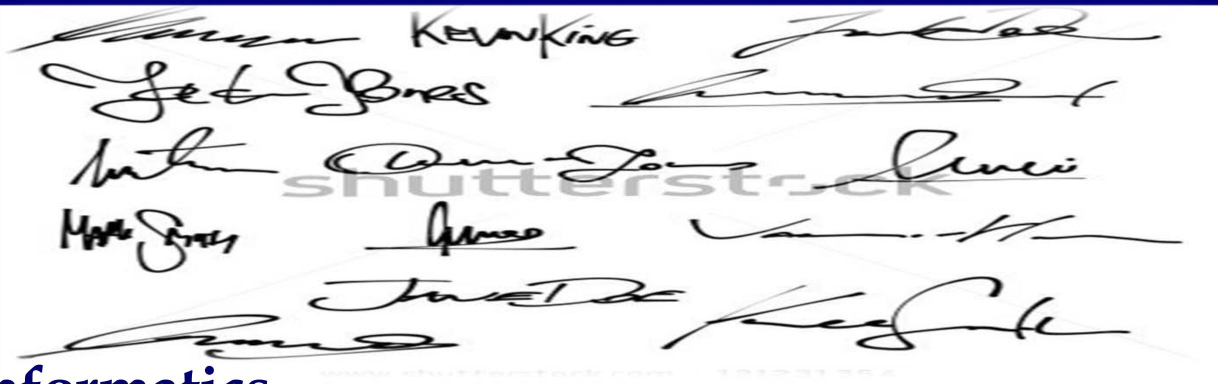
Precision Medicine: Deciphering patient-specific signatures



Radha Nagarajan, Ph.D.
Associate Professor
Division of Biomedical Informatics
University of Kentucky

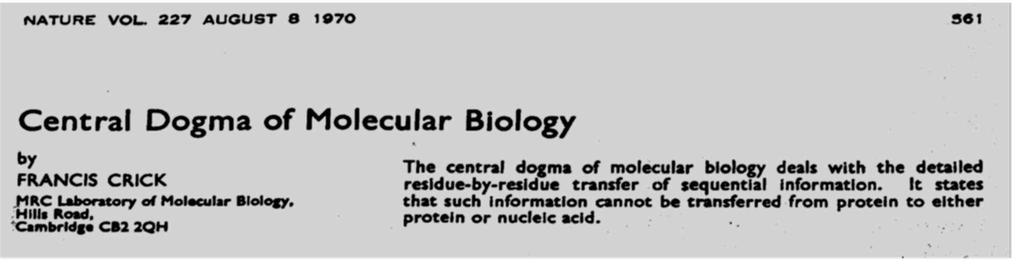
Current Research

- Broad theme: Knowledge discovery from high-throughput molecular and observational healthcare data sets using machine learning and network science approaches.
- Specific Biomedical Informatics Areas
  - Precision Medicine
  - Translational Biomedical Informatics
  - Systems Biology

Note: Some of the concepts are generic and can be extended readily to other settings (e.g. evolution of team science).

Central Dogma of Molecular Biology (Crick, 1970)

Handshake between **three macromolecules** (DNA, RNA, Protein) and **three critical biological processes** (Replication, Transcription, Translation).



NATURE VOL. 227 AUGUST 8 1970 561

Central Dogma of Molecular Biology

by
FRANCIS CRICK
MRC Laboratory of Molecular Biology,
Hills Road,
Cambridge CB2 2QH

