Mining Omics Data Helps Understand Phenotype-specific Biological Mechanisms

11. 8. 2013

Sun Kim
Department of Computer Science and Engineering
Bioinformatics Institute
Interdisciplinary Program in Bioinformatics
Seoul National University

Outline

- Some research questions omics projects
- Omics data?
- Genomics and epigenomics
- Breast cancer project
- Drought resistant rice project
- Discussion

Drug Resistant Cancer

- The OSU-IU Center for Cancer Systems Biology has been investigating the mechanism of developing drug resistance in

  breast, prostate, and ovarian cancer.
- In particular, we are interested in investigating

  changes in *epigenetic mechanisms* in terms of

  gene regulation and pathway activation

  while in transition to a hormone-/chemo-sensitive

  to ***a hormone-/chemo-insensitive phenotype*** in cancer.

Breast Cancer Subtypes

Cancer. 2010 Jan 15;116(2):486-96.

Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness.

Pedraza V, Gomez-Capilla JA, Escaramis G, Gomez C, Torné P, Rivera JM, Gil A, Araque P, Olea N, Estivill X, Fárez-Vidal ME.

- Human biology underlying breast cancer subtypes.
- A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell.* 2006. 10:515–527

The Breast Cancer Research Foundation.

Prevention and A Cure in Our Lifetime.
www.bcrfcure.org

Drought Resistant Rice

- Hunting for genes that can be modified -- hopefully safely – for more drought resistant rice

These Questions Can Be Better Answered by Constructing Networks of Biological Entities of Multiple Types.

- (Current research)
  genomic variations
  → cancer susceptibility, etc

- **(New direction)**
  Networks of genetic and epigenetic elements
  → cancer susceptibility, etc

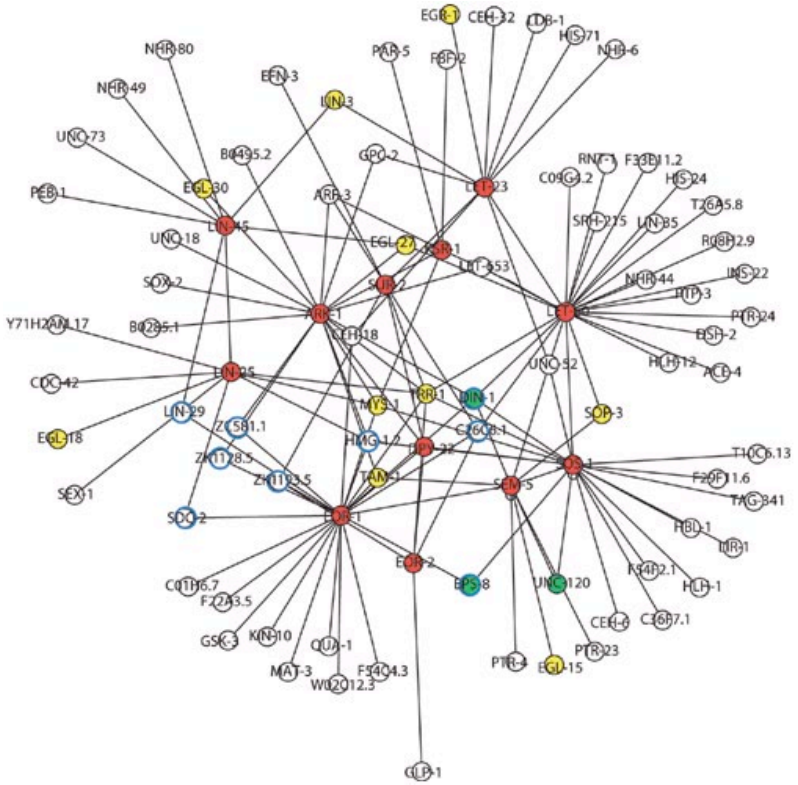


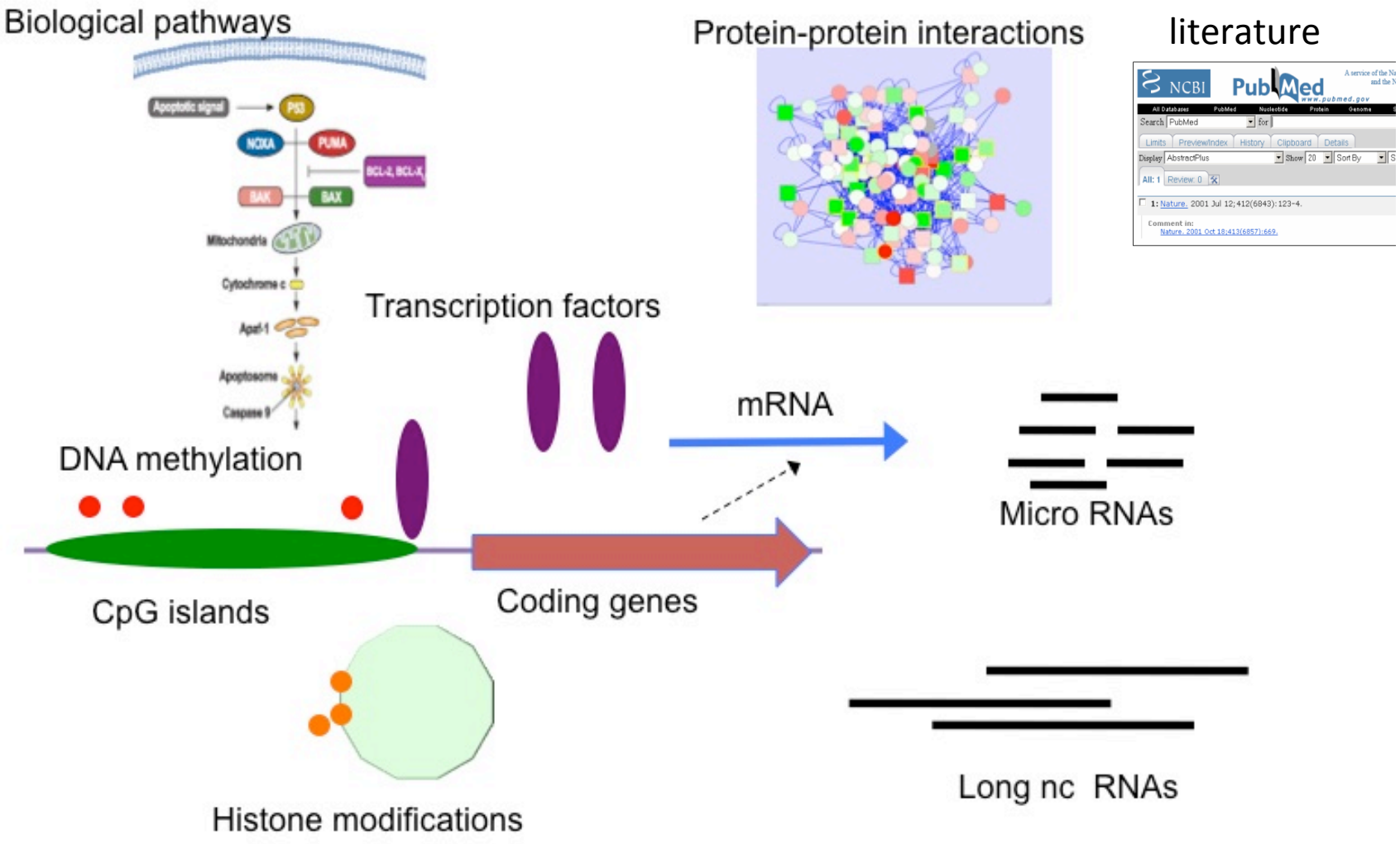
Image from Nature Reviews Genetics 7, 664-665

Genomics, epigenomics, and phenotypes

Omics

- Complete measurement of biological events in the whole cell.
- **Genomics**
  - Genome sequence
  - Expression of genes
- **Epigenomics**
  - "Epi" means "on" or "upon", thus control mechanisms for genetic elements
  - DNA methylation
  - Histone modification
  - Non-coding RNA interference with coding genes.

Genetic and Epigenetic Elements



Genomics and Disease

Centers for Disease Control and Prevention
Your Online Source for Credible Health Information

All CDC Topics
Choose a topic above

A-Z Index A B C D E F G H I J K L M N O P Q R S T U V W X Y Z #

Public Health Genomics

- **Genomics**
- About Us
- Impact Update
- **Genomics and Health**
- Population Research
- Genomics Translation
- Family Health History
- Genetic Testing
- Genomic Resources
- Site Map

