High throughput comparative genomics using a Chado backend

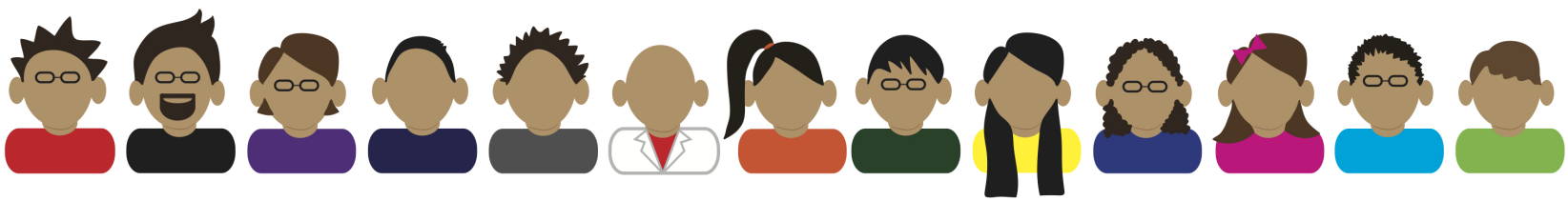
Mara Kim
mara.kim@vanderbilt.edu

GMOD 2014
Vanderbilt University

Jan 16, 2014

The Rokas Lab

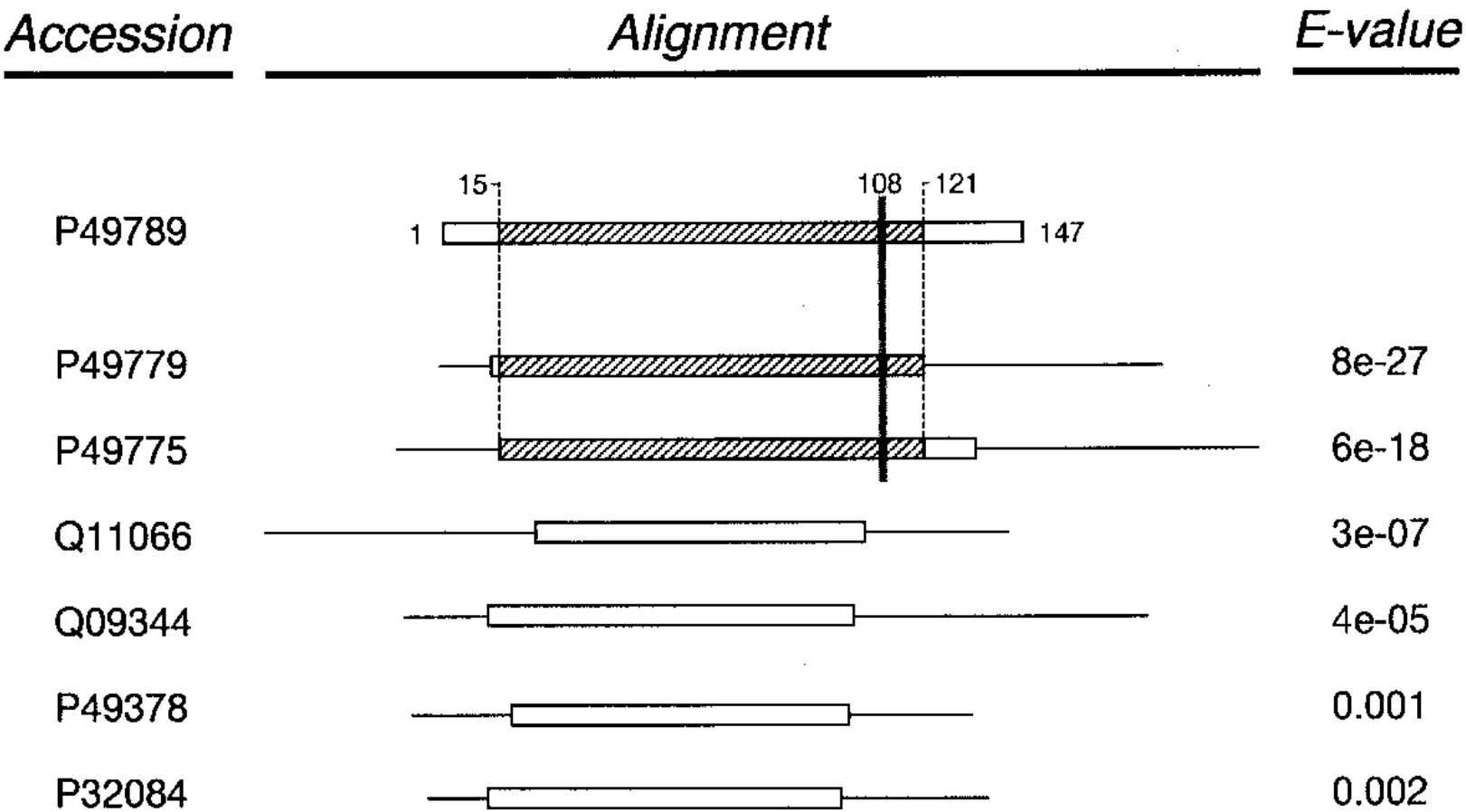
- Computational Genomics laboratory
  - Fungal genetics
  - Human preterm birth
- RokasDB - Comparative genomics database
  - 200+ eukaryote genomes
  - Modified Chado schema



Credit: Haley Eidem

Measures of genomic similarity

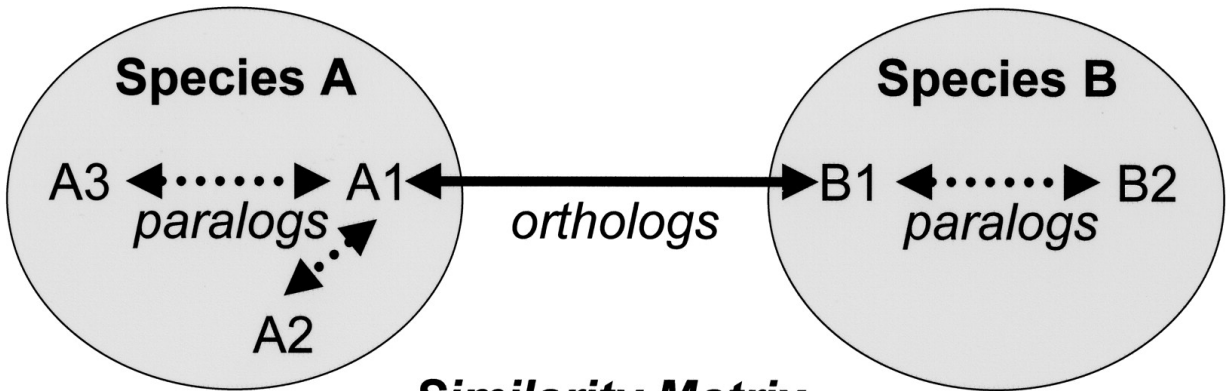
- Sequence similarity



S. F. Altschul et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic Acids Res.* 25.17 (1997), pp. 3389–3402

Measures of genomic similarity

- Ortholog estimation



Similarity Matrix

	A1	A2	A3	B1	B2
A1		***300***	***152***	***61***	**29**
A2	***233***		***150***	**60**	**29**
A3	***118***	***117***		**40**	**30**
B1	***69***	**68**	**45**		***88***
B2	**33**	**33**	**34**	***100***	

