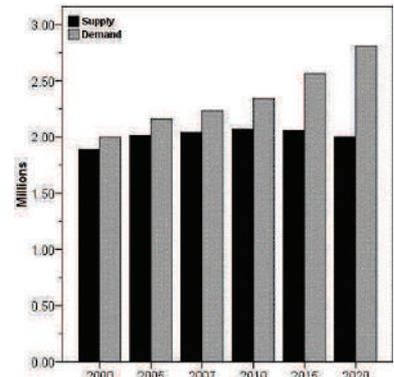
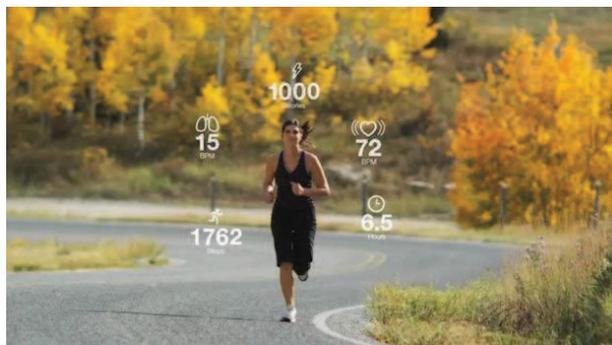
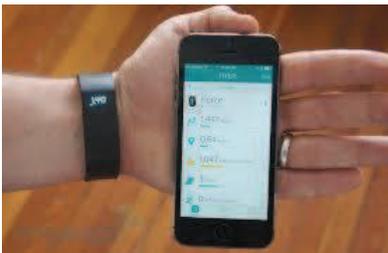


Figure 2. Demand and supply of nurses (in millions).

# Pervasive Devices of the Future

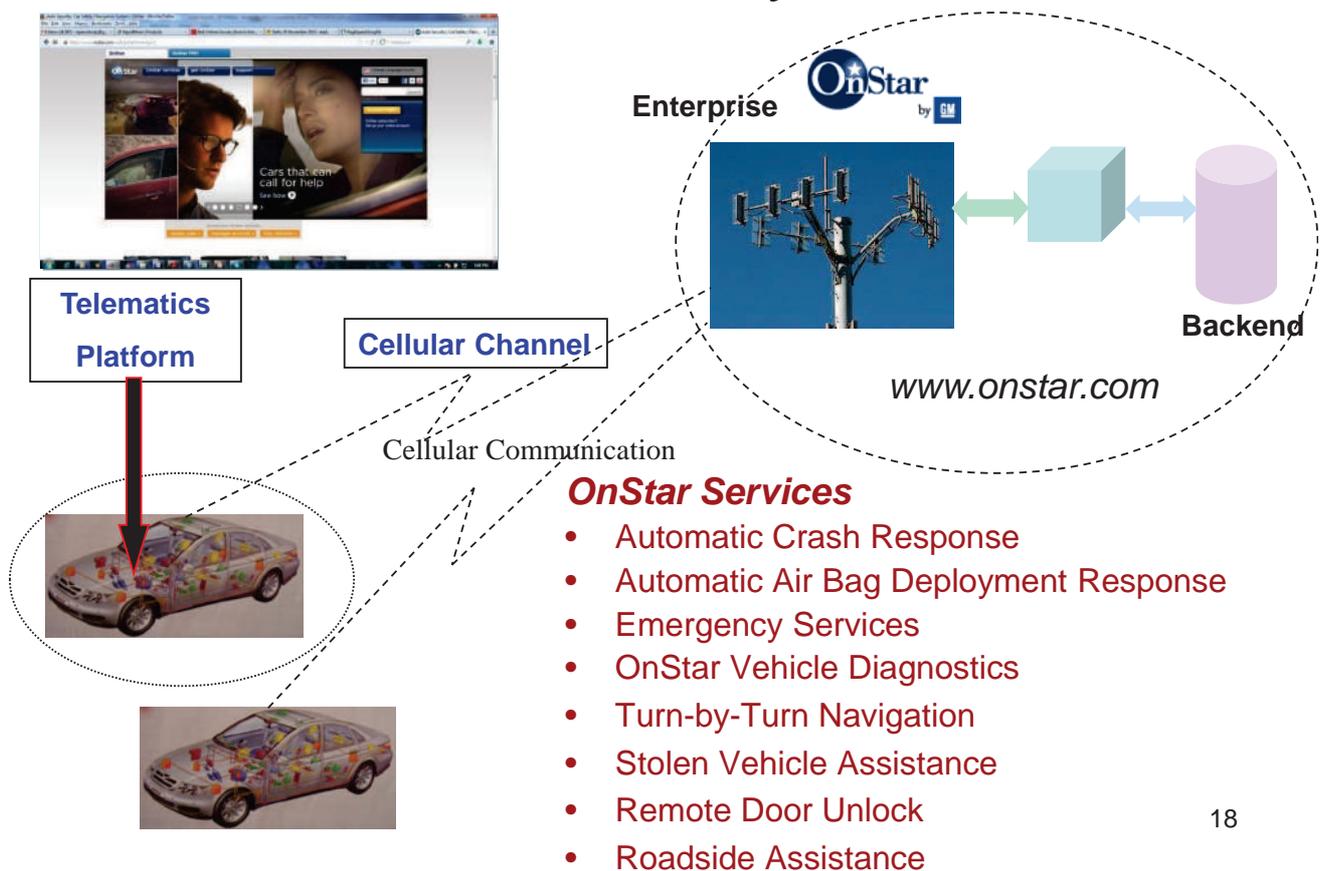
*fitbit.com*



# Parallels with the Automotive Sector

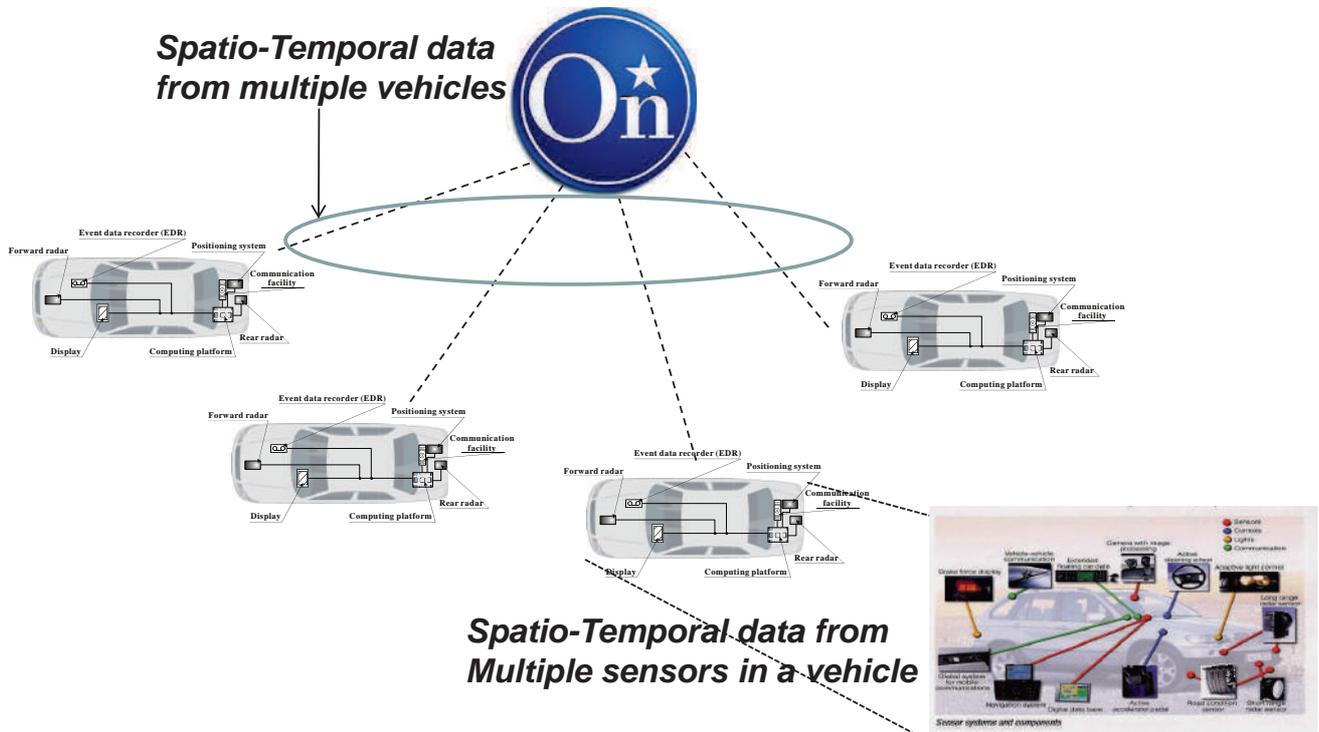
17

## The OnStar System



18

# Onstar Diagnostics



## Engineering Challenges

# Engineering Challenges

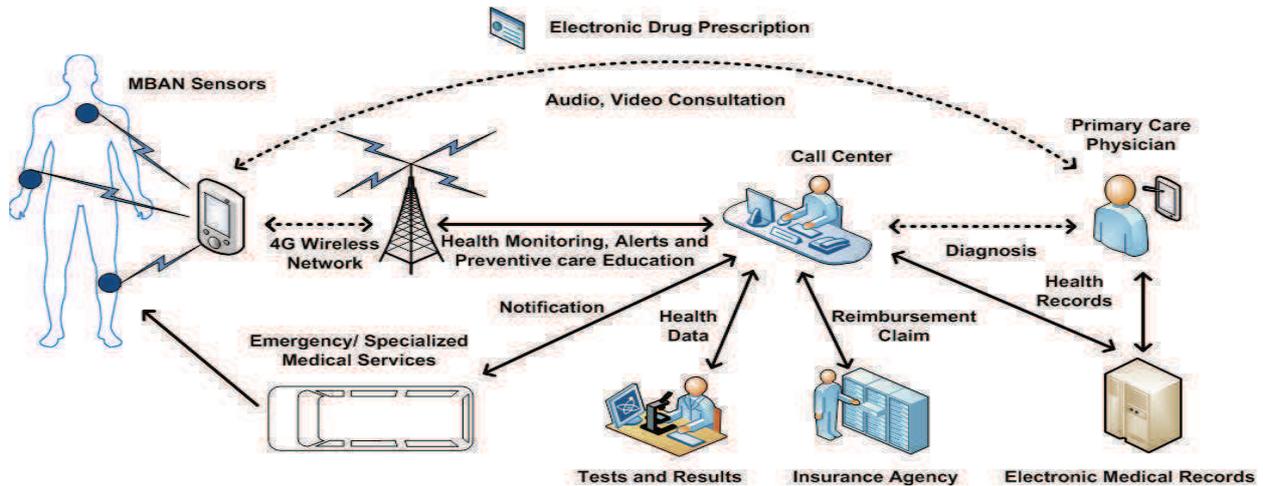
- *Wireless & Pervasive Networks*
  - *Lossy links: packet errors are random*
- *Robust Devices & Networks*
- *Reliability*
- *Distributed Algorithms*
- *Dire need for Standardization*
- *Need for Novel Protocols*
  - *Lightweight Protocols with Low Foot-Print*
  - *E.g., Transport Layer (TCP)*
- *Privacy & Security issues*
- *Big Data Analytics*

