





# UM Rhino Challenge Data Analysis

### Summit

April 14, 2008

A. Hero, P. Wolfe

Y. Huang, A. Rao, P. Harrington, M. Kliger

Depts of EECS and Bioinformatics Program

University of Michigan

### Outline

- Recapitulation of milestones
- Time course analysis (EDGE, RF)\_
- Differential motif analysis (DMDA)\_
- Path forward

### **Recapitulation of milestones**

subject	#sxbegan	sx peak	info	0.1onset	0.2onset	0.8onset	0.1maxT	0.2maxT	0.8maxT	
	1 control									
	2 control									
	3 54 hrs	day 4	moderate	5.4	10.8	43.2	9.6	s 19.2	2 76.8	
	4 60 hrs	day5	severe	6	12	48	12	2 24	4 96	
	5 asx and sl	hedding								
	6 36 hrs	day 3	severe sx	3.6	7.2	28.8	7.2	2 14.4	4 57.6	
	7 day 2 42 l	nday 2 at	56severe sx	4.2	8.4	33.6	5.6	5 11.3	2 44.8	
	8 control									
	9 day 3	day 3	mild sx	7.2	14.4	57.6	7.2	2 14.4	4 57.6	Pd>0.7 at Pf<0.05 at 0.2T
	10 control									
	11 48 hrs	4	8 mild sx	4.8	9.6	38.4	4.8	9.6	38.4	
	12 asx and sl	hedding								Pd>0.95 at Pf<0.01 at 0.81
	13 control									
	14 control									
	15 day 2 48 l	hiday 2 54	hsevere sx	4.8	9.6	38.4	5.4	11.3	2 43.2	
	16 day 2 42 I	h54 hours	moderate	4.2	8.4	33.6	5.4	i 11.3	2 43.2	
	17 asx and sl	hedding								
	18 control									
	19 day 5	day 5	mild	12	24	96	12	2 24	1 96	
:	20 44 hours	day 4	severe sx	4.4	8.8	35.2	9,6	3 19.2	2 76.8	

# Main findings

>mRNA data quality is generally high but exhibits high biological variability

>A small panel discriminatory biomarkers has been found

»Several of these biomarkers appear to be in good agreement with immune and inflammatory pathways

Benchmarks can be attained (caveat: small sample)

>mRNA and immunoassay markers seem to have highest predictive value

»Differential S vs A molecular signatures are strong at 0.8T, weaker at 0.1T

## Analysis

»Methods adapted to different fundamental assumptions about the

"signal" model:



- > 1) Linear (fixed effects) regression model
- > 2) Random effects model
- > 3) Mixed random effects model

# Linear fixed effects of Analysis applied

•SAM/PAM analysis (TimeSlice)\_

•EDGE/Co-cluster Analysis (TimeCourse)\_

•Random effects

•LDA/PCA (TimeSlice)\_

•Pareto sample depth distributions (PSDD)

(TimeCourse)\_

Mixed Effects

•Random forest regression (Time\_Slice)\_

Differential Matif Discovery and Analysis (TimeCourse)

### SAM/PAM/EDGE Analysis

Differential Expression Analysis At Baseline / Pre-Challenge Asymptomatic (n=22) vs. Symptomatic (n=21)\_



#### **Training and Testing 18-Gene Classifier Using LDA-based PAM Method**

#### Top 673 unique significant SAM genes (q-value <25%) are used to build the predictor



Baseliı	ne / Prech	allenge	(Training)_			0.1T				0.8T	
True\Pred	Asymp	Symp	<b>Detection Rate</b>	True\Pred	Asymp	Symp	Detection Rate	True\Pred	Asymp	Symp	<b>Detection Rate</b>
Asymp	22	0	100%	Asymp	9	1	90%	Asymp	9	1	90%
Symp	1	20	95.2%	Symp	1	9	90%	Symp	1	8	88.9%

Diagnosis of multiple cancer types by shrunken centroids of gene expression. Robert Tibshirani, Trevor Hastie. Balasubramanian Narasimhan. and Gilbert Chu

