



The Centre for Molecular Medicine and Therapeutics



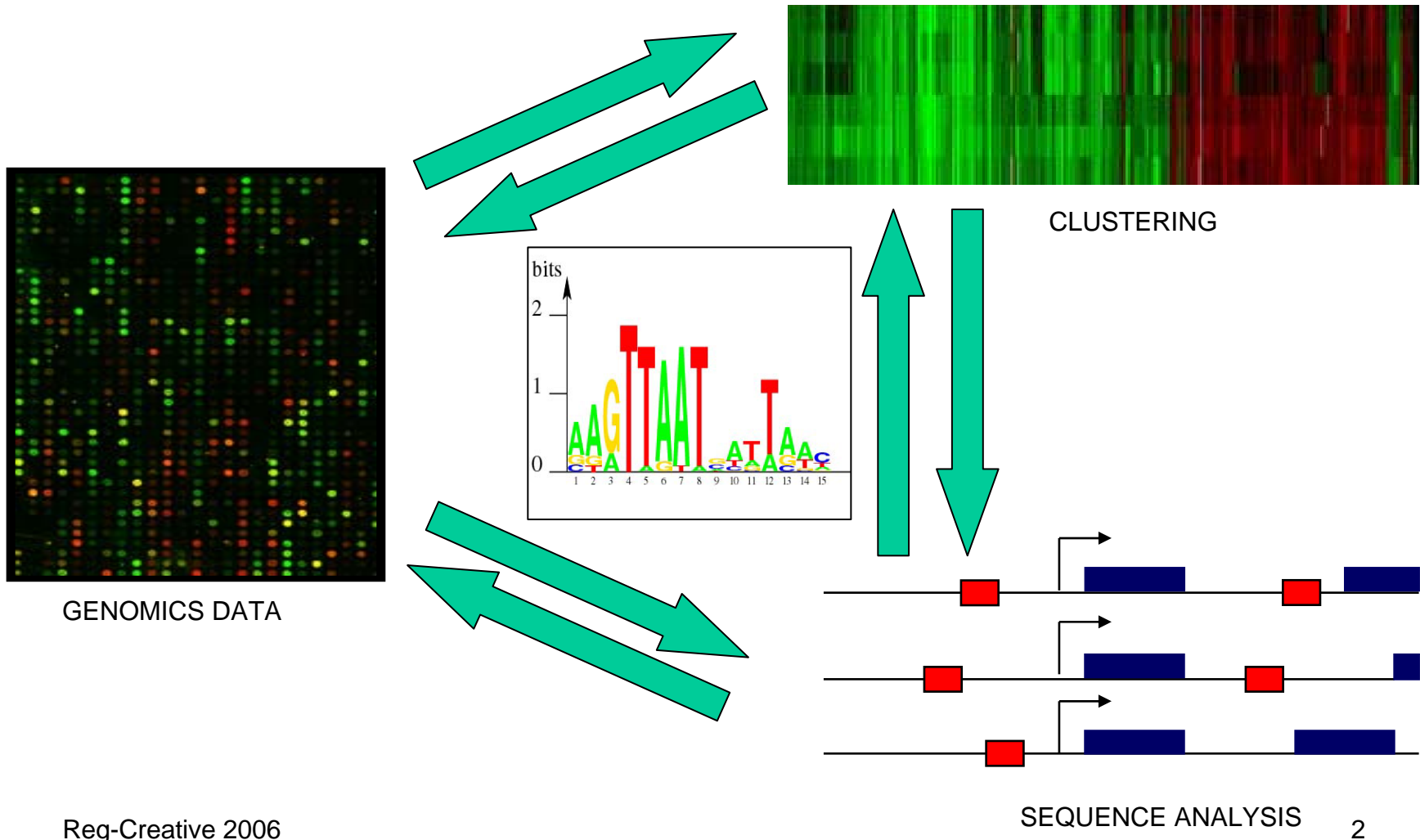
# JASPAR, TFCAT and PAZAR

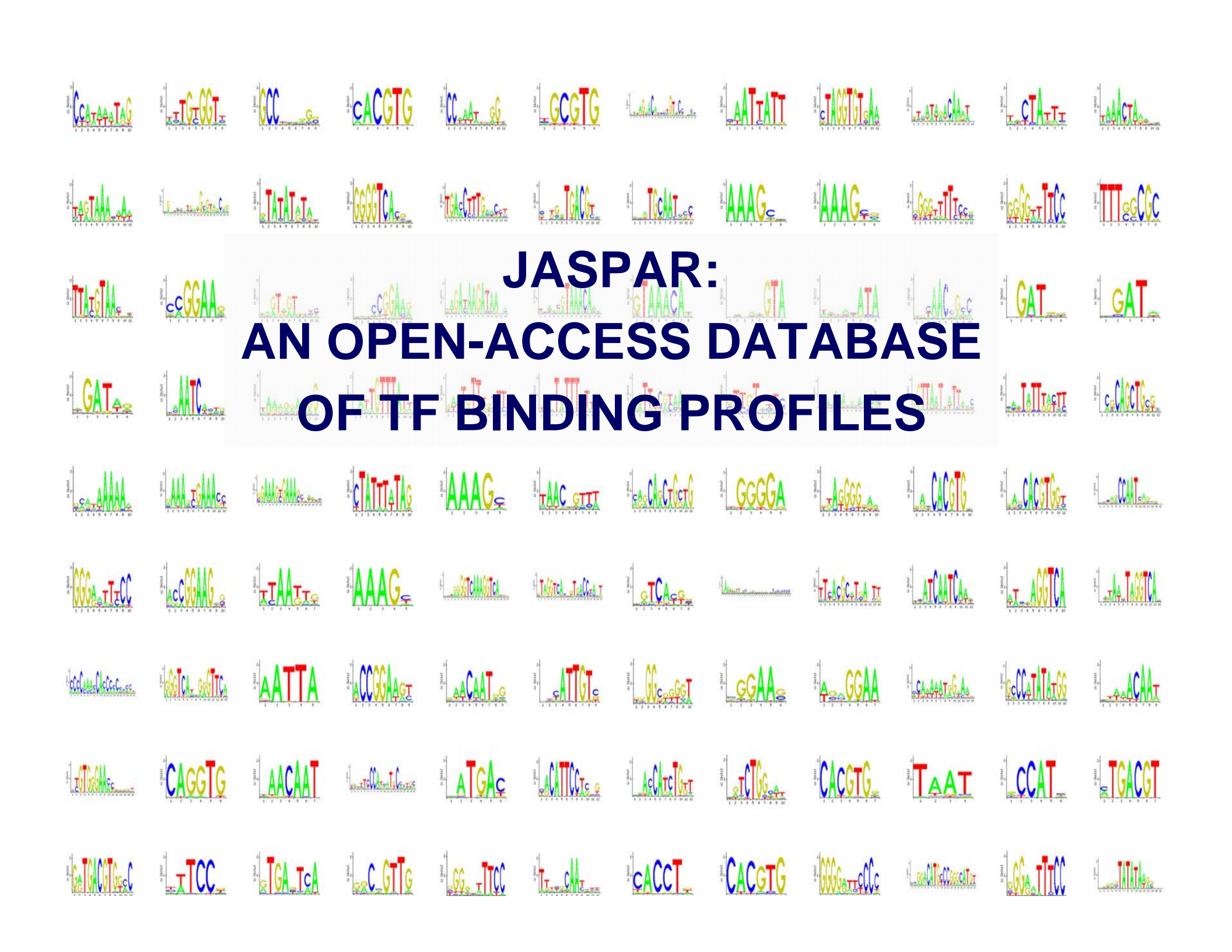
**Wyeth W. Wasserman**

University of British Columbia

[www.cisreg.ca](http://www.cisreg.ca)

# Defining Cis-Regulatory Mechanisms for Co-Expressed Genes

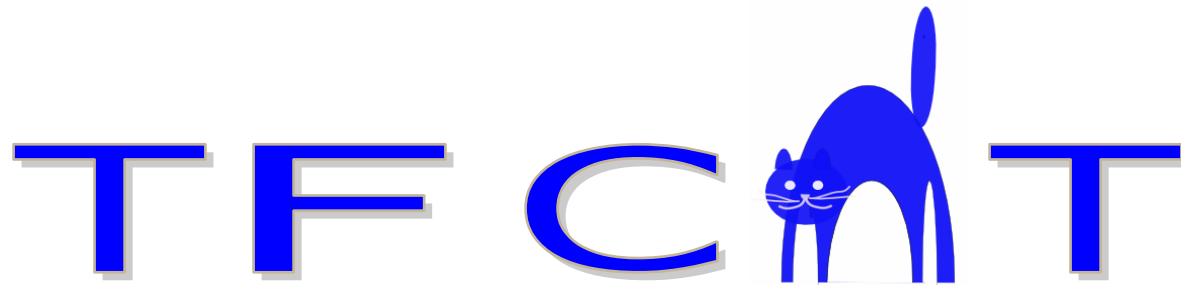




# JASPAR: AN OPEN-ACCESS DATABASE OF TF BINDING PROFILES

# Data Challenges

- Need larger and more complete collections of TFBS Profiles and Regulatory Sequence Annotation
- Need annotated catalog of TFs both for evaluation of results and for selection of candidate members from families of TFs with similar target site recognition
- Need larger compendium of reference collections for evaluation of system performance



