

The National Center for Integrative Biomedical Informatics (NCIBI)

2nd Annual NCBC All Hands Meeting

Brian D. Athey
H.V. Jagadish
David J. States
Gilbert S. Omenn
Daniel L. Kiskis
Thomas A. Finholt
James D. Cavalcoli
Violet A. Elder
and NCIBI Team

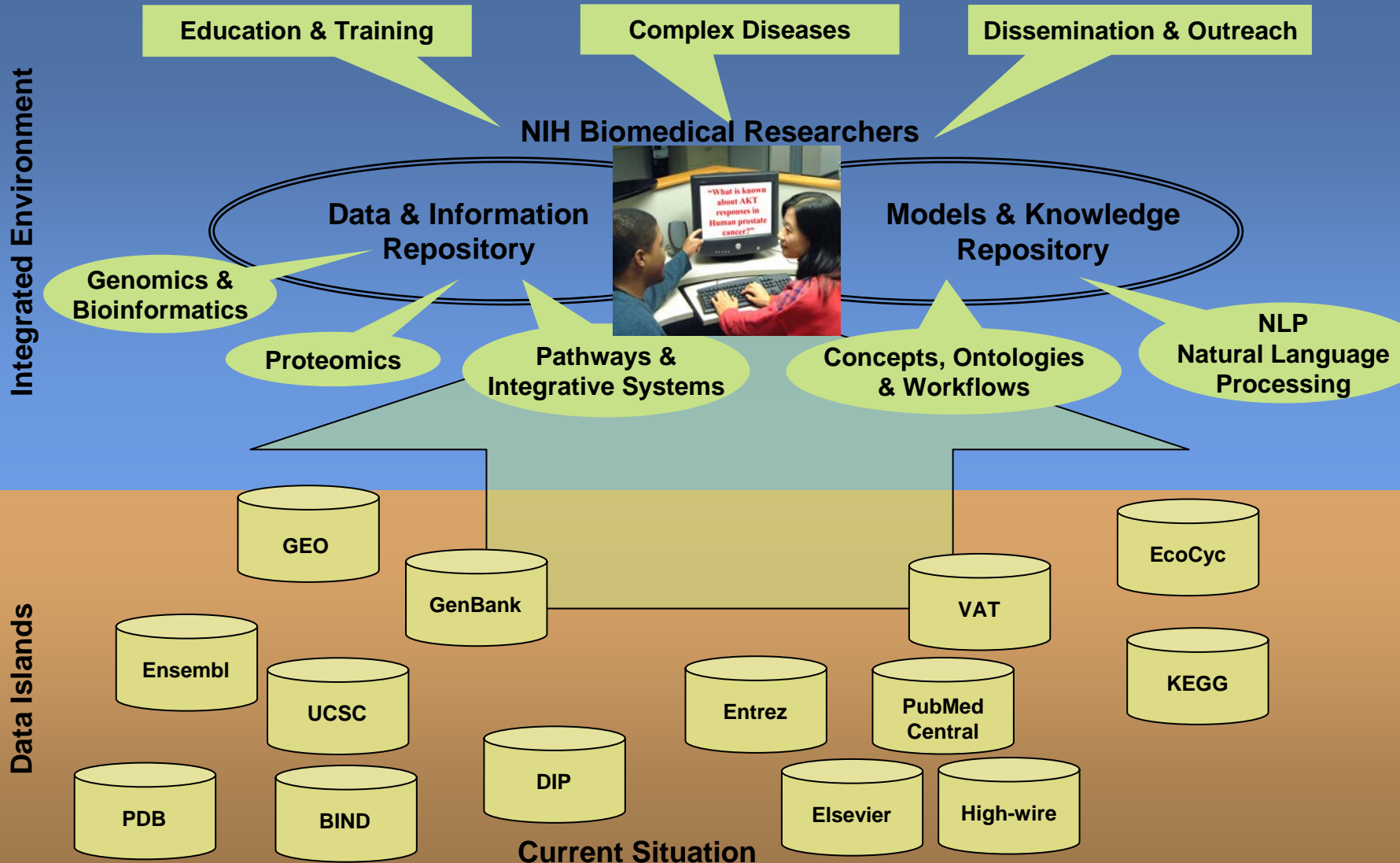


Outline

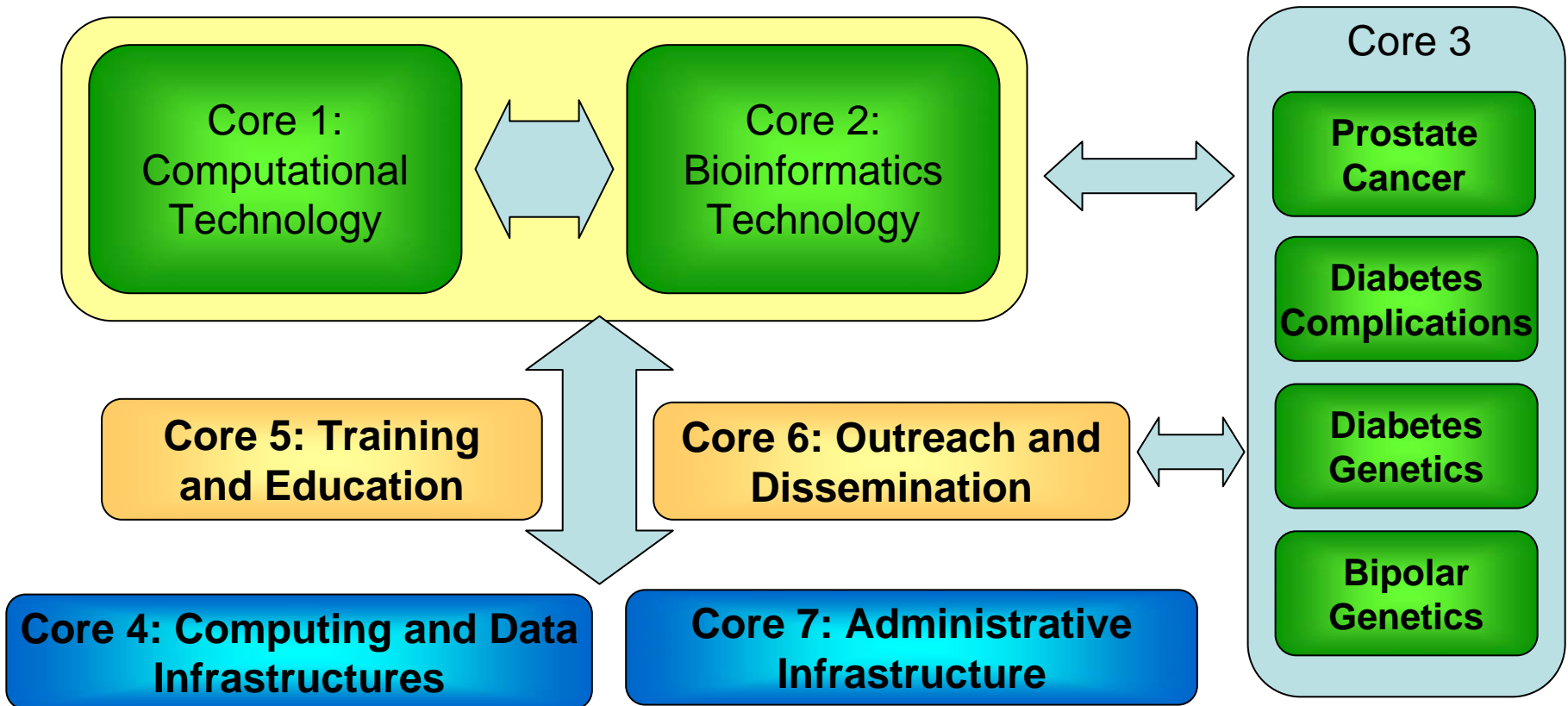
- NCIBI Vision, Structure and Team
- NCIBI Technologies: Cores 1 and 2
- NCIBI Central Focus--The Driving Biological Problems (DBPs): Core 3
- The Glue: Cores 4,5,6 and NCIBI Partners
- NCIBI “Hot Topics”—Clear Evidence of Early Success
- Scaling and Sustaining the Center and Building Bridges



Vision of the NIH National Center for Integrative Biomedical Informatics (NCIBI)



Simplistic Overview of NCIBI Structure



NCIBI Leadership Team Members



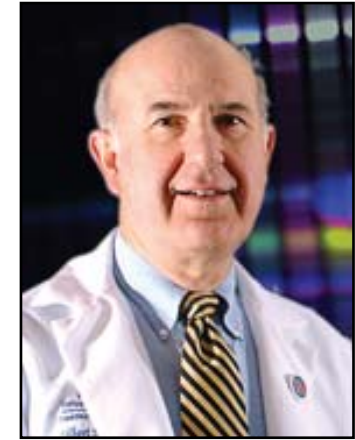
Brian D. Athey, Ph.D.
PI and Chair
Senior Scientific Director
Cores 5 & 6 Co-Director
Core 7



H.V. "Jag" Jagadish, Ph.D.
Senior Scientific Director
Core 1 Director



David J. States, M.D., Ph.D.
Senior Scientific Director
Cores 2 and 5 Co-Director



Gilbert S. Omenn M.D., Ph.D.
Senior Scientific Director
Core 3 Director



Daniel L. Kiskis, Ph.D.
Core 4  *Director*



Thomas A. Finholt, Ph.D.
Core 6 Co-Director



James D. Cavalcoli, Ph.D.
NCIBI Project Manager



Violet A. Elder, M.P.P.
Administration & Budget Director



Core 3: Driving Biological Projects and PIs

Arul Chinnaiyan, M.D., Ph.D.
Prostate Cancer
Oncomine



Michael Boenhke, Ph.D.
Diabetes T2 Genetics
FUSION PI



Melvin McInnis, M.D.
Bipolar Disorder Genetics

Eva Feldman, M.D., Ph.D.
Diabetes T1
Complications



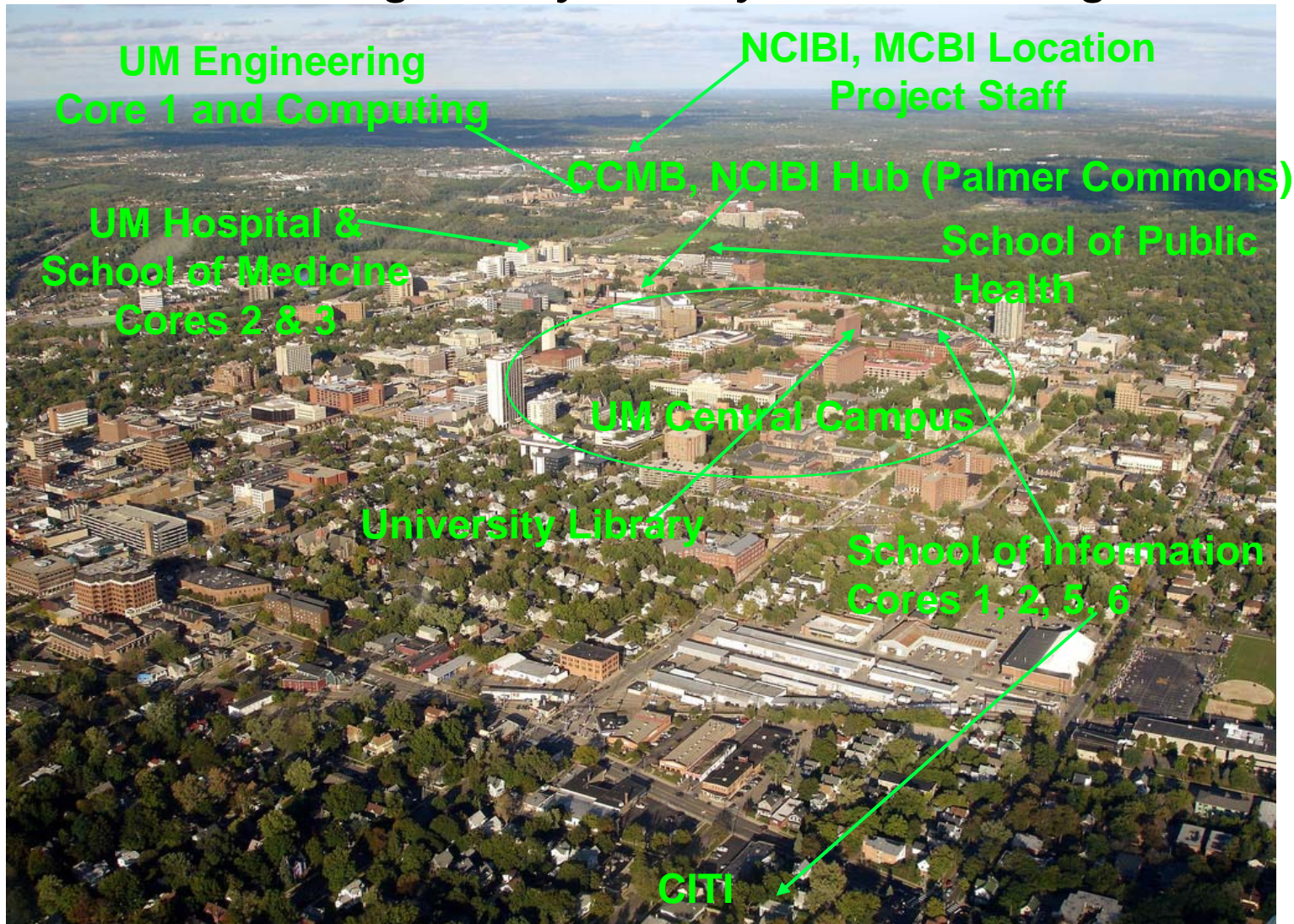
NCIBI Applications-Oriented Interdisciplinary Research (IDR) and Implementation Team

University of Michigan

- Clinical Applications and Basic Medical Sciences; Bioinformatics and Computational Biology: Medical School
- Computational Genomics and Population Genetics: School of Public Health
- Database and Algorithm Development, Architectures, and Machine Learning: College of Engineering, Several Departments
- Human-Computer Interfaces (HCI), User Requirements, and Evaluation; Natural Language Processing: School of Information and the Center for Information Technology Integration (CITI)
- Access to Full Text Literature for Natural Language Processing: University Library



The NCIBI Hub is Distributed Across the University of Michigan Campus *It is Functioning as a Hybrid Physical-Virtual Organization*



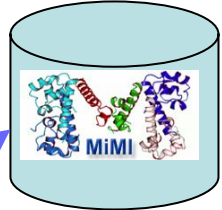
Granular Overview of NCIBI Activities Underway: Cores 1-3

User Environment, Evaluation & Feedback

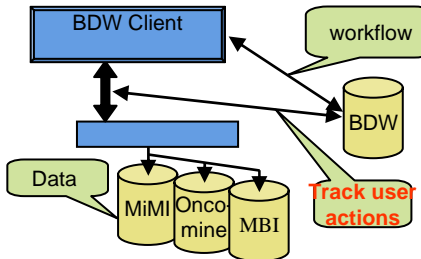
User Environment Analyses:
Mirel, Ackerman



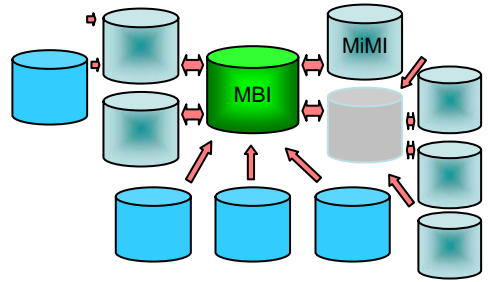
Database Technologies for Deep Integration of Biological Information



Platform Interoperability Project:
States and Athey



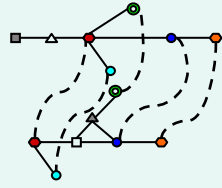
MiMI: Michigan Molecular Interactions DB:
Jagadish lab, Tarcea