#### **Case-by-case Prediction Results Report at 0.1T and 0.8T Milestone**

0.1T	Subject	Hr	Pheno	Predicted	Posterior Asymp	Posterior Symp	0.8T	Subject	Hr	Pheno	Predicted	Posterior Asymp	Posterior Symp
C01H04	C01	4	А	А	75.2%	24.8%	C01H48	C01	48	А	А	75.9%	24.1%
C02H04	C02	4	А	А	72.6%	27.4%	C02H48	C02	48	А	А	73.1%	26.9%
C05H04	C05	4	А	А	58.2%	41.8%	C05H48	C05	48	А	А	52.1%	47.9%
C08H04	C08	4	А	А	62.3%	37.7%	C08H48	C08	48	А	А	55.5%	44.5%
C10H08	C10	8	А	А	62.2%	37.8%	C10H48	C10	48	А	А	66.7%	33.3%
C12H04	C12	4	А	А	62.6%	37.4%	C12H48	C12	48	А	А	70.6%	29.4%
С13Н04	C13	4	А	А	57.3%	42.7%	С13Н48	C13	48	А	А	67.7%	32.3%
C14H04	C14	4	Α	8	42.3%	57.7%	C14H48	C14	48	Α	S	48.2%	51.8%
C17H04	C17	4	А	А	62.8%	37.2%	C17H48	C17	48	А	А	64.2%	35.8%
C18H04	C18	4	А	А	50.3%	49.7%	C18H48	C18	48	А	А	56.3%	43.7%
С03Н04	C03	4	S	S	28.8%	71.2%	C03H48	C03	48	S	S	29.2%	70.8%
C04H08	C04	8	S	S	39.1%	60.9%	C04H48	C04	48	S	S	41.6%	58.4%
С06Н04	C06	4	S	Α	51.5%	48.5%	C06H30	C06	30	S	S	47.9%	52.1%
C07H04	C07	4	S	S	41.1%	58.9%	C07H42	C07	42	S	S	34.5%	65.5%
C09H08	C09	8	S	S	42.8%	57.2%	C09H48	<b>C09</b>	48	S	Α	50.8%	49.2%
C11H04	C11	4	S	S	27.5%	72.5%	C11H48	C11	48	S	S	19.8%	80.2%
C15H04	C15	4	S	S	43.7%	56.3%	C15H48	C15	48	S	S	42.3%	57.7%
C16H04	C16	4	S	S	31.4%	68.6%	C19H96	C19	96	S	S	24.1%	75.9%
C19H04	C19	4	S	S	35.2%	64.8%	C20H42	C20	42	S	S	41.6%	58.4%
C20H04	C20	4	S	S	31.8%	68.2%	Subject	16 Hrs42	chip	has qual	lity issues.		

#### Table 1: List of 18-Gene Predictor

Probeset	Symbol	Cytoband	Score	Location	Туре
213478_at	KIAA1026	chr1p36.21	-3.234	Plasma Membrane	other
209396_s_at	CHI3L1	chr1q32.1	-3.221	Extracellular Space	enzyme
205033_s_at	DEFA1	chr8p23.1	-3.005	Extracellular Space	other
205040_at	ORM1	chr9q31-q32	-2.994	Extracellular Space	other
203936_s_at	MMP9	chr20q11.2-q13.1	-3.015	Extracellular Space	peptidase
207269_at	DEFA4	chr8p23	-2.802	Extracellular Space	other
203691_at	PI3	chr20q12-q13	-2.879	Extracellular Space	other
219594_at	NINJ2	chr12p13	-3.140	Plasma Membrane	other
205844_at	VNN1	chr6q23-q24	-2.762	Plasma Membrane	enzyme
221491_x_at	hCG_1998957	chr6p21.3	2.538	Plasma Membrane	transmembrane receptor
218272_at	FLJ20699	chr22q13	3.167	Unknown	other
202869_at	OAS1	chr12q24.1	2.992	Cytoplasm	enzyme
209433_s_at	PPAT	chr4q12	2.996	Cytoplasm	enzyme
203290_at	HLA-DQA1	chr6p21.3	2.647	Plasma Membrane	transmembrane receptor
218638_s_at	SPON2	chr4p16.3	3.039	Extracellular Space	other
212070_at	GPR56	chr16q12.2-q21	3.008	Plasma Membrane	G-protein coupled receptor
204439_at	IFI44L	chr1p31.1	3.037	Unknown	other
219269_at	HMBOX1	chr8p21.1	3.588	Nucleus	transcription regulator

