UNIVERSIDAD DE CHILE

Predicting the metagenomic content using multiple CART trees

Rodrigo Assar
School of Medicine, U. Chile.

October 28, 2014
BMTL 2014, Naples

Co-authors: Dante Travisany, Diego Galarce, Alejandro Maass

CMM
Center for Mathematical Modeling

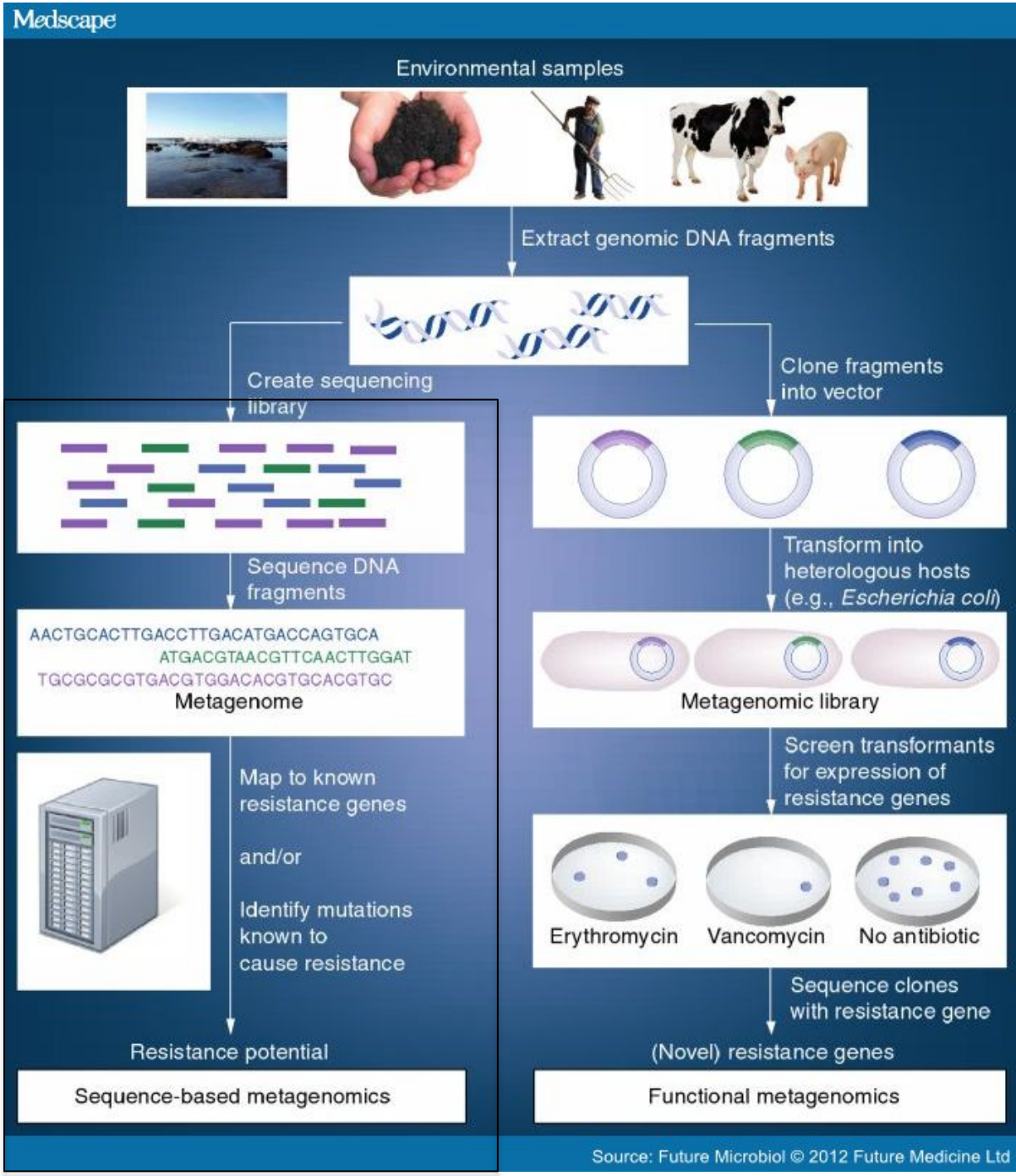
Topics

- Introduction:
  - Metagenomics and Microbiome
- Metagenomic Pipeline:
  - Classic steps
  - Including Multiple CART trees
- Results & Conclusions

Metagenomics

Given an environment and sample DNA sequences:

- Present organisms?
- Concentrations?



http://www.medscape.com/

Why using sequence-based Metagenomics?

- >99% of the microbials in nature are non-culturable by available techniques.
- Metagenomics is cultivation-independent.

