Hsu100 Conference, Beijing University, July 5-7 2010

Liquid association and multivariate analysis of complex data

Ker-Chau Li

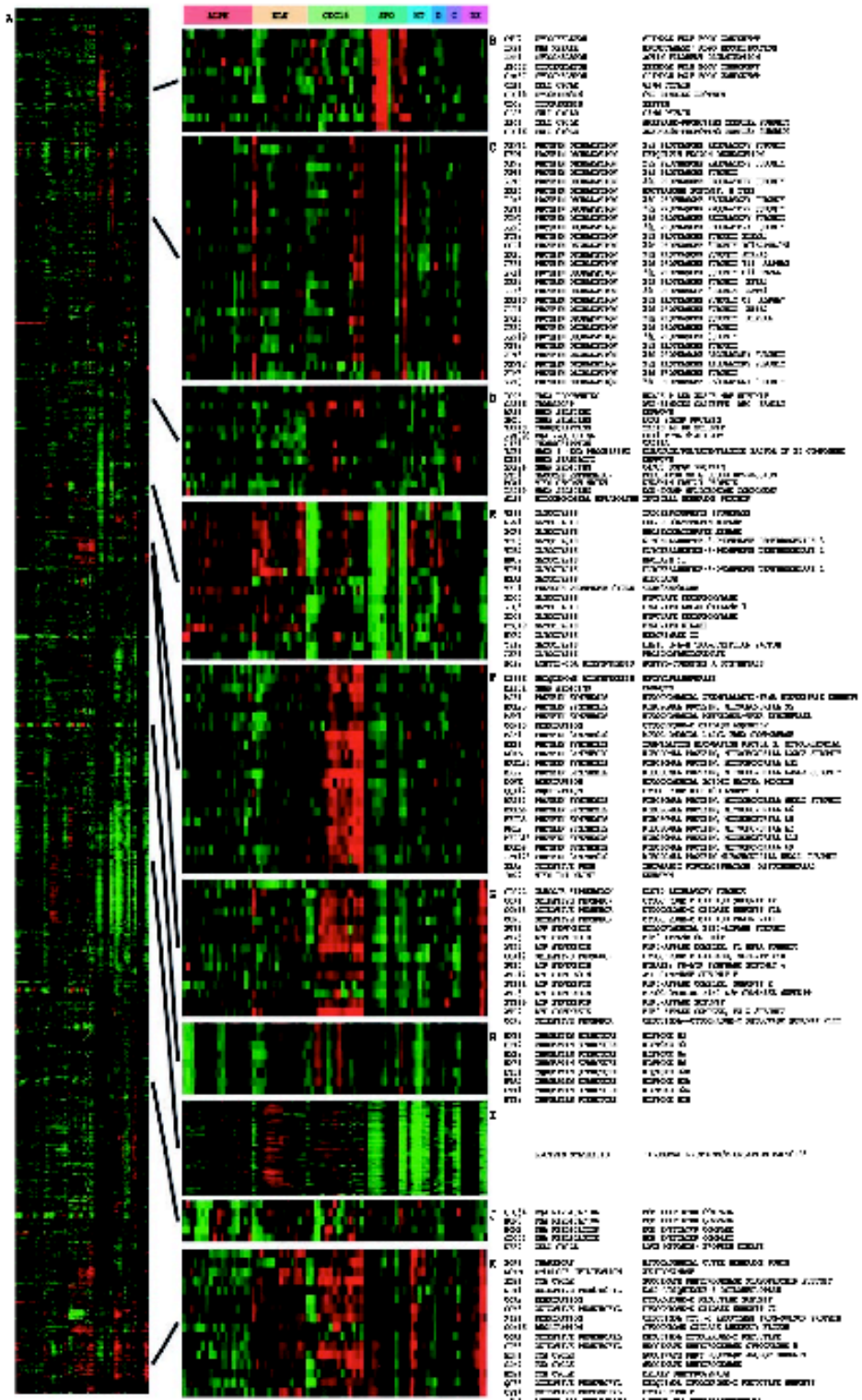
Institute of Statistical Science, Academia Sinica

Department of Statistics, UCLA

-
- High dimensionality and nonlinearity are two important issues in large scale, complex genomic data. In microarray gene expression studies, various clustering, dimension reduction, and variable selection procedures have been successfully applied. Gene clustering is often accomplished by measuring the similarity between gene expression profiles. The commonly-used similarity measure, correlation coefficient, is inadequate for discovering nonlinearity. Recently we have developed an on-line bio-computing system based on the new statistical concept of liquid association (LA). LA measures the change in correlation between a pair of variables as mediated by a third variable. Statistical properties of LA and the related extensions under the broader context of multivariate analysis will be discussed.

gene-expression data

	cond1	cond2	……..	condp
gene1	x11	x12	……..	x1p
gene2	x21	x22	……..	x2p
		…	…	
gene n				



Why clustering make sense biologically?

The rationale is

Genes with high degree of expression similarity are likely to be functionally related.

may form structural complex,

may participate in common pathways.

may be co-regulated by common upstream regulatory elements.

Simply put,

Profile similarity implies functional association

However, the converse is not true

The expression profiles of majority of functionally associated genes are indeed uncorrelated

- Microarray is too noisy
- Biology is complex

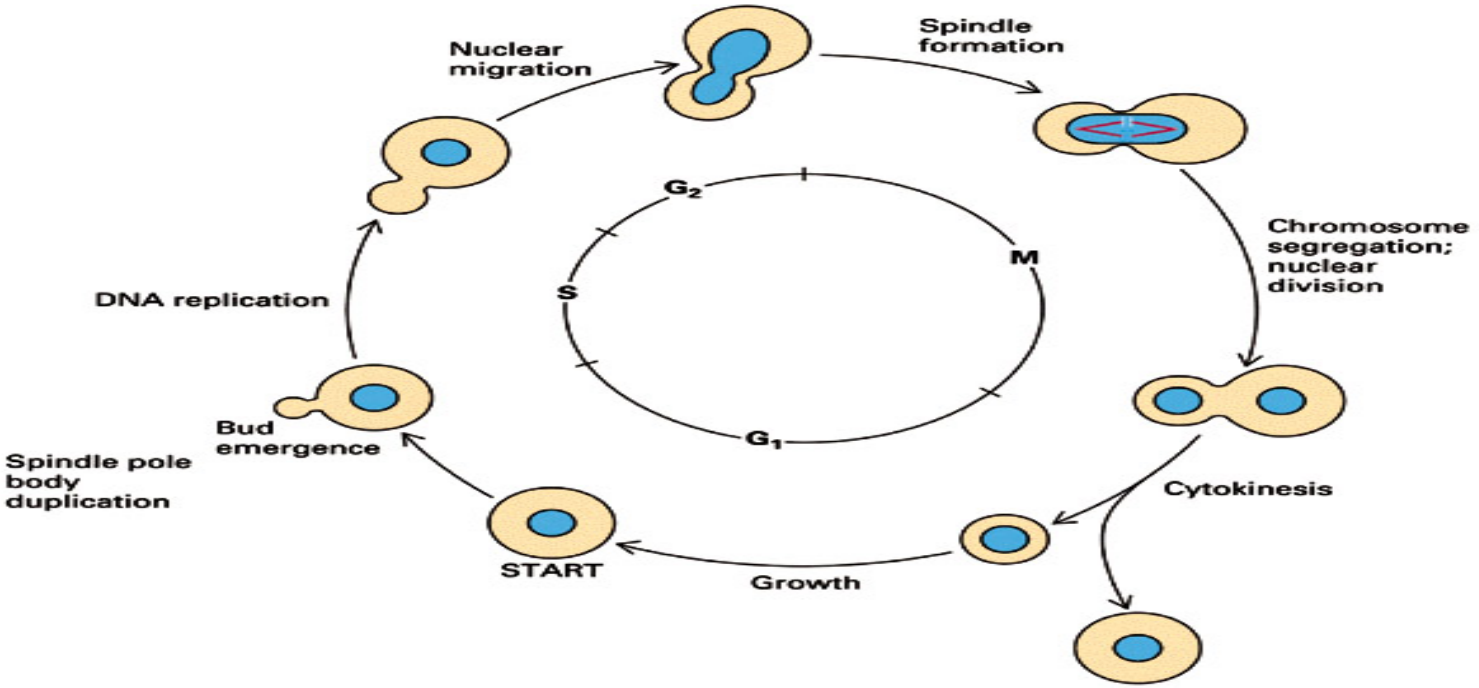
Patterns of Coexpression for Protein Complexes by Size in Saccharomyces Cerevisiae

NAR 2008, Ching-Ti Liu, Shinsheng Yuan, Ker-Chau Li

- Many successful functional studies by gene expression profiling in the literature have led to the perception that profile similarity is likely to imply functional association. But how true is the converse of the above statement? Do functionally associated genes tend to be co-regulated at the transcription level? In this paper, we focused on a set of well-validated yeast protein complexes provided by Munich Information Center for Protein Sequences (MIPS). Using four well-known large-scale microarray expression datasets, we computed the correlations between genes from the same complex. We then analyzed the relationship between the distribution of correlations and the complex size (the number of genes in a protein complex). We found that except for a few large protein complexes such as mitochondrial ribosomal and cytoplasmic ribosomal proteins, the correlations are on the average not much higher than that from a pair of randomly selected genes. The global impact of large complexes on the expression of other genes in the genome is also studied. Our result also showed that the expression of over 85% of the genes are affected by six large complexes: the cytoplasmic ribosomal complex, mitochondrial ribosomal complex, proteasome complex, F0/F1 ATP synthase (complex V) (size 18), rRNA splicing (size 24), and H+- transporting ATPase, vacular (size 15).

Yeast Cell Cycle

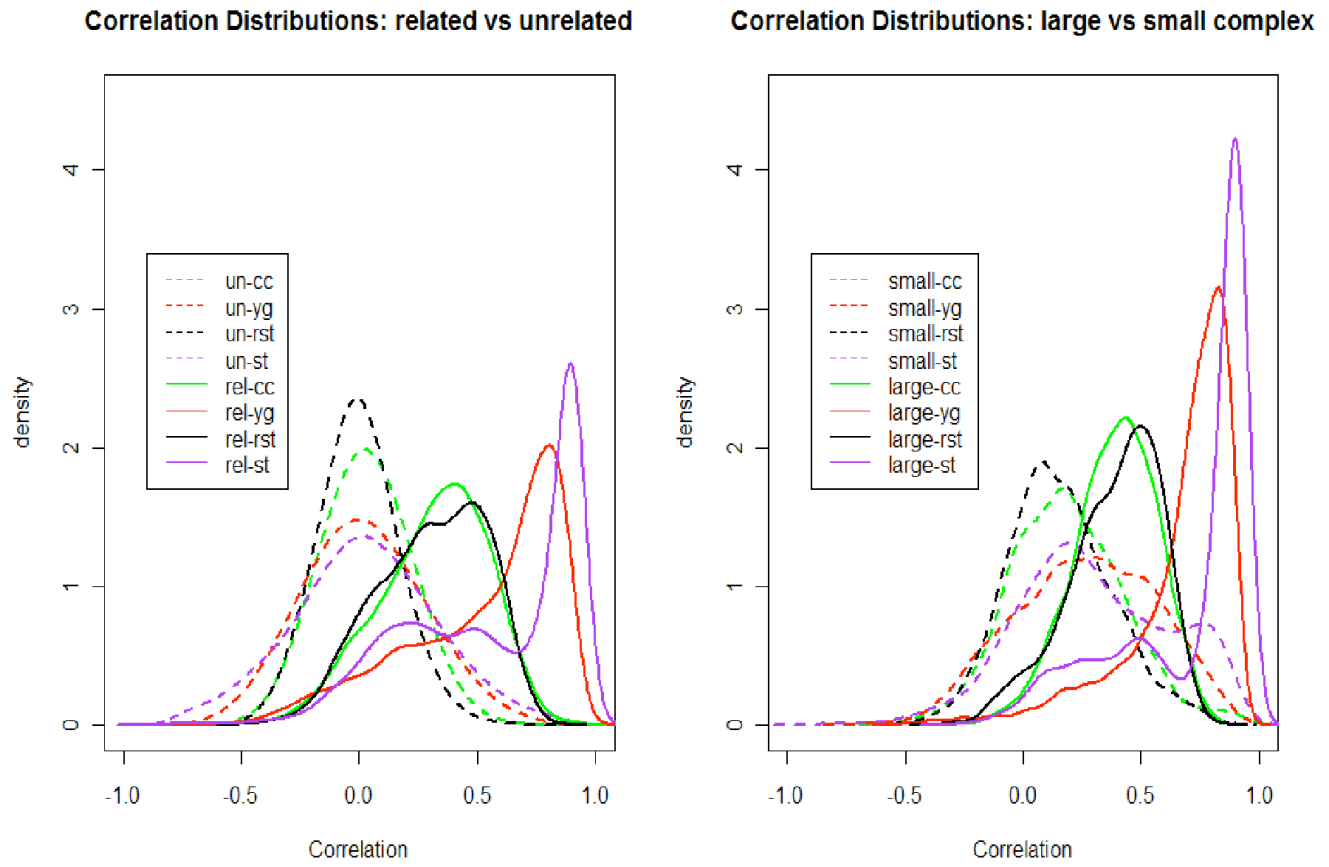
(adapted from Molecular Cell Biology, Darnell et al)



gene-expression data

	cond1	cond2	……..	condp
gene1	x11	x12	……..	x1p
gene2	x21	x22	……..	x2p
		…	…	
gene n				

Four large scale datasets



- Figure 1. Comparison of correlation distributions for protein pairs with respect to functional association (shown in left panel) and complex size (shown in right panel). The terms “cc”, “yg”, “rst” and “st1” represent four different data sets: cellcycle, segregation genetics, rosetta and stress data, respectively. Protein complex pairs are abbreviated as “rel” and unrelated pairs are abbreviated as “unrel”.

Why no correlation?

- Protein rarely works alone
- Protein has multiple functions
- Different biological processes or pathways have to be synchronized
- Competing use of finite resources : metabolites, hormones,
- Protein modification: Phosphorylation, proteolysis, shuttle, …

  Transcription factors serving both as activators and repressors

The **thyroid hormone receptor** differs functionally from glucocorticoid receptor in two important respects :

it binds to its DNA response elements **in the absence of hormone**, and **the bound protein represses transcription** rather than activating it.

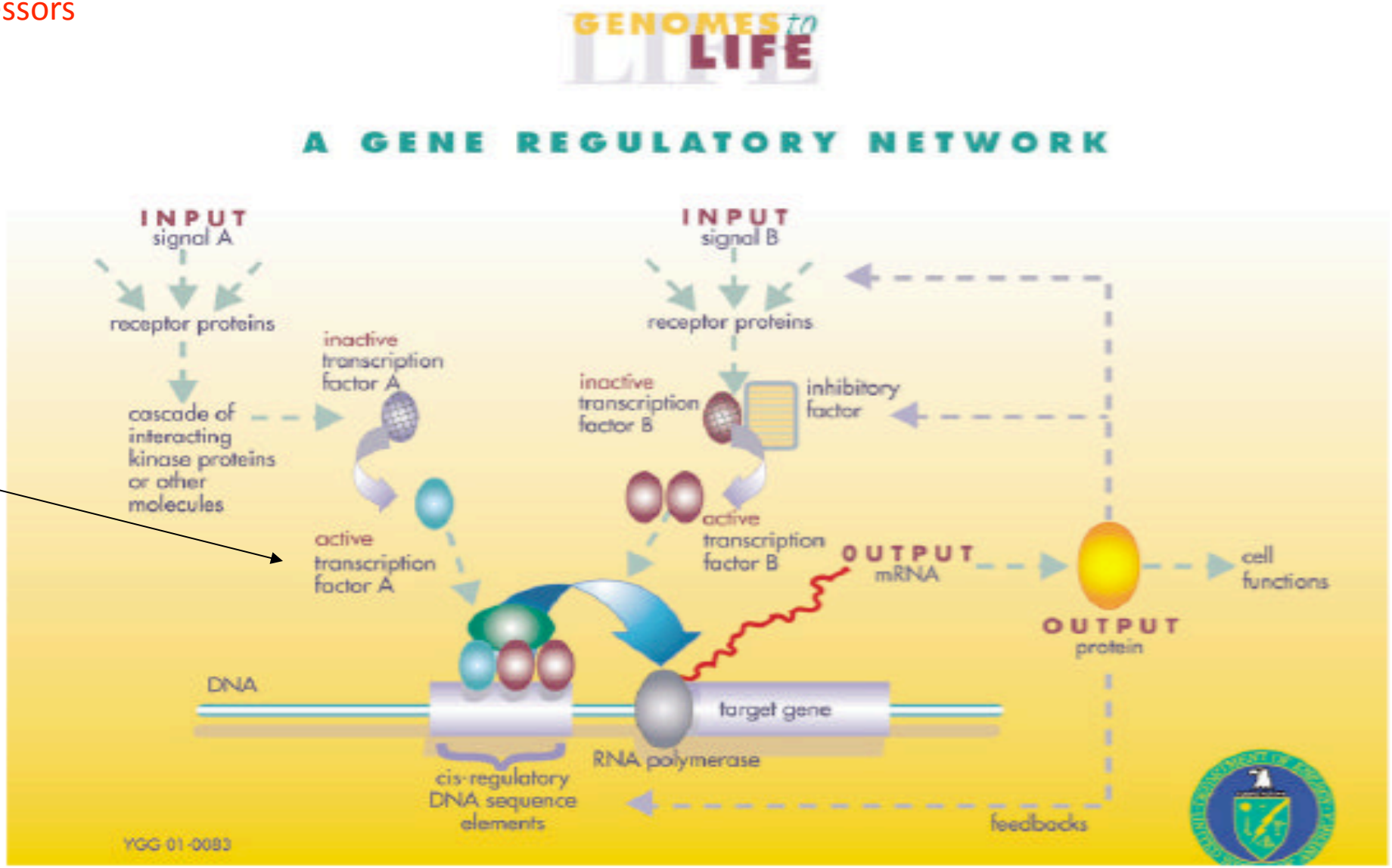
When thyroid **hormone binds to the thyroid hormone receptor**, **the receptor is converted** from a repressor **to** an **activator**.

