Comparative analysis of microbial genomes

Focus on *Mycobacterium tuberculosis*

Nicky Mulder

University of Cape Town

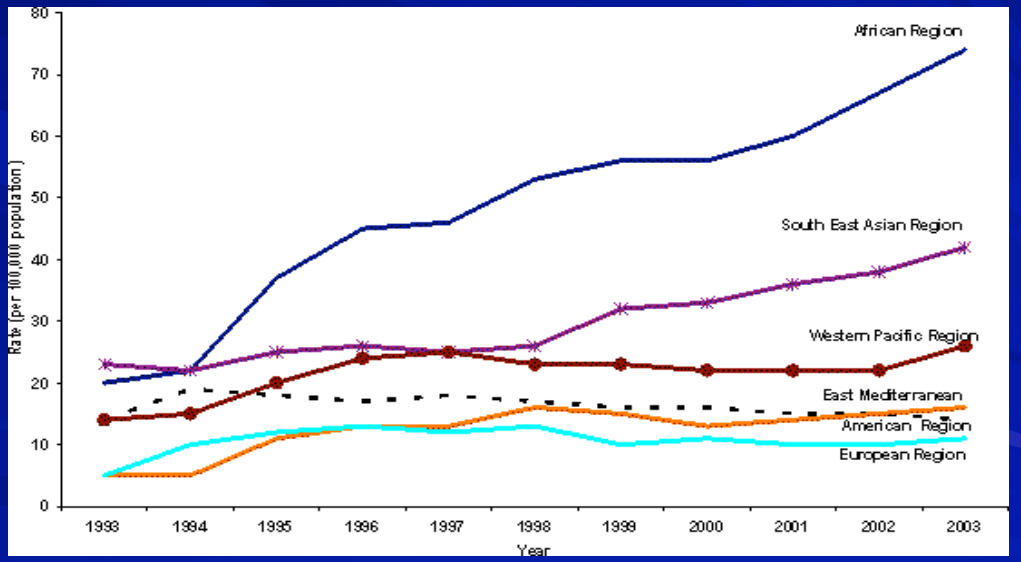
European Bioinformatics Institute

Outline

- Tuberculosis background
- Mycobacterial genomics
- Comparative analysis of microbial genomes
- Identification of expanded families in *M. tuberculosis*
- Future work

Background

- 1/3 world's population is infected with tuberculosis
- ~1000 people die of TB in South Africa daily
- Increased incidence due to AIDS
- Caused by *Mycobacterium tuberculosis*



World TB notification rate per 100,000 population size (http://www.afro.who.int/tb/notificationtrend.html)

Mycobacterial genome sequencing

- Hundreds of microbial genomes completely sequenced
- Mycobacteria sequenced:
  - Mycobacterium leprae
  - Mycobacterium tuberculosis (H37Rv, Oshkosh, F11)
  - Mycobacterium bovis (& BCG)
  - Mycobacterium paratuberculosis
  - Mycobacterium (sp. MCS, KMS, JLS)
  - Mycobacterium smegmatis
  - Mycobacterium avium
  - Mycobacterium ulcerans
  - Mycobacterium vanbaaleni

Mycobacterium tuberculosis H37Rv

- 4.4Mb genome
- ~4000 proteins
- 51% have predicted functions
- Approximately half the genome is made up of duplicated families
- Has many lipid and fatty acid metabolism pathways
- Still so much not known about the remaining genes and pathogenesis

Comparative genomics

- **Aim: generate phylogenetic profiles to identify potential virulence genes**
- 84 genomes, ~250,000 genes
- 56 pathogens 28 non-pathogens
- BLASTClust at different score densities
- Score density 0.5 and 1.0

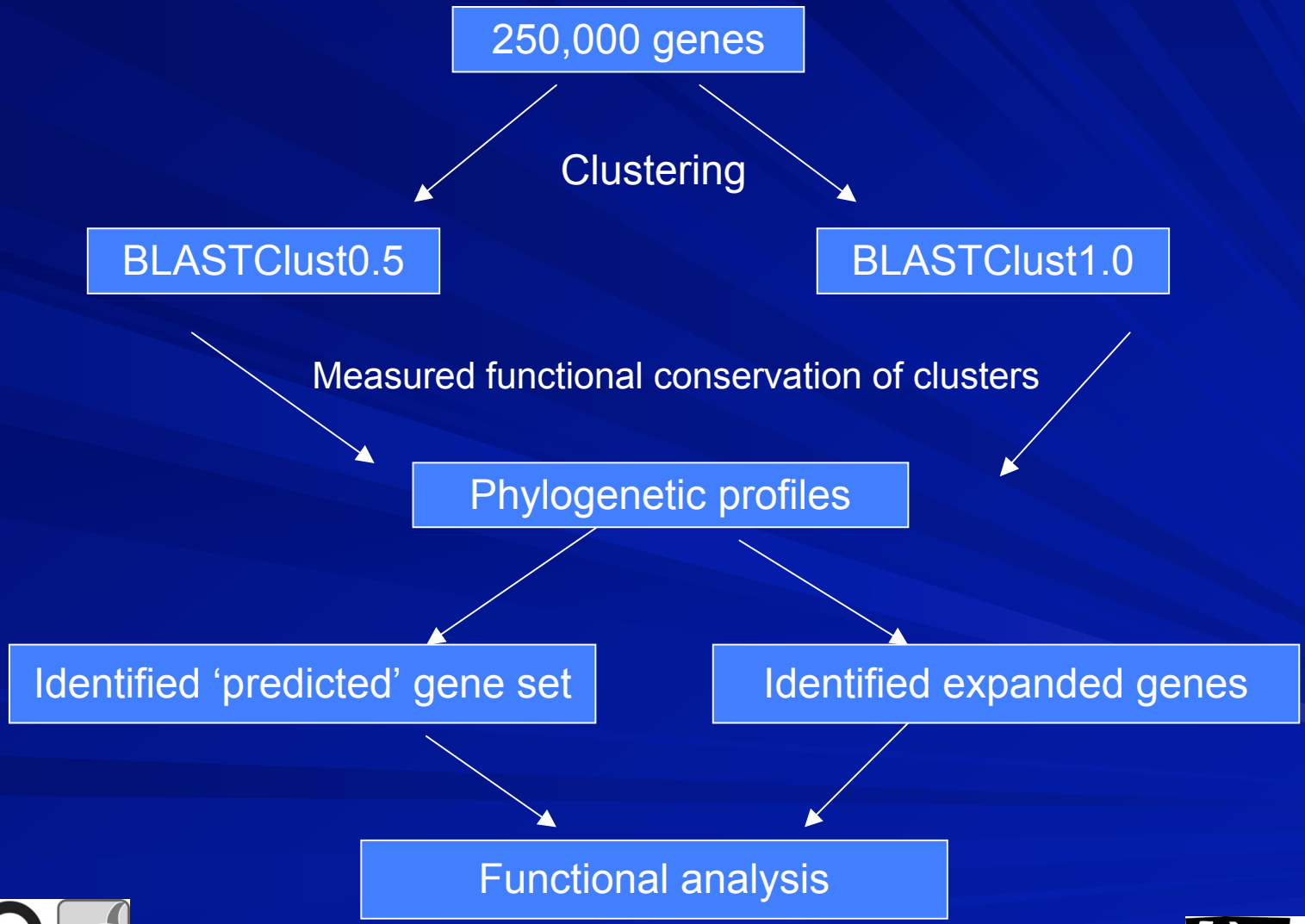
  S = -N/AL'

  L' is length of sequence in the alignment, L
  N is the number of identical residues
  A is the total alignment length

CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Methods



CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

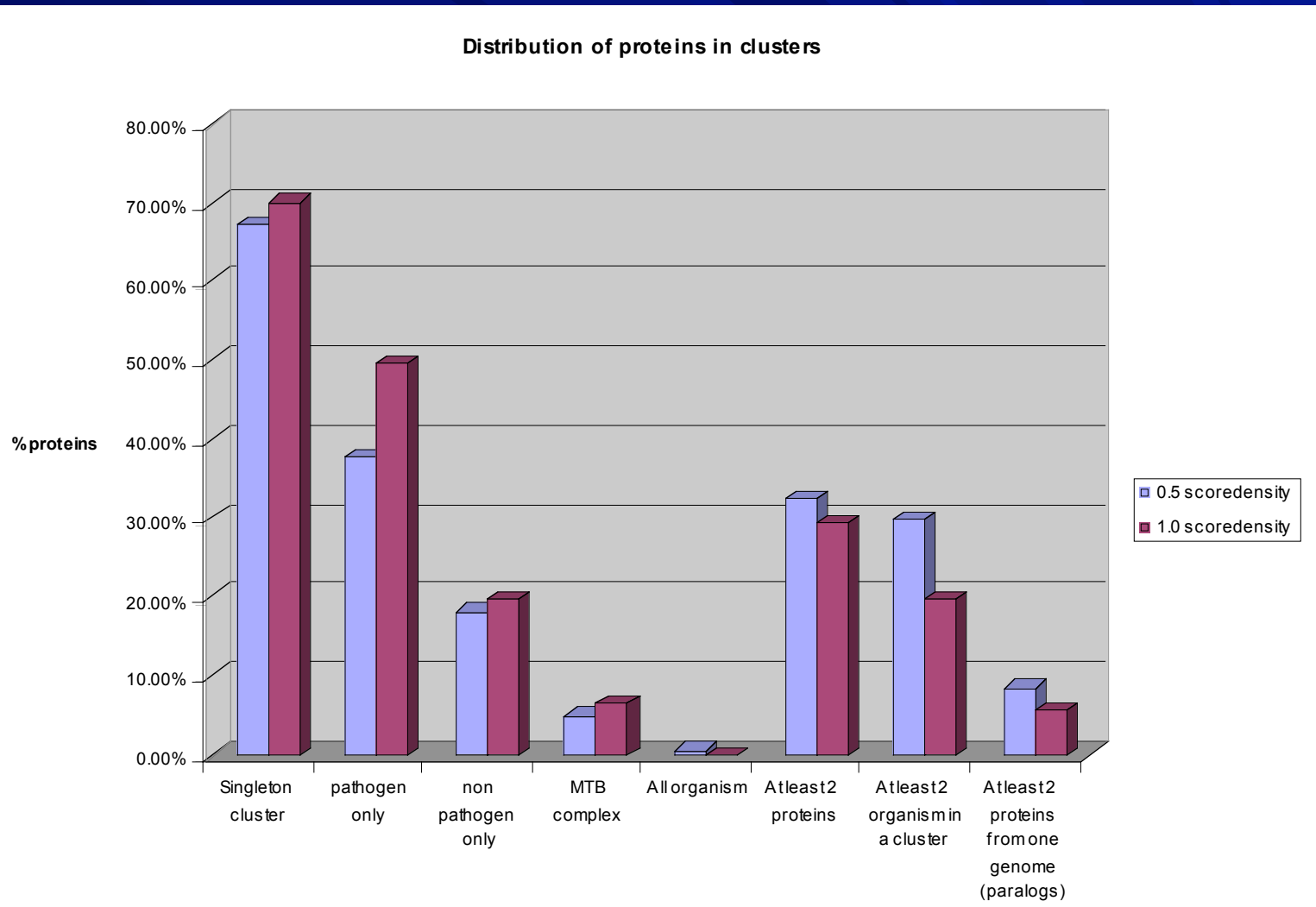
Summary of clustering results

Experiments	0.5 score density	1.0 score density
Total clusters created	63819	121905
Total Singletons (one protein families)	42770	85735
Total more than one protein clusters	20849	36170
Total Pathogen only clusters	7904	18022
Total non Pathogen only clusters	3824	7165
Total MTB Complex only clusters	1060	2362
Number with at least two genomes per cluster	19056	24110
Number of clusters with at least 2 proteins (paralogs) in one genome	1794	2061
All organisms (84 genomes)	10	2
At least 80 genomes	135	39
All organism (at least 60 genomes)	448	130

CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Distribution of proteins in cluster types



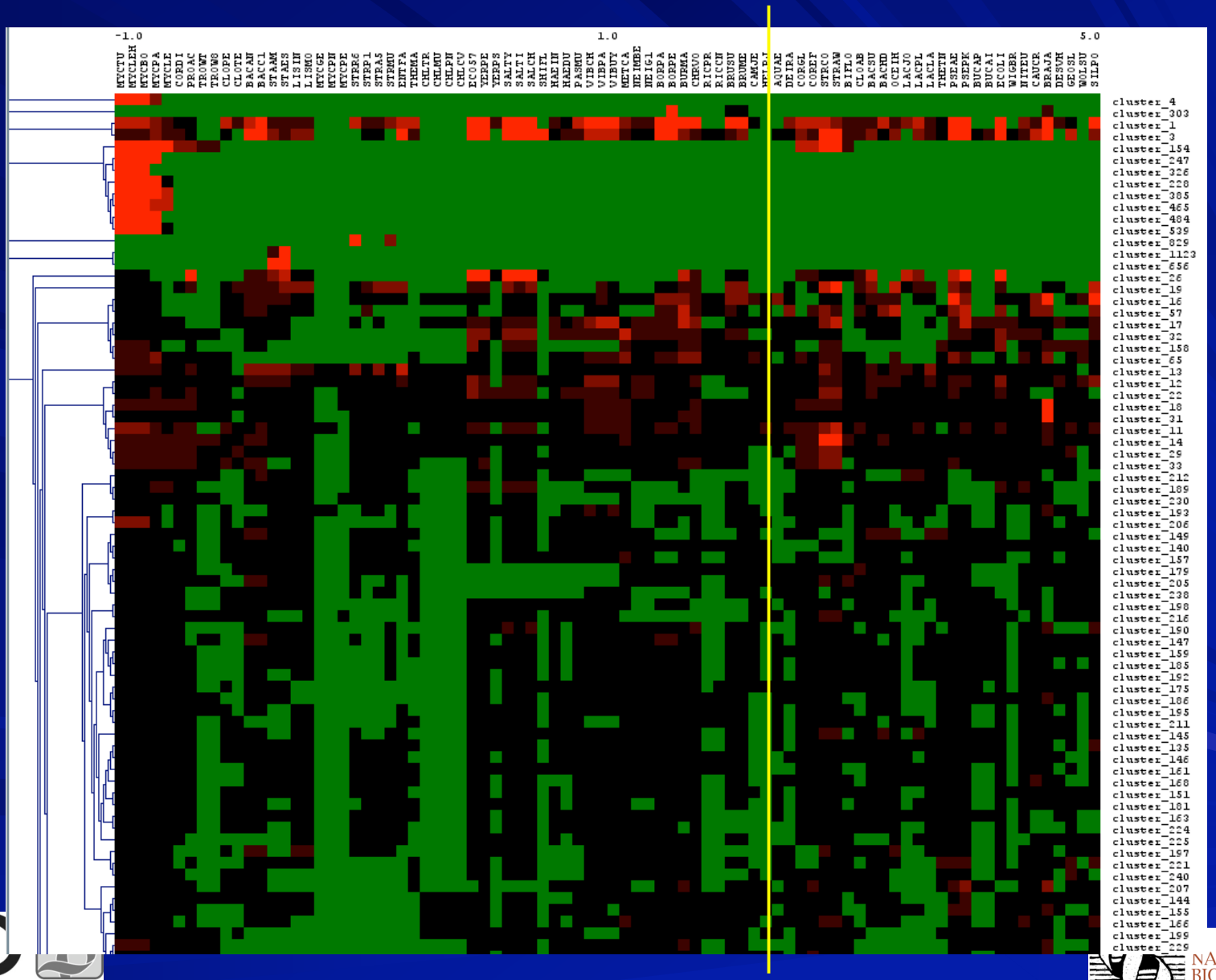
CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Measuring functional diversity within clusters

- Used DE lines, InterPro matches and GO Slim
- Can only be done for those covered: ~70% InterPro, 50% GO slim
- BlastClust1.0 greatest conservation –most >90% conservation of InterPro domains and annotation
- In 1.0 PE/PPE families agreed with previous reports