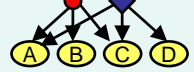
MBI: Molecular Biology Integration DB: States and Athey

Tools for large-Scale analysis

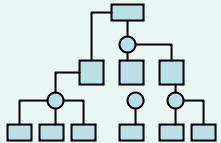
SAGA:
Patel Lab



Bayesian Nets:
Woolf Lab



Workflows and Interoperability:
NCIBI Programming Team; Broad Institute; MAGNet



Natural Language Processing: States, Meng, Radev Labs



Concept Mapper:
Rhodes, Patel, Woolf

Driving Biological Problems (Omenn)

Feldman Lab:
Diabetes Type 1, NRF2 signaling T1DM pathways

McInnis lab:
Bipolar Disorder Interactions Between Genetic Linkage Peaks, WNT signaling

Boehnke Lab:
Diabetes Type II SNP WGA workflow Genetic Interaction linkage workflow

Chinnaiyan and Omenn Labs:
Prostate Cancer Bayesian, NLP, Onco mine



Data and Information Repositories and Knowledge Bases

• Challenges

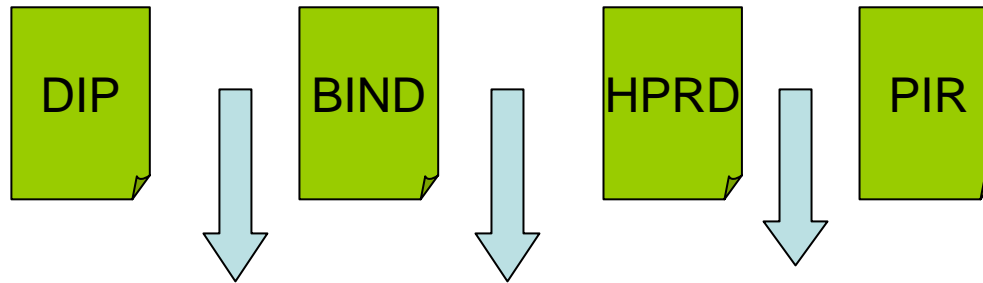
- Capturing and representing biological variation
 - Fundamental to genetic studies (Core 3c and 3d)
 - Variation at multiple levels (SNPs, alternative splicing, post translational modification, stochastic cellular processes, etc.)
- Ambiguous data
 - Many named entities (e.g. “PCR” phosphocreatine) are domain specific
 - Entities not well defined at a molecular level (“AKT”) are widely used
- The updates problem
- The Sheer mass of data and data sources
- Data merging from heterogeneous experimental and computational sources

• Value added

- Making connections is what science is all about
- Precompute linking paths greatly accelerates user work
- Quality assurance
 - Identify inconsistent data from input sources
 - Suggest data sources
 - Educate users about “what is out there”
 - Emphasize reliable sources



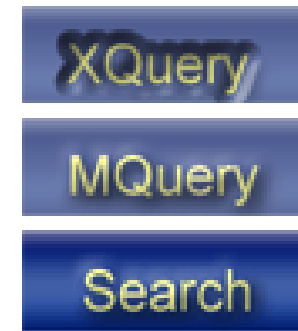
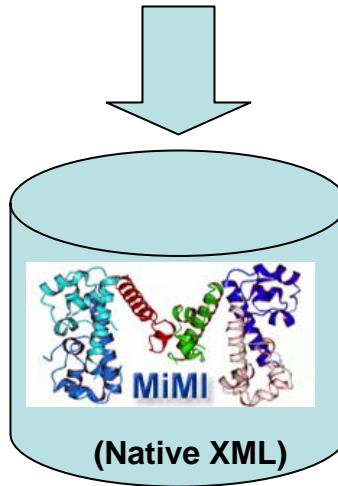
MiMI Data Sources, Deep Data Integration and Access



Deep Integration Techniques

Deep Integration Techniques:

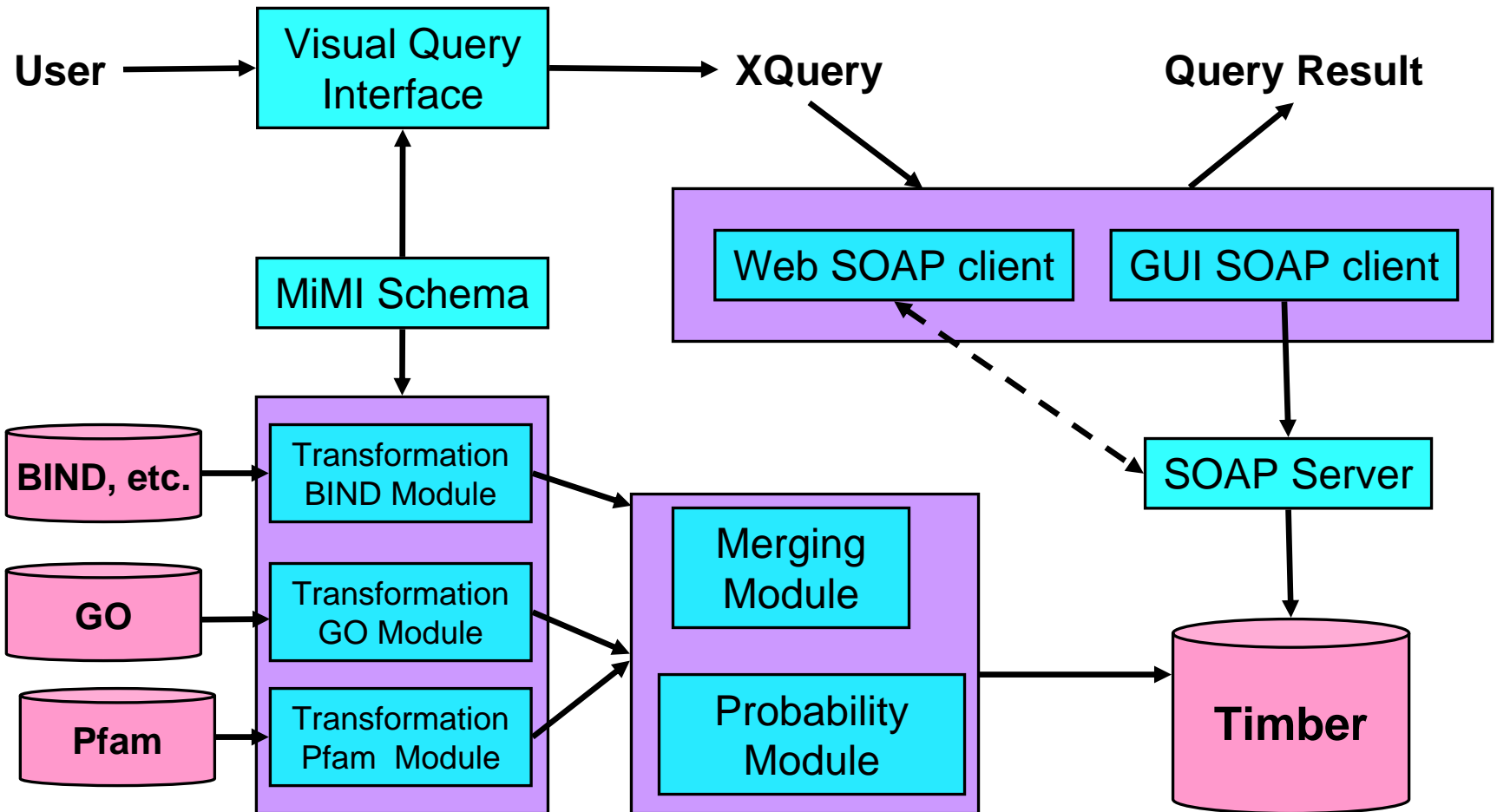
- Preserves provenance and conflicts
- Uniquely identify Proteins across multiple sources
- Integrates knowledge bases



MiMI is accessible to both expert and novice users by providing:

- Intelligent search & form-based queries
- XQuery

MiMi System Architecture



Sequence Alignment by Approximate Subgraph Alignment (SAGA): A Fast and Flexible Graph Matching Tool

- Motivation

- Large amount of biological graph data: e.g. KEGG, GenMAPP, BIND
- Graph database sizes are large and increasing in size
- Graph querying is a common requirement for many of the NCIBI DBPs
- Datasets are noisy/incomplete: so exact matching is not very useful

- Challenge: Graph Matching is a Hard Problem

- Even the simpler task of finding all subgraphs in the database that *exactly match* the query graph is NP complete
- Here we have a harder problem – *approximate* Subgraph Matching
 - Allow approximate matching of node/edge labels, and structural differences (e.g. allow node/edge deletion and addition)
 - A powerful mechanism for dealing with noise/partial information

- The database-centric SAGA approach

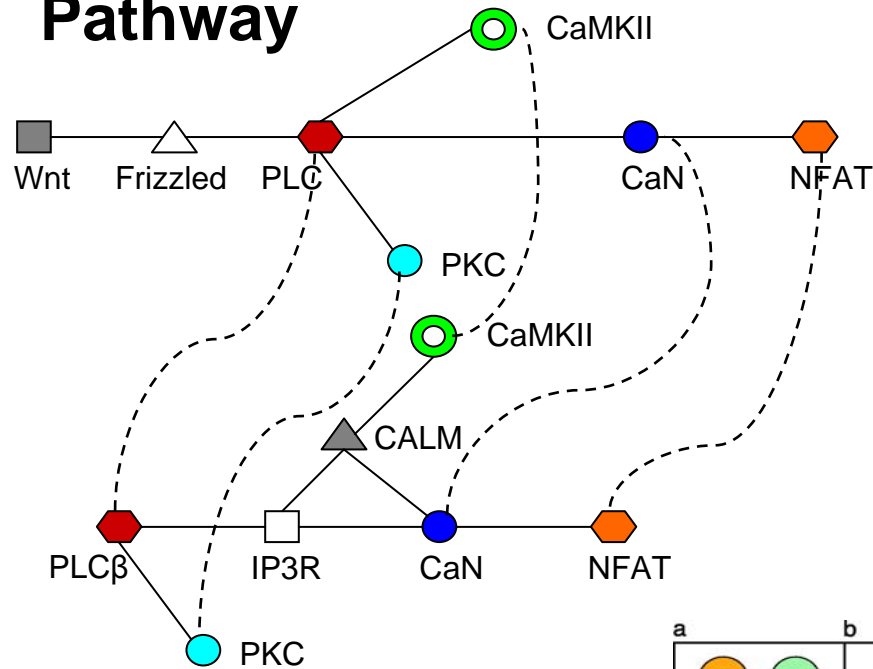
- Build an index on small graph substructures in the database
- Use the index to match fragments of the query with fragments in the database, allowing for various types of mismatches
- Assemble larger matches using a graph clique detection algorithm



SAGA Results: Query KEGG with Wnt/CA2+ Signaling Pathway

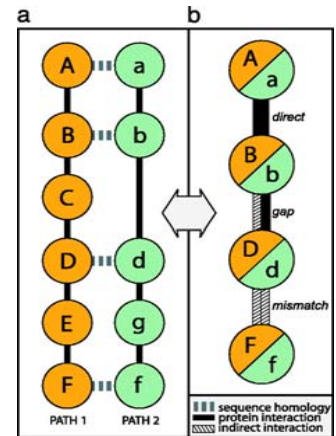
Query: Wnt/Ca2+ Signaling
KEGG id: 04310hsa

Match: Calcium Signaling
KEGG id: 04020hsa



Limitations of Existing Methods:

- Gindex & GraphGrep: only perform exact matching
- Grafil & PIS: no gap nodes are allowed
- PathBlast: only matches paths; edge alignment only tolerates one gap nodes, e.g. (B,D) with (b,d) and (D, F) with (d, f)
- None of the existing methods can detect this match



Kelley et. al. PNAS(2003)
PathBlast Example



MBI: Molecular Biology Integration Database

Applications

- Support for cross-domain global analysis (genomics, proteomics, metabolomics, networks)
- Central, stable molecular sequence repository
- Reference point for sequence links to literature
- Sequence history and changes tracking
- Curation of highly similar sequence groups

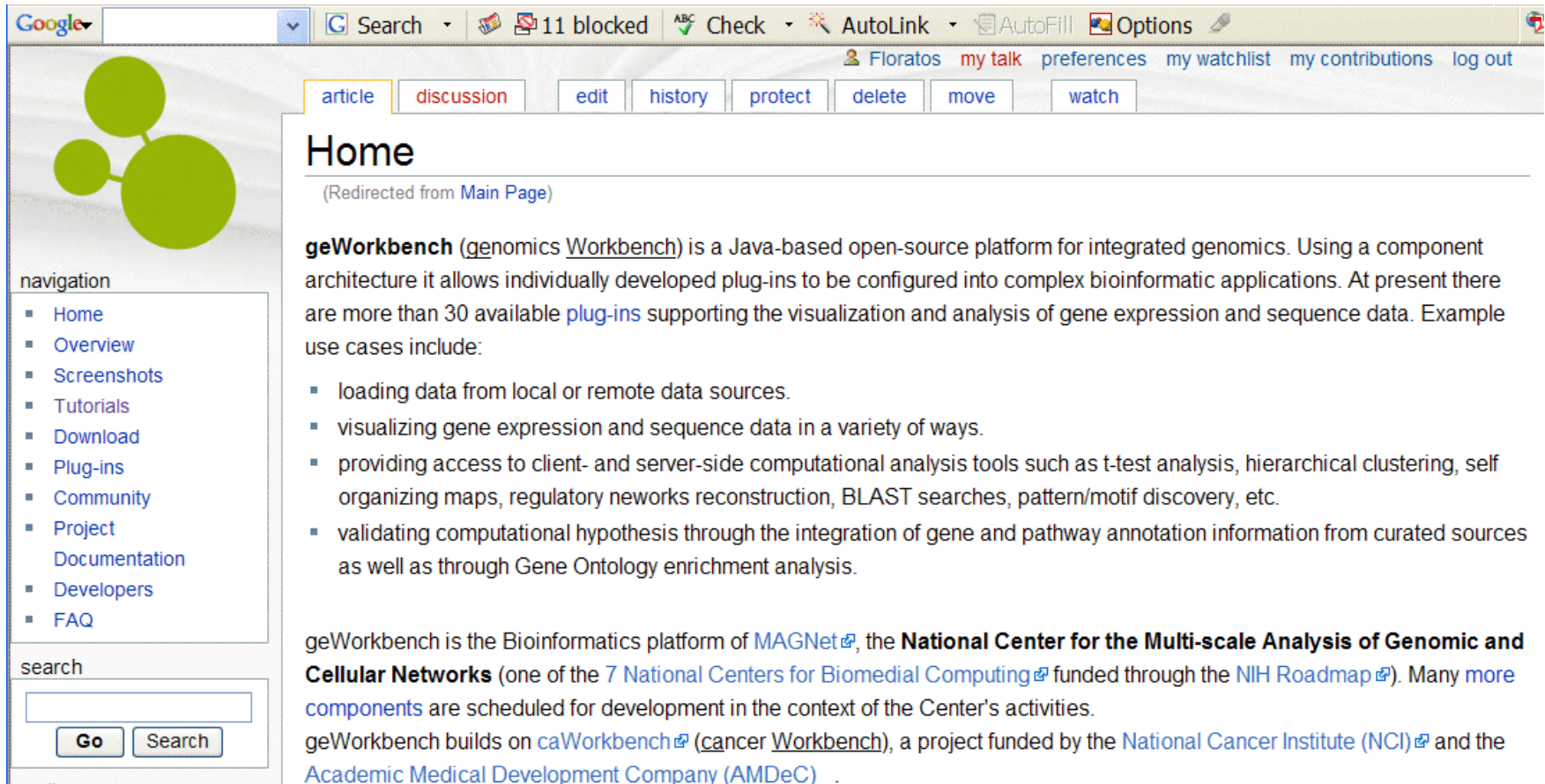
Data Sources

- Swiss-prot
- IPI, HUPO PPP
- PIR
- Ensembl
- UniGene
- RefSeq
- Entrez Gene
- Affymetrix
- MGI
- TransFac
- Oncomine