Gene A = gene **produces THR**

Gene B= gene **regulated by THR**

THR alone represses B
THR+ HM activates B

Transcription factors: proteins that bind to DNA Activator; repressors



Expression levels of A and B can be either **positively correlated** or **negatively correlated**, depending on thyroid **hormone level**.

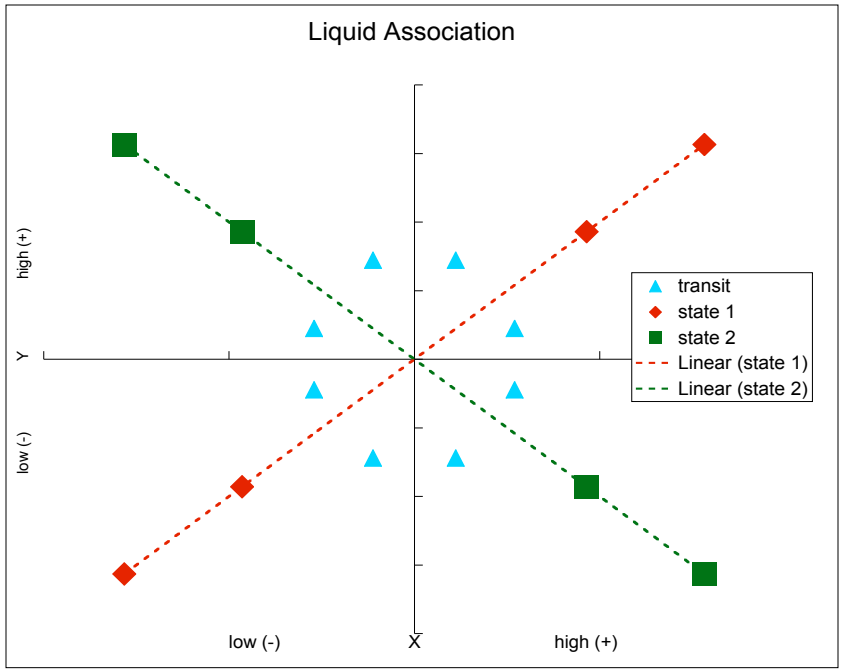
THR alone represses B
THR+ HM activates B

If during an experiment, hormone level **fluctuated** as organisms try to accomplish different tasks and if we cannot tell what tasks are, then .....

Of course, the book is not talking about yeast there. However,
Pairwise similarity is not enough!

Liquid Association (LA)

- LA is a generalized notion of association for describing certain kind of ternary relationship between variables in a system. (Li 2002 PNAS)



- Green points represent four conditions for cellular state 1.
- Red points represent four conditions for cellular state 2.
- Blue points represent the transit state between cellular states 1 and 2.
- (X,Y) forms a LA.

Profiles of genes X and Y are displayed in the above scatter plot.

Important! Correlation between X and Y is 0

Statistical theory for LA

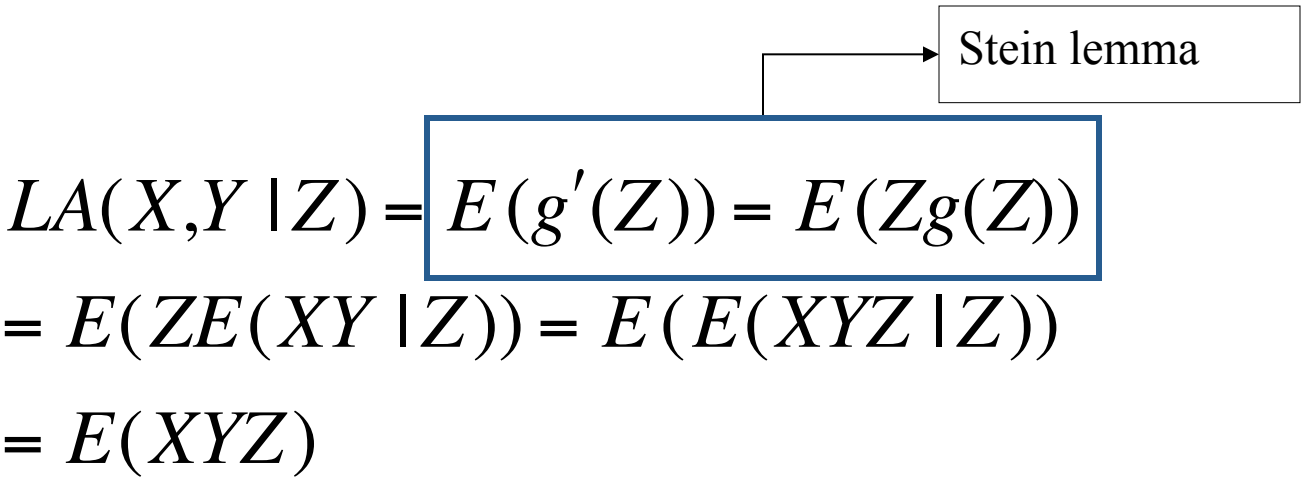
- X, Y, Z random variables with mean 0 and variance 1
- Corr(X,Y)=E(XY)=E(E(XY|Z))=Eg(Z)
- g(z) an ideal summary of association pattern between X and Y when Z =z
- g′(z)=derivative of g(z)
- Definition. The LA of X and Y with respect to Z is LA(X,Y|Z)= Eg′(Z)

Statistical theory-LA

- **Theorem**. If Z is standard normal, then LA(X,Y|Z)=E(XYZ)
- Proof. By Stein's Lemma : $Eg'(Z)=Eg(Z)Z$
- $=E(E(XY|Z)Z)=E(XYZ)$
- Additional math. properties:
- bounded by third moment
- =0, if jointly normal
- transformation

Stein Lemma

- To compute E(g'(Z)) is not easy. With help from mathematical statistics theory, the LA(X,Y|Z) can be simplified as E(XYZ) when Z follows normal distribution.



Normality ?

- Convert each gene expression profile by taking normal score transformation
- LA(X,Y|Z) = average of triplet product of three gene profiles:

$(x_1y_1z_1 + x_2y_2z_2 + \ldots) / n$

-
-

8th place negative

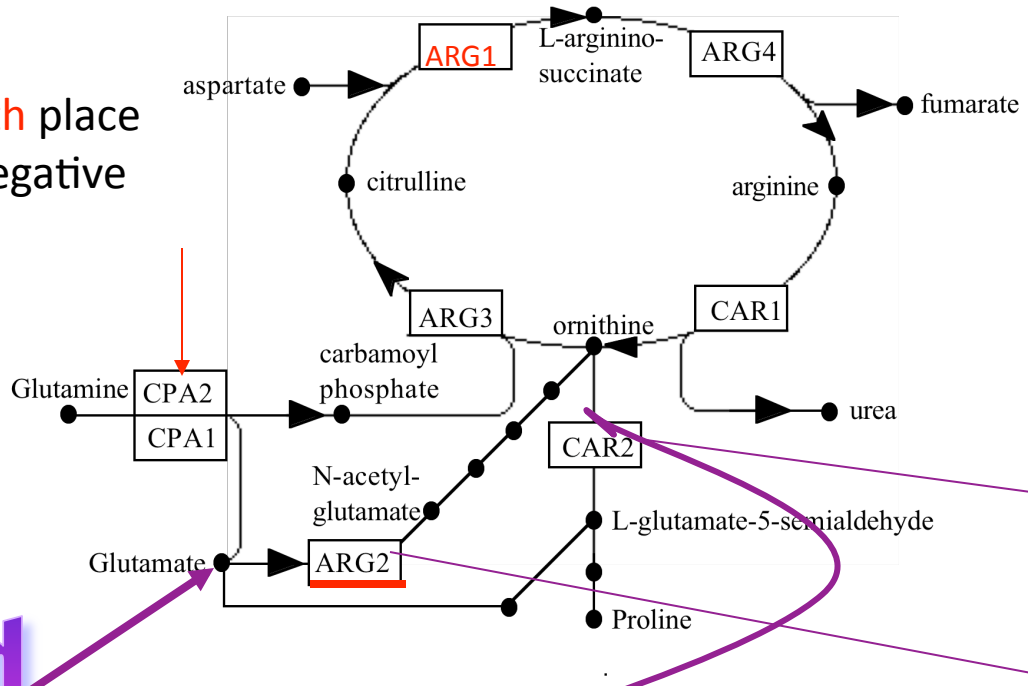


Figure 2. The four genes in the urea cycle are coded by ARG3, ARG1, ARG4, and CAR1 in S. Cerevisiae.
ARG2 enocodes acetyl-glutamate synthase, which catalyzes the first step of ornithine biosynthesis. CPA1 and CPA2 enocode small and large units of carbamoylphosphate synthetase. CAR2 encodes ornithine aminotransferase. This chart is adapted from KEGG.

Head

Backdoor

Y

X

Compute LA(X,Y|Z) for all Z

Rank and find leading genes

Adapted from KEGG

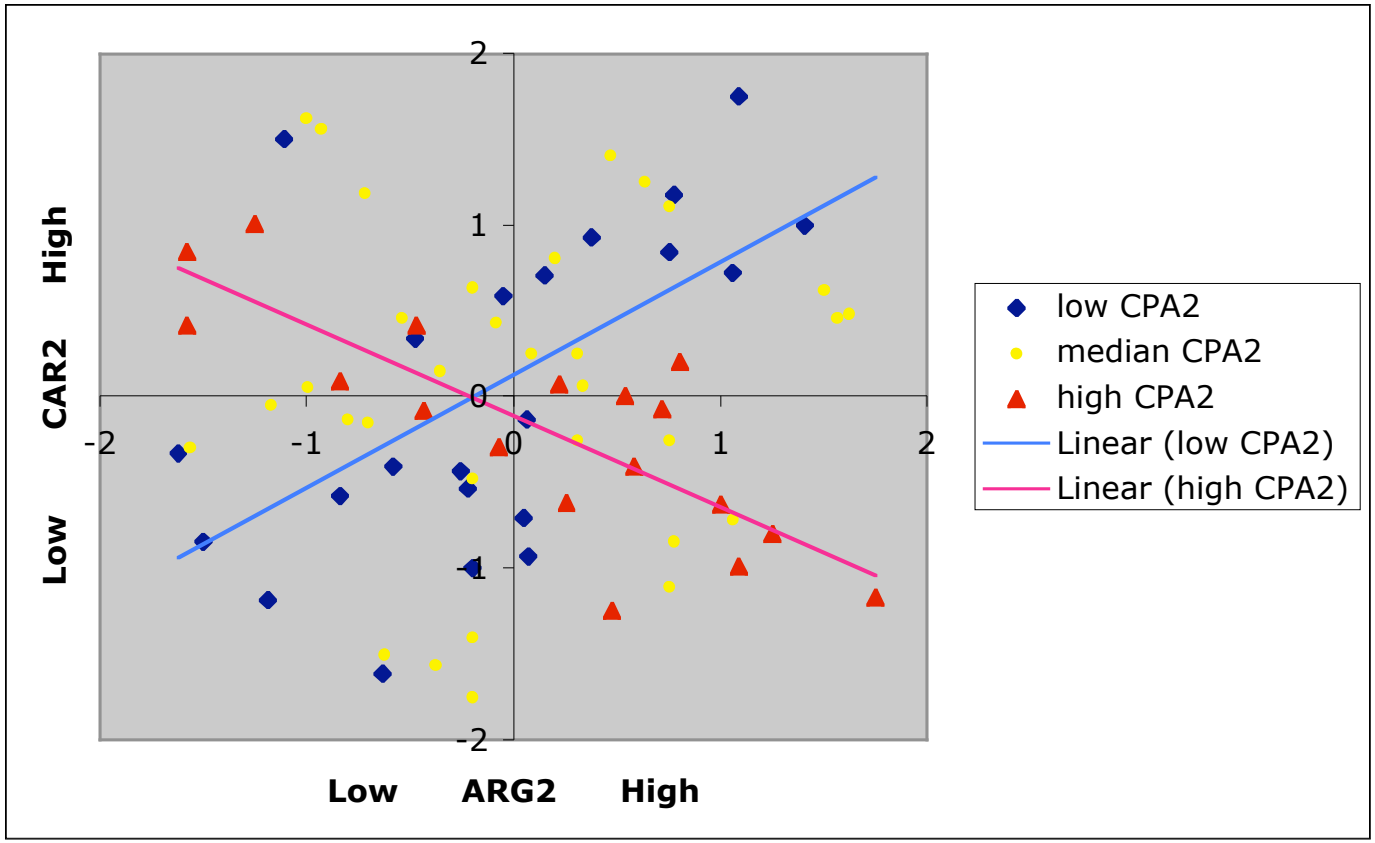
Figure 4. Urea cycle/ argininbiosynthesis pathway. ARG2 encodes acetyl-glutamate synthase, which catalyzes the first step in synthesizing ornithine from glutamate. Ornithine and carbamoyl phosphate are the substrates of the enzyme ornithine transcarbamoylase, encoded by ARG3. Carbamoyl phosphate synthetase is encoded by CPA1 and CPA2. ARG1 encodes argininosuccinate synthetase, ARG4 encodes argininosuccinase, CAR1 encodes arginase, and CAR2 encodes ornithine aminotransferase.

Why negative LA?

high CPA2 : signal for arginine demand.
up-regulation of ARG2 concomitant with down-regulation of CAR2
prevents ornithine from leaving the urea cycle.

When the demand is relieved, CPA2 is lowered, CAR2 is up-regulated,

opening up the channel for orinthine to leave the urea cycle.



Statistical significance

- P-value can be calculated by permutation test or by large sample approximation
- Plot of liquid association is provided by two methods:

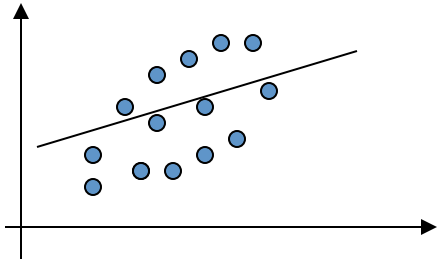
  MLE for mixture model

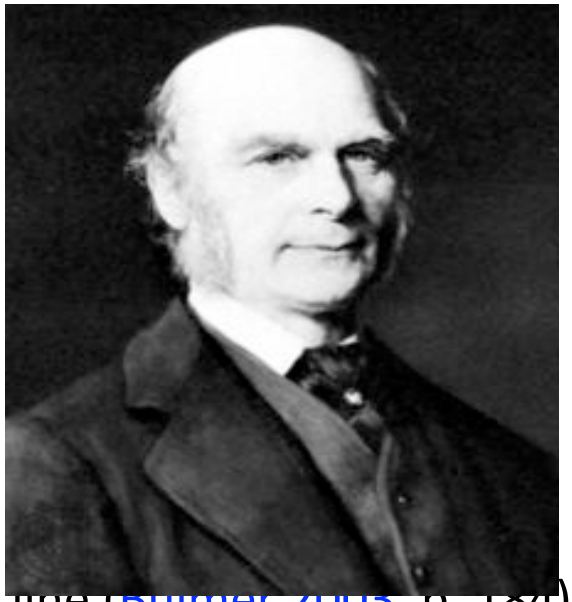
  discrete method

Historical review of correlation

- Why using mean ?
- Why using standardization?
- Outliers
- Rank correlation
- Normal score transformation (Fisher-Yates)

Regression, correlation





- Sir **Francis Galton** (1822 - 1911),
- half-cousin of Charles Darwin,
- was an English Victorian polymath, anthropologist, eugenicist, tropical explorer, geographer, inventor, meteorologist, proto-geneticist, psychometrician, and statistician.

He was knighted in 1909.

- Galton invented the use of the regression line (Bulmer 2003, p. 184), and was the first to describe and explain the common phenomenon of regression toward the mean, which he first observed in his experiments on the size of the seeds of successive generations of sweet peas.

Bivariate normal

Pearson correlation

- Corr (X,Y) = E[(X-E(X))(Y-E(Y)]/ SD(X)SD(Y)

X,Y are two random variables

Corr (often denoted ρ , or r) is between -1 and 1

r= 0 , uncorrelated

>0, positive correlation

<0, negative correlation

Intuitive illustration: larger value of X correlates with

Larger value of Y (r>0)

Larger value of X correlates with smaller value of Y (r<0)

Choice of average value : E(X), E(Y) { why not using median ??? Galton did so}

Application in microarray gene expression analysis

Galton's discovery of regression

ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

[With Plates IX and X.]

This memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

DIAGRAM BASED ON TABLE I.