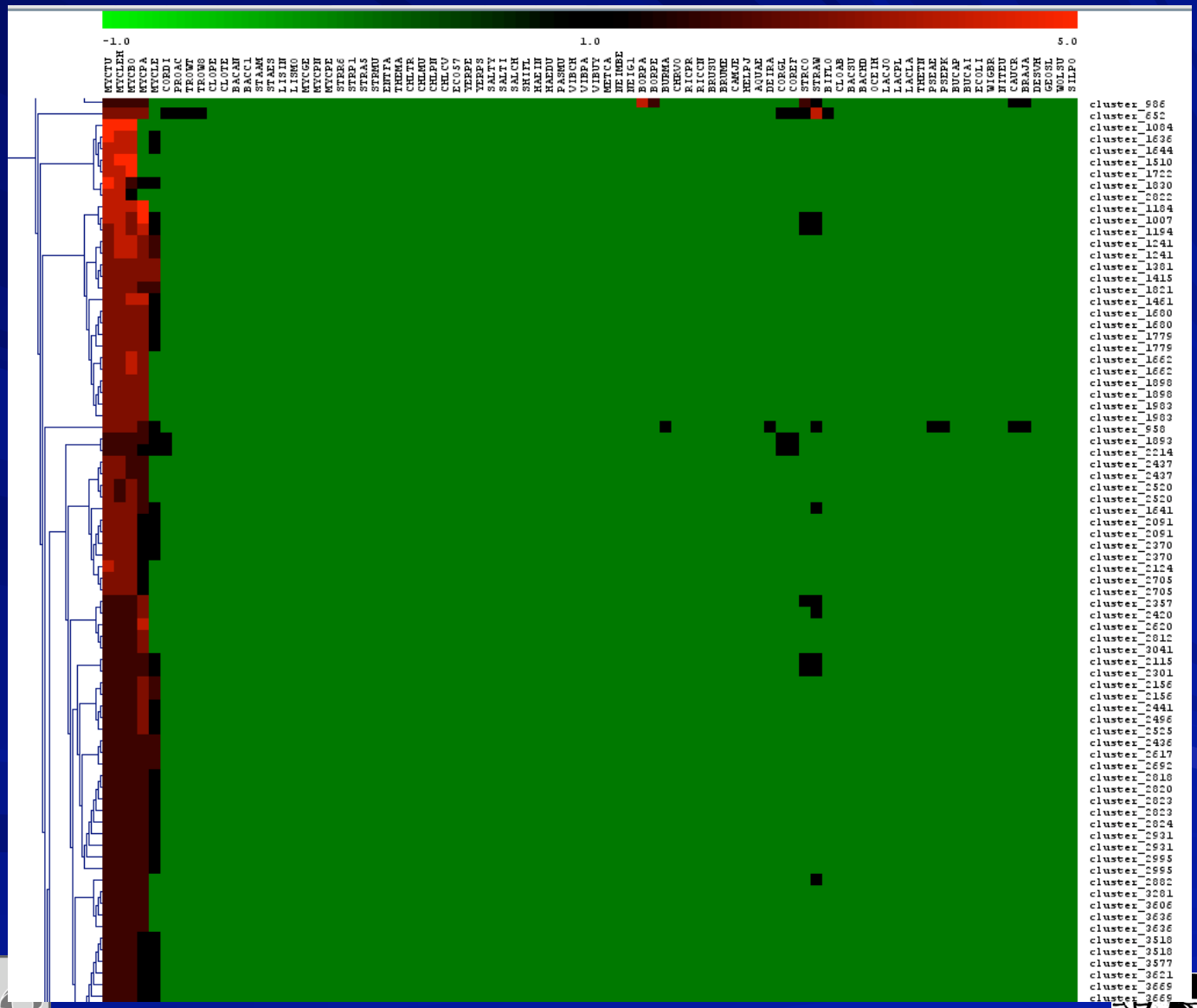
Heatmap of results 1



CBIC Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Heatmap of results 2



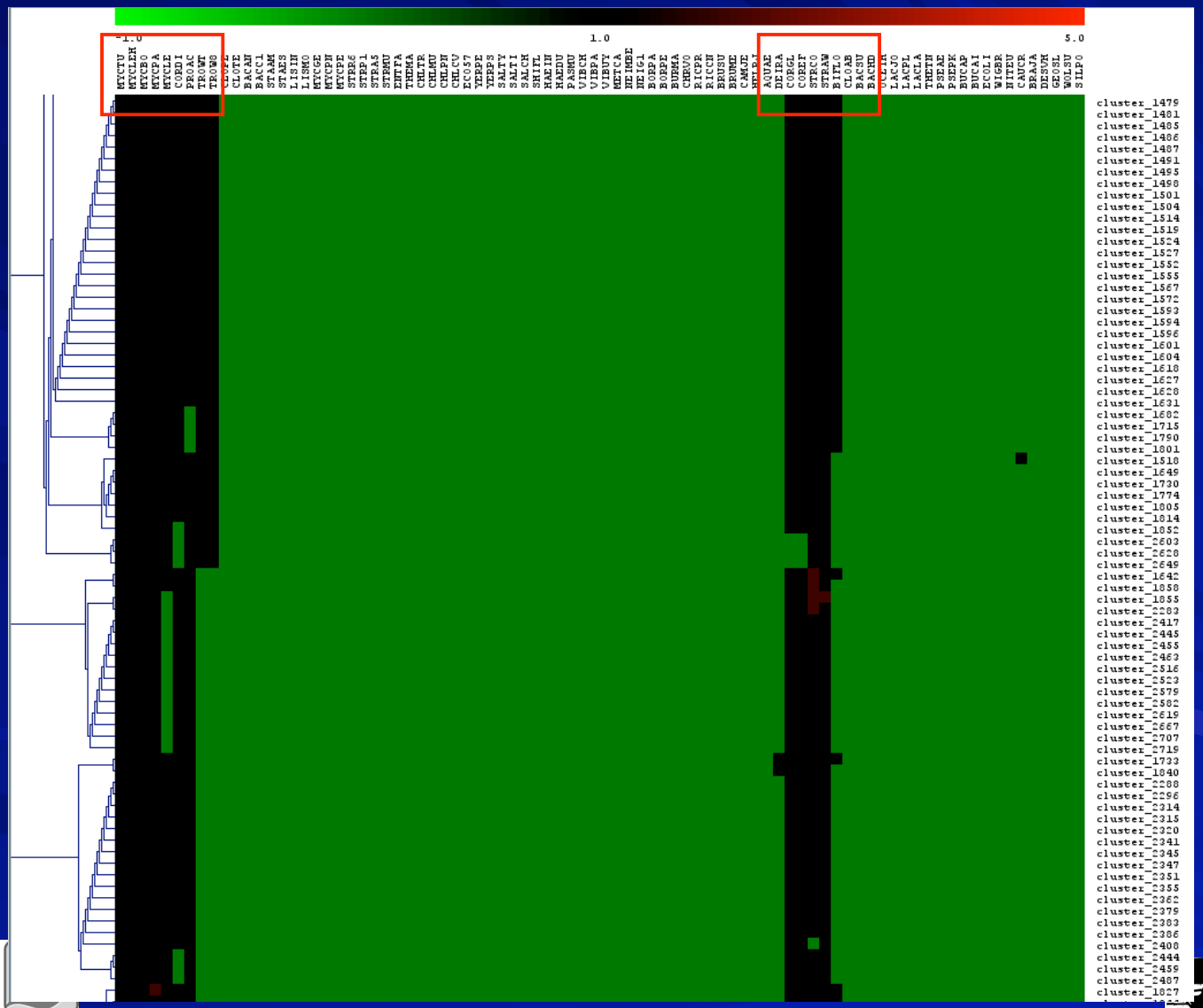
CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Heatmap of results 3