NCIBI is Using geWorkbench (MAGNet Center) as one of its Problem Solving Platforms

www.geworkbench.org



Google Search 11 blocked Check AutoLink AutoFill Options

Floratos my talk preferences my watchlist my contributions log out

article discussion edit history protect delete move watch

Home

(Redirected from [Main Page](#))

geWorkbench ([genomics Workbench](#)) is a Java-based open-source platform for integrated genomics. Using a component architecture it allows individually developed plug-ins to be configured into complex bioinformatic applications. At present there are more than 30 available [plug-ins](#) supporting the visualization and analysis of gene expression and sequence data. Example use cases include:

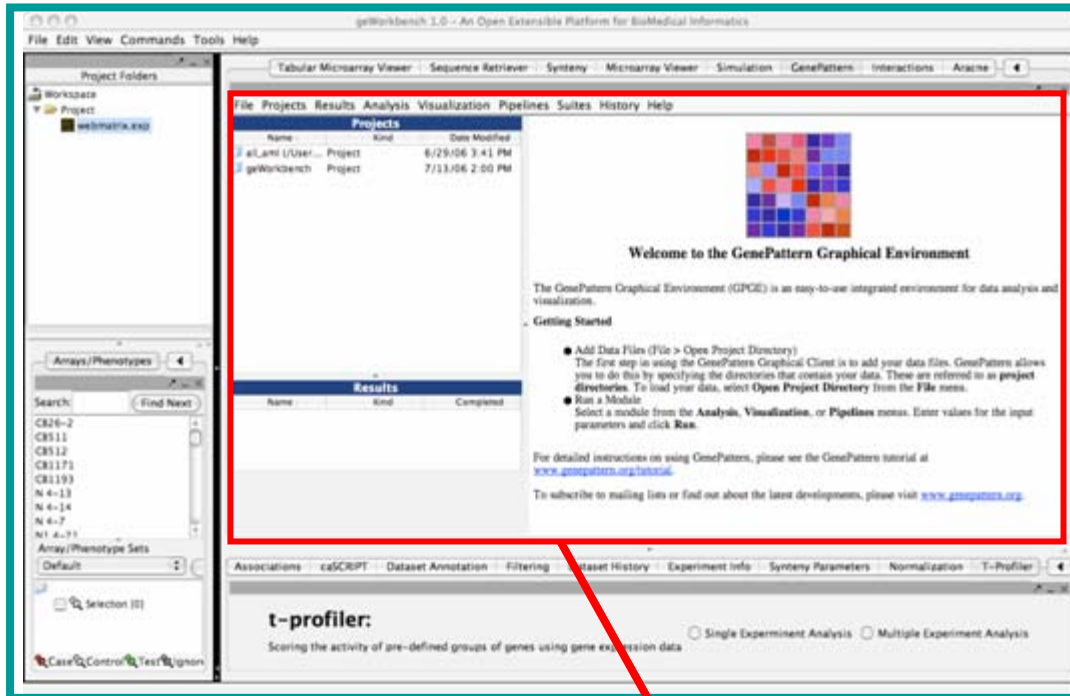
- loading data from local or remote data sources.
- visualizing gene expression and sequence data in a variety of ways.
- providing access to client- and server-side computational analysis tools such as t-test analysis, hierarchical clustering, self organizing maps, regulatory networks reconstruction, BLAST searches, pattern/motif discovery, etc.
- validating computational hypothesis through the integration of gene and pathway annotation information from curated sources as well as through Gene Ontology enrichment analysis.

geWorkbench is the Bioinformatics platform of [MAGNet](#), the **National Center for the Multi-scale Analysis of Genomic and Cellular Networks** (one of the [7 National Centers for Biomedical Computing](#) funded through the [NIH Roadmap](#)). Many [more components](#) are scheduled for development in the context of the Center's activities.

geWorkbench builds on [caWorkbench](#) ([cancer Workbench](#)), a project funded by the [National Cancer Institute \(NCI\)](#) and the [Academic Medical Development Company \(AMDeC\)](#).

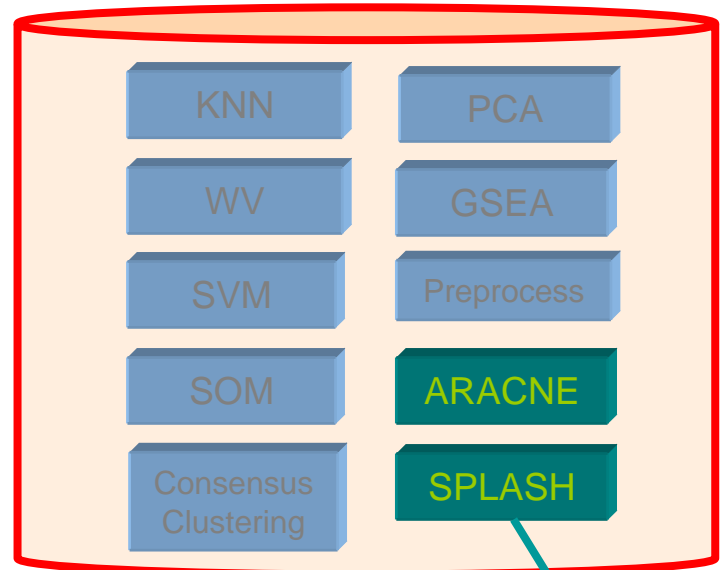


NCIBI is Actively Catalyzing GenePattern/geWorkbench Interoperability



geWorkbench application

GenePattern UI plug-in



GenePattern module repository

geWorkbench modules

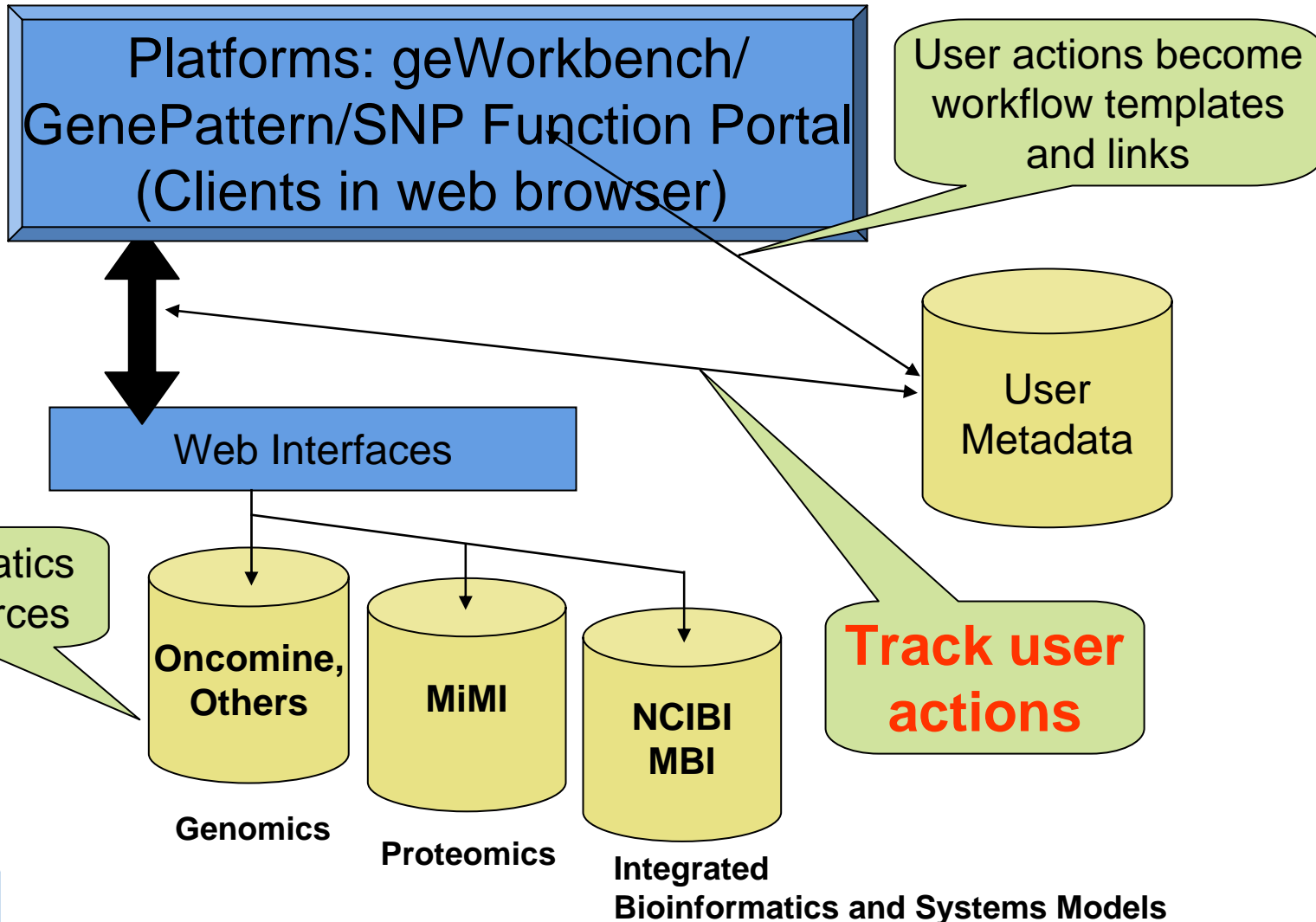
Execute GenePattern modules from within geWorkbench

Wrap geWorkbench modules as GenePattern tasks



Vision: Learning from experts to train novices

Human Computer Interface (HCI)-level tracking of NCIBI Users



Goal: “Tune” the DBP User Experience with User Evaluation Studies

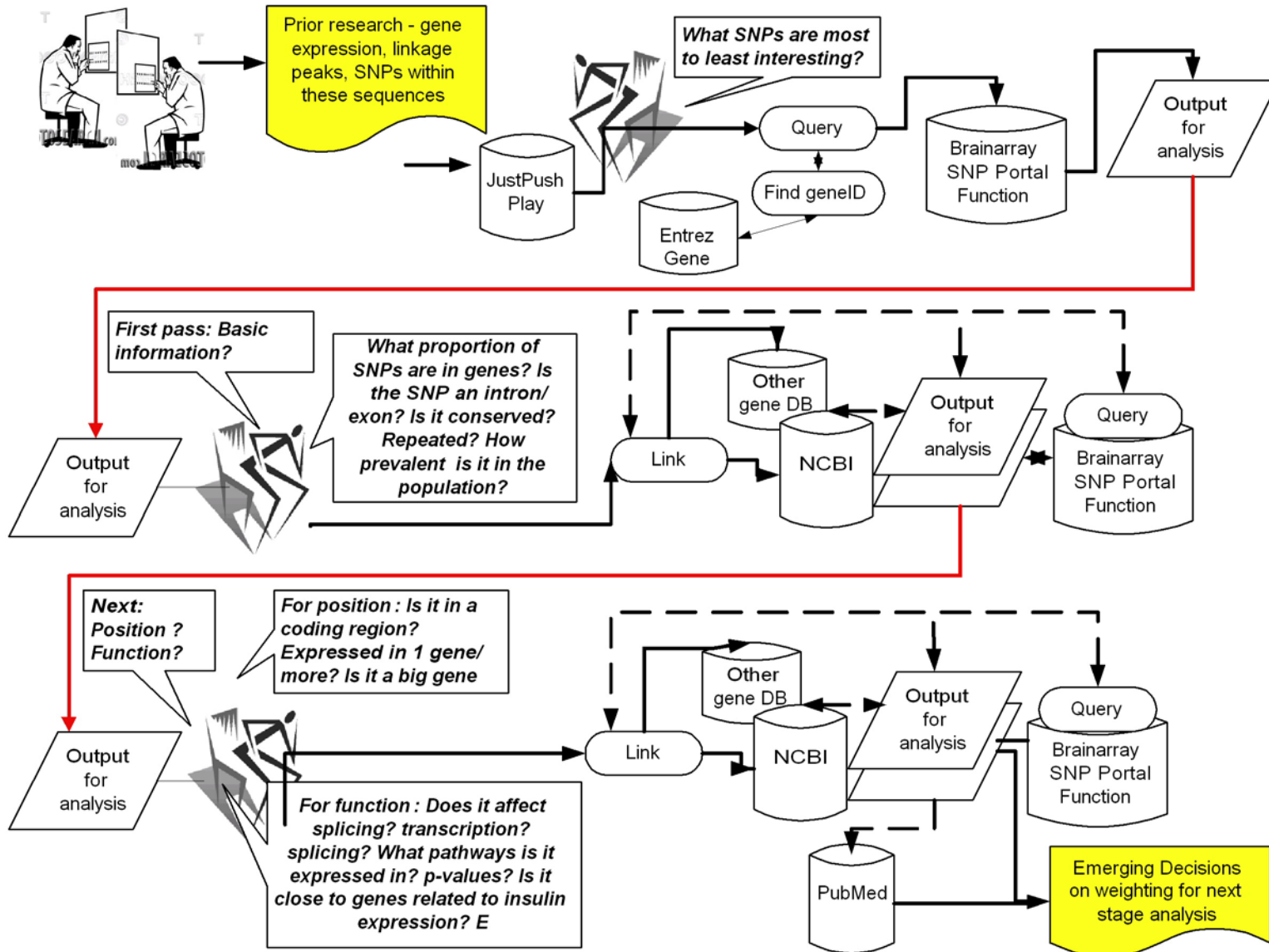
The Source of Feedback for Enhancement

- **Methods:** Naturalistic observations, interviews, transcripts, validation
 - Bioinformatics-savvy DBP researchers – hypothesizing multifaceted aspects of work: including uses of systems for generating and investigating hypotheses; exchanges with remote and onsite collaborators, and interacting with developers/support
 - Bioinformatics-savvy DBP researchers – modeling (Woolf lab)
 - Inexperienced/ “bioinformatics un-savvy” DBP researchers (e.g. T2DM)
 - DBP researchers from diverse laboratories (other labs)
 - Scientific-users’ feedback on similar systems (Meng lab)
- **Expected Outcomes:** User models, scenarios, requirements sensitive to context, experience, purpose, cognitive models of their science