Review

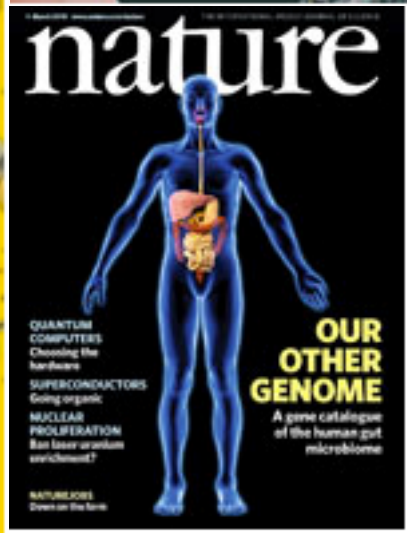
Nature Reviews Genetics **6**, 805-814 (November 2005) | doi:10.1038/nrg1709

Metagenomics: DNA sequencing of environmental samples

Susannah Green Tringe & Edward M. Rubin

Although genomics has classically focused on pure, easy-to-obtain samples, such as microbes that grow readily in culture or large animals and plants, these organisms represent only a fraction of the living or once-living organisms of interest. Many species are difficult to study in isolation because they fail to grow in laboratory culture, depend on other organisms for critical processes, or have become extinct. Methods that are based on DNA sequencing circumvent these obstacles, as DNA can be isolated directly from living or dead cells in various contexts. Such methods have led to the emergence of a new field, which is referred to as metagenomics.

Microbiome studies: Literature explosion



You are what you ~~EAT~~ HOST

Cell

Volume 156, Issue 3, 30 January 2014, Pages 408–411

Minireview

You Are What You Host: Microbiome Modulation of the Aging Process

Caroline Heintz[1], William Mair[1,*]

Show more

DOI: 10.1016/j.cell.2014.01.025

The critical impact that microbiota have on health and disease makes the interaction between host and microbiome increasingly important as we evaluate therapeutics. Here, we highlight growing evidence that, beyond disease, microbes also affect the most fundamental of host physiological phenotypes, the rate of aging itself.

Impact on Health

http://www.human-microbiome.org

MetaHIT
Metagenomics of the Human Intestinal Tract
European research project

INTRODUCTION

Since 2008, researchers with the European consortium MetaHIT have been analyzing the collected genomes of the microorganisms present in our intestine : the microbiota.

RESEARCH

Little understood until now, the intestinal microbiota interests researchers as an avenue of inquiry to explain the evolution of chronic diseases.

FINDINGS

The MetaHIT consortium published two major findings in the scientific journal Nature : an established catalog of bacterial genes in the intestine; and the discovery of enterotypes.

PERSPECTIVES

MetaHIT opens avenues for further efforts in the field of human microbiome research : early detection of chronic diseases, personalized medicine and more healthful food.

Budget

22 million euros

The 4 year program was financed in large part by the European Union under the FP7 (*7th Framework Programme*).

Laboratories

8 countries
14 research & industrial

institutions are involved in the consortium, with more than 50 researchers and cooperation between Europe and China.

The microbiota

The microbiota is an ecosystem composed of billions of bacteria that make up a veritable "organ." Within 24 hours of birth, these bacteria colonize our digestive tract to form our intestinal microbiota (2kg for adults). MetaHIT focuses on the digestive tract since it is where the largest and most diversified bacterial community lives in our body.

Observations

chronic diseases
infectious diseases

Observations made in the past 50 years cannot be solely explained by variations of our genome.

Research themes

Nutrition. Better knowledge of the intestinal microbiota of individuals will enable the nutritional needs to adapt to everyones specific nutrient needs.

Medicine. With the study of the microbiota and the established catalogue of genes, we can have an unprecedented overview of the microbiota in healthy individuals and in patients. With the discovery of enterotypes we can imagine the upcoming development of new diagnostic or even prognostic tools for human health.

DEFINITION

***Enterotypes**

There are three in the world's population, each characterized by a predominant bacterial population.

Genome sequencing

3,3 million genes

The gut bacterial gene catalog, which can be compared to a *molecular scanner*, was established by metagenomic high throughtput sequencing and allows the observation of the human gut microbiome.

Discovery of the 3 enterotypes*

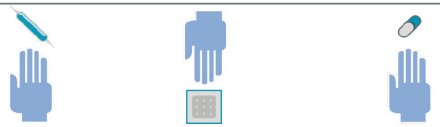
Chronic diseases

Disturbances in the microbiota can be early warning signs for certain diseases like Crohn's disease or diabetes.