Mycobacteria

Corynebacteria
Tropheryma



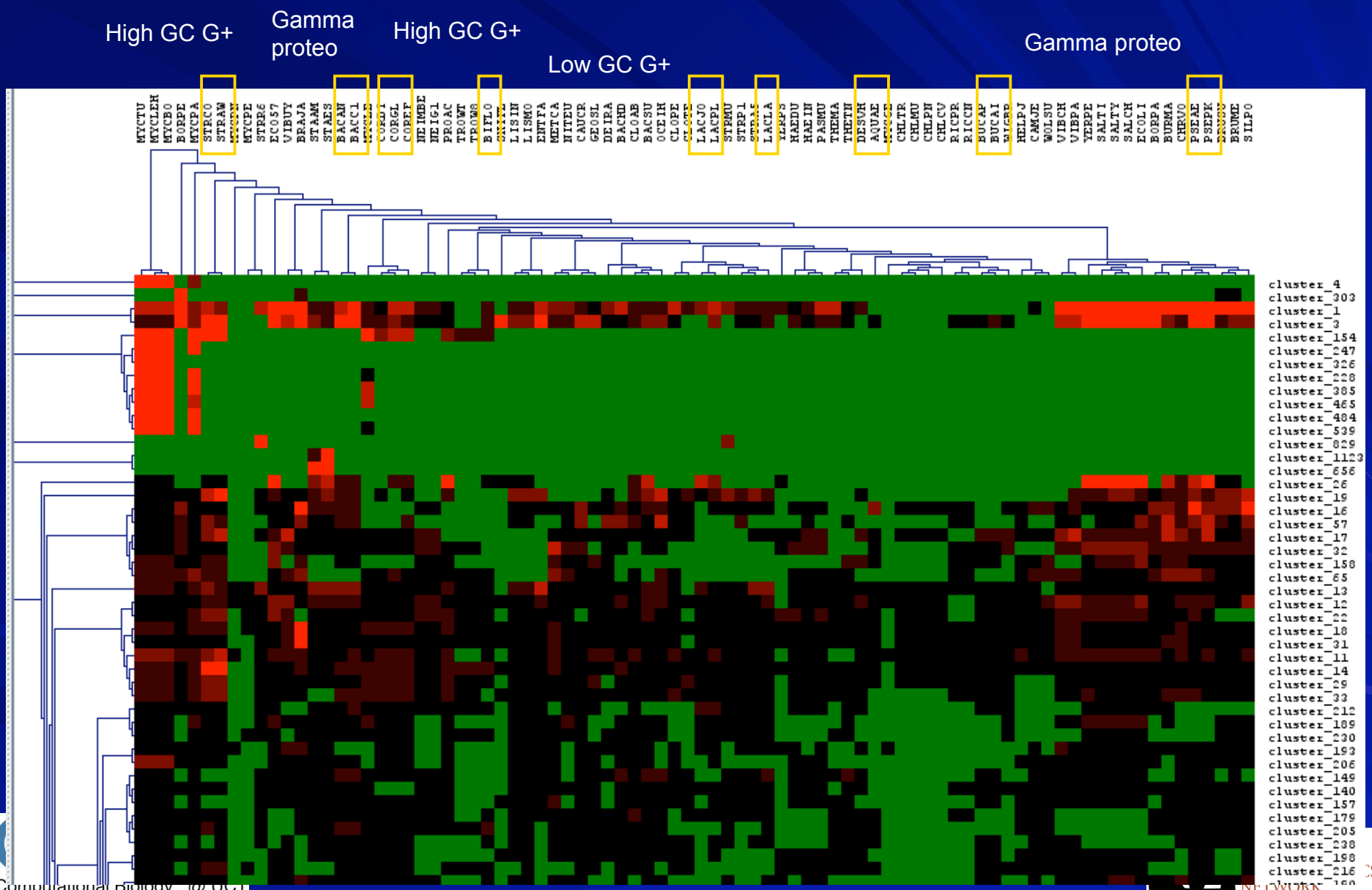
Streptomyces

Corynebacteria

CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Clustering by organism

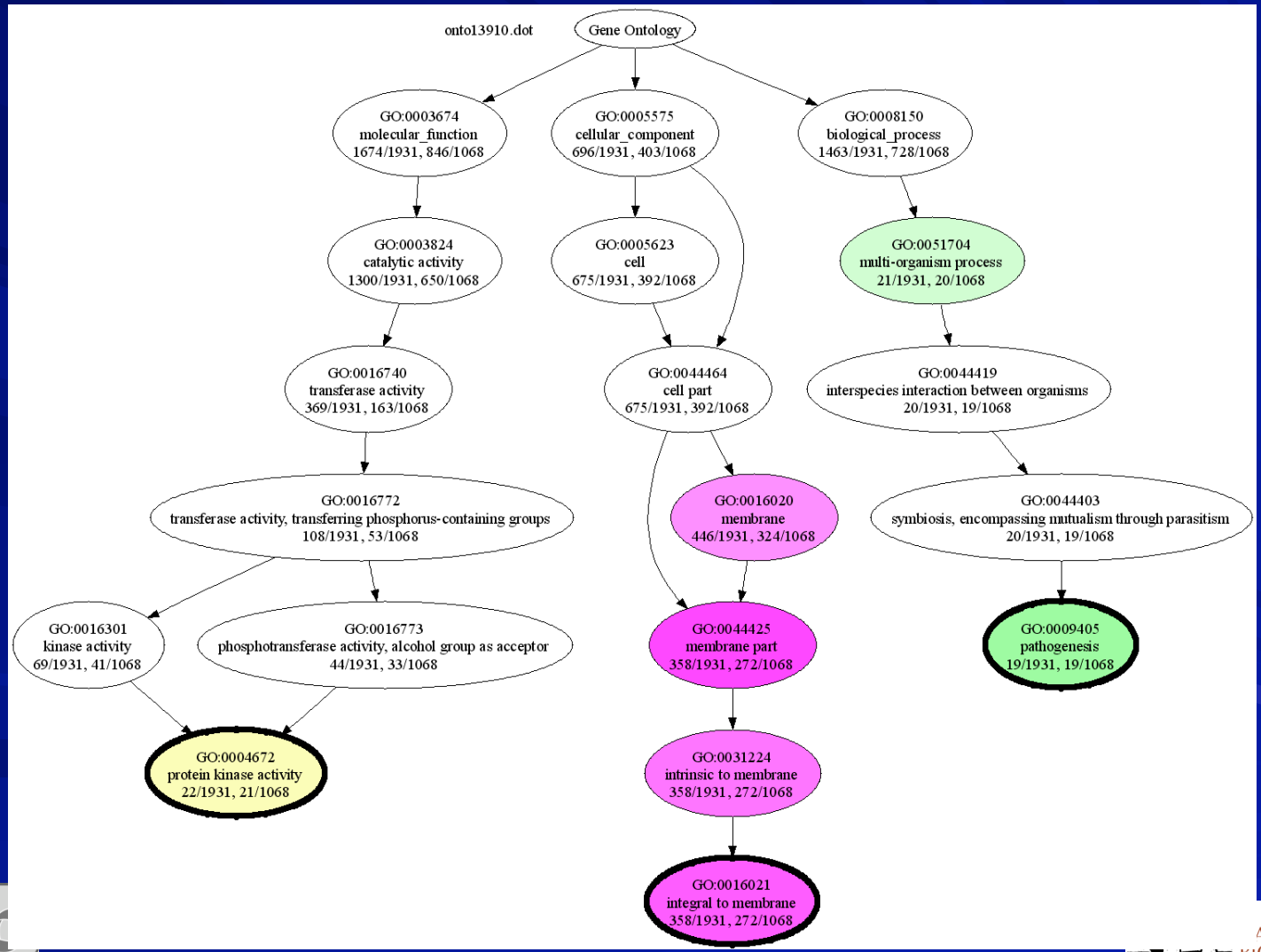


Known virulence genes

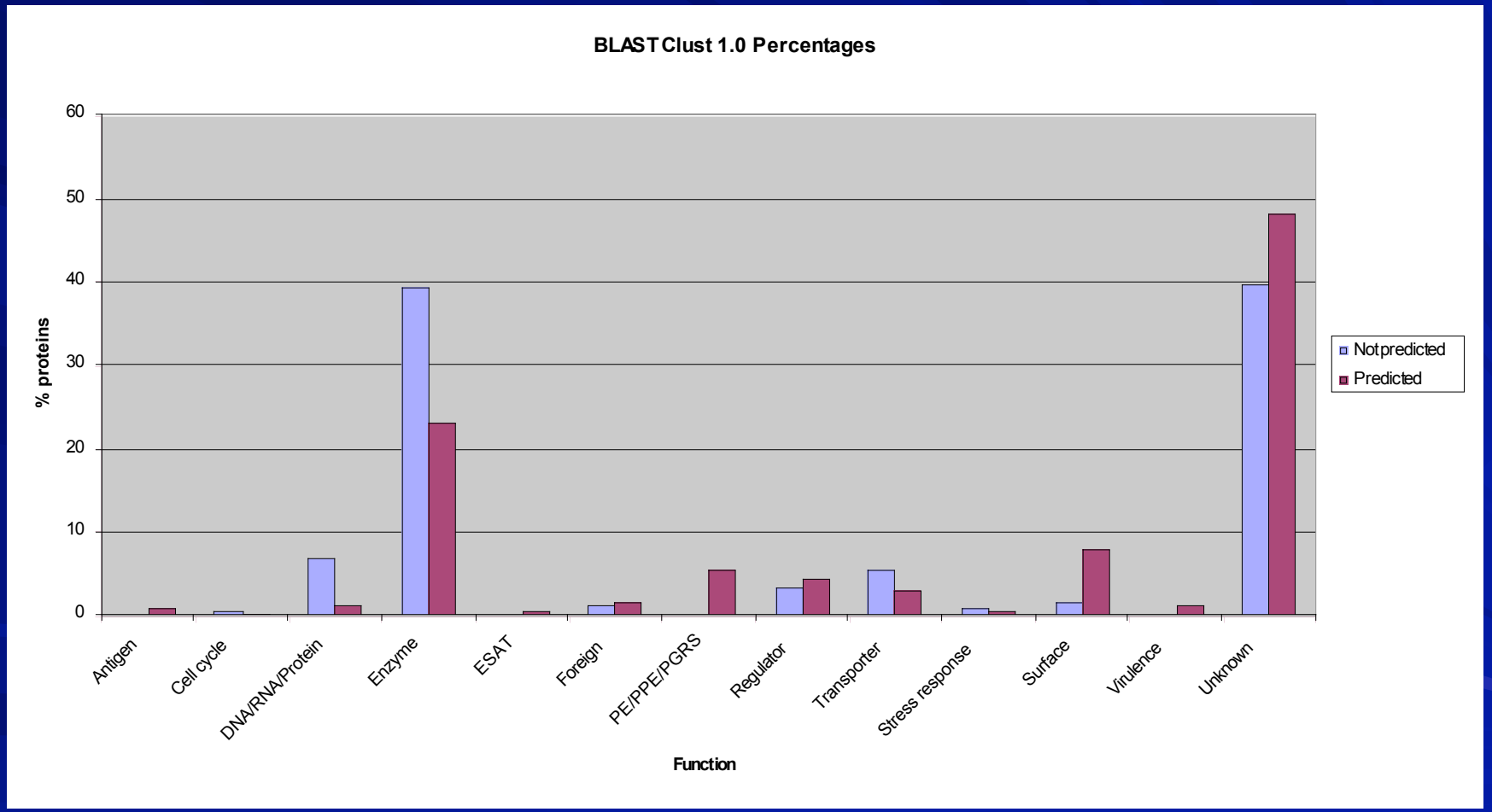
Cluster No	Proteins in pathogens (no. organisms)	Proteins in non pathogens (no. organisms)	Common DE line
652	16 (8)	8 (5)	DNA-binding_ response_regulator
1583	7 (7)	7 (4)	Mycobacterial persistence regulator MRPA
2441	10 (5)	0	Virulence_factor/mce-family_protein
2613	6 (6)	4 (4)	Cytotoxin_/haemolysin_ homologue
2818	9 (5)	0	Virulence_factor_mce_ family_protein
2820	9 (5)	0	Mce-family_protein_mce2d
2822	9 (3)	0	Phospholipase_C_1_precursor_ (EC_3.1.4.3)_(MTP40_antigen)
2823	9 (5)	0	Virulence_factor_mce_ family_protein
2931	9 (5)	0	Mce-family_protein_mce1b_
3332	8 (5)	0	Virulence_factor_mce_ family_protein
7120	5 (5)	0	Heparin-binding_ hemagglutinin_(Adhesin)
8232	4 (4)	0	Sulfatase family protein
8234	4 (4)	0	Virulence_factor_mce_ family_protein
8935	4 (4)	0	Exported_repetitive_protein_precursor_(Cell_surface_protein_pirG)
11724	3 (3)	0	Virulence_factor_mce_ family_protein
1094	3 (3)	0	Virulence-regulating_ (arac/xyls_fa[illegible])

Functional analysis

Ontologizer –statistically over-represented GO terms in predicted set



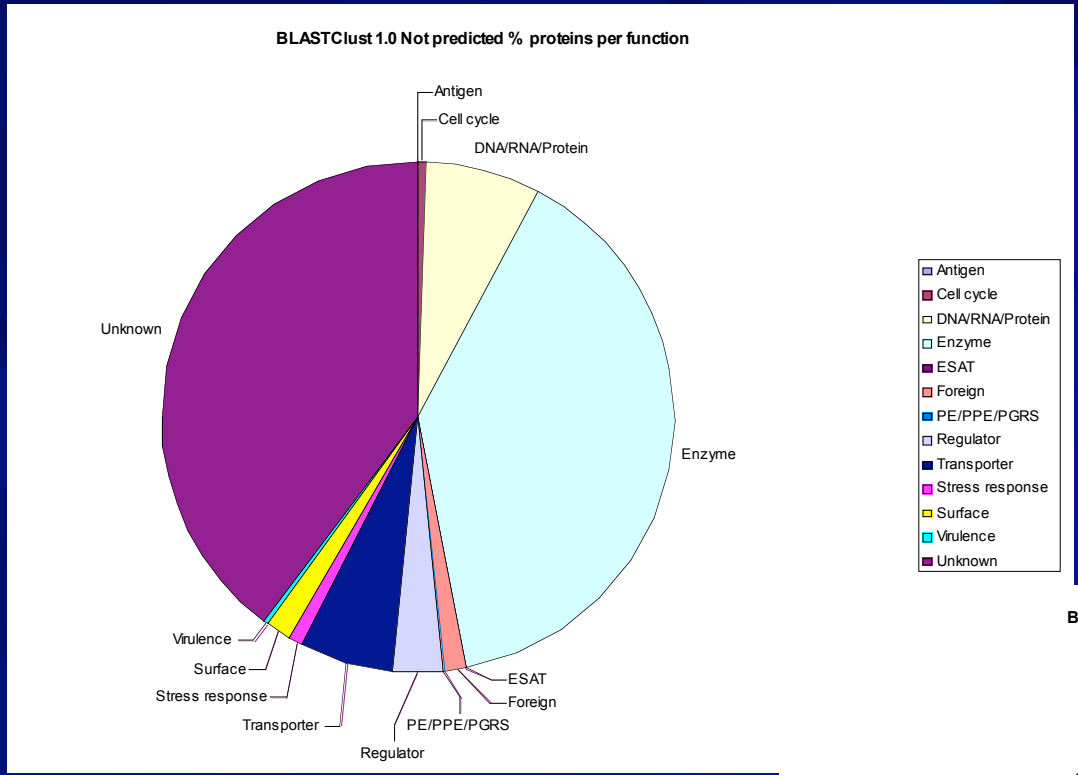
Functional analysis continued



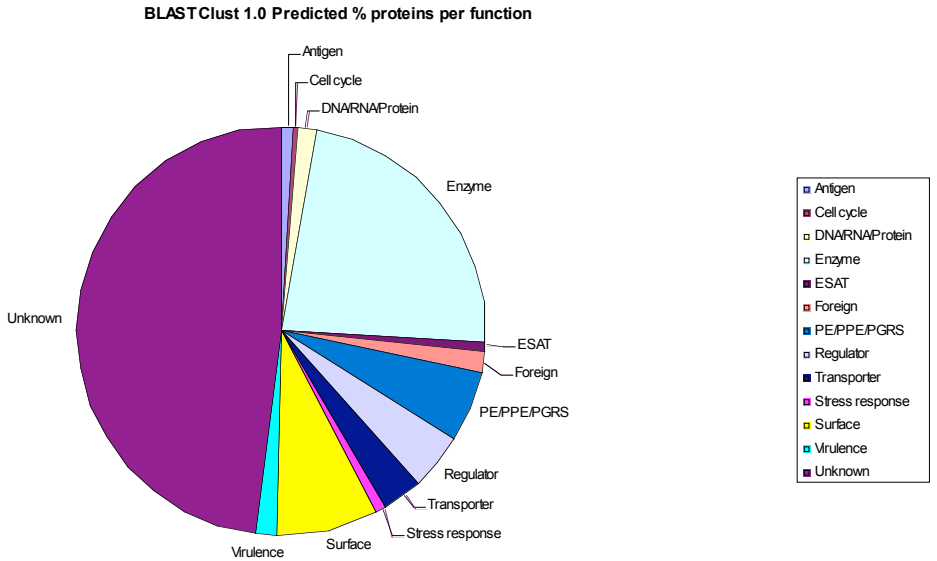
CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Functional composition of protein sets