(all female heights are multiplied by 1·08)

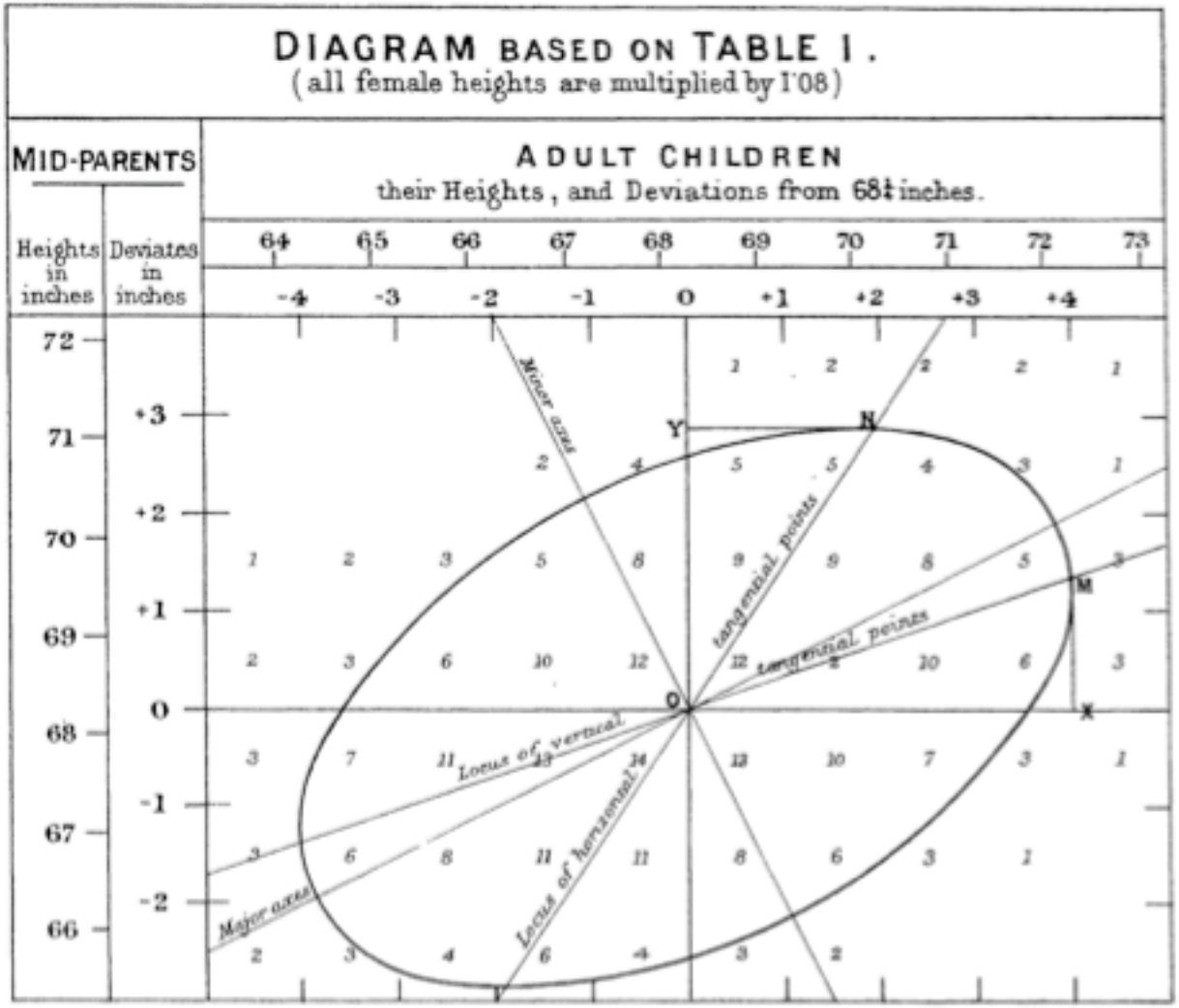
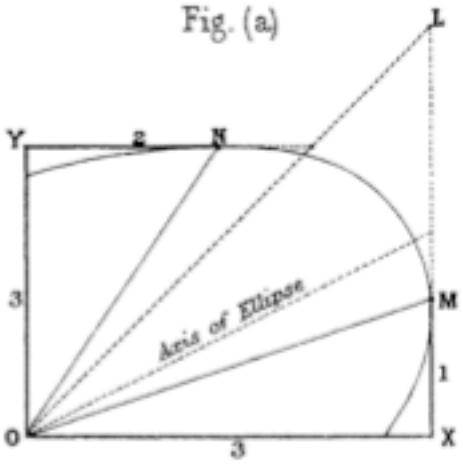


Fig. (a)



Journ. Anthropolog. Inst., Vol. XV., Pl. X.

J.P. & W.R. Emslie, lith.

Elliptically contoured bivariate distribution

- The set of points with equal frequencies of occurrence lie on concentric ellipses with common center, shape and orientation

Under the additional assumption that the marginal distributions of X and Y are normal and the , it can be proved that

the joint distribution of X,Y must follow a bivariate normal distribution, joint density function : f(x,y)= a little bit complex

Parameters of mean of X, mean of Y, SD of X, SD of Y, correlation, μ_1, μ_2,σ_1,σ_2, ρ

Non Gaussian distribution

- Only marginal normal distribution is required
- If X,Y,Z follows a joint normal distribution, then LA(X,Y|Z)=0
- LA explores non-linearity of high dimensional data
- For some applications, X,Y need not be transformed to normal distribution

Liquid Association is not Partial correlation

- X, Y, Z
- Z->Y, Z->X (Causal analysis)
- X=aZ+b+error
- Y=a'Z+b'+error'

Partial correlation of X,Y = corr (error, error')

If Z causes X and Y, then partial correlation=0

(X=Coke sale, Y=eye disease incidence rate, Z=season)

Start with a pair of correlated genes X, Y, find Z to minimize partial correlation.

This is very different from LA.

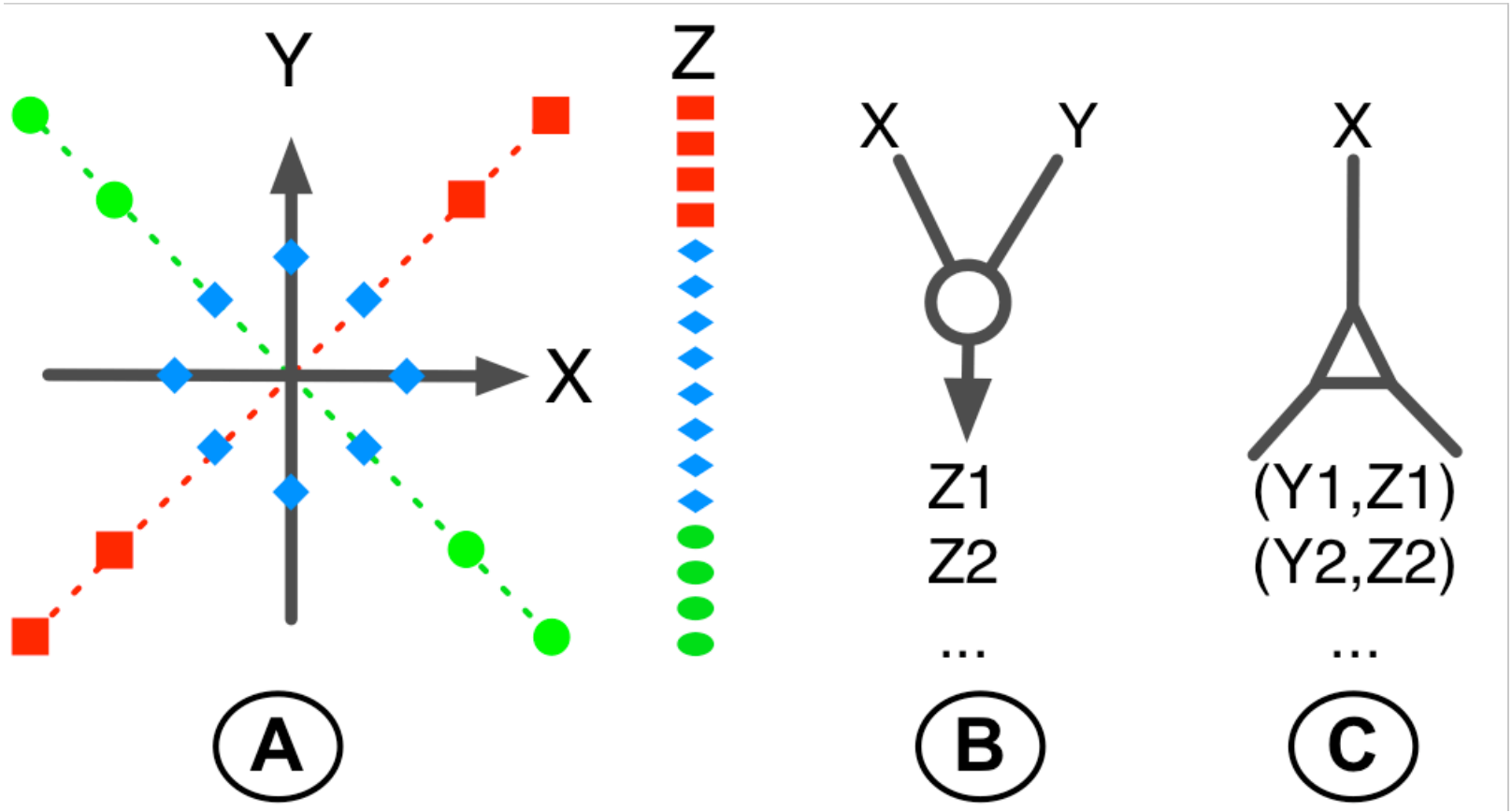
Partial correlation analysis often requires joint normal distribution of X,Y,Z

- Why not considering
- changes in Corr(X,Y|Z)?
- Harder to explain the biological meaning

(see binary case of Z)

Liquid association:

A method for exploiting lack of correlation between variables



Generalization of liquid association

- Binary Z
- Transforming X and Y
- A paradigm of using liquid association
- LAP web application

http://kiefer.stat2.sinica.edu.tw/LAP3

(Note: case sensitive)

III Correlating gene-expression with gene markers

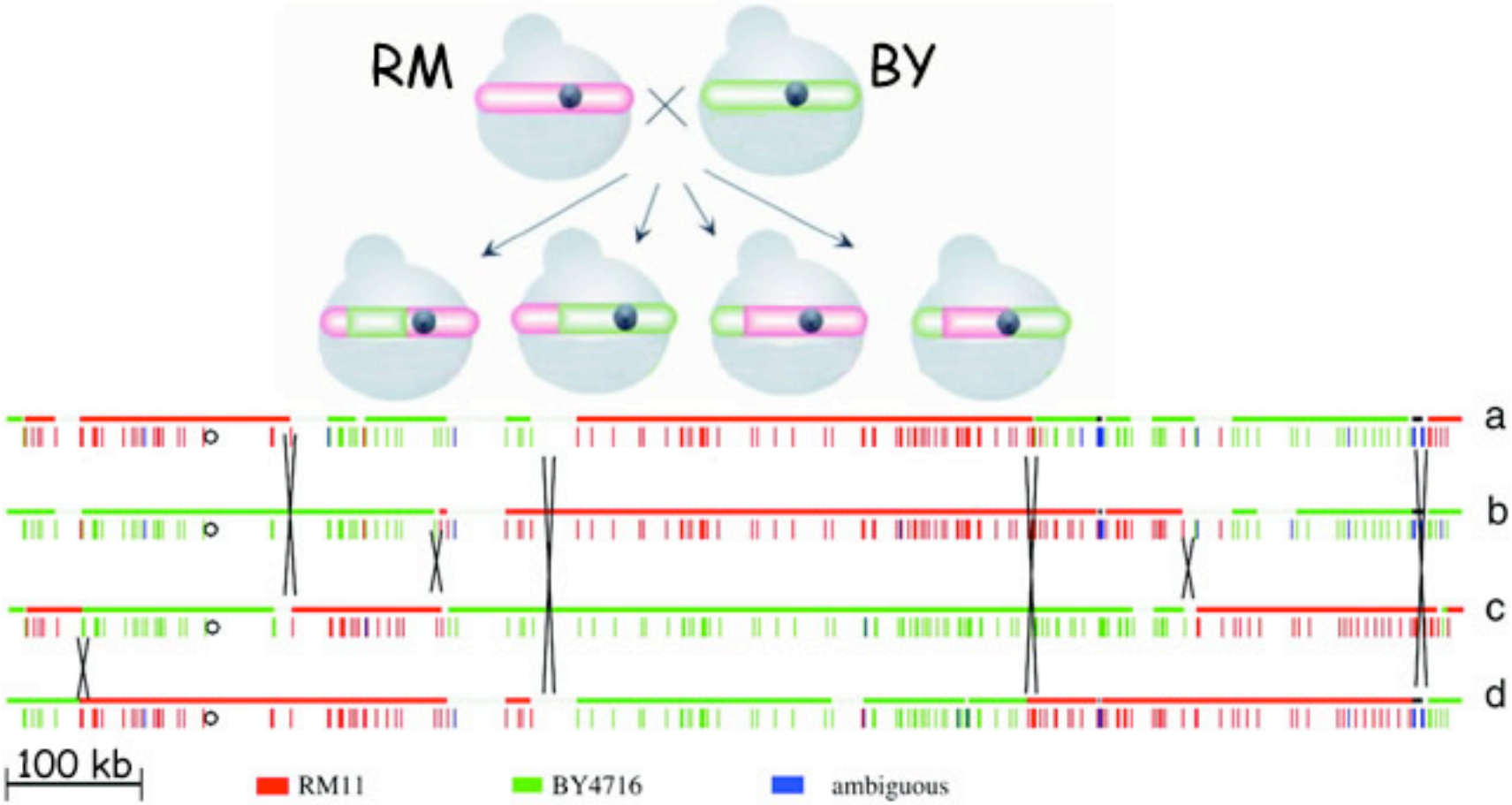
	c.line1	c.line2	.....	C.linep
gene1	x11	x12	.....	x1p
gene2	x21	x22	.....	x2p
		...	...	

Marker 1	y11	y12	.....	y1p
Marker 2	y21	y22	.....	y2p
	.....			

Brem, R., Yvert, G., Clinton, R, Kruglyak, L. (2002) Science Vol 296, 752-755. Yeast segregation

Genetic study of gene expression (eQTL)

Two parents strains: RM11 and BY 4716 are crossed to generate some offspring with diverse genetic make-ups



Chromosome XII genotypes of four segregants from one tetrad

Figure 1: A schematic diagram of eQTL studies. An eQTL dataset consists of marker genotype profiles and gene expression profiles. (a) 1D-trait mapping: conventional eQTL mapping compares one gene expression profile against one marker genotype profile to find significant differential expression. (b) 2D-trait mapping: co-expression trait mapping compares the co-expression pattern between two genes' expression profiles against one marker genotype profile.

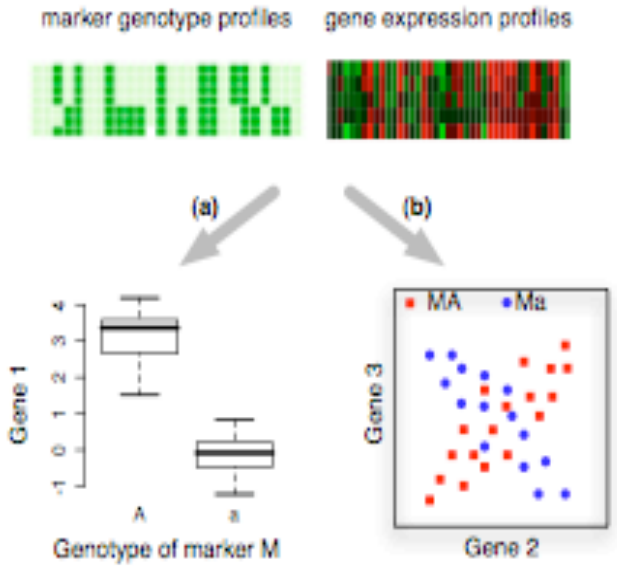


Table 1: Co-expression of leucine biosynthesis genes mediated by the eQTL of *LEU2*

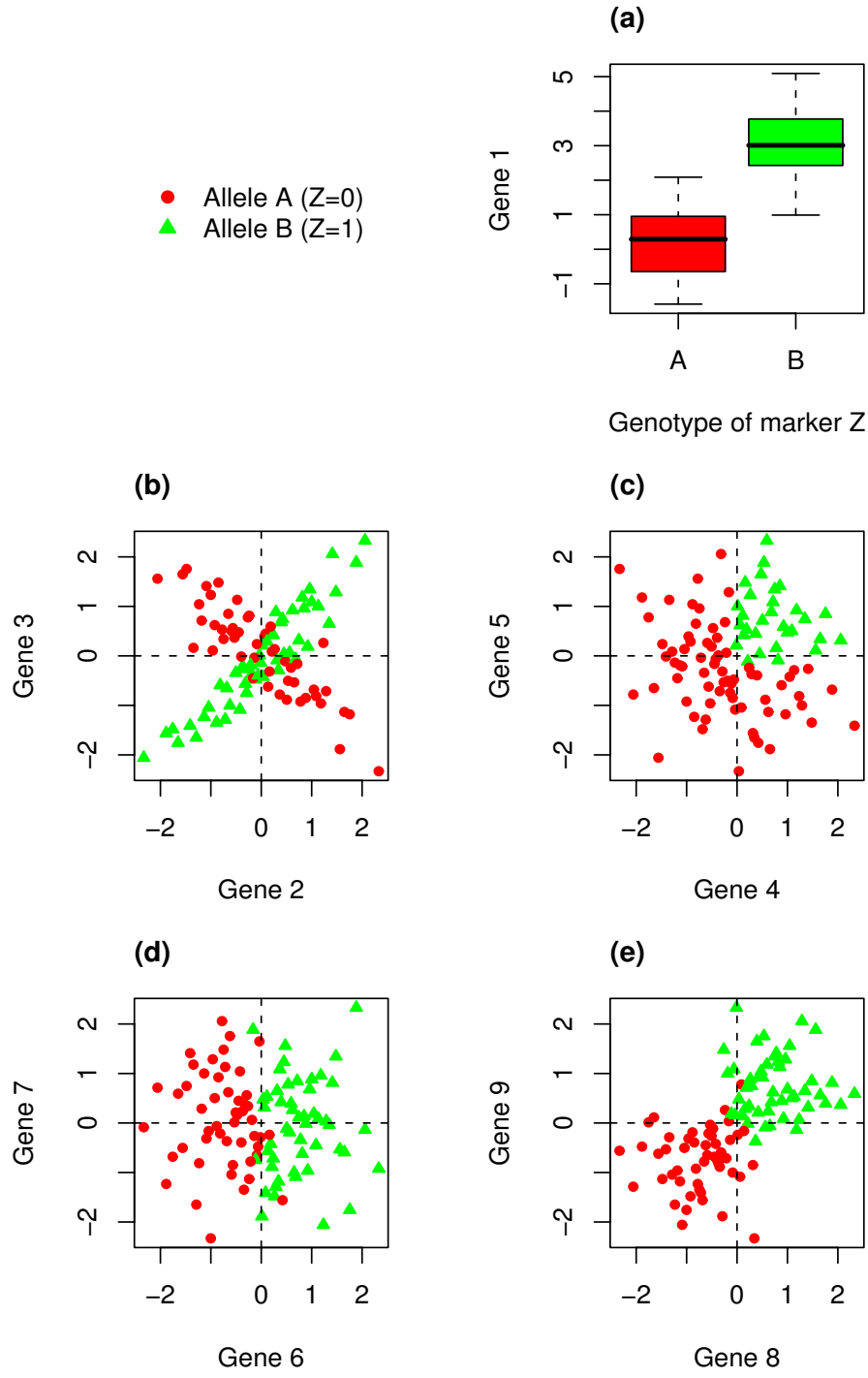


Figure 1

A schematic diagram of eQTL studies. Suppose the eQTL is captured by marker Z with allele A and B. We code Z as 0 or 1 if the inherited allele is A or B, respectively. (a) 1D-trait mapping: conventional eQTL mapping compares one gene expression profile against one marker genotype profile to find significant differential expression. (b) 2D-trait mapping: co-expression trait mapping aims at the detection of changes in the co-regulation pattern by comparing two genes' expression profiles against one marker genotype profile. (c) Two genes are co-up-regulated under allele B(Z = 1). (d) The expression of one gene has a shift in the marginal distribution, detectable by 1D mapping. (e) The expression of both genes are shifted in the marginal distributions, detectable by 1D mapping. Only (b) and (c) are detectable by our 2D-trait mapping method.