# TF Catalog – Taking inventory of mouse and human TFs

Debra Fulton and Wyeth Wasserman (UBC)

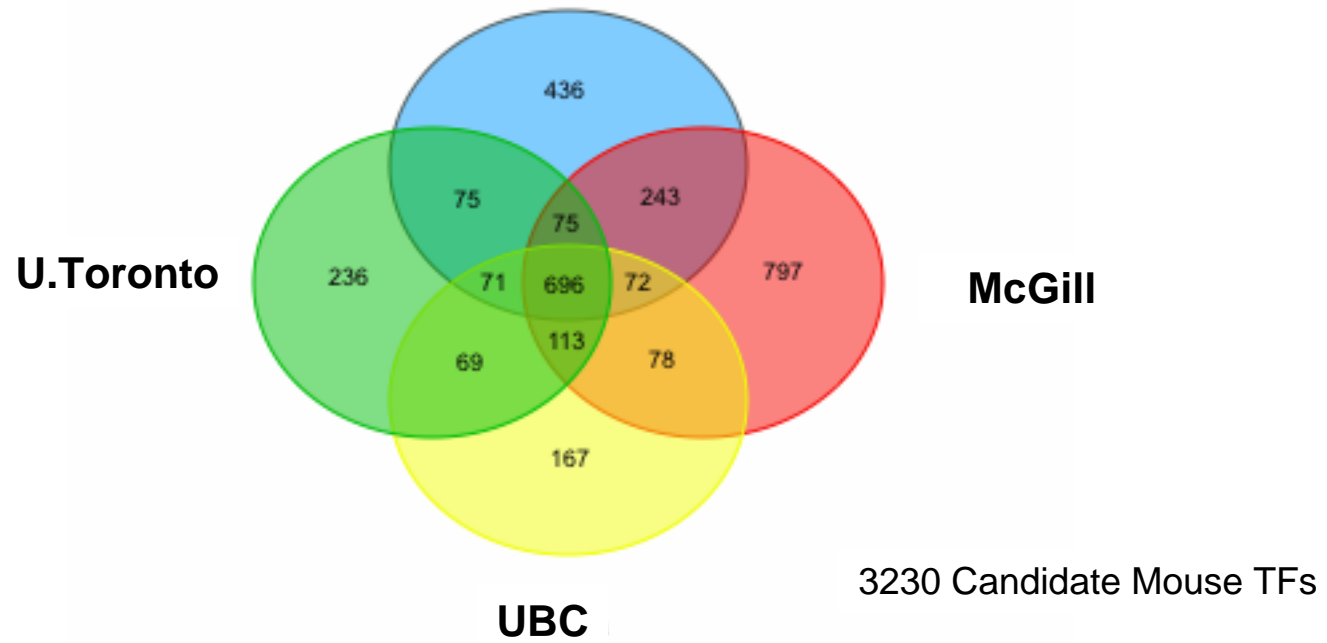
Jared Roach (ISB)

Gwenael Breard and Tim Hughes (UoT)

Sarav Sundararajan and Rob Sladek (QGC/McGill)

# T F C T

ISB



# TFCat Review Process

- Genes reviewed: 841
  - Assign category/judgement
  - Link PMIDs for category basis
  - Set biased for TFs with available literature
- Positive TF 82%
- DNA Binding 63%
  - Sequence-specific subset 92%
- Independent re-review process

# DBD Super Class Taxonomy

(Luscombe/Thornton)

**BASIC DOMAIN** (BD) proteins which include a basic DNA binding domain region;

**BETA SCAFFOLD** (BS) characterized by large beta sheets structures used to bind DNA ;

**ZINC CLUSTERING** (ZC) composed of tetrahedral coordination of 1 or 2 zinc ions by conserved cysteine and histidine residues;

**HELIX TURN HELIX**(HTH) two alpha helices connected by a beta turn or longer linkers such as loops;

**WINGED HELIX TURN HELIX** (WHTH) extension of HTH but includes a third alpha helix and an adjacent beta sheet;

**OTHER ALPHA HELIX** (OAH) all proteins that use alpha-helices as method for DNA binding;

**OTHER** (O) this superclass accommodates all other DNA-binding structures

# Extensions to Luscombe Taxonomy

- 1.1) Homeodomain-like
  - 100) Myb Domain Family
- 1.1) Helix-Turn-Helix
  - 101) GTF2I
- 1.2) Winged Helix-Turn-Helix
  - 102) Forkhead Domain Family
- 1.2) Winged Helix-Turn-Helix
  - 103) RFX Domain Family
- 2.1) Zinc-coordinating Group
  - 104) GATA Domain Family
- 2.1) Zinc-coordinating Group
  - 105) Glial Cells Missing (GCM) Domain Family
- 2.1) Zinc-coordinating Group
  - 106) SMAD MH1 Domain
- 4) Other Alpha-Helix Group
  - 28) High Mobility Group-Box Family
- 4) Other Alpha-Helix Group
  - 107) Sand Domain Family
- 6) Beta Hairpin\_Ribbon Group
  - 108) Methyl-CpG-binding domain, MBD family
- 7) Other
  - 109) High Mobility Group HMG-AT-hook Family
- 7) Other
  - 110) Runt Domain Family
- 7) Other
  - 111) IPT/TIG Domain Family

