KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

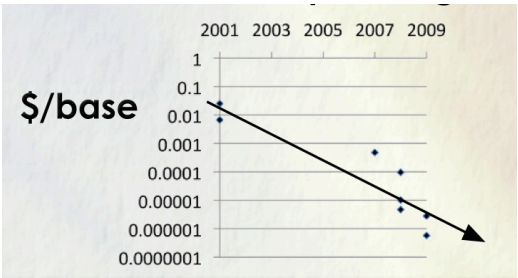
DOE Systems Biology Knowledgebase

KBase

Integration and modeling for predictive biology

Communities and Metagenomics

- Cheap DNA sequencing revolutionizes microbial community study
- Separate tools provide different analysis capabilities
- We lack tool integration
- Currently IMG/M, MG-RAST, Meta Microbes Online
- There are several MISSING CRITICAL PIECES THAT ALLOW INTEGRATION!!



Source: Rob Knight, UColorado

1	MG-RAST metagenomics analysis server	10.87 TBp DNA 100 billion sequences
2	metamicrobesonline	Gene families and gene trees
3	img/m	11.6 million protein coding genes

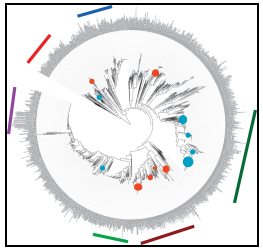
- Integration via API
- new functions via API
- Generate consistent view on microbial communities
- Expose data and services via API (for 1 and 2)

Bioprospecting

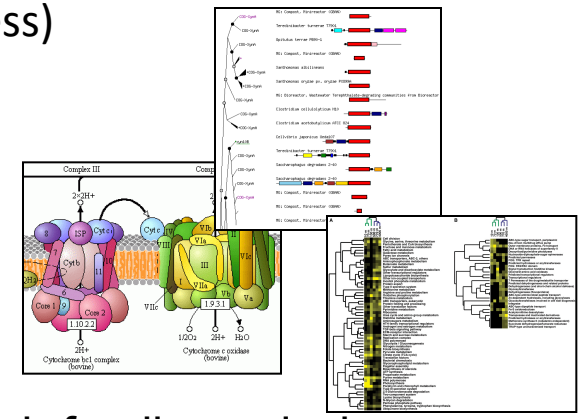
- Integration of MG-RAST and MetaMO (DONE)
- Provide novel functionality
  - Combining existing features
- Hires gene/protein trees for metagenomes
- Via mapping into existing protein trees for proteins form complete organisms
- **Work:**
- 1) Retrieve members of gene family from MG-RAST e.g. "glycoside hydrolase" for several metagenomes (DONE)
- 2) Integrate reads into existing protein trees (in progress)
- 3) Sub family analysis (exists in MetaMO)

```
/Communities/
reads_for_func
tion()
```

