

Microbial Genome & Metagenome Analysis: Computational Challenges

Natalia N. Ivanova *
Nikos C. Kyrpides *
Victor M. Markowitz **

* Genome Biology Program, Joint Genome Institute
** Lawrence Berkeley National Lab

Microbial genome & metagenome analysis

- ❖ General aims
 - Understand microbial life
 - Apply to agriculture, bioremediation, biofuels, human health
- ❖ Specific aims include
 - Predict biochemistry & physiology of organisms based on genome sequence
 - Explain known biochemical & physiological properties

☞ Metabolic reconstruction

```

50CTGGAAAAGGACGACTAAGCA
IACCGGTATACGGGTGAGGTAGT
50TGGTGGGATGAGGTGAGACTC
50TCAAGTGGTTCAGGTATGGT
50TCCCTGCTGCGGGCAGCGATT
50CGGAACAAGATCGCTGTGGC
50GAAGGAAAGGCTGGGTATGCTY
IAGCGCGAGGTGATCACGGGGGT
50CACCGCGAGTTCATCAAGGAK
TGAAGGTCGCAATGGGAAGACI
TGGCAAGCACATGATCTCGTAAI
TATCGTGGCGAGATGGTGGCAI
TCTGGGACACACAGGTGGTGGI
50GCTGCGGATGGGCGCTGGI
50TATGGGAAGCGGGGTGGAK
IATTAAGCGGAGGTGATTTGTY
50AGCGGAGGTCCTGCTGCTGCI
50AGCTCAACCGGACTGGTGGKI
50TGAAGCGGAGTTCCTGTAI
50AGGAGCTGGATGCTGCTGAI
50TGGAGACTGGTACTGACCI
50CGGACAGGCTGGGCGGAI
50AGCGGAGCTGGTGGTGGCI
50AGCGGAGCTGGTGGTGGCI
50AGCGGAGCTGGTGGTGGCI

```

```

graph TD
    GS[Genome sequence] --> SIM[Sequence interpretation module]
    EMP[Empty metabolic pathways] --> MIM[Metabolic information interpretation module]
    AS[Annotated sequences] --> SIM
    EI[Experimental information] --> MIM
    SIM --> PDI[Pathway dependencies]
    SIM --> PWE[Pathways with evidence]
    MIM --> PWE
    PWE --> RP[Reconstructed pathways]
    PDI --> RP
  
```

* Ivanova & Lykidis (2009) Metabolic reconstruction. *Encyclopedia of Microbiology*, Elsevier: 607-621.

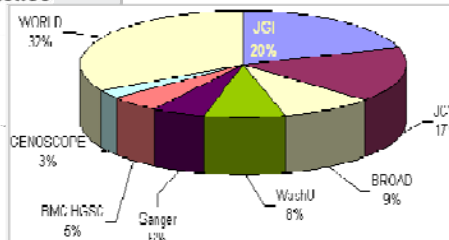
Genome sequence data size

Now

- ~1,400 microbial genomes = 5.5 mil genes
- ~ 100 metagenomes samples = 2.5 mil genes