Trait-trait dynamic interaction: 2D-trait eQTL mapping for genetic variation study

- Wei Sun
- Shinsheng Yuan
- Ker-Chau Li
- (***)To whom correspondence should be addressed. E-mail: kcli@stat.ucla.edu

  kcli@stat.sinica.edu.tw

Direct difference of Correlation Cofficient

- For the protocol case as depicted by the schematic Figure 1(b), the two
- measures are equivalent because E(X|Z = 1) = E(X|Z = 0), E(Y|Z = 1) = E(Y|Z = 0), SD(X|Z = 1) = SD(X|Z = 0), and SD(Y|Z = 1) = SD(Y|Z = 0).

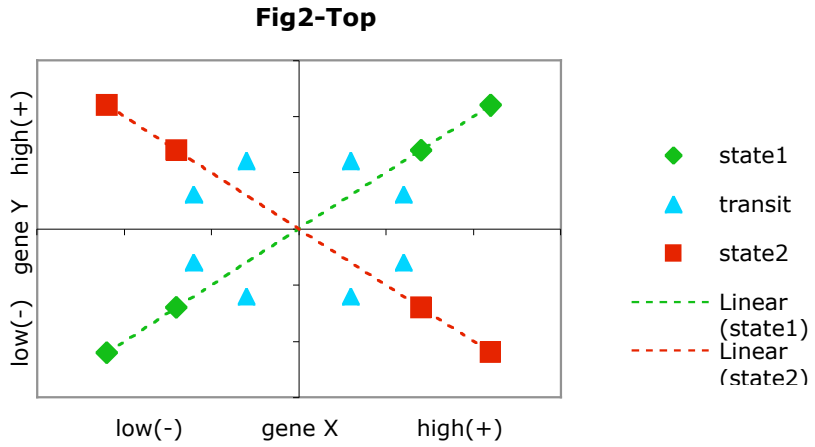
- Be aware of the different biological interpretation of what it means by co-expression/coregression of two genes.
- (i) two different baseline expressions are used in defining up-regulation
- or down-regulation of a gene
- (ii) two different scales are used in defining the strength of up-regulation or
- down-regulation of a gene.

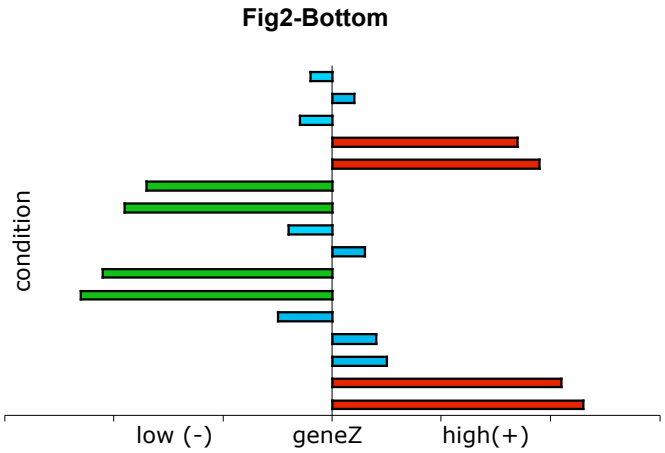
In contrast, LA method uses only one common baseline and only one common scale.

- Another difference between LA and DCC is that while LA
- can be applied to both discrete and continuous Z, it is not easy to obtain an implementable version of DCC for a continuous Z.

Projective LA (PLA)

Schematic illustration of LA





Projection-based LA for studying a group of genes

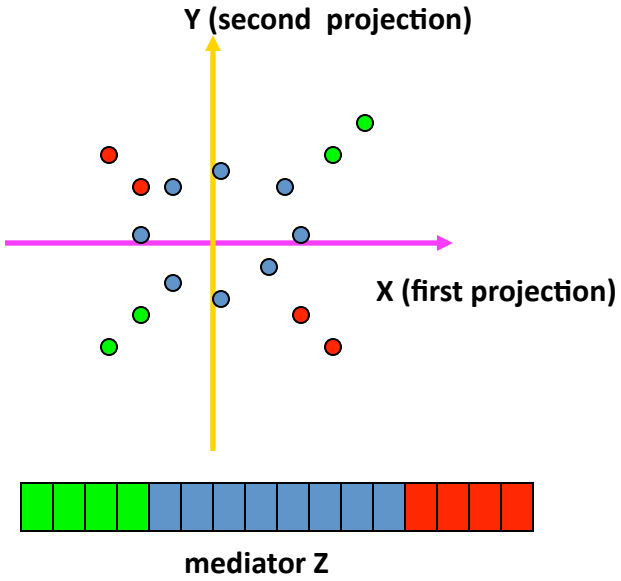


Figure 2 (a)

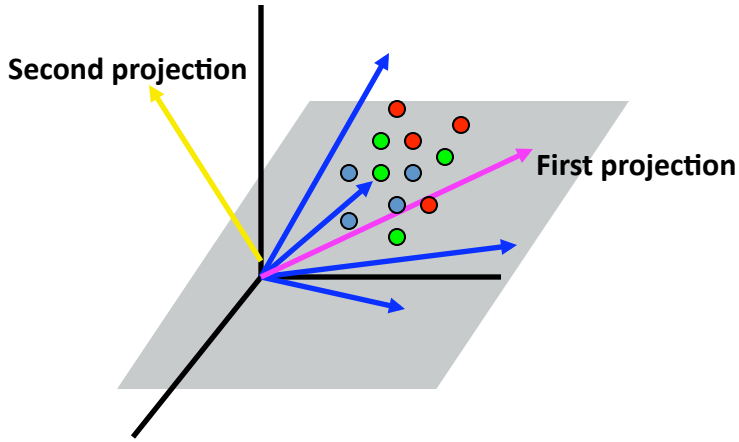


Figure 2(b).

$\boldsymbol{X}$: vector of p variables, $X_1,\ldots X_p$, each measures the expression level of one gene. one-dimensional projection: $a'\boldsymbol{X} = a_1X_1+..+a_pX_p$ with norm $||a||=1$. 2-D projection: a, b be orthogonal : $a'b=a_1b_1+..+a_pb_p=0$.

Liquid association between $a'\boldsymbol{X}$ and $b'\boldsymbol{X}$ as mediated by Z is $LA(a'\boldsymbol{X}, b'\boldsymbol{X}|Z)= E(a'\boldsymbol{X}\, b'\boldsymbol{X}\, Z)=E(a'\boldsymbol{X}\boldsymbol{X}'b\, Z)=a'E(Z\boldsymbol{X}\boldsymbol{X}')b$.

Most informative 2-D projection :

maximize $|a'E(Z\boldsymbol{X}\boldsymbol{X}')b|$ over any pair of orthogonal projection directions a, b.

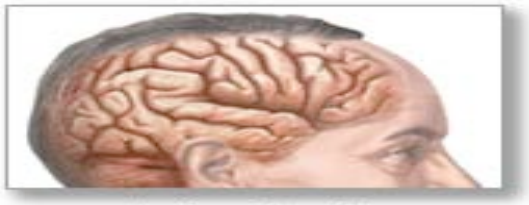
solution : eigenvalue decomposition of the matrix $E(Z\mathbf{X}\mathbf{X}')$:

$$E(Z\mathbf{X}\mathbf{X}')\, v_i = l_i\, v_i\ , \quad l_1 \geq \ldots\ldots\ldots \geq l_p$$

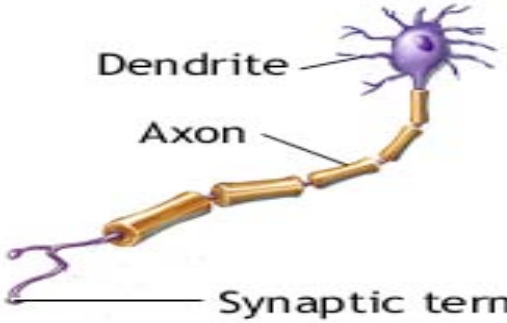
v_i are eigenvectors and l_l are eigenvalues.

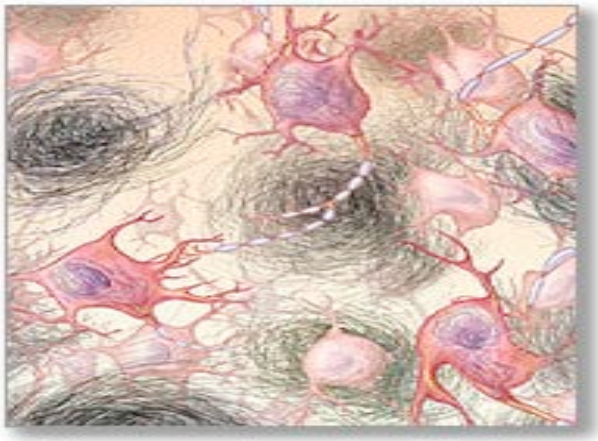
Theorem. Assume Z is normal with mean 0, SD= 1. Subject to $||a||=||b||=1$ and $a'b=0$, the maximum for the absolute value of $LA(a'X, b'X|Z)$ is $(l_1 - l_p)/2$. The optimal 2-D projection is given by $a=(v_1+v_p)/\sqrt{2}$ (or $-a$), $b= (v_1 - v_p)/\sqrt{2}$ (or $-b$).

Alzheimer's disease



Aging brain





Neurons in aging brain

A.D.A.M.

The brain tissue shows "neurofibrillary tangles" (twisted fragments of protein within nerve cells that clog up the cell), "neuritic plaques" (abnormal clusters of dead and dying nerve cells, other brain cells, and protein), and "senile plaques" (areas where products of dying nerve cells have accumulated around protein). Although these changes occur to some extent in all brains with age, there are many more of them in the brains of people with AD. The destruction of nerve cells (neurons) leads to a decrease in neurotransmitters (substances secreted by a neuron to send a message to another neuron). The correct balance of neurotransmitters is critical to the brain.

Amyloid beta peptide is the predominant component of senile plagues in brains of MD patients.

It is derived from

Amyloid-beta precusor protein (APP) by consecutive proteolytic cleavage of

Beta-secretase and

gamma-secretase

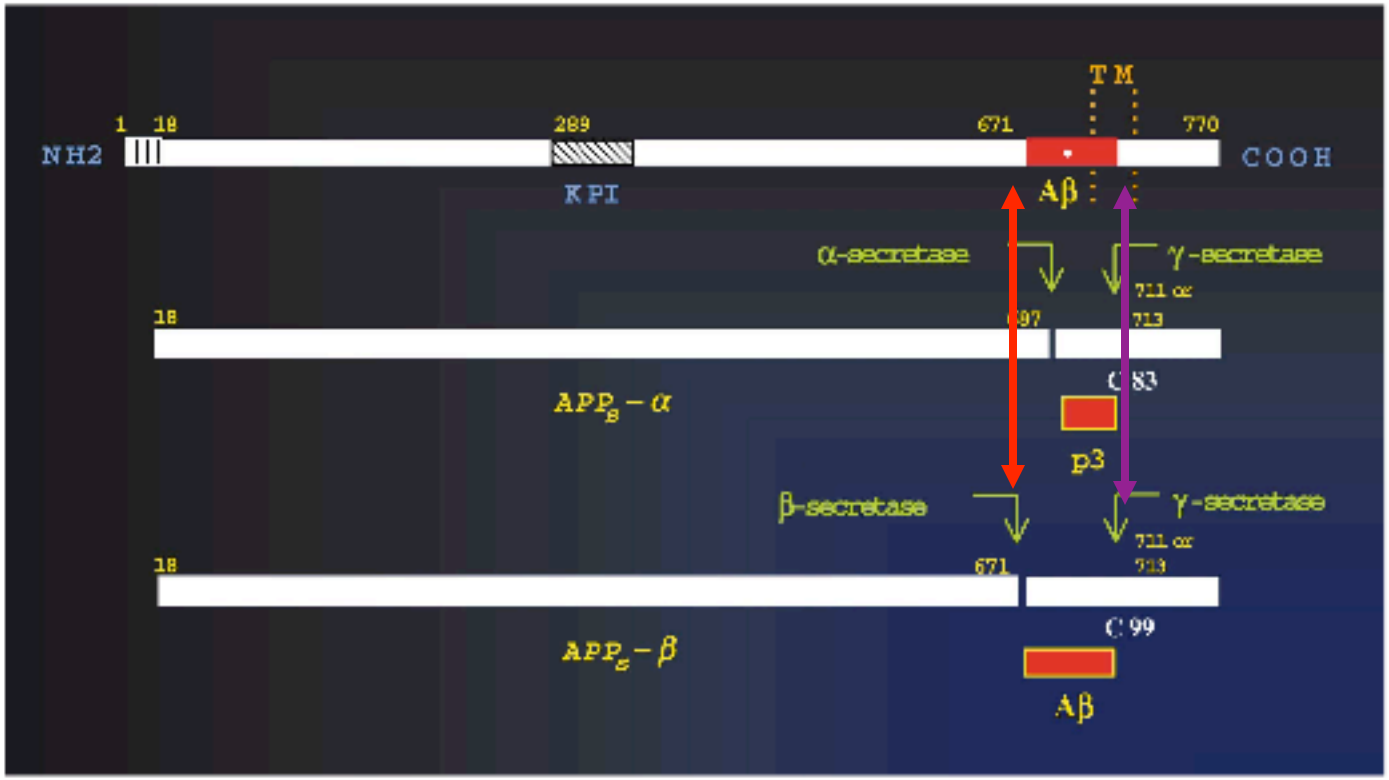


FIG. 1. Schematic diagrams of the β-amyloid precursor protein (APP) and its principal metabolic derivatives. Top diagram depicts the largest of the known APP alternate splice forms, comprising 770 amino acids. Regions of interest are indicated at their correct relative positions. A 17-residue signal peptide occurs at the NH_2 terminus (box with vertical lines). Two alternatively spliced exons of 56 and 19 amino acids are inserted at residue 289; the first contains a serine protease inhibitor domain of the Kunitz type (KPI). A single membrane-spanning domain (TM) at amino acids 700–723 is indicated by the vertical dotted lines. The amyloid ß-protein (Aβ) fragment includes 28 residues just outside the membrane plus the first 12–14 residues of the transmembrane domain. In the middle diagram, the arrow indicates the site (after residue 687; same site as the white dot in the Aβ region of APP in the upper diagram) of a constitutive proteolytic cleavage made by protease(s) designated α-secretase that enables secretion of the large, soluble ectodomain of APP (APP_s-α) into the medium and retention of the 83-residue COOH-terminal fragment in the membrane. The C83 fragment can undergo cleavage by a protease(s) called γ-secretase at residue 711 or residue 713 to release the p3 peptides. The bottom diagram depicts the alternative proteolytic cleavage after residue 671 by a protease(s) called β-secretase that results in the secretion of the slightly truncated APP_s-β molecule and the retention of a 99-residue COOH-terminal fragment. The C99 fragment can also undergo cleavage by γ-secretase to release the Aβ peptides.

tors (KPI), indicating one potential function of these longer APP isoforms. Indeed, the KPI-containing forms of APP found in human platelets serve as inhibitors of factor

modified ("matured") through the secretory pathway. Its acquisition of *N*- and *O*-linked sugars occurs rapidly after biosynthesis, and its half-life is relatively brief (~45–60

What is the physiological role of APP?

Cao X, Sudhof TC.

A transcriptionally active complex of APP with Fe65 and histone acetyltransferase Tip60.

Science. 2001 Jul 6;293(5527):115-20.

Abstract of Cao and Sudhof

Amyloid-beta precursor protein (APP), a widely expressed cell-surface protein, is cleaved in the transmembrane region by gamma-secretase. gamma-Cleavage of APP produces the extracellular amyloid beta-peptide of Alzheimer's disease and releases an intracellular tail fragment of unknown physiological function. We now demonstrate that the ***cytoplasmic tail of APP forms a multimeric complex with the nuclear adaptor protein Fe65 and the histone acetyltransferase Tip60.*** This complex potently stimulates transcription via heterologous Gal4- or LexA-DNA binding domains, suggesting that ***release of the cytoplasmic tail of APP by gamma-cleavage may function in gene expression***.