21

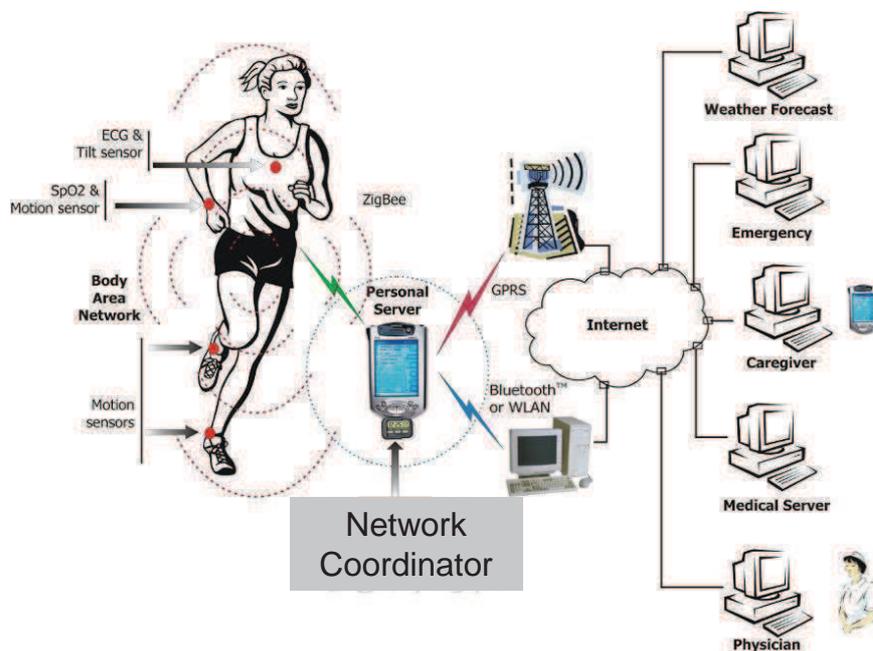
Healthcare Networks are  
Complex Networks

22

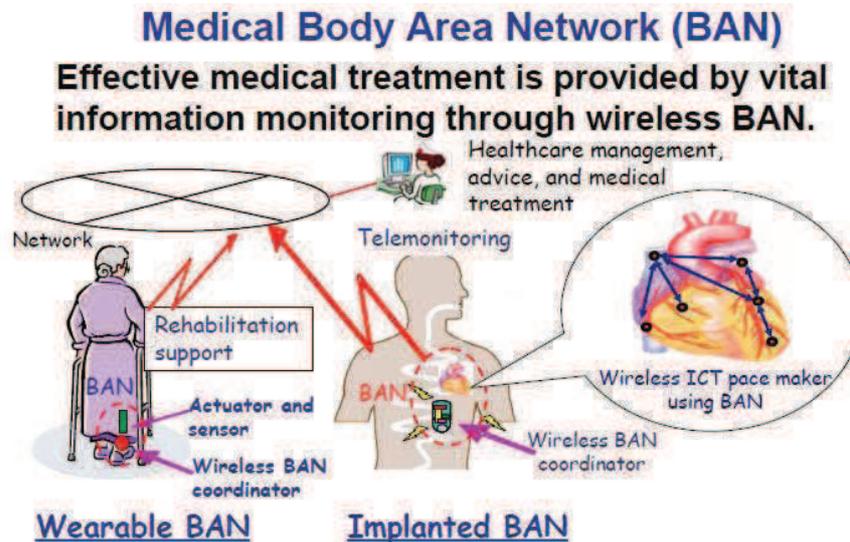
*Mobile & Wireless Everywhere*  
*Heterogeneous Devices/Systems*  
*Distributed & Complex Software*  
*Reliability and Robustness*  
*Multiple Interfaces*  
*Security, Privacy*



## Body Sensor Networks



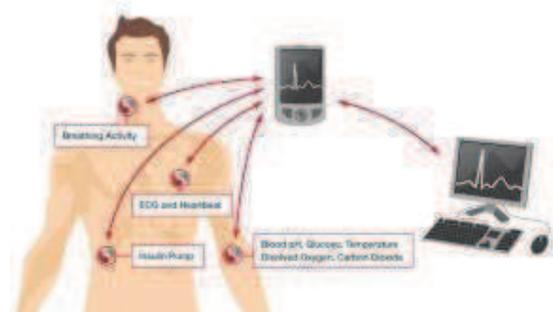
# Medical Body Area Networks



25

## Technical Challenges Facing BSN

- *Energy Supply and Demand*
- *System Security and Reliability*
- *Context Awareness*
- *Improved Sensor Design*
- *Biocompatibility*
- *Integrated Therapeutic Systems*
- ...

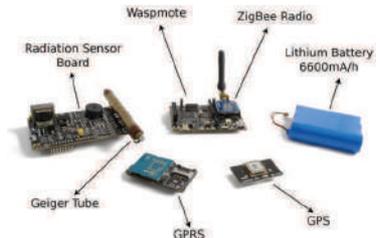


# Energy Supply & Demand

- Key considerations for BSNs:
  - Power consumption
  - Need to reduce battery consumption

- Solution

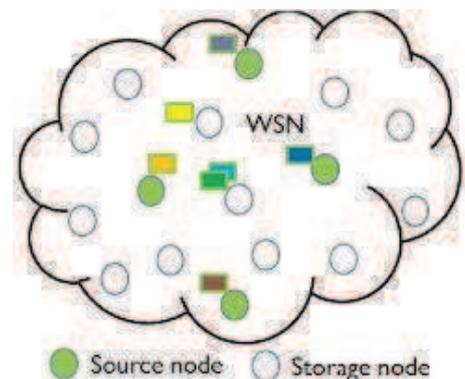
- Development of micro-fuel cells
- Use of biocatalytic fuel cells
- Reducing battery consumption through increased use of power scavenging from on-body sources
  - Vibration
  - Temperature



27

# System Security & Reliability

- Infrastructure required for the practical deployment of BSN applications
  - Robust
  - Efficient
  - Lightweight security



272

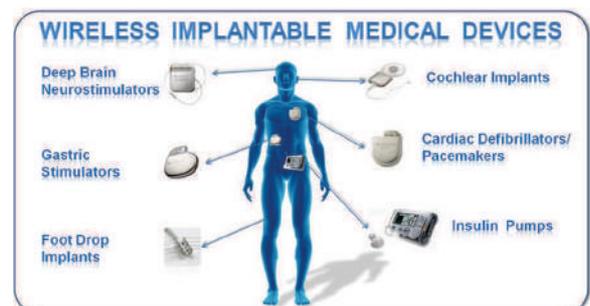
# Context Awareness

- Importance of context (environment)
- Example
  - Sleeping
  - Walking
  - Exercising, ...
- Techniques used for activity recognition and tracking daily activities
  - Naïve Bayesian classifiers
  - Hierarchical hidden semi-Markov models

29

## Improved Sensor Design

- Improvements in sensor manufacturing and nano-engineering techniques
- Advances in MEMS technology