Reads for Protein
1



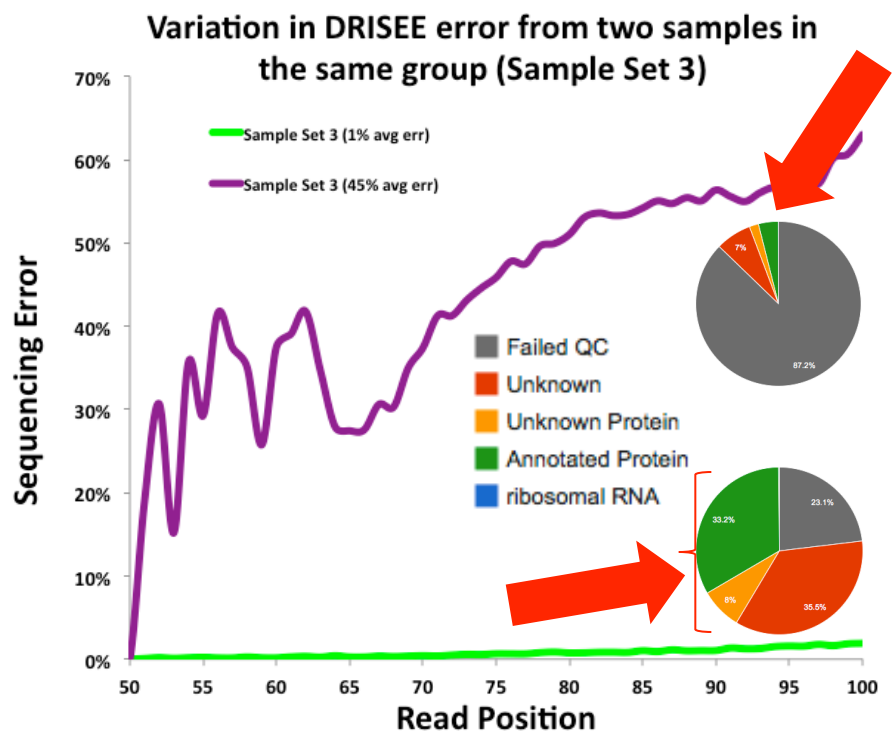
Protein tree with reads
2



Sub family analysis
3

However quality is all important

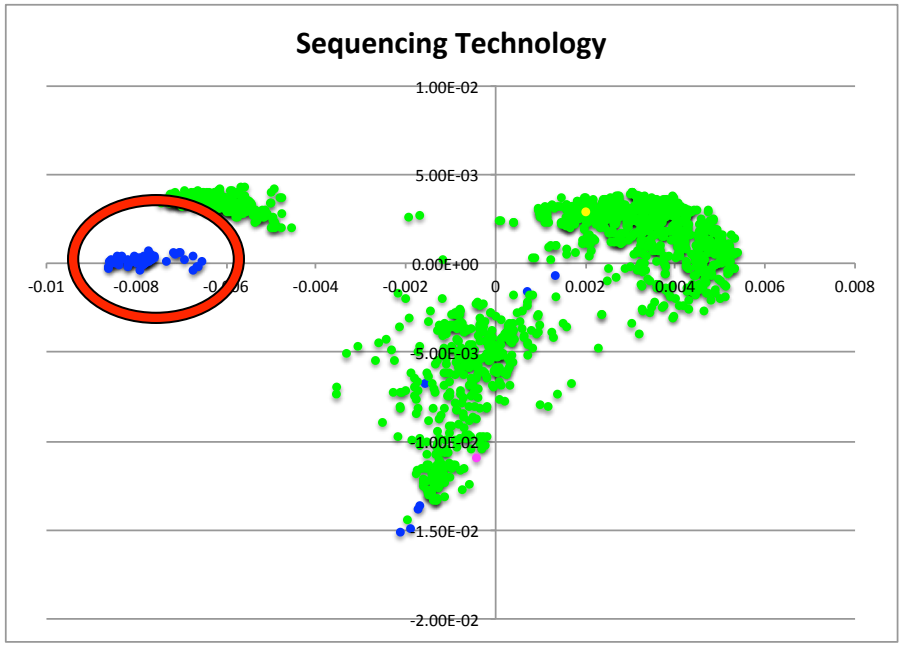
- DNA data comes from many sources
  - Often unassembled
  - Many formats
  - Lacks metadata
- Result: Data quality is mostly ignored
- DNA sequence quality can easily outweigh biological signal
- **Example**: fraction of predicted genes (green+orange) for two **real** samples with 1% and 45% average predicted error!
- This is not biology, this is error!



Metagenomic data quality

- For large scale comparison or integration studies, normalizing samples is important
- Example:
  - Real data, ~1500 HiSeq lanes, 10GBp per sample

- New Cluster emerges that is based solely on sequencing technology used
- Sample origin (body site) plays a less significant role than technology

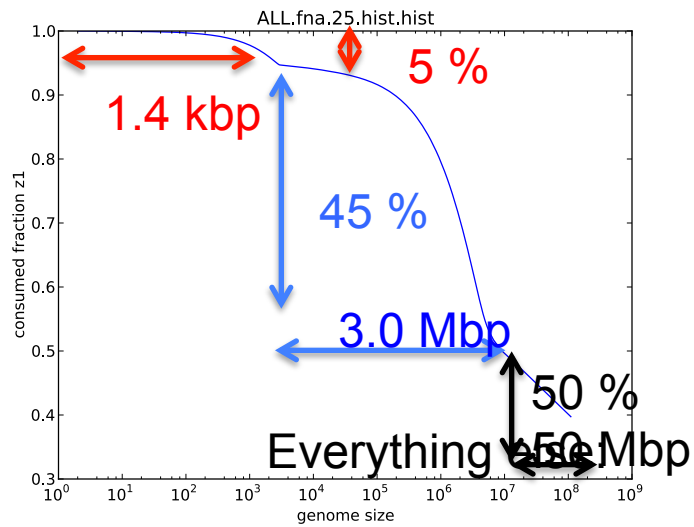


Showing PCoA
normalized euclidian distance
Relative Subsystem abundance

Coloring:
Green = Illumina data
Blue = 454 data

Experiment Design Wizard

- Provide QC for existing data
  - MG-RAST QC (pre-existing)
  - Measure read quality (DONE)
  - Measure other noise (e.g. adapters) (DONE)
- Reconstruct community
  - Derive alpha diversity (DONE)
  - Predict capability to assemble (in progress)
    - E.g. is there a dominant species
    - Low diversity