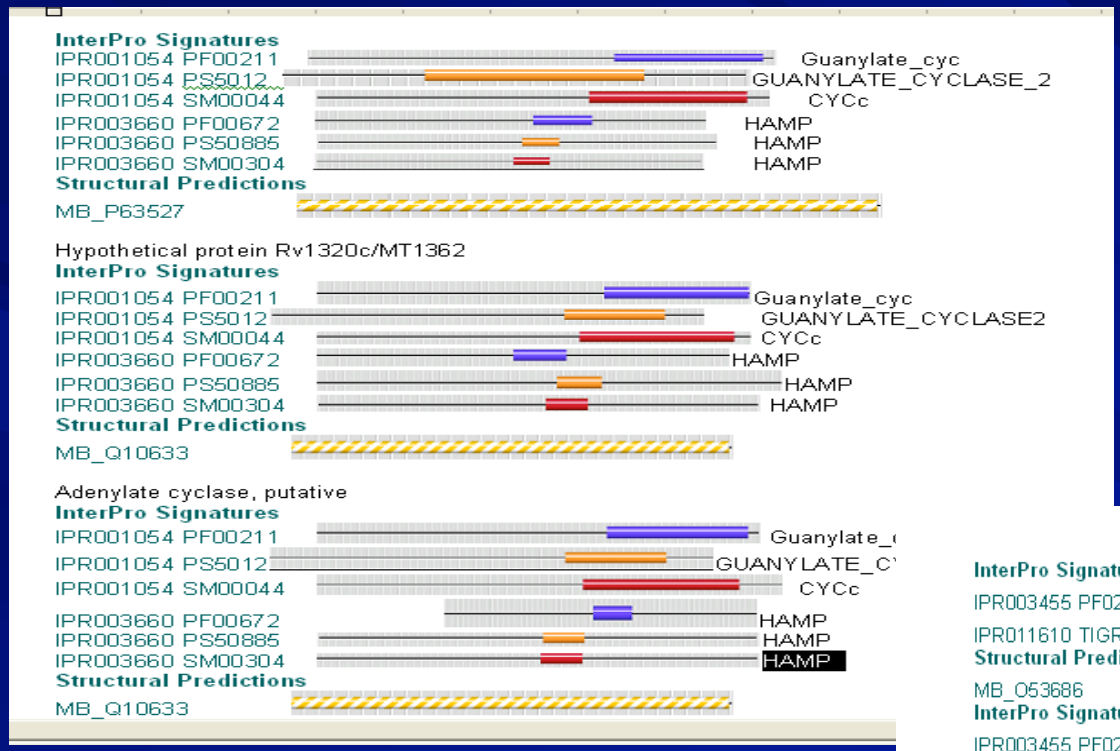


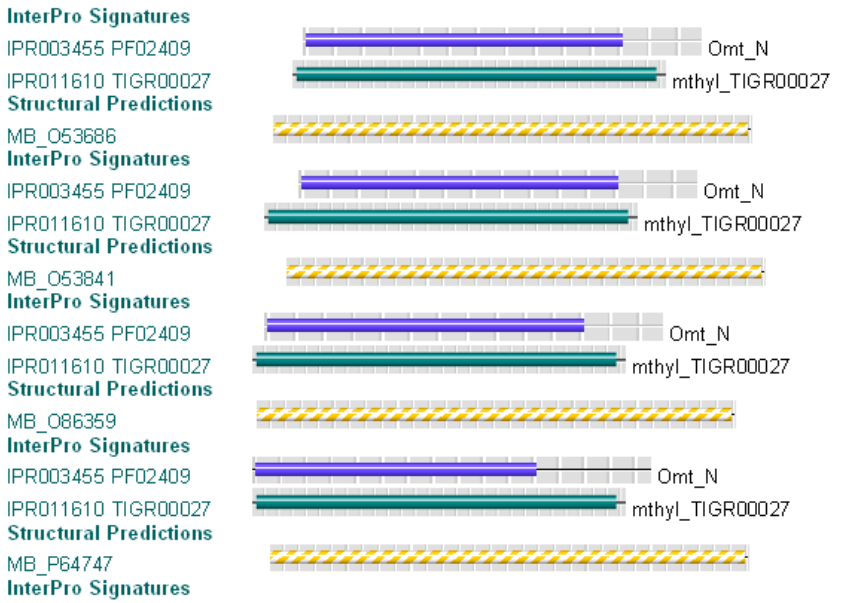
Both have high proportion of hypothetical proteins