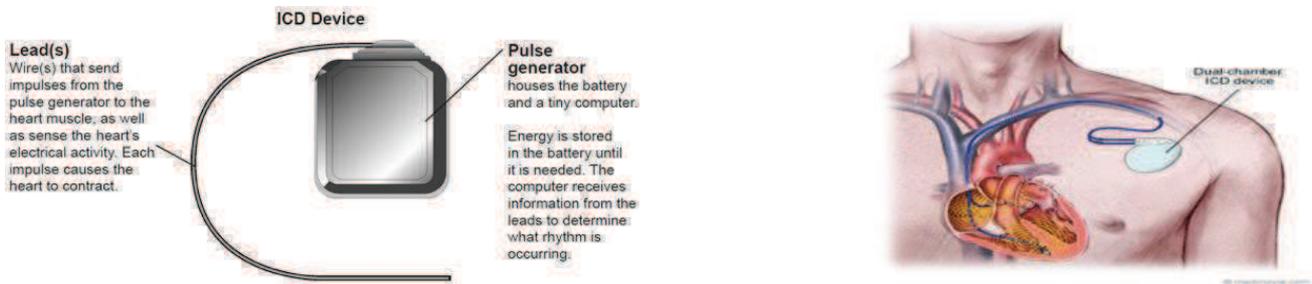


Smaller Implantable and Attachable Sensors

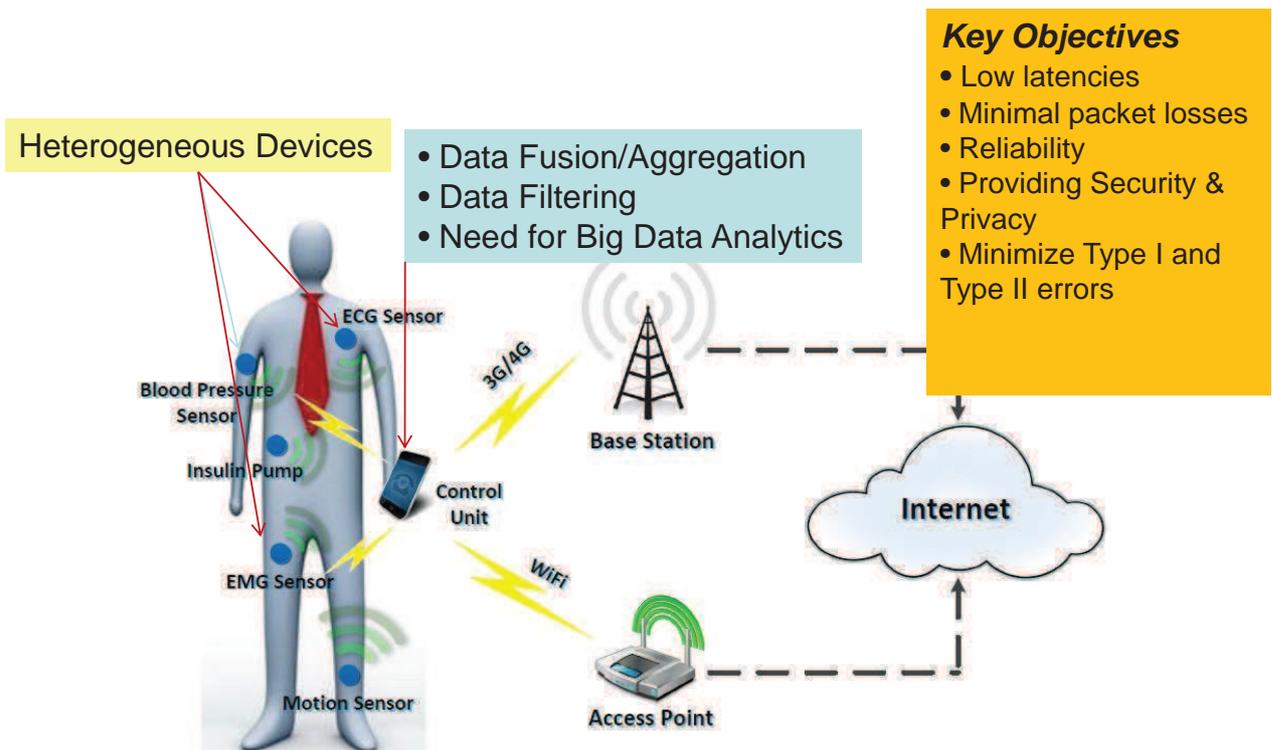
30

# Biocompatibility

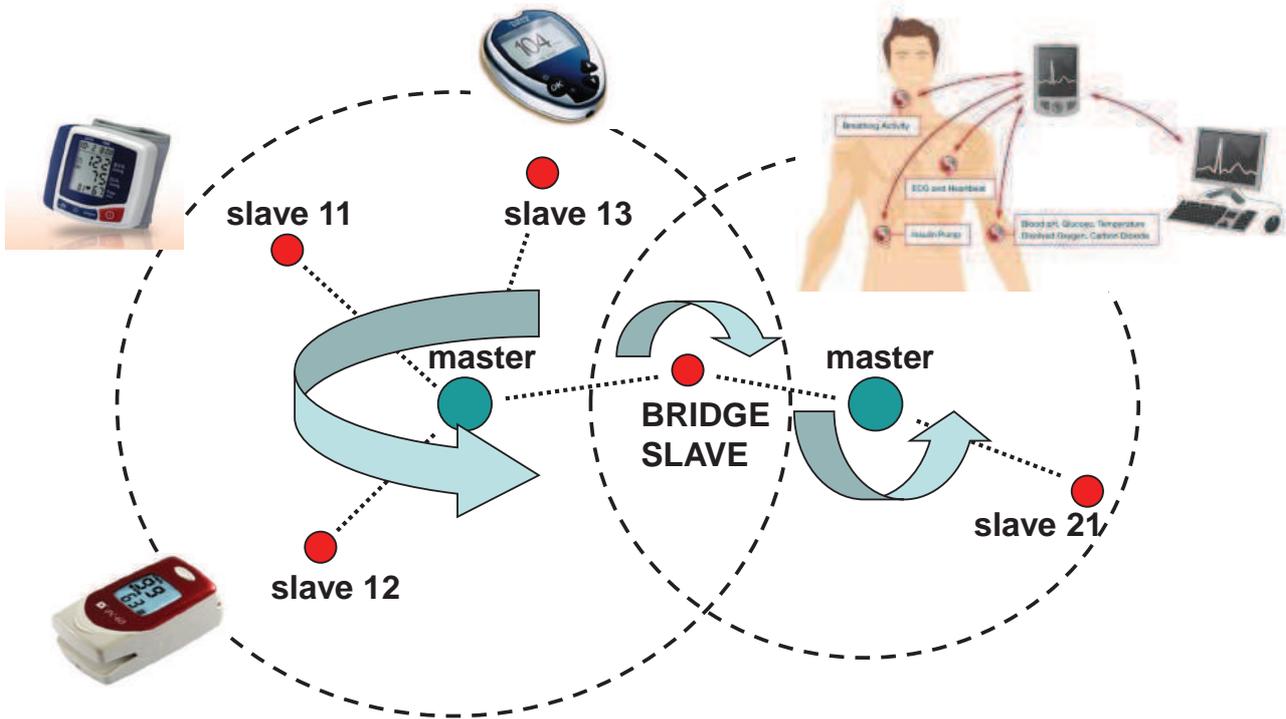
- Implantable sensors and stimulators have had to overcome the problems of long-term stability and biocompatibility
- Examples
  - Cardiac Pacemaker
  - Implantable Cardioverter-Defibrillator (ICD)



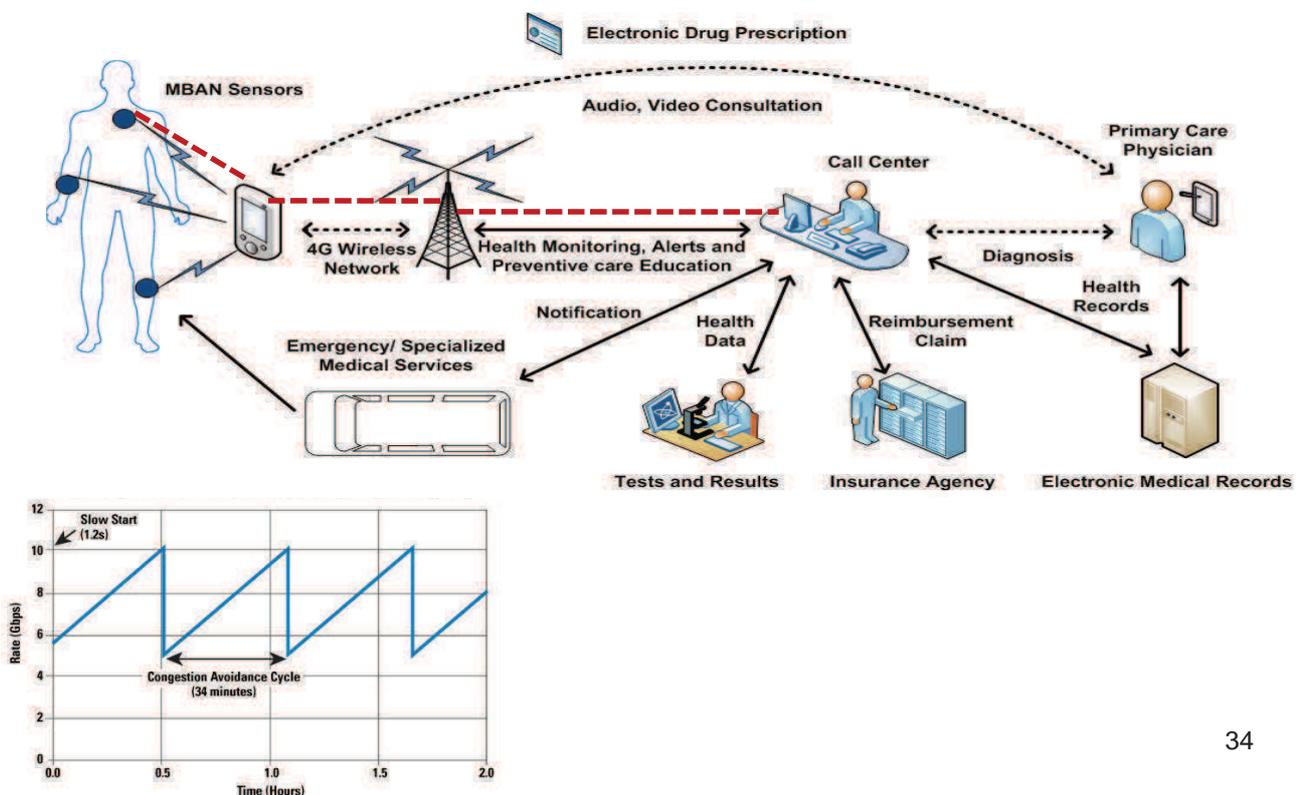
# Robust End-to-End Performance



# Scheduling in Personal Area Networks



# Performance of Transport Layer Protocols (e.g., TCP)

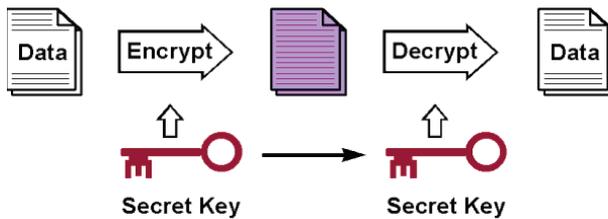


# Mobile & Wireless Everywhere

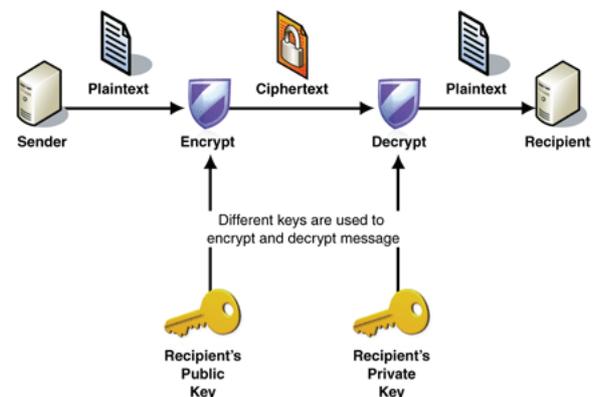
## Heterogeneous Systems

### Distributed & Complex Software

### Multiple Interfaces



**SECURITY IS A HUGE CHALLENGE !**



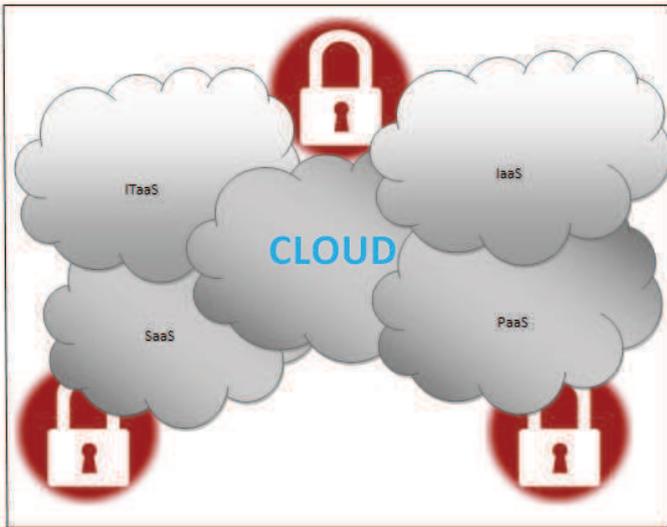
## Security & Privacy in Healthcare Networks