Up-regulated genes in symptomatic cases Down-regulated genes in symptomatic cases

#### Table 1 (Cont'd): Function of 18-Gene Predictor

Probeset	Symbol	Gene Title	Relevant Function
213478_at	KIAA1026	kazrin	Cellular development and organ morphology
209396_s_at	CHI3L1	chitinase 3-like 1 (cartilage glycoprotein-39)_	A tiny variation in a gene known as CHI3L1 increases susceptibility to asthma, bronchial hyperresponsiveness and decline in lung function (NEJM Apr 2008)_
205033_s_at	DEFA1	defensin, alpha 1	Innate immune response to pathogen
205040_at	ORM1	orosomucoid 1	Inflammation modulating molecules
203936_s_at	MMP9	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)_	Regulator of cellular migration, likely plays a role in recruitment of leukocytes other than chemokines
207269_at	DEFA4	defensin, alpha 4, corticostatin	Innate immune response to pathogen
203691_at	PI3	peptidase inhibitor 3, skin-derived (SKALP)_	Elafin reduced the activation of inflammatory TF NFKB (anti-inflammatory effect)_
219594_at	NINJ2	ninjurin 2	Cell adhesion
205844_at	VNN1	vanin 1 (Pantetheinase precursor)_	homolog to mouse VNN1, involved in lymphocyte migration in cell adhesion and in the colonization of the thymus by hematopoietic precursor cells
221491_x_at	hCG_19989 57	major histocompatibility complex, class II, DR beta 1	Antigen presentation
218272_at	FLJ20699	hypothetical protein FLJ20699	
202869_at	OAS1	2',5'-oligoadenylate synthetase 1, 40/46kDa	Interferon signaling (inhibition of virus replication)_
209433_s_at	PPAT	phosphoribosyl pyrophosphate amidotransferase	
203290_at	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	Antigen presentation
218638_s_at	SPON2	spondin 2, extracellular matrix protein	Critical in antigen presentation and initiation of innate immune response
212070_at	GPR56	G protein-coupled receptor 56	Signal transduction
204439_at	IFI44L	interferon-induced protein 44-like	
219269_at	HMBOX1	homeobox containing 1	Transcription repressor

### Functional Analysis Using Ingenuity Reveals Major Involvement by 18 Predictor Genes in Canonical Inflammatory and Infectious Disease Pathway



to RSV infection

Another PA inhibitor SERPINE2 (PN-1) is significantly

#### Legend of Pathway Analysis Diagram

Relationships

#### Network Shapes Chemical or Drug Cytokine Enzyme G-protein Coupled Receptor Group or Complex Growth Factor Ion Channel $\langle \rangle$ Kinase Ligand-dependent Nuclear Receptor Peptidase Phosphatase Transcription Regulator Translation Regulator Transmembrane Receptor Transporter

Other

Α B binding only A В inhibits в acts on A B inhibits AND acts on (A В leads to (A в translocates to (A В reaction catalysis enzyme (A reaction direct interaction indirect interaction

Note: "Acts on" and "inhibits" edges may also include a binding event.