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

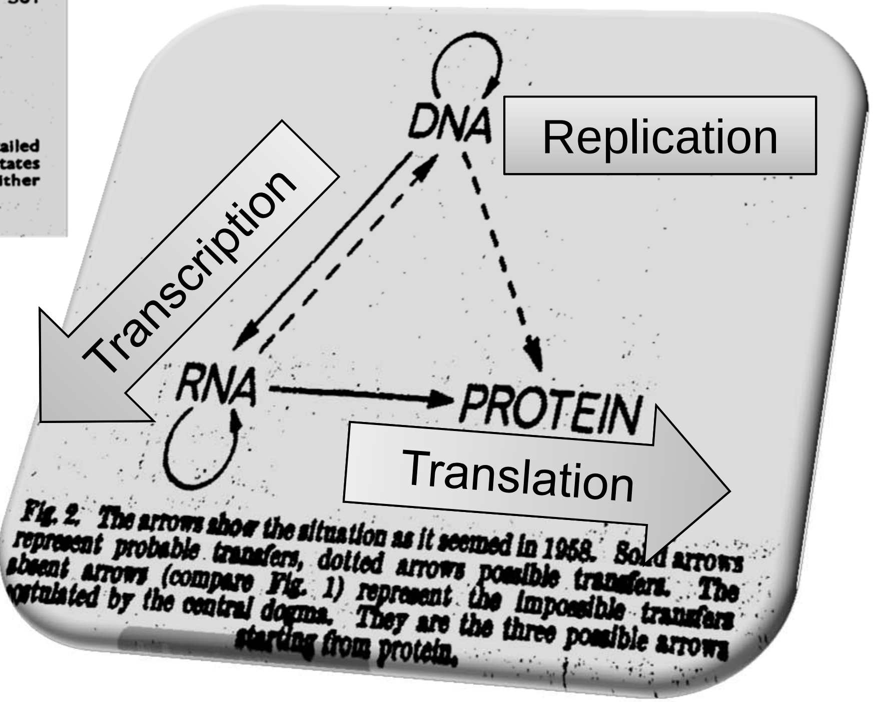


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

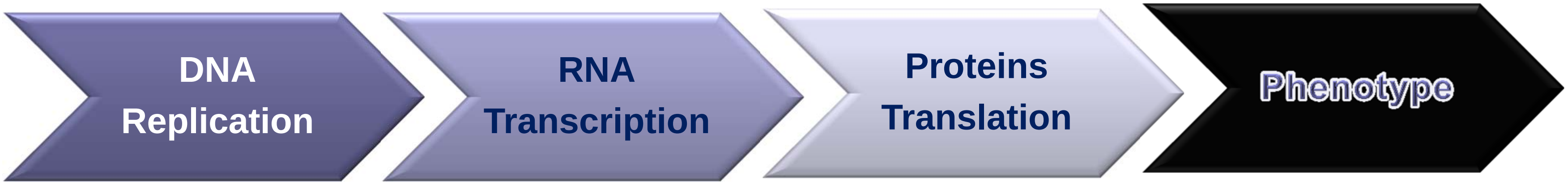
francis crick

- Stated by Crick in 1958
- Detailed treatment can be found @ (Crick, F. 1970, Nature)

More importantly, these three processes precede phenotype formation.

Phenotype? Observable trait (e.g. physiological,
morphological, behavioral)

Proximity to Phenotype



DNA Replication

RNA Transcription

Proteins Translation

Phenotype

Bottom line Understanding the molecular signatures across these molecules/processes can provide novel insights into mechanisms underlying phenotype formation.

Profiling DNA, RNA, Protein:
Sequencing (Whole Genome)

- Macromolecule of interest: DNA
- Can reveal differences in sequence composition across subjects.

  Note: Establishing causality between sequence differences and the observed phenotype can be tricky!

eric lander

Question Can we get closer to the phenotype?

Molecular Profiling

- Macromolecules of Interest: RNA, Protein
- Processes of Interest: Transcription and Translation
- High-throughput Assays: Microarrays, Protein Arrays, RNA Sequencing

AFFYMETRIX®

illumina®

pat brown

mike snyder

So we have ability to profile critical molecules.

Question

Can we use discern distinct phenotype from their molecular profiles? (**Molecular Diagnostics)**

Classical workflow

- Given Expression of molecules across two groups/classes (e.g. normal subjects (-), lung cancer subjects(+))
- Objective
  - Discern these groups using classification algorithms.
  - Evaluate the performance of the classifier using established performance measures (e.g. Sensitivity (tp/p), Specificity (tn/n), Accuracy (tp+tn)/(p+n), ……………).
  - Predict the label of a new incoming samples.

Routine classification while useful subscribe to “One-size fits all”

Sits below the waist
Slim through seat & thigh
Skinny, 14.75" leg opening
Sits below the waist
Slim fit
Super skinny, 13.25" leg opening
Spray on ultra-skinny fit
Stretch denim for comfort
Super skinny, 11.5" leg opening

© Levis

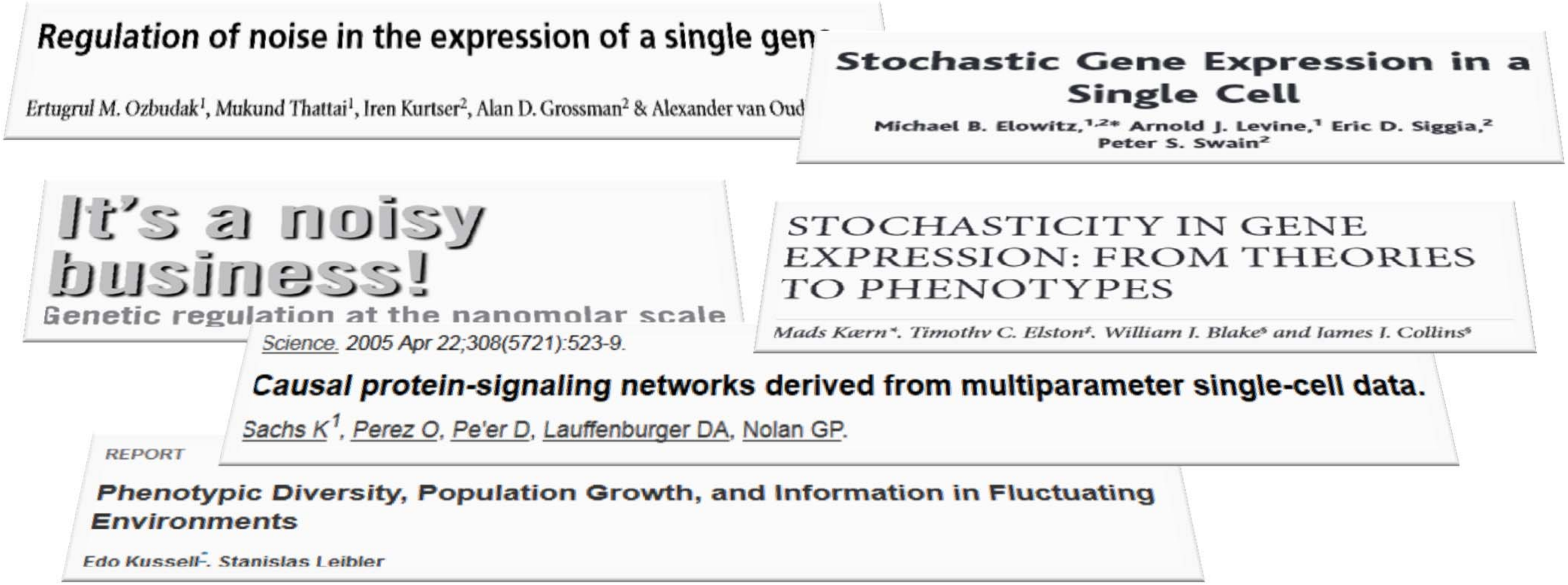
- Assumption Samples within a given disease phenotypes are homogeneous.
- Assumption Samples that don’t fit a pre-defined cut-off or profile are often deemed as an outlier and filtered out prior to classification (is this ethical?)
- Assumption Same fixed set of molecular markers is used as representative of each sample within and between the groups

Homogeneity assumption are routinely violated

Heterogeneity in fact manifests across multiple scales.