Take X=APP, Y=APBP1

- APBP1 encodes FE65
- Find BACE2 from our short list of LA score leaders.
- BACE2 encodes a beta-site APP-cleaving enzyme

Take X=APP, Y=HTATIP

HTATIP encodes Tip60

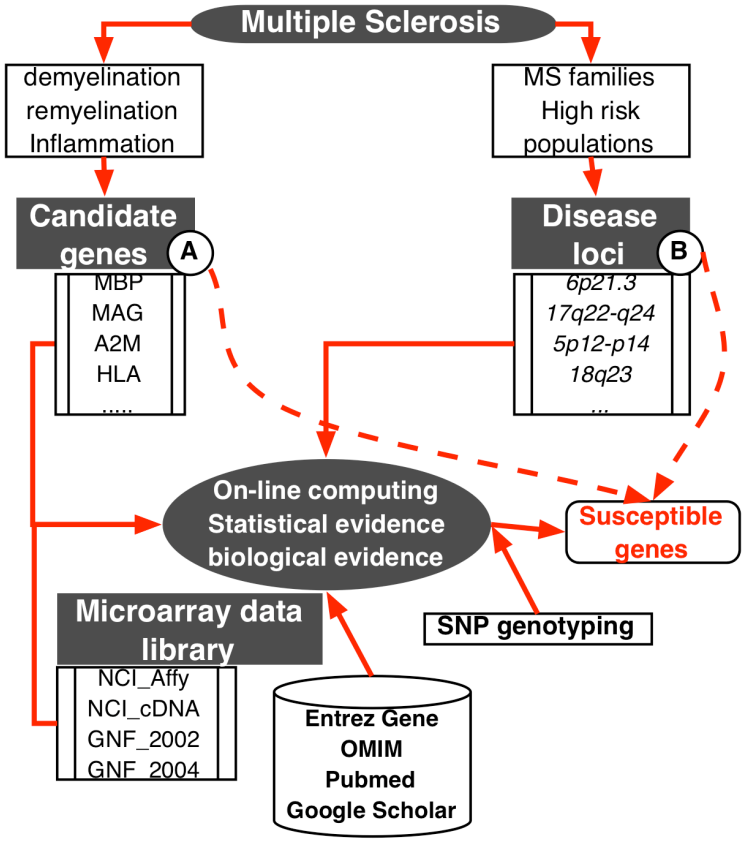
Finds PSEN1 (second place positive LA score leader)

Which encodes presenilin 1,

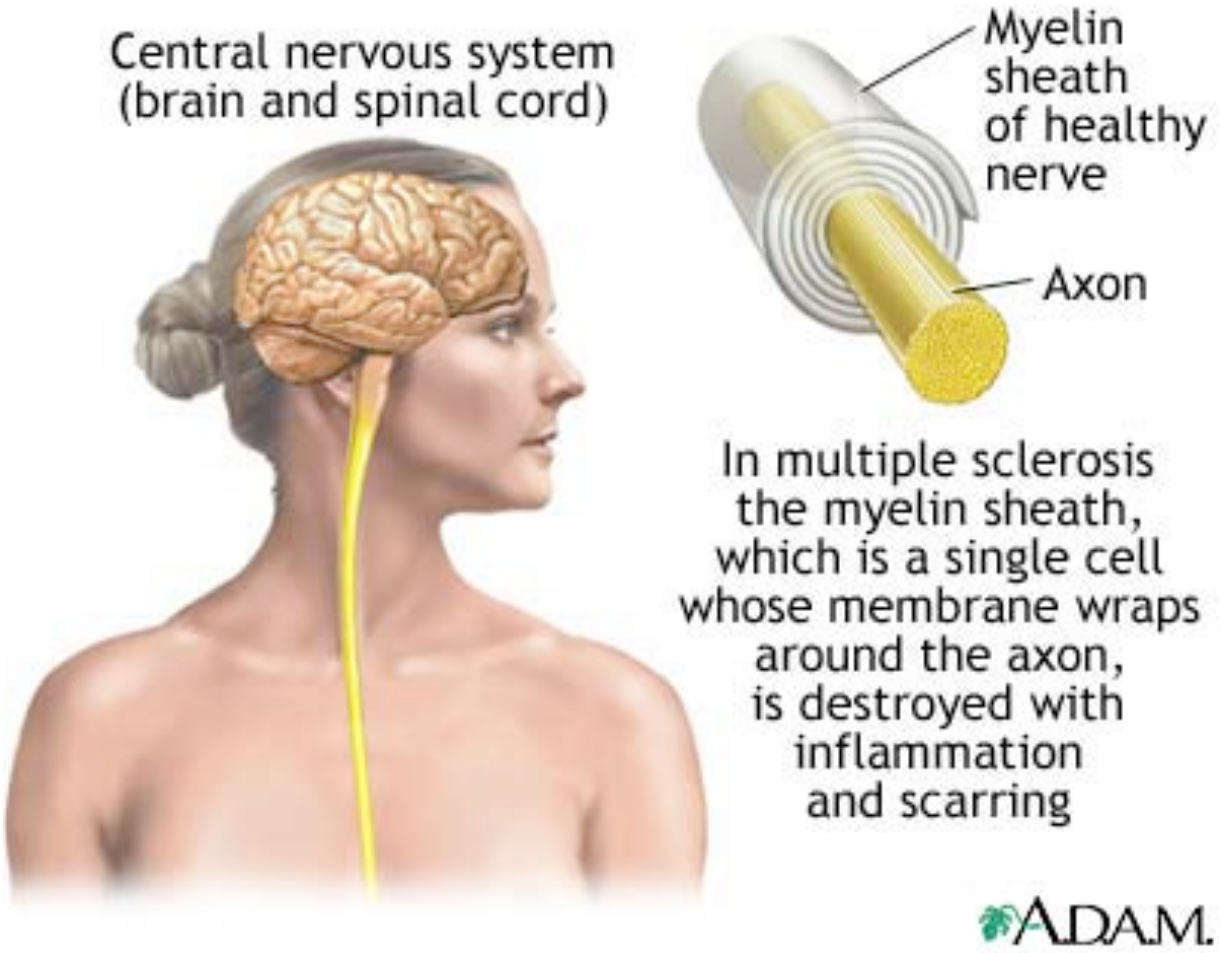
a major component of

gamma-secretase

Application: finding candidate genes for Multiple sclerosis

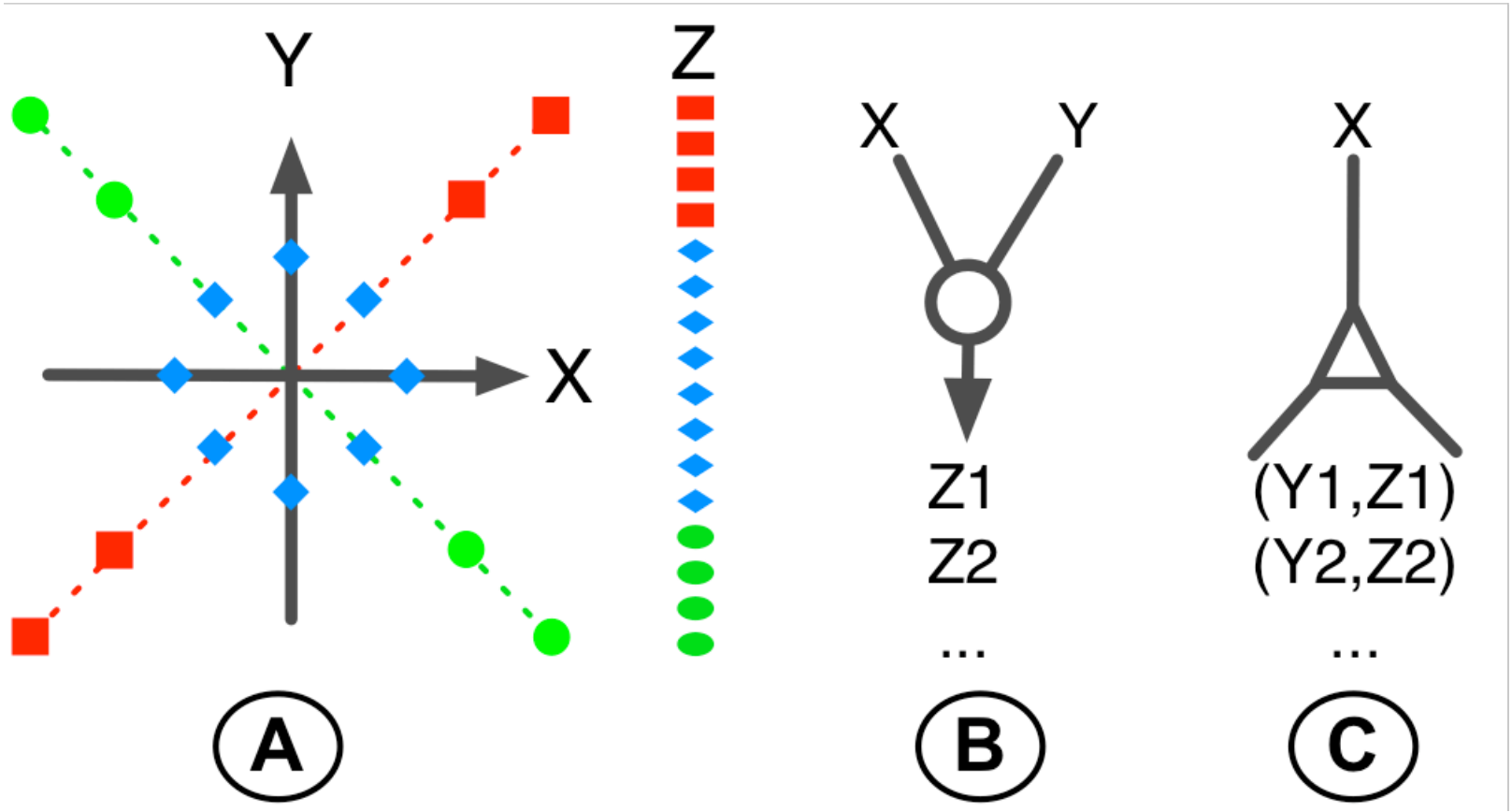


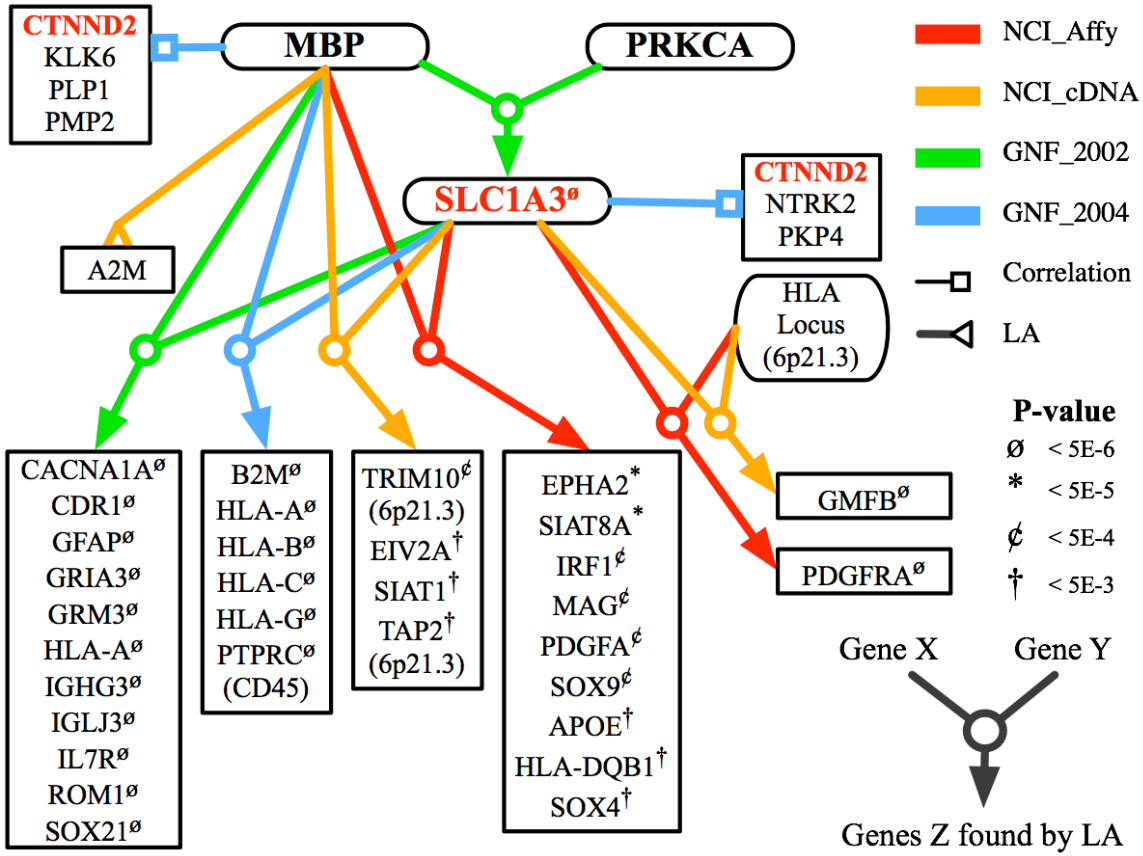
Multiple Sclerosis



Liquid association:

A method for exploiting lack of correlation between variables




CTNND2
KLK6
PLP1
PMP2
MBP
PRKCA
SLC1A3ø
CTNND2
NTRK2
PKP4
A2M
HLA
Locus
(6p21.3)
CACNA1Aø
CDR1ø
GFAPø
GRIA3ø
GRM3ø
HLA-Aø
IGHG3ø
IGLJ3ø
IL7Rø
ROM1ø
SOX21ø
B2Mø
HLA-Aø
HLA-Bø
HLA-Cø
HLA-Gø
PTPRCø
(CD45)
TRIM10¢
(6p21.3)
EIV2A†
SIAT1†
TAP2†
(6p21.3)
EPHA2*
SIAT8A*
IRF1¢
MAG¢
PDGFA¢
SOX9¢
APOE†
HLA-DQB1†
SOX4†
GMFBø
PDGFRAø
NCI_Affy
NCI_cDNA
GNF_2002
GNF_2004
Correlation
LA
P-value
ø < 5E-6
* < 5E-5
¢ < 5E-4
† < 5E-3
Gene X
Gene Y
Genes Z found by LA

glutamate-induced excitotoxicity

SLC1A3 is highly expressed in various brain regions including cerebellum, frontal cortex, basal ganglia and hippocampus. It encodes a sodium-dependent glutamate/aspartate transporter 1 (GLAST). Glutamate and aspartate are excitatory neurotransmitters that have been implicated in a number of pathologic states of the nervous system. Glutamate concentration in cerebrospinal fluid rises in acute MS patients whilst glutamate antagonist amantadine reduces MS relapse rate. In EAE, the levels of GLAST and GLT-1 (SLC1A2) are found down-regulated in spinal cord at the peak of disease symptoms and no recovery was observed after remission. We consider highly encouraging that several lines of evidence including both genetic association and gene expression association, would be consistent with the glutamate-induced excitotoxicity hypothesis of the mechanisms resulting in demyelination and axonal damage in MS.

Validation for the genetic relevance of *SLC1A3 to MS*. We set to test if there is any genetic relevance of SLC1A3 to MS. Before *SLC1A3* was brought up by the LA method, our fine mapping effort focused on a more telomeric region (between 10.3 and 17.3 Mb) of 5p, which had provided the highest two-point lod scores in Finnish MS families (22) . Guided by the LA findings, we further included five SNPs flanking the *SLC1A3* gene (Table 2) to be genotyped in our primary study set consisting of 61 MS families from the high risk region of Finland. The most 5' SNP, rs2562582, located within 2kb from the initiation of the *SLC1A3* transcript showed initial evidence for association to MS (p=0.005) in the TDT analysis, suggesting a possible functional role of this variant in the transcriptional regulation of this gene. Moreover, as shown in Table 2, stratification of the Finnish MS families according to the strongest associating SNP on the HLA region23, rs2239802, strengthened the association between the *SLC1A3* SNP and MS (p=0.0002, TDT). ***Thus, based on LA, and supported by association analyses in an MS study sample, the presence of SLC1A3 serves to connect all four major MS loci identified in Finnish families, elucidating a potential functional relationship between genetically identified genes and loci.***

The NEW ENGLAND JOURNAL of MEDICINE

Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study

The International Multiple Sclerosis Genetics Consortium*

Abstract

Background

Multiple sclerosis has a clinically significant heritable component. We conducted a genomewide association study to identify alleles associated with the risk of multiple sclerosis.

Methods

We used DNA microarray technology to identify common DNA sequence variants in 931 family trios (consisting of an affected child and both parents) and tested them for association. For replication, we genotyped another 609 family trios, 2322 case subjects, and 789 control subjects and used genotyping data from two external control data sets. A joint analysis of data from 12,360 subjects was performed to estimate the overall significance and effect size of associations between alleles and the risk of multiple sclerosis.

Results

A transmission disequilibrium test of 334,923 single-nucleotide polymorphisms (SNPs) in 931 family trios revealed 49 SNPs having an association with multiple sclerosis ($P<1\times10^{-4}$); of these SNPs, 38 were selected for the second-stage analysis. A comparison between the 931 case subjects from the family trios and 2431 control subjects identified an additional nonoverlapping 32 SNPs ($P<0.001$). An additional 40 SNPs with less stringent P values (<0.01) were also selected, for a total of 110 SNPs for the second-stage analysis. Of these SNPs, two within the interleukin-2 receptor α gene (*IL2RA*) were strongly associated with multiple sclerosis ($P=2.96\times10^{-8}$), as were a nonsynonymous SNP in the interleukin-7 receptor α gene (*IL7RA*) ($P=2.94\times10^{-7}$) and multiple SNPs in the HLA-DRA locus ($P=8.94\times10^{-81}$).

Conclusions

Alleles of *IL2RA* and *IL7RA* and those in the HLA locus are identified as heritable risk factors for multiple sclerosis.

The writing group (David A. Hafler, M.D., Alastair Compston, F.Med.Sci., Ph.D., Stephen Sawcer, M.B., Ch.B., Ph.D., Eric Lander, Ph.D., Mark J. Daly, Ph.D., Philip L. De Jager, M.D., Ph.D., Paul I.W. de Bakker, Ph.D., Stacey B. Gabriel, Ph.D., Daniel B. Mirel, Ph.D., Adrian J. Ivinson, Ph.D., Margaret A. Pericak-Vance, Ph.D., Simon G. Gregory, Ph.D., John D. Rioux, Ph.D., Jacob L. McCauley, Ph.D., Jonathan L. Haines, Ph.D., Lisa F. Barcellos, Ph.D., Bruce Cree, M.D., Ph.D., Jorge R. Oksenberg, Ph.D., and Stephen L. Hauser, M.D.) assume responsibility for the overall content and integrity of the article.

*The affiliations of the writing group and other members of the International Multiple Sclerosis Genetics Consortium are listed in the Appendix.

This article (10.1056/NEJMoa073493) was published at www.nejm.org on July 29, 2007.