CBIO Computational Biology @ UCT

Hypothetical proteins

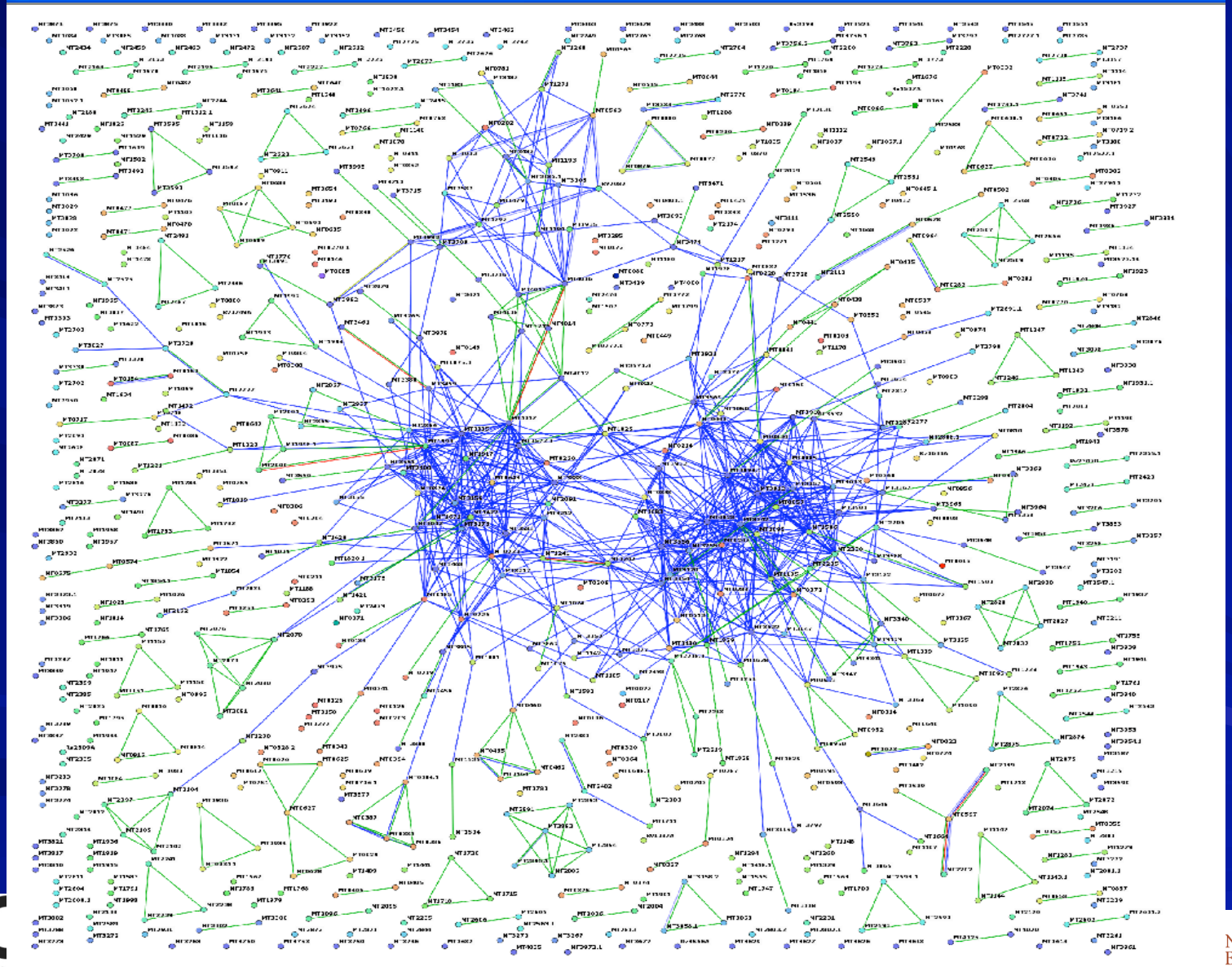




CBIO Computational Biology @ UCT

NETWORK

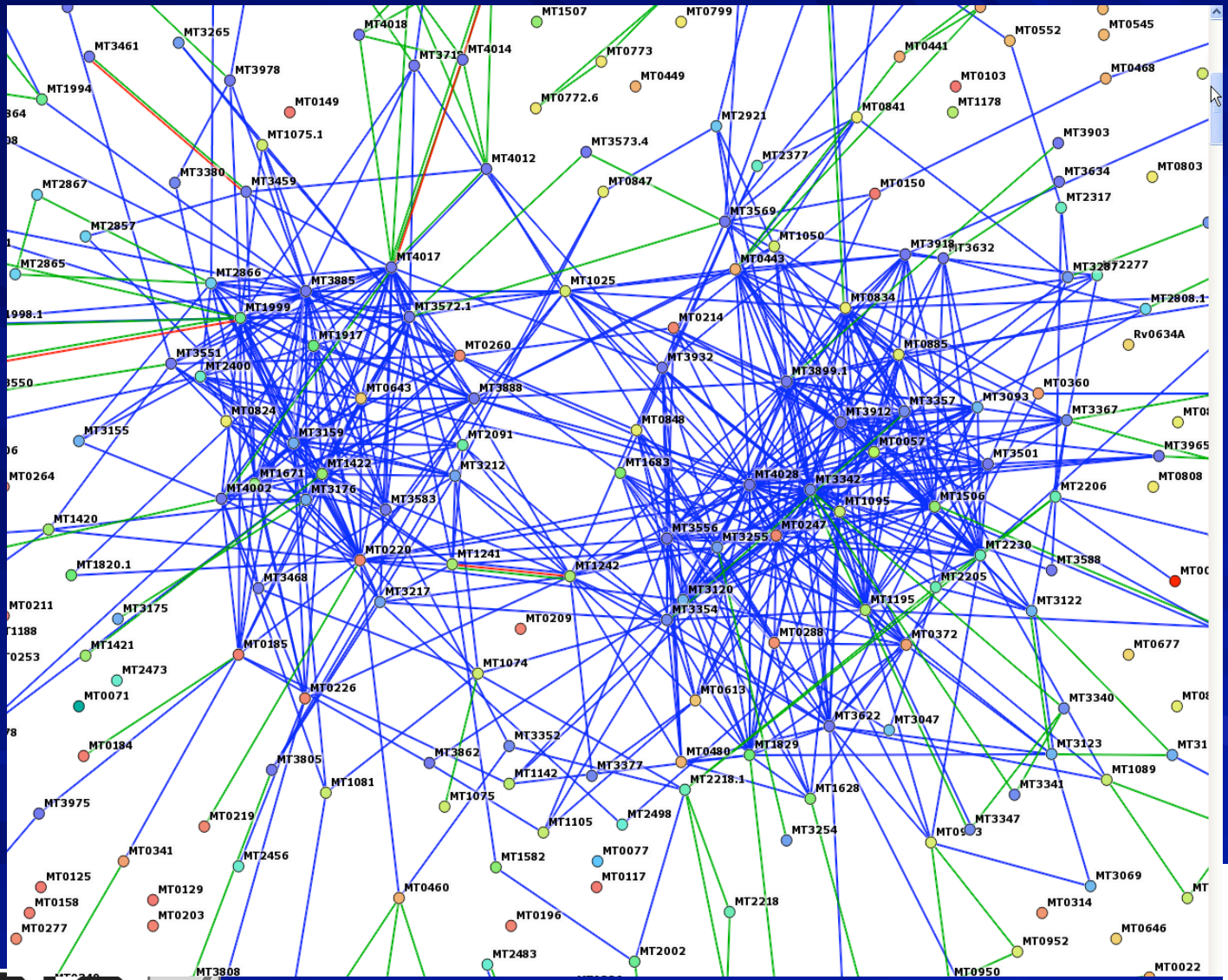
Connections between hypotheticals using STRING



CBIC Computational Biology @ UCI

NATIONAL BIOINFORMATICS NETWORK

Zooming in

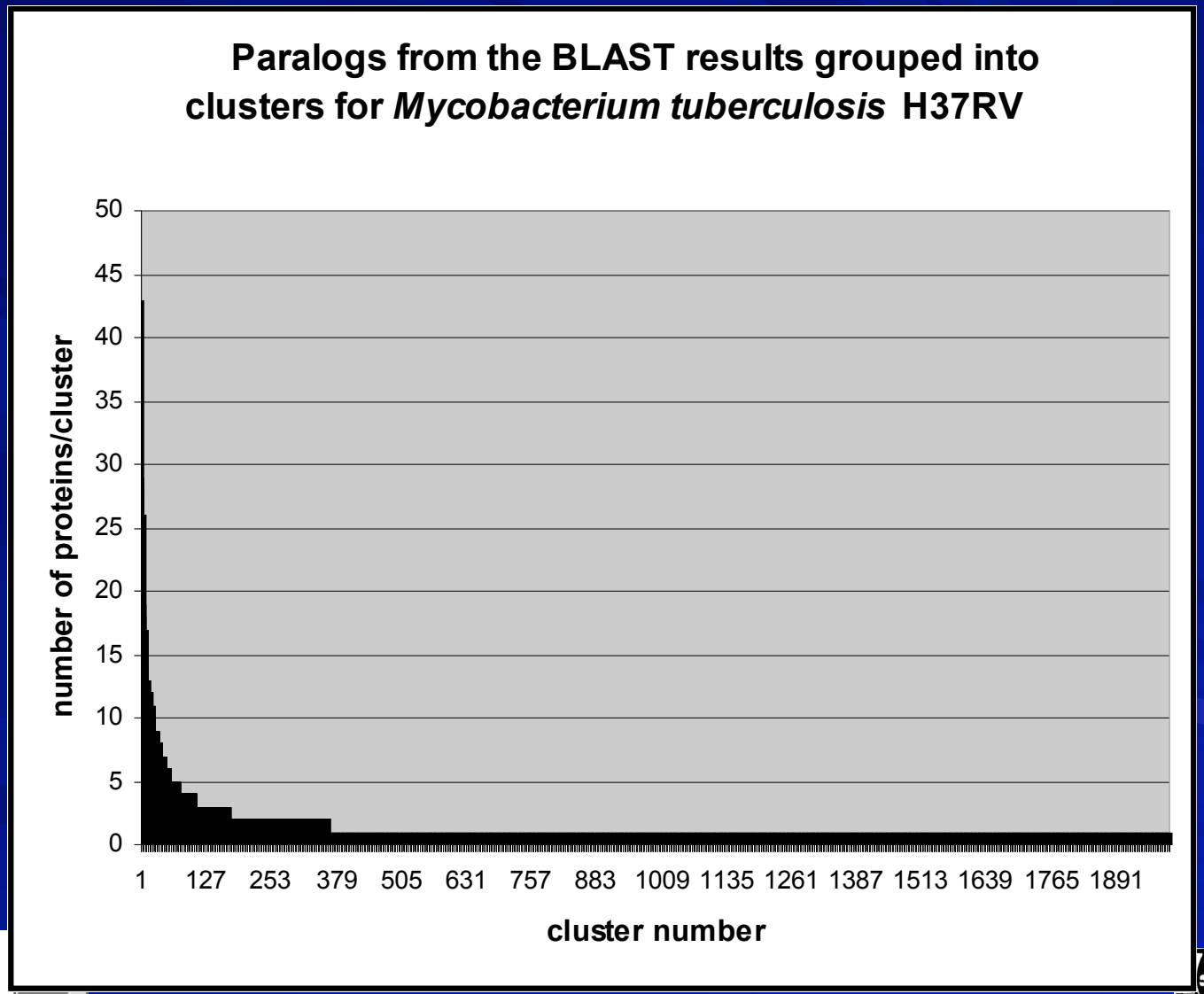


CBIO Computational Biology @ UCT

Resources available to find info on hypotheticals

Neighborhood
Gene Fusion
Cooccurrence
Coexpression
Experiments
Databases
Textmining
[Homology]