Genomics

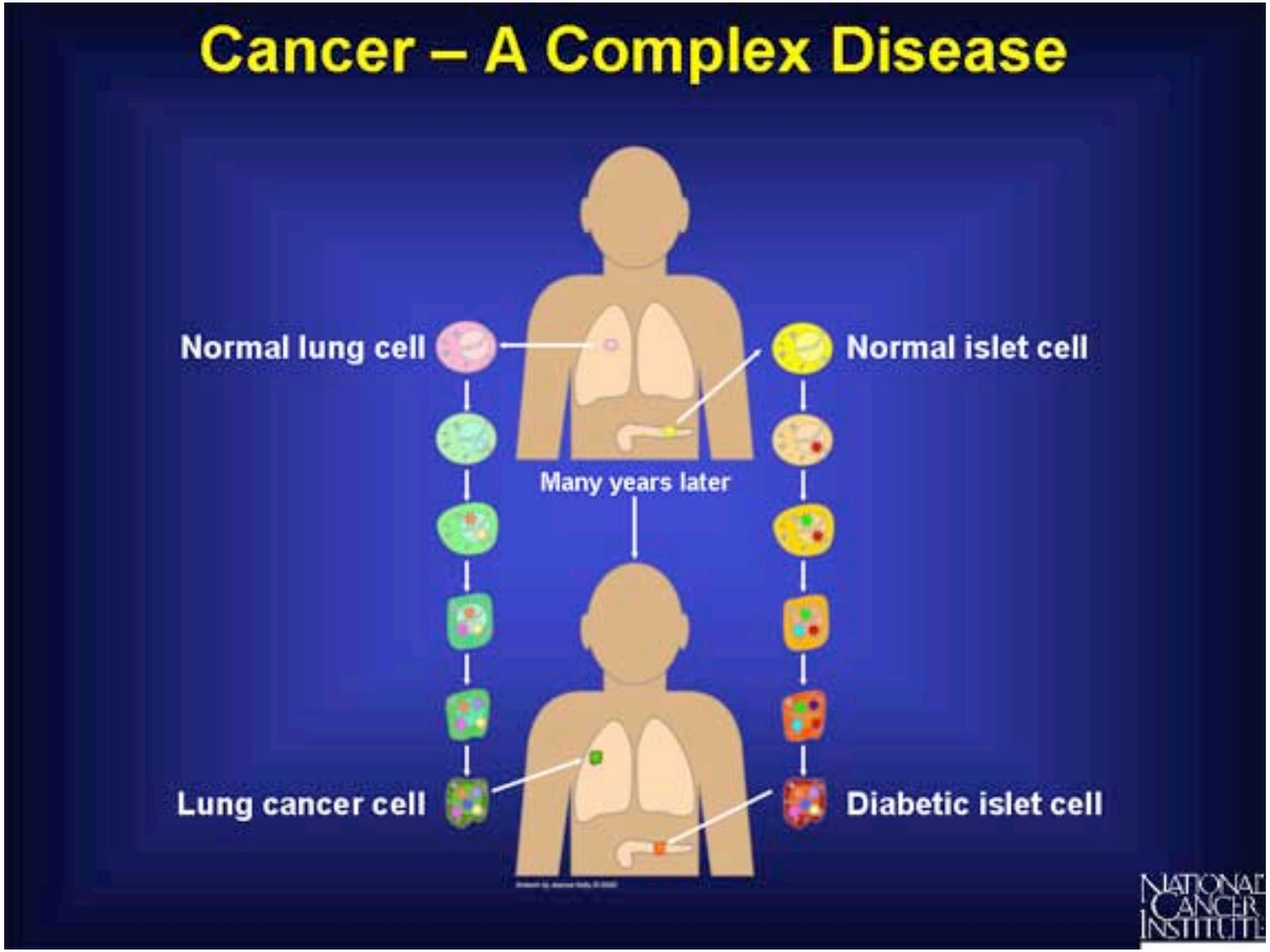
Genomics and Health

Genomics plays a role in nine of the Ten Leading Causes of Death in the United States, most notably cancer and heart disease. These diseases are partly the result of how genes interact with environmental and behavioral risk factors, such as diet and physical activity. Also, a large fraction of children's hospitalizations are due to diseases that have genetic components.

By studying the relationship between genes, environment, and behaviors, researchers and practitioners can learn why some people get sick, while others do not. Family health history information can also help to identify people who may have a higher risk for certain diseases. Better understanding of genetic and family history information can help researchers and practitioners identify, develop, and evaluate screening and other interventions that can improve health and prevent disease. Individuals can contribute to their health by keeping records of their family health information and sharing this information with their doctor and with other family members.

