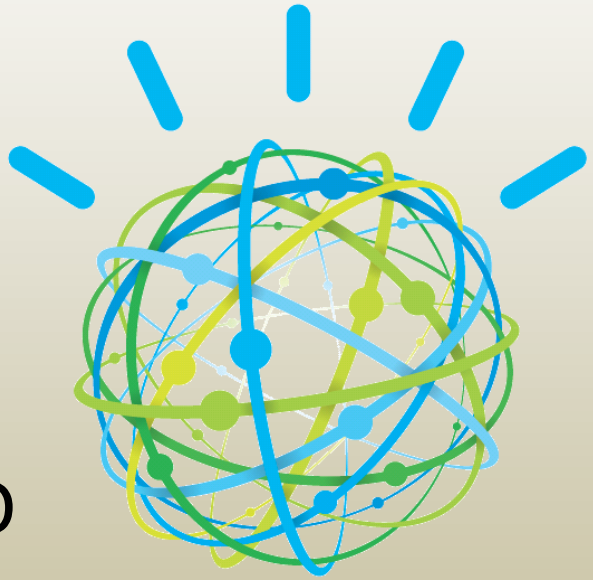
Using Data Analysis to Gain a Competitive Advantage in the Life Sciences



Gail K. Naughton PhD

8th Annual ACSESS Conference

March 11, 2011

Lots of Data to Analyze!

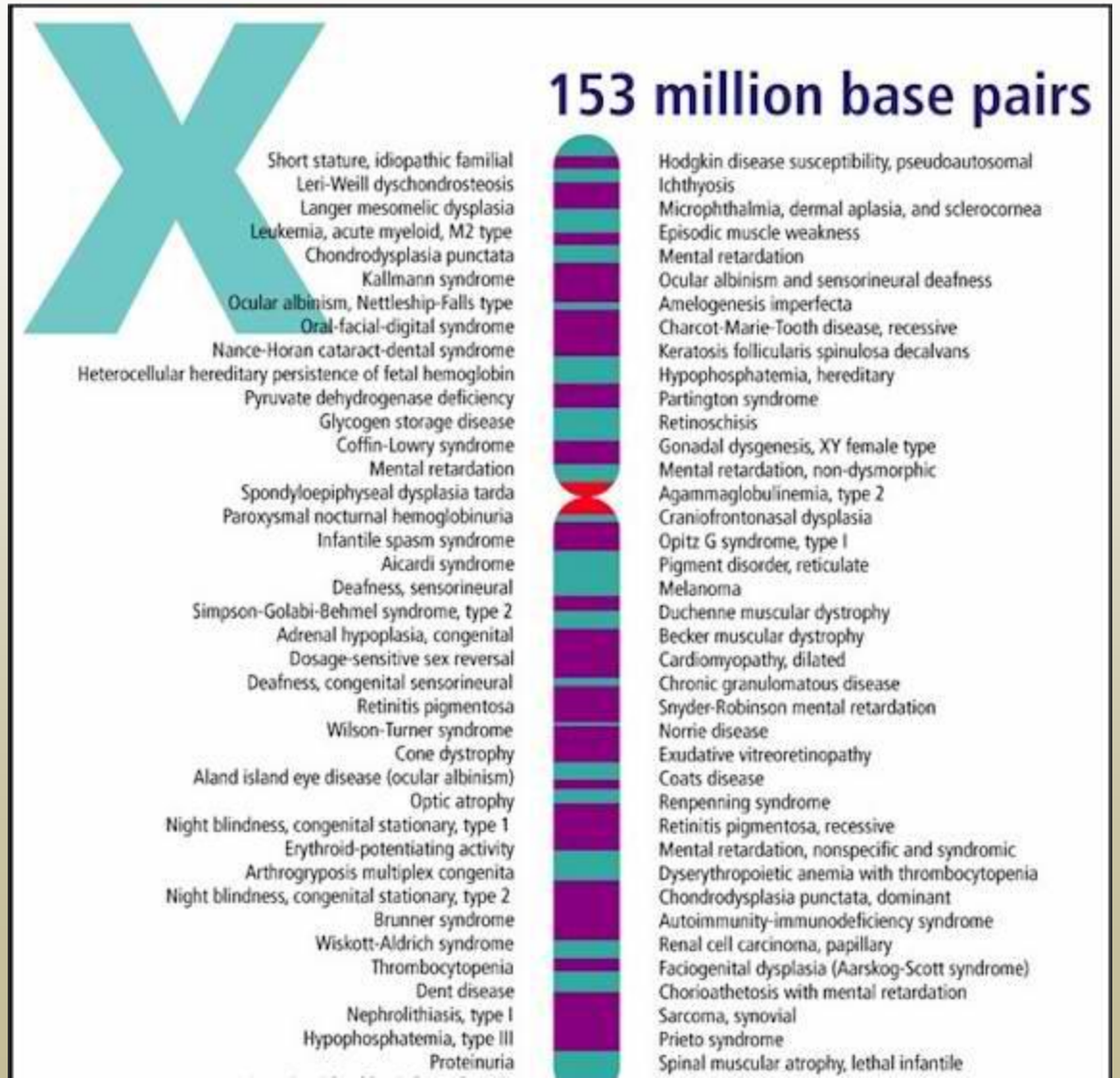
- Data intensive problems exist in every business, with more data being generated every second
- Great need for business analytics, data mining
- Industry specific

  -Insurance: predicting risk

  -Medical: predicting patient outcomes
- Progress in neural networks, mathematic modeling, and bioinformatics

Map of the human X chromosome

Assembly of the human genome is one of the greatest achievements of bioinformatics.