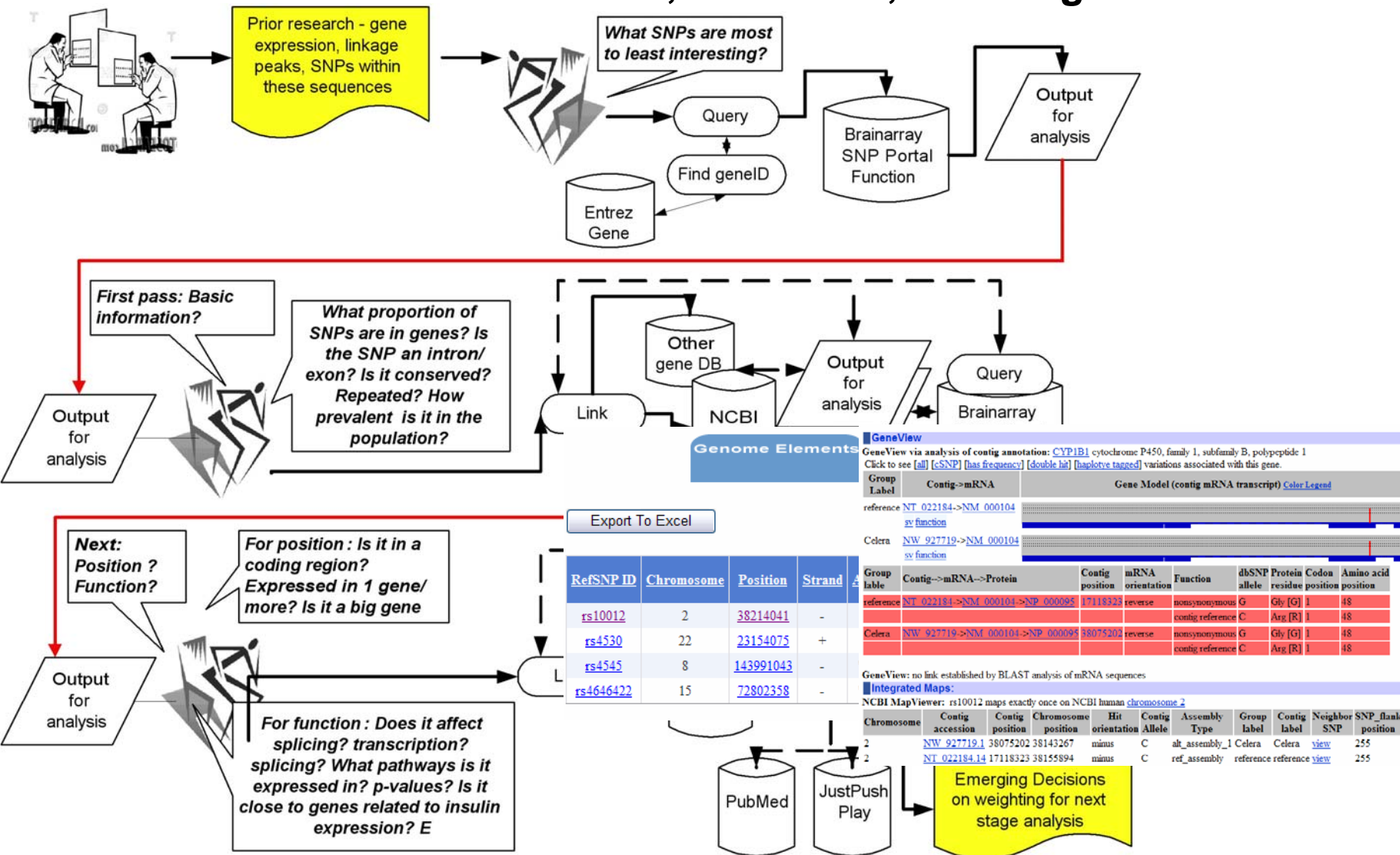
Use-in-context Narrative (1)

Mirel, Ackerman, and Wright



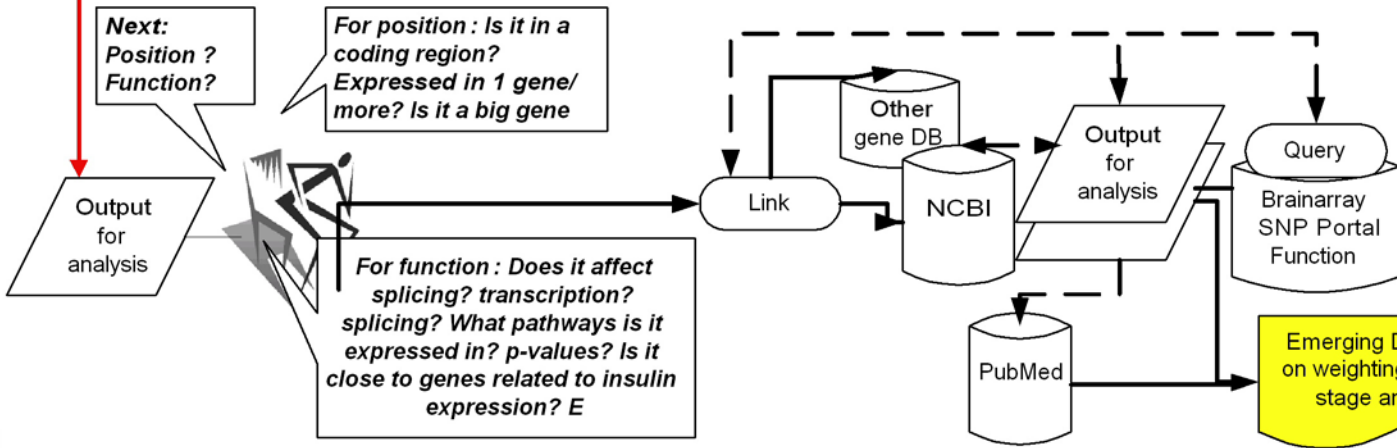
Use-in-context Narrative (2)

Mirel, Ackerman, and Wright



(3)

Mirel, Ackerman, and Wright

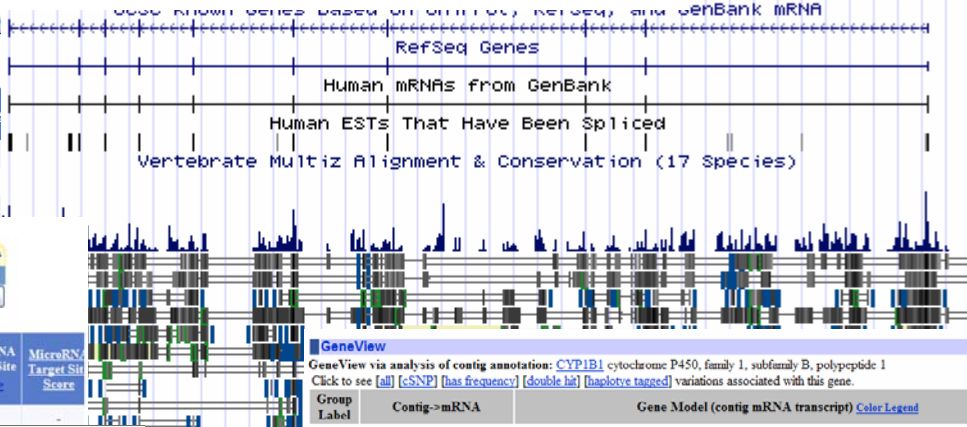


RefSNP ID	Chromosome	Position	Strand	Alleles	Conserved Element	Conserved Element Score	Repeat Mask	Repeat Mask Score
rs10012	2	38214041	-	'C/G'		>>		
rs4530	22	23154075	+	'C/T'		Y	553	
rs4545	8	143991043	-	'A/G'				

RefSNP ID	Transfac Binding Site	Transfac Matrix	Splice Site	Essential Splice Site	Frame Shift	Regulatory Region	Stop Gained	Stop Loss	CpG Island	CpG Ratio	DNAse I Hypersensitive Site	DNAse I Site Score	MicroRNA Target Site	MicroRN Target Site Score
rs10012						Y				0.87				

Matching Transcription Factor Matrix

Allele	Matrix ID	Position	Strand	Core Match Score	Matrix Match Score	Sequence
C	ISDL_02	2	+	1	0.96	tgtGAAAAGct
T	ISDL_02	2	+	1	0.96	tgtGAAAAGct
T	ISUBX_01	18	+	1	0.928	gtgctgcTAATGcccact
T	VSIPF1_Q4_01	20	-			



GeneView

GeneView via analysis of contig annotation: [CYP11B1](#) cytochrome P450, family 1, subfamily B, polypeptide 1

Click to see [\[all\]](#) [\[cSNP\]](#) [\[has frequency\]](#) [\[double hit\]](#) [\[haplotype tagged\]](#) variations associated with this gene.

Group Label	Contig->mRNA	Contig position	mRNA orientation	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
reference	NT_022184->NM_000104							
Celera	NW_927719->NM_000104							
reference	NT_022184->NM_000104	NP_000095	reverse	nonsynonymous	G	Gly [G]	1	48
contig reference					C	Arg [R]	1	48
contig reference					G	Gly [G]	1	48
contig reference					C	Arg [R]	1	48

Information Extraction vs. Information Retrieval

Information Retrieval

- Article Retrieval (publishers)
- Term-Based Queries (e.g. Pubmed)
- Structured Databases (e.g. BIND)
- Canonical Resources (e.g. STKE)

Information Extraction and Analysis

- Database Integration
- Full and Partial Parsing
- Statistical Text Processing
- Assist Model Building (e.g. ODE)

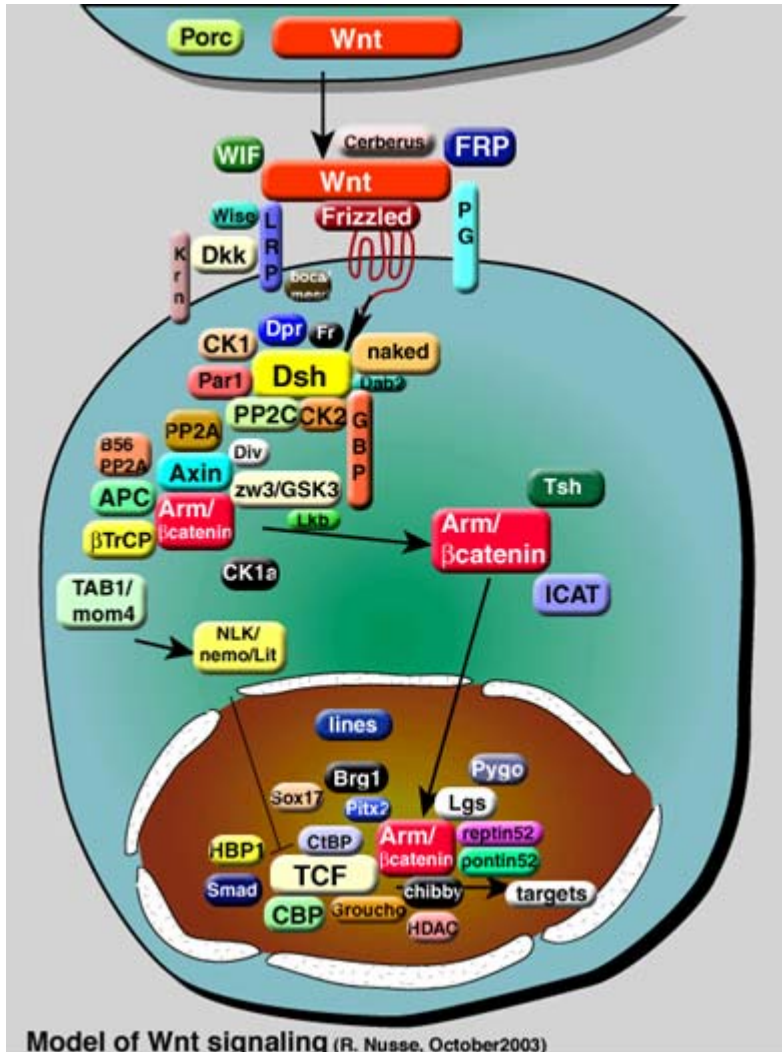
Pilot Project: Wnt Signal Pathway Reconstruction

- full parse vs. human expert curation
- good performance, can we expand it?



D. J. States

Wnt Pathway Project: Human Curation vs. NLP



Cerberus -> Wnt Wnts 10067895
 WIF <-> Wnt 10201374
 Dickkopf Dkk <-> LRP 11357136 11433302 11448771
 Dickkopf Dkk <-> Kremen Krm 11357136
 Wise <-> LRP 12900447
 Wnt <-> Frizzled 8717036
 Wnt <-> FRP Frp 8717036
 LRP <-> Wnt Wnts 11029006 11029007 11029008
 LRP <-> boca mesd 12581525 12581524
 Proteoglycans PG <-> Wnt 2158444
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> CK1e CKI 105176 3210535959
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> CK2 CKII 9214626 12700239
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> GBP Frat1 Frat-1 10428961 10882137 10684251
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Par-1 11433294
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> PP2C 10644691
 PP2C <-> Axin 10644691
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Frodo 11941372
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> naked cuticle gene naked 10693810 11274052
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Axin 10329628 10882137 9920888
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Dapper Dpr 11970895
 Dishevelled Dishevelled dishevelled disheveled Dsh Dvl <-> Disabled-2Dab-2
 Disabled2 Dab2 12805222
 Disabled-2Dab-2 Dab2 Disabled2 <-> Axin 12805222
 LKB1 XEEK1 <-> GSK 12973359
 Armadillo beta-catenin <-> zw3 GSK-3b GSK3 GSK3beta 9554852 9601644 10073940
 11927557 12000790
 Armadillo beta-catenin <-> Casein Kinase 1 casein kinase 1 CK1a CKI CKIalpha 955485
 2 9601644 10073940 11927557 12000790
 Armadillo beta-catenin <-> APC 9554852 9601644 10073940
 Armadillo beta-catenin <-> Axin 9554852 9601644 10073940
 Armadillo beta-catenin <-> Slitlim b-TrCP 9461217 9784611 10072378
 Axin <-> PP2A 9920888
 Axin <-> LRP 11336703
 Axin <-> GSK-3b GSK3 GSK3beta 9482734 9501208 9601644
 Axin <-> APC 9482734 9501208 9601644
 PP2A <-> APC 10082233
 Axin <-> Diversin 12183362
 beta-catenin <-> TCF 0000000
 TCF <-> Groucho 9783586
 Groucho <-> HDAC 10485845
 beta-catenin <-> Legless Bcl9 11955446 11967528 12015286
 beta-catenin <-> Pygopus pygopus pygo 11955446 11967528 12015286
 beta-catenin <-> Chibby 12712206
 TCF <-> CBP P300 9774110 10775268 10769018
 beta-catenin <-> Ptb2 12464179
 beta-catenin <-> Brg-1 11532957
 beta-catenin <-> Pontin52 Pontin pontin 11080158
 beta-catenin <-> Reptin52 reptin Reptin 11080158
 beta-catenin <-> XSox17 10549281
 beta-catenin <-> Smad4 10693808
 TCF <-> CIBP 10375506
 TCF <-> HBP1 11500377
 TCF <-> Lit1 NLK Nemo 10380924 10391247 10391246
 Lit1 NLK Nemo <-> TAB1 TAK1 MOM-4 10380924 10391247 10391246
 Teashirt Tsh <-> beta-catenin 10205174
 beta-catenin <-> ICAT 10896789

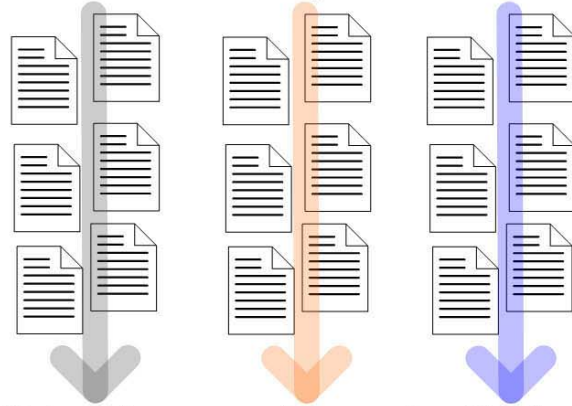


Model of Wnt signaling (R. Nusse, October 2003)

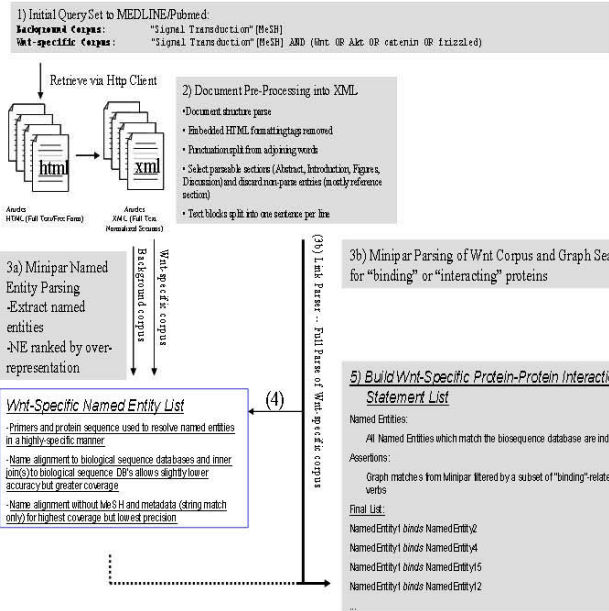


NCIBI NLP Processing Overview: Linking Biomedical Literature to DBPs

Background Literature Disease-Specific Corpus Candidate Genes Corpus



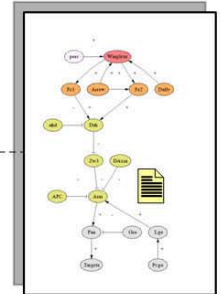
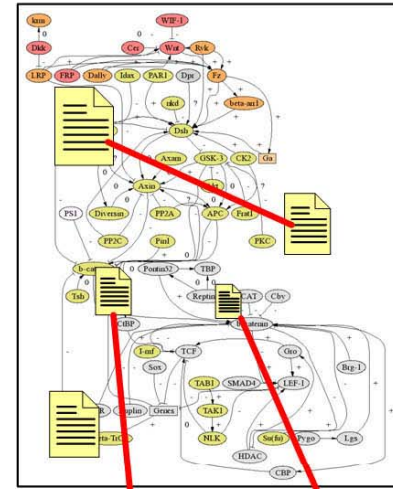
Natural Language Processing Pipeline



Link Discovery Between Pathways via Literature (Terms, Relationships)

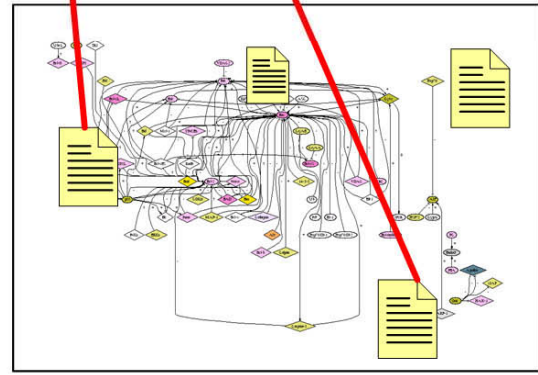


Pathway Resources (KEGG, STKE, etc.)