#### **Time Series Analysis Using EDGE ---- Two Densely Sampled Subjects**











Pathway Analysis of 41 Top Significant Genes (*q*-value < .5%)

\* 15 out of 41 genes are directly related to immune response (colored in pink to red depending on its significance)

#### Pathway Analysis of 41 Top Significant Genes (*q*-value < .5%) Reveals Direct Involvement in Cancer or Organismal Injury and Abnormalities



\* 18 out of 43 genes are directly related to cancer or organismal injury and abnormalities (colored in pink to red depending on its significance)

#### Time Course Analysis Using 5 Samples Per Individual at BL/0.1T/0.8T/T





**Expression Pattern of 7 Predictor Genes Whose Time Course** 

1. Inside each box, 5 time points are shown: BL/T0/T0.1/T0.8/T

2. Individual 16 missing 42hrs and individual 14 missing 96hrs. data imputed using closest time point

### Expression Pattern of 7 Predictor Genes Whose Time Course Analysis Results Are Also Significant (*q*-value <25%)\_



1. Inside each box, 5 time points are shown: BL/T0/T0.1/T0.8/T

2. Individual 16 missing 42hrs and individual 14 missing 96hrs. data imputed using closest time point



#### **Expression Pattern of 7 Predictor Genes Whose Time Course**

Analysis Results Are Also Significant (*q*-value <25%)

1. Inside each box, 5 time points are shown: BL/T0/T0.1/T0.8/T

2. Individual 16 missing 42hrs and individual 14 missing 96hrs, data imputed using closest time point



- 1. Inside each box, 5 time points are shown: BL/T0/T0.1/T0.8/T
- 2. Individual 16 missing 42hrs and individual 14 missing 96hrs, data imputed using closest time point

#### Major Canonical Signaling Pathways and Metabolic Pathways Involved by Top Significant Time Course Genes (q-value<5%)\_



Ratio

Pathway	Significant Genes
Interferon Signaling	IFIT1,IFIT3,OAS1,IFITM1,IFI35,MX1,PSMB8
Complement System	CFD,CD59,SERPING1,C1QA,C1QB,C3AR1,C2
Glycosphingolipid Biosynthesis - Globoseries	NAGA,GLA,GM2A,HEXB
IL-10 Signaling	CCR1,FCGR2C,MAPK14,BLVRA,IL1B,CD14
Aminosugars Metabolism	NAGK,GM2A,HEXB,HK3,NANS
Antigen Presentation Pathway	PSMB9,HLA-DRA,PSMB8,MR1

#### Major Disease and Disorders Involved by Top

#### Significant Time Course Genes (q-value<5%)



#### Canonical Pathway Involvement by Significant Genes: Cellular Growth and Proliferation / Organism Injury



© 2000-2008 Ingenuity Systems, Inc. All rights reserved.

### Canonical Pathway Involvement by Significant Genes: Immunological Diseases



### Canonical Pathway Involvement by Significant Genes: Immunological Diseases



### Canonical Pathway Involvement by Significant Genes: Inflammatory Diseases



### Random Forest Analysis

- Random Forests: Ensembles of tree based classifiers using lasso selection (Breiman 1990).
- Ys=a1\*f1(Xs) +...+an\*fn(Xs) + I1(a1...an)
- For classification Y is symptom 3-level sx of each chip
- For prediction Y is the end state (sx or asx) of each chip
- RF trained on Y and X=genes, metabolites, AA/AC and immunoassay data over days 0-3).
- Stratified sampling used to compensate for unbalanced groups

### Forward Prediction at 0.2T



0
0
·····
0
0
0
0
-
<u>_</u>
0
0
0
0
0
0
O
0
0
0
0
<u> </u>
0
0
0
 O
0

### Forward Prediction at 0.8T

\_\_\_\_\_

0
0
0 
0
0
0
0
0
0
0
0
0
-
2
0
-
0
0
0
0
0
0
_ D
0
0

------

	0
	0
	0
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
• • • • • • • • • • • • • • • • • • •	
····· 0	
0	
0	
0	
0	
0	
0	
0	

### **Observations from RF analysis**

- GO information embedding was applied to mRNA data prior to RF
- Perfect forward predictions of symptom onset from 0.2T and 0.8T.
- Biological validation is necessary (Gene, Metabolomic Ontology)\_
- mRNA is enhanced by other biomarkers, esp. immunoassay
- Current state estimation may allow the registration of patient state along their symptom/viral titer curve (see Extra slides)
- Cross-platform data normalization does not appear to add much value.