What is Bioinformatics?

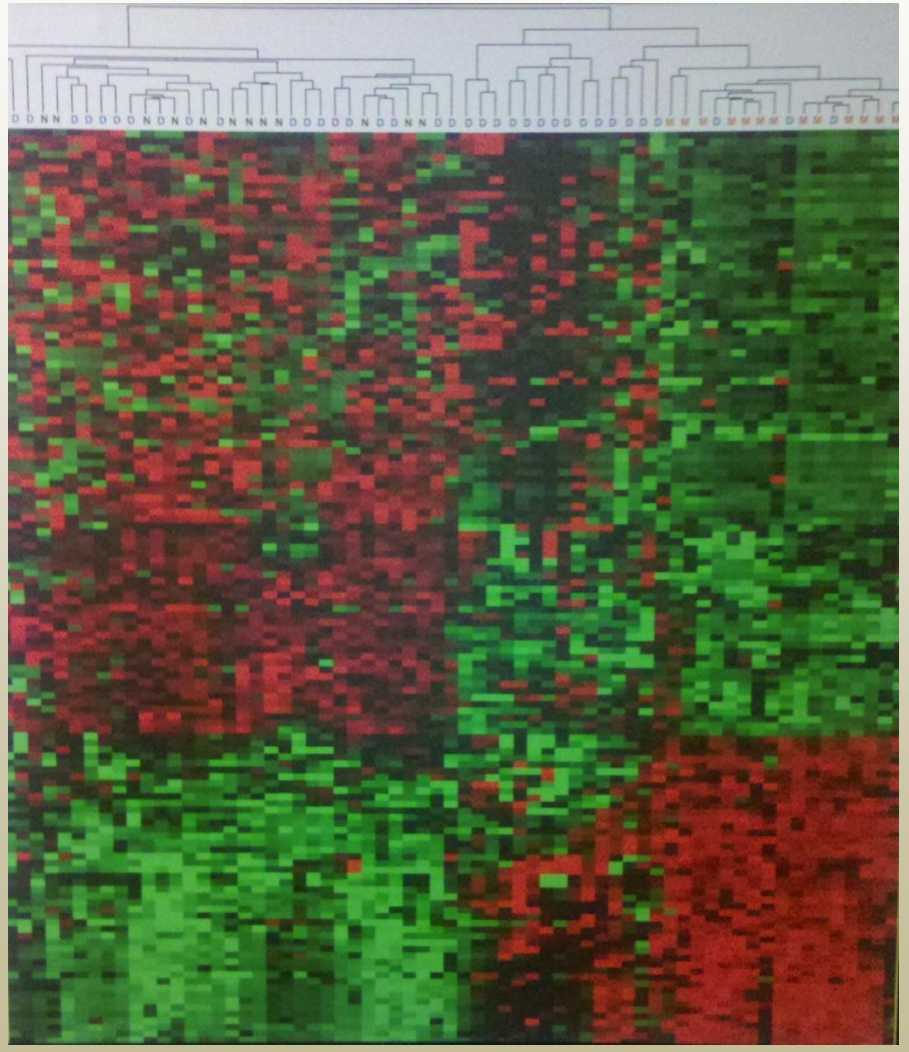
- Entails the creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data.
- Mapping and analyzing DNA and protein sequences
- Aligning different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures.

Bioinformatics

- Focus on developing and applying computationally intensive techniques pattern recognition, data mining, machine learning algorithms, and visualization
- Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression, and protein-protein interactions, genome-wide association studies and the modeling of evolution.