# C - I - a - s - s - i - f - i - c - a - t - i - o - n

Protein Group	Protein Group Description	Family	Family Description	TF Count
1.1	Helix-Turn-Helix	101	GTF2I	6
1.1	Helix-Turn-Helix Group	100	Myb Domain Family	19
1.1	Helix-Turn-Helix Group	2	Homeodomain Family	122
1.2	Winged Helix-Turn-Helix	102	Forkhead Domain Family	19
1.2	Winged Helix-Turn-Helix	103	RFX Domain Family	2
1.2	Winged Helix-Turn-Helix	13	Interferon Regulatory Factor	6
1.2	Winged Helix-Turn-Helix	15	Transcription Factor Family	8
1.2	Winged Helix-Turn-Helix	16	Ets Domain Family	15
2	Zinc-coordinating Group	104	GATA Domain Family	8
2	Zinc-coordinating Group	105	Glial Cells Missing (GCM Domain Family)	2
2	Zinc-coordinating Group	106	SMAD MH1 Domain	5
2	Zinc-coordinating Group	17	BetaBetaAlpha-zinc finger family	370
2	Zinc-coordinating Group	18	Hormone-nuclear Receptor Family	34
2	Zinc-coordinating Group	19	Loop-Sheet-Helix	1
3	Zipper-Type Group	21	Leucine Zipper Family	53
3	Zipper-Type Group	22	Helix-Loop-Helix Family	44
4	Other Alpha Helix Group	29	MADS Box Family	4
4	Other Alpha-Helix Group	107	Sand Domain Family	3
4	Other Alpha-Helix Group	28	High Mobility Group (HMG-box Family)	18
5	Beta-sheet group	30	TATA box-binding family	2
6	Beta Hairpin_Ribbon Group	108	Methyl-CpG-binding domain, MBD family	1
6	Beta-Hairpin_Ribbon	34	Transcription Factor T-Domain	10
7	Other	109	High Mobility Group HMG-AT-hook Family	1
7	Other	110	Runt Domain Family	2
7	Other	111	TIG Domain Family	8
7	Other	37	Rel Homology Region Family	7
7	Other	38	Stat Protein Family	5
8	Enzyme Group	47	DNA Polymerase-Beta Family	7

15%

47%

04%

12%

# TFCat Summary

- Collection available
  - Ongoing curation
- Website release pending
  - Building WIKI to collect user feedback
- Linking to PAZAR
- Questions? Debra Fulton is here



## Open-access regulatory sequence repository – an information mall

Elodie Portales-Casamar  
Jonathan Lim  
Stefan Kirov  
Jay Snoddy  
Wyeth Wasserman