Nutritional impact

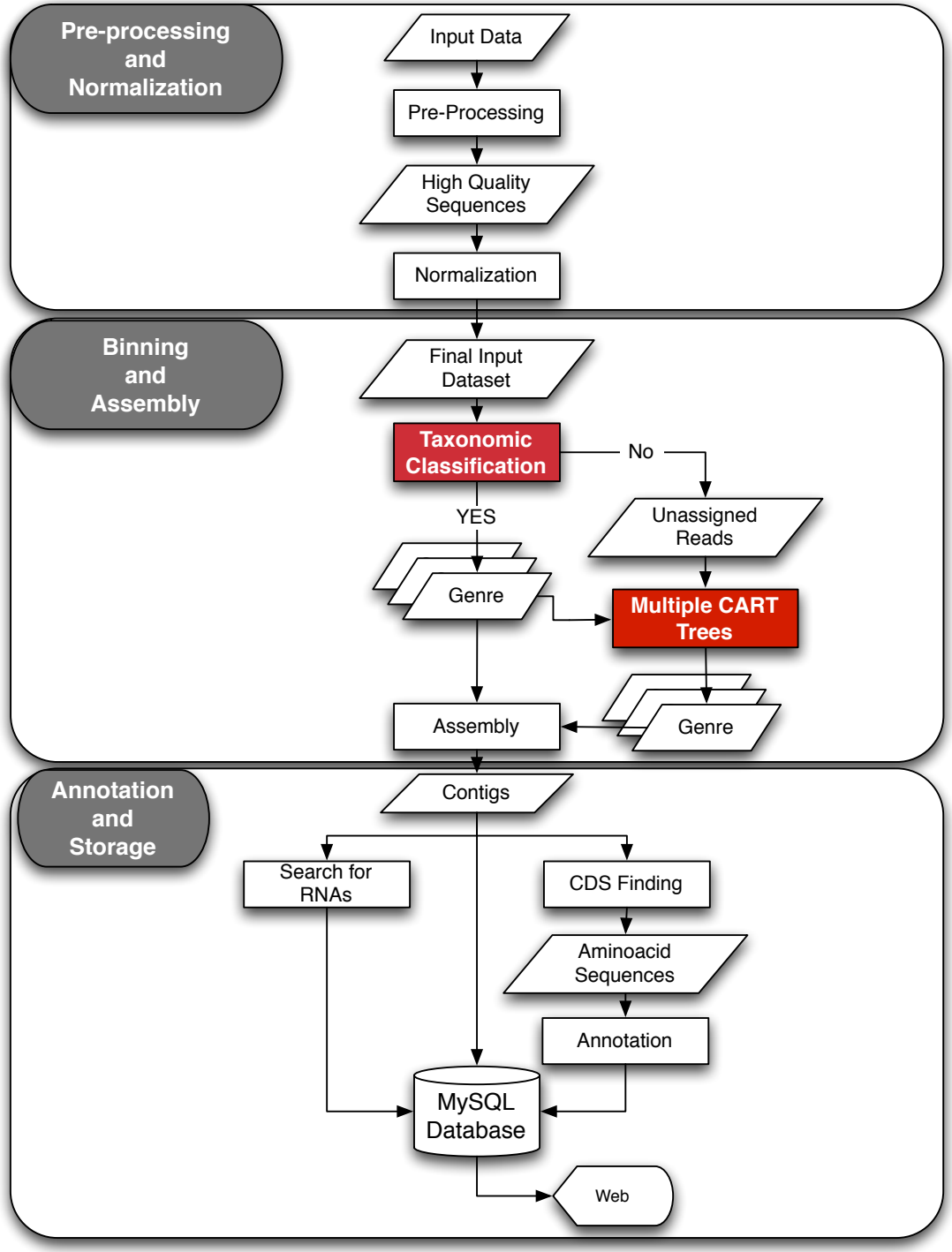
If it is possible to reveal early warning signs of obesity, one can imagine nutritional intervention and diet advice being used to reestablish a healthy microbiota. The possibility of intervening directly in the flora, in the case of disturbance to the intestinal ecosystem, could also be envisioned.