Predicting replicons from k-mers

New tool will:
•Inform user of data analysis options
Or
•More sequencing options
E.g. add a mate pair run for $1000 and preform assembly

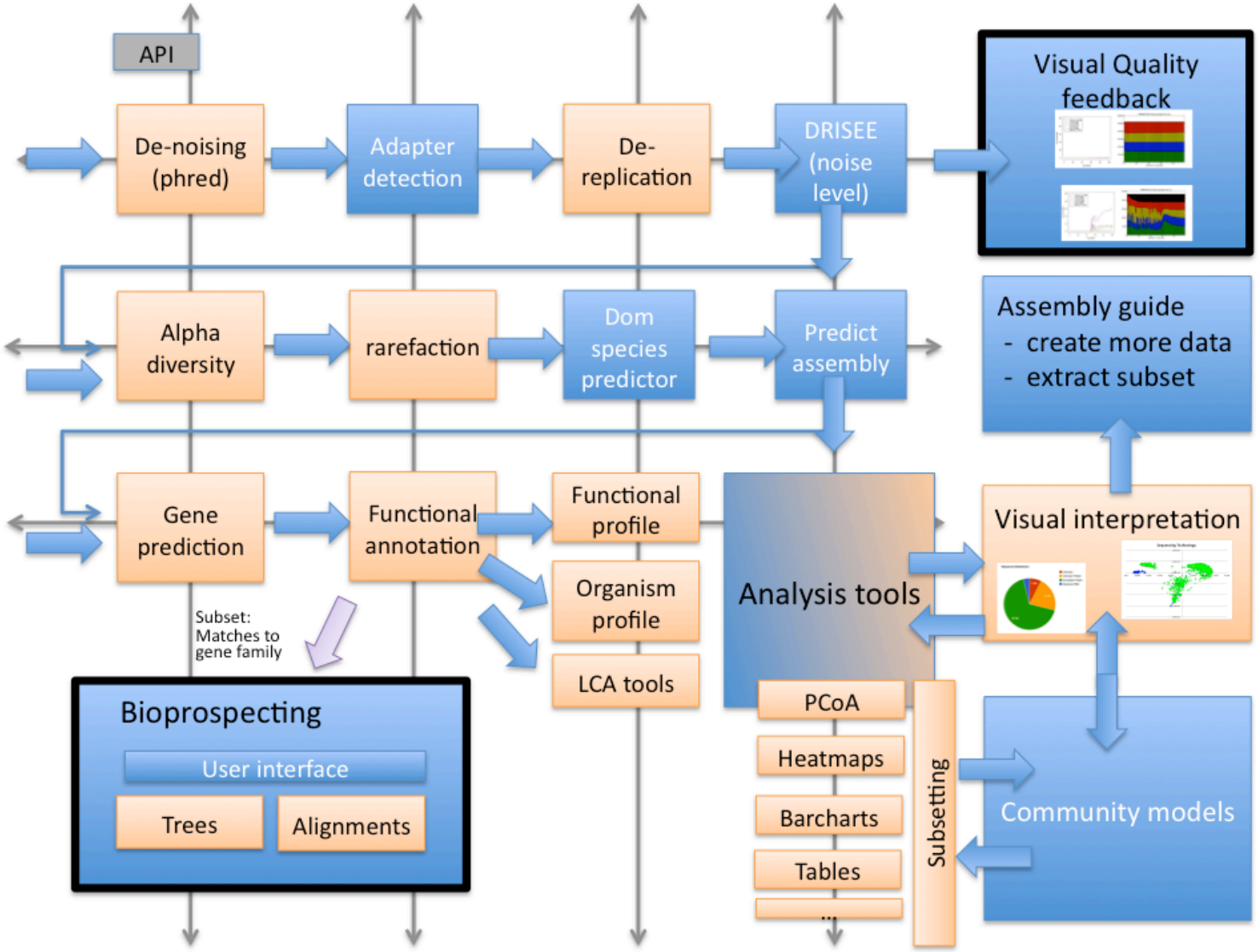
Integration within Communities

- One common API (as part of the overall KBASE API)
- One data access paradigm
- All data QC'ed with common standard an quality information is available