# Numerous Regulatory Databases – No Coordination



Transcriptional Regulatory Element Database



TRANSCRIPTION REGULATORY  
REGIONS DATABASE

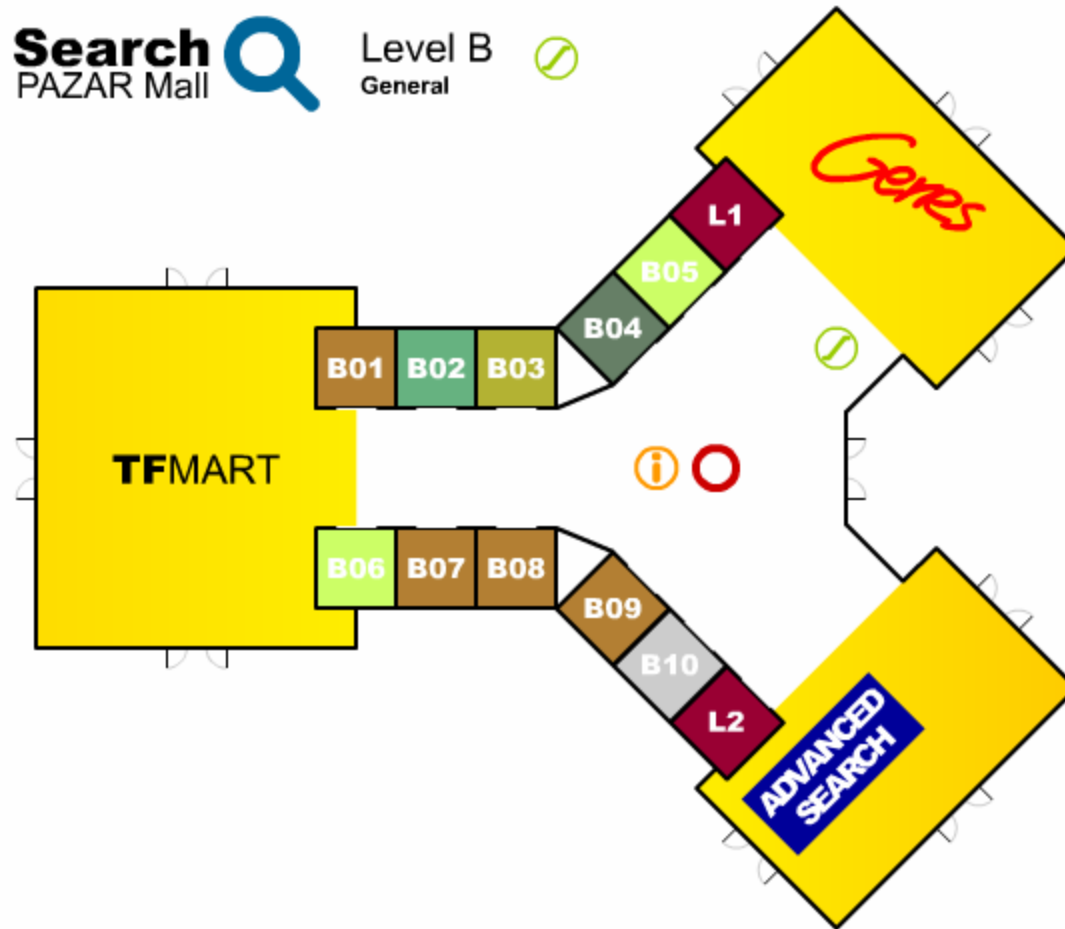


# PAZAR



Grand Bazaar, Istanbul

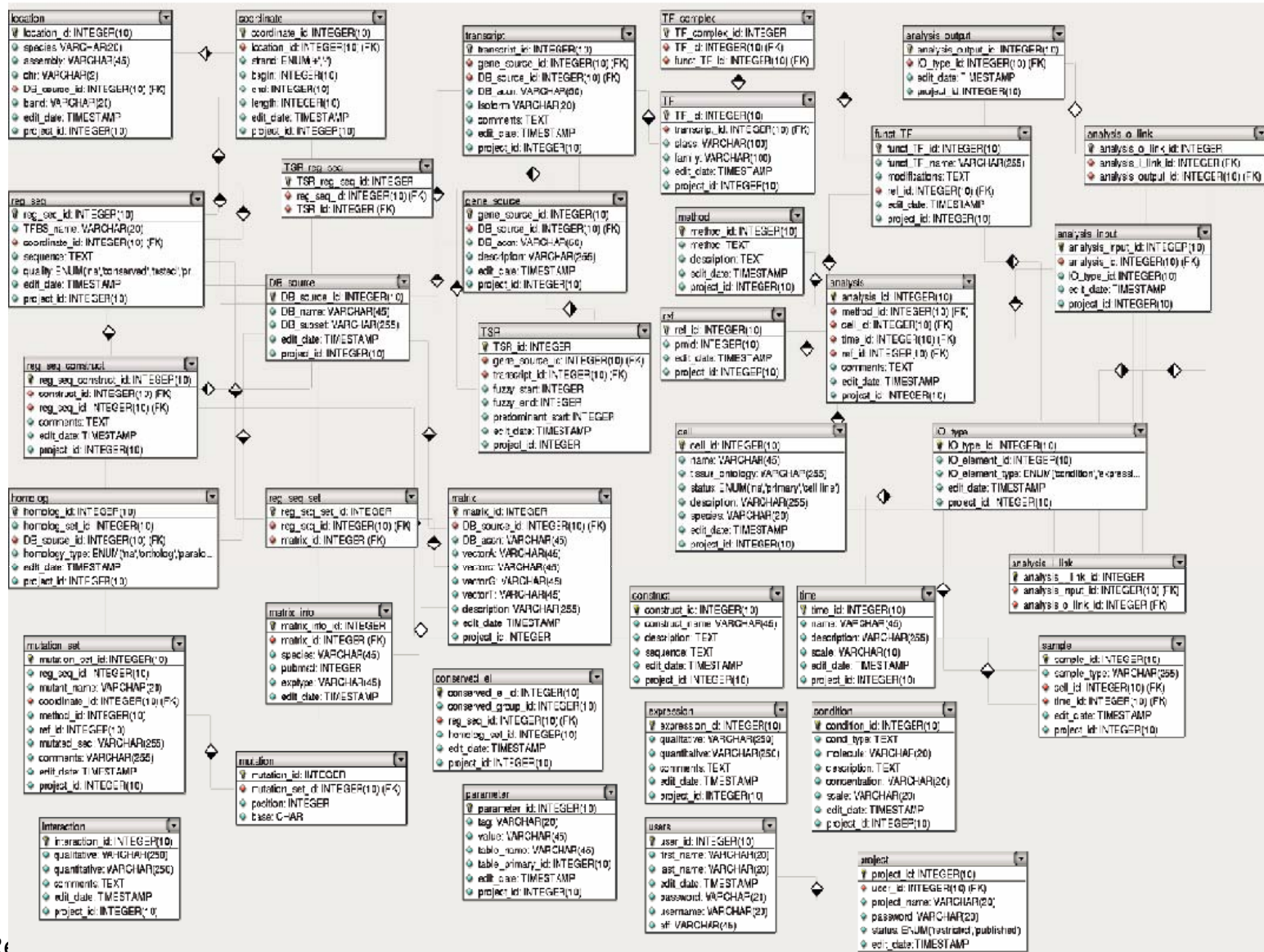
# Retrieval/Browsing Interface



# Highlights

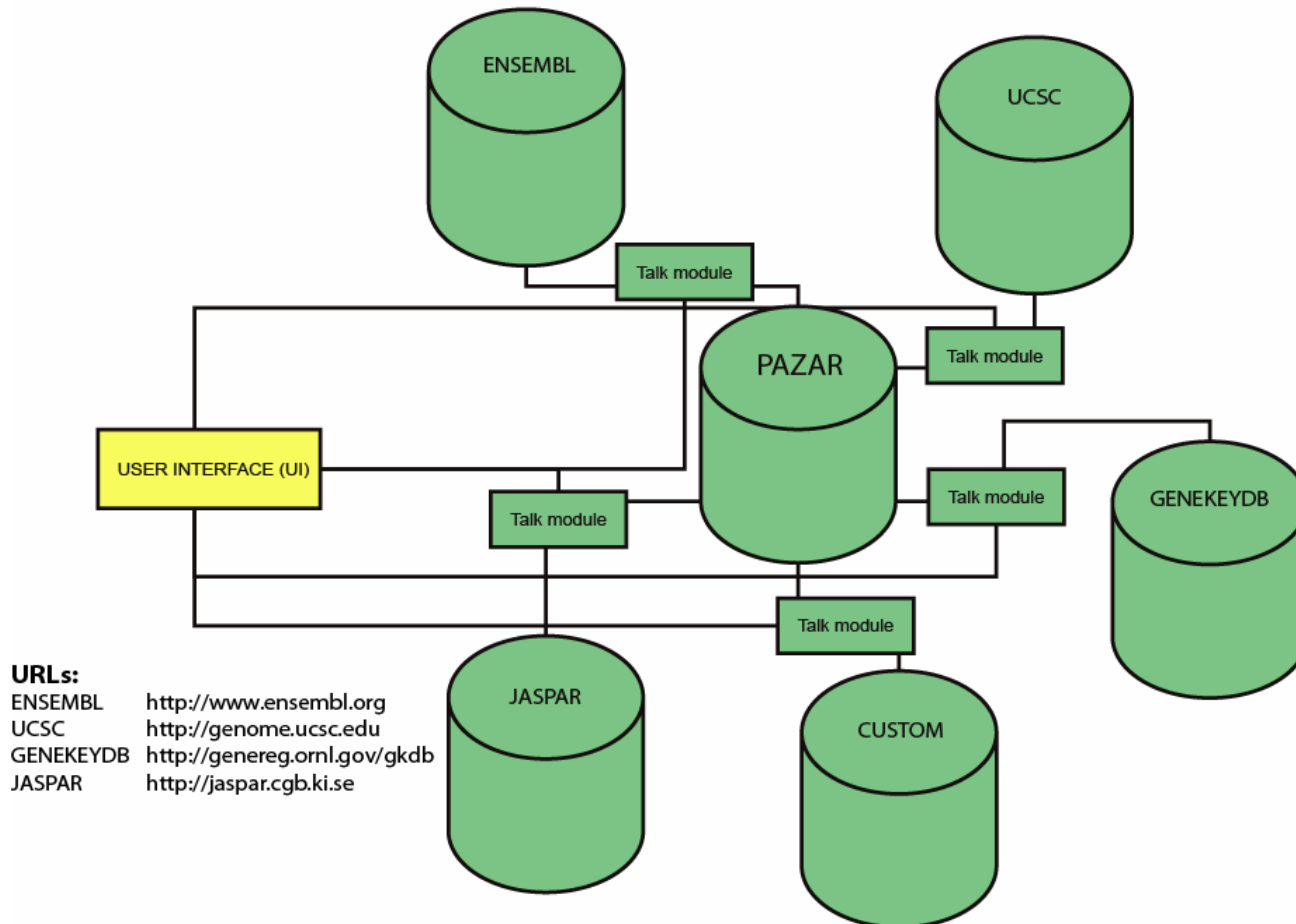
- Available: [www.pazar.info](http://www.pazar.info)
- All data linked to genome assemblies available in Ensembl (limiting species)
- Three project classes
  - Open – you can modify data
  - Published – you can read (and copy) everything
  - Restricted – only owner-approved users
- Open-Access/Open-Software
  - Code in sourceforge
  - Data can be extracted from “open” and “published” projects