Personalized medicine



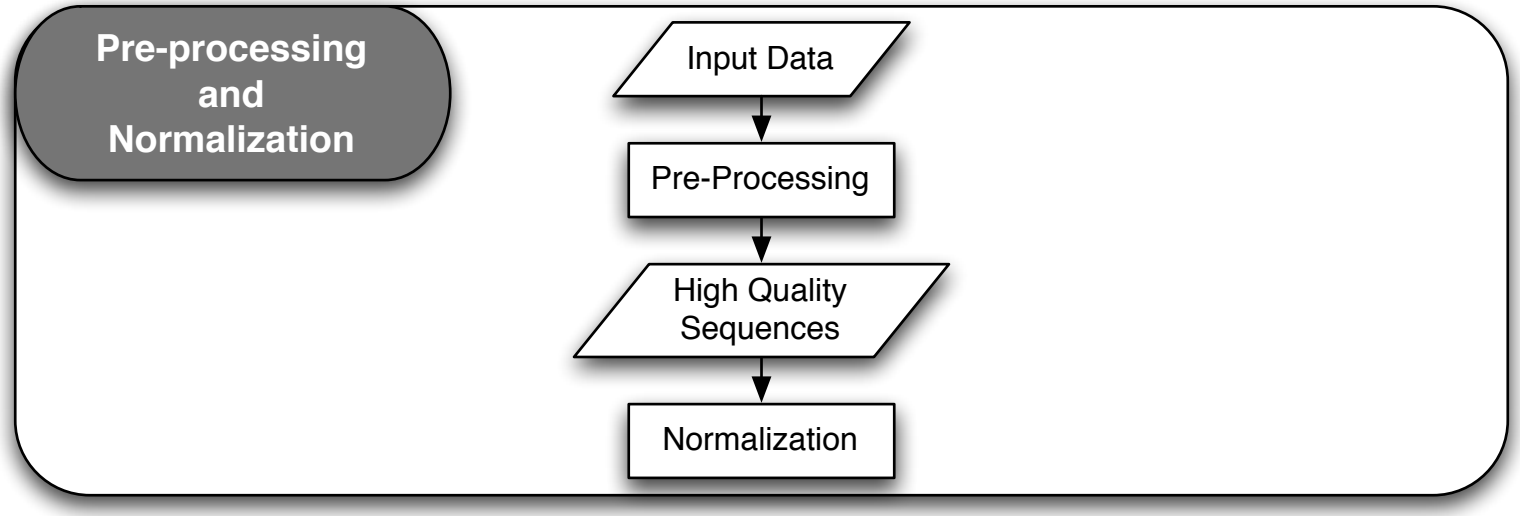
Classification by enterotype will help in the development of diagnostic tools able to reveal cases where a planned treatment would not be effective, and to adapt it accordingly.