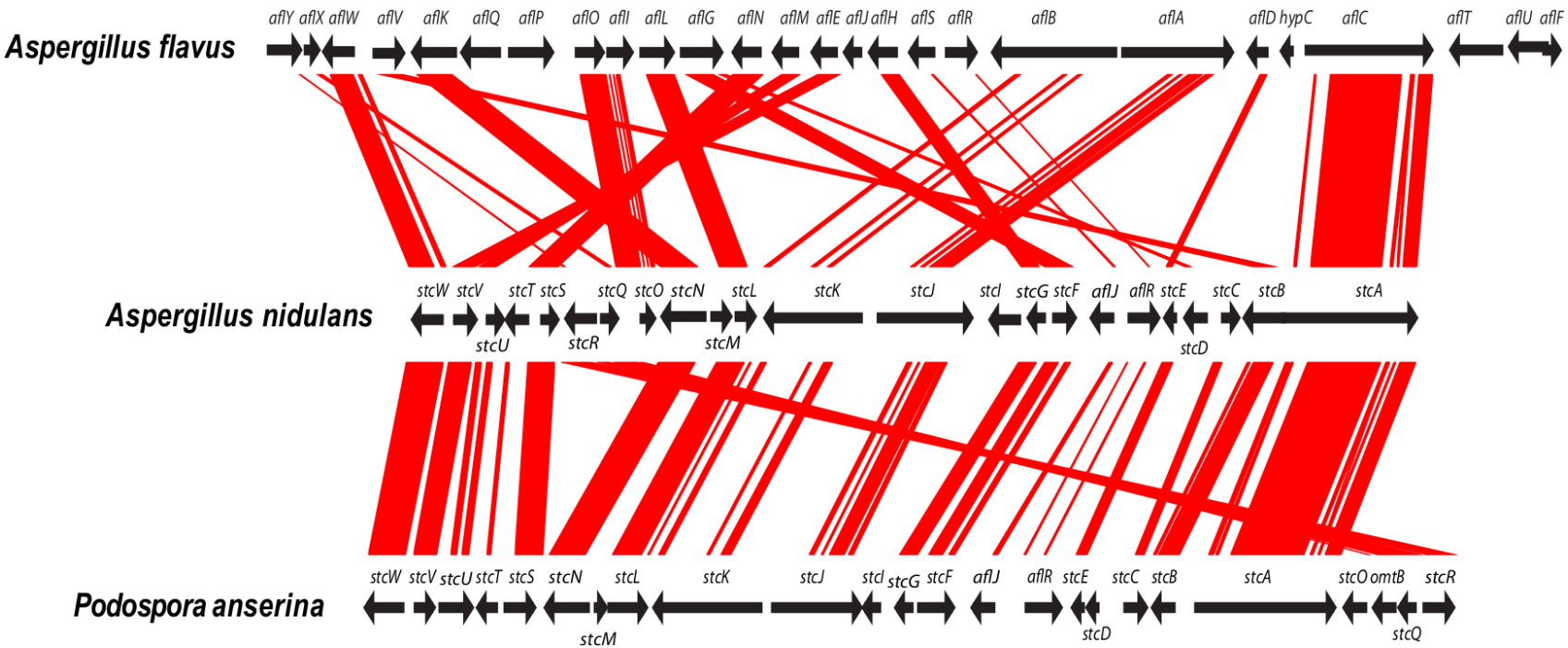
Normalized weights

$-\log_{10}$ (P-value)

L. Li, C. J. Stoeckert, and D. S. Roos. "OrthoMCL: identification of ortholog groups for eukaryotic genomes". In: *Genome Res.* 13.9 (2003), pp. 2178–2189

Measures of genomic similarity

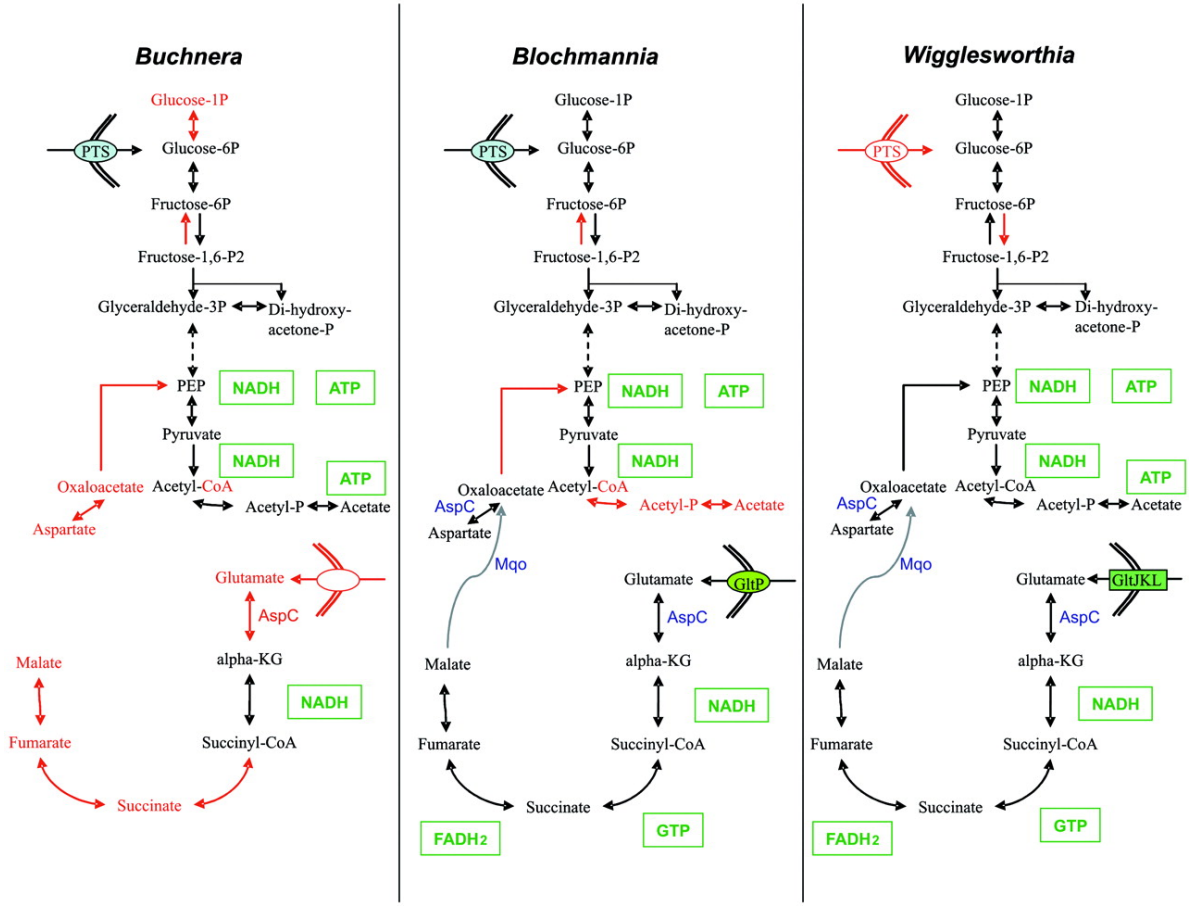
- Conservation of Synteny



J. C. Slot and A. Rokas. "Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi". In: *Curr. Biol.* 21.2 (2011), pp. 134–139