# DMDA

- Differential motif discovery and analysis (Hero:08) discriminates between graphical model for A and S interactions (gene-gene, gene-protein, proteinprotein)\_
  - Guiding principle: gene regulation network adapts in response to infection
- Method: detect consistent changes in loci of a small number of genes by comparing within-chip mRNA correlations in A and S groups
  - 1. Normalize mean and variance within each group at T0, 0.1T, 0.8T, T
  - 2. Sparse approximation to whole-chip (23Kx23K) covariance matrix
  - 3. Sparse matrix manipulation to detect pairs of "pivotal" mRNA probes with high correlations of opposite sign in A and S.

– 4. From these pairs construct interaction graphs under A and S
(A. Hero, report in preparation, 2008)

### Rhino-data at 0.8T DMDA -> IFIT5+SRGAP2 loci



Asymptomatic subjects

Symptomatic subjects

### Rhino-data at 0.2T PLEK,MMP11,TRIM3, ANKRD40 loci



#### **Functional Analysis of High Score Genes From DMDA Analysis**

prof.her<del>p motif</del> 3.5 3.0 (anlav-d)bol-2.0 1.5 1.0 0.5 0.0 ase ase Cell Death velopment Hematological System Development and Function Transport ase Connective Tissue elopment and Function Cellular Development Organismal Development Cancer Tissue Development Immune and Lymphatic System Development and Function Gene Expression Tissue Morphology Lipid Metabolism Small Molecule Biochemistry Cell Signaling elopmental Disorder Respiratory Disease Post-Translational Modification Repair Amino Acid Metabolism System Development and Function and Skeletal and Muscular System Development and Function Cellular Growth and Proliferation Neurological Dise Dise stern Dise Dermatological Diseases Conditions and Gastrointestinal Molecular Ō Damage š Embryonic ø Reproductiv RNA Ó Nervous Dev

Analysis: prof.hero motif

#### Functional Analysis of High Score Genes From DMDA Analysis Pathway Related to Respiratory Disease

Network 5 : prof.hero motif : prof.hero.motif



© 2000-2008 Ingenuity Systems, Inc. All rights reserved

### Functional Analysis of High Score Genes From DMDA Analysis Top Rank #1 Pathway Related to Tissue Development



© 2000-2008 Ingenuity Systems, Inc. All rights reserved.

### Path Forward

- Integration/completion of all analyses
  - CV, integrate results, include missing chips
- Integrated pathway analysis (genetic, metabolic)\_
- Identify stable targets for on-chip diagnostic testing
- Refine and validate results on new mRNA chips
- Further refine for cohorts in Challenge II and III

### Extras

RF predictor with Z-score normalization RF for current state estimation DMDA: Under the hood Sampling matrices and timing

# Forward prediction at 0.8T (scaled)\_

0
0
- -
~ 
· · · · · ·

CV error rate: 45%. Prediction accuracy 100%.

· · · · C

# Current state estimation (PAM+DMDA genes)

0
 С
0
0
 О
6
0
0
0
0
0
0
0
0
0
0
0
0
0
<u> </u>
0
- 0

			· · · · · · · · · · · · · · · · · · ·
		0	
		O	
		0	
		0	
		0	
	· · · · · O		
	····· 0··		
	n		
	- n		
Ő			
0			
0			
-			
0			
0			
0			
0			

### PAM+DMDA genes (scaled)\_

	0
	0
	0
	_
	0
	0
	0
٩ د	5
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
- 0	
0	
0	
0	
0	
=	

				0
				0
			0	
		·····	Ŭ	
		~		
		-		
		-		
		0		
	0			
	0			
	0			
	0			
	• • • • • • • • • • • • • • • • • • •			
	0			
	0			
	o			
	0			
	0			
	0			
	0			
 ŏ				
0				
0				
_				
 0				
 0				
 0				
		1		

# Current State Estimation (Metabolites)\_

0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
- 0
0

				· · · · · · c
			0	
		0		
		0		
	····· (			
	·····G			
	0			
	0			
	0			
	0			
	0			
	0			
	0			
	0			
	0			
	0			
	0			
	0			
	0			
	)			
0				
0				
0				
	1			