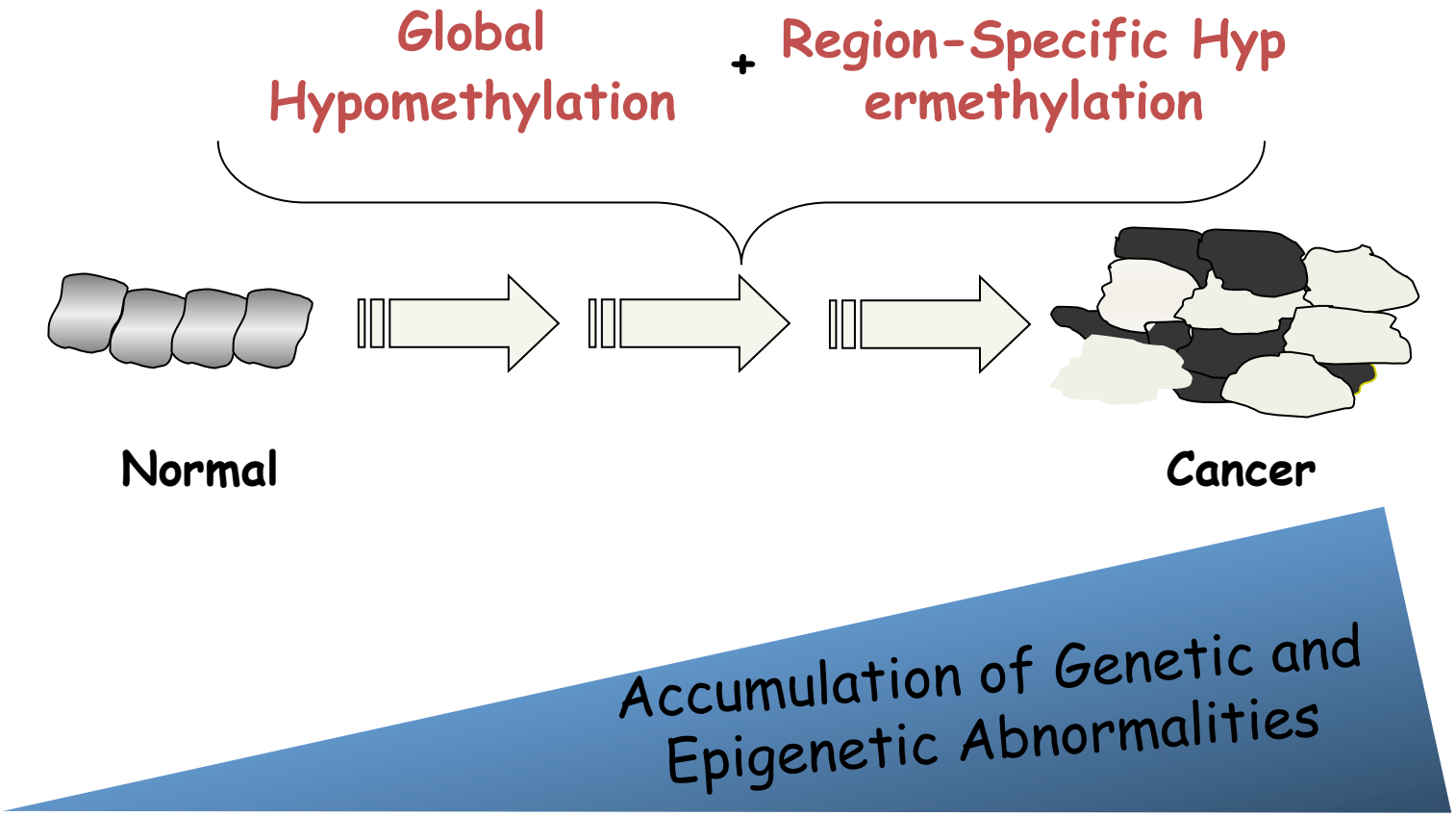
Learn More About Genomics and Health

http://www.cdc.gov/genomics/public/index.htm

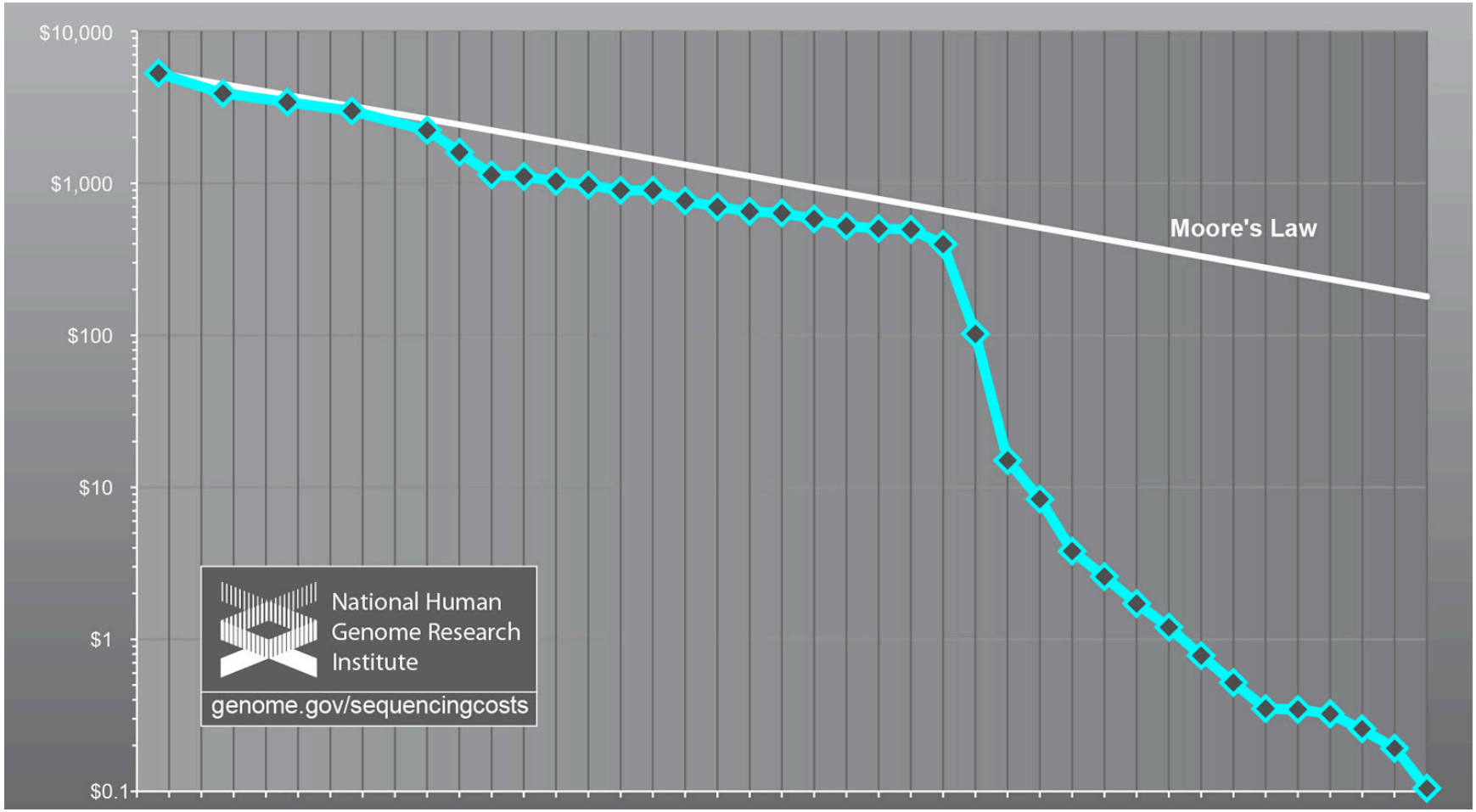


http://www.cancer.gov/cancertopics/understandingcancer/geneticvariation/page2

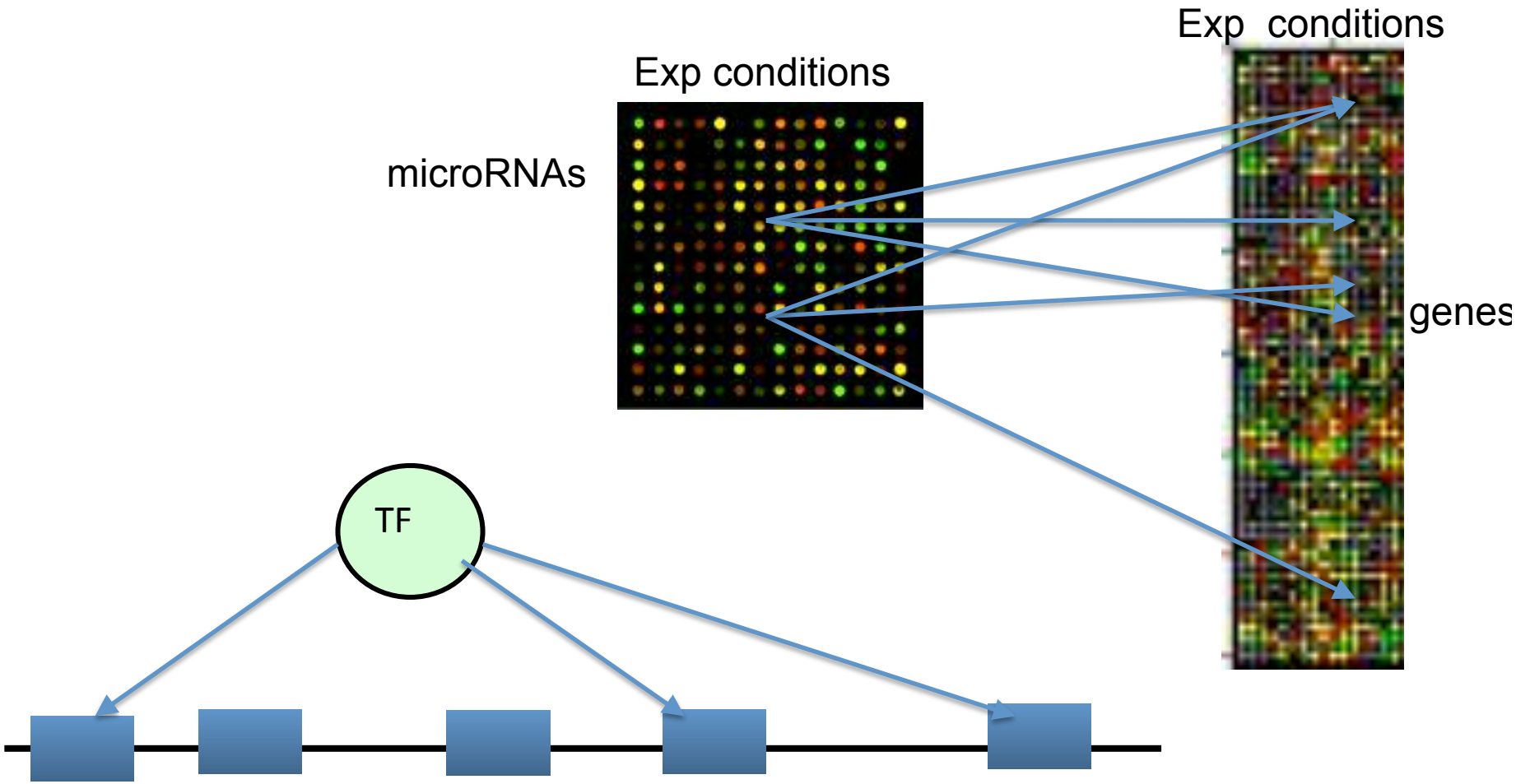
Progressive Accumulation of DNA Methylation in Cancer



The 1st and 2nd Revolution in Sequencing Technologies



TF and microRNA Regulates Many Genes



Breast cancer subtypes (since 2005)

Breast Cancer Project

- A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell.* 2006. 10:515–527

- Already available
  - Genome-wide DNA methylation data
  - Gene expression data
- We plan to do:
  - Sequencing of selected genes and CpG islands.
  - TF ChIP-seq
  - RNA-seq

The Breast Cancer Research Foundation
Prevention and A Cure in Our Lifetime.
www.bcrfcure.org

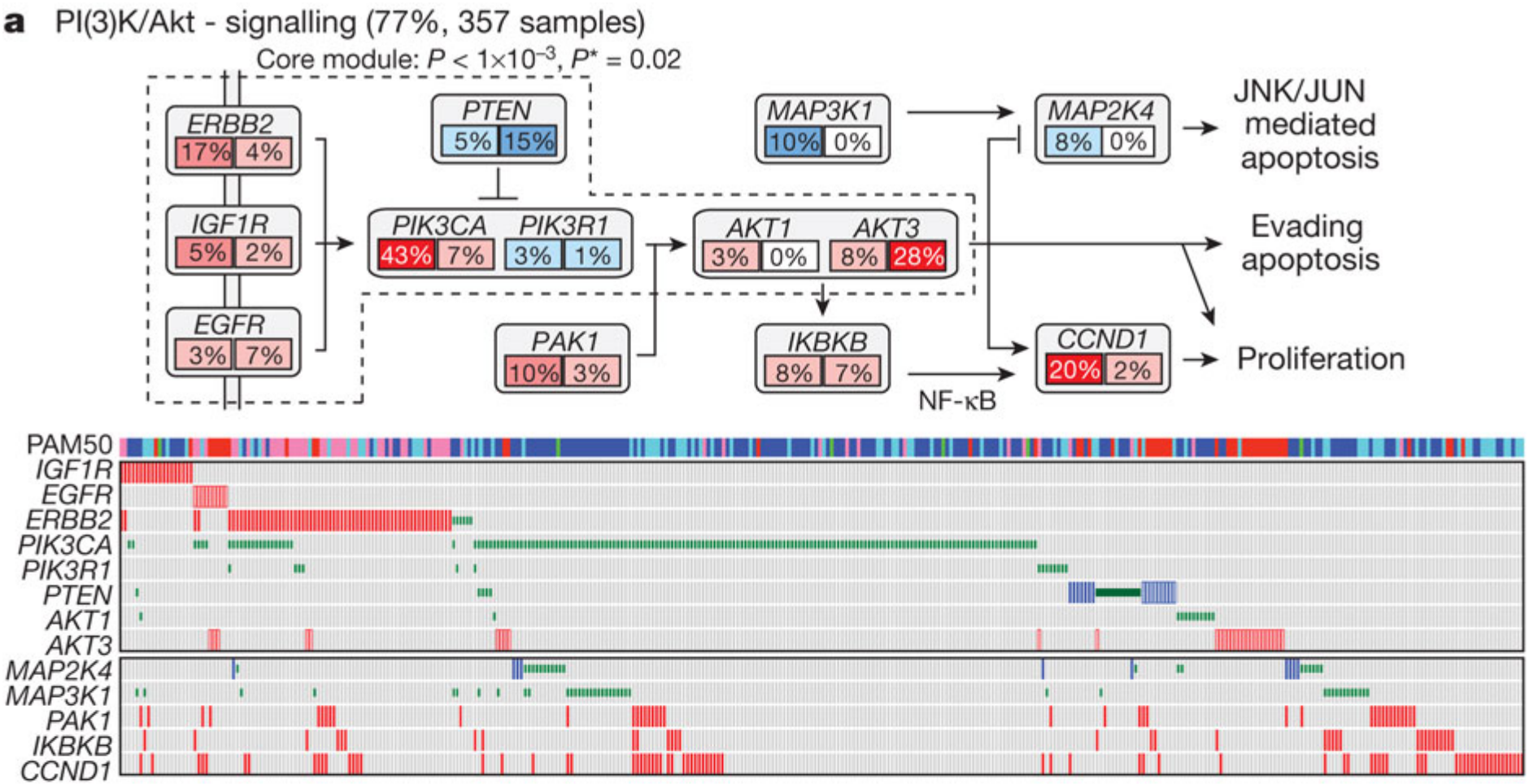
The Cancer Genome Atlas

Understanding genomics to improve cancer care

- http://cancergenome.nih.gov/
- To chart the genomic changes involved in more than 20 types of cancer.
- To date, TCGA has achieved comprehensive sequencing, characterization, and analysis of the genomic changes in the brain cancer, glioblastoma multiforme, and ovarian cancer.

Mutual exclusively modules in cancer (MEMo) analysis

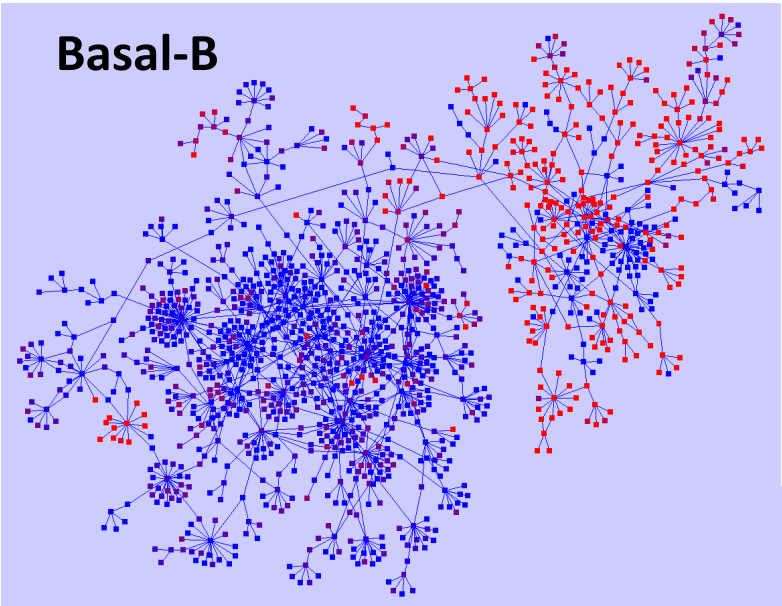
DC Koboldt *et al. Nature* **000**, 1-10 (2012) doi:10.1038/nature11412