Model Organism-Specific Pathways (Drosophila/Human/Mouse)

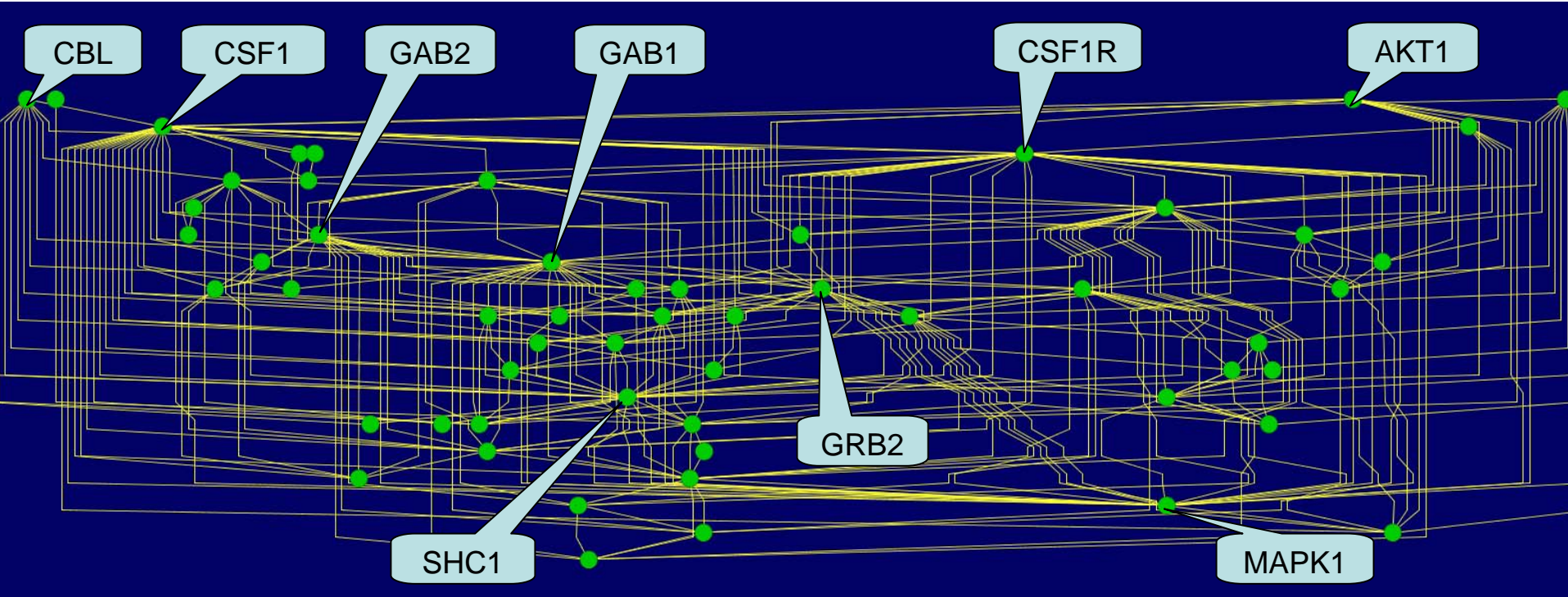
External Canonical Pathway (e.g. Apoptosis or Disease-specific pathway)



NCBI MeSH, GO, etc. Database

Biosequence Database (MBI)

Graphical Text Summarization



Nodes => genes

Edges => sentences referring to multiple genes

Genes and relationships in Lee AW, States DJ (2000) Mol Cell Biol. 2000 Sep;20(18):6779-98.



D. J. States

National Center for Integrative Biomedical Informatics

Overview: Text Processing in SLIF

- Find *entity names* in text, and *panel labels* in text and the image.
- *Match* panels labels in text to panel labels on the image.
- *Associate* entity names to textual panel labels using *scoping* rules.

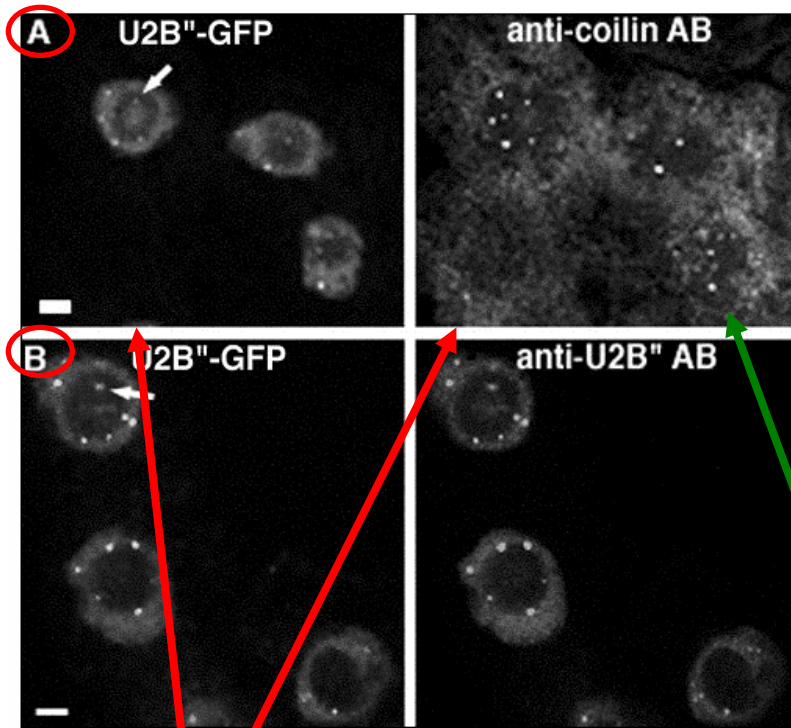


Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B⁰-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright foci of anti-coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. (B) Single confocal optical section of BY-2 cells expressing U2B⁰-GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies are in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B⁰ also express the U2B⁰-GFP fusion. Bars, 5 μ m.

Core 3: Driving Biological Problems (DBPs) – Criteria

- Common diseases with complex and heterogeneous etiology—many unknowns
- Extensive complex datasets available
- Experienced NIH PIs with well-funded programs generating new data and capable of testing models
- Commitment to Core 1 and 2 interactions, problem-posing sessions, and open access
- Interest in cross-DBP analyses and outreach



3.1: Prostate Cancer Progression

- Focus on clinically-critical androgen-dependent to androgen-independent switch
- Explore newly discovered androgen-driven TMPRSS2/ETS fusion gene translocation (transcription factor) phenotypes (Science 28, Oct 2005)
- Create “smart” parsers for text mining
- Integrate gene expression, proteomics, protein-protein interaction data, starting with Oncomine 3.0, into a clinically-heuristic systems model
 - Link Oncomine 3.0 data and tools to other NCIBI resources and platforms to enhance systems integration





- A resource for examining gene expression in cancer.
- Collect, standardize, analyze, and deliver published cancer gene expression data to the research community.
- Probe the expression of a gene across thousands of cancer samples or explore genes, processes, and pathways deregulated in a particular type of cancer.
- Oncomine pre-computes cancer profiles, clusters, and gene set modules so you can focus on discovery. Read more here.

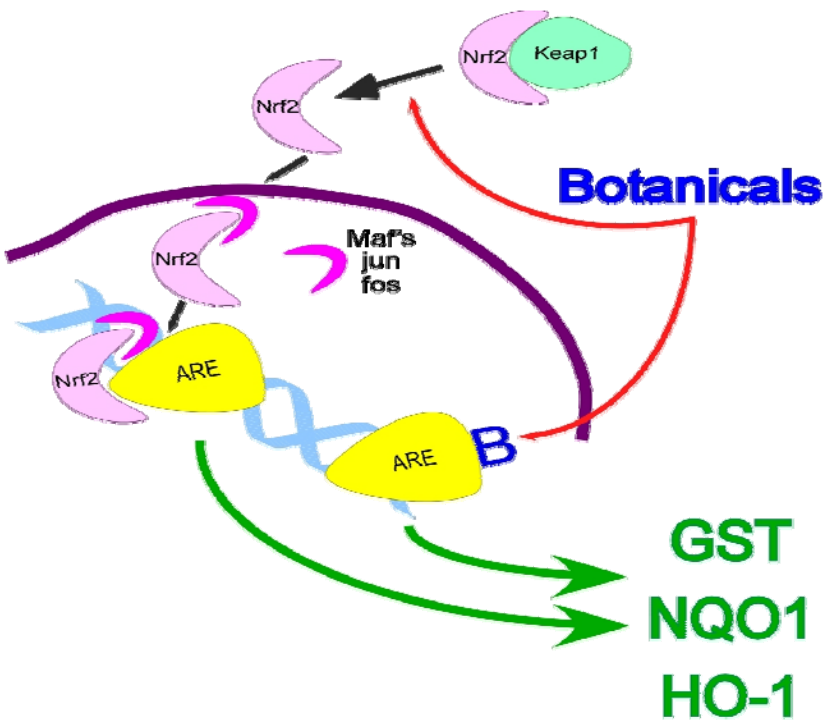
Statistics

Studies - Catalog: 896
Studies - Analyzed: 149
Microarrays: 16656
Data points: 308226536
Cancer Types: 49
Registered Users: 8416



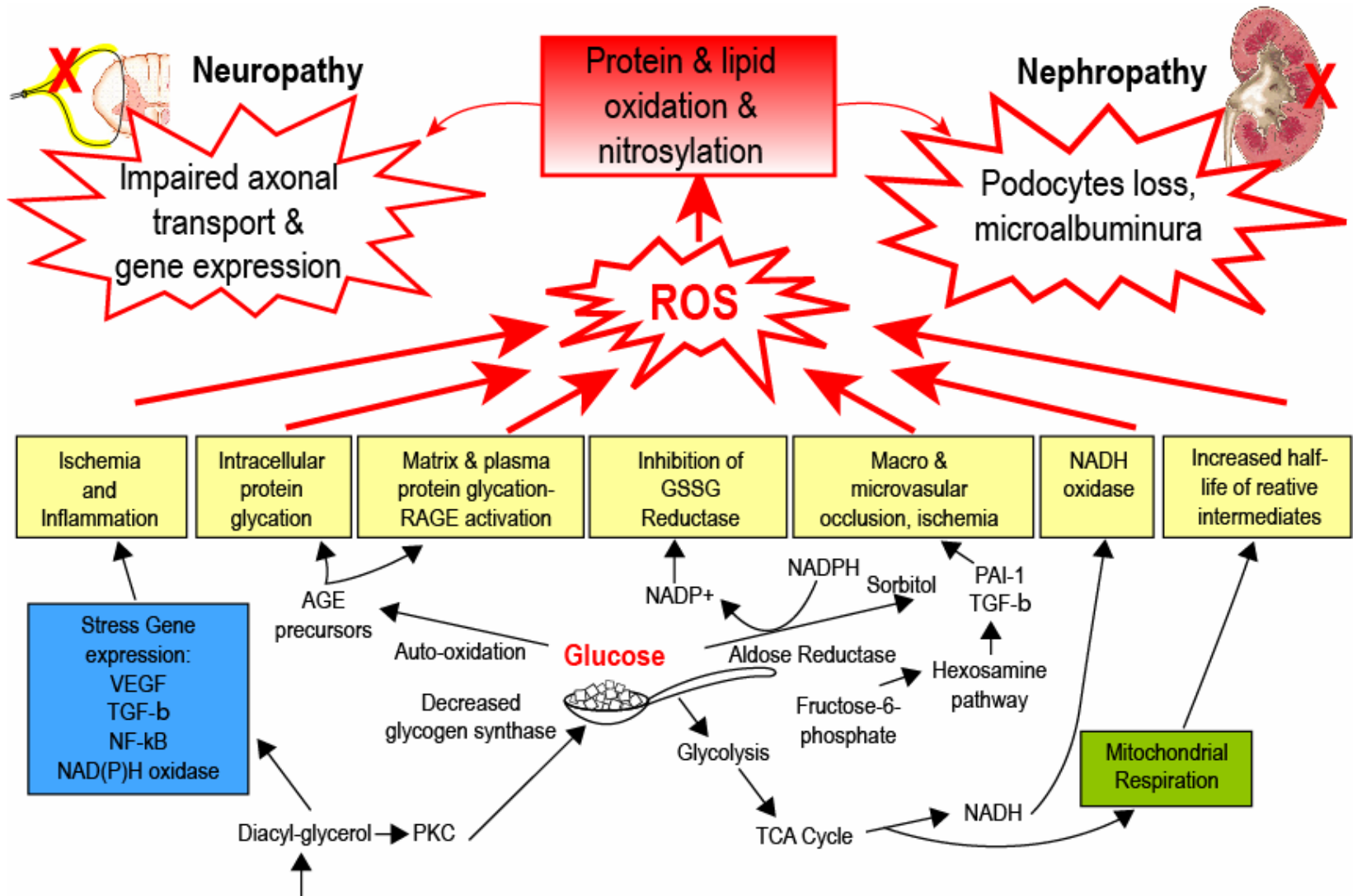
A. Chinnaiyan

3.2: Type 1 Diabetes Neuropathy and Nephropathy: Mechanisms and Models



- Compare Nrf2 neuroprotection signaling to other pathways via SAGA tool and database workflow (Patel lab)
- Identify antioxidant response elements and tie to signaling pathways and metabolites
- Model T1DM-associated pathways via Bayesian Network analysis (Woolf lab)
- Collaborators will test therapies against reactive oxygen species in animal model (resveratrol trial)

Unifying hypothesis of diabetic complications



DiabetesT1 Complications: Integration and outlook

- NCIBI Data Warehouse (Core 1,2):
 - Gene expression data base (TED, States)
 - Metabolic parameters (Pennathur)
 - GC/MS: Full Scan Vs SIM mode
 - LC/MS and QQQ: Centroid Vs Profile mode
 - ESI/MALDI/Q-TOF: Integration of m/z over time
 - Pattern Recognition Mass spectrometry
- NCIBI data processing (Core 1,2):
 - Promoter modeling to ARE (States)
 - Integration with clinical parameters (States, Clauw)
 - Bayesian Networks of oxidative stress (Woolf)
 - Prediction of metabolomic read-out parameters as non-invasive screening tools
 - Integration with genetic linkage workflow of DM II (Boehnke)
- Community outreach (Core1-3):
 - Nephromine (Chinnaiyan):
 - Use established world-wide networks of research centers