Metagenomic Pipeline



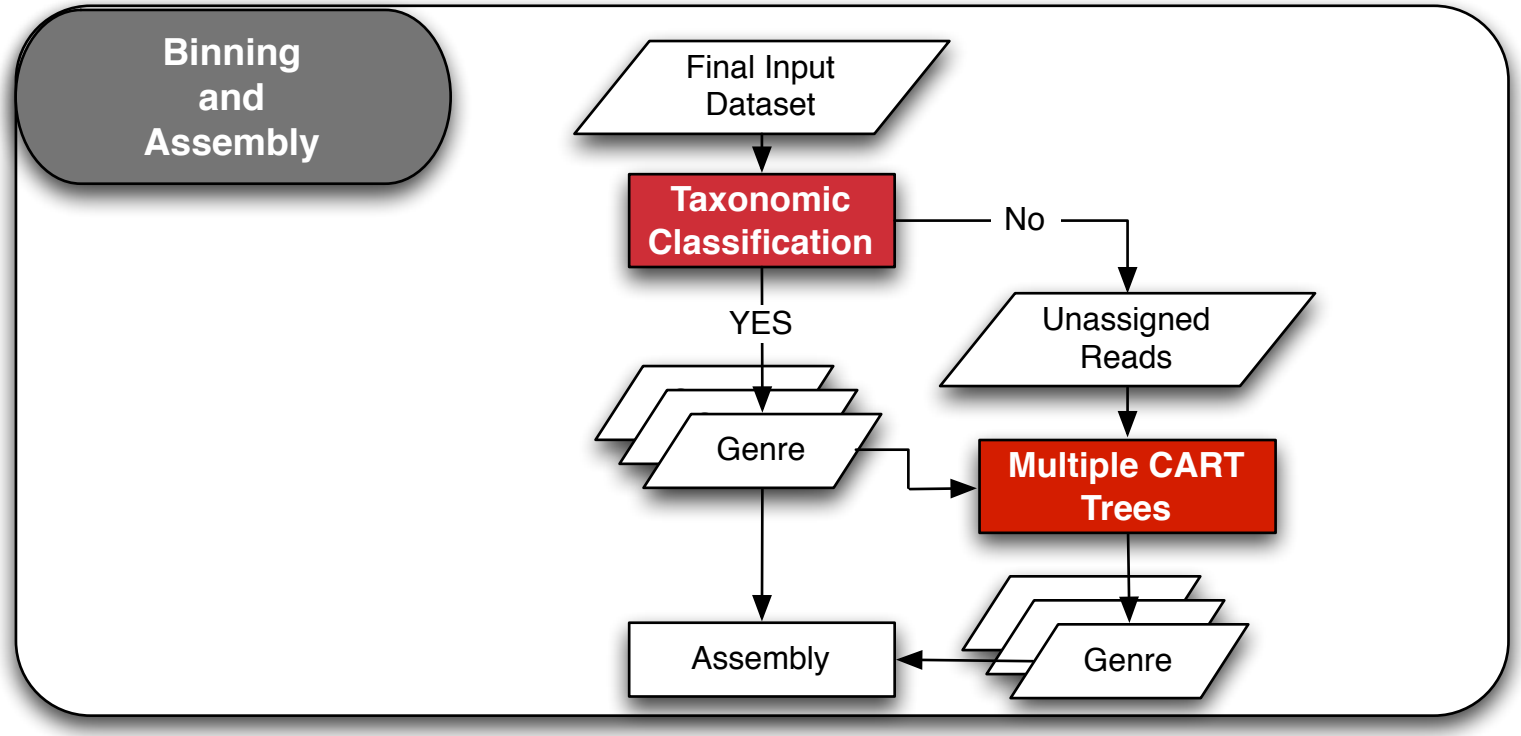
Metagenomic Pipeline

First step: Pre-processing and Normalization



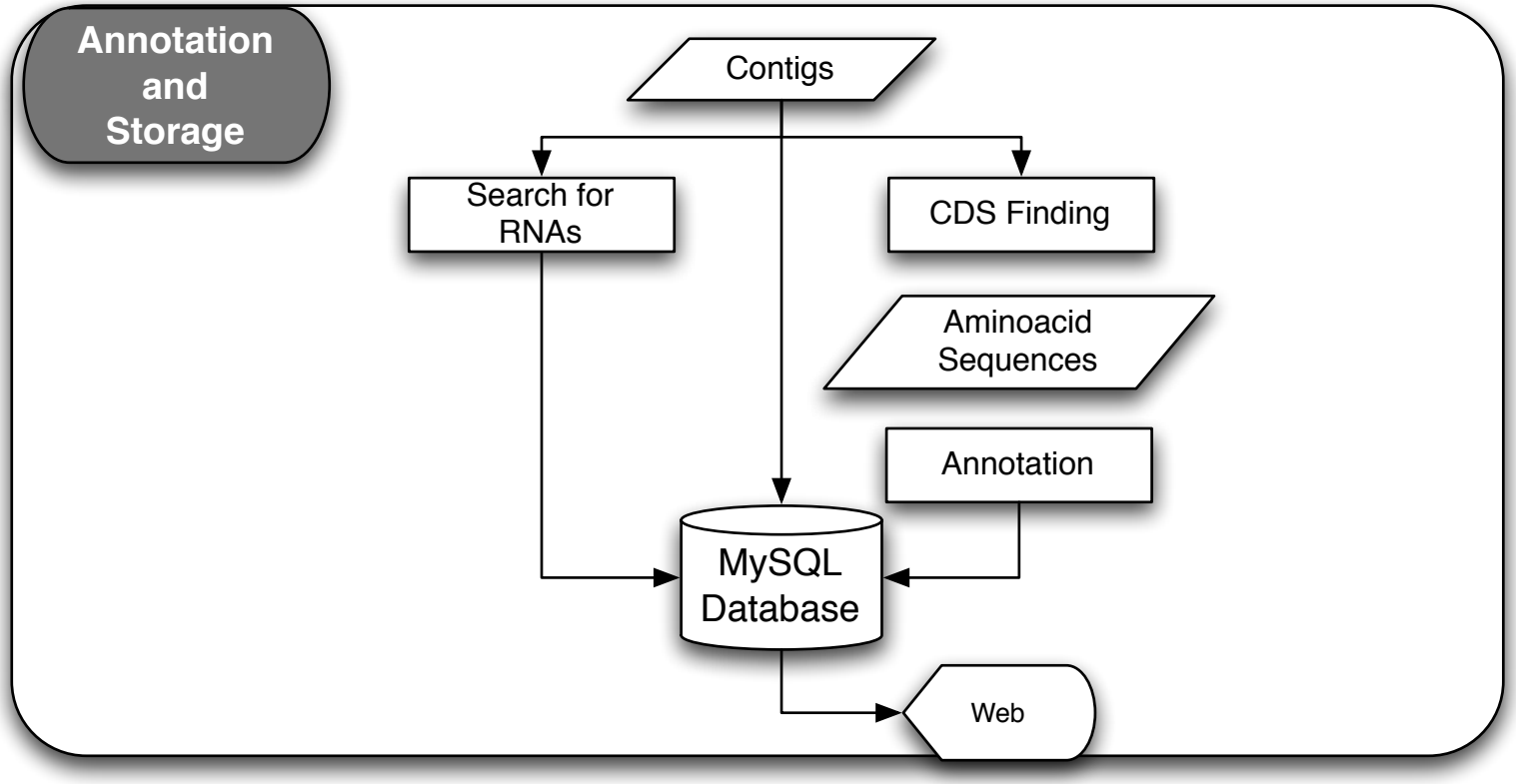
Metagenomic Pipeline