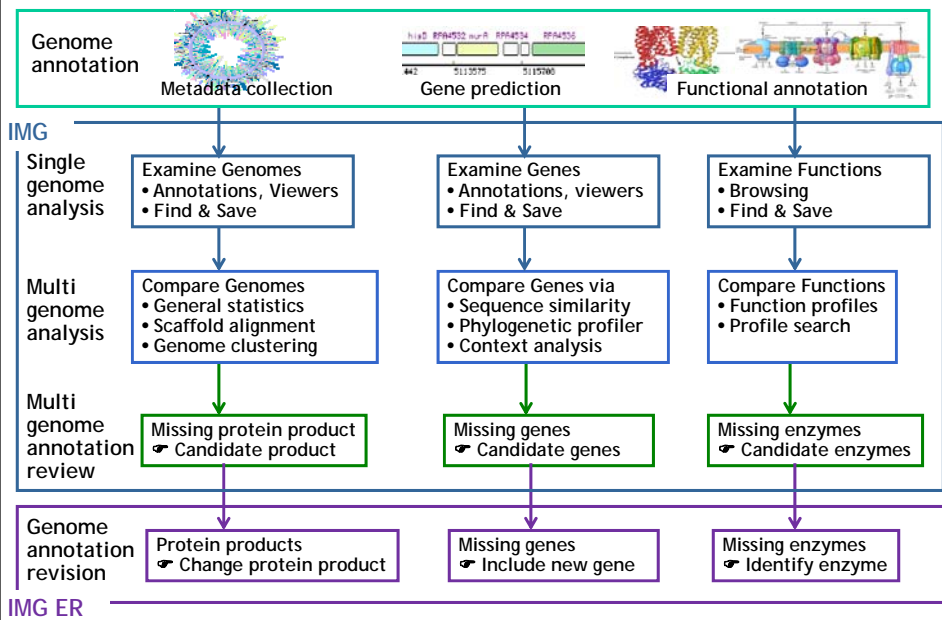
Next 3 Years

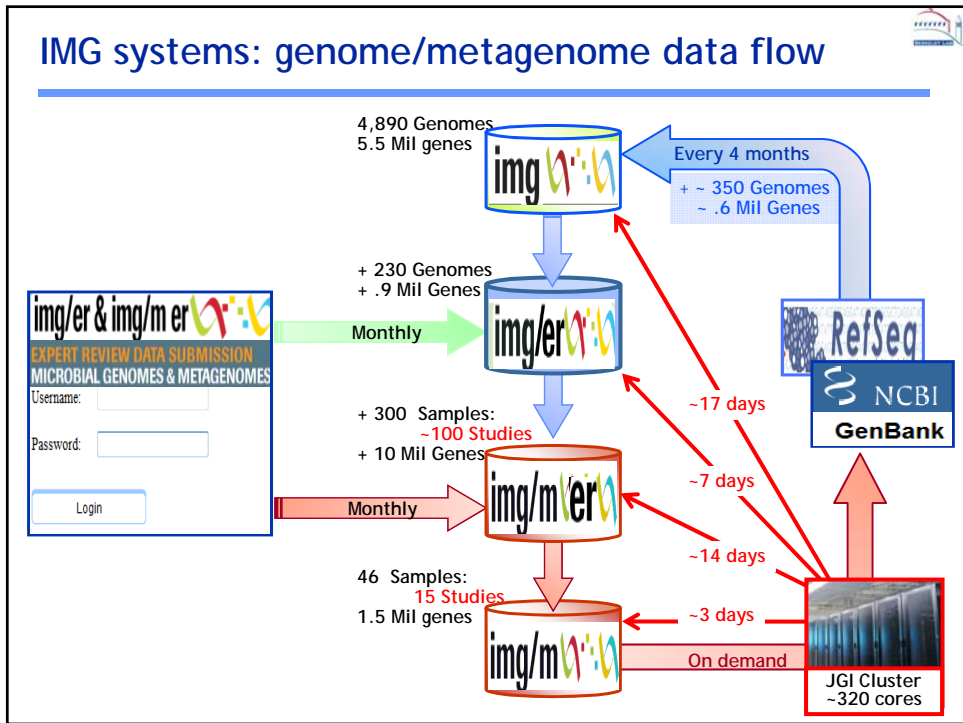
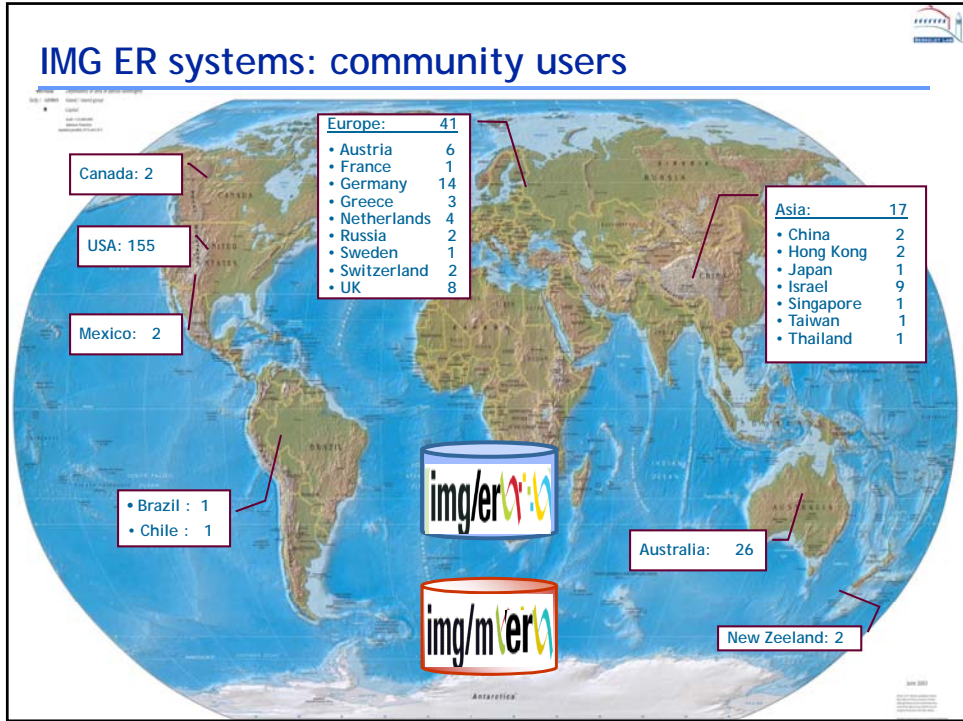
- ~10,000 microbial genomes = 40 mil genes (4K genes/genome)
- ~ 400 fungal genomes = 4 mil genes (10K genes/genome)
- ~ 200 plant genomes = 6 mil genes (30K genes/genome)
- ~ 500 metagenome samples = 50 mil genes (100K genes/sample)