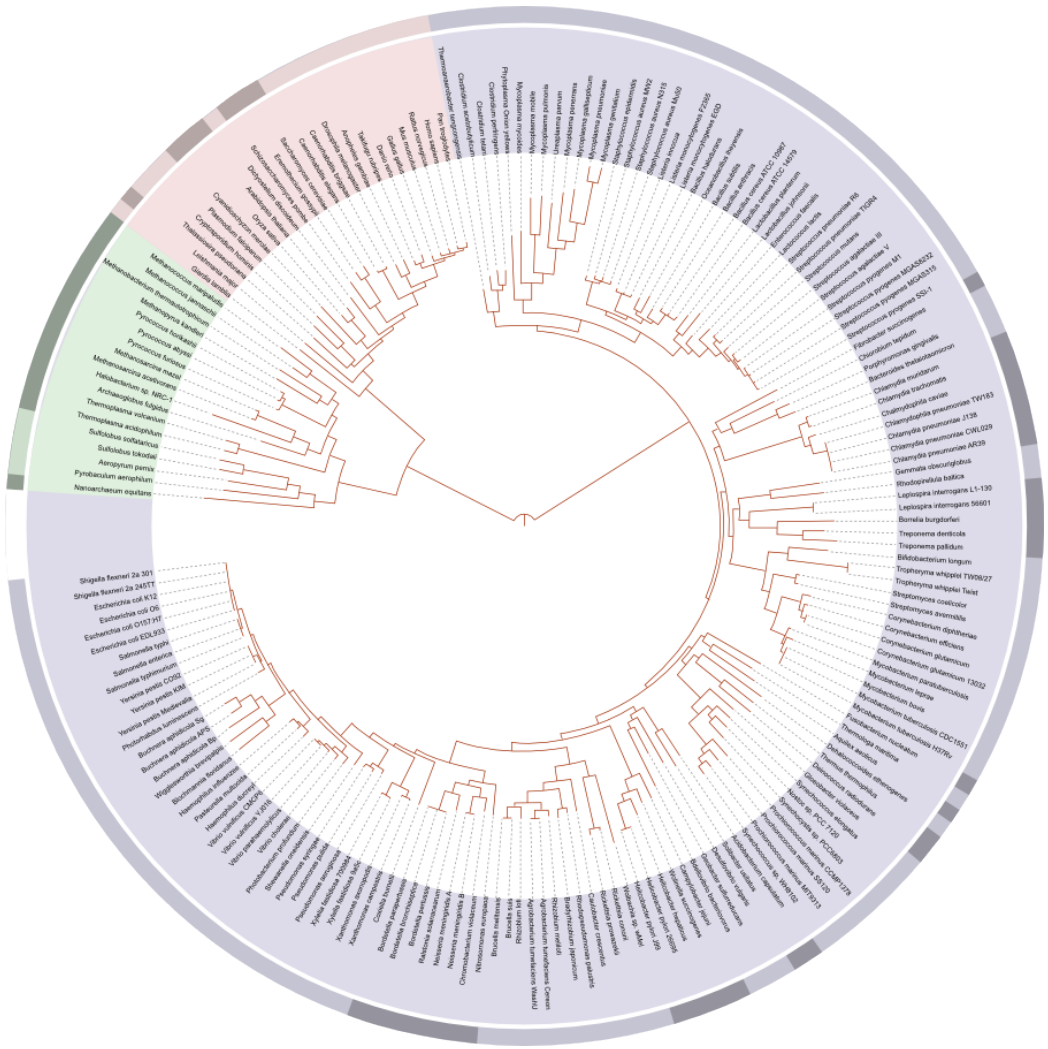
Measures of genomic similarity

- Functional similarity



E. Zientz, T. Dandekar, and R. Gross. "Metabolic interdependence of obligate intracellular bacteria and their insect hosts". In: *Microbiol. Mol. Biol. Rev.* 68.4 (2004), pp. 745–770

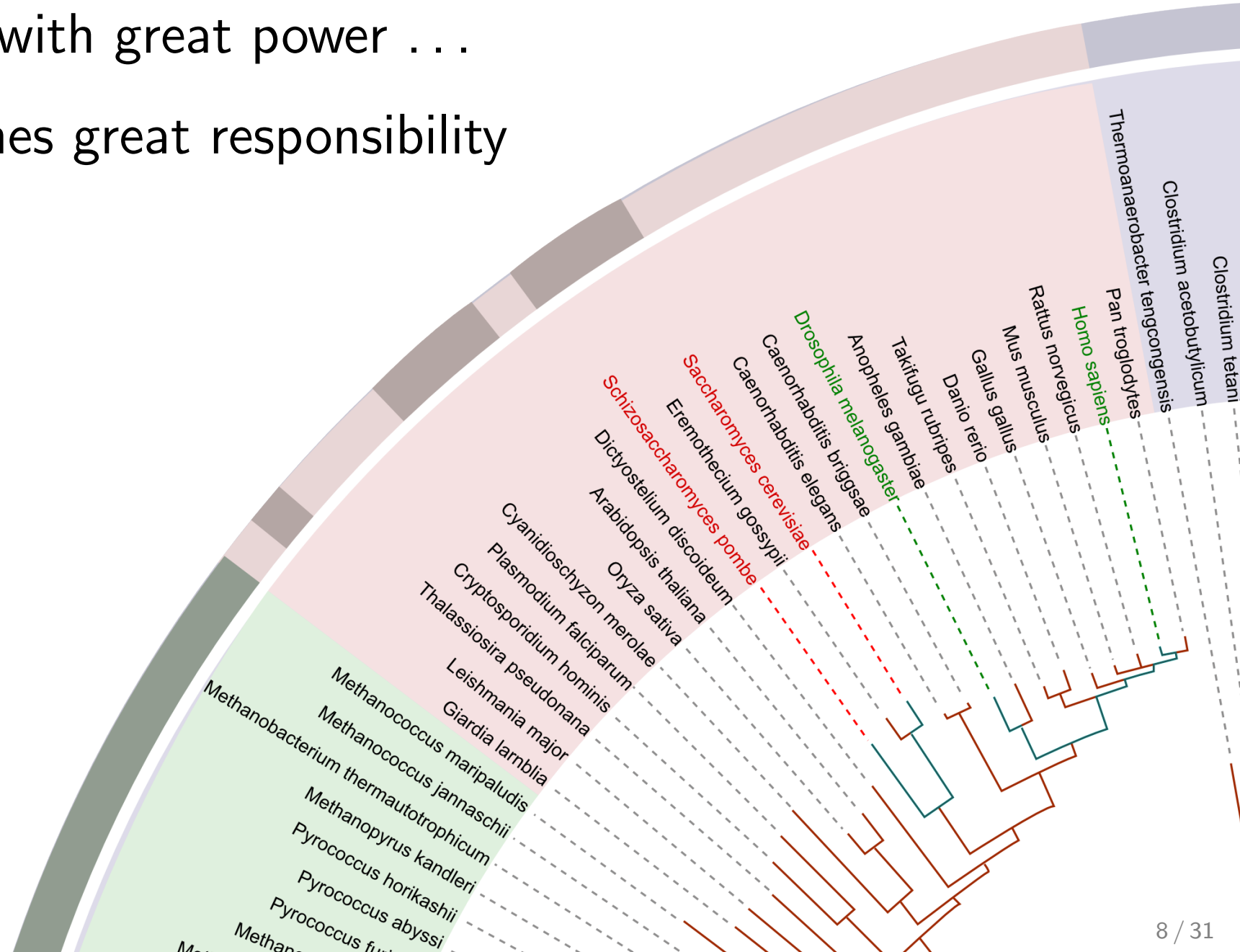
Trees: with great power . . .



F. D. Ciccarelli et al. "Toward automatic reconstruction of a highly resolved tree of life". In: *Science* 311.5765 (2006), pp. 1283–1287

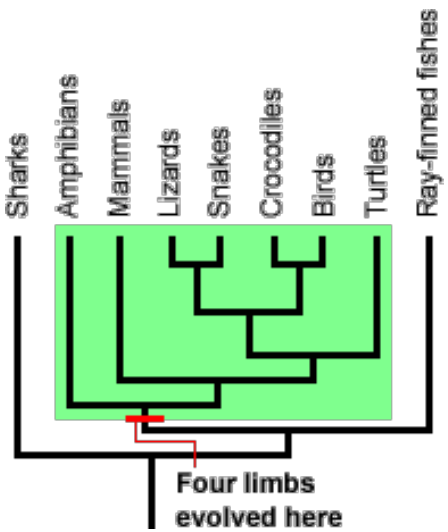
Trees: with great power . . .

comes great responsibility



Phylogeny vs. Similarity

- Phylogeny: shared ancestry (Darwinian homology)
- Similarity: common function (Owenian homology)
- Phylogenetic trees represent evolutionary history
- Relevance of model organisms depends on *similarity*
  - How do we quantify and visualize genomic similarity?



http://evolution.berkeley.edu/evolibrary/article/evo_08

Introducing. . .

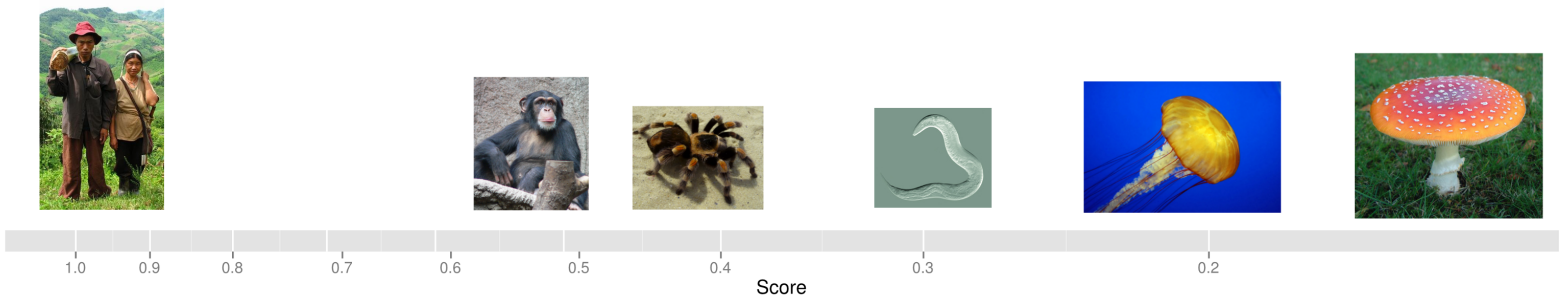
The Genome Yardstick

What is a Yardstick?