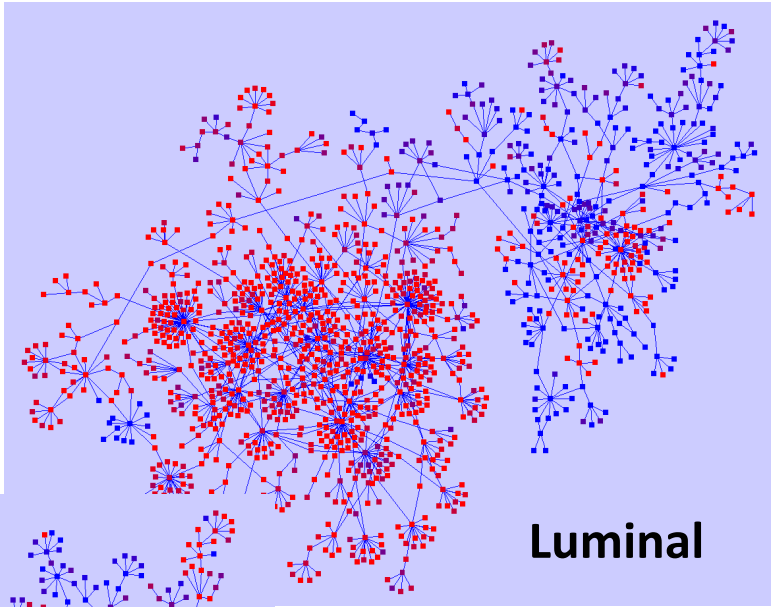


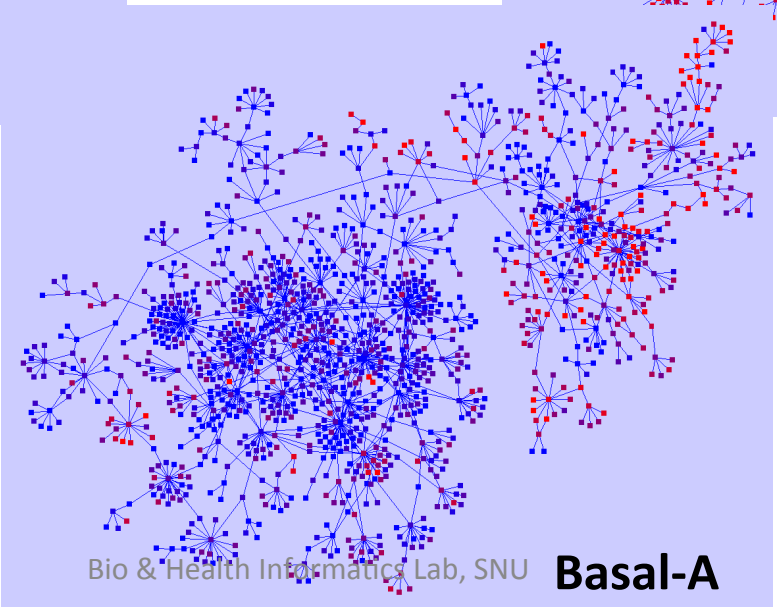
3 Research Questions

- Towards transcriptional network biomarkers.
- TF-DNA interactions and their effect on downstream gene regulation.
- Integrated analysis of microRNA and mRNA omics data.

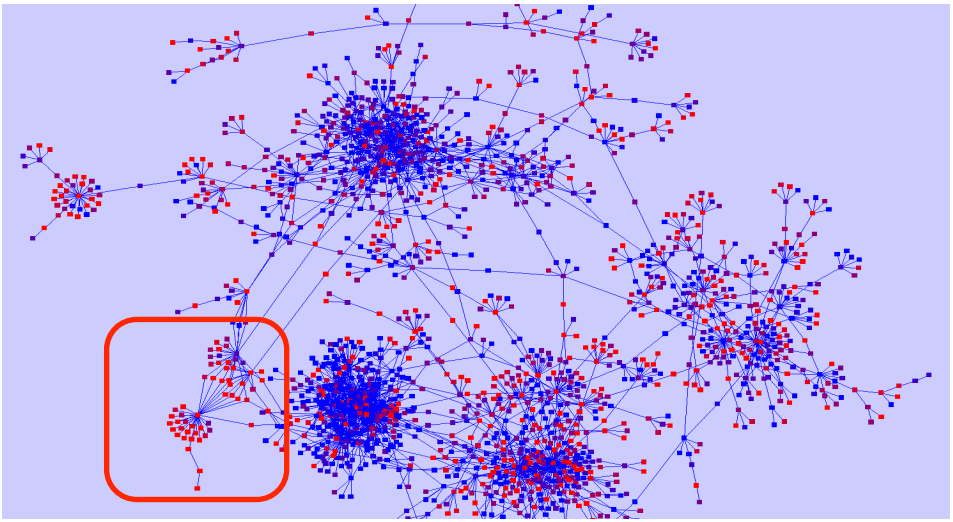
TF Networks of 30 Breast Cancer Cells



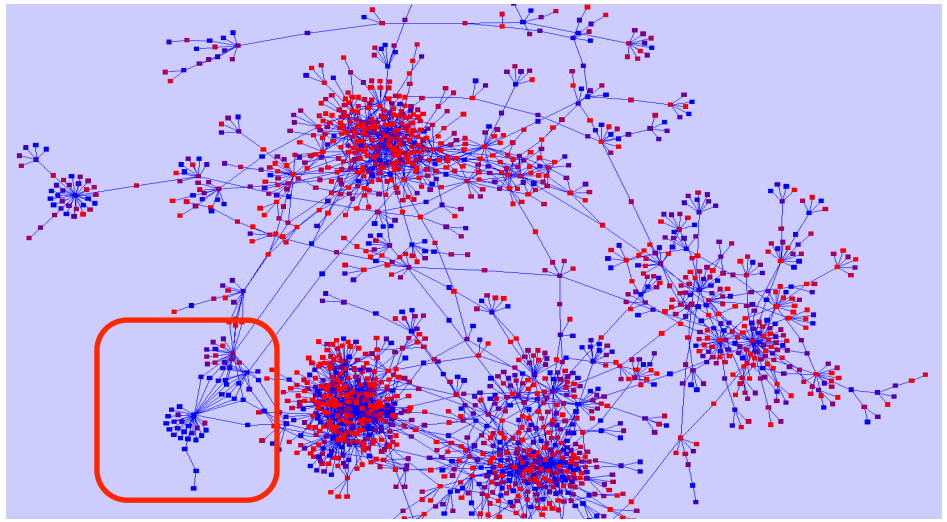




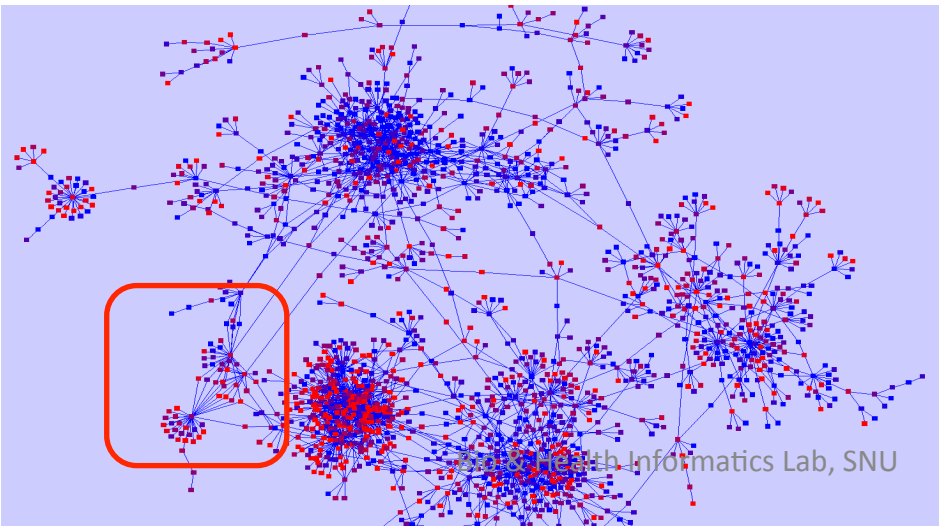
Network Constructed from TCGA BRC Normal Data