Second Step: **Classification** and Assembly



Metagenomic Pipeline

Final step: Annotation and Storage



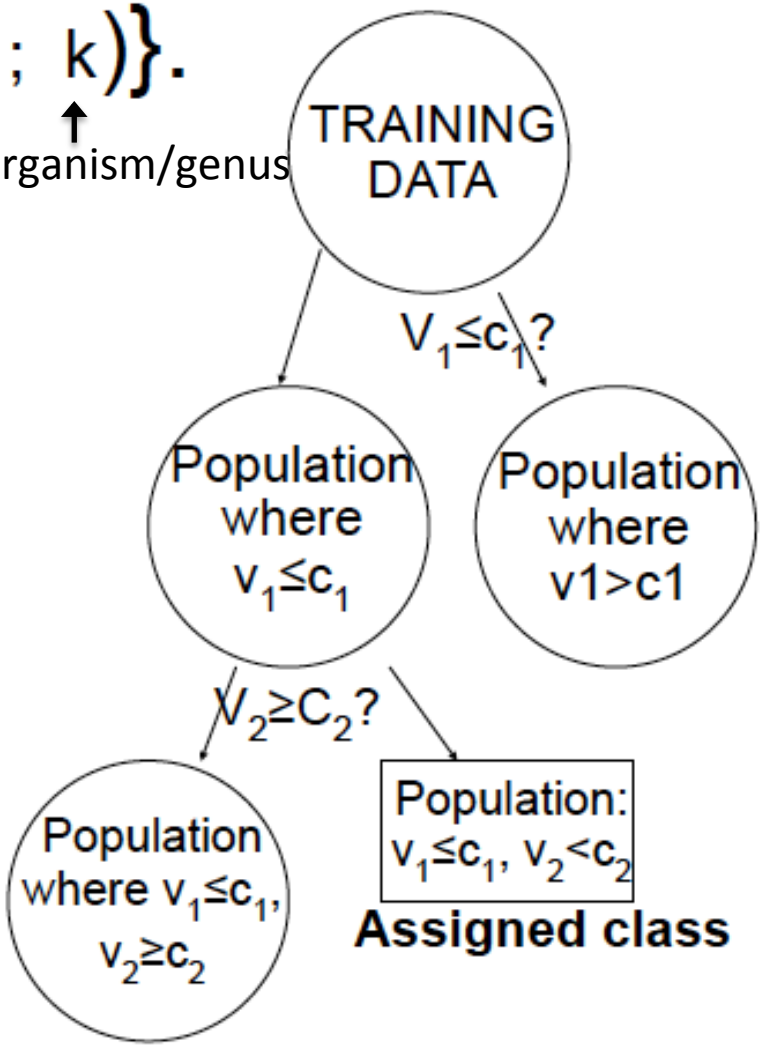
Building and pruning each tree

Training, Validation data $\{(v_1, \ldots, v_n; k)\}$.