- Security and Performance trade-offs
  - Strong Security may lead to poor performance
  - Example: High latencies
    - Unacceptable in healthcare !
- Data Privacy is critical

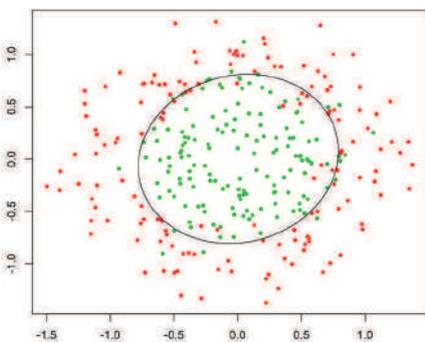
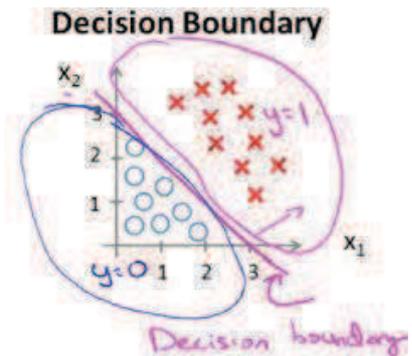
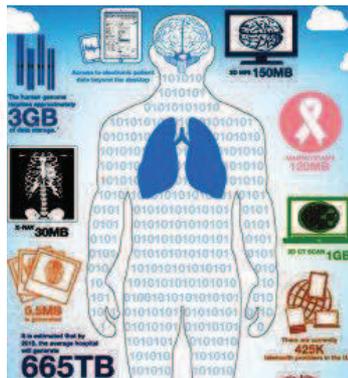


# Cloud Security

- Cloud-based tele-health and medical security and privacy protection

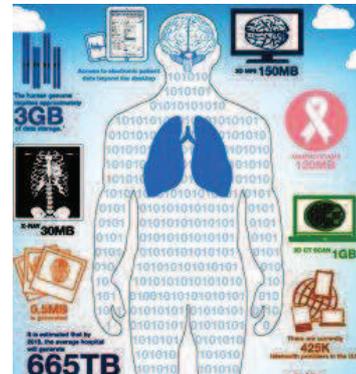


# Big Data Analytics in Healthcare



# Key Trends

- **The Patient Data Warehouse**
- **Predictive Medicine**
- **Wellness Maintenance**
- **Just-in-Time-Medicine**



<http://www.forbes.com/sites/netapp/2013/04/17/healthcare-big-data/>

39

## The Patient Data Warehouse

- By 2015, the average hospital will have two-thirds of a *petabyte* (665 terabytes) of patient data
- 80% of which will be unstructured data
  - CT scans
  - X-rays
- *It's eye opening that the human body needs so much storage !*

40

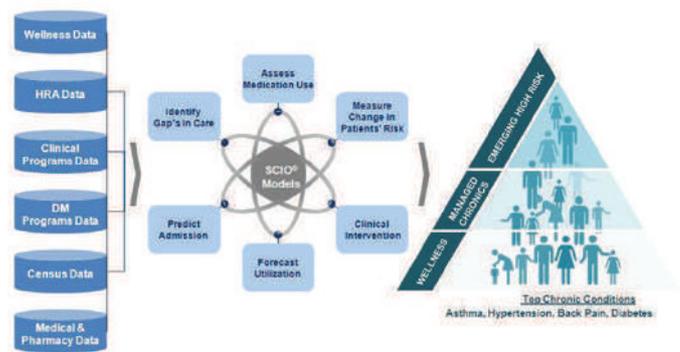


# Just-in-Time Medicine

- Treating patients at the wrong time and in the wrong place is costly
- Scheduled care is much cheaper than unscheduled care
- Optimizing patient discharge timing could save up to \$70 billion [McKinsey]



*Insights from  
Big Data Analytics could be  
the prescription for  
better care, lower costs and  
higher productivity !*



# Industry Standards for Healthcare

- **HL7**
  - Health Layer 7
- **HIPAA**
  - Health Insurance Portability & Accountability Act
- **IHE**
  - Integrating the Healthcare Enterprise
- **SNOMED**
  - The Systematized Nomenclature of Medicine
- **PACS**
  - Picture Archiving and Communication Systems
    - Allow scans and X-rays to be shared seamlessly across departments

45

## Research Projects in the Pipeline

- *Transmission of Stethoscopic Sounds*
- *Portable Cost-Effective Ultrasound Machine and Image Reconstruction/Transmission*
- *Mobile Software Development and Internetworking*
- *Mobile Health Record System*
- *Geographic and Health Information Fusion*
- *Simulation and Modeling*

46

# The Future of Healthcare ...



47

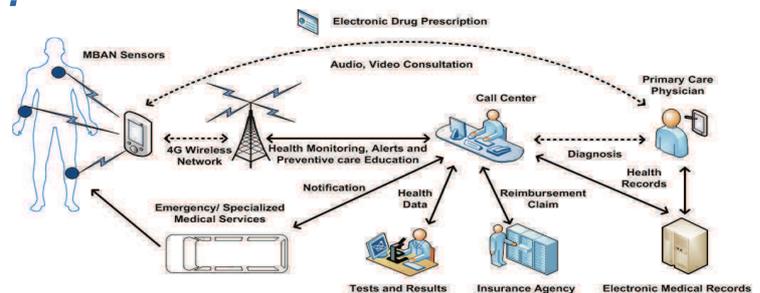
## Technologies Likely to Change the Future of Healthcare

- Digital Diagnostics ([www.neurotrack.com](http://www.neurotrack.com))
- The Cloud
- Ultra-Fast Scans
- Wearables
- Health Informatics
- Digital Therapy
- Concierge Medical Services
- Networks and Coaching
- Self Insurance

48

# Concluding Remarks

- *End-to-end Pervasive Healthcare Scenario*
- *Closed Loop Feedback*
- *Health For All, Anywhere, Anytime, ...*
- *Convergence of Healthcare and Telecom*
- *Major role of Big Data Analytics*
- *Emerging Killer Services & Killer-Apps in the Healthcare World !*



## Example of a 'Killer' Service

A Device that keeps Track of our Driving !  
*Usage-based Insurance Program*



<http://www.progressive.com/auto/snapshot-how-it-works/>

# Thank You

rajeevshorey@gmail.com

51

52

# Laboratory Visits and Discussions



DE LA RECHERCHE À L'INDUSTRIE



DSV/IG/GENOSCOPE/LABGEM



# LABORATOIRE D'ANALYSE BIOINFORMATIQUES EN GENOMIQUE et METABOLISME (LABGeM)

Claudine Médigue



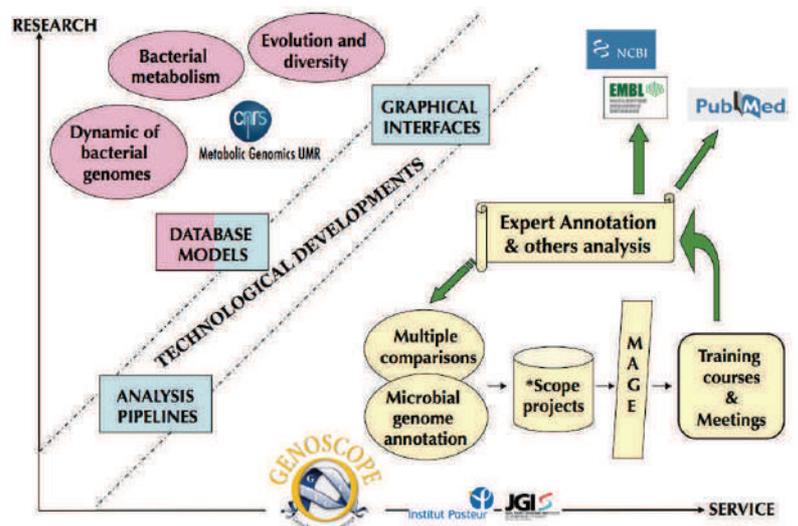
www.cea.fr



DE LA RECHERCHE À L'INDUSTRIE

## cea LABGEM ACTIVITIES

- Tools for bacterial genomes annotation
- Next Generation Sequencing data analysis
  - Quantitative transcriptomics via RNA-seq
  - Resequencing of evolved strains, SNPs/indels
  - (Meta)genomics analyses, taxonomic assignments
- Bacterial Metabolism
  - Curation of metabolic data
  - Metabolic network reconstruction
  - Discovery of new enzymatic activities



- Microbial genome annotation service: **MicroScope platform**

[www.genoscope.cns.fr/agc/microscope](http://www.genoscope.cns.fr/agc/microscope)



=> Predict small and atypical genes

## AMIGene: Annotation of Microbial Genes

Stéphanie Bocs, Stéphane Cruveiller, David Vallenet, Grégory Nuel<sup>1</sup> and Claudine Médigue\*

Génomoscope/UMR-CNRS 8030 and <sup>1</sup>Laboratoire Statistique F-91034 Evry, France

*Nucleic Acids Research*, 2003, Vol. 31, No. 13 3723-3726  
DOI: 10.1093/nar/gkg290

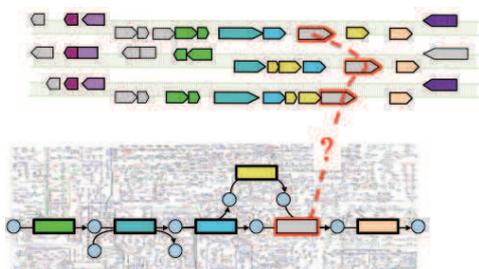
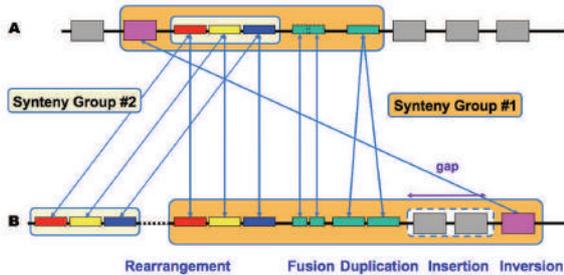


## MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes

Stéphane Cruveiller, Jérôme Le Saux, David Vallenet, Aurélie Lajus, Stéphanie Bocs and Claudine Médigue\*

Génomoscope/UMR-CNRS 8030, Atelier de Génomique Comparative, 2 rue Gaston Crémieux, F-91006 Evry, France