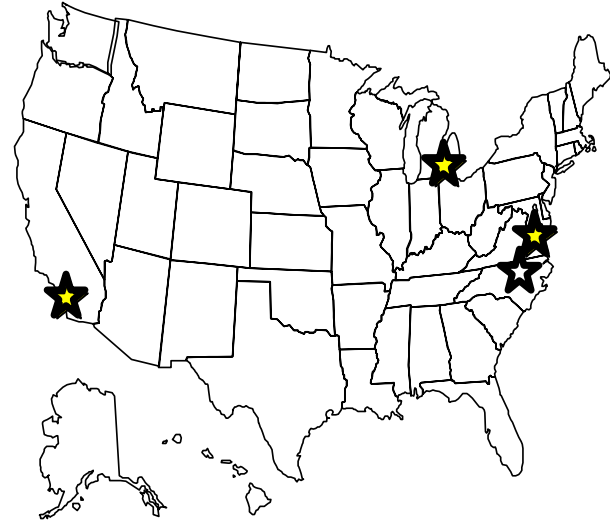
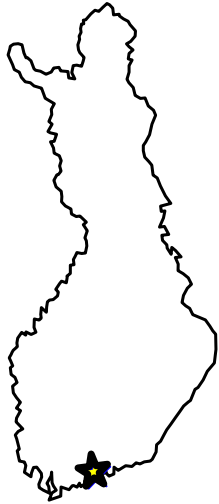
3.3: Type 2 Diabetes: Genetic and Phenotypic Heterogeneity

- Exploit ongoing FUSION study datasets
- Implement SNP analysis workflow for whole genome association (WGA) study
 - For each SNP (and then haplotype), screen databases to identify: coding/noncoding, effect on protein, evolutionary conservation, splice variant, transcriptional regulation, microRNA binding, CpG island, DNase hypersensitivity, disease associations
 - Devise algorithms with much higher throughput
 - Organize parallel analyses for DBP 4 (bipolar) seeking more homogeneous subgroups of patients



M. Boehnke, A. Jackson, L. Scott, and F. Meng

FUSION Study: Finland-United States Investigation of NIDDM Genetics



- National Public Health Institute, Helsinki
- National Human Genome Research Institute, Bethesda
- University of North Carolina School of Medicine, Chapel Hill
- USC Keck School of Medicine, Los Angeles
- University of Michigan School of Public Health, Ann Arbor



FUSION Genomewide Association Study

- Stage 1: Genotype 1200 cases, 1200 controls on Illumina 317K SNP platform (CIDR)
- Stage 2: Genotype remaining 1500 cases, 1500 controls on best 1-2% of Stage 1 SNPs
 - SNPs associated with T2D, related traits
 - consider also genome annotation
- Follow up: genotype additional markers, additional samples, refine disease-marker association
- Two-stage designs can maintain power, reduce cost



M. Boehnke, A. Jackson, L. Scott, and F. Meng

National Center for Integrative Biomedical Informatics

SNP Annotation

- Not all 317K SNPs equally interesting:
 - non-synonymous, splice sites, conserved regions, “gene deserts”
 - transcription factor binding sites, enhancers, promoters, deletion associated, protein binding site
 - previously associated with T2D, related traits, pathways
- For one or few candidate genes, can annotate by hand; harder genome wide

M. Boehnke, A. Jackson, L. Scott, and F. Meng



SNP Annotation (continued)

- Use annotation to:
 - select SNPs for stage 2
 - weight results of stage 1+2 joint analysis
- Annotate not just based on SNP itself, but also based on SNPs it “tags”
- SNP Function Portal helping provide the relevant information

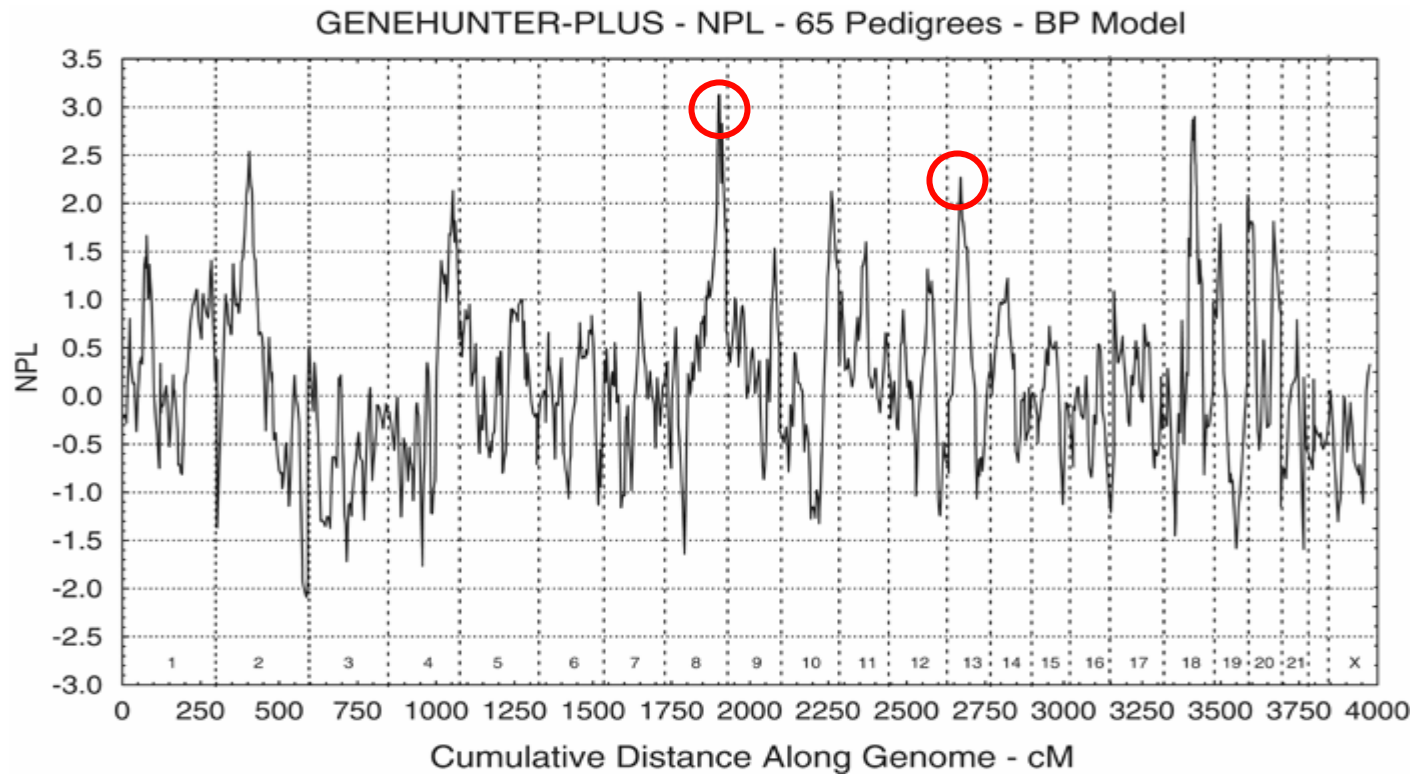
M. Boehnke, A. Jackson, L. Scott, and F. Meng



3.4: Bipolar Disorder

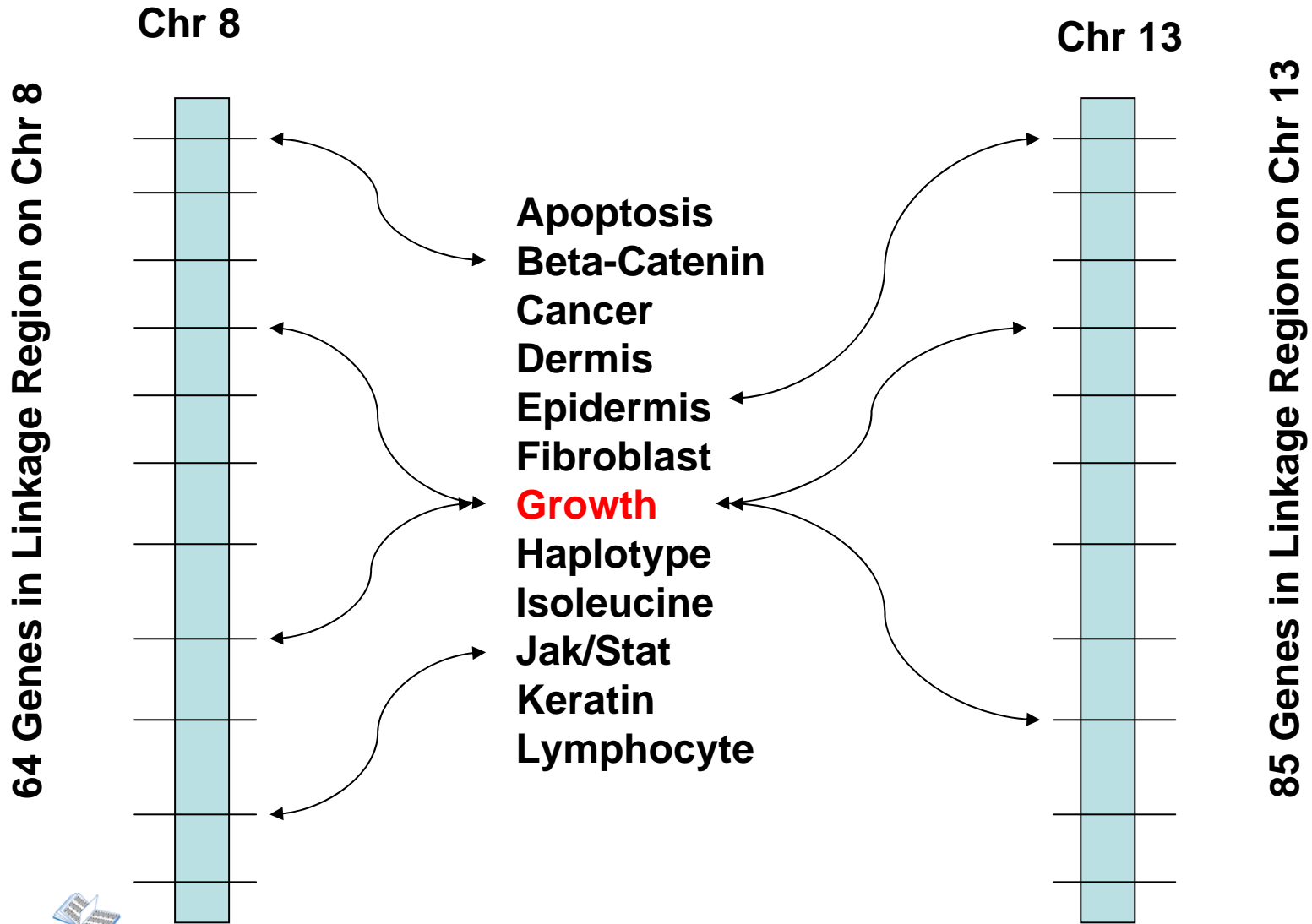
- Exploit ongoing NIMH studies of genetic, psychological, and psychiatric heterogeneity
- Import data from brain imaging
- Generate faster algorithms for analysis of pathways for candidate genes
- Use flexible tools to search pathways for similarities
- Model interactions of variants at two different gene loci (two different chromosomes)

Genetic Interaction of Loci on 8q24 and 13q12 Alter Susceptibility to Bipolar Disorder



Nonparametric linkage analyses using GENEHUNTER-PLUS for BPI, BPII, and SAM.
Chromosomes are indicated along the bottom of the figure.
McInnis et al. *Molecular Psychiatry* 8:3288-98 and 9:191-6, 2003

3.4 MeSH Keywords



Bayesian Network (BN) Modeling: Three Directions

BNs are graphical learning algorithms that detect causal or apparently causal relationships from experimental data.

1. Biomarker Identification
2. Static Bayesian Networks
3. Dynamic Bayesian Networks

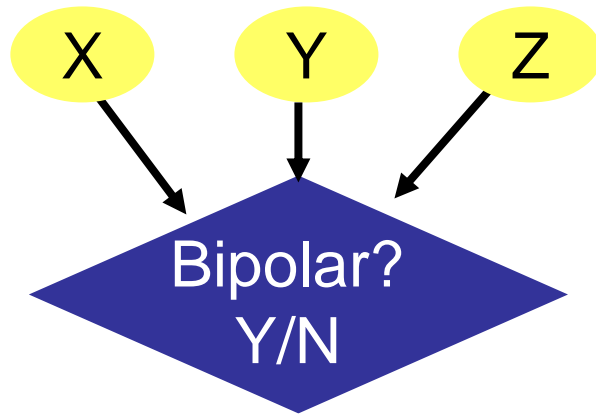
Common Features to be Integrated

1. Relevant User Feedback
2. Optimal Experimental design
3. Guided Model Expansion
4. Prediction / Instantiation Engine



Goal: Accurately define biomarkers for Bipolar disorder

Gene candidates: **A** **B** **C** **D** ... ~15,000 in total



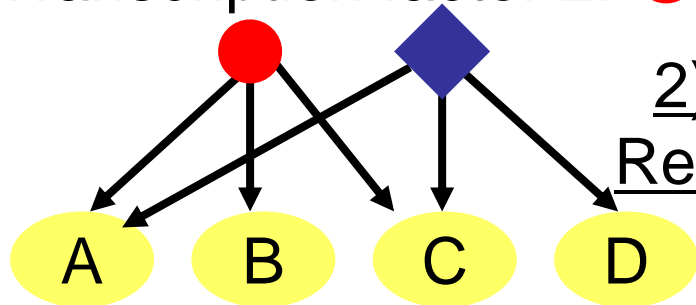
Detect high scoring Bayesian networks that predict the disorder. Captures nonlinear, and complex logical relationships that are apparently causal.

Status:

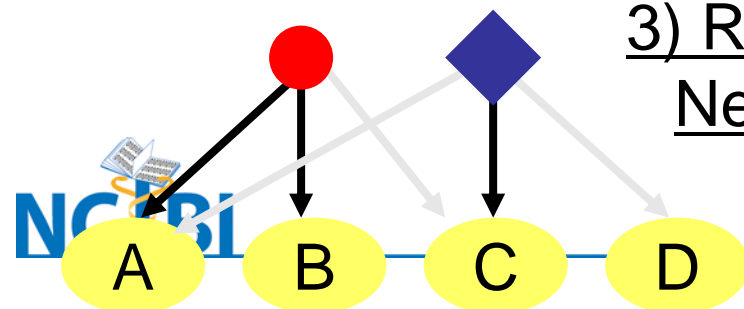
- Affymetrix Data for 64 patients has been collected and being reformatted for preliminary analysis.
- Network search algorithm being modified to efficiently scan this space

Goal: Accurately define which transcription factors are responsible for governing gene expression.