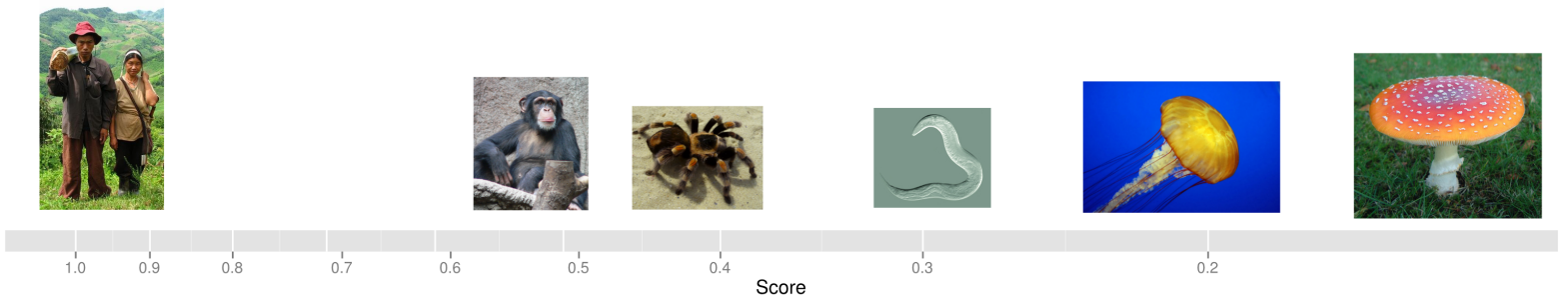
> **yard-stick** *n.* a rigid yard-measure; also *fig.*, **a standard of comparison**.
>
> Oxford English Dictionary. "yard, n.2". Oxford University Press.
> http://www.oed.com/view/Entry/231201?redirectedFrom=yardstick

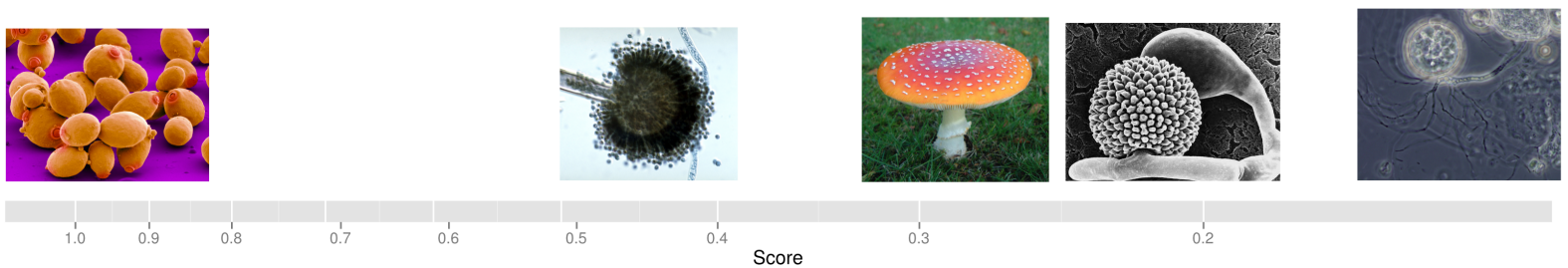
Our *standard of comparison* is the similarity between the human genome and other species which is used to gauge the similarity between two arbitrary genomes.

What is a Yardstick?



What is a Yardstick?