10 Basal-B Cell Lines



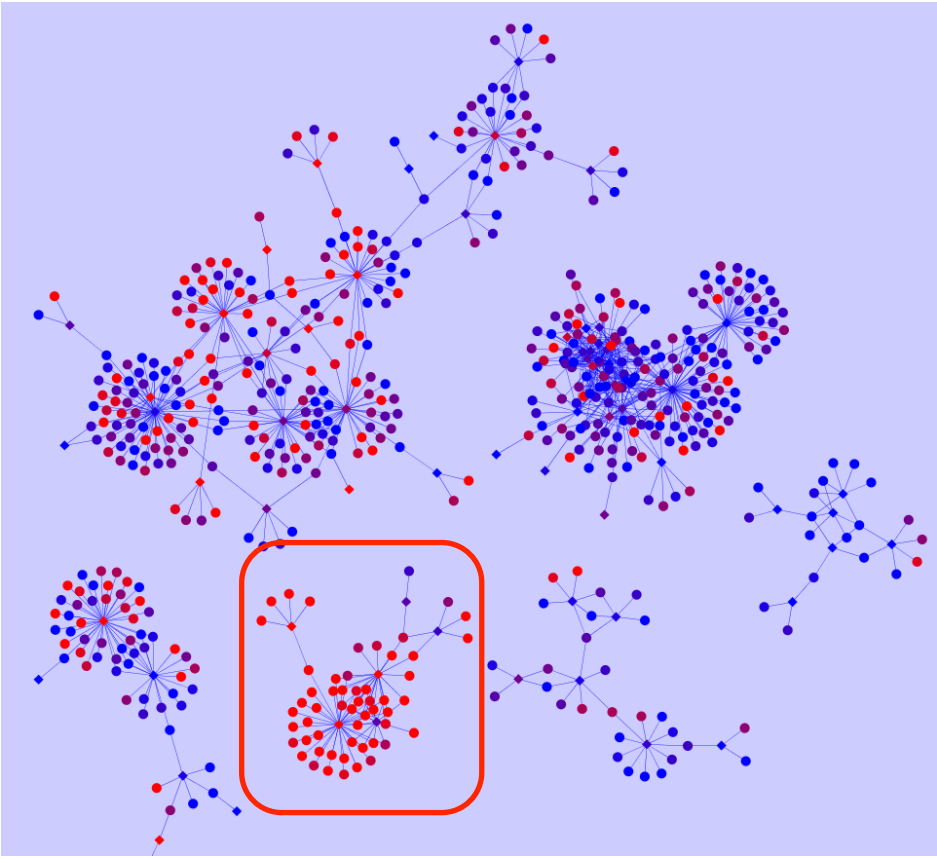
10 Luminal Cell Lines



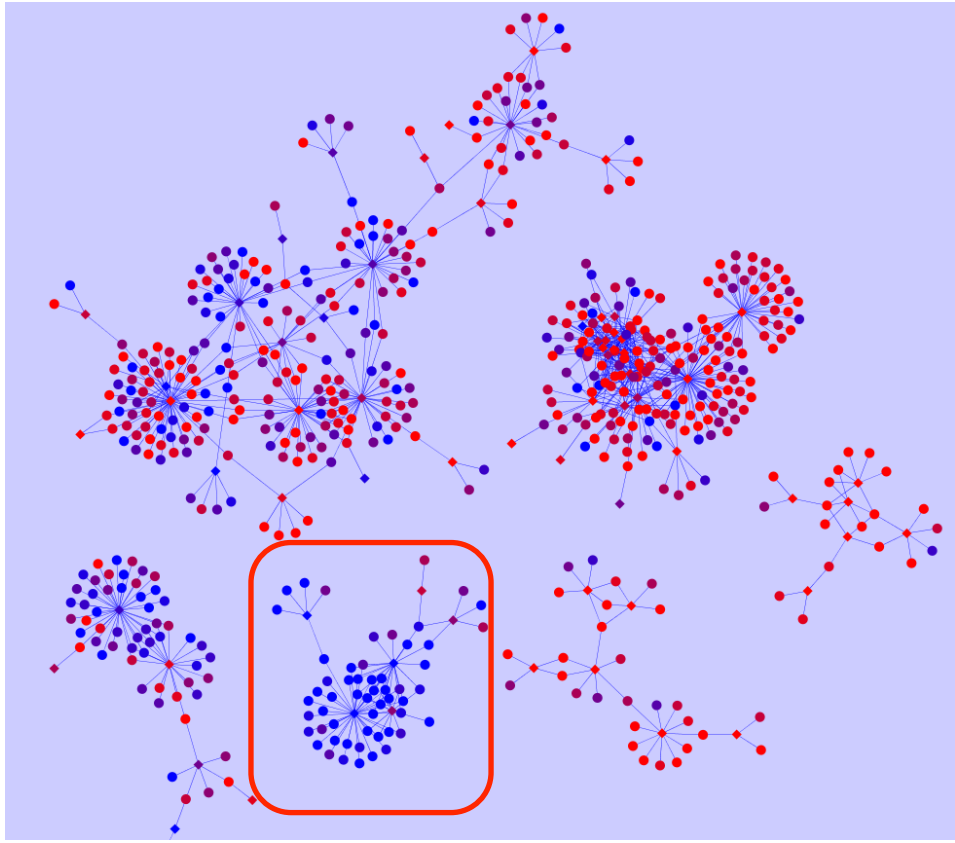
7 Basal-A Cell Lines

Network Constructed from TCGA BRC Tumor Data

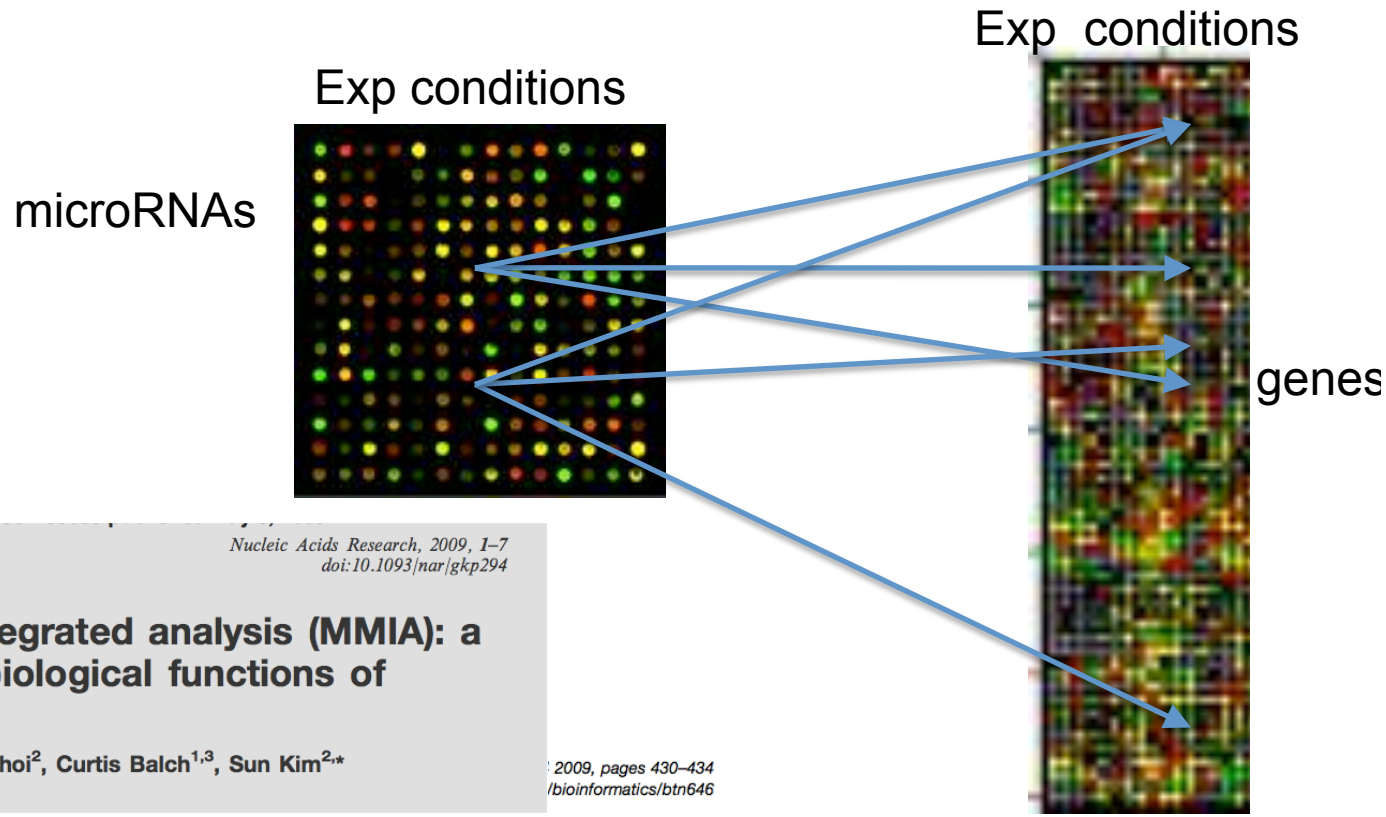
10 Basal-B Cell Lines



10 Luminal Cell Lines



Network of microRNAs and Genes



Nucleic Acids Research, 2009, 1–7
doi:10.1093/nar/gkp294

MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression

Seungyoon Nam[1], Meng Li[1], Kwangmin Choi[2], Curtis Balch[1,3], Sun Kim[2,*] and Kenneth P. Nephew[1,3,*]

2009, pages 430–434
/bioinformatics/btn646

Gene expression

Computational analysis of microRNA profiles and their target genes suggests significant involvement in breast cancer antiestrogen resistance

Fuxiao Xin[1], Meng Li[1,2,3], Curt Balch[2,4], Michael Thomson[5], Meiyun Fan[6], Yunlong Liu[7], Scott M. Hammond[8], Sun Kim[1,9,*] and Kenneth P. Nephew[2,3,4,6,10,*]

microRNA Networks

- Computational Analysis of MicroRNA Profiles and Their Target Genes Suggests Significant Involvement in Breast Cancer Antiestrogen Resistance. *Bioinformatics*. 2009 Feb 15;25(4):430-4.
- Nam S, Li M, Choi K, Balch C, Kim S, Nephew KP. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. Nucleic Acids Res. 2009 May 6.
- An integrative analysis of cellular contexts, miRNAs and mRNAs reveals network clusters associated with antiestrogen-resistant breast cancer cells. *BMC Genomics*. 2012 Dec 27;13:732.

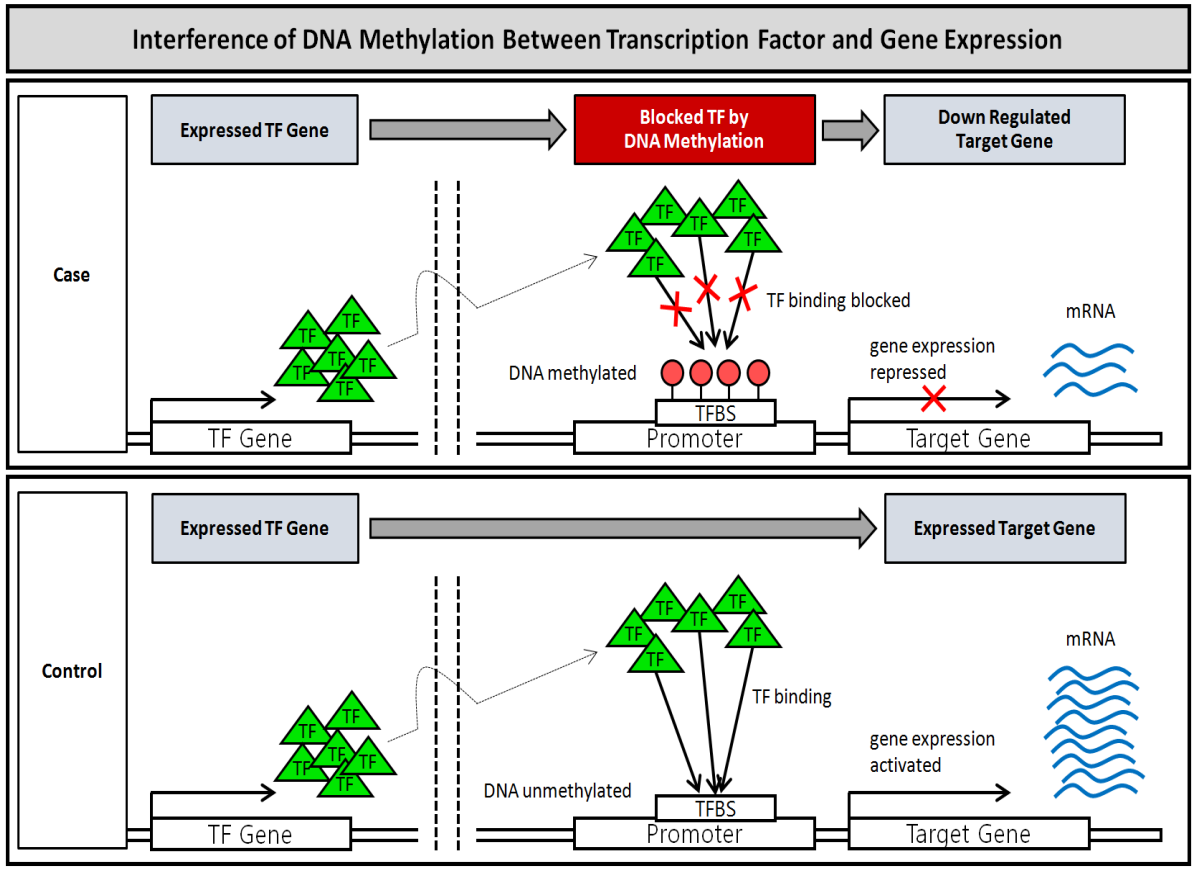
- We are expanding MMIA for the sequencing data and also developing new algorithms.