Transcription factor 1: 
Transcription factor 2: 



Reduce graph based on consistency with expression data in Oncomine



1) Genomic Representation



2) Bayesian Representation



3) Reduced Network

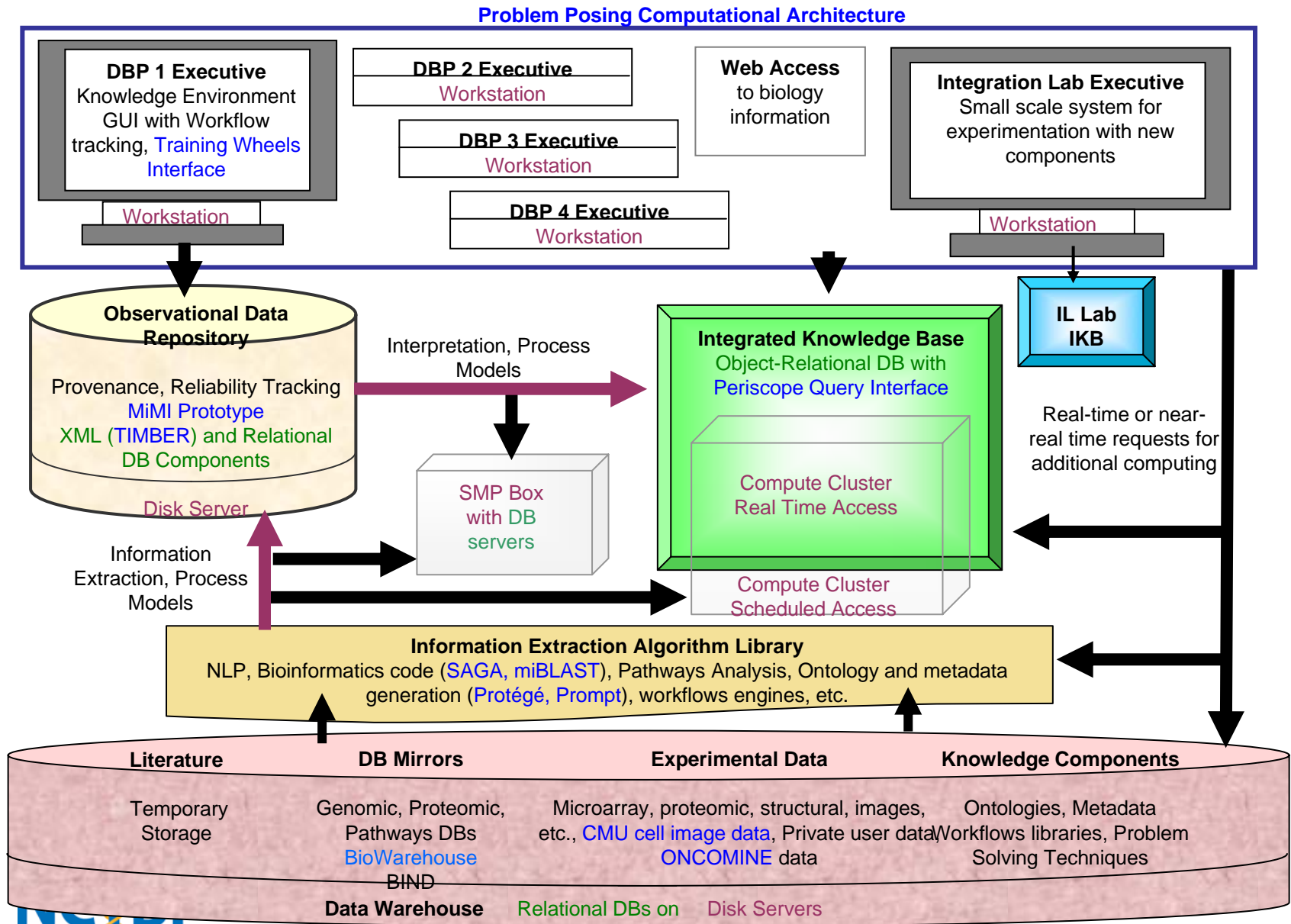


Core 3: Ongoing Steps

- Integrate these tools into the larger process (e.g., feed results into SAGA Graph Matching tool)
- Problems Built into workflow and implemented in geWorkbench and/or GenePattern
- Apply across Driving Biological Problems (DBPs)
- Guide next revision of tools in Cores 1 & 2 through interactive user feedback (fail early/fail often model)
- Evaluate an additional DBP for years 4 and 5+
- Collaborate with appropriate R01 and R21 applicants
 - 4 proposals submitted last round, several more in pipeline



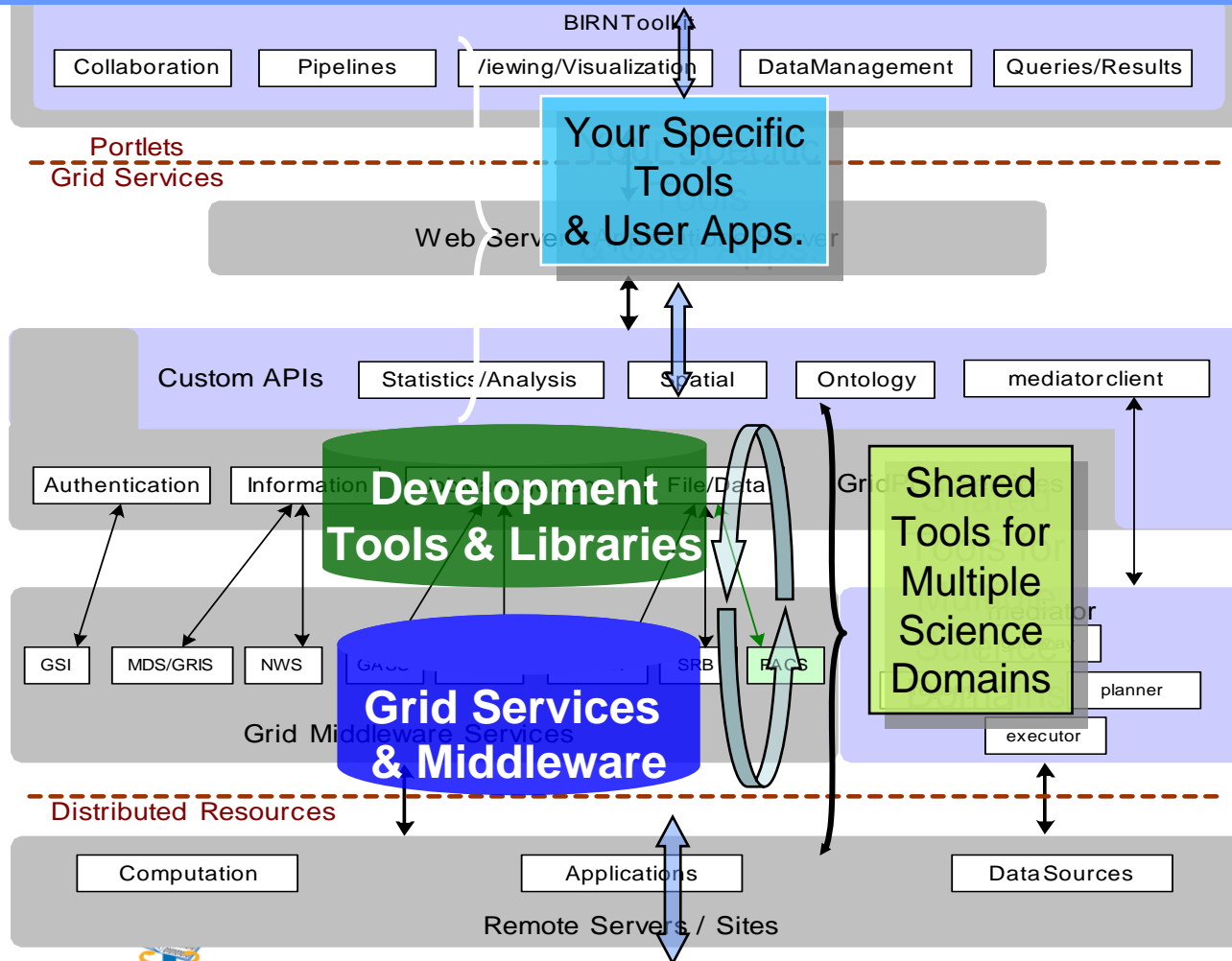
Core 4 Infrastructure: The components are built and will be integrated by December 2006



NCIBI is Adapting the BIRN Core Cyberinfrastructure Model

Friendly Work Facilitating Portals

Authentication - Authorization - Auditing - Workflows - Visualization - Analysis



- BIRN builds on evolving community standards for middleware
- Adds new capabilities required by projects
- Does System Integration of domain-specific tools building a distributed infrastructure
- Utilizes commodity hardware and stable networks for baseline connectivity

Distributed Computing, Instruments and Data Resources

Core 5: Education and Training

- **First teach ourselves**

- “Tools and Technologies” lunchtime demonstration series (launched). Will move to streaming video over the intra and Internet in 2006
 - Presentations about current development and tools being used by various NCIBI components
 - Informal and highly interactive
 - Recording as the basis for external training and evaluation
 - Video, archived PowerPoint slides, Wiki notes
 - Fully Interdisciplinary
- This effort will allow us to naturally shift to teaching others, especially DBP Researchers

- **Leverage the environment**

- New UM Center for Computational Medicine and Biology (CCMB)
- Bioinformatics and Computational Biology Graduate Training Program
 - Many Trainees participating in the NCIBI NIGMS T32
 - Support in years prior to joining NCIBI projects
 - Enhanced training and research opportunities for T32 trainees
 - Synergistic infrastructure
 - Courses, seminars, journal clubs, facilities

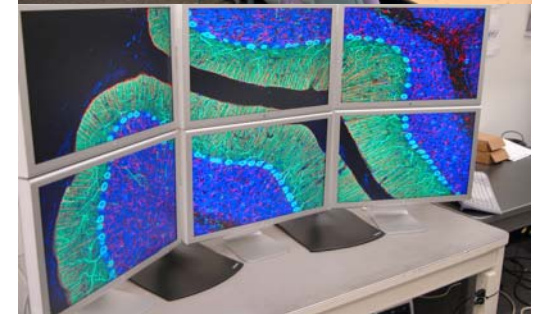


NCIBI First Annual Research Conference: June 2006



NCIBI is Leveraging the UM Bio-cyberinfrastructure Build-out: 'The Connection Project'

- Effort at the School of Information (SI) to prototype and deploy next-generation, real-time collaboration systems
 - Tom Finholt, Project Director
 - Erik Hofer, Technical Director
 - Emilee Rader, PhD Student
 - David Lee, MSI Student
 - Ted Hanss, Medical School Lead
- Research and development focuses on real-time Interdisciplinary Research collaborations and communication. Hard link to UCSD Biomedical Informatics Research Network (BIRN)
- Sponsored by UM Office of the Provost and the UM Medical School
 - Additional equipment provided by M-GRID and Sun Microsystems
 - NCIBI and its host the UM Center for Computational Medicine and Biology (CCMB) are early adaptors
- Already being used for NCSA collaboration (beginning with UM Proteomics 551 offered to UIUC students next term)



Path to Sustainability: NCIBI has Established a Strong Set of Strategic partners

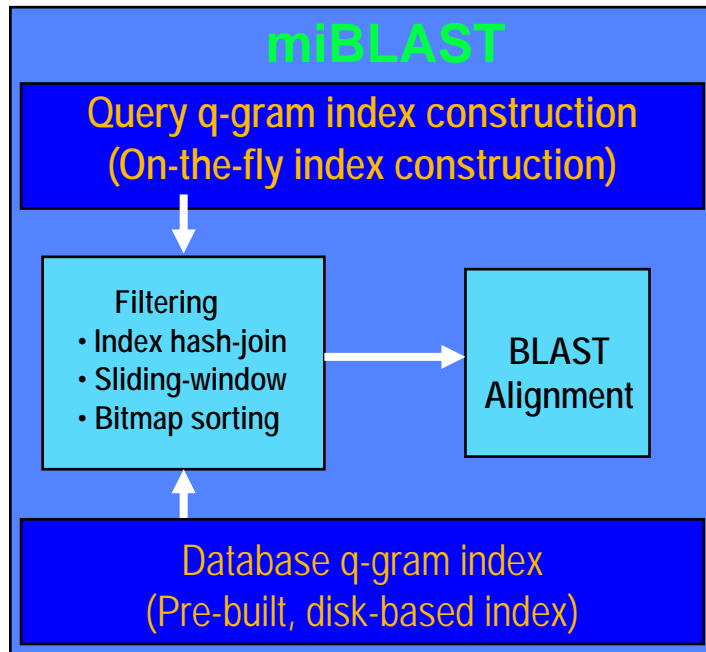


NCIBI Hot Topic (1)

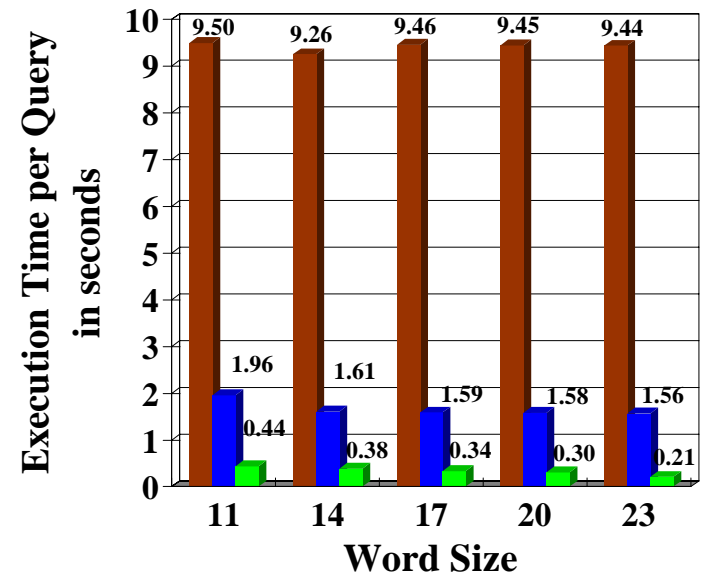
miBLAST: Scalable BLAST for Batch Workloads

- A common task is to search a large sequence database using a “set” of query sequences.
 - E.g. Validation of the Affymetrix probe set against UniGene.
- Approach: A novel database inspired “join” algorithm which indexes both the data and the query sets.

Query the entire Affy probe set against Human UniGene.



naïve BLAST batch BLAST miBLAST



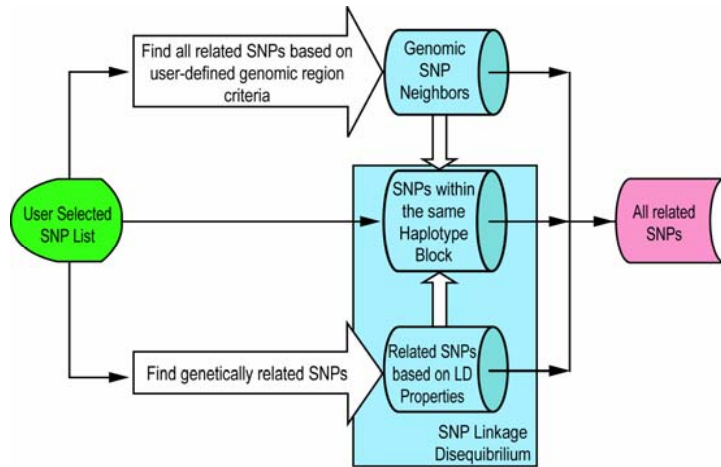
miBLAST is 22X faster than BLAST

• miBLAST: www.ncibi.org/resources

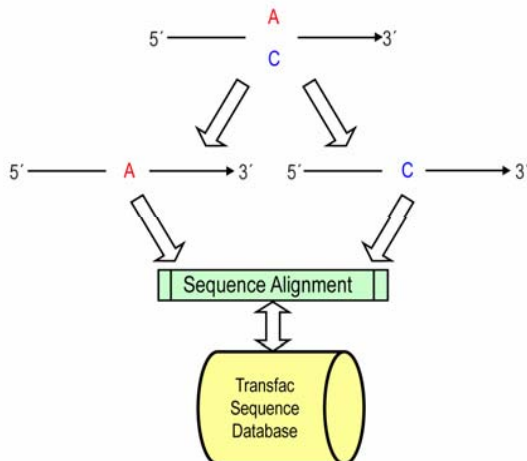


NCIBI Hot Topic (2)

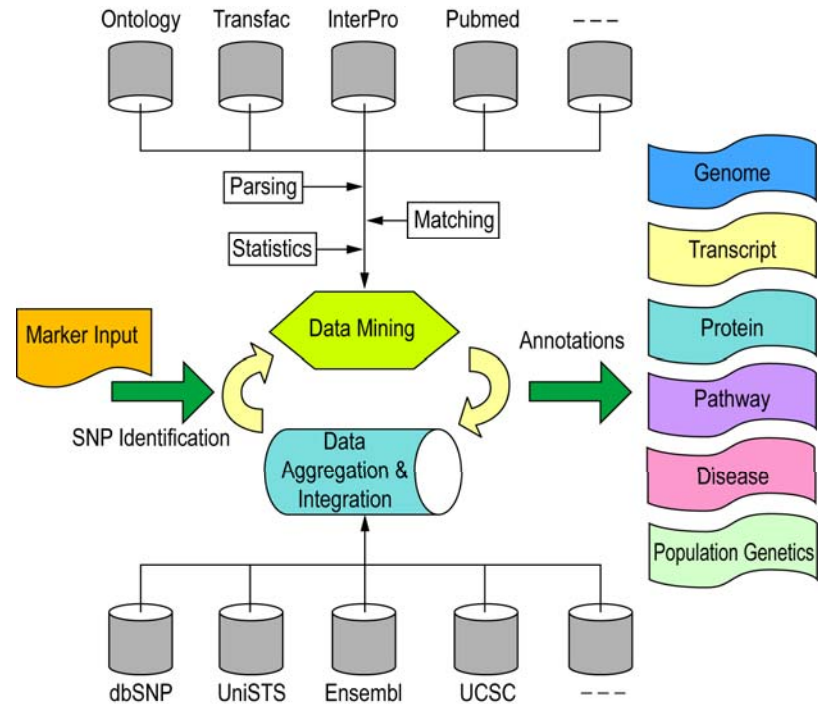
Innovations in SNP Analysis



Determining all related SNPs by genomic location



Mapping SNPs variation to Transcription factor binding sites.