*Nucleic Acids Research*, 2005, Vol. 33, Web Server issue W471-W479  
doi:10.1093/nar/gki498



## BIOINFORMATICS ORIGINAL PAPER

Genome analysis

### Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data

Frédéric Boyer<sup>1</sup>, Anne Morgat<sup>1</sup>, Laurent Labarre<sup>3</sup>, Joël Pothier<sup>4</sup> and Alain Vian<sup>1\*</sup>

<sup>1</sup>INRIA Rhône-Alpes, HELIX Group, 655 avenue de l'Europe, 38334 Montbonnot Cedex, France;

Vol. 21 no. 23 2005, pages 4209-4215  
doi:10.1093/bioinformatics/bti711

OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

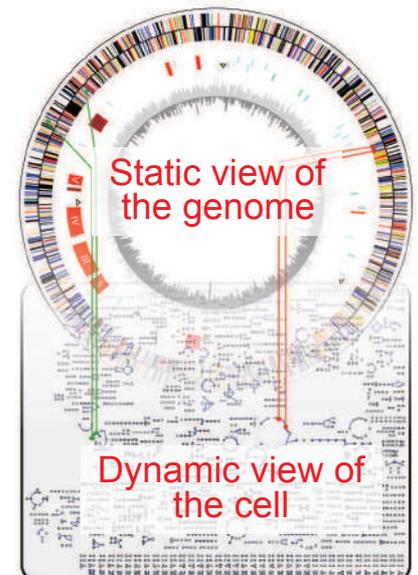
## The CanOE Strategy: Integrating Genomic and Metabolic Contexts across Multiple Prokaryote Genomes to Find Candidate Genes for Orphan Enzymes

Adam Alexander Thil Smith<sup>1,2,3\*</sup>, Eugeni Belda<sup>1,2,3</sup>, Alain Vian<sup>1</sup>, Claudine Médigue<sup>1,2,3</sup>, David Vallenet<sup>1,2,3\*</sup>

=> candidate genes for global/local orphan enzymatic activities may be found in the “gaps”

## An integrated environment for:

- Microbial genome (re)annotation
- Comparative analyses
- Function and biological process predictions
- Expression and evolutionary studies by NGS data integration



First publication in 2006

Release 1 in 2009

*Nucleic Acids Research*, 2006, Vol. 34, No. 1 53-65  
doi:10.1093/nar/gkj456

**MaGe: a microbial genome annotation system supported by synteny results**

David Vallenet<sup>1</sup>, Laurent Labarre, Zoé Rouy, Valérie Barbe<sup>1</sup>, Stéphanie Bocs, Stéphane Cruveiller, Aurélie Lajus, Géraldine Pascal, Claude Scarpelli<sup>1</sup> and Claudine Médigue

Atelier de Génomique Comparative, CNRS-UMR8030 and <sup>1</sup>Génomoscope, 2 rue Gaston Crémieux, 91097 Evry Cedex, France

=> Synteny maps & curation of gene function

**DATABASE**

October, Vol. 2008, Article ID: bsp021, doi:10.1093/dzbaa/bfp021

**Original article**

**MicroScope: a platform for microbial genome annotation and comparative genomics**

D. Vallenet<sup>1</sup>, S. Engelen, D. Mormico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli and C. Médigue

=> Comparative genomics tools & gene carts

*Database*, 2008, Vol. 2008, Article ID: bsp021, doi:10.1093/dzbaa/bfp021

**MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data**

David Vallenet<sup>1,2,3\*</sup>, Eugeni Belda<sup>1,2,3</sup>, Alexandra Calteau<sup>1,2,3</sup>, Stéphane Cruveiller<sup>1,2,3</sup>, Stefan Engelen<sup>1</sup>, Aurélie Lajus<sup>1,2,3</sup>, François Le Fèvre<sup>1,2,3</sup>, Cyrille Longin<sup>1,2,3</sup>, Damien Mormico<sup>1,2,3</sup>, David Roche<sup>1,2,3</sup>, Zoé Rouy<sup>1,2,3</sup>, Grégory Salvignol<sup>1,2,3</sup>, Claude Scarpelli<sup>1</sup>, Adam Alexander Thil Smith<sup>1,2,3</sup>, Marion Weiman<sup>1,2,3</sup> and Claudine Médigue<sup>1,2,3\*</sup>

Release 2 in 2013

=> Curation and analysis of metabolic data

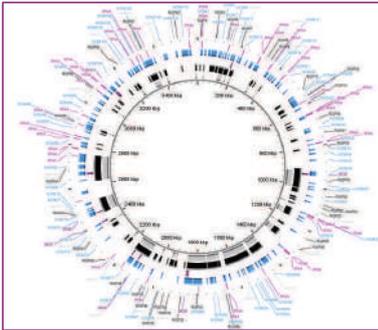
Welcome guest (Lost password?)   LOGIN OR SIGN UP

Acinetobacter baylii ADP1  chromosome ACIAD

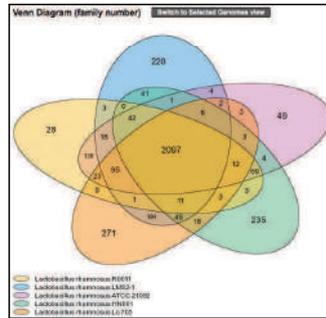
MaGe Genomic Tools **Comparative Genomics** Metabolism Searches Export Experimental Data User Panel About

Gene PhyloProfile  
Regions of Genomic Plasticity  
LinePlot  
Fusion / Fission  
PKGBD Synteny Statistics  
RefSeq Synteny Statistics  
Pan/Core-Genome

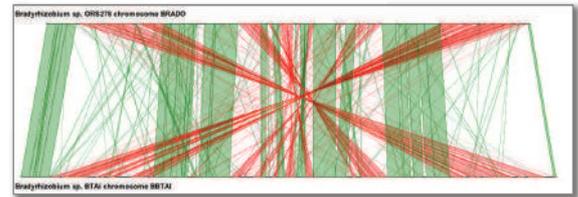
### RGP finder



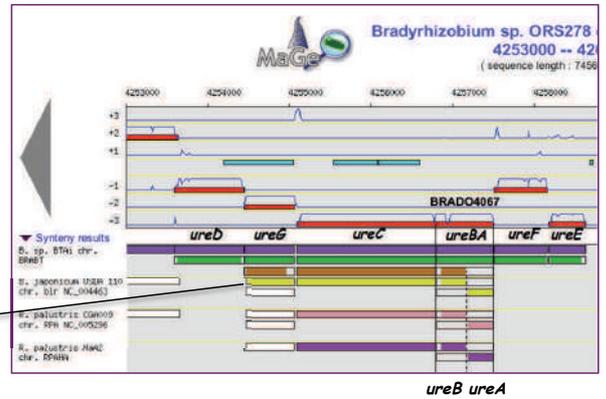
### Pan/core Genome



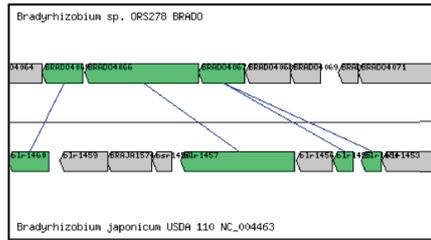
### LinePlot



### Fusion/Fission



### Synton visualization



Welcome guest (Lost password?)   LOGIN OR Register new account

MaGe Genomic Tools Comparative Genomics **Metabolism** Searches Export Experimental Data User Panel About

Kegg  
MicroCyc  
Metabolic Profiles  
Pathway Synteny  
Pathway Curation  
CanOE

Reference pathways: KEGG, EcoCyc, MetaCyc

Cupriavidus taiwanensis LMG19424 Pathway: cysteine biosynthesis I

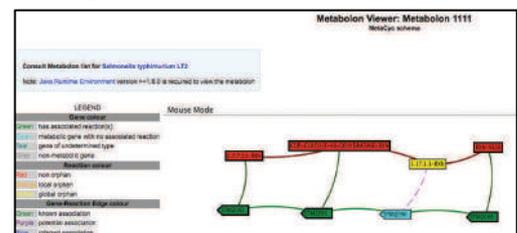
BioCYC Database Collection

Extrane Glutamic cycle (TCA cycle) pathway Acinetobacter baylii ADP1

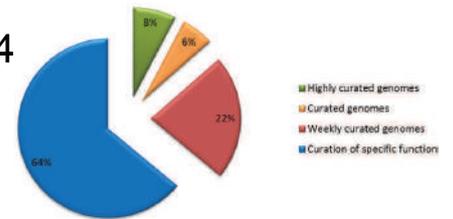
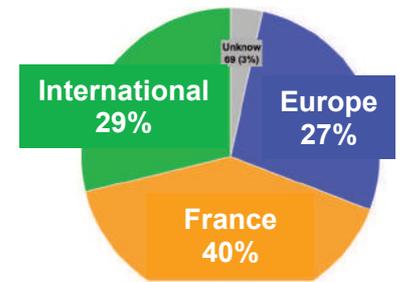
### Metabolic phyloprofile

Aromatic Compounds Degradation <sup>(1)</sup>	Reactions nb	Acinetobacter baylii ADP1	Pseudomonas entomophila L48
3-phenylpropionate degradation	10	0	0.40
4-hydroxyphenylacetate degradation	9	0.33	0.67
siloxime degradation	3	1	0.33
anthranilate degradation I (aerobic)	1	1	0
atrazine degradation I (aerobic)	3	0.33	0.33
benzoate degradation I (aerobic)	2	1	1
benzoyl-CoA degradation I (aerobic)	8	0.38	0.25
benzoyl-CoA degradation II (anaerobic)	6	0.17	0
benzoyl-CoA degradation III (anaerobic)	9	0.11	0
catechol degradation to β-ketoadipate	4	1	1
chlorogenic acid degradation	1	1	0
cyanurate degradation	3	0.33	0.33
ethylbenzene degradation (anaerobic)	5	0	0.20
ferulate degradation	3	1	0
galate degradation II	4	0	0.25

### CanOE



- >3,100 genomes; analysis of 4 genomes a day in 2014
- More than 300 citations
- 2,000 user accounts - 400 active accounts per month
- 2,400 human-expertized annotations a month in 2014  
=> 445 genomes have at least one gene being manually annotated
- Since 2008, >320 persons have been trained in France and abroad



- MicroScope is a key partner for several academic labs and companies

➤ Certifications:



*since january 2012 for our research, development and service activities*

# The France Génomique infrastructure

Technology & Healthcare Symposium  
Evry – October 15<sup>th</sup>, 2014



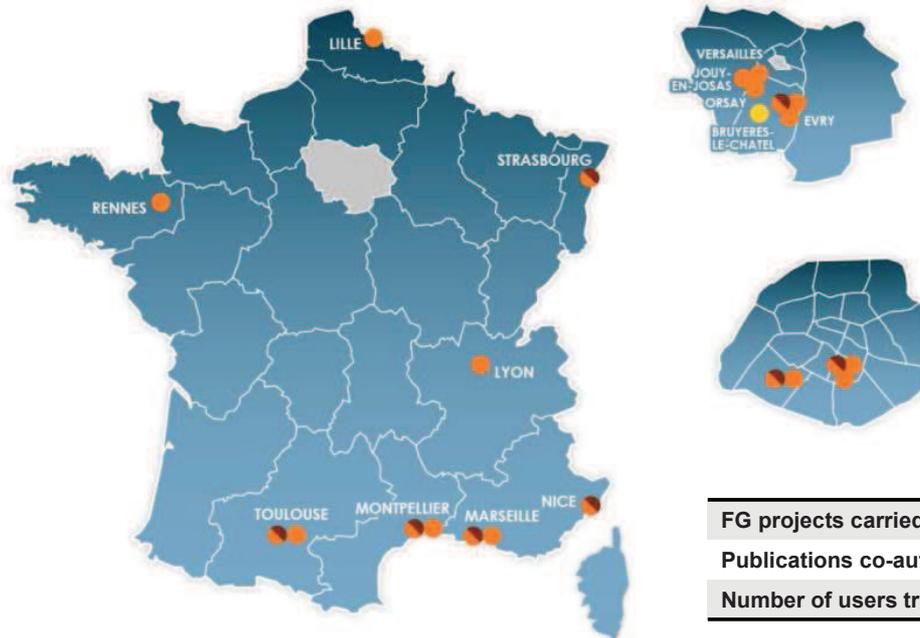
## What is France Génomique ?

***A national genomics and bioinformatics infrastructure, spread out on the French territory***

- **All the platforms already existed under the “IBiSA” label**
- **The IA funding (60 M€) “only” covers 10% to 15% of the platforms’ budget over the 8 years period. It does however leverage our ability to:**
  - **structure and reinforce our technological offer, improve its visibility**
  - **maintain / reinforce the global expertise and ensure its rapid diffusion**
  - **invest for the medium term in new technologies and developments**

**The leverage effect is very high, in terms of competitiveness**

## Localisation of the FG platforms



FG projects carried out in 2013	726
Publications co-authored by FG	185
Number of users trained in 2013	1284

<http://www.france-genomique.org>

## A wide range of genomic resources and expertise

***Genomics has become an indispensable component in all fields of life sciences research: the increasing diversity of technologies and applications involved cannot be mastered by any individual platform***

- Each platform has its own skills and specific field of genomic competence and applications
- The platforms' activities are coordinated and supervised by the FG governing bodies
- The projects are submitted through the FG Web portal
- The users have access to state-of-the-art technology & expertise and their project is carried out on the most suitable platform

<http://www.france-genomique.org>

## A critical mass in terms of capacity and expertise

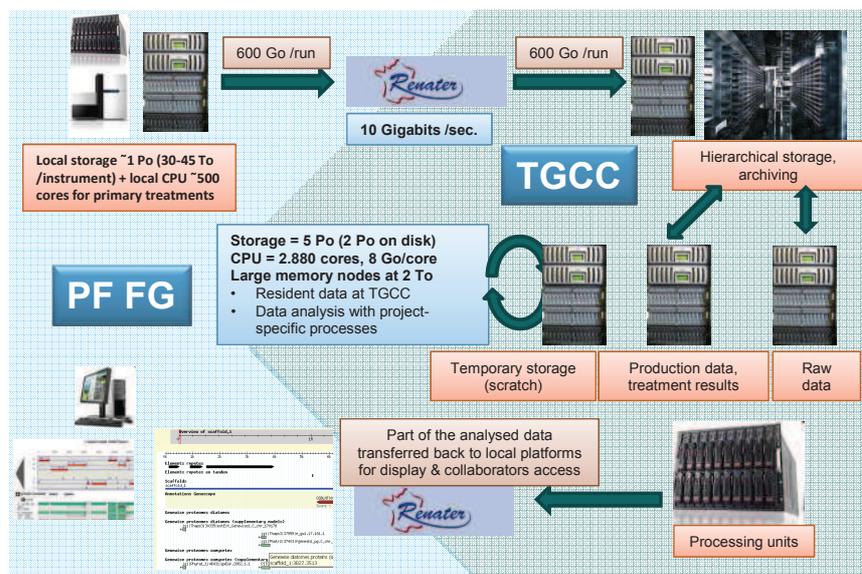
*What is at stake is the French presence and sovereignty in a field (genomics) of increasing strategic importance for the future of research, medicine and life sciences industry in general*

- Coordinated / synergised technology survey & developments (bioinformatics, "wet lab"...), widely diffused to the community
- Capability to undertake very large projects with a high scientific impact: annual CFP funded by FG
- Improved national and international visibility, early access to new technologies (e.g. 3<sup>rd</sup> generation NGS)
- Dedicated data storage and computing capacities deployed at the TGCC (CEA/DAM) in Bruyères-le-Châtel

<http://www.france-genomique.org>

## Central IT resources: e-infrastructure

**5 M€ investment at the TGCC/CCRT (CEA/DAM) in Bruyères-le-Châtel**



- The change of scale had become unavoidable
- The running costs of the e-infrastructure should be taken into account in the projects' budget
- Proper balance between local and central resources to be kept

<http://www.france-genomique.org>

## "Large projects" call for proposals (CFP)

- **Focused on “very large” projects, otherwise difficult to fund, and giving to FG a high visibility**
- **First and second CFP in 2012 and 2013: 96 projects were submitted, 26 selected by the scientific review board**
- **Average funding per project (direct costs): 200 to 300 k€**
- **Next CFP to be opened early (March) 2015**

TITRE DU PROJET	DOMAINE	PORTEURS	AFFILIATION DES PORTEURS
Contribution to decipher cancer heredity	Génétique Humaine	Brigitte BRESSAC-de PAILLERETS Daniel GAUTHIERET	Institut Gustave Roussy - Villejuif
The Pea genome sequencing project	Biodiversité	Judith BURSTIN	INRA UMR AgroEcologie - Dijon
Metagenomics of ancient eric soils	Biodiversité	Jean-Michel CLAVERIE	UMR7256 CNRS-AMU - Marseille
Genomics in hematological malignancies and stem cells	Génétique Humaine	Thérèse COMMES	CHU Montpellier - Université Montpellier 2
Epigenetic inheritance and speciation in Paramecium	Biodiversité	Sandra DUHARCOURT	UMR7592 CNRS / Université Paris Diderot
Meta-OMICS from the world plankton - TARA OCEANS	Biodiversité	Eric KARSENTI Patrick WINCKER	EMBL - Heidelberg CEA / Genoscope - Evry
Mycocaptura : novel genes for mycoses	Génétique Humaine	Jocelyn LAPORTE	IGBMC U964, UMR7104, Université de Strasbourg
The GENESIS exome sequencing project	Génétique Humaine	Fabienne LESUEUR	INSERM U900 - Institut Curie
Microbial biogeography of French soils	Biodiversité	Lionel RANJARD	UMR1347 INRA / AgroSup / Université de Bourgogne - Dijon
IRIGIN - International Riso Genome Initiative	Biodiversité	François SABOT	UMR DIADE, IRD France SUD - Montpellier
1002 yeast genome	Biodiversité	Joseph SCHACHERER Gianni LITI	UMR7158 CNRS / Université Strasbourg UMR7284 CNRS / INSERM / Université de Nice Sophia Antipolis
Exome sequencing of schizophrenia patients	Génétique Humaine	Michel SIMONNEAU Jean-François DELEUZE	INSERM U894 - Paris CEA / CNG - Evry

<http://www.france-genomique.org>

## Developments & technology survey

- **Mission: to maintain and reinforce the French cutting-edge expertise in genomics, bioinformatics and their applications, and to increase the proficiency of the life sciences community in general**
- **45 engineers recruited (fixed-term contracts) for developments, 2/3 of them in bioinformatics**

**WP1.1 Implementation of our tools in the HPC environment at TGCC**  
**WP1.2 Adaptation of the sequencing platforms bioinformatics environment**  
**WP2.1 Quality control and technology evaluation**  
**WP2.2 Evaluation of sequence mapping software**  
**WP2.3 Evaluation of sequence assembly software**  
**WP2.4 Variant detection in genomic data**  
**WP2.5 Analysis of gene expression level (RNAseq)**  
**WP2.6 Regulation of gene expression (incl. epigenomics)**  
**WP2.7 Genomics and metagenomics data analysis**

- **The knowledge, know-how and tools developed by FG should be diffused towards the life sciences community**

<http://www.france-genomique.org>

**Thank you for your attention**

**pleber@genoscope.cns.fr**  
**www.france-genomique.org**



## BIOINFORMATICS AT CNG

François Artiguenave, PhD, eMBA  
DSV/IG/CNG

## BIOINFORMATICS STEPS FOR GENOME ANALYSIS

### (1) Genome analysis:

DNA and RNA sequences from different experimental sources and using various technologies and platforms

### (2) Consequences of mutations and genomic alterations:

Mutations in coding and non-coding regions, changes in gene expression, genome structure alterations