### Metabolite (scaled)\_

0
0
0
0
0
0
0
0
0
0
0
0
0
0
- 0
0
0
- -
0
0
6
~ 
- -
0
0

			·····
		····· 0 ·····	
		• O • • • • • • •	
	C		
	0		
	0		
	0		
	0		
	_		
	0		
	0		
o			
····· 0 ··			
0			
0			
0			
0			
0			
0			

### Current State Estimation (AA/AC)\_

							0
							0
						0	
					· · · · · · c		
					c		
					0		
					<b>.</b>		
				0			
			0				
			0				
			0				
			0				
			0				
			0				
		0.0					
		0					
	· · · · · · · · · · · · · · · · · · ·						
	0						
	0.00						
	0						
	0						
	0.00						
	0						
	0						
0							
0							

			0	
		0		
	o			
	0			
	0			
	0			
	0			
	0			
	0			
	0			
····· C	)			
0				
•••••				
0				
0				
0				
0				
-				
0				
~				
0				
0				
0				
0				

### AA/AC (scaled)\_

		0
		0
	0	
	• • • • • • • • • • • • • • • • • • • •	
	0	
	•••••	
	0	
	0	
0		
0		
0		
0		
-		
0		
0		
0		
0		
0		
0		
0		
0		
0		
0		
0		
0		
0		
0		

	0
	0
o	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
0	
o	

# Current State Estimation (Immunoassay)\_

		~	0
		_	
		0	
		0	
	0		
	·····		
	·····		
	,		
····· 0 ··			
0			
0			
o			
õ			
0			
-			
0			
0			
0			
0			
0			
·····			
<u> </u>			
õ			
-			
0			
0			
0			
0			



### Immunoassay (scaled)\_

	····· 0
	····· 0···
	· · · · · O · · · ·
	····· 0 · ····
	••••••
	0
	0
0	
0	
0	
····· 0	
0	
0	
0	
0	
<u> </u>	
0	
0	
0	
0	
0	
n	
. n	
_	
0	
0	

0
······O····
0
0
0
0
0
0
n n
0
O
0
0
O.
0
-
0
0
0
 O
0
0

### **Current state estimation**

	PAM+DMD A genes (12)_	Metabolites (5)_	AA (7)_	ImmunoAssay (9)_	All (15)_
Scaled	32.08	37.74	42.45	22.6	32.8
Unscaled	31.13	40.57	46.23	23.38	31.2

### DMDA - Under the Hood





### **Revised sampling protocol**

Subject	30-Nov	1- Dec Pre	1-Dec 4 hr	1-Dec 8 hr	1-Dec 12 hr	2-Dec 20 hr	2-Dec 24 hr	2-Dec 30 hr	2-Dec 36 hr	3-Dec 42 hr	3-Dec 48 hr	4-Dec	5-Dec	6-Dec	Sx Onset	T (hours)	N tubes
Symptoms																	
3 6 7 9* 10* 15 16 20	$\begin{array}{c} x \\ x $	X XX X X X X X X X	X X X X X X X	x xx x x	XX	xx	XX	XX	X XX	XX X X	x xx x x x x x x x x x x	x xx x x x x x x x x x x	x xx x x x x x x x x x x		12 noon 12/3 5 PM 12/3 12 noon 12/3 12/3/2008 15:00 12/4/2008 6:30 12/4/2008 6:30 12 noon 12/3 3 AM 12/2 7 AM 12/3	53 58 53 54 71 71 53 44 48	6 26 7 6 6 7 7 7
No Symptoms																	
1 5 8 11 12 13 14 17 18 19	× × ×× × × × × × × × × ×	X X XX X X X X X X X X	× × × × × × × × × ×	X X XX X X X X X X	XX	XX	XX	XX	XX	хх	x x xx x x x x x x x x x x x x	X X XX X X X X X X X X X	x x xx xx x x x x x x x x x x x x x x				7 7 26 7 7 7 7 7 7 7 7 7
* = marginal	sx																173

 $\bigcirc$ 

173