- **Sequence microRNA Champiagn!**

Transcription Factor & DNA Methylation Analysis in Cancer

- A mixture model-based discriminate analysis for identifying ordered transcription factor binding site pairs in gene promoters directly regulated by estrogen receptor-alpha. *Bioinformatics*. 2006;22(18): 2210-6.
- dPattern: transcription factor binding site (TFBS) discovery in human genome using a discriminative pattern analysis. *Bioinformatics*. 2007 Oct 1;23(19):2619-21.
- Predicting DNA methylation susceptibility using CpG flanking sequences. *Pac Symp Biocomput*. 2008:315-26.
- Enriched transcription factor binding sites in hypermethylated gene promoters in drug resistant cancer cells. *Bioinformatics*. 2008 Aug 15;24(16):1745-8.
- Integrated Analysis of DNA Methylation and Gene Expression Reveals Specific Signaling Pathways Associated with Platinum Resistance in Ovarian Cancer. *BMC Medical Genomics*, 2009, 2:34
- Genome-wide DNA methylation maps in follicular lymphoma cells determined by methylation-enriched bisulfite sequencing, *PLoS ONE*, 2010 Sep 29;5(9)
- Oncogenic ETS proteins mimic activated RAS/MAPK signaling in prostate cells. *Genes and Development*, 2011. 25: 2147-2157.
- A Novel K-mer Mixture Logistic Regression for Methylation Susceptibility Modeling of CpG Dinucleotides in Human Gene Promoters, *BMC Bioinformatics*, 2012, 13(suppl 3)
- CpG island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene*, 2012
- Genome-wide analysis and modeling of DNA methylation susceptibility in 30 breast cancer cell lines by using CpG flanking sequences. *Journal of Bioinformatics and Computational Biology*. in press.
- Integrated Analysis of Genome-wide DNA Methylation and Gene Expression Profiles in Molecular Subtypes of Breast Cancer, *NAR*, in press.

Step 1. TF network construction
Step 2. Adding DNA methylation ➔ 중요 TF 발굴
Step 3. 중요 TF의 TF-Chip Sequencing



Schematic overview of the phenotype-comparative analysis for interference of TF binding by DNA methylation resulting in the suppression of downstream gene expression

Target gene	Binding TF	TFBS support rate
CDH1	SMAD1	100.0
CDH1	FOXO1	100.0
CLDN4	CEBPA	62.5
CLDN4	CEBPB	62.5
CLDN4	CEBPD	62.5
CLDN4	CEBPE	62.5
CLDN4	CEBPG	62.5
ESRP1	CUX1	90.0
GRHL2	PDX1	100.0
KRT19	PAX6	60.0
PRR15L	IKZF1	50.0
AKR1B1	E2F1	91.7
PLOD2	PAX3	100.0

Downregulated target gene with TFBS on h ypermethylated region

NAR, 2013

Are These TF Related to Breast Cancer?

- Genes *CDH1*, *ESRP1* and *GRHL2* have been shown to play critical roles in epithelial-mesenchymal transition (EMT), a process associated with metastatic events in cancer and also highly relevant to tumor progression (32,33).
- A study by Dumont *et al.* (35) showed that the induction of EMT was accompanied by repression of *CDH1* expression and subsequent DNA hypermethylation at its promoter in basal-like breast cancer.
- Lombaerts et al. (34) reported that CDH1 is downregulated by promoter methylation and related to EMT in breast cancer cell lines.

Are These TF Related to Breast Cancer? (continued)

- Additionally, recent studies showed that *GRHL2* and *CDH1* in human breast cancer cells were highly correlated and suppressed EMT by repressing expression of the *ZEB1* gene (36,37).
- *ESRP1* was shown to regulate a switch in CD44 alternative splicing, an event required for EMT and breast cancer progression (38).
- Moreover, there might be potential interplay between target genes. Overexpression of *GRHL2* upregulated *ESRP1* expression (36) and *GRHL2* was shown to be essential for adequate expression of the *CDH1* and *CLDN4* (39).

Complex Relationship Analysis Among Transcription Factor, DNA Methylation, Mutation, and Gene Expression

- **Integrated Analysis of Genome-wide DNA Methylation and Gene Expression Profiles in Molecular Subtypes of Breast Cancer, *NAR*, in press**.

- mCpG-SNP-EXPRESS (http://biohealth.snu.ac.kr/mcpg-snp-express/)
- **mCpG-SNP-*EXPRESS*: An Integrated Analysis of DNA Methylation, Sequence Variation (SNPs), and Gene Expression for Distinguishing Cellular Phenotypes**

- **Modeling DNA methylation susceptibility and CpG islands.**

- **A long long way to go!**

A 3-Step Strategy to Construct TF Networks

- **(1st Step)**: Putative TF network construction by analyzing transcriptome data from cell lines of multiple phenotypes and also from data in the public domain.
- **(2nd Step)**: DNA methylation and/or histone data to identify and rank DNA-TF interactions.
  - Suggest TF ChIP-seq experiments.
- **(3rd Step)**: Reconstruct TF network utilizing transcriptome, methylome, and TF ChIP-seq data.

Drought resistant rice (since 2012)

Drought Resistant Rice

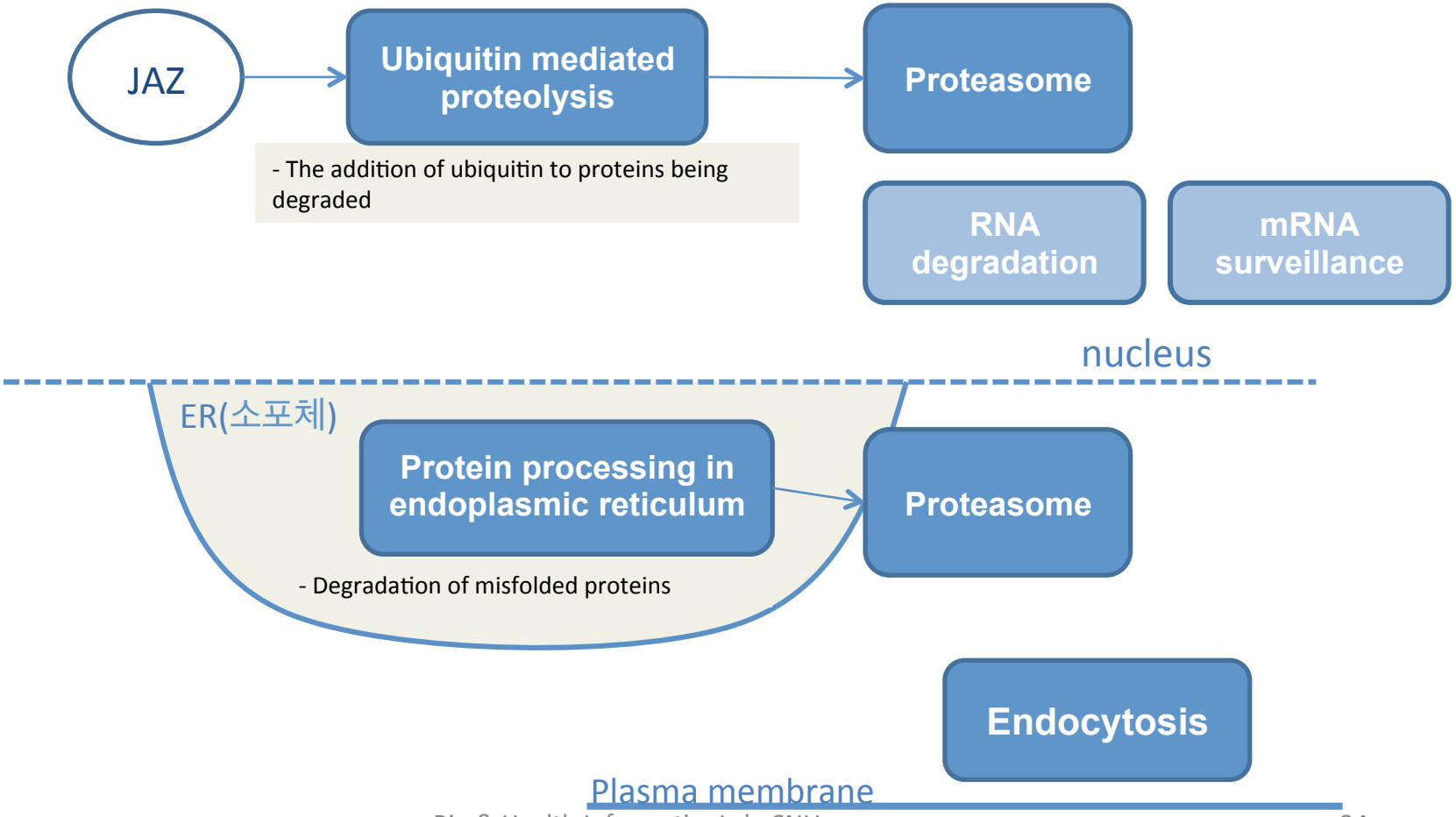
- **Difference in biological mechanisms between drought resistant rice and non-drought resistant rice**