@ molecular scale

- Molecular mechanisms are inherently stochastic.
- Several studies have demonstrated inherent noisiness and heterogeneity in expression across ”isogenic” (identical DNA sequence representation) single cells



Regulation of noise in the expression of a single gen

Ertugrul M. Ozbudak[1], Mukund Thattai[1], Iren Kurtser[2], Alan D. Grossman[2] & Alexander van Oud

Stochastic Gene Expression in a Single Cell

Michael B. Elowitz,[1,2*] Arnold J. Levine,[1] Eric D. Siggia,[2] Peter S. Swain[2]

It’s a noisy business!

Genetic regulation at the nanomolar scale

STOCHASTICITY IN GENE EXPRESSION: FROM THEORIES TO PHENOTYPES

Mads Kærn*, Timothy C. Elston‡, William J. Blake§ and James J. Collins§

Science. 2005 Apr 22;308(5721):523-9.

Causal protein-signaling networks derived from multiparameter single-cell data.

Sachs K[1], Perez O, Pe'er D, Lauffenburger DA, Nolan GP.

REPORT

Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments

Edo Kussell*, Stanislas Leibler

What we need

- Approach that accommodate inherent heterogeneity and patient-specific variations with potential to reveal sub-populations

Seemingly Homogeneous samples within a group

Potential sub-populations

© pink floyd

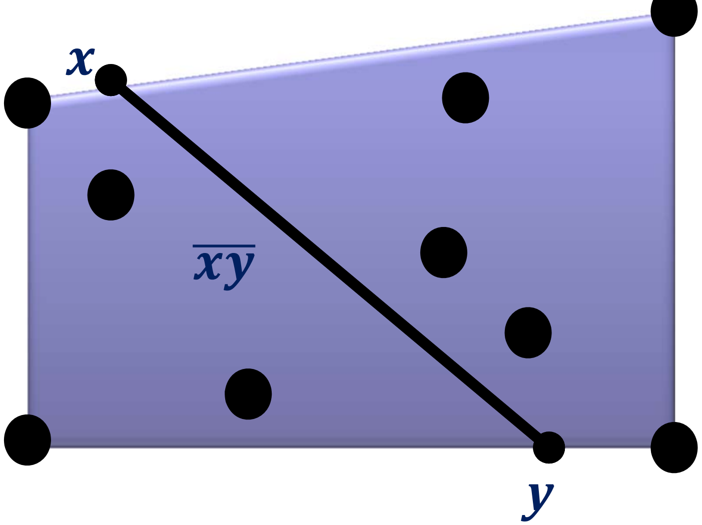
What we proposed:

Convex Hull Selective Voting Algorithm

Convex Hull

- **Convex combination**: Linear combination of points in a plane ($\boldsymbol{x, y \epsilon R^2}$) is given by

$$(1 - \alpha) * x + \alpha * y \text{ where } 0 \leq \alpha \leq 1.$$



- **Convex Subset:** A subset $\boldsymbol{\Sigma}$ of a plane is convex iff for any pair of points $\boldsymbol{x, y \epsilon \Sigma}$ the line segment $\boldsymbol{\overline{xy}}$ is completely contained in $\boldsymbol{\Sigma}$.
- **Convex hull $\boldsymbol{\Psi(\Sigma)}$**: is the smallest convex set containing $\boldsymbol{\Sigma}$.

Several algorithms have been proposed in the literature for generating the convex hull of a given set of points.

Why convex hull?

- Incorporates the geometry of the points in a plane (plane? since we consider two-dimensional projections where each point is represented by a pair of molecules)
- Does not impose any constraint on the distributional profiles or functional form

Convex Hull Selective Voting Algorithm

Given: Molecular expression profiles of m genes across two disease groups (classes) n samples con n_{con} and exp n_{exp} such that $n = n_{con} + n_{exp}$ represented by the matrix X_{mxn}.

Initialize the control and exp voting matrices as follows:

$$v^{con}(r,s) \leftarrow 0; v^{exp}(r,s) \leftarrow 0; s = 1 \dots n, r = 1 \dots N_b$$

- Step 1: Set $r \leftarrow 1$

Choose a window length $k1$ and divide the given samples X_{mxn} into training V_{mxk2} and test U_{mxk1} sets such that $k2 = n \backslash k1$ and $k2 \gg k1$.

Let the $clinical\ labels*$ of the n samples in X_{mxn} be stored in the vector L_{1xn} where

$L(k)$ = 1 implies the k^{th} sample is control

= 2 implies the k^{th} sample is cancer.

clinical labels? these labels are traditionally based on clinical criteria at the point of care.

- Step 2: Choose the gene pair (i, j), $i, j = 1 \dots \mathrm{m}, i \neq j$.
- Project the training data $(V_{ik2,}\ V_{ik2})$ onto a plane.
- Generate the convex-hulls (Ψ_{con}, Ψ_{exp}) corresponding to the control con and cancer exp training samples.
- If $\Psi_{con} \cap \Psi_{exp} \neq \Phi$ (i.e. the convex hulls overlap), we iteratively prune the hulls as described below.