N Engl J Med 2007;357.

nature genetics

Interleukin 7 receptor a chain (*IL7R*) shows allelic and functional association with multiple sclerosis

Simon G Gregory[1,9], Silke Schmidt[1,9], Puneet Seth[2], Jorge R Oksenberg[3], John Hart[1], Angela Prokop[1], Stacy J Caillier[3], Maria Ban[4], An Goris[5], Lisa F Barcellos[6], Robin Lincoln[3], Jacob L McCauley[7], Stephen J Sawcer[4], D A S Compston[4], Benedicte Dubois[5], Stephen L Hauser[3], Mariano A Garcia-Blanco[2], Margaret A Pericak-Vance[8] & Jonathan L Haines[7], for the Multiple Sclerosis Genetics Group

Multiple sclerosis is a demyelinating neurodegenerative disease with a strong genetic component. Previous genetic risk studies have failed to identify consistently linked regions or genes outside of the major histocompatibility complex on chromosome 6p. We describe allelic association of a polymorphism in the gene encoding the interleukin 7 receptor *a* chain (IL7R) as a significant risk factor for multiple sclerosis in four independent family-based or case-control data sets (overall P • 2.9 ✖ 10^{-7}). Further, the likely causal SNP, rs6897932, located within the alternatively spliced exon 6 of IL7R, has a functional effect on gene expression. The SNP influences the amount of soluble and membrane-bound isoforms of the protein by putatively disrupting an exonic splicing silencer.

Multiple sclerosis is the prototypical human demyelinating disease, which requires multiple sources and types of positive evidence to and numerous epidemiological, adoption and twin studies have implicate a candidate gene, can be used to integrate both statistical provided evidence for a strong underlying genetic liability[1].The and functional data. disease is most common in young adults, with more than 90% of Using genomic convergence[3], we identified 28 genes that were affected individuals diagnosed before the age of 55, and fewer than 5% differentially expressed in at least two of nine previous expression diagnosed before the age of 14.[2] Females are two to three times more studies (Supplementary Table 1 online). We focused on three genes frequently affected than males[2], and the disease course can vary (interleukin 7 receptor alpha chain (*IL7R*) [MIM: 146661], matrix substantially, with some affected individuals suffering only minor metalloproteinase 19 (*MMP19*) [MIM: 601807] and chemokine (C-C disability several decades after their initial diagnosis, and others motif) ligand 2 (*CCL2*) [MIM: 158105]) that had a previously reaching wheelchair dependency shortly after disease onset. The published or inferred functional role in multiple sclerosis and that complex etiology of the disease and the currently undefined molecular were not located within the MHC. Two of the three genes localize to mechanisms of multiple sclerosis suggest that moderate contributions previous regions of genetic linkage on 17q12 (*CCL2*)[4] and 5p13.2 from multiple risk loci underlie the development and progression of (*IL7R*)[5]. We analyzed a large data set of 760 US families of European the disease. descent, including 1,055 individuals with multiple sclerosis, and

Many different approaches, including genetic linkage, candidate identified a significant association with multiple sclerosis susceptibility gene association and gene expression studies, have been used (sum-only for a nonsynonymous coding SNP (rs6897932) within a key marized in ref. 2) to identify the genetic basis of multiple sclerosis. transmembrane domain of *IL7R*. We subsequently replicated this However, genetic linkage screens have failed to identify consistent initial significant association in three independent European popular regions of linkage outside of the major histocompatibility complex tions or populations of European descent: 438 individuals with (MHC). Candidate gene studies have suggested over 100 different multiple sclerosis and 479 unrelated controls ascertained in the United associated genes, but there has not been consensus acceptance of any States, 1,338 individuals with multiple sclerosis and their parents such candidates. Similarly, gene expression studies have identified ascertained in northern Europe (the UK and Belgium) and 1,077 hundreds of differentially expressed transcripts, with little consistency individuals with multiple sclerosis and 2,725 unrelated controls also across studies. The alternative approach of genomic convergence[3], ascertained in northern Europe. We show that rs6897932 affects

[1]Center for Human Genetics, and [2]Center for RNA Biology and Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina 27710, USA. [3]Department of Neurology, University of California, San Francisco, California 94143, USA. [4]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QQ, UK. [5]Section for Experimental Neurology, Katholieke Universiteit Leuven, 3000 Leuven, Belgium. [6]School of Public Health, University of California, Berkeley, California 94720, USA. [7]Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA. [8]Miami Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, Florida 33136, USA. [9]These two authors contributed equally to this work. Correspondence should be addressed to J.L.H. (jonathan@chgr.mc.vanderbilt.edu) or M.A.P.-V. (MPericak@med.miami.edu).

nature genetics

Variation in interleukin 7 receptor a chain (*IL7R*) influences risk of multiple sclerosis

Frida Lundmark[1], Kristina Duvefelt[2], Ellen Iacobaeus[3], Ingrid Kockum[1,3], Erik Wallstro¨m[3], Mohsen Khademi[3], Annette Oturai[4], Lars P Ryder[5], Janna Saarela[6], Hanne F Harbo[7,8], Elisabeth G Celius[8], Hugh Salter[9], Tomas Olsson[3] & Jan Hillert[1]

Multiple sclerosis is a chronic, often disabling, disease of the central nervous system affecting more than 1 in 1,000 people in most western countries. The inflammatory lesions typical of multiple sclerosis show autoimmune features and depend partly on genetic factors. Of these genetic factors, only the HLA gene complex has been repeatedly confirmed to be associated with multiple sclerosis, despite considerable efforts. Polymorphisms in a number of non-HLA genes have been reported to be associated with multiple sclerosis, but so far confirmation has been difficult. Here, we report compelling evidence that polymorphisms in IL7R, which encodes the interleukin 7 receptor a chain (IL7Ra), indeed contribute to the non-HLA genetic risk in multiple sclerosis, demonstrating a role for this pathway in the pathophysiology of this disease. In addition, we report altered expression of the genes encoding IL7Ra and its ligand, IL7, in the cerebrospinal fluid compartment of individuals with multiple sclerosis.

IL7Ra (also known as CD127), encoded by *IL7R*,is a member of the type I cytokine receptor family and forms a receptor complex with the common cytokine receptor gamma chain (CD132) in which IL7 is the ligand[1]. The IL-7–IL7Ra ligand-receptor pair is crucial for proliferation and survival of T and B lymphocytes in a nonredundant fashion, as shown in human and animal models, in which genetic aberrations lead to immune deficiency syndromes[2]. *IL7R* is located on chromosome 5p13, a region occasionally suggested to be linked with multiple sclerosis[3].

We have considered *IL7R* a promising candidate gene in multiple sclerosis, and we have recently reported genetic associations with three *IL7R* SNP markers in up to 672 Swedish individuals with multiple sclerosis and 672 controls, as well as two associated haplotypes spanning these markers[4]. To confirm these associations, we assessed an independent case-control group consisting of 1,820 individuals with multiple sclerosis and 2,634 healthy controls from the Nordic countries (Denmark, Finland, Norway and Sweden), independent from the data set analyzed in ref. 4 (Supplementary Table 1 online). Of these, 91% of the affected individuals had experienced an initially relapsing-remitting course of multiple sclerosis (RRMS), whereas 9% had a primary progressive course (PPMS). In addition, we analyzed the expression of *IL7R* and *IL7* in the peripheral blood as well as in cells from the cerebrospinal fluid (CSF).

We genotyped the three previously associated SNPs, located in intron 6 (rs987106 and rs987107) and exon 8 (rs3194051). rs987106 and rs3194051 were in high linkage disequilibrium (LD) (r^2 • 0.99, |*D*¢ • 1.00|), and rs987107 was in partial LD (r^2 • 0.29, |*D*¢ • 0.99|) with rs987106 and rs3194051, as they were located in the same haplotype block. The size of the study allowed full power (100%) to detect an odds ratio (OR) of 1.3. All SNPs were genotyped using the Sequenom hME assays. The observed control genotypes conformed to Hardy-Weinberg equilibrium. All three SNPs confirmed significant association with multiple sclerosis in this nonoverlapping case-control group, with very similar ORs as in the previous study (rs987107, *P* • 0.002; rs987106, *P* • 0.001; rs3194051, *P* • 0.002). A test for heterogeneity between the data sets from Norway, Denmark, Finland and Sweden showed no evidence of stratification, thus permitting a combined analysis (Mantel-Haenszel–corrected and crude ORs are shown in Table 1). We estimated the three-marker haplotype frequencies using the EM algorithm in Haploview[5].The estimated distribution of haplotypes differed significantly between affected individuals and controls (*P* • 0.001), with two haplotypes associating with multiple sclerosis, one conferring an increased risk of disease (*P* • 0.0004) and the other conferring a decreased risk of disease (*P* • 0.003) (Supplementary Table 2 online), in accordance with previous results[4].

According to data from the HapMap CEU population, *IL7R* is located within a tight LD block containing no additional genes. To

[1]Division of Neurology, Department of Clinical Neuroscience, Karolinska Institutet at Karolinska University Hospital–Huddinge, SE-141 86 Stockholm, Sweden. [2]Clinical Research Centre, Mutation Analysis Facility, Karolinska University Hospital, SE-141 86 Huddinge, Sweden. [3]Neuroimmunology Unit, Department of Clinical Neuroscience, Karolinska Institutet at Karolinska University Hospital–Solna, SE-171 76 Stockholm, Sweden. [4]Danish Multiple Sclerosis Research Centre and [5]Department of Clinical Immunology, Rigshospitalet, Copenhagen University Hospital, DK-2100 Copenhagen, Denmark. [6]Department of Molecular Medicine, National Public Health Institute, FI-00290 Helsinki, Finland. [7]Institute of Immunology, University of Oslo, N-0027 Oslo, Norway. [8]Department of Neurology, Ullevaal University Hospital, N-

International MS whole genome association study(2007).

- Affymetrix 500K to screen common genetic variants of 931 family trios.
- Using the on-line supplementary information provided, we found two SNPs, rs4869676(chr5:36641766) and rs4869675(chr5: 36636676) with TDT p-value 0.0221 and 0.00399 respectively, are in the upstream regulatory region of the SLC1A3 gene.
- In fact, within the 1Mb region of rs486975, there are a total 206 SNPs in the Affymetrix 500K chip. No other SNPs have p-value smaller than that of rs486975.
- The next most significant SNPs in this region are rs1343692(chr5:35860930), and rs6897932(chr5:35910332; the identified MS susceptibility SNP in the IL7R axon).
- The MS marker we identified rs2562582(chr5: 36641117) is , less than 5K apart from rs4869675, but was not in the Affymetrix chip.

A little bit late

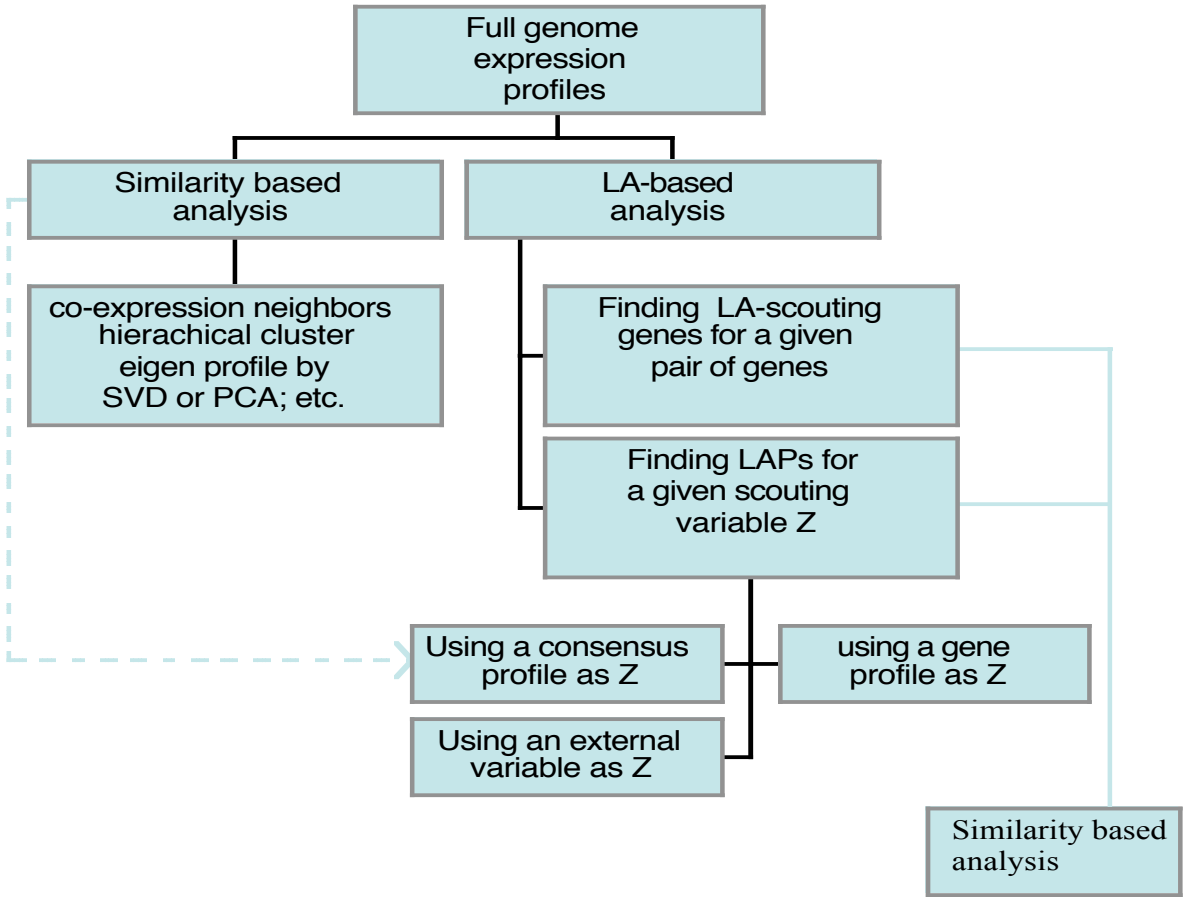
- IL7R was found long time ago before by LA !!!See the attached the e-mail/\ sent more than two years ago in 2005 !!!
- Begin forwarded message:**From:** Ker Chau Li (local) <kcli@stat.ucla.edu>**Date:** March 28, 2005 10:17:51 AM PST**To:** Robert Yuan <syuan@stat.ucla.edu>, Aarno Palotie <APalotie@mednet.ucla.edu>, Daniel Chen <pharmacogenomics@yahoo.com>, Denis Bronnikov <denis@ucla.edu>, Palotie Leena <leena.peltonen@ktl.fi>**Cc:** Ker Chau Li (local) <kcli@stat.ucla.edu>
- **Subject: IL7R**

(I thought this e-mail should have been sent out already; but it has not)I take X=SLC1A3, Y=MBP, Z= any gene, using 2002 Atlas data. Two genes are from the short list of genes with highest LA scores.IL7R interleukin 7 receptor and HLA-A

IL7R is at 5p13. Interesting coincidence??

other interesting findings include GFAP glial fibrillary acidic protein on 17q21 (Alexander disease)GRM3 (glutamate receptor, metabotropic)**CDR1 (**cerebellar degeneration-related protein 1)Ighg3 (immunoglobulin heavy constant gamma 3)Iglj3 (immunoglobulin lambda joining 3)

Figure 3. Organization chart for incorporating LA with similarity based methods. Co-expressed genes found by profile similarity analysis can be pooled together to obtain a consensus profile for LA-scouting. Likewise, the genes identified through LA system can be further analyzed for patterns of clustering. For some applications, the scouting variable may come from external sources related to the expression profiles. SVD: singular value decomposition; PCA: principal component analysis.



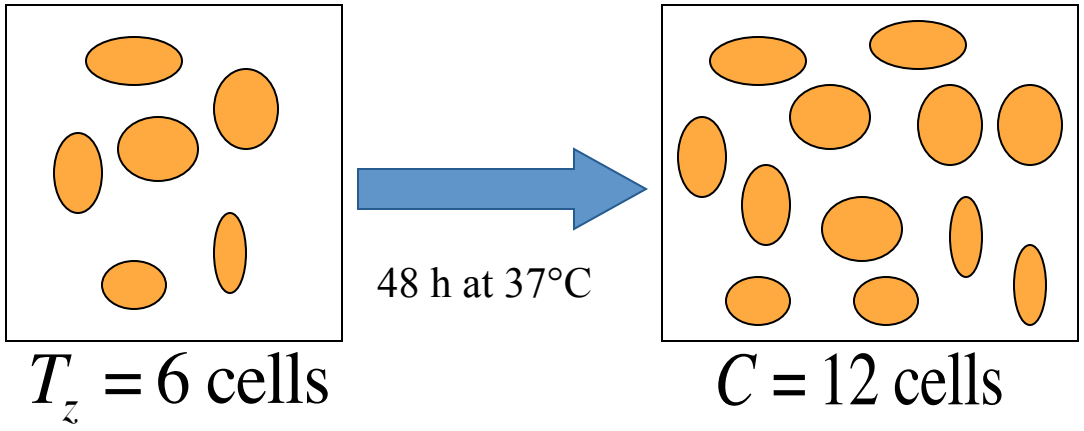
III Correlating gene-expression with drug-responsiveness

	c.line1	c.line2	.....	C.linep
gene1	x11	x12	.....	x1p
gene2	x21	x22	.....	x2p
		...	...	

---

drug1	y11	y12	y1p
drug2	y21	y22	y2p
	.....		

Experiment procedure

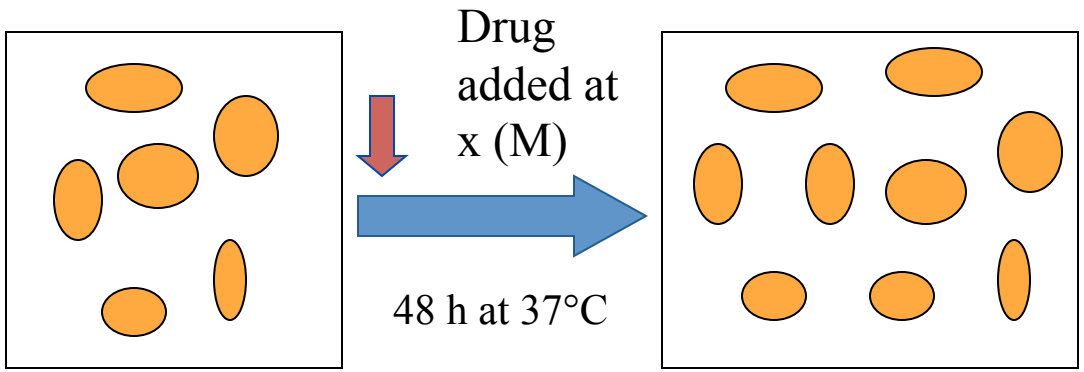


$T_z = 6$ cells

$C = 12$ cells

Rate of inhibition

$$\frac{(T_i - T_z)}{(C - T_z)} \times 100$$



$T_z = 6$ cells

$T_i = 9$ cells

E.g.:
(9-6)/(12-6)*100=50
GI_{50} is x

GI_{50} is the concentration of the drug needed to inhibit the growth of the cells up to 50%

Drug sensitivity is defined as:
-log(GI_{50})

Drug Sensitivity profile

- For each chemical compound, the tests are done for all 60 cell lines listed previously.
- For each cell line and each compound, there were multiple experiments performed to obtain the average drug concentration.