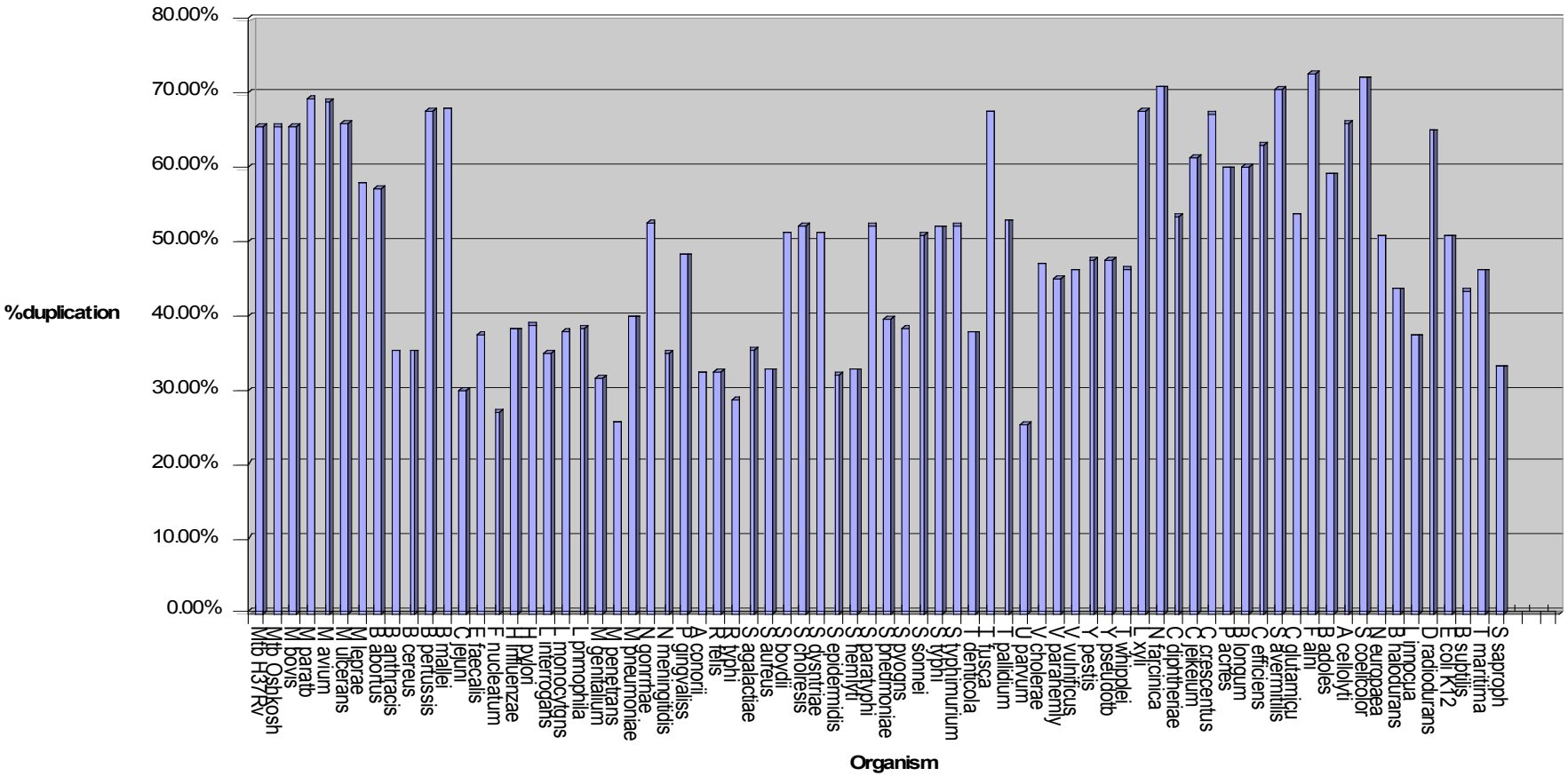
Expanded families in H37Rv



CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

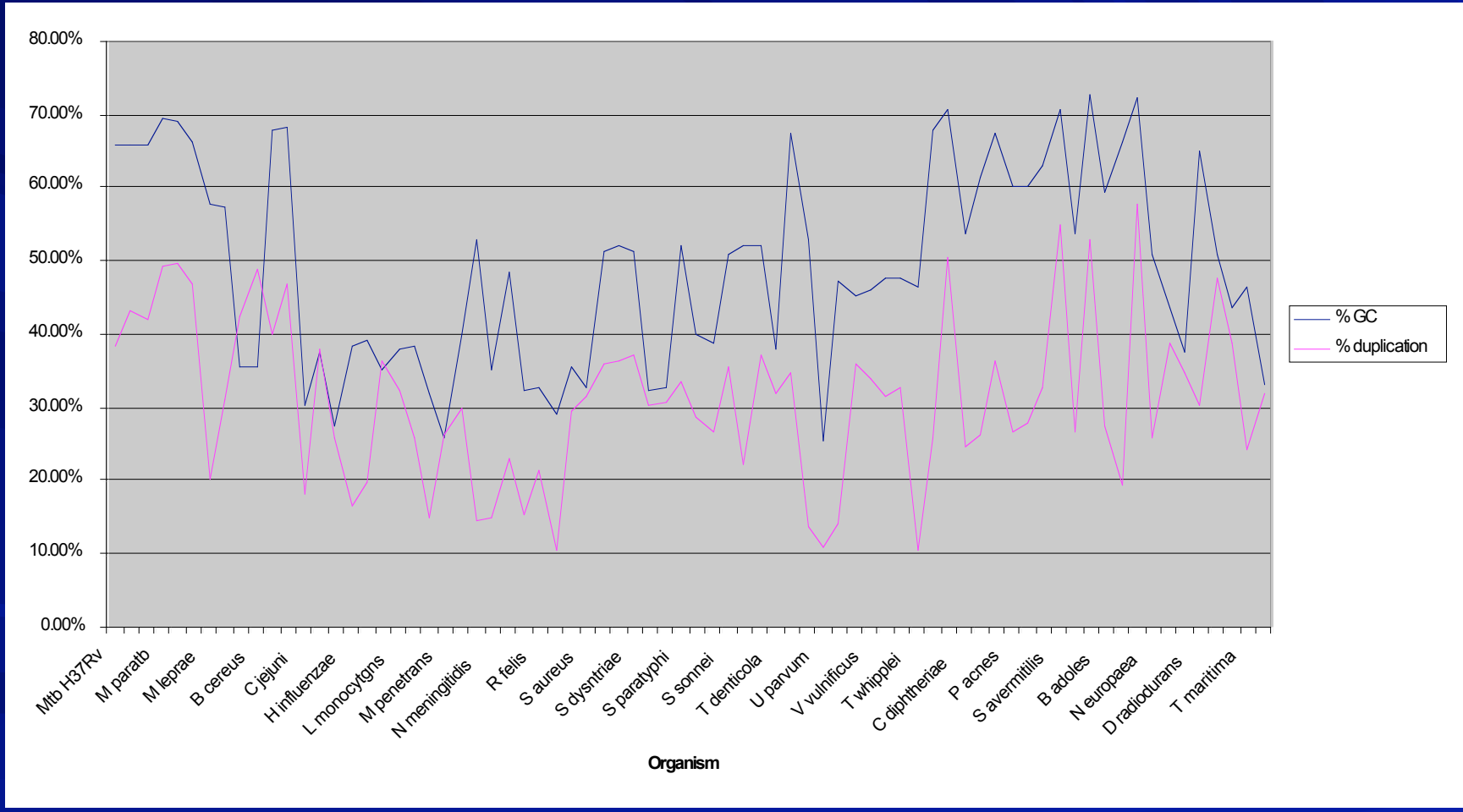
Duplication in multiple genomes



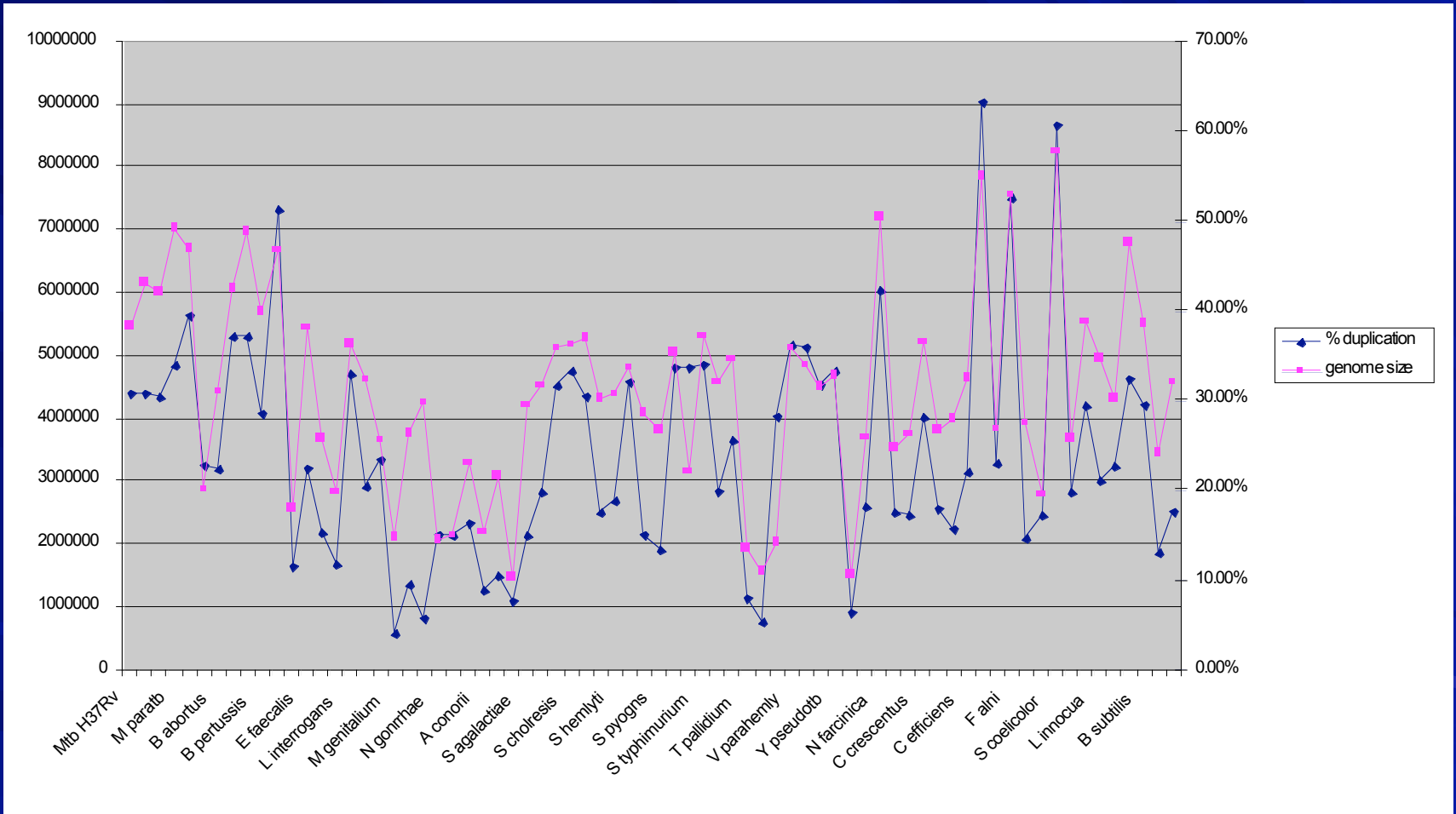
CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Relationship between GC content and % duplication