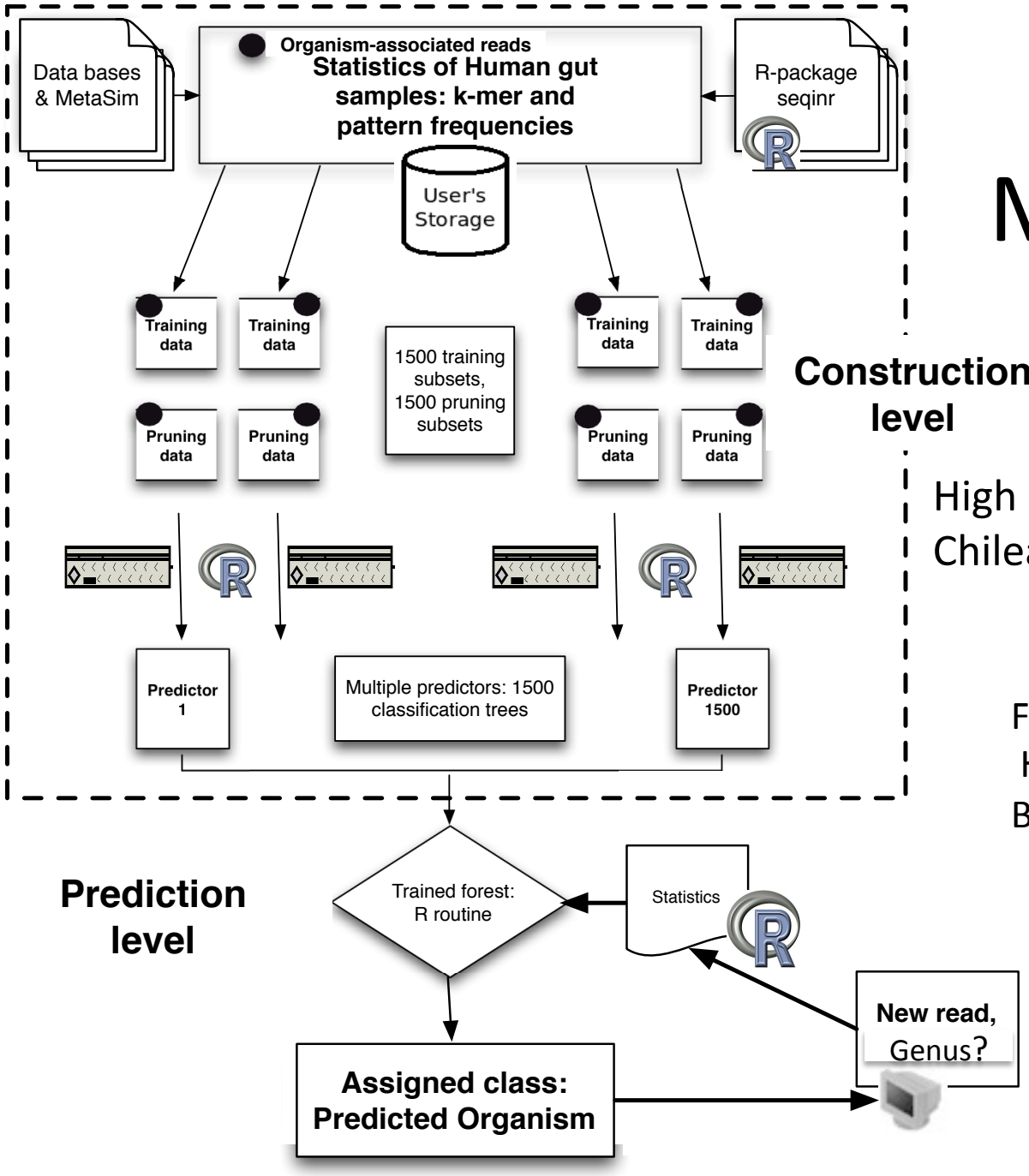
K-mer and DNApattern frequencies

Organism/genus

- Construction:
  - Start at root, continue until high separation degree.
  - Step: partition with "class homogeneity" maximization.
  - At each terminal node : class minimizing the expected misclassification.
- Validation: pruning to improve results on validation data.