Drug Sensitivity profile (cont.)

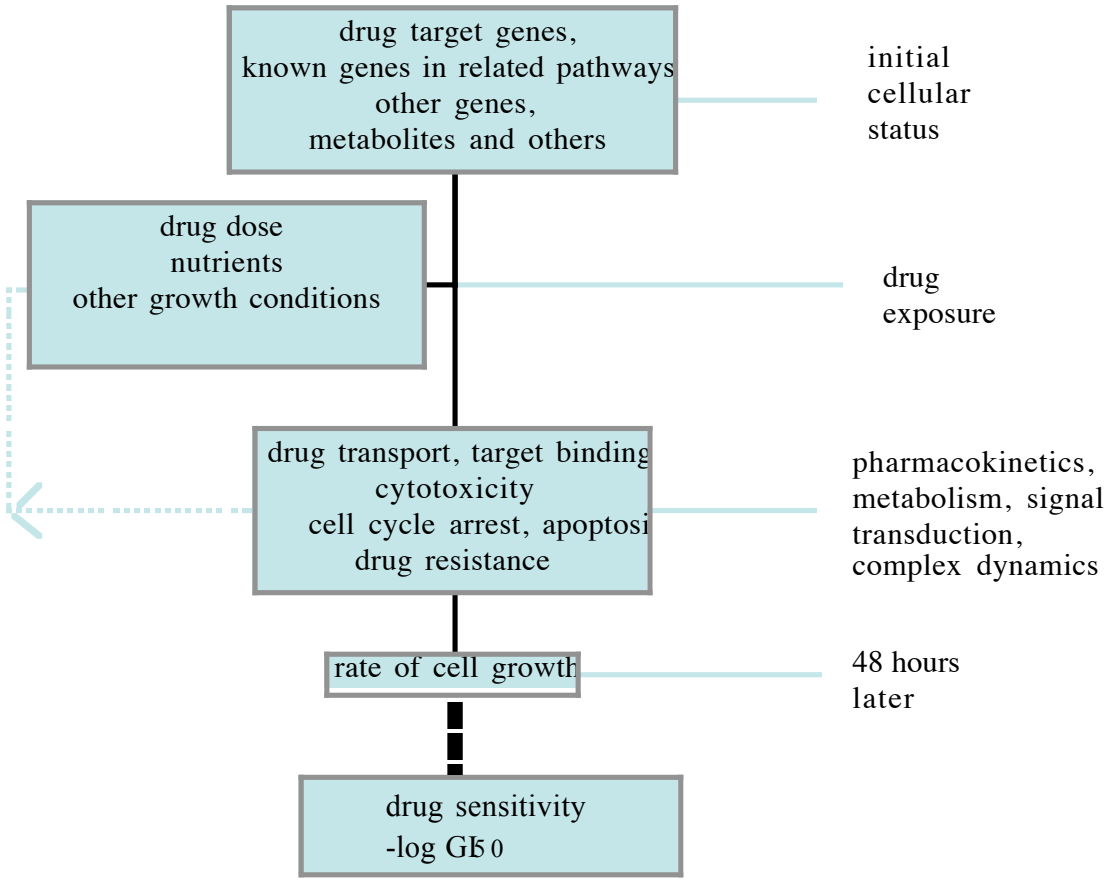
- Drugs with similar profiles usually share similar molecular structures and biochemical mechanism of actions. (R. Bai et al. 1991, K. D. Paull et al. 1992, H. N. Jayaram et al., 1992)
- Similarity is measured by Pearson's correlation coefficients.

- COMPARE (gateway to NCI's anticancer screen database)

Compare Drug sensitivity and Gene expression profiles

- Sherf et. al. (2000) compare gene expression profile with the drug sensitivity profile by computing correlation coefficient.
- 9706 genes in the expression data set.
- 118 chemotherapy agents with known molecular mechanism of drug action
- Genes whose profiles correlate well with drug sensitivity profiles are thought to be related to drug functioning. Why?

Figure 1. Gene-drug interaction



In the NCI anticancer screen, each candidate agent is tested for a broad concentration range against each of the 60 cell lines in the panel. More than 60,000 agents have been screened.

Similarity comparison : Computer program "COMPARE:each compund is compared with others and a list of agents with
similar patterns in responsiveness in inhibiting growth inhibition is given,

Limitation of Pearson's Correlation Coefficient (CC)

- Many drugs do not correlate well their molecular target genes.
- The pair, MTX and DHFR, has only 0.071.

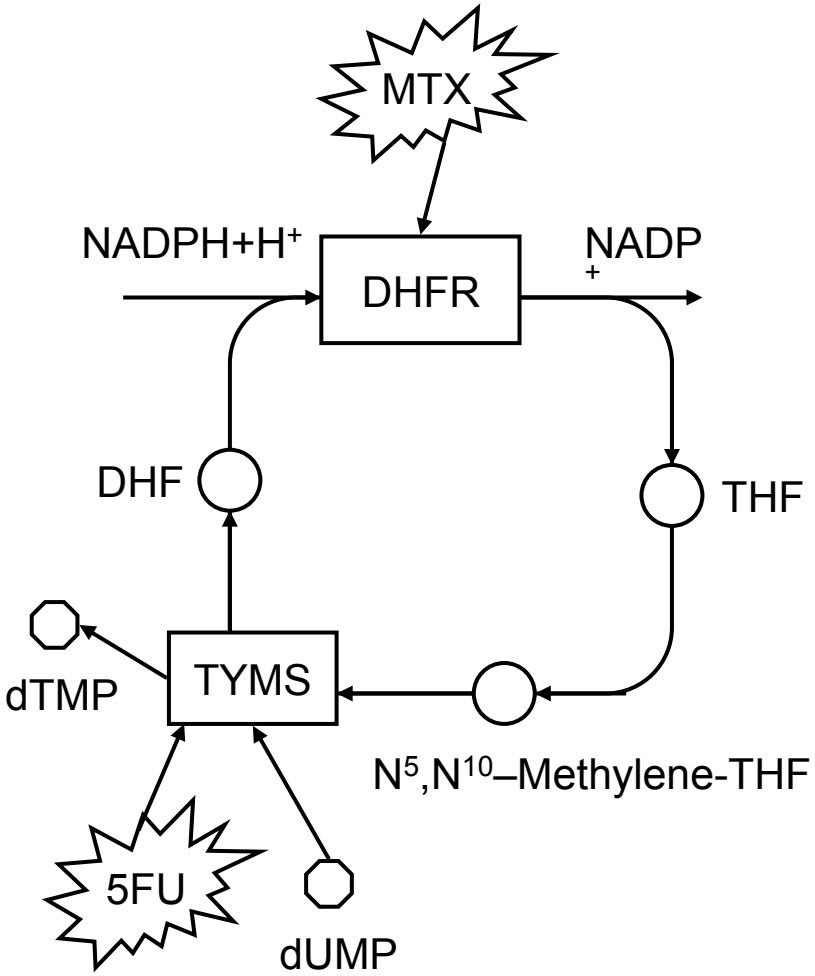
Liquid Association (LA) is used to understand the relationship between MTX and DHFR.

Methotrexate

MTX has been used for treatment of childhood acute lymphoblastic leukemia (ALL), non-Hodgkin's lymphoma, osteogenic sarcoma, chorocarcinoma and carcinomas of breast, head and neck[7]. The binding of MTX and its polyglutamated forms to its primary target DHFR inhibits the reduction of folate and 7,8-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate(THF). The inhibition in turn results in the quick conversion of all of a cell's limited supply of THF to DHF by thymidylate synthase (encoded by *TYMS*) reaction (Figure 2, left panel). This prevents further dTMP synthesis.

Inhibition of DNA component synthesis

- X is chosen to be a drug. Y is chosen to be the target gene which encodes the protein known to participate in drug's activity.
- E.g. MTX (Methotrexate) and DHFR (Dihydrofolate Reductase).

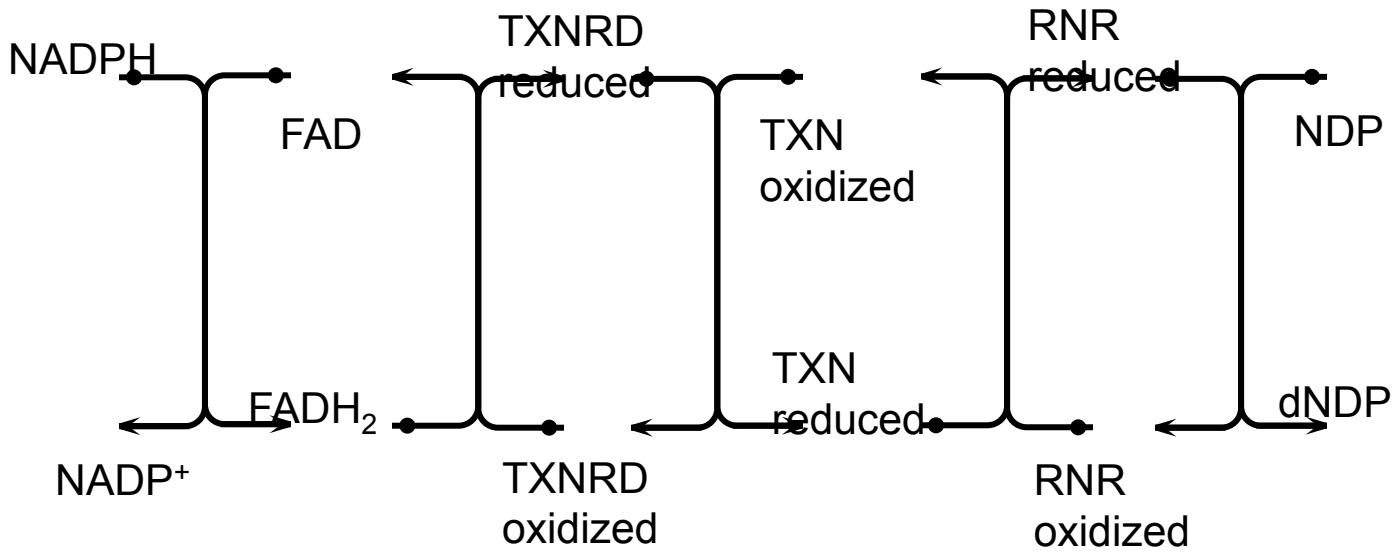


RNA to DNA
U to T

MTX, DHFR,TYMS

Table S.1 Genes with high LA scores for MTX, DHFR, TYMS												
Table S.1 (A)				Table S.1 (B)				Table S.1 (C)			
MTX				MTX				DHFR			
DHFR				TYMS				TYMS			
Z	LAP	Z	LAP	Z	LAP	Z	LAP	Z	LAP	Z	LAP
MDA5	0.4062	KIAA1706	-0.4263	DGSI	0.391	NA-6043	-0.3016	NA-1837	0.4559	CSTB	-0.3975
MEF2D	0.3755	CCNH	-0.3878	RANBP1	0.3566	NA-4504	-0.3	NA-2696	0.4421	CRYL1	-0.3925
NA-6915	0.3746	NA-3315	-0.3869	MAPK1	0.3412	KIAA0276	-0.2998	NA-2283	0.4276	NA-1556	-0.3903
NA-1878	0.3699	TXN	-0.3832	UFD1L	0.3375	NA-3333	-0.2969	TEM8	0.415	GJA5	-0.3903
NA-1684	0.3659	FLJ10035	-0.3832	HTF9C	0.3298	ATP2B1	-0.2889	TOX	0.4147	HPX	-0.3868
MFGE8	0.3655	EIF4E	-0.3675	NA-1390	0.324	RAB31	-0.287	SLC9A6	0.4048	MYO15B	-0.3848
KIAA0337	0.3647	COX11	-0.359	KPNA1	0.3225	CRIP2	-0.2768	NA-9327	0.3995	M17S2	-0.3837
TESK1	0.3612	NA-1802	-0.3581	PPIL2	0.3035	TXNRD1	-0.2752	MSN	0.397	NA-6537	-0.3827
KIAA0555	0.3585	PCSK7	-0.3505	NA-6953	0.3021	MGC2721	-0.2732	TMEFF1	0.3931	MRPL12	-0.3788
ZFPL1	0.3548	MGC21874	-0.3487	ATP6V1E1	0.2934	TNFRSF19L	-0.2732	NA-6473	0.3831	ICA1	-0.3757
EHD2	0.3521	KIAA1354	-0.343	GHR	0.2931	CYP2C9	-0.2689	TBC1D5	0.3803	GPCR1	-0.3721
RER1	0.3435	C2	-0.3428	IFRD2	0.285	ABI-2	-0.2686	SUSP1	0.3736	HRASLS3	-0.3704
DDAH2	0.3343	NA-1844	-0.3422	CDK4	0.2835	AEBP1	-0.2674	IGFBP3	0.3675	MAD	-0.3702
NA-728	0.3327	DPP4	-0.3324	BHC80	0.2802	SPAG9	-0.2619	EDN3	0.3643	FLJ22283	-0.3692
NA-671	0.3227	WHSC2	-0.3321	MRPL49	0.2776	NA-2843	-0.2614	FLJ10392	0.3633	LYZ	-0.3679
SPTB	0.3208	C9orf10	-0.3294	SERHL	0.2742	APPBP2	-0.2603	NA-5323	0.3629	TXN	-0.3666
APOC1	0.3196	FLJ20156	-0.3294	HIRA	0.2685	FOXP1	-0.2601	FST	0.3601	RNF13	-0.3665
RAB2L	0.3117	KIAA1078	-0.3289	HSRTSBETA	0.2647	GSTZ1	-0.259	PRKACB	0.3582	LOC51133	-0.3661
CNTNAP1	0.3103	MST4	-0.3263	COPA	0.2645	HAN11	-0.257	FAF1	0.3581	SERPINA6	-0.3656
NA-4135	0.3102	ABCE1	-0.3255	NA-1676	0.2639	CORO2B	-0.2558	MSCP	0.3581	DAB2IP	-0.3638

Pathway for TXN and TXNRD



MTX-sensitivity and the co-expression pattern of DTTT subsystem.

- ***TXN* and *TXNRD1* encode thioredoxin and thioredoxin reductase respectively.**
- **Together with DHFR and TYMS, they form a subsystem (abbreviated DTTT) that critically regulates the biosynthesis of DNA components[8]**
- **The role of the thiol-disulfide redox regulation in tumor growth and drug resistance is reviewed in Reference (9).**

LAP Web Application

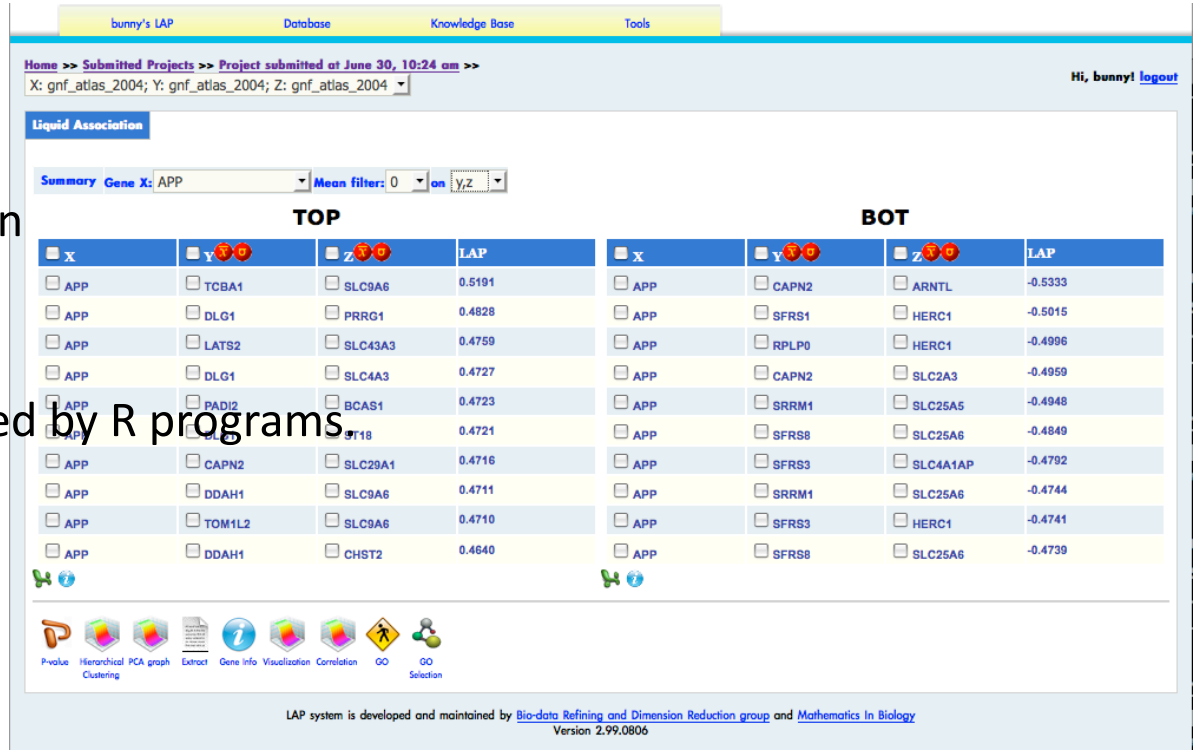
- A web application for biological researchers
  - http://kiefer.stat2.sinica.edu.tw/LAP3/
  - More than a web database
    - Search, compute, and data analysis
    - Create, save, and share LAP projects.
  - No dedicated platform, no installation
    - You may use LAP Web Application on any device with web browser

Technologies behind LAP Web App.