The Revolutionary Gene Chip

- Affymetrix led the way
- Shows what genes are upregulated/downregulated
- Tests for thousands of genes
- RNA data can be correlated with secreted proteins via ELISA and proteomics



rmat Tools Data Window Help

Type

60% Arial 12

33

B	C	D	E	F	G	H	I	J	K	L
Neural tubulin	0					Neural tubulin	Nervous Syster	#N/A	#N/A	#N/A
Neurofilament (NF)	#N/A	#N/A	#N/A	#N/A		Neurofilament (	Nervous Syster	#N/A	#N/A	#N/A
Neuron-specific enolase						Neuron-specific	Nervous Syster	#N/A	#N/A	#N/A
Neurosphere	0					Neurosphere	Nervous Syster	#N/A	#N/A	#N/A
Noggin	0					Noggin	Nervous Syster	#N/A	#N/A	#N/A
O4	0					O4	Nervous Syster	#N/A	#N/A	#N/A
O1	0					O1	Nervous Syster	#N/A	#N/A	#N/A
Synaptophysin	6855	SYP	NM_003179	SYP		Synaptophysin	Nervous Syster	0.9993333	3.47626	3.478579169
Tau	0					Tau	Nervous Syster	#N/A	#N/A	#N/A
						0	0	#N/A	#N/A	#N/A
rly Mesoderm Markers	0					Early Mesoderr	Early Mesoderr	#N/A	#N/A	#N/A
Gata 6	2627	GATA6	NM_005257	GATA6		Gata 6	Early Mesoderr	1.1057996	0.15071732	0.136297137
Brachyury	6862	#N/A	#N/A	#N/A		Brachyury	Early Mesoderr	#N/A	#N/A	#N/A
Tbx6	5634	PRPS2	BX456116	PRPS2		Tbx6	Early Mesoderr	1.1190424	0.3118011	0.27863207
MHOX	0					MHOX	Early Mesoderr	#N/A	#N/A	#N/A
Nodal	4838	NODAL	NM_018055	NODAL		Nodal	Early Mesoderr	1.0034579	3.1214073	3.11065098
CD31 (PECAM1)	5175	#N/A	#N/A	#N/A		CD31 (PECAM	Early Mesoderr	#N/A	#N/A	#N/A
CD325 (M-Cadherin)	0					CD325 (M-Cad	Early Mesoderr	#N/A	#N/A	#N/A
D34 (Mucosialin, gp 105-120)	947	#N/A	#N/A	#N/A		CD34 (Mucosia	Early Mesoderr	#N/A	#N/A	#N/A
NF-YA	4800	NFYA	AK025201	NFYA		NF-YA	Early Mesoderr	1.1875716	0.1982198	0.166911873
Sca-1 (Ly6A/E)	0					Sca-1 (Ly6A/E)	Early Mesoderr	#N/A	#N/A	#N/A
						0	0	#N/A	#N/A	#N/A
-MSC-Positive Markers	#N/A	#N/A	#N/A	#N/A		BM-MSC-Posi	BM-MSC-Posi	#N/A	#N/A	#N/A
CD 105	2022	#N/A	#N/A	#N/A		CD 105	BM-MSC-Posi	#N/A	#N/A	#N/A
CD 106	7412	VCAM1	NM_001078	VCAM1		CD 106	BM-MSC-Posi	0.27310297	0.995578	3.645430879
CD 120	7132	#N/A	#N/A	#N/A		CD 120	BM-MSC-Posi	#N/A	#N/A	#N/A
CD 124	3566	#N/A	#N/A	#N/A		CD 124	BM-MSC-Posi	#N/A	#N/A	#N/A
CD 29	3688	ITGB1	NM_133376	ITGB1		CD 29	BM-MSC-Posi	1.2972282	0.07294504	0.056231463
CD 44	960	CD44	BE764339	CD44		CD 44	BM-MSC-Posi	0.86515576	2.019786	2.334592328
CD 54	3833	#N/A	#N/A	#N/A		CD 54	BM-MSC-Posi	#N/A	#N/A	#N/A
CD 71	7037	TFRC	NM_003234	TFRC		CD 71	BM-MSC-Posi	0.99994075	0.09514309	0.095148728
CD 73	4907	#N/A	#N/A	#N/A		CD 73	BM-MSC-Posi	#N/A	#N/A	#N/A
CD 90	7070	THY1	#N/A	#N/A		CD 90	BM-MSC-Posi	#N/A	#N/A	#N/A
Collagen type I	1277	#N/A	#N/A	#N/A		Collagen type I	BM-MSC-Posi	#N/A	#N/A	#N/A
Fibronectin	2335	FN1	AA293788	FN1		Fibronectin	BM-MSC-Posi	1.1195177	3.387947	3.02625586
Integrin beta-1	0					Integrin beta-1	BM-MSC-Posi	#N/A	#N/A	#N/A
Sca1	0					Sca1	BM-MSC-Posi	#N/A	#N/A	#N/A
SH 2	0					SH 2	BM-MSC-Posi	#N/A	#N/A	#N/A
SH 3	0					SH 3	BM-MSC-Posi	#N/A	#N/A	#N/A
CD10	4311	MME	NM_007289	MME		CD10	BM-MSC-Posi	0.999779	0.08908006	0.089099751
CD105 (Endoglin)	2022	ENG	#N/A	#N/A		CD105 (Endoglin	BM-MSC-Posi	#N/A	#N/A	#N/A
CD106 (VCAM-1)	7412	VCAM1	NM_001078	VCAM1		CD106 (VCAM-	BM-MSC-Posi	0.27310297	0.995578	3.645430879
CD117 (SCF R, c-kit)	3815	KIT	NM_000222	KIT		CD117 (SCF R,	BM-MSC-Posi	0.99988496	0.4625312	0.462584416
D120a (TNF Receptor Type I)	7132	TNFRSF1A	NM_001065	TNFRSF1A		CD120a (TNF F	BM-MSC-Posi	0.96589136	0.40460536	0.418893239