Pruning the Hulls

i. Drop the vertices of the hulls in the overlap region

ii. Generate the pruned hulls $(\Psi'_{con}, \Psi'_{exp})$

iii. Set $(\Psi_{con} = \Psi'_{con}, \Psi_{exp} = \Psi'_{exp})$

iv. If $\Psi_{con} \cap \Psi_{exp} = \Phi$, then quit else go to (i).

- ❑ Step 3: Vote the samples in the test data represented by $(U_{ik}, U_{ik}), k = 1 \ldots k1)$ as follows:

  If $(U_{ik}, U_{ik}) \epsilon \Psi_{con}$, then the vote the k^{th} sample as con, increment the vote $\upsilon^{con}(r, k) \leftarrow \upsilon^{con}(r, k) + 1$

  If $(U_{ik}, U_{ik}) \epsilon \Psi_{exp}$, then the vote the k^{th} sample as exp, increment the vote $\upsilon^{exp}(r, k) \leftarrow \upsilon^{exp}(r, k) + 1$

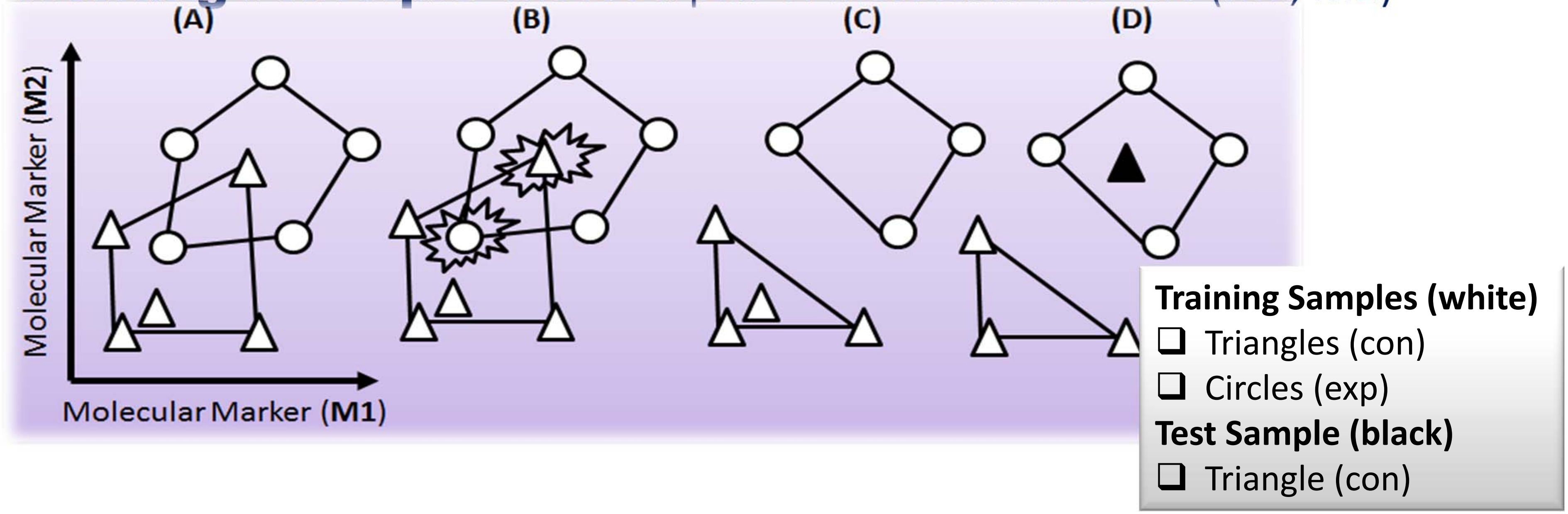
- ❑ Step 4: Repeat Steps 2 and 3 for each pair of genes (i, j). The number of times each sample is voted as con and exp is stored in υ^{con} and υ^{exp} respectively

- ❑ Step 5: Repeat Steps 2–4 using the next window of $k1$ samples as test sample.

- Step 6: Repeat Step 5 till each sample is voted on as a test case once.
- Step 7: Set $r \leftarrow r+1$. Repeat Steps 1–6, by randomly permuting the columns of X_{mxn}.
- Step 8: class label $L^{*}(k)$ of the k^{th} sample is
  - **con**: if $v^{con}(,k)$ and $v^{can}(,k)$ are significantly separated and $E[v^{con}(,k)] > E[v^{can}(,k)]$
  - **exp**: if $v^{con}(,k)$ and $v^{can}(,k)$ are significantly separated and $E[v^{can}(,k)] > E[v^{con}(,k)]$
  - **neither**: if $v^{con}(,k)$ and $v^{can}(,k)$ are not significantly separated

- Determine the classification performance measures [e.g. Accuracy (ACC), Sensitivity (SEN), Specificity (SPC)] by using the clinical label $L(k)$ as the ground truth.

Working Principle Choose a pair of molecular markers (M1, M2)



Training Samples (white)

- □ Triangles (con)
- □ Circles (exp)

Test Sample (black)

- □ Triangle (con)

A. Generate the convex-hulls for the two classes using the training samples.

B. Prune the samples in the overlap region.

C. Obtain the pruned non-overlapping hulls.

D. Vote on the test samples.

In the above case, the test sample (con) is voted as (exp) by (M1, M2).
The exp ensemble set of the test sample consists of (M1, M2).