# A complex database schema to allow flexibility

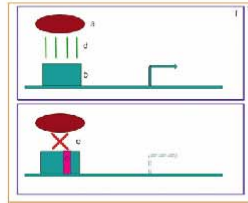


## PAZAR can be linked to external data resources (ensembl, genekeydb) using a “talk” module

PAZAR is confined to the description of regulatory sequence features. There is often need for other information, such as gene identifiers, genomic DNA sequence, etc. The API talk module grants access to external resources. It is easily extensible to support other databases, including new “malls”, while providing standard accessor methods.



## XML exchange format



Relationship Between Entities and Tables	
Entity	Table(s) involved
a Transcription Factor (TF)	TF, transcript, gene_source
b Transcription Factor Binding Site (TFBS)	reg_seq, TSR, gene_source
c TFBS Mutation	mutation, mutation_set
d TF - TFBS Interaction (induction of expression)	interaction / expression
e No TF - TFBS Interaction (no induction of expression)	interaction / expression
f Experiment Description	method, ref, cell, time, condition
g Project Description	project, users

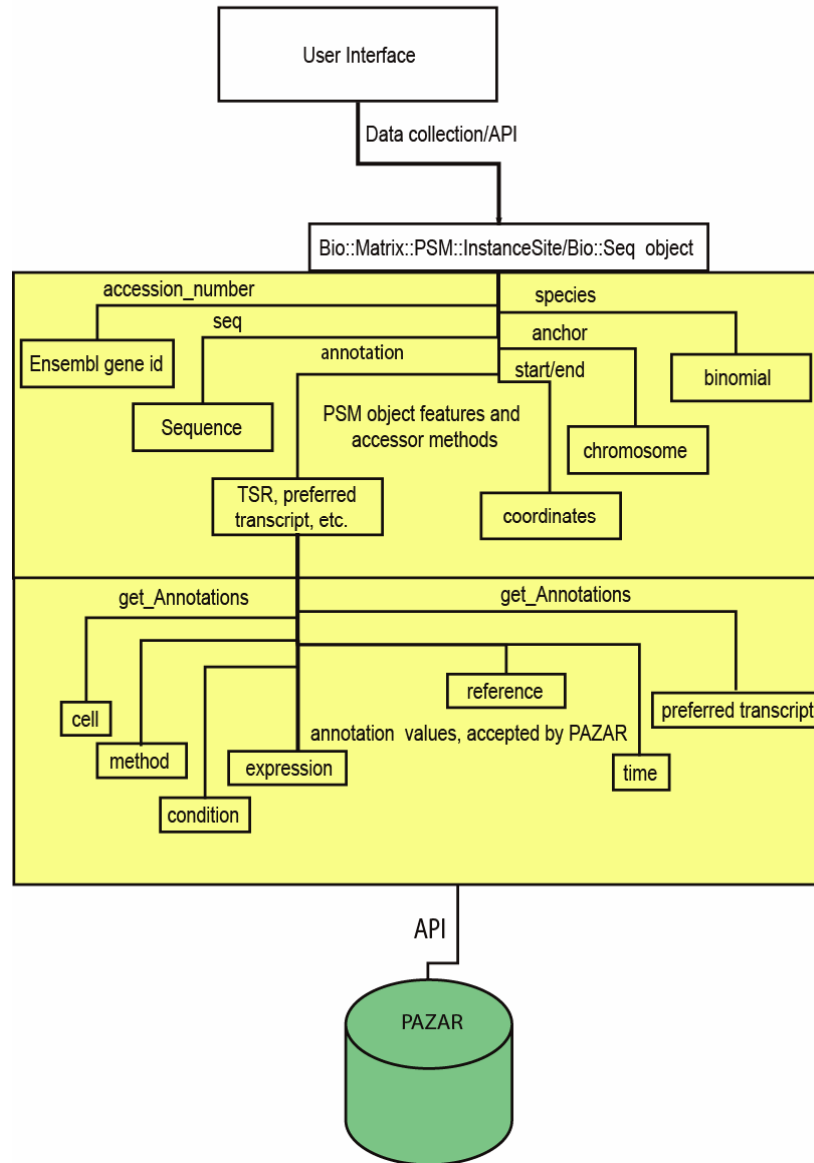
```

<pazar>
  <project name="example_project" pazar_id="project_0" status="restricted">
    <user affiliation="affiliation" first_name="first_name" last_name="last_name"
pazar_id="user_0" username="username"/>
  </project>
  <data>
    <gene_source description="PDE6B" pazar_id="gene_0">
      <db_accession db_accn="ENSG00000133256" db_name="ensembl" />
      <TSR fuzzy_start="609373" fuzzy_end="609373" pazar_id="TSR_0">
        <transcript pazar_id="transcript_0">
          <db_accession db_accn="ENST00000255622" db_name="ensembl" />
        </transcript>
        <reg_seq TFBS_name="NRE" quality="tested" pazar_id="reg_seq_0"
sequence="ATTGTAGGAGTGAGTCAGCTGACCCGC">
          <coordinate begin="609283" end="609310" length="28" strand="+">
            <location assembly="NCBI 35" band="4p16.3" species="human">
              <db_accession db_name="ensembl" />
            </location>
          </coordinate>
          <mutation_set pazar_id="mutation_set_0">
            <coordinate begin="609294" end="609299" length="6" strand="+">
              <location assembly="NCBI 35" band="4p16.3" species="human">
                <db_accession db_name="ensembl" />
              </location>
            </coordinate>
            <mutation base="g" position="1" pazar_id="mutation_0"/>
            ...
            <mutation base="a" position="6" pazar_id="mutation_5"/>
          </mutation_set>
        </reg_seq>
      </TSR>
    </gene_source>
    <gene_source description="NRL" pazar_id="gene_1">
      <db_accession db_accn="ENSG00000129535" db_name="ensembl" />
      <transcript pazar_id="transcript_1">
        <db_accession db_accn="ENST_00000250471" db_name="ensembl" />
        <tf class="bZIP" family="MAF" pazar_id="tf_0"/>
      </transcript>
    </gene_source>
    <funct_tf pazar_id="funct_tf_0" tf_ids="tf_0"/>
    <interaction qualitative="yes" pazar_id="interaction_0"/>
    <interaction qualitative="no" pazar_id="interaction_1"/>
  </data>
  <analysis >
    <method method="EMSA"/>
    <ref pmid="11438531"/>
    <cell name="Y79" species="human" status="cell_line"/>
    <input_output >
      <input inputs="funct_tf_0"/>
      <output outputs="interaction_0"/>
    </input_output>
    <input_output>
      <input inputs="mutation_set_0"/>
      <output outputs="interaction_1"/>
    </input_output>
  </analysis>
</pazar>