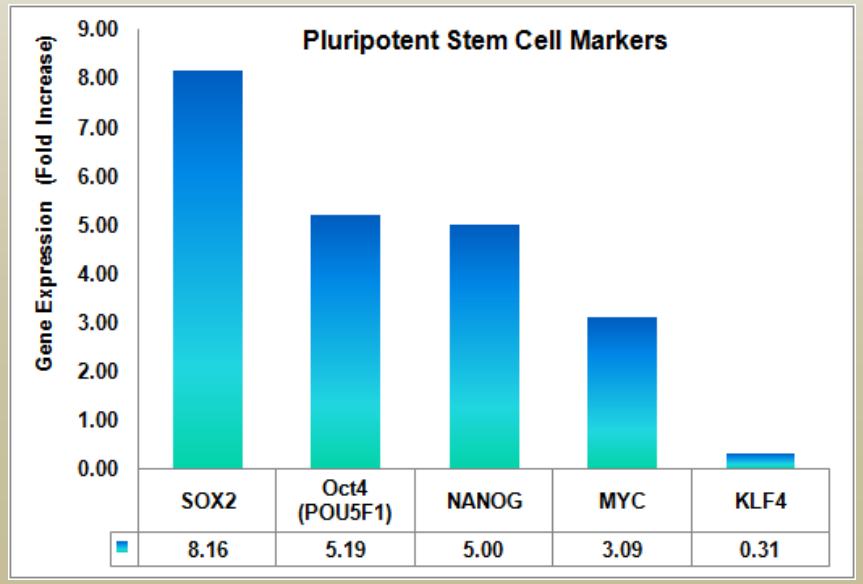
Graphs Sheet1

Internet Explorer ca...

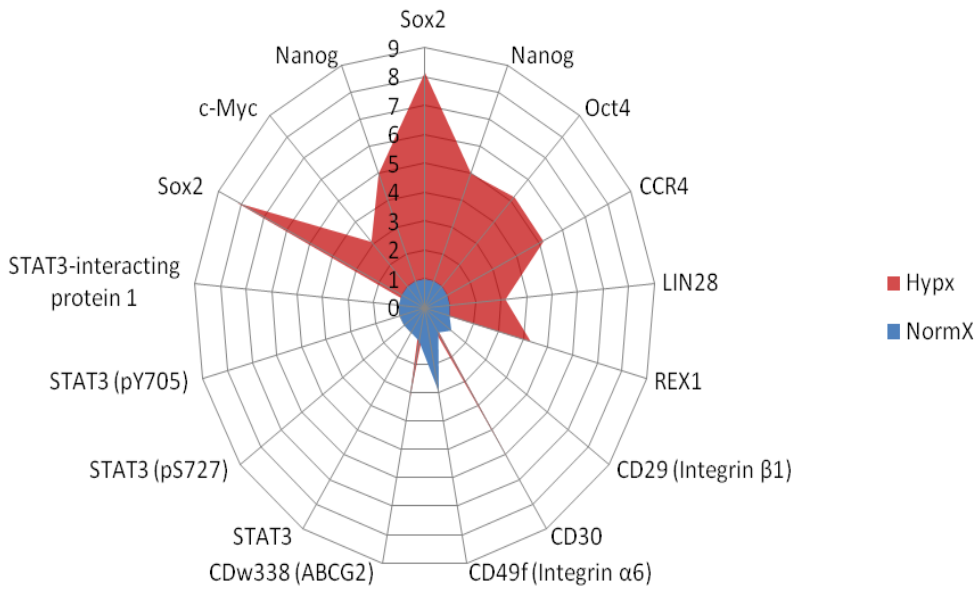
Effects of Hypoxia on Gene Expression

GENE	FOLD INCREASE	FOLD DECREASE
Wnt 4	5.94	
Wnt 7a	5.43	
Wnt 7b	4.05	
Wnt 2b	3.95	
Wnt 10a	3.86	
Wnt 8b	3.48	
Wnt 6	3.36	
Wnt 3a	3.19	
Wnt 9b	3.06	
Wnt 9a	3.02	
Wnt 11	2.89	
Wnt 5a		8.33
Wnt 2		7.14
Wnt 5b		5.26