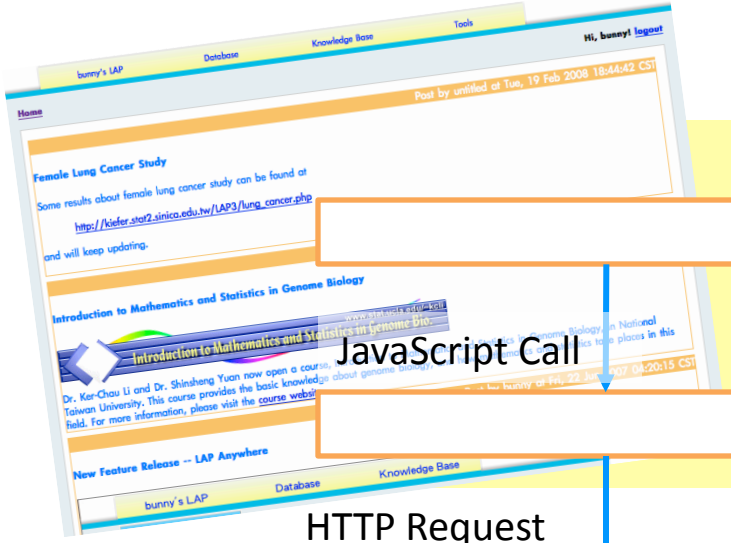
- Distributed computing
  - Computation is performed by the cluster of computers in MIB group
    - User does not need a powerful machine
- AJAX(Asynchronous JavaScript and XML, intensively used in Web 2.0 applications)
  - Provide “application-like” web services, better user experience than traditional web services.
    - Enable interactions between front-end user and back-end server facilities.

LAP Web App. Features

- LAP is more than a biological database
  - Search and compute!
    - Supports the following computation methods:
      - LAP
      - PLA
      - Correlation
      - Clustering
      - Cox regression
      - PCA
      - P-value
      - Graphs created by R programs.



LAP Web App. Architecture



Web browser (user client)

User Interface

JavaScript Call

HTML+CSS data

Ajax Engine

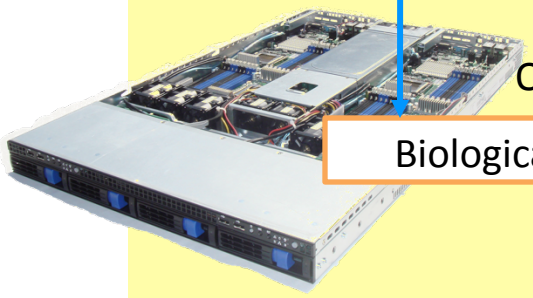
HTTP Request

XML data

Web Server + PHP engine

SQL Query

Commands



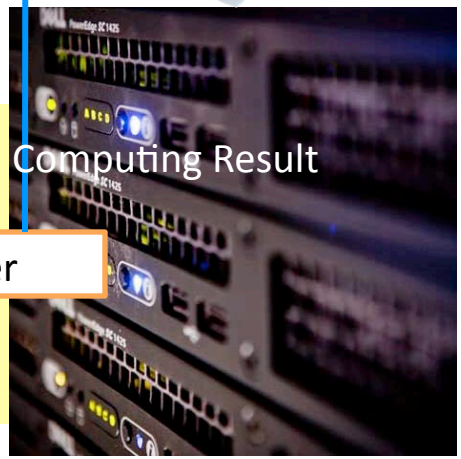
Query Result

SQL Query

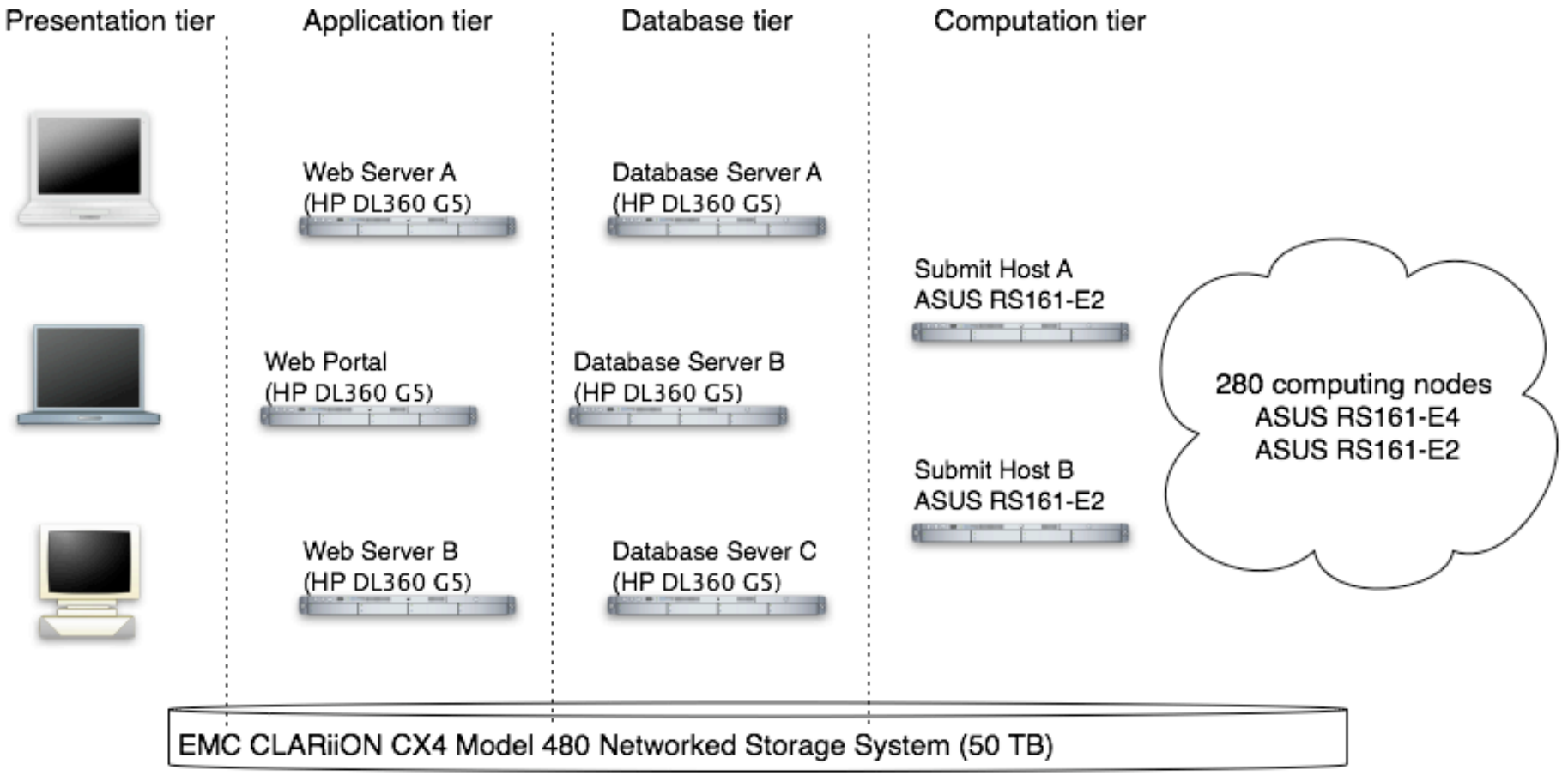
Biological Databases

Computer Cluster

Query Result

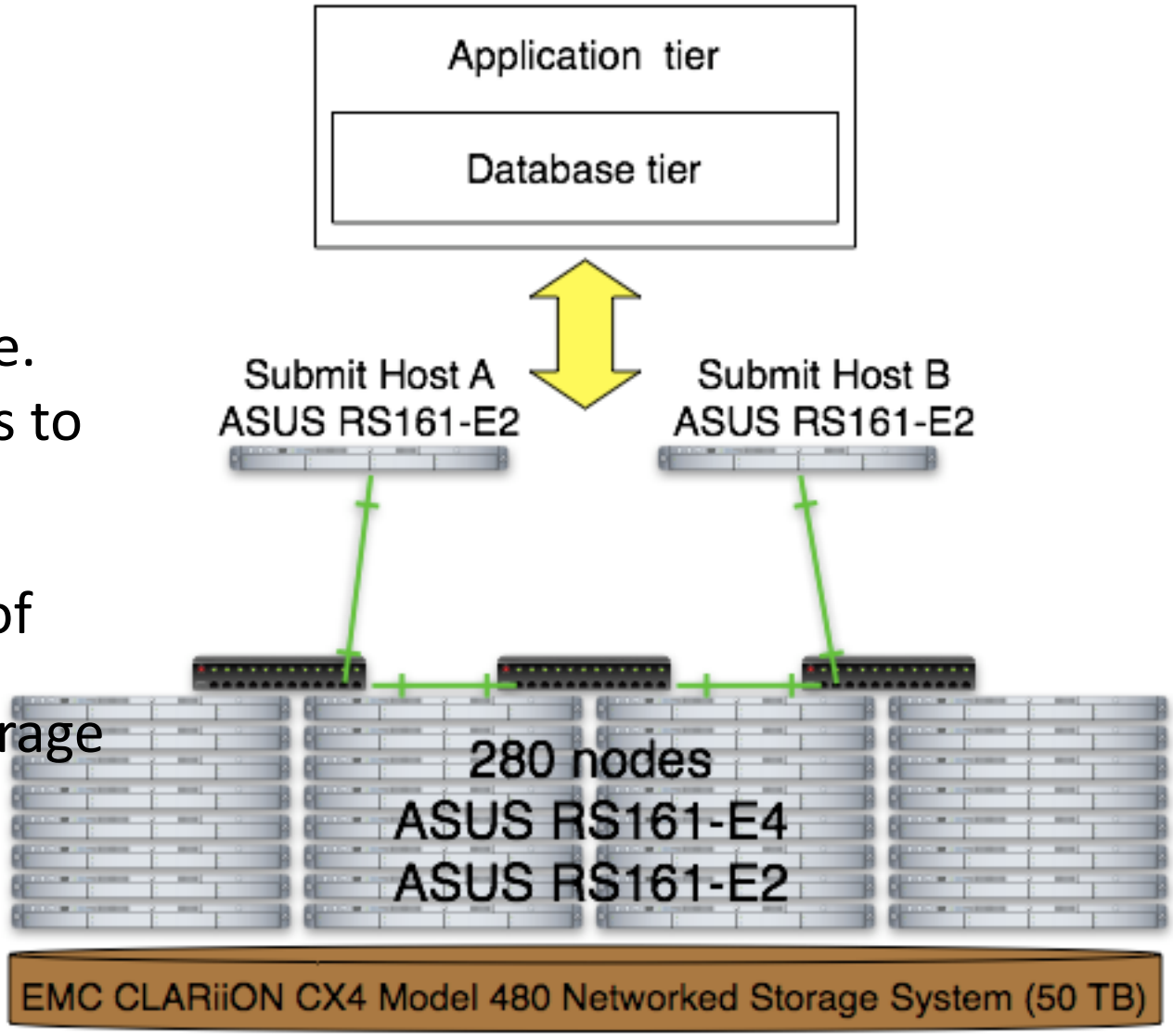


Four-tier architecture for LAP



Computation tier

- 280 computing nodes for pre-computing large scale data and serving end-user from the web site.
- Two submit hosts to increase host availability
- Routine backup of the end-user's projects with storage system of 50 terabytes



Future Work — Performance improvement

- Improve computing performance
  - Optimization/Parallelization of algorithms
- Improve database performance
  - Database structures
  - Parallelization of database queries

Future Work – Cloud Computing

- Toward cloud computing
  - More computation facilities open to the research community.
  - Better separation/protection of computing resources

Supporting various in house research projects

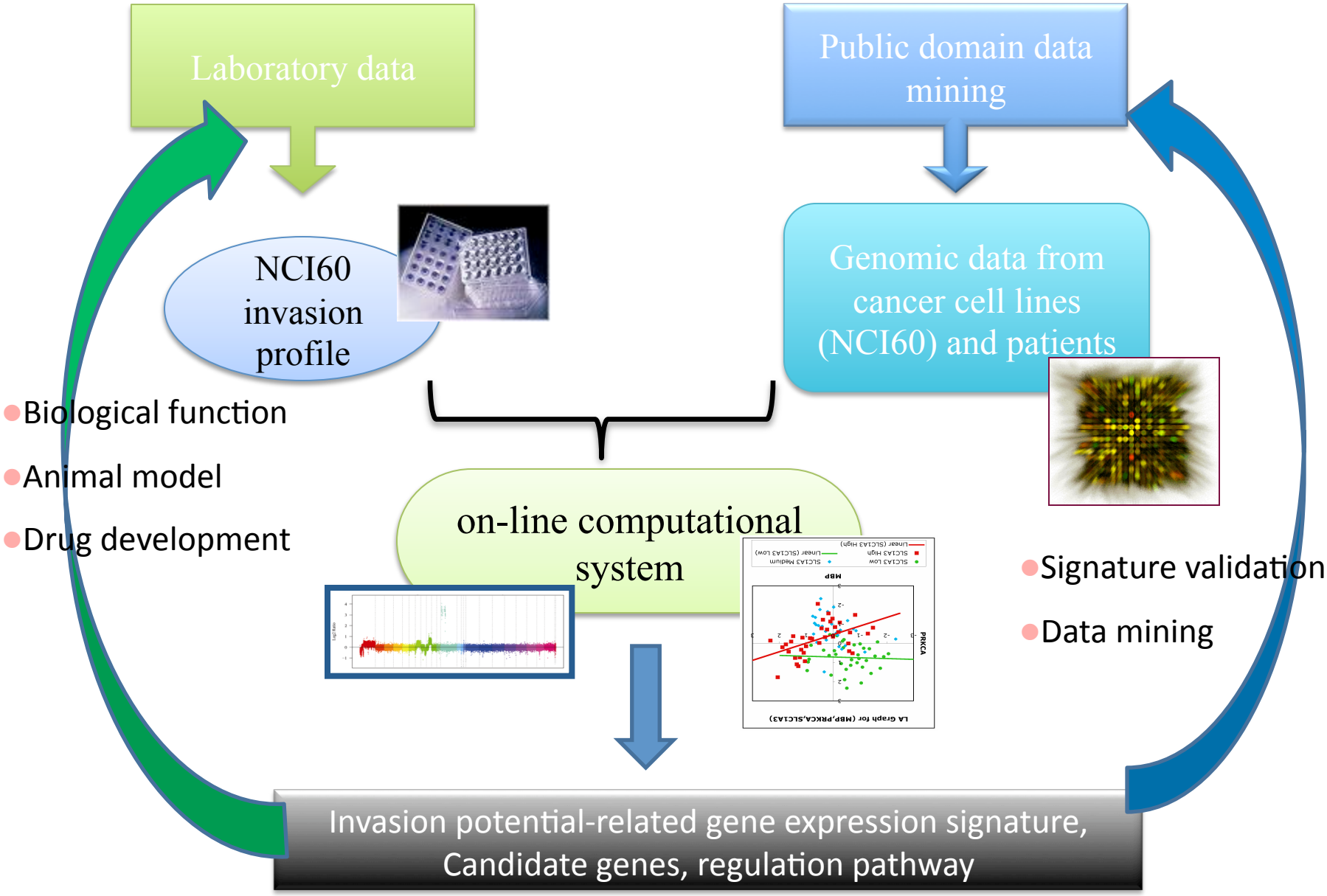
- Lung cancer
- Tumor invasion study using NSC60 cell lines

NCI-60 cell line based integrative computational system for tumor invasion- related genes

許藝瓊博士

中研院統計所

syic@stat.sinica.edu.tw


Laboratory data
Public domain data mining
NCI60 invasion profile
Genomic data from cancer cell lines (NCI60) and patients
Biological function
Animal model
Drug development
on-line computational system
Signature validation
Data mining
Invasion potential-related gene expression signature, Candidate genes, regulation pathway

LA related References

- **Li, K.C. (2002) Genome-wide co-expression dynamics: theory and application.** ***Proceedings of National Academy of Science*** **. 99, 16875-16880.**
- Li, K.C., and Yuan, S. (2004) A functional genomic study on NCI's anticancer drug screen. *The Pharmacogenomics Journal*, **4**, 127-135.
- **Li, K.C., Ching-Ti Liu, Wei Sun, Shinsheng Yuan and Tianwei Yu (2004). A system for enhancing genome-wide co-expression dynamics study.** ***Proceedings of National Academy of Sciences*** **. 101 , 15561-15566**.
- Yu , T., Sun, W., Yuan , S., and Li, K.C. (2005). Study of coordinative gene expression at the biological process level. *Bioinformatics* **21** 3651-3657.
- Yu, T., and Li, K.C. (2005). Inference of transcriptional regulatory network by two-stage constrained space factor analysis. *Bioinformatics* **21**, 4033-4038.
- Wei Sun; Tianwei Yu; Ker-Chau Li (2007). Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics*; doi: 10.1093/bioinformatics/btm327 (correspondence author: Li)
- Yuan, S., and Li. K.C. (2007) Context-dependent Clustering for Dynamic Cellular State Modeling of Microarray Gene Expression. Bioinformatics 2007; doi: 10.1093/bioinformatics/btm457 (correspondence author: Li)
- **Li, KC, Palotie A, Yuan, S, Bronnikov, D., Chen D., Wei X., Choi, O., Saarela J., Peltonen L. (2007) Finding candidate disease genes by liquid association. Genome Biology (in Press).**
- **Wei, S., Yuan,S., and Li, K.C. Trait-trait interaction: 2D-trait eQTL mapping for genetic vriation study. BMC Genomics 2008, 9:242**

Acknowledgements

- Robert Yuan
- Former UCLA students : Wei Sun, Ching-Ti Liu, Xuelian Wei, Tianwei Yu
- Current biology collaborators: Pan-Chyr Yang (Dean of Medical School, NTU), S.L.Yu, H.W. Chen
- Post Docs : Yi-Chiung Hsu, Pei-Ing Hwang, Shang-Kai Tai
- Mission specific Research assistants:

Guan I Wu, Ying-Fu Ho, (hardware)

Hung Wei Tseng (Web 2.0, Ajax)

Cin Di Wang, Chia Hsin Liu, Cheng-Tao Chen, Chia Hung Lin, Kang Chung Yang, Shiao Bang Chang, Shian-Lei Ho