- Repeat for all pairs of markers and all test samples.
- Repeat multiple times by randomly assigning the samples to the test and training sets.

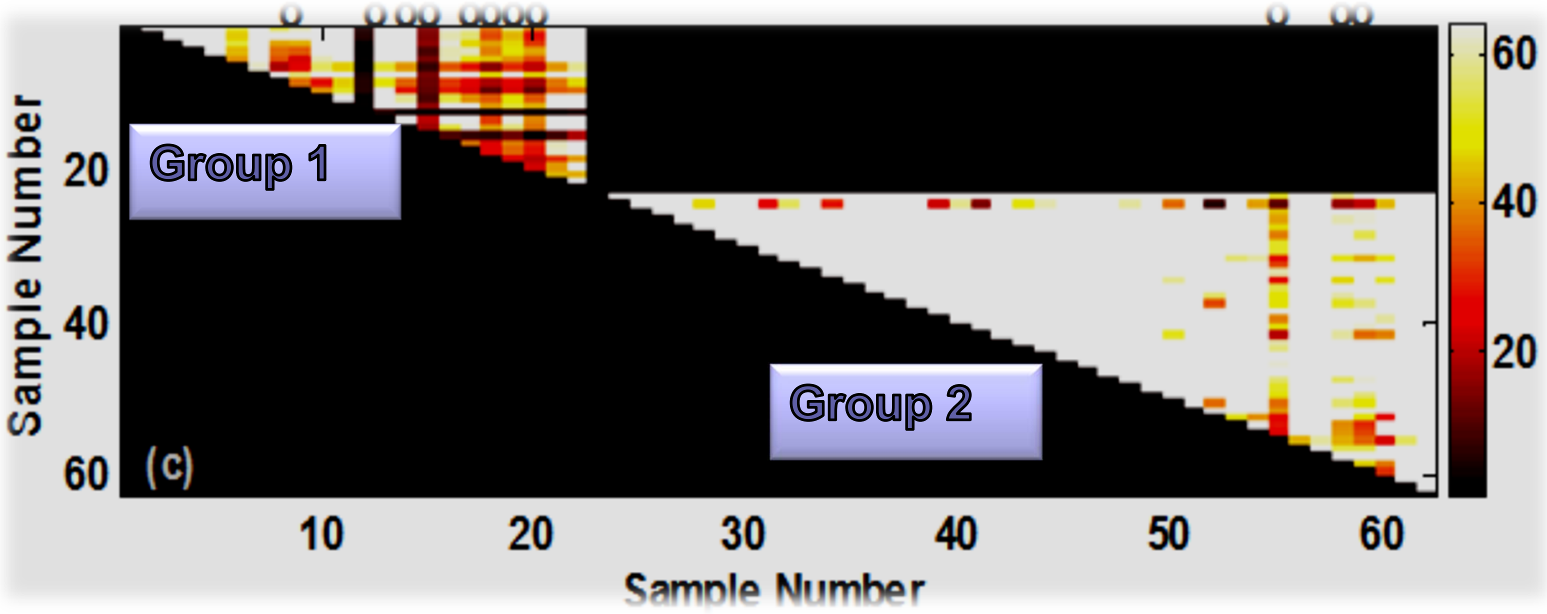
Personalized Ensemble Sets

Pairs of genes that vote on a given sample represents its ensemble set. Ensemble sets may not necessarily be identical across samples within a group.

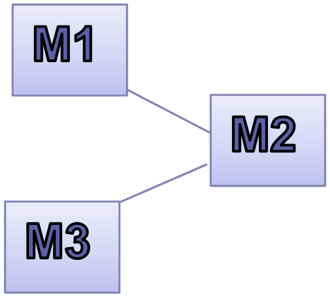
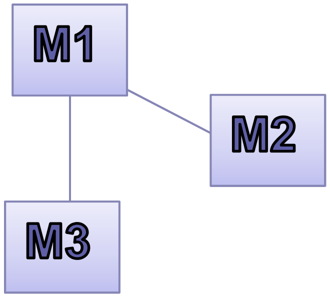
Ensemble Set Sample 1	Ensemble Set Sample 2
(M1, M2); (M2, M3)	(M1, M2); (M1, M3)

Consensus Map consensus between ensemble sets

- Dark streaks in the consensus reveal lack of consensus between the ensemble sets across the respective samples.