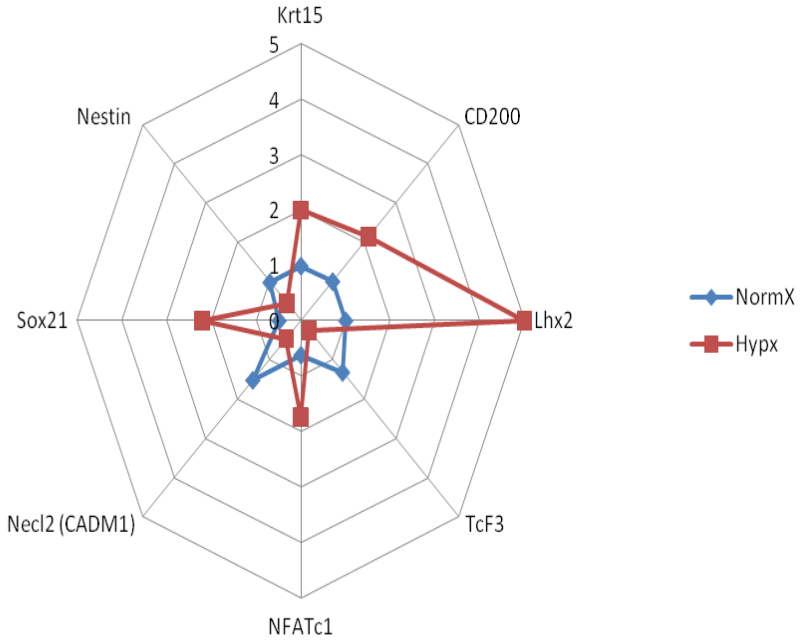
GENE	FOLD INCREASE
LAMG3	3.99
LAMA5	3.37
LAMG2	2.92
LAMA1	2.01
LAMB1	21.45
LAMG1	19.95
LAMA4	17.77
LAMA2	9.58
COL4A5	17.86
CXCL12	47.8
NID2	8.87
NID1	4.45
NOTCH2	15.7



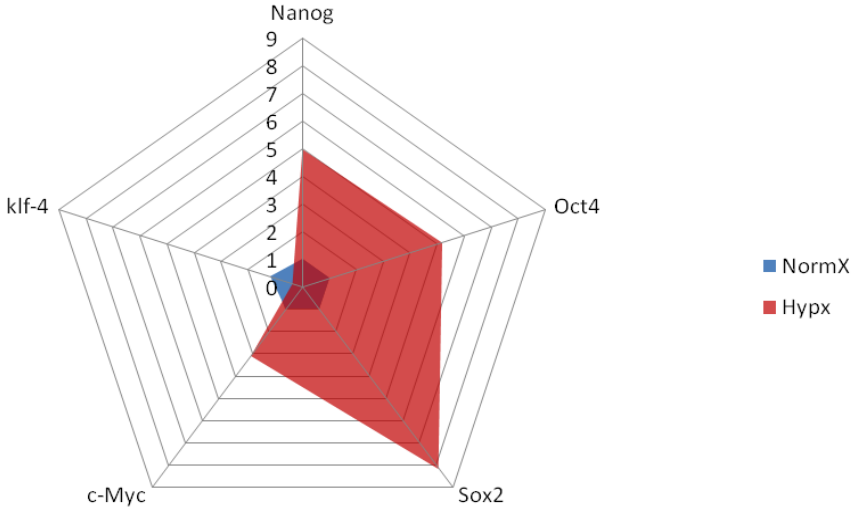
Embryonic Stem Cell Markers



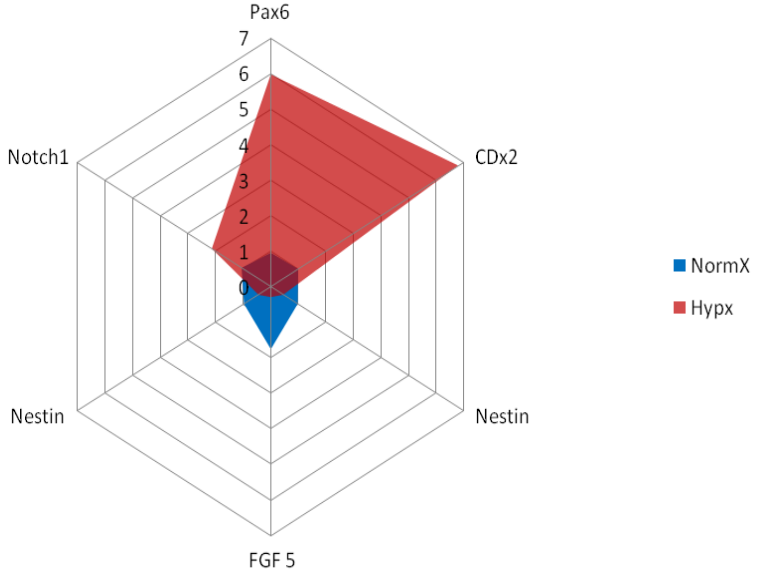
Hair Follicle Cells



Induced Pluripotent Genes



Early Ectoderm



New Technologies Can Improve Decision-Making & Trial Design

- Average cost of developing successful new drug exceeds $802 million, 8 years
- Molecular data can:

  -Help design better-powered trials, cut phase times & clinical costs

  - Identify failing candidates earlier

  - Improve clinical success rates

  - Enhance intellectual property position and/or marketing claims

RNA profiling by RT-PCR and microarray reveals biologically meaningful information on skin pathophysiology

Tracking Therapeutic Effect: Enbrel Study Results

Decreases in IFNγ, IL-12B, IL23A mRNA correlated with clinical improvement over time (trend seen with TNFα). Evidence that early molecular changes correlate with future clinical effect

Summary of correlation coefficients and p values for predictive comparisons between early change in gene expression and long-term clinical response.