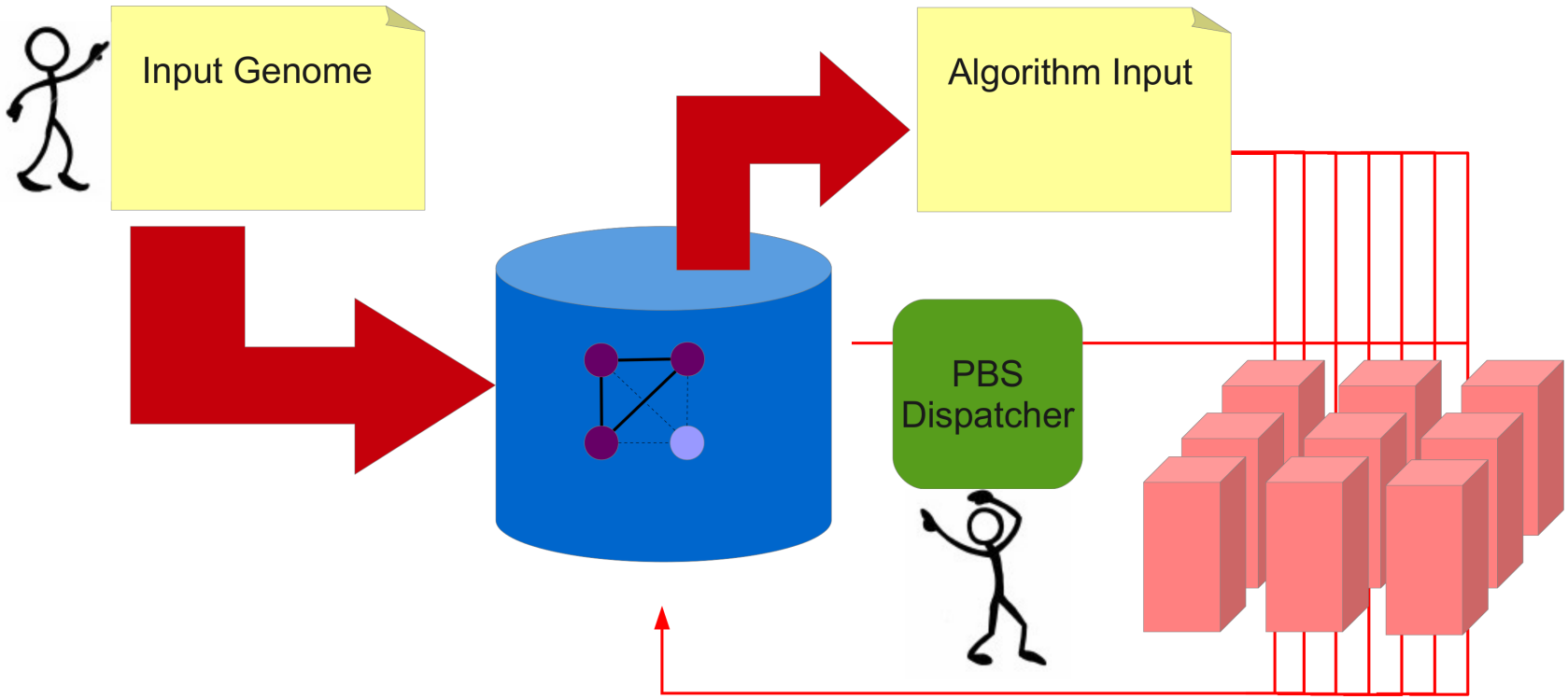




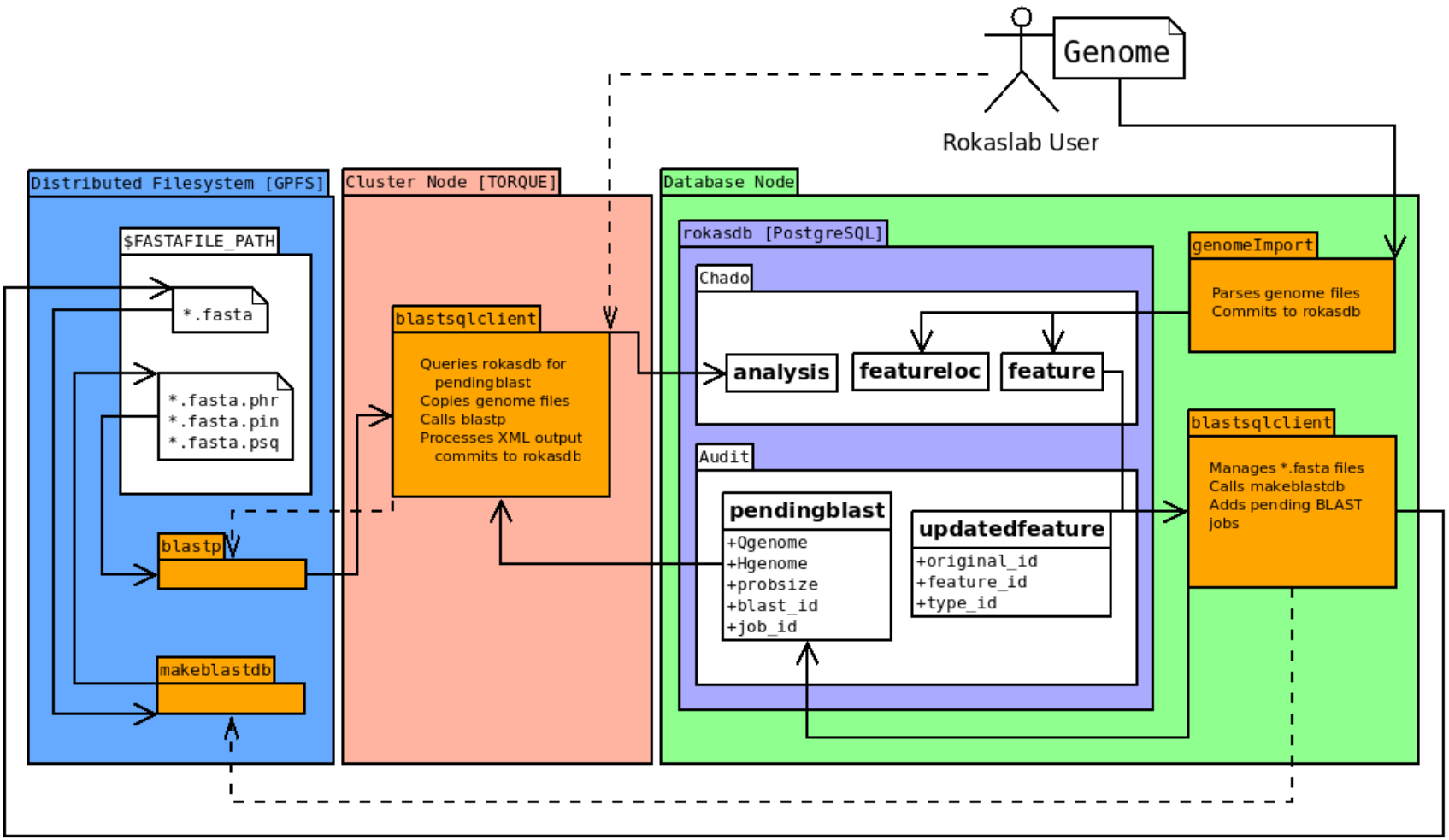
Goal

Quantify the molecular phenotype of the organisms on the yardstick

General Datapath



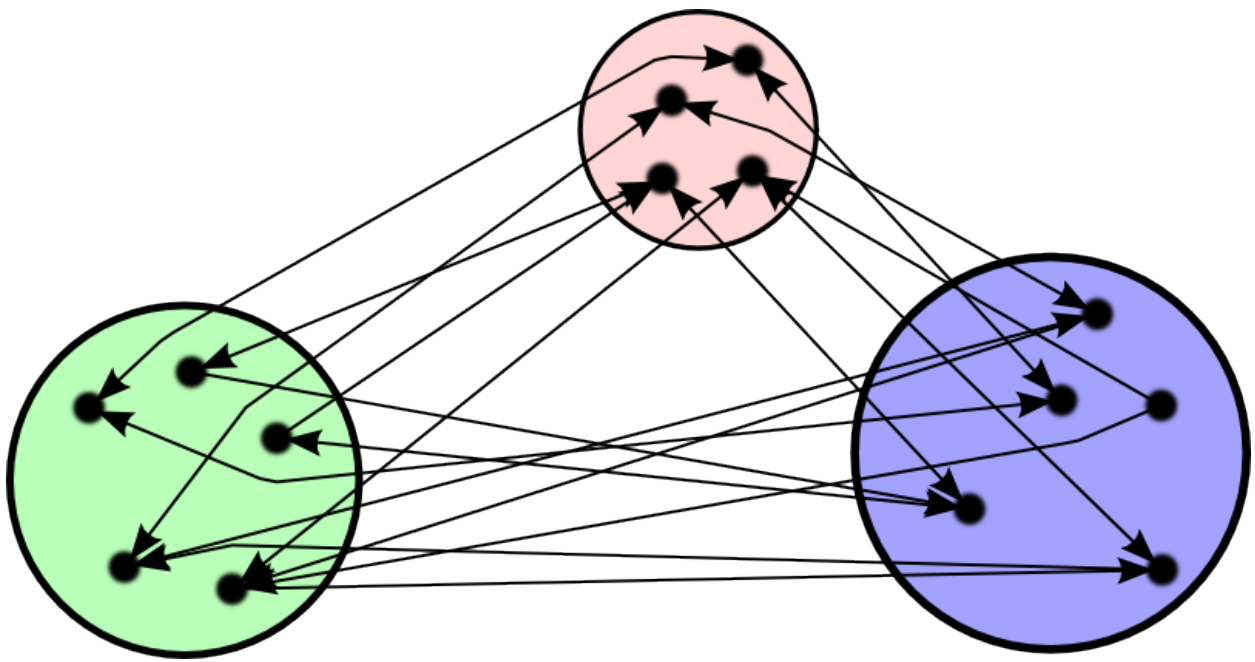
Blast Datapath



Score best hits

Score the best hit of every gene to every other genome

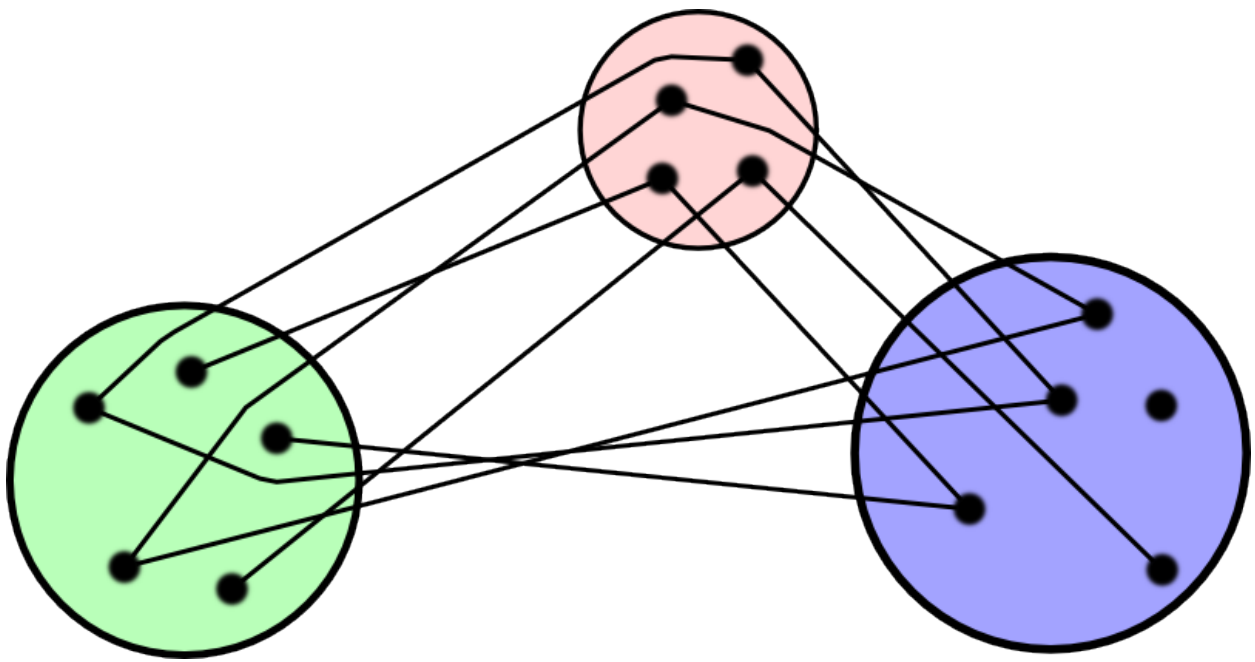
- Protein sequence identity
- Matrix adjusted protein identity (BLAST bitscore)