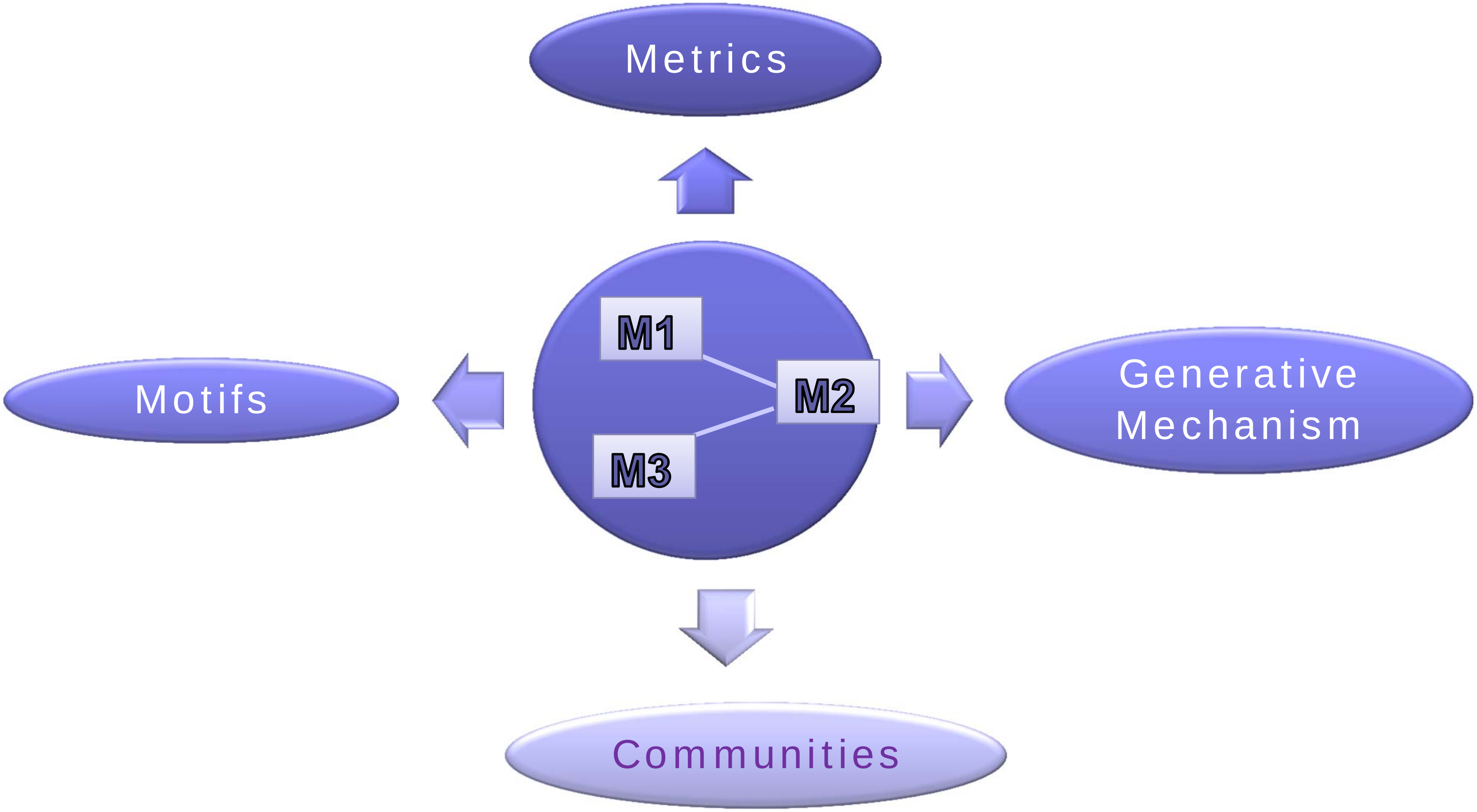
Network abstractions of the ensemble sets

Ensemble Set Sample 1 (M1, M2); (M2, M3)	Ensemble Set Sample 2 (M1, M2); (M1, M3)
	

Note: Same set of molecules (M1, M2, M3) vote for these two patients but their wiring is markedly different across the samples.

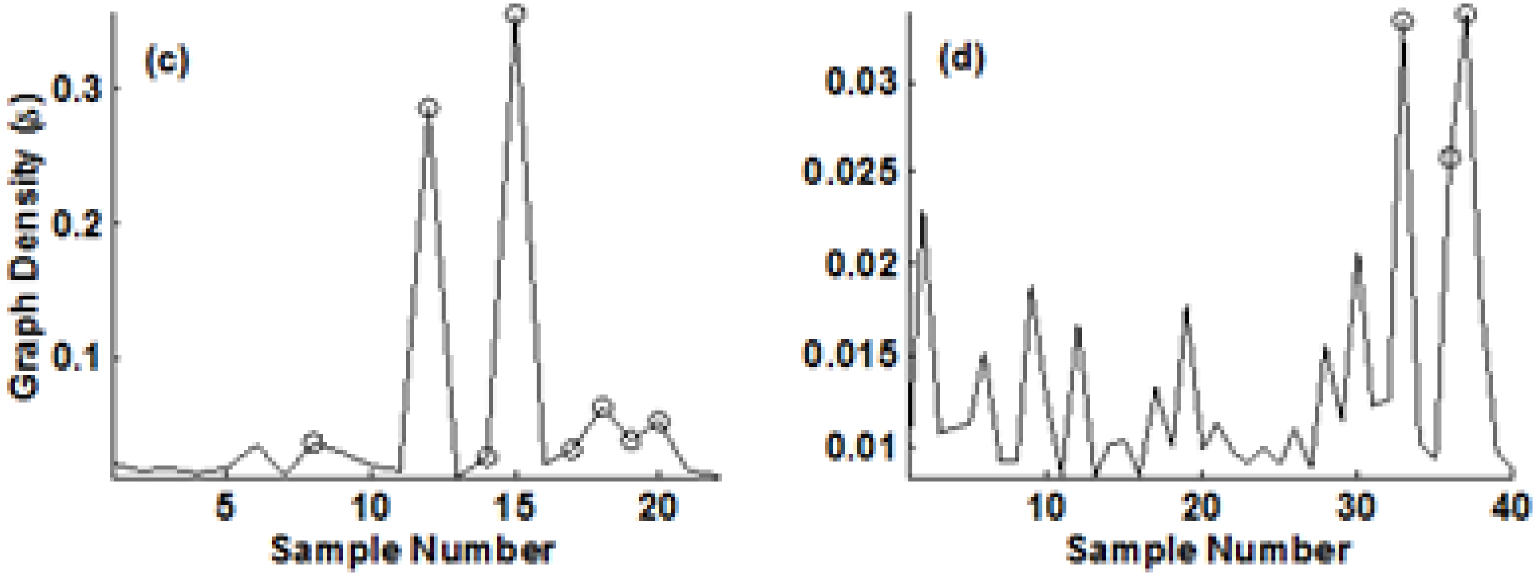
Question: Do the above patterns have any significance?