	Time interval and correlation data			
	Week (0, 1) mRNA vs. Week (0, 8) NPF		Week (0, 4) mRNA vs. Week (0, 8) NPF	
Gene	Correlation coefficient: R	One-sided t-test	Correlation coefficient: R	One-sided t-test
TNFα	–	–	0.69	0.05 < P < 0.1
IFNγ	–	–	0.694	0.05 < P < 0.1
Krt-16	**-0.856**	**0.025 < P < 0.05**	–	–
IL-12B	–	–	0.796	0.05 < P < 0.1
IL-23A	**0.8**	**0.025 < P < 0.05**	**0.784**	**0.025 < P < 0.05**

New Era In Drug Development

Theranostics defines the development of diagnostic tests that identify patients most suitable for a drug and provide feedback as to how the drug is doing

"The utilization of diagnostic assays that go alongside treatment with Herceptin represents a dramatic success story in the area of pharmacogenetics and gives us some idea of what the process of parallel development of diagnostic assays and drugs will look like down the road."

Oren Cohen, M.D., CMO, CSO, QuintilesTransnational Corp

- HercepTest and PathVysion for the assessment of HER2 over expression for assessment of patients with metastatic breast cancer suited for Herceptin (Genentech);2003 sales $425M
- VentanaDx c-Kit for identification of patients to receive Gleevec for CML (Novartis); 2003 sales $1.13B

Companies in Theranostics

- Abbott Labs: Vysis, a genomic-disease-management company
- Genentech: HercepTest and PathVysion
  - Every therapeutic paired with a companion diagnostic
- Wyeth: Mylotarg for treatment of CD33 positive acute myeloid leukemia
- Roche diagnostic division
- Sanofi-Aventis: Enox test card for detection of anticoagulation activity of Lovenox (enoxaparin)
- ImClone: EGFR pharmDx to qualify patients for Erbitux

Powered by Affymetrix, Celera, Illumina, Life Technologies, and many others

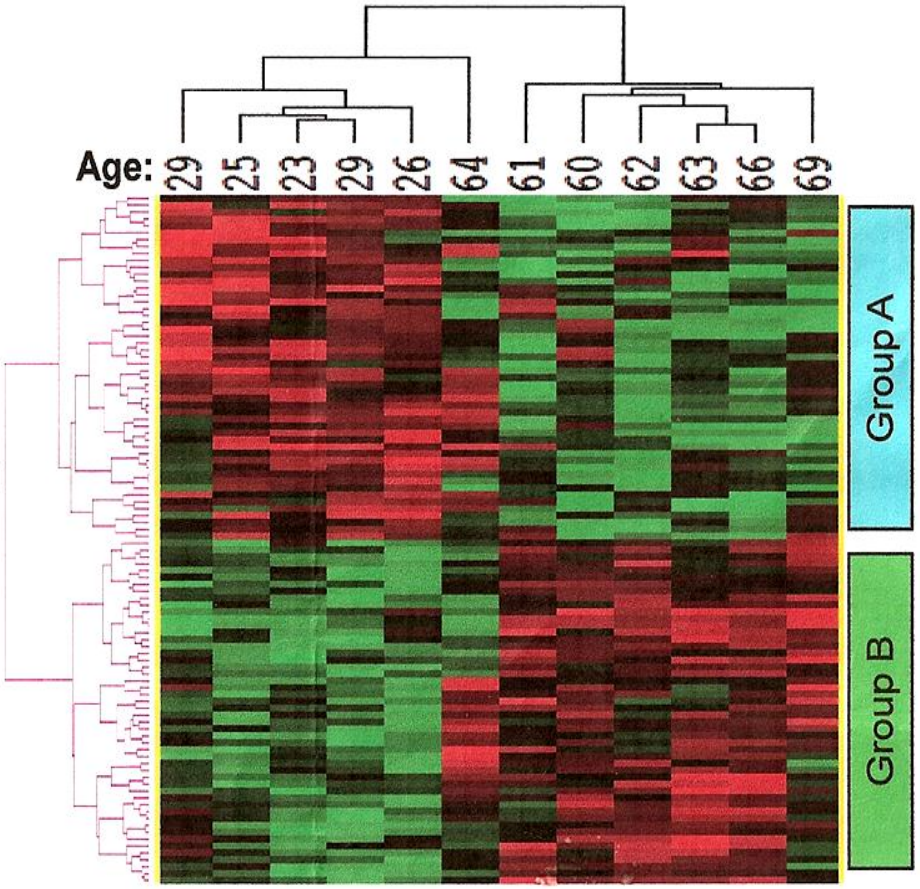
Gene Expression Profiles Distinguish "Youth" and "Old" Groups