Annotating SNPs: datasources and workflow



•SNP Portal: www.ncibi.org/resources

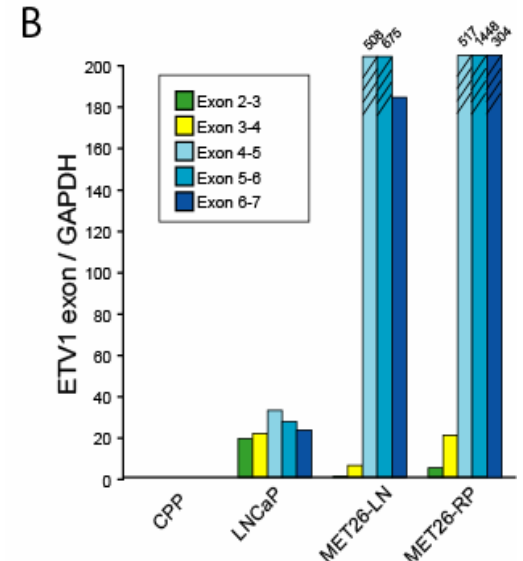
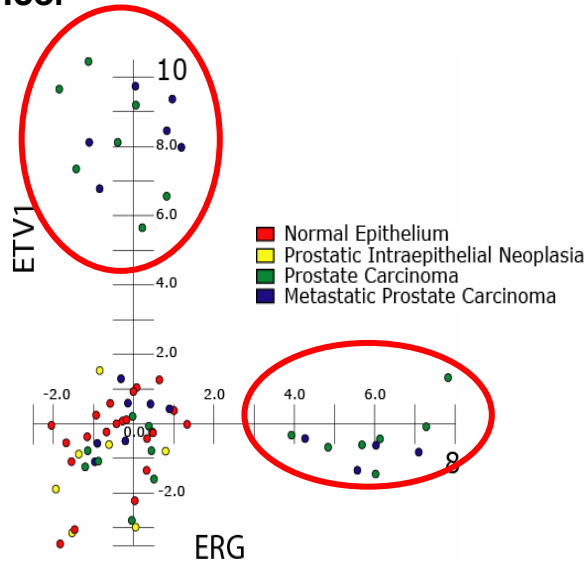
National Center for Integrative Biomedical Informatics

NCIBI Hot Topic (3)

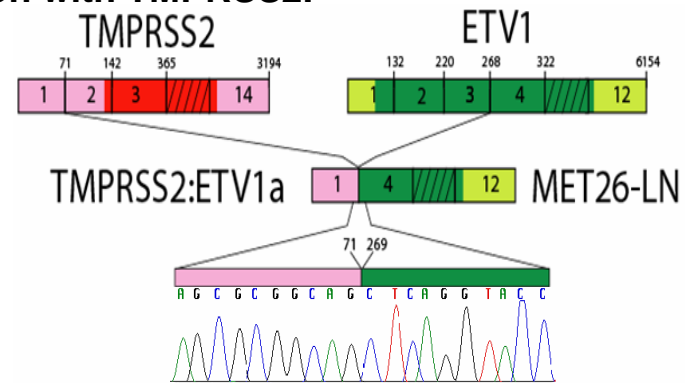
Discovery of a Common Fusion Gene in Prostate Cancers

Rank	%	Score	Study	Cancer	Gene	Evidence
1	95	20.056	Valk et al.	Leukemia	<i>RUNX1T1</i>	XX
1	95	15.4462	Vasselli et al.	Renal	<i>PRO1073</i>	X
1	90	12.9581	Ross et al.	Leukemia	<i>PBX1</i>	XX
1	95	10.03795	Lapointe et al.	Prostate	<i>ETV1</i>	**
1	90	9.1163	Tomlins et al.	Prostate	<i>ETV1</i>	**

Bioinformatics approach yields a list of genes differentially expressed in stages of prostate cancer



5' exons showed reduction in expression in Cancer progression. Led to discovery of 5' gene fusion with *TMPRSS2*.



Oncomine: www.ncibi.org/resources

NCIBI Hot Topic (4)

The Prechter Bipolar Genetic Repository



Heinz C. Prechter

- Will contain data collected in partnership with the University of Michigan, Johns Hopkins University (JHU), Stanford University and Cornell University
- Samples housed at the University of Michigan (UM) Depression Center
- Maintained as a nationally accessible database and linked to analysis by the National Center for Integrative Biomedical Informatics (NCIBI)

Background--Johns Hopkins University Bipolar family samples:

- The collection of families began in 1986
- The families were assessed by psychiatrists, blood samples taken and lymphocytes used to make immortalized cell lines stored at JHU
- Formal NIMH funding started 1988 and has continued. DANA foundation support in the early to mid '90s
- Recent JHU projects use the Rutgers University NIMH repository, but initial samples are stored only at JHU

The University of Michigan (UM) Prechter Bipolar Genetic Repository will:

- Receive JHU samples: ~ 1,500 samples from 140 bipolar families
- Maintain the JHU immortalized bipolar families cell lines and prepare DNA for scientific study
- Make the samples available to scientists world-wide; A joint JHU-UM accession committee will review requests to receive and study the DNA



UM National Center for Integrative Biomedical Informatics (NCIBI) Participation:

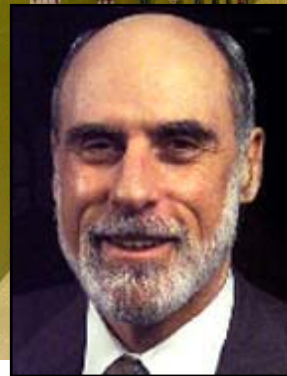
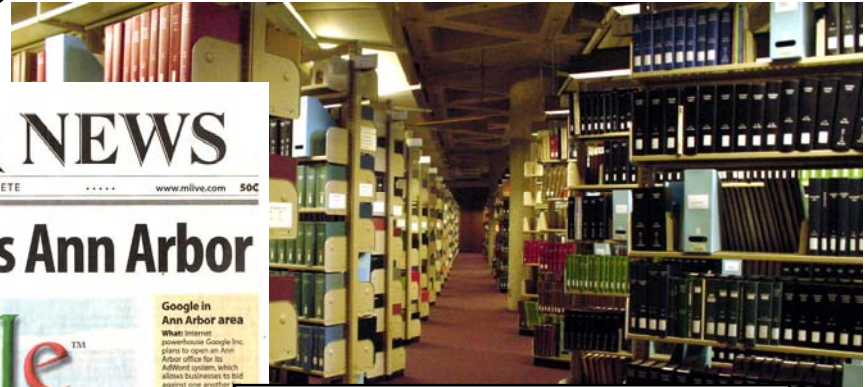
- NCIBI will host the genome-wide microsatellite data from JHU and NIMH samples
- NCIBI will expand and improve data presentation and analysis
- NCIBI will host the CHR 8q24 SNP data and allow the searching of results
 - Washington University also hosts the NIMH specific samples; will make data available



*Funded by the Heinz C. Prechter Bipolar Research Expendable Fund
and the Heinz C. Prechter Bipolar Research Endowed Fund.*

**University of Michigan
Depression Center**

NCIBI is actively leveraging the University of Michigan's Special Relationship with Google



NCIBI “Hot Topics” Publications

miBLAST

- Kim YJ, Boyd A, Athey BD, Patel JM. miBLAST: scalable evaluation of a batch of nucleotide sequence queries with BLAST. Nucl Acids Res 2005;33:4335-4344.
- Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucl Acid Res 2005;33:e175.

Innovations in SNP Analysis

- Wang P, Dai M, Xuan W, McEachin RW, Jackson AU, Scott LJ, Athey B, Watson SJ, Meng F. SNP Function Portal: a web database for exploring the function implication of SNP alleles. 2006; ISMB2006/Bioinformatics (In press).
- Mohlke KL, Jackson AU, Scott LJ, Peck EC, Suh YD, Chines PS, Watanabe RM, Buchanan TA, Conneely KN, Erdos MR, Narisu N, Enloe S, Valle TT, Tuomilehto J, Bergman RN, Boehnke M, Collins FS. Mitochondrial polymorphisms and susceptibility to type 2 diabetes-related traits in Finns. Human Genetics 2005;118:245-254.

TMPRSS2/Prostate Cancer Progression

- Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310:644-648.
- Tomlins SA, Mehra R, Rhodes DR, Smith LR, Roulston D, Helgeson BE, Cao X, Wei JT, Rubin MA, Shah RB, Chinnaiyan AM. TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer. Cancer Res 2006;66:3396-3400.



See <http://www.ncibi.org/publications> for a list of 32 recent publications

National Center for Integrative Biomedical Informatics (NCIBI) External Advisory Board



Don Detmer, M.D.
*President and CEO
American Medical Informatics Association*



Mark H. Ellisman, Ph.D.
*Professor and Director
University of California, San Diego*



Franklyn G. Prendergast, M.D., Ph.D.
*Director, Mayo Clinic Center for the Research of
Individual*



Edward Lazowska, Ph.D.
*Bill & Melinda Gates Chair of Computer Science
University of Washington*



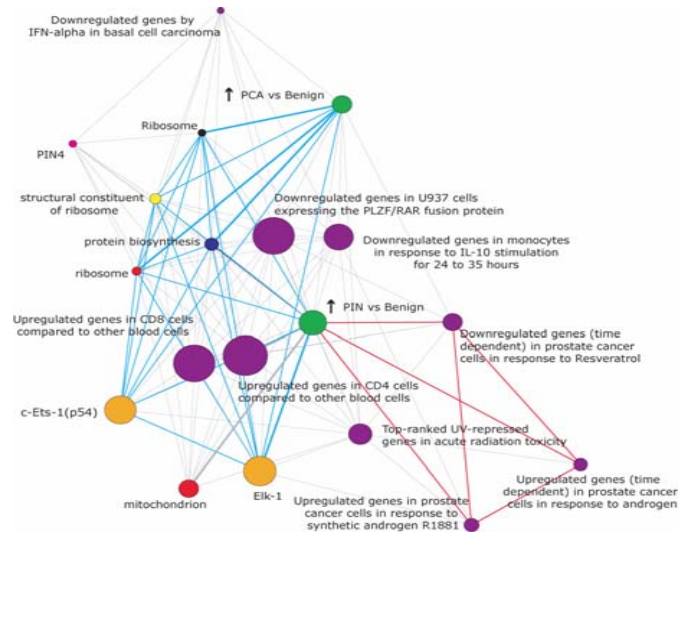
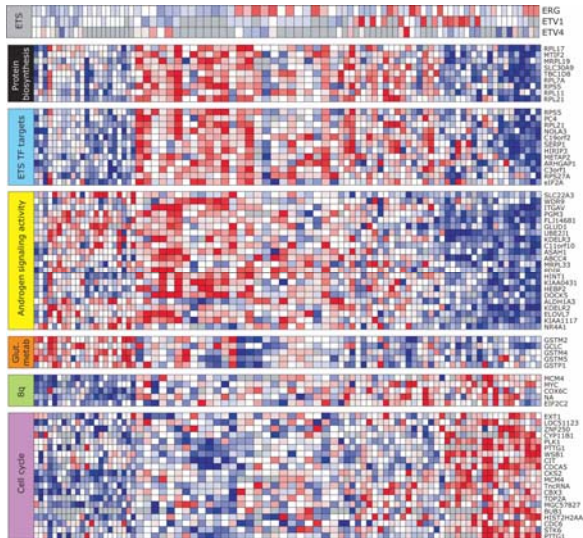
D.E. Shaw, Ph.D.
*Chairman of D. E. Shaw & Co., Inc.
Professor, MIT*



Opportunities to Discuss in More Depth

1) Tuesday 10AM – Noon “Applications of Systems Biology, Modeling, and Analysis” Work Group Meeting

2) NCIBI Dissemination Event “Computational Systems Biology to Accelerate Research in Complex Diseases; Diabetes and Prostate Cancer”



Back to the Future and the Challenges Ahead: From the NIH Roadmap Web Site

As the Centers begin to generate the software and data management tools to serve as fundamental building blocks for 21st century medical research, individual scientists will be funded to work together with the centers. **"Big science" and "small science" will work hand-in-hand** to advance all of biomedical research. Through these efforts, researchers will be able to share data gathered from large experiments. **The best minds will be able to work together effectively to tackle unsolved mysteries**, such as the role of heredity in individuals' different responses to medicines and the complex interplay of genetic and environmental factors in common diseases such as Alzheimer's disease, heart disease, cancer, and diabetes.

The Bioinformatics and Computational Biology initiatives will create a national software engineering system. Through a computer-based grid, biologists, chemists, physicists, computer scientists, and physicians anywhere in the country will be able to share and analyze data using a common set of software tools. **Developers of the project envision that the system will resemble that of the integrated software packages for office tools installed on most home computers today, in which information can be traded seamlessly between software such as spreadsheets, word-processing and e-mail programs.**

The URL for the NIH Roadmap Web site is nihroadmap.nih.gov. For more information on the Bioinformatics and Computational Biology initiatives, contact C. John Whitmarsh, Ph.D., National Institute of General Medical Sciences, (301) 451-6446,



Special Thanks

- NCIBI Program Officer (PO) – Dr. Karen Skinner, NIDA
- NCIBI Lead Science Officer (LSO) – Dr. Donald Jenkins, NLM
- SDIWG Chair and NCBC Leader, Dr. Peter Lyster
- Acting Director, Center for Bioinformatics and Computational Biology, NIGMS; Dr. John Whitmarsh



NCIBI Team Members Present



Robert Murphy, Ph.D.
Carnegie Mellon University
NCIBI Subcontractor



Scott A. Tomlins, Ph.D.
University of Michigan
NCIBI Core 3



Jinesh Patel, Ph.D.
Co-I Core 1, NCIBI
University of Michigan



Peter J. Woolf, Ph.D.
Core 2 & 3, NCIBI
University of Michigan



Michael Reich, Ph.D.
Broad Institute, MIT



NCIBI Team Members Present (cont.)



Matthias Kretzler, M.D.
Core 3, Diabetes complications



Fan Meng Ph.D.
Co-I: Cores 2 and 3



Jill Mesirov, Ph.D.
*NCIBI Subcontractor
MIT/Broad Institute*



Barbara Mirel, Ph.D.
Co-I: Core 5, Evaluation