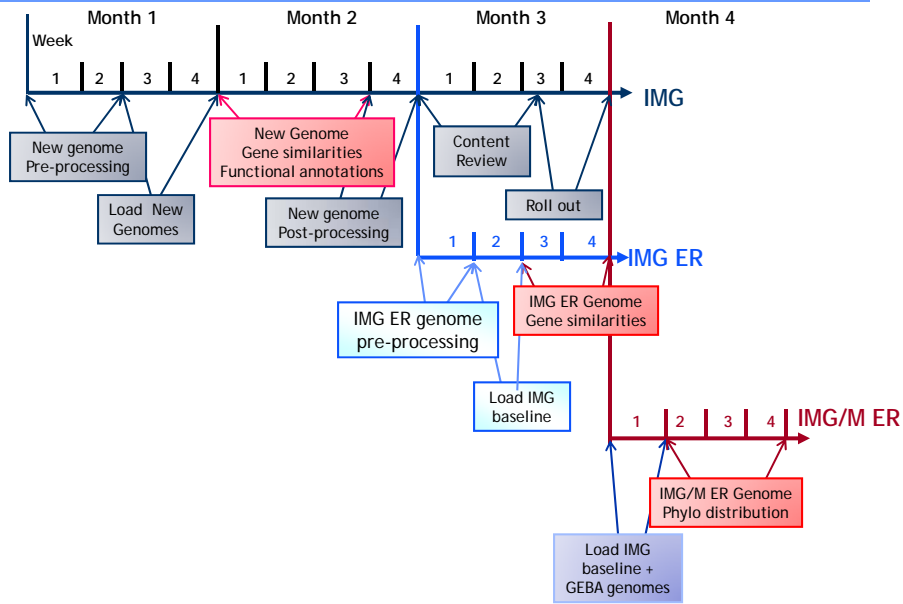
Major Sequencing Centers
(Jan 2009: 4,370 projects)

Genome sequence data processing & analysis





IMG systems: maintenance process



Current requirements

- ❖ Architectures: Linux cluster
- ❖ Typical production run:
 - 232 cores, 4 GB/core, 500 GB data read/written
 - 20 TB on-line storage
- ❖ Data read/written: ~ 1 GB
- ❖ Current primary codes and their methods or algorithms
 - Various flavors of BLAST
 - HMM protein searches
- ❖ Known limitations/obstacles/bottleneck
 - Managing, reformatting, reordering computation results