- System is scalable
- Extensible (easy to build stand alone tools consuming the API)
- Allows many group involvement

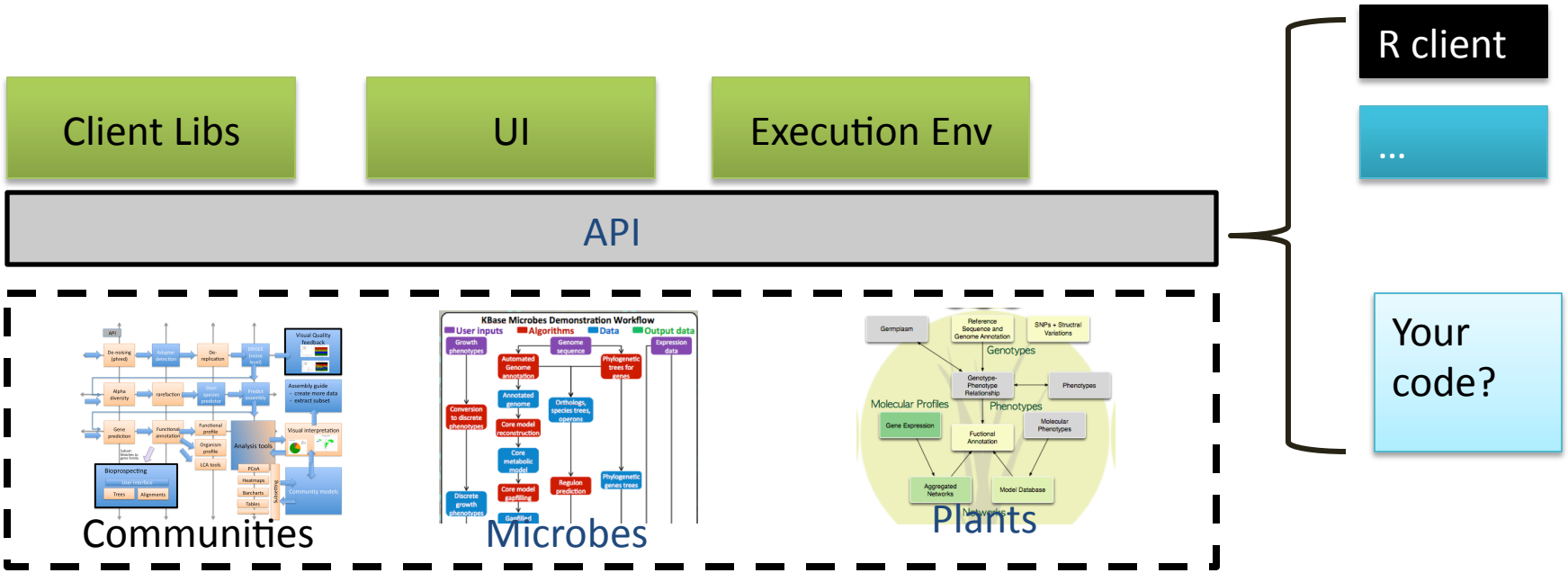
- Already has first customers
  - ➔ There is a real need in the community for this kind of integration.
- We are also eating our own dog food (consuming our own APIs)

Workflow/Plan



Integrated into KBASE runtime and other domains

- API allows integration into existing tools
  - e.g. R client for microbial community analysis
- API also contains microbes and plants
- Integration of all 3 domains inside a single environment
  - Plus runtime environment (Magellan)



Acknowledgements

- Tom Brettin, ORNL
- Dylan Chivian, LBNL
- Travis Harrison, ANL
- Narayan Desai, ANL
- Kevin Keegan, ANL
- Keith Keller, LBNL
- Tobias Paczian, ANL
- Will Trimble, ANL
- Andreas Wilke, ANL
- Jared Wilkening, ANL

@ 1st communities hackathon
LBNL, Jan 2012

@ 1st integration meeting
ANL, Feb 2012