### (3) Network level analysis:

metabolic and signaling pathways, gene control networks and functional classes

### (4) Drug

related to proteins and pathways

### (5) Collaborative interfaces

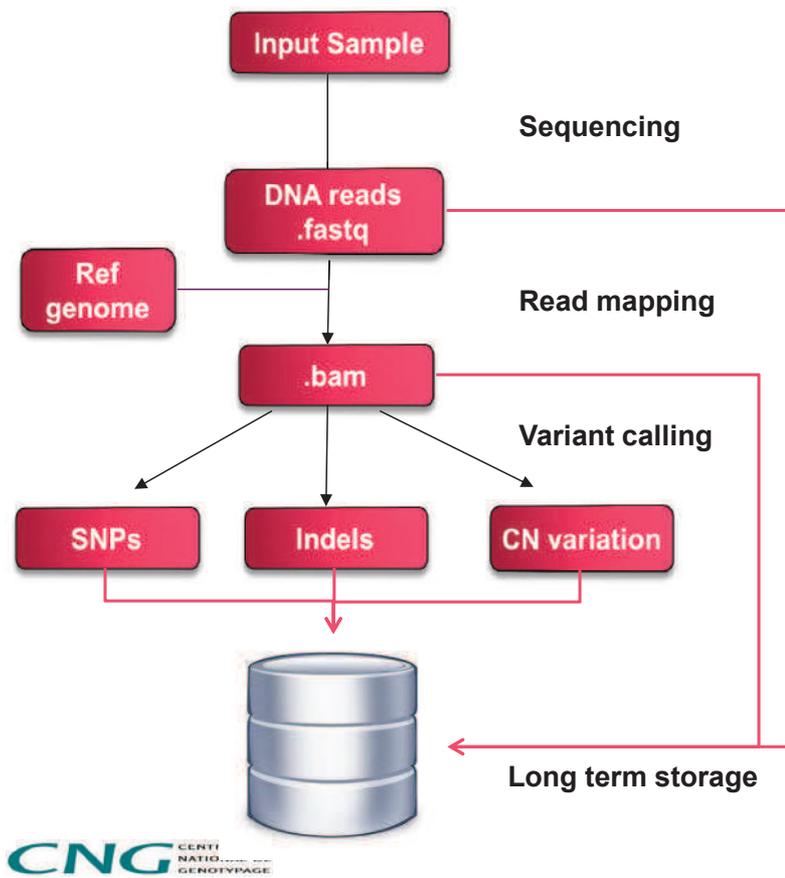
### Principal developments

- Software for management and analysis of NGS and other high-throughput genomic data
- Methods for the prediction of consequences of mutations at protein family level
- Methods for the prediction of binding sites for transcription factor and miRNAs
- Analysis of alterations in splice sites and splice factors
- Prediction of the consequences of alterations in epigenetic markers
- Databases and analysis platforms for pathways and networks
- Systems linking disease and symptoms to molecular entities
- Emerging simulations of biological networks and pathways
- Public repositories of drugs and small molecules
- Systems linking protein networks with drug targets
- Initial systems able to display genomic information and complex analysis in biomedical/clinical environments

### Key problems

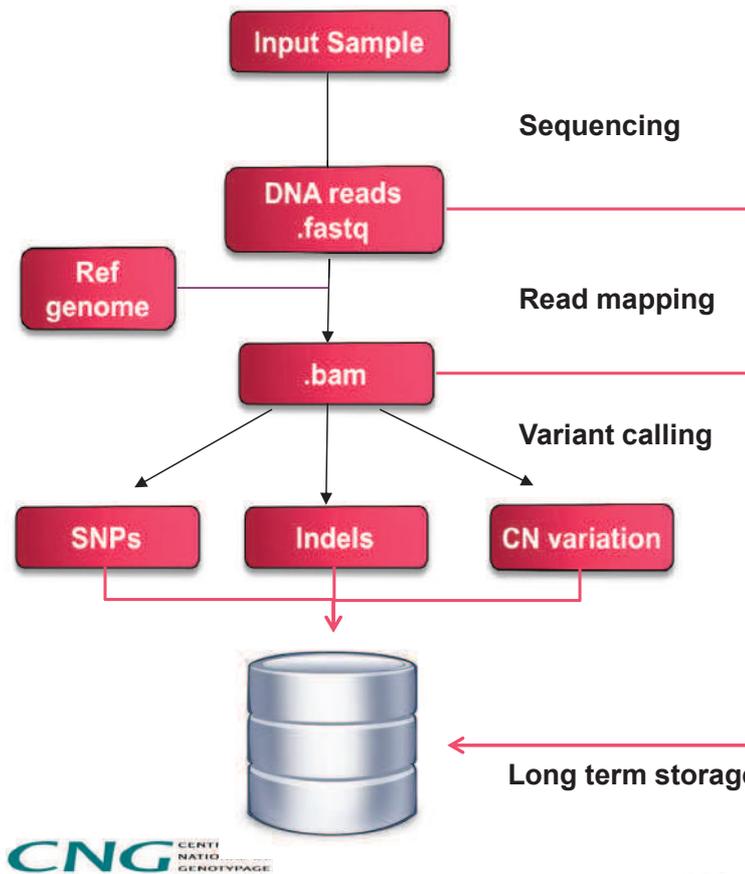
- Data storage and organization
- Linking heterogeneous data types
- Adapting algorithms, methods and tools to the fast evolution of the experimental techniques
- Definition of function at various levels, from biochemical to cellular
- Collection of accurate experimental data on consequences of mutations
- Prediction of activating and deactivating mutations at a quantitative level
- Integration of pathway information
- Completing pathway and network data with experimental information
- Standards for pathway simulation and validation
- Accessibility of medical information and toxicology reports
- Adaptation of text and data base mining system
- Extraction of accurate information on *in vivo* drug targets
- Data protection
- Data provenance and traceability
- Robustness and performance
- Adequate information for users and different level of access

# Varscope : sequencing and polymorphisms



Software	QC Metrics
Casava	Reads quality
Bwa Samtools PicardTools GATK	Mapping report Duplicates Coverage...
GATK	Variants quality - mapping, coverage Variant frequency
SNPeff SNPsift VAAST	Localisation Conservation Functional Impact score

# Scaling up TGCC – portage AS+

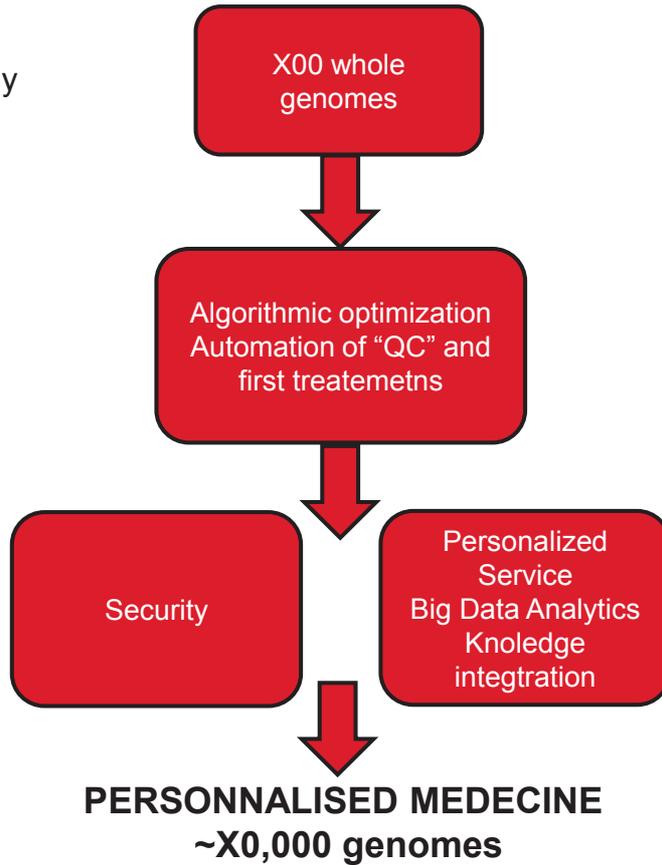


CNG	TGCC
Samples	
500	2500
Mapping BWA (*30) obtenu par distribution du calcul	
75 millions Reads / hour = 50 % prod CNG (150 millionps reads / h))	900 millions reads / day  >> prod CNG
Temps de restitution SNP Calling (*10)	
20 exomes / 24 heures = 30 % prod CNG	200 exomes / 24 h *3 prod CNG
Storage	
10 TB	50 TB

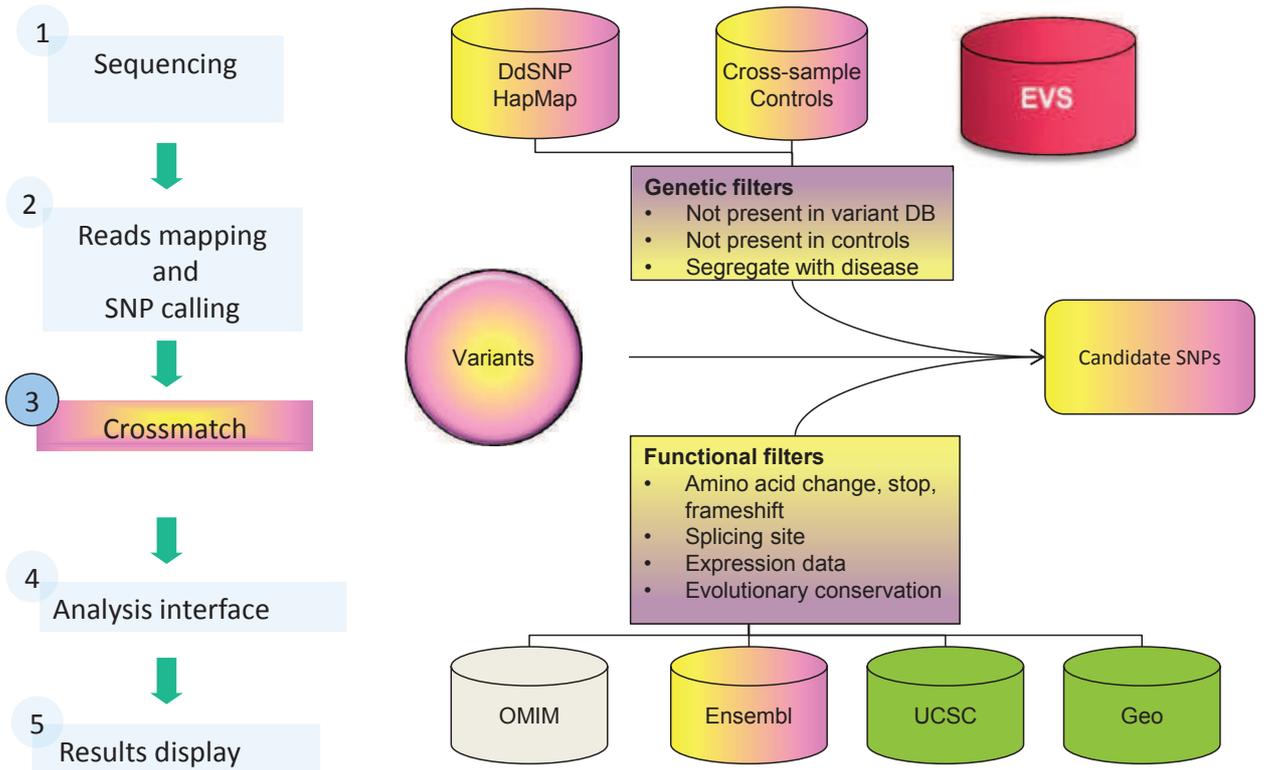
Today  
Poorly automatized  
Heterogeneous technology  
Not an strong demand