Score best hits

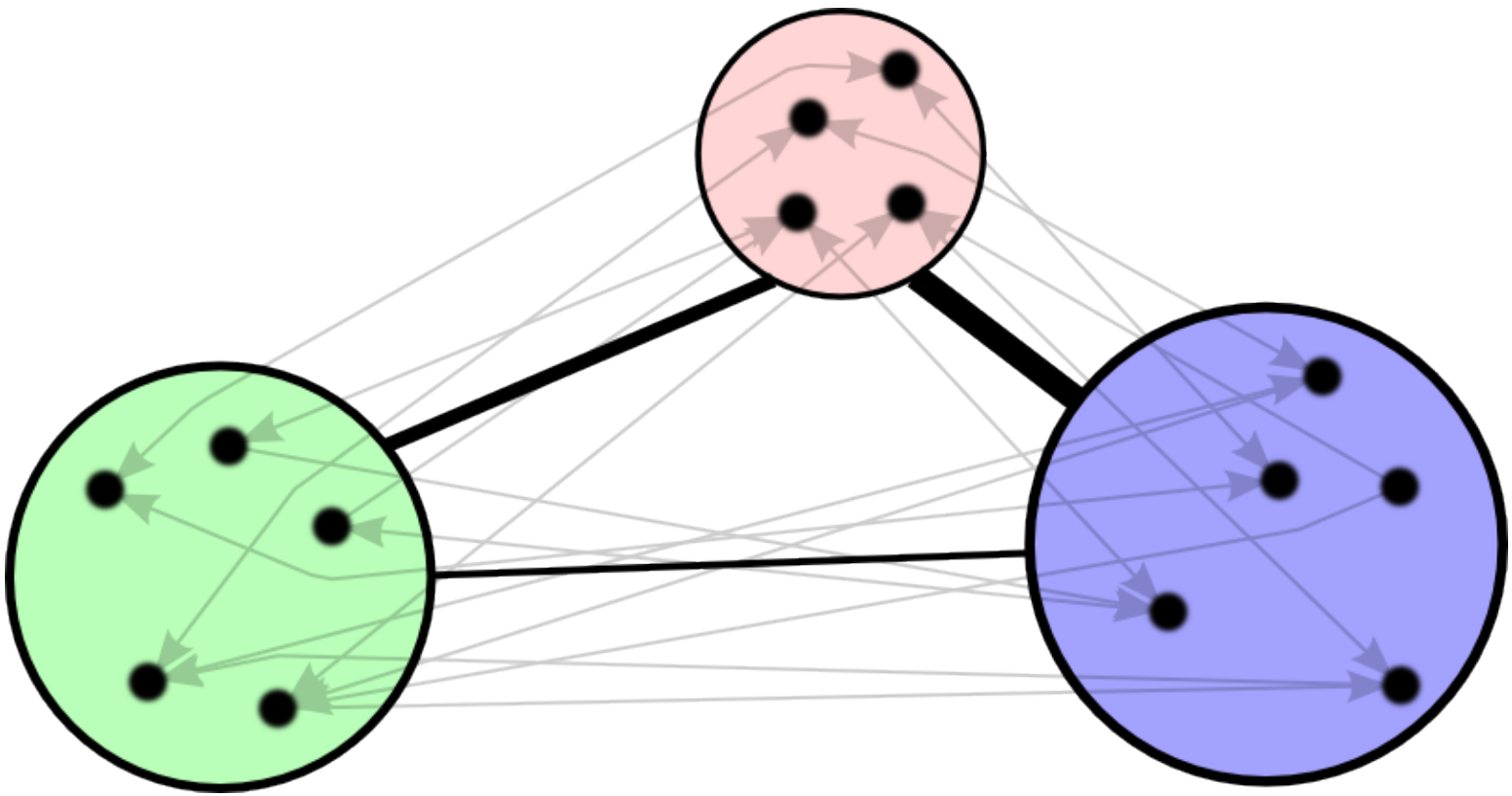
Score the best hit of every gene to every other genome

- Protein sequence identity
- Matrix adjusted protein identity (BLAST bitscore)
- Orthology estimation using reciprocal best hit