```

## API data structure

The API is based on existing Bioperl data structures and methods. Using Bioperl allows the PAZAR project to use standardized procedures.



# Some Statistics

- “Restricted” but going public soon
  - “PLEIADES PROJECT” NEURO GENES
    - Regulated Genes: 77
    - Regulatory sequence (genomic): 303
    - Transcription Factors: 78
    - Annotated Publications: 143
- “Published” projects include
  - JASPAR
  - Muscle
  - Liver
  - ARE collection

# Current Efforts

- Three full-time annotators at work
  - Pleiades collection
- Improving annotation interface
- Ontology links for expression
- TFCat integration
- Graphical display of annotations

# PAZAR and OREGANNO

- Different systems and intentions
  - PAZAR allows private curation projects
  - Differ in style of annotations
  - PAZAR data is not validated – you must choose data collections that you trust
  - PAZAR is a mall; OREGANNO is a super-store
- PAZAR allows for broad range of data
  - SELEX
  - Promoter deletion experiments
  - TF Complexes
  - Mutations
  - TSS definition/Alternative Promoters
- Working together
  - Ontologies
  - Data exchange

# Help?

- Text mining tools to accelerate annotation
- Graphical display of information in database
- Ontology building expertise
- Collaborative projects
  - Open to expansion and improvements to facilitate research projects
- Questions? Elodie Portales-Casamar is here

# Putting It All Together



# Thanks!

## THE AMAZING PEOPLE WHO DID THE WORK!

- **Elodie Portales-Casamar**
- **Debra Fulton**
- Jonathan Lim
- Stuart Lithwick
- Magdalena Swanson
- Amy Ticoll
- David Martin
- David Arenillas
- Jochen Brumm
- Alice Chou
- Shannan Ho Sui
- Andrew Kwon
- Dimas Yusuf
- Miroslav Hatas
- Dora Pak



- James Mortimer
- Brian Kennedy



- Jay Snoddy
- Stefan Kirov (BMS)

## FUNDING

- CIHR
- IBM
- MSFHR
- MerckFrosst
- GenomeBC
- GenomeCanada
- CFI
- BC Children's Hospital Foundation