Massive genome data



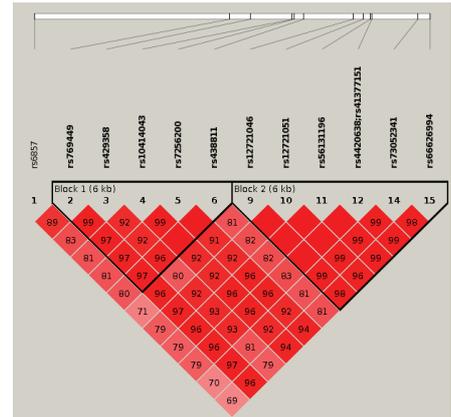
## Polymorphism detection pipeline



- Variants identified by WGS (809 subjects : 40 millions variants)
- Data integration:
  - annotation of variants in coding regions
  - functional annotation of variants in non-coding regions using epigenetic data
  - deleteriousness/conservation scores of variants
- Association tests with the phenotype, at the level of:
  - each variant (plink)
  - each gene / genetic region (SKAT)
- Exploration of combined effects within a subset of associated variants using machine learning methods

- 46% of SNPs appear in only 1 individual: 18 400 000 SNPs
- 20000-25000 SNPs of an individual are specific
- <1% of the SNPs of an individual are specific
- Similar results on 1000 genomes
- SNP set tests may have low power
- SKAT (logistic regression) not good when several rare variants
- Interaction effects hard to detect
- Only very strong associations detected

- Comparison of AD patients (188) versus others (621) with Plink
- **15/16 SNPs with a significant p-value** after Bonferroni correction (Fisher's exact /  $\chi^2$  test) :  $10^{-18} < p < 10^{-9}$   
**in the APOE region (36kb)** including PVRL2, TOMM40, APOE, APOC1, APOC1P1 genes
- **Top associated SNP:**  
rs429358 (missense) :  $p=5 \times 10^{-18}$   
One of the 2 SNPs of **APOE $\epsilon$ 4 allele**
- **Frequent variants (MAF 20-40%)**



François Artiguenave, CEA

Vincent Meyer, CEA  
Florian Sandron, CEA  
Lilia Mesrob, INSERM

Edith Le Floch, CEA

Christophe Battail, CEA  
Solène Julien, Univ Orsay

Nizar Touleimat, CEA  
Xavier Benigni, CEA

Aurélie Leduc CEA

Polymorphism

Statistical analysis

Gene expression

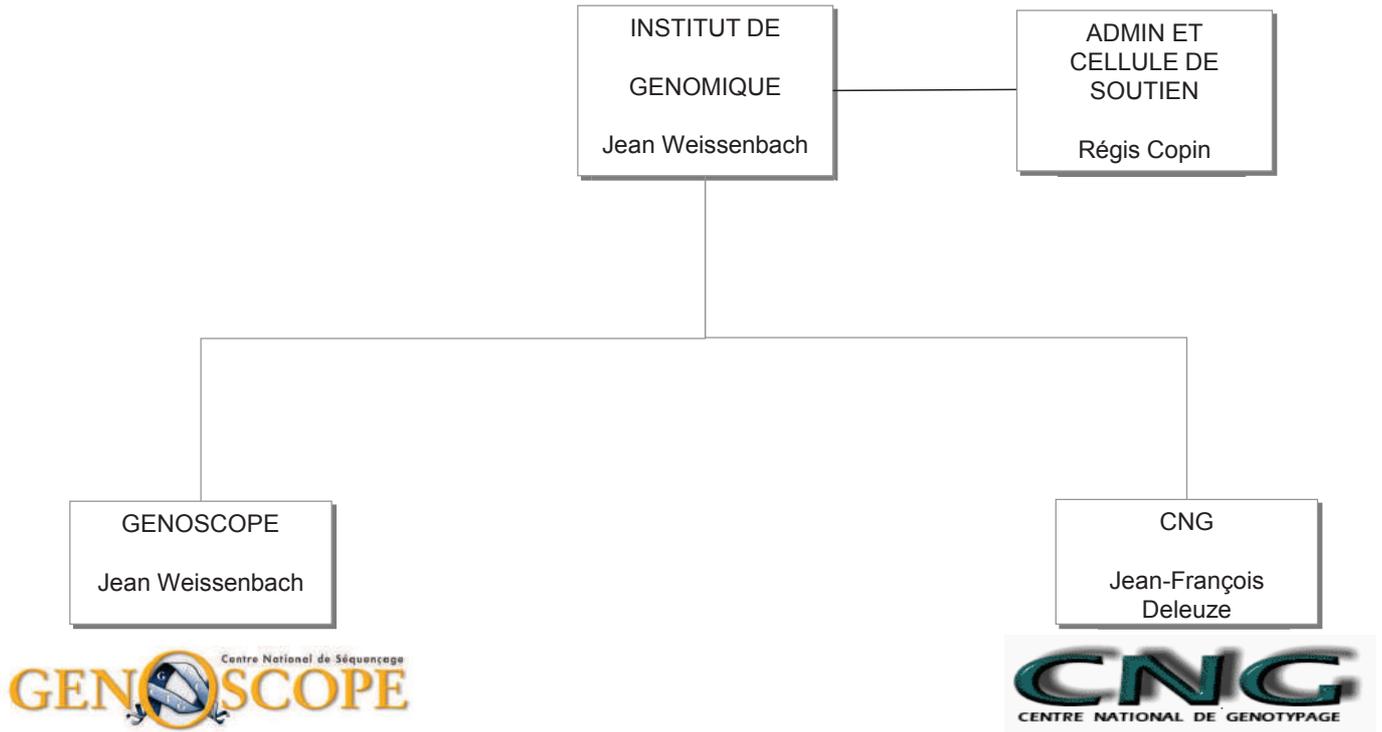
Epigenetics

Informatics Development





# Institut de Génomique



CEA/FAR – COMET | 10 FEVRIER 2014 | PAGE 1



# Institut de Génomique

- Some historical considerations
- Missions
- Overall organization



## Some historical considerations

- Genoscope was created in 1996
- CNG was created in 1997
- initial missions
  - take part in the Human Genome Project (Genoscope)
  - service the French scientific community
  - maintain state of the art technology in sequencing and genotyping
  - run in house projects providing international visibility
  - take IP
- evolution
  - creation of a CNRS research unit (2000) within Genoscope
  - integration in the Direction des sciences du vivant (DSV) of CEA (2007)



## Present missions (general)

- Direction des Sciences du vivant (DSV) from CEA has a commitment towards Ministry of research:
  - to meet the needs of the scientific community in terms of production of sequence and genotyping data.

These needs are taken in charge by the Institut de Génomique whose missions are:

- to produce for the scientific community sequence data from the genomes of various organisms
- to analyze and exploit sequence data for its own needs and those of its partners
- to analyze and interpret human genome sequence with the aim of understanding human diseases with a genetic component

## Genetics of human diseases

To identify and understand the molecular bases of human diseases in order to improve their diagnostic and treatment.

Identification is based on whole genome molecular analyses

## Genome analyses on all types of organisms

prokaryotes  
economic interest  
evolution

## Develop new tools for industrial biotechnology

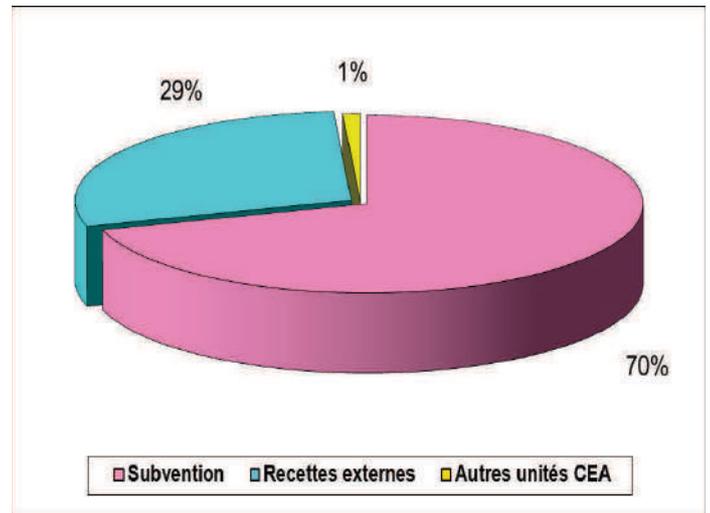
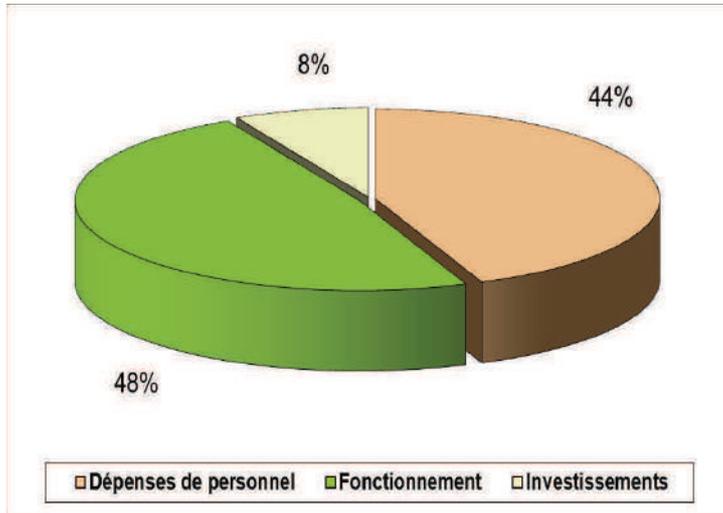
exploiting biodiversity explored by sequencing  
using synthetic biology approaches

## Budget Programmes et Effectifs IG 2014 (par natures)

Dépenses de personnel	15,6
Fonctionnement	17,2
Investissements	2,7
<b>Total</b>	<b>35,6</b>

En M€

Subvention	25,0
Recettes externes	10,2
Autres unités CEA	0,4
<b>Total</b>	<b>35,6</b>



CEA/FAR – COMET | 10 FEVRIER 2014 | PAGE 7

## Salariés à l'IG par statut au 31/12/2013

Staff 31/12/2013

	EC	CNG	CNS	Total
Permanent	16	62	126	204
Non-permanent	1	8	30	39
External collaborators		13	20	33
Scientific advisors	1		1	2
	18	83	177	278

# Computation

- In house (Genoscope + CNG)
  - computation 1200 cores
  - storage 2 Pbytes
- Remote France-Genomique resource at TGCC (DAM/CEA)
  - computation 3000 cores
  - storage 5 Pbytes

## Organisation générale (2014) Centre National de Séquençage