GeneChip data analysis

- GCRMA normalization
- ~14,000 genes after eliminating very low expressed genes (expression level < 100 across all samples)
- t-test was performed to compare "young" (23 ~ 29) and "old" (60 ~ 69), ($p<0.05$)
  - Multiple testing correction
    - Benjamini and Hochberg
  - False discovery rate ($q < 0.05$)
  - Results in finding of 483 differentially expressed genes identified
- PAM (Prediction Analysis for Microarrays) used to rank these genes: between "young" and "old"
- 100 highest ranked genes selected
  - Hierarchical clustering analysis
  - "Young" (N=5)
  - "Old" (N= 7)
- Develop "Skin age index" to predict aging

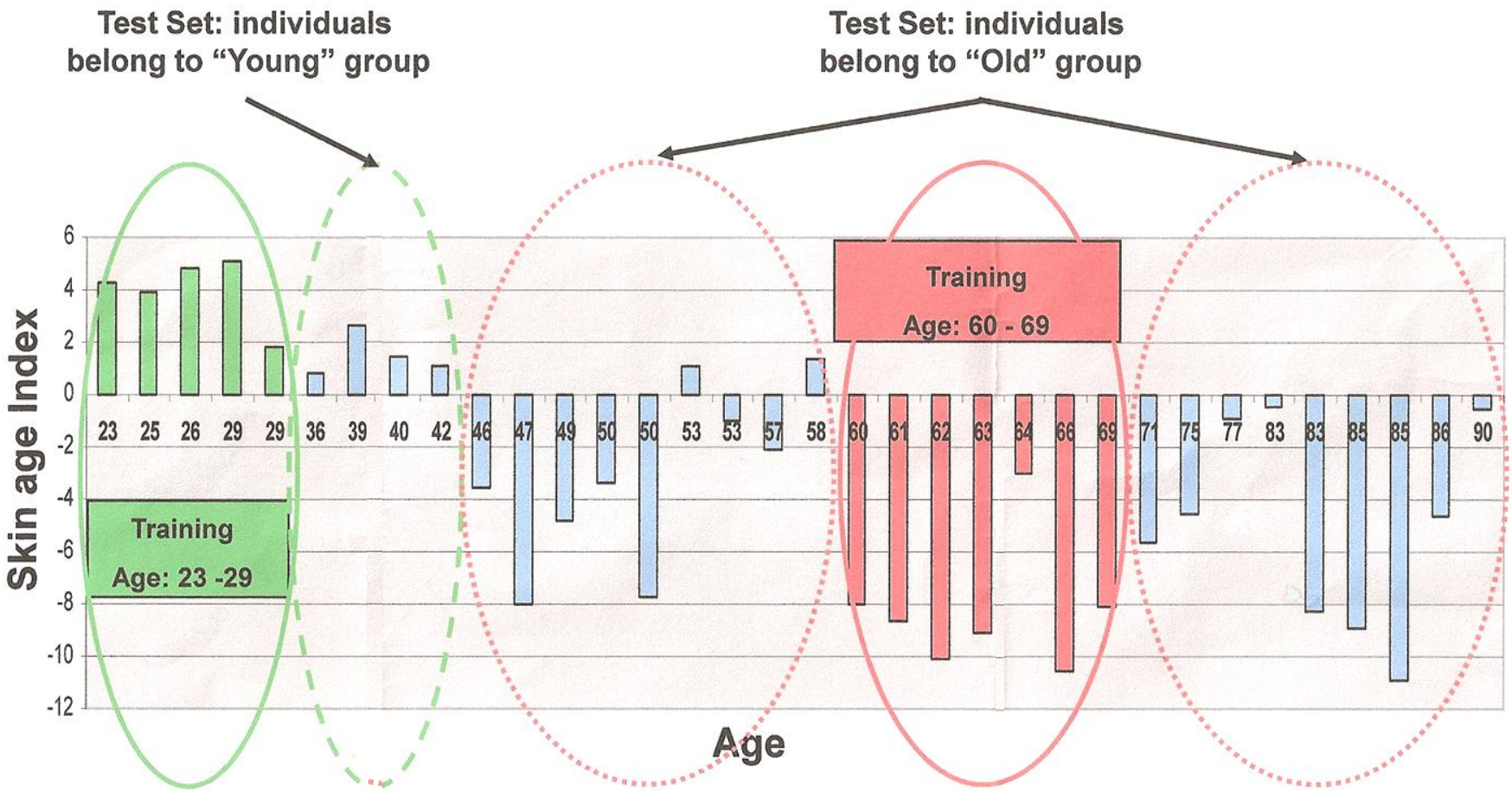
Preliminary Conclusion

- 100-gene classifier developed that discerned skin from "young" and "old" individuals