Network Abstractions of the Ensemble Sets



there's more …

Graph density estimates of the networks show marked deviation for the misclassified samples.



Application to periodontal disease – preliminary results

Periodontal assessment and classification at the point of care:

- Based on three clinical variables

Clinical variables

BOP: Bleeding on Probing

PD : Pocket Depth

CAL: Clinical Attachment Loss

- Classifies the patients into dichotomous silos based on three critical clinical variables with pre-defined cut-off.

Gingivitis

- BOP at ≥ 20% of sites
- < 10% of the sites have PD ≥ 4 mm
- No site with CAL ≥2 mm.

Periodontitis

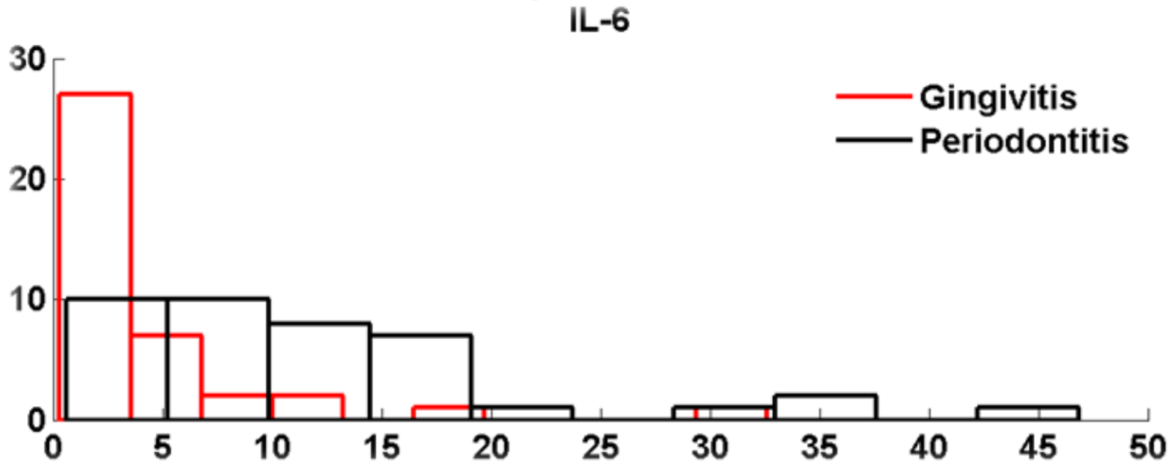
- BOP at > 20% of sites
- > 10% of the sites had PD ≥ 4 mm
- > 10% of the sites had CAL ≥2 mm

Couple of things

- Uses pre-defined cut-off The same set of variables (PD, BOP, CAL) with pre-defined cut-off is used as the rule of thumb to discern the disease groups.
- Does not provide insights into patient-specific variations Classifies a sample into either of these groups (Gingivitis or Periodontitis). Does not provide insights into possible heterogeneities within these two groups.

Extracting patient-specific profiles based on salivary biomarkers

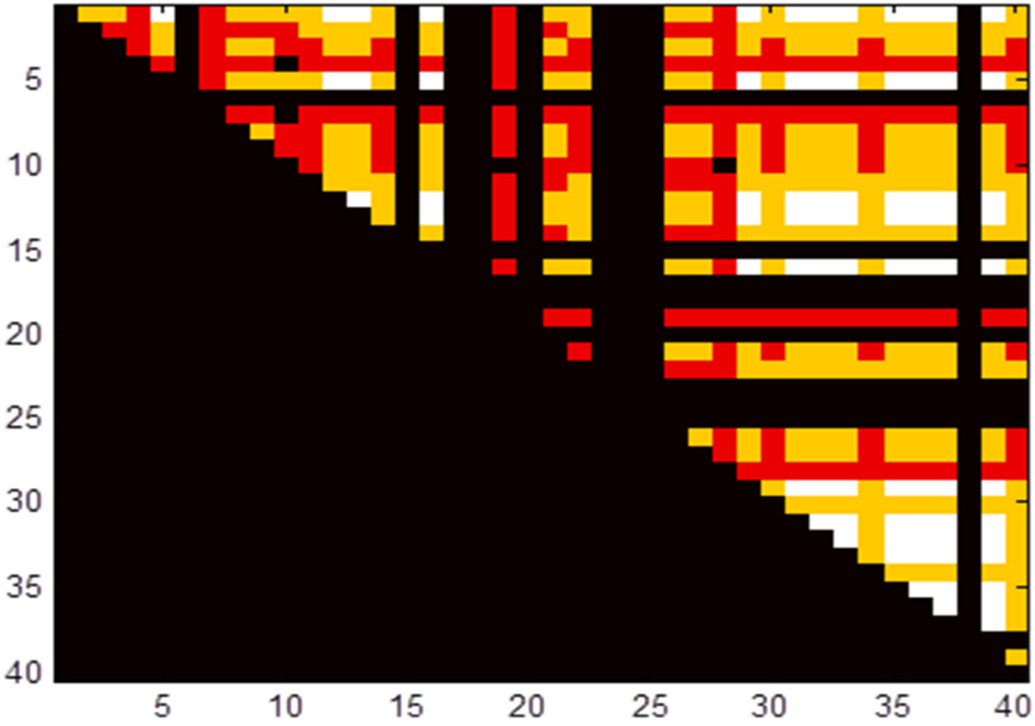
- Salivary Biomarkers Whole saliva was analyzed via Luminex/ELISA across four established salivary molecular markers IL-1ß, IL-6, MMP-8, MIP-1**a** corresponding to inflammation, soft tissue destruction and bone remodeling.