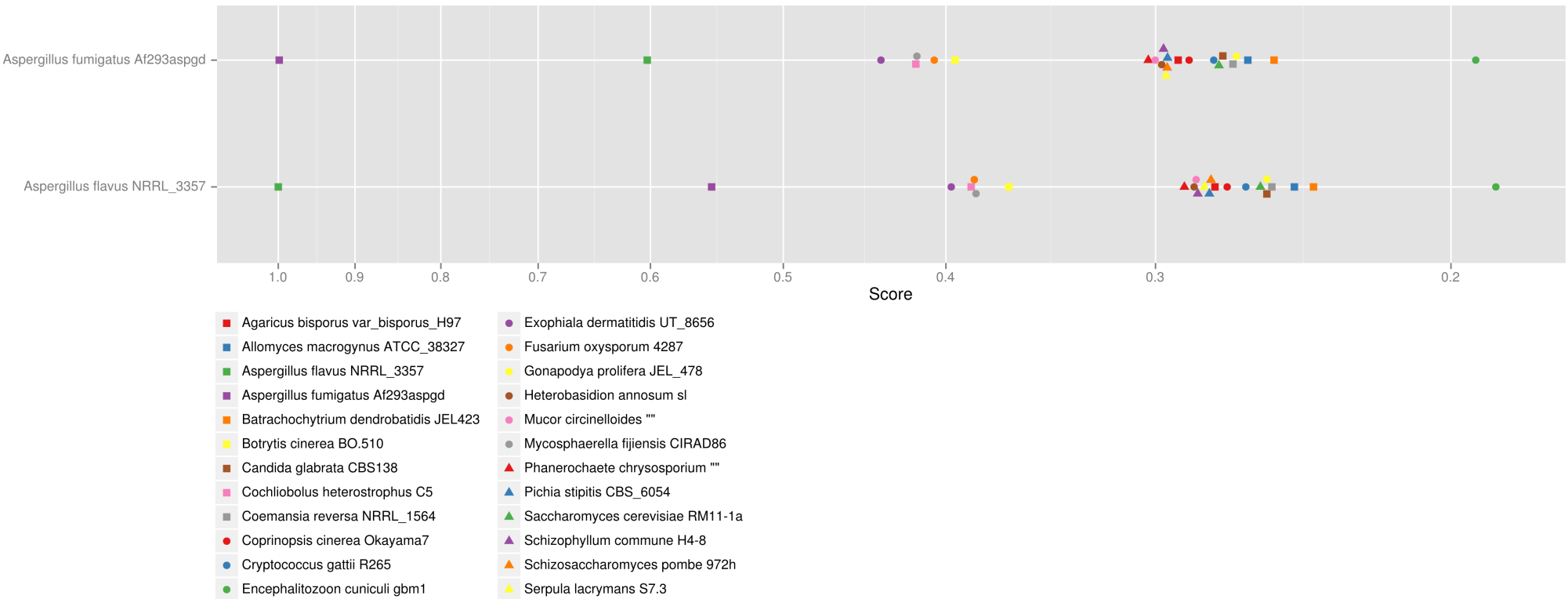


Calculate Aggregate Proteome Similarity

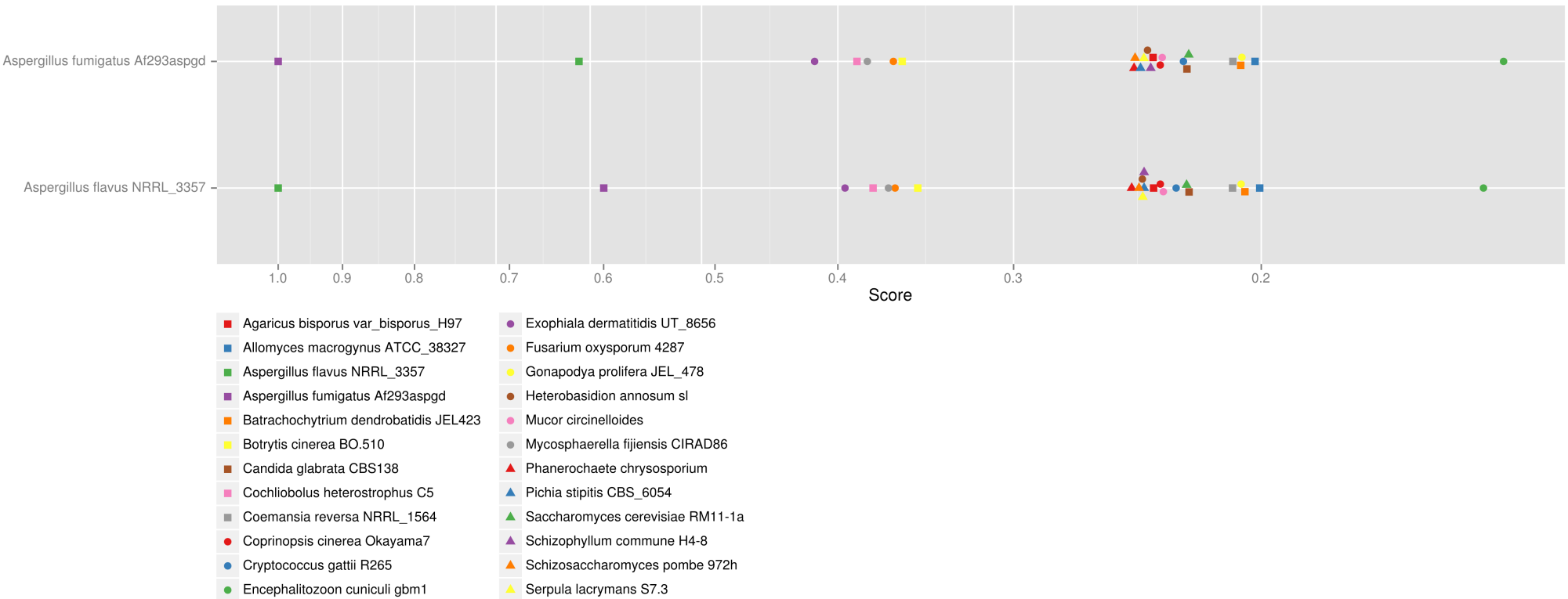
For each pair of genomes, find the average score by each measure



Aggregate Protein Sequence Identity Yardstick



Aggregate Bitscore Yardstick (BLOSUM62)



Specialized yardsticks

Generate yardstick using only a subset of the genome

- Metabolic Genome Yardstick
- Developmental Genome Yardstick

Weight genes by expression

- Placental Genome Yardstick
- Embryonic Genome Yardstick

Pathway centric yardsticks – Gene Set Enrichment Analysis

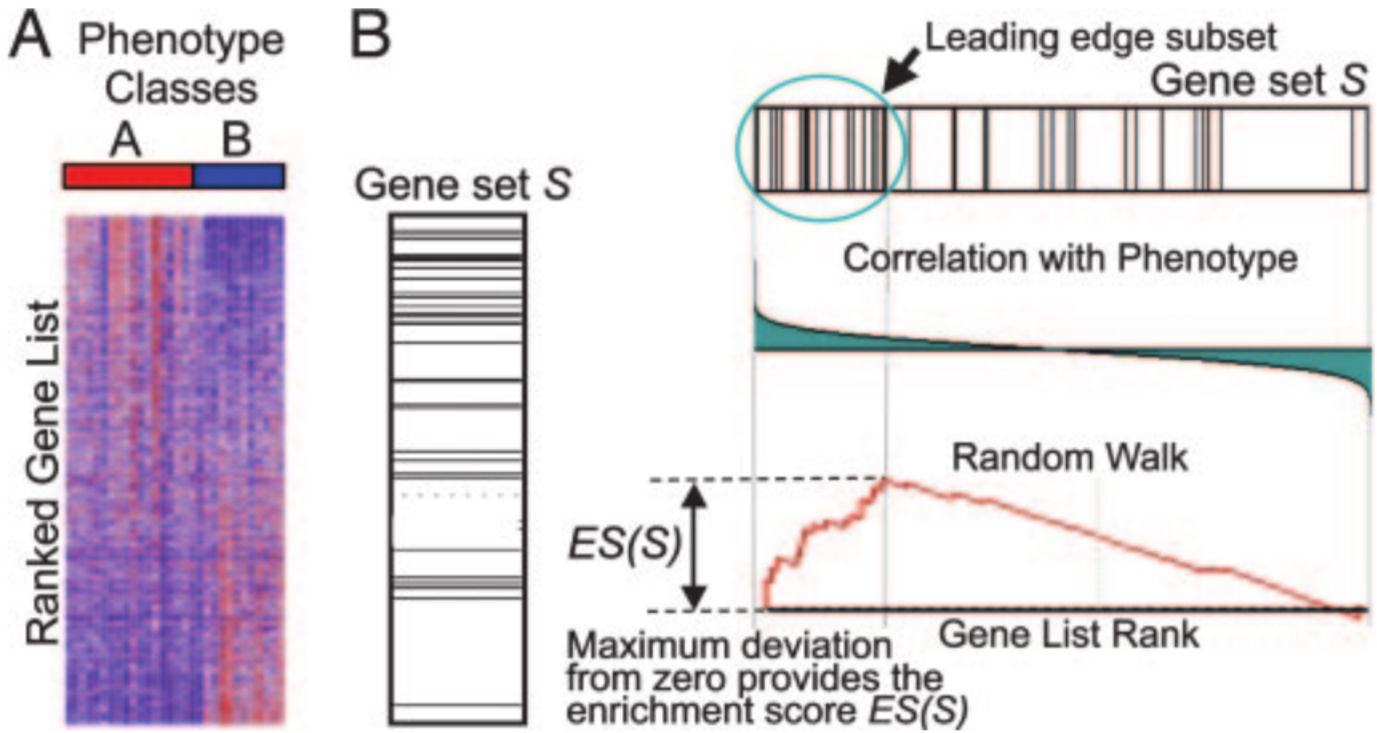


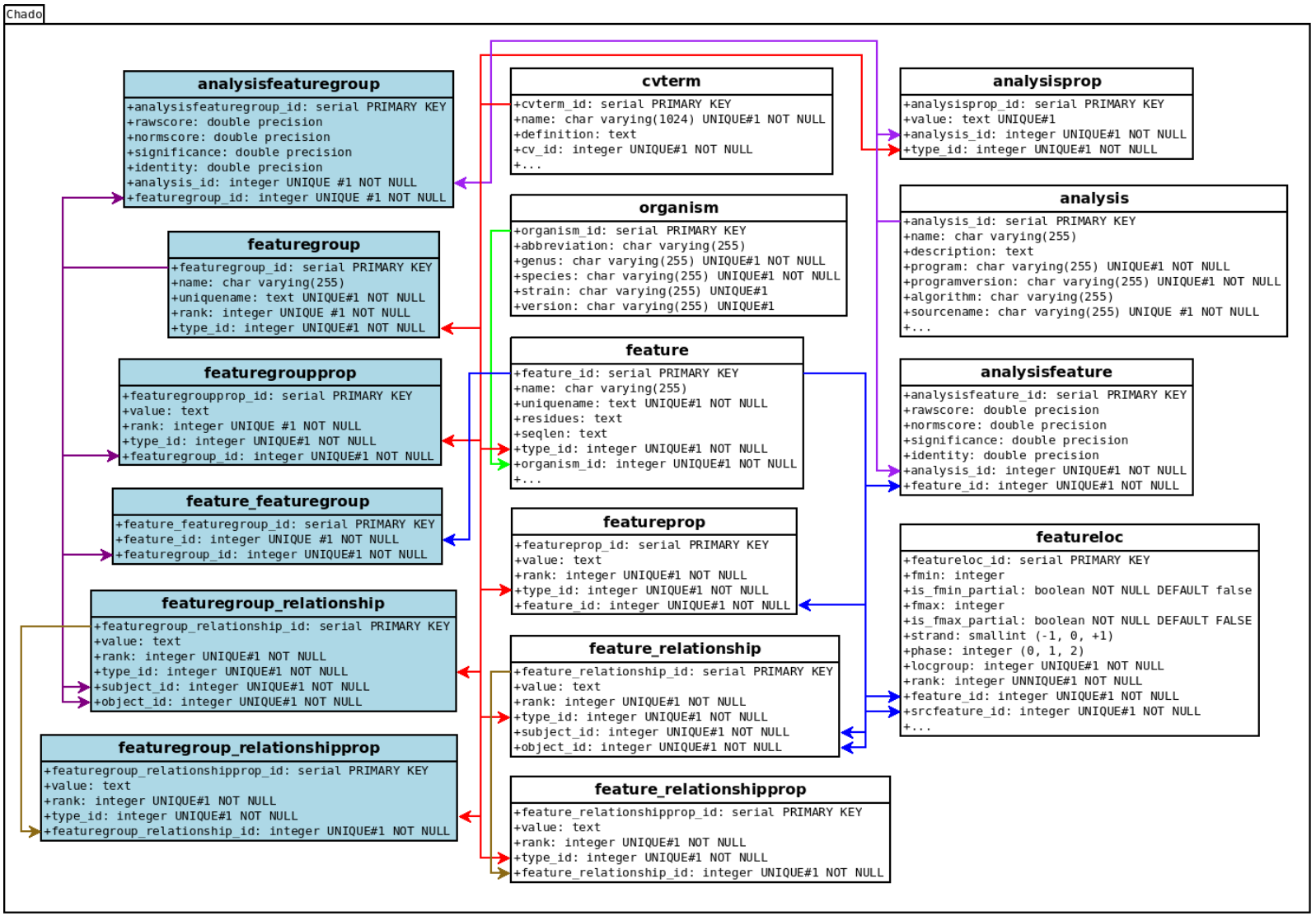
Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the ''gene tags,'' i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

A. Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proc. Natl. Acad. Sci. U.S.A.* 102.43 (2005), pp. 15545–15550

Comparative module (featuregroups)

- Represent and annotate sets of features
- Avoids denormalization of ortholog annotation
- Allows access to sets via Foreign Keys

Proposed module structure



Complexity analysis

- featuregroup
  - Insert group: $O(\log n)$
  - Insert member: $O(\log n)$
  - Select members: $O(\log n)$
  - Annotate group: $O(\log n)$
  - Delete group: $O(\log n)$
- Tagging via featureprop
  - Insert group(!): $O(\log n)$
  - Insert member: $O(\log n)$
  - Select members: $O(n)$
  - Annotate group(!): $O(n)$
  - Delete group: $O(n)$

Summary

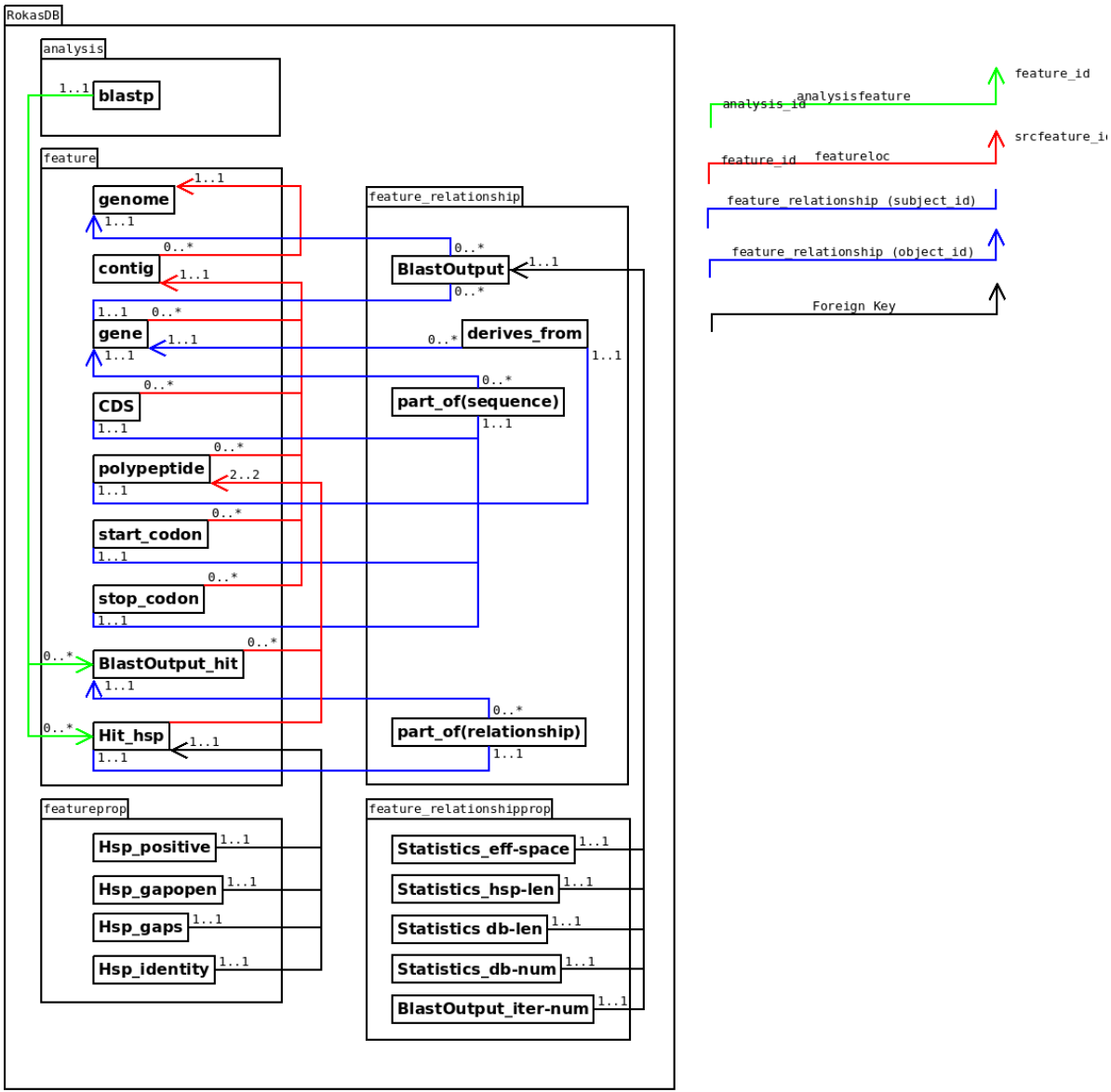
- Genome Yardsticks quantify similarity between organisms
- Specialized yardsticks test relevance of model systems
- Standard way to represent sets of genomic features is needed

Awknowledgements

- Dr. Antonis Rokas
- Dr. Kris McGary
- Dr. Jennifer Wisecaver

- GMOD Project
- March of Dimes

BLAST in Chado



Proposed SQL

```
--Groups
CREATE TABLE featuregroup (
    featuregroup_id serial PRIMARY KEY,
    name varchar(255),
    uniquename text NOT NULL,
    rank integer NOT NULL DEFAULT 0,
    type_id integer NOT NULL REFERENCES cvterm
      ON UPDATE CASCADE ON DELETE CASCADE
      DEFERRABLE INITIALLY DEFERRED,
    is_analysis boolean NOT NULL DEFAULT false,
    UNIQUE(uniquename, rank, type_id)
);

--Group members
CREATE TABLE feature_featuregroup (
    feature_featuregroup_id serial PRIMARY KEY,
    featuregroup_id integer NOT NULL REFERENCES featuregroup
      ON UPDATE CASCADE ON DELETE CASCADE
      DEFERRABLE INITIALLY DEFERRED,
    feature_id integer NOT NULL REFERENCES feature
      ON UPDATE CASCADE ON DELETE CASCADE
      DEFERRABLE INITIALLY DEFERRED,
    UNIQUE(featuregroup_id, feature_id)
);

--Group annotation
CREATE TABLE featuregroupprop (
    featuregroupprop_id serial PRIMARY KEY,
    value text,
    rank integer NOT NULL DEFAULT 0,
    type_id integer NOT NULL REFERENCES cvterm
      ON UPDATE CASCADE ON DELETE CASCADE
      DEFERRABLE INITIALLY DEFERRED,
    featuregroup_id integer NOT NULL REFERENCES featuregroup
      ON UPDATE CASCADE ON DELETE CASCADE
      DEFERRABLE INITIALLY DEFERRED,
    UNIQUE(rank, type_id, featuregroup_id)
);

--Analysis featuregroups
CREATE TABLE analysisfeaturegroup (
    analysisfeaturegroup_id serial PRIMARY KEY,
    rawscore double precision,
    normscore double precision,
    significance double precision,
    identity double precision,
    analysis_id integer NOT NULL REFERENCES analysis
      ON UPDATE CASCADE ON DELETE CASCADE
      DEFERRABLE INITIALLY DEFERRED,
    featuregroup_id integer NOT NULL REFERENCES featuregroup
      ON UPDATE CASCADE ON DELETE CASCADE
      DEFERRABLE INITIALLY DEFERRED,
    UNIQUE(analysis_id, featuregroup_id)
);

--TODO? featuregroup_pub
--      featuregroup_relationship, etc.
```

```
MaraKim_GMOD2014 1.0
e6d7383f6f2a69db45040ec596d254b1d5a460a8
## master
```