Skin age index = Sum of "Group A" – Sum of "group B" + α (constant)

100-Gene Classifier Distinguishes “Young” and “Old” Individuals



Knowledge Management
Gaining Greater Importance

- The Practice of selectively applying knowledge from previous experiences of decision making to current and future decision making activities with the express purpose of improving the organization's effectiveness
  - Decision making processes (Go/No GO)
  - Improving decisions (based on time, cost)
  - Knowledge in Action (touches every step)
- KM success is linked to improving organizational and individual performance
  - Improve effectiveness/margins
  - Improve product cycle times
  - Reduce rejects
  - Improve customer satisfaction

Organizations that invest in analytics are seeing the results

Organizations who are leading in analytics outperform those just beginning by 3X

Top performers are 5.4X more likely to use an analytic approach over intuition

, MIT Sloan Management Review /IBM Institute of Business Value study

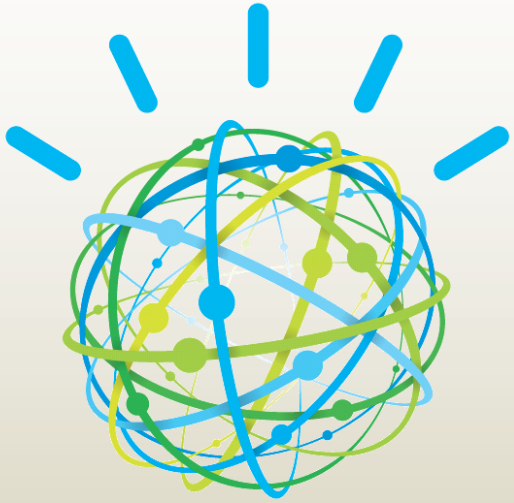
IBM

Watson answers a grand challenge

What is Watson?

- **Watson is a workload optimized computer system that battled Jeopardy's best human players**
- **A computing system built by a team of IBM scientists who set out to create a system that rivals a human's ability to answer questions posed in natural language with speed, accuracy and confidence**
- **A system that rapidly processes information to find precise answers to complex questions – which holds enormous potential for businesses**

What is Watson?



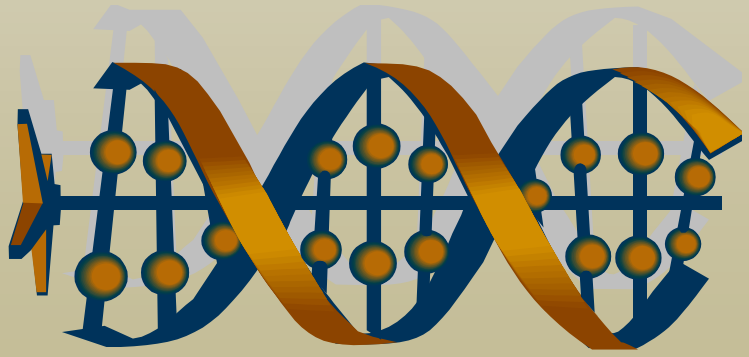
"Of course a lot more is at stake than just a game show victory…Researchers have their sights set on applying the technology in fields from health care to help desks…It's easy to see how such a breakthrough could be put to good use there, helping doctors to accurately diagnose patients' conditions by sifting through mountains of data in mere seconds."
USA Today, 1/14/11

A new paradigm in IT

- The computing paradigm for business has changed – new systems underlie every business process, and the hardware and software performance in those systems is closely tied to actual business performance
- Watson harnesses IBM's workload optimized POWER7 system which can process thousands of simultaneous tasks at rapid speeds
- Such technology can be adapted to diagnose disease, handle online technical support, parse vast tracts of legal documents and impact healthcare, government, transportation, and other industries

The future of business is in the data

- Watson-like analytics applied to business can provide answers with a confidence ranking that can be gleaned from both structured and unstructured data by running hundreds of different kinds of analytical queries across all different kinds of information
- Applying those innovations to an organization can help transform business models – meaningful insights from information can help anticipate and shape better business outcomes, improve business operations and boost service to customers



“Real Time” Data Analysis

- Computing underlies operations—from supply chain management, to human resources and payroll, to financial management, security and risk.
- Instrumented roadways, power grids, consumer goods and food—businesses need the ability to analyze the data coming from these sources
- Today’s businesses require the same kind of information consumption, advanced analytics and real-time response that is needed to answer questions on Jeopardy!

Computational Science and Engineering is Essential!

- Solving data intensive problems “real time”
- Bioinformatics rapidly growing in use
- Computational localization/targeting and robotics expanding
- Businesses of all types need reliable and fast analysis to help direct their business operations and strategy

YOU ARE THE SOLUTION!