- Positiv**ely skewed** Distributions of the expression were positively skewed indicative of heterogeneity within Gingivitis as well as Periodontitis subjects.
- Overlap between the distribution Lack of a fixed threshold separating the two groups.

Deciphering patient-specific variations using the selective voting approach

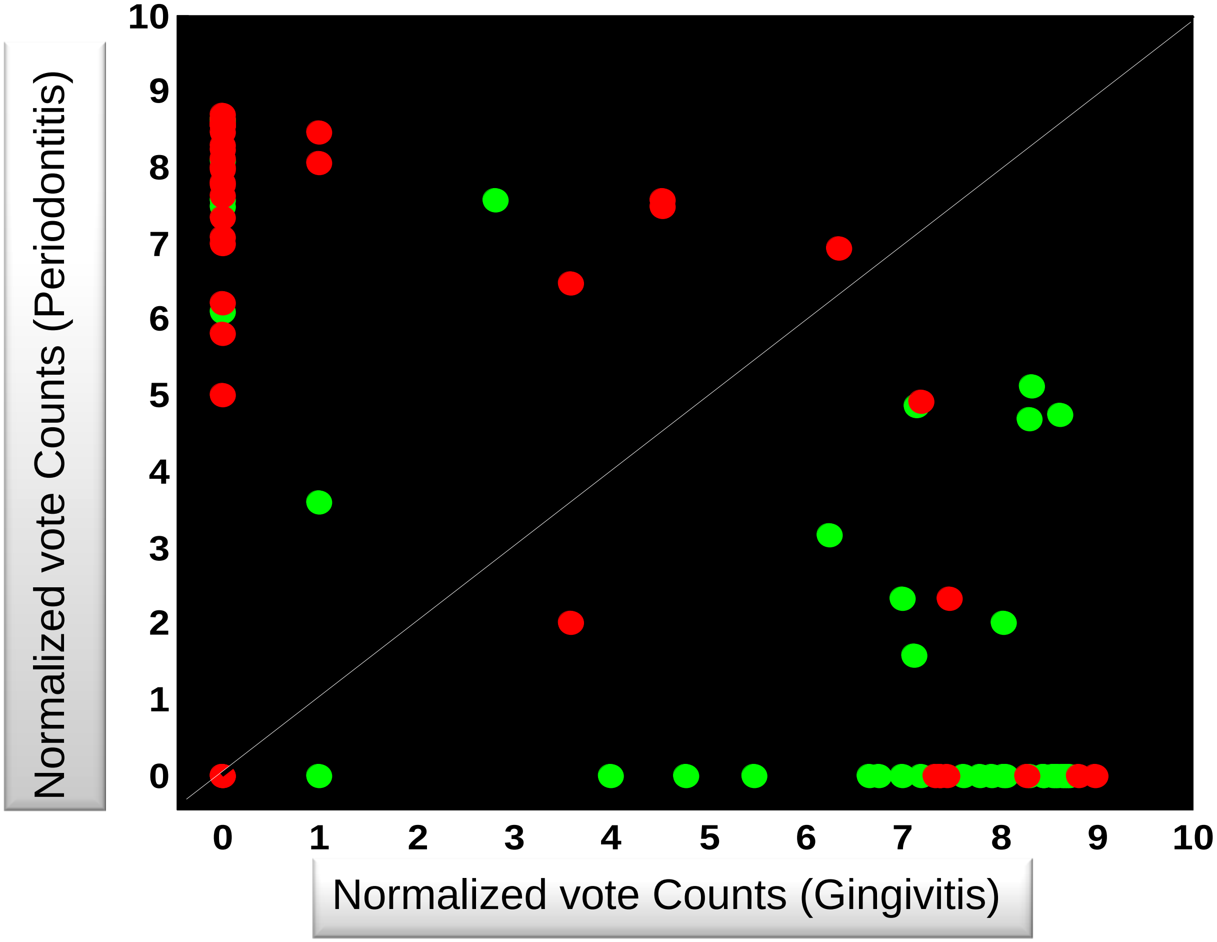
- Consensus Map Pronounced dark streaks in the consensus map revealed lack of consensus between these samples and the rest of the gingivitis samples.



- Performance Mesures Similar performance metrics (accuracy, sensitivity, specificity) across a battery of classification algorithms (LDA, QDA, NB, SVM). A subset of samples were routinely misclassified by all algorithms.

Varying proclivity of subjects to the disease groups

- Considerable variation in the proclivity of the patients to gingivitis (green, abscissa) and periodontitis (red, ordinate) revealed by the vote counts.
- Diagonal represents the line of separation between the disease phenotypes.



❑ Collaborators

▪ Jeffrey Ebersole, Craig Miller	Center for Oral Health Research, College of Dentistry
▪ Chi Wang, Sally Ellingson	Markey Cancer Center

❑ Funding Sources

▪ **P20GM103538/NIGMS**	(PI: Ebersole)
▪ **UL1TR000117**	(PI: Kern)

❑ Acknowledgements

- Google Images/Wiki

Questions?