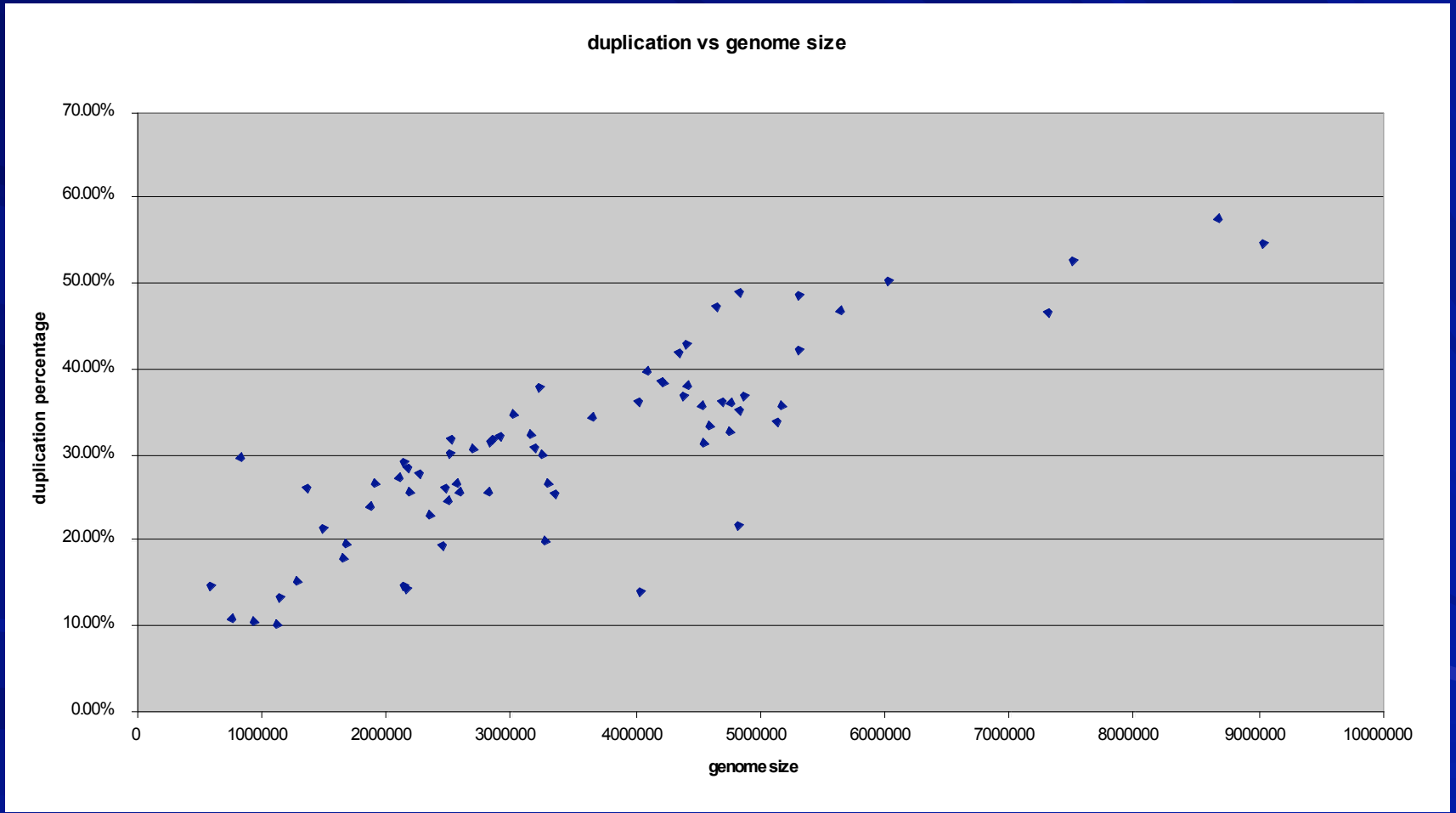
Relationship between genome size and % duplication



CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Relationship between genome size and % duplication



CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

Duplicated families in *M.tb* and *M. leprae*

Combined InterPro and BLAST results gave 354 clusters

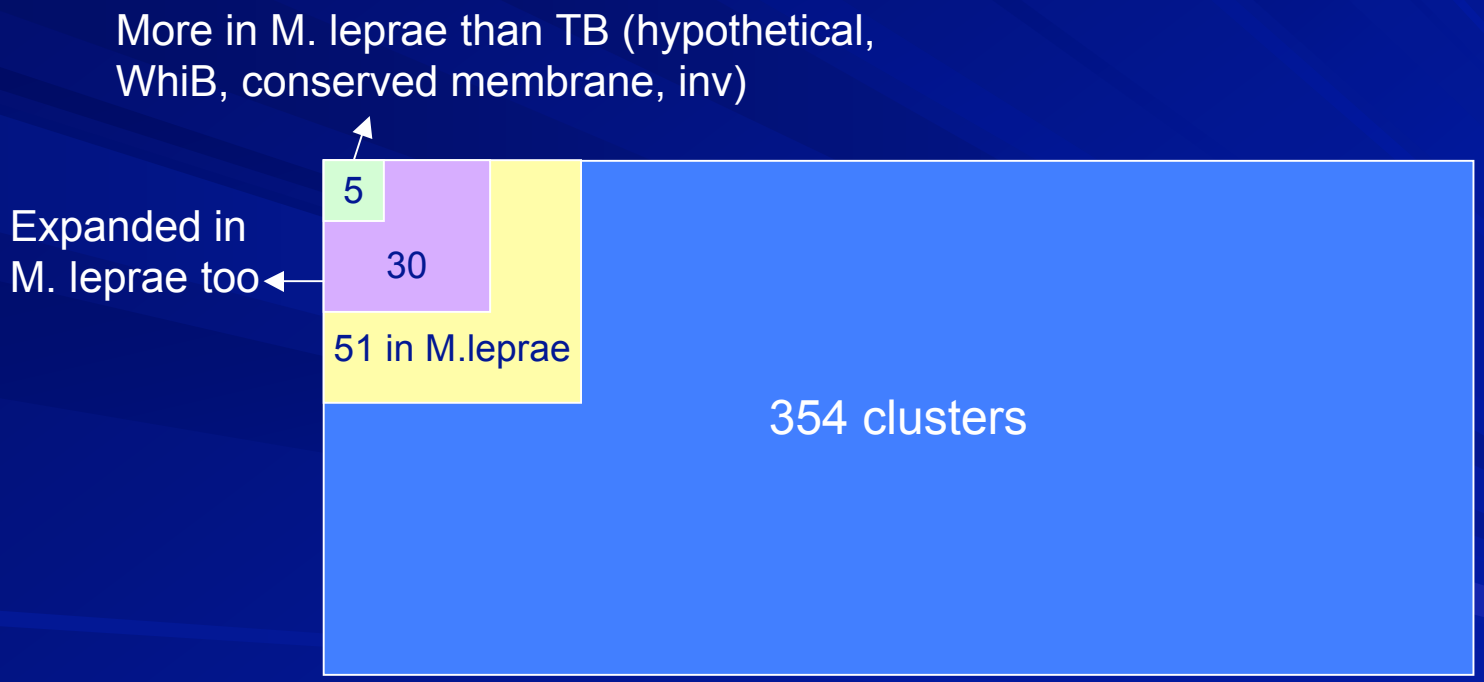
354 clusters

CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK

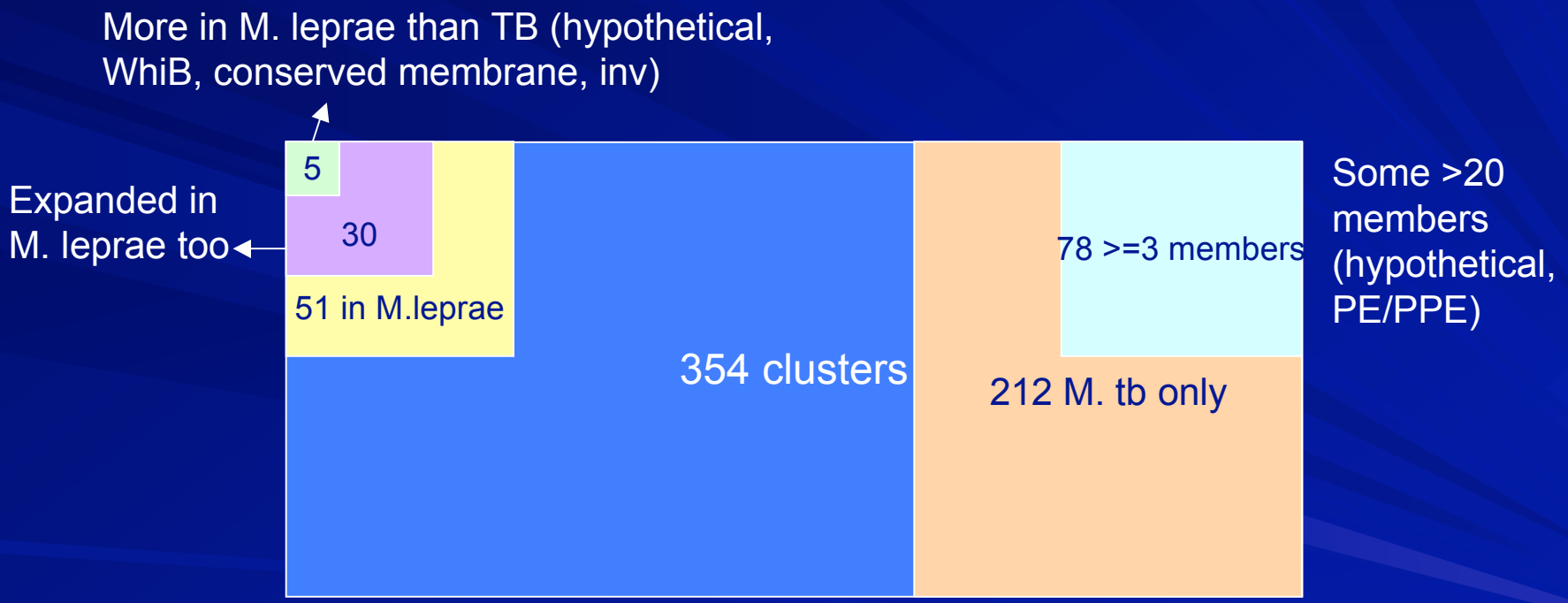
Duplicated families in *M.tb* and *M. leprae*

Combined InterPro and BLAST results gave 354 clusters



Duplicated families in *M.tb* and *M. leprae*

Combined InterPro and BLAST results gave 354 clusters



Future work

- Investigate hypothetical protein families further
- Determine functional fate of expanded families that are unique to TB or TB complex
- Measure diversity vs cluster size
- Compare diversity in upstream and coding regions
- Look at expression correlation of expanded families

Acknowledgements

- Halimah Rabiu
- Venu Vuppu
- Gordon Jamieson
- NBN funding

CBIO Computational Biology @ UCT

NATIONAL BIOINFORMATICS NETWORK