- mRNA-seq
- MicroRNA
- DNA methylation
- TF ChIP-seq

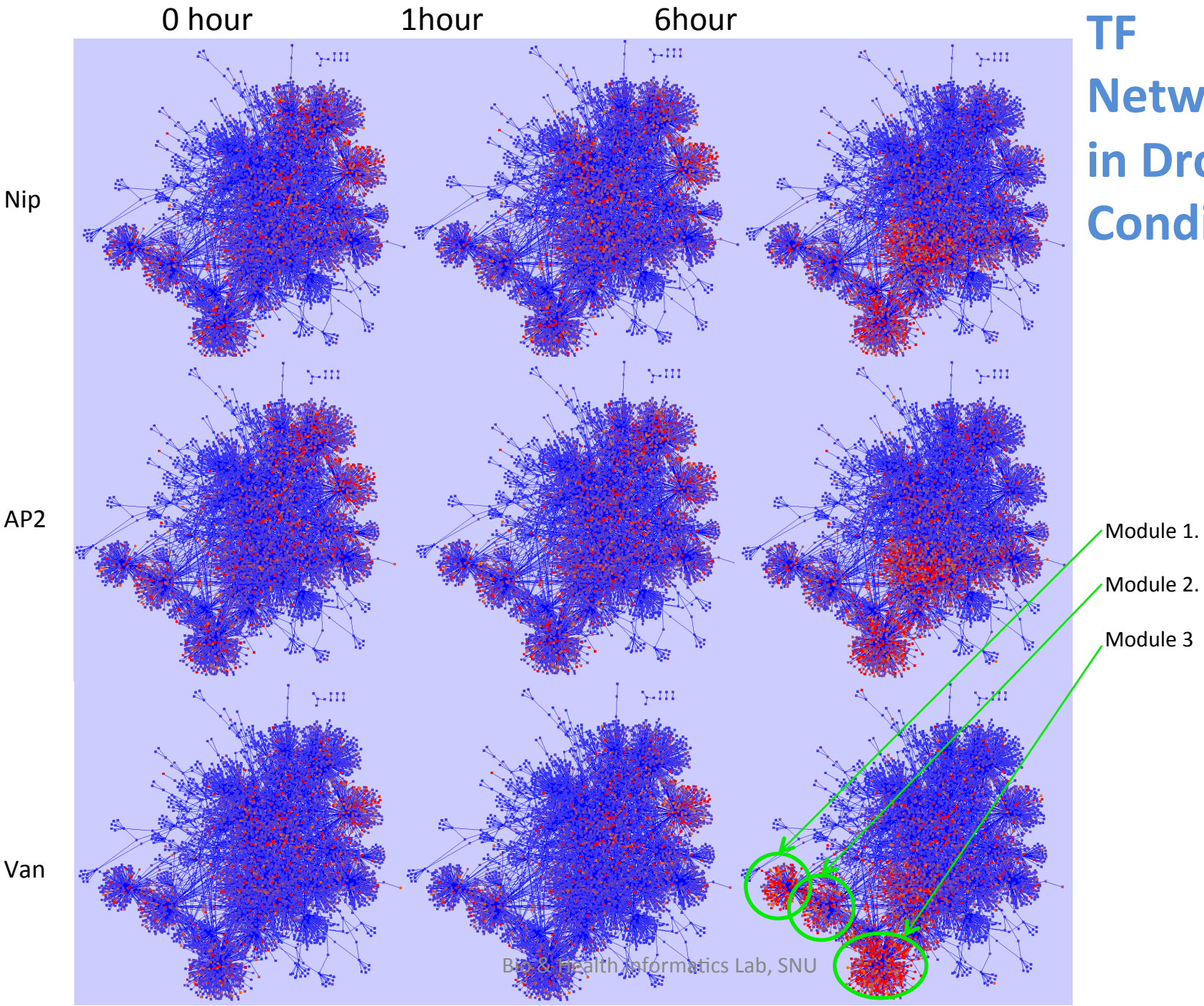
The Same Biological Mechanisms Were Highlighted by Four Omics Data Analyses

- Differentially expressed pathways in drought resistant conditions
- TF network analysis in drought resistant conditions
- Gene fusion, alternative splicing, and isoforms in drought resistant conditions
- Non-coding RNA interference with coding genes in drought resistant conditions

Pathways Interactions in Drought Conditions



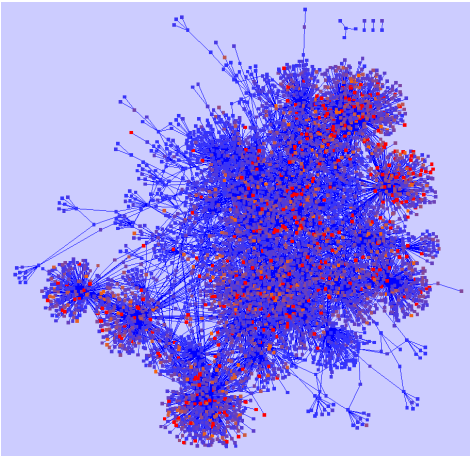
TF Networks in Drought Conditions



Other Omics Projects

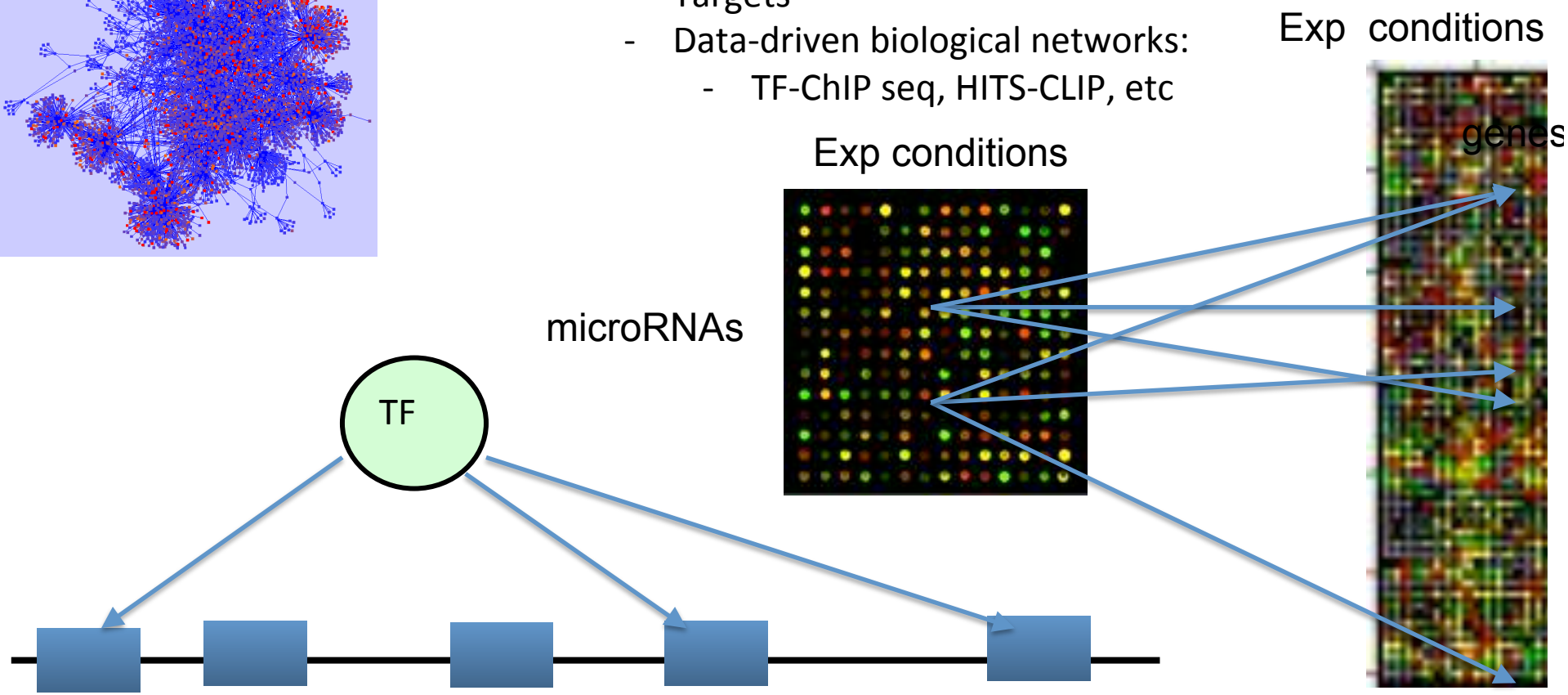
- Prostrate cancer
- Ovarian cancer
- Autoimmune disease
- Organ transplant and immune systems
- Huntington disease
- Global warming and plants

Nice Marriage Between Computational and Experimental Biology



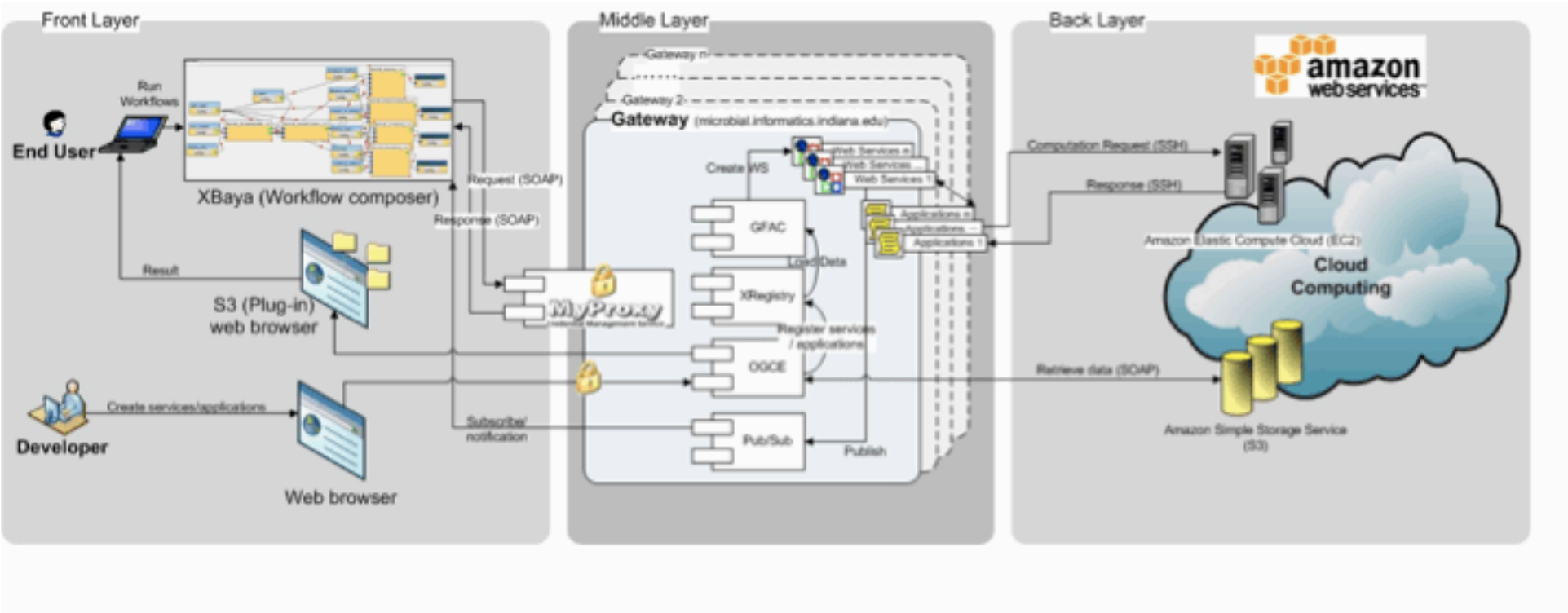
Next step experiments:

- Knocking down genes including TFs
- Targets
- Data-driven biological networks:
  - TF-ChIP seq, HITS-CLIP, etc



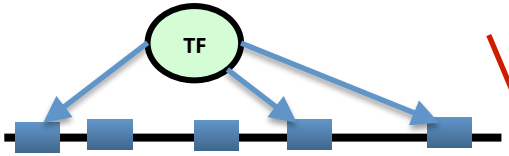
BioVLAB: A Reconfigurable Cloud Infrastructure for Bioinformatics

- Collaboration between biologists and informaticians.



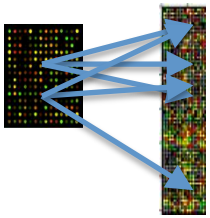
Putting Altogether

- BioVLAB을 통한 의학/생물 연구자들과의 공동 연구 인프라구축



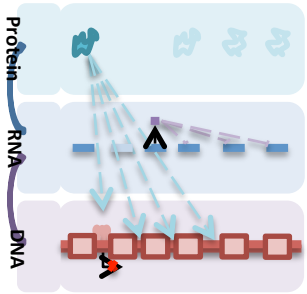
TF 유전자 조절 네트워크

+



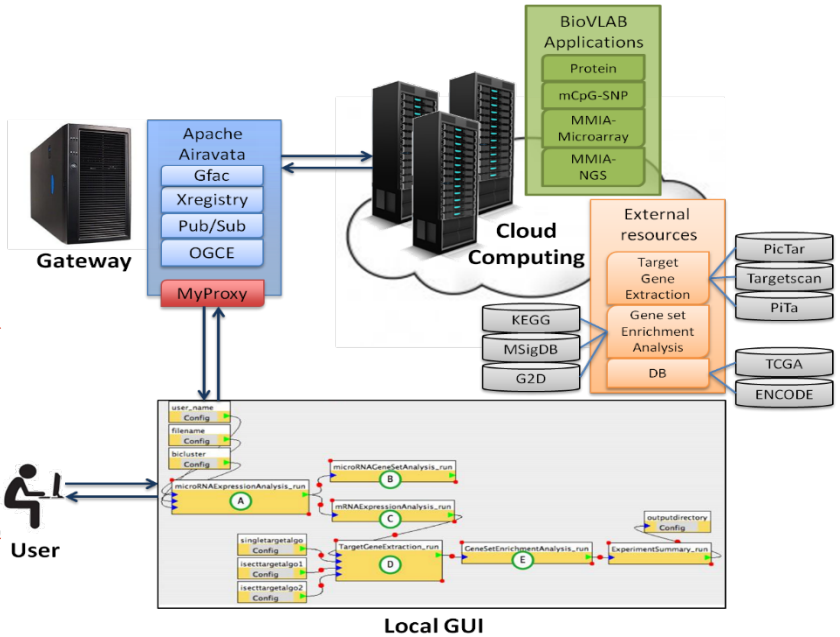
miRNA 유전자 조절 네트워크

(+ DNA 메틸화 정보)

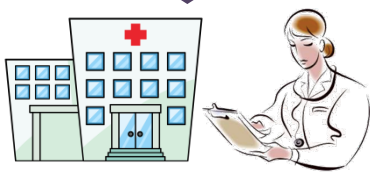


TF 와 miRNA 가 통합된 유전자 조절네트워크

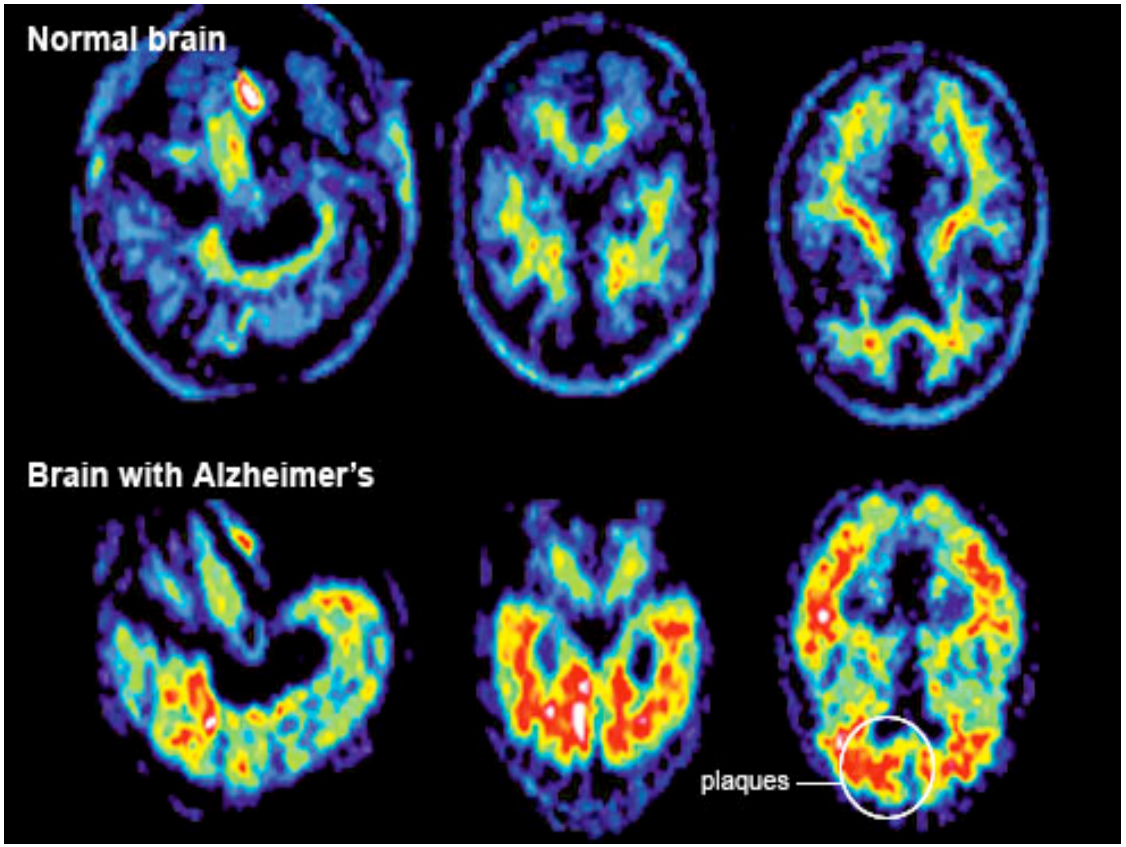
유전자 조절 네트워크와
바이오마커 정보를
클라우드 상에 업로드



클라우드 시스템을 활용한
전문가들의 연구

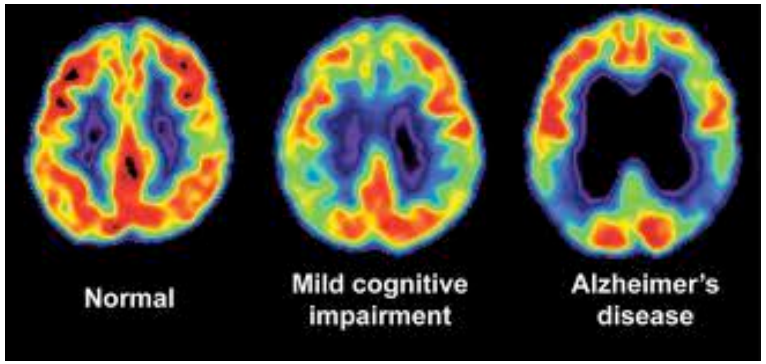


Imaging + Genetics + Epigenetics



Unfortunately,
Altzheimer.

Why?



http://medimoon.com/2012/11/amyloid-imaging-helps-in-evaluating-possible-alzheimer-disease/
http://coloradodementia.org/alzheimers-disease-in-colorado/

Proteins, Enyzmes, Motifs

- Gene Cluster Profile Vectors: a method to infer functionally related gene sets by grouping proximity-based gene clusters. *BMC Genomics* 2011
- GeneclusterViz: a tool for conserved gene cluster visualization, exploration and analysis, *Bioinformatics*, 2012
- Building Interacting Partner Predictors Using Co-varying Residue Pairs Between Histidine Kinase and Response Regulator Pairs of 48 Bacterial Two-Component Systems, *Proteins*, 2011
- Sequence-Based Enzyme Catalytic Domain Prediction Using Clustering and Aggregated Mutual Information Content. *Journal of Bioinformatics and Computational Biology*. Vol. 9, No. 5 (2011) 1–15
- And motif discovery algorithms …

Acknowledgements

- **Seoul National University**
  Je-keun Rhee
  Kwangsoo Kim
  JeaHyun An
  Keoyri Jo
  Sungmin Rhee
  Jinwoo Park
  HongRyul Ahn
  Heejoon Chae
  Inuk Jung
  Minsoo Kim
  Seyoon Ko
  Yoonjeong Cha (MIT)

- **IU Bloomington**
  Kenneth Nephew
  Heejoon Chae

OSU ICBP center
Pearlly Yan
Tim H-M. Huang

- **Funding Agencies**

Korea
NRF-2012M3C4A7033341
NRF-2011-0031935
Next-Generation BioGreen 21 Program (No.PJ009037022012)

US
NCI U54 CA11300
NCI R01 CA85289

Thank you!

Questions, please!