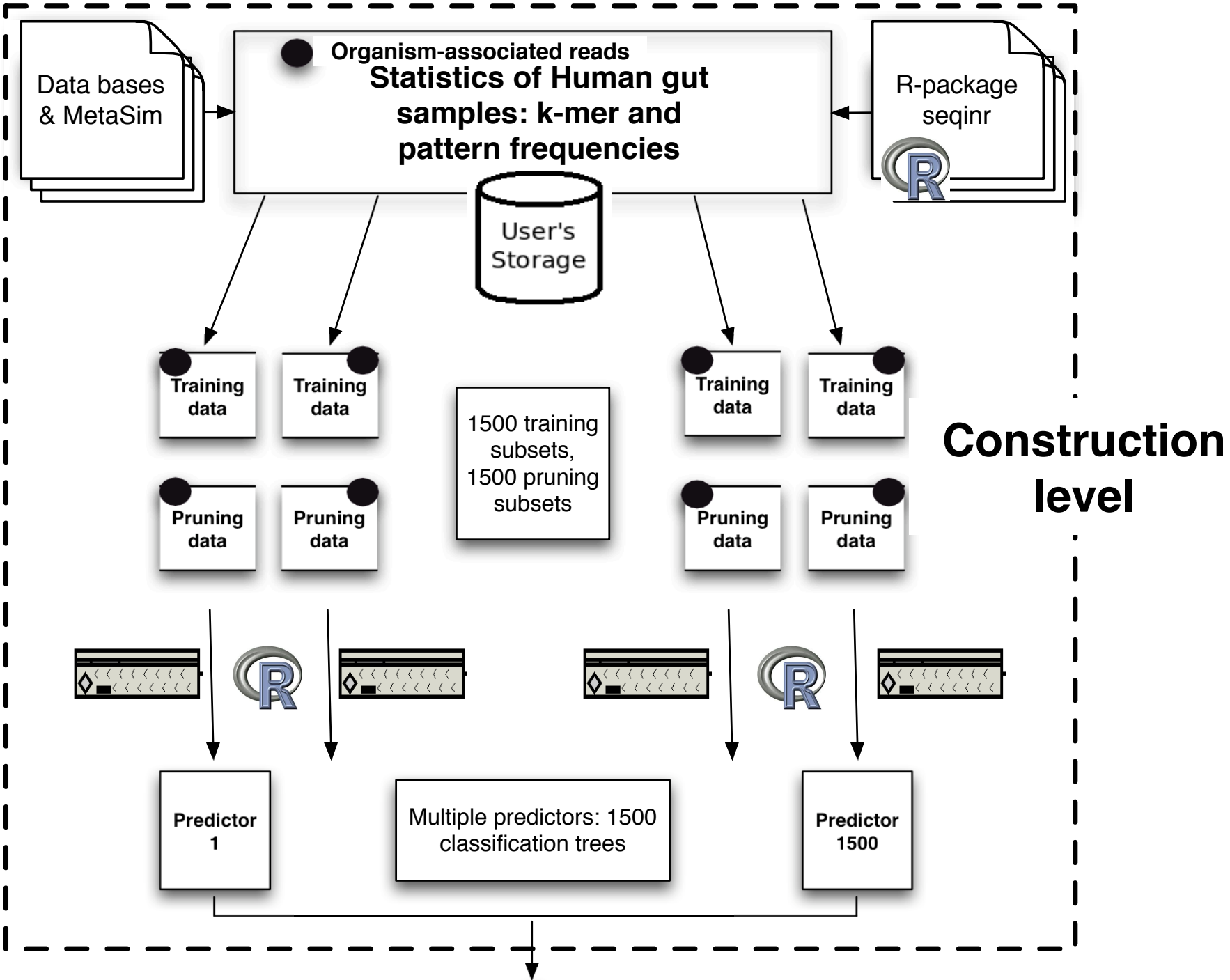


Multiple CART trees

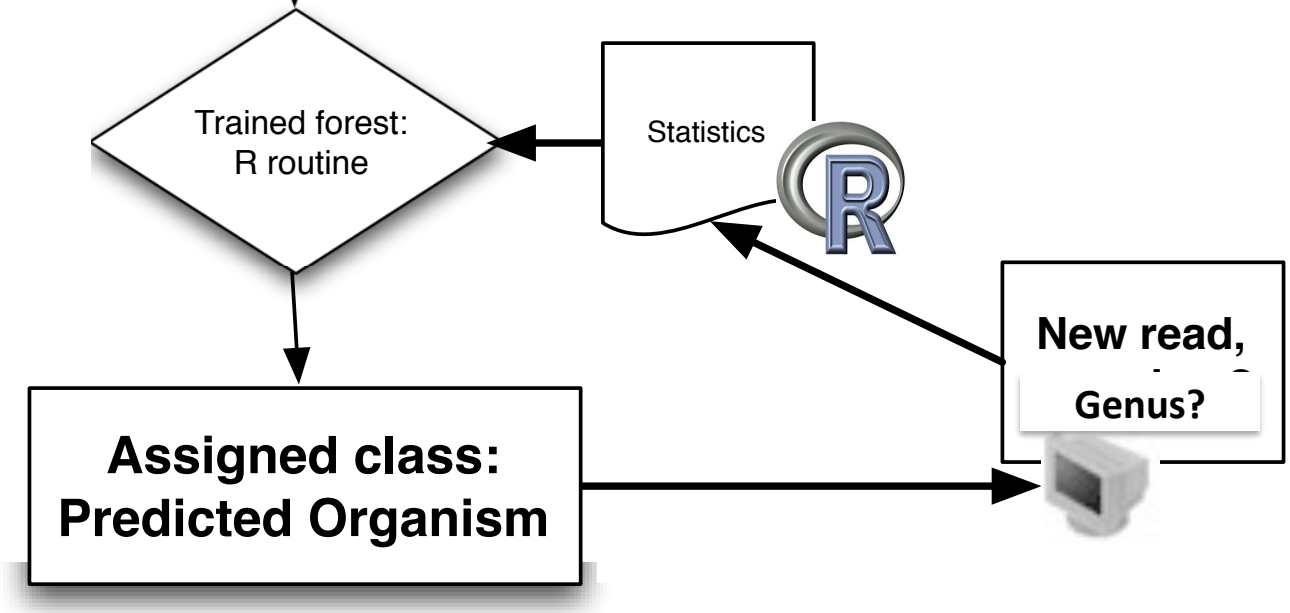


High Performance Computing: Chilean HPC-BioMed Network

Figure adapted from
Hödar, Assar, Colombres et al.
BMC Genomics (2010), 11(1): 348


Data bases & MetaSim
Organism-associated reads
Statistics of Human gut samples: k-mer and pattern frequencies
R-package seqinr
User's Storage
Training data
Training data
Training data
Training data
1500 training subsets, 1500 pruning subsets
Pruning data
Pruning data
Pruning data
Pruning data
Construction level
Predictor 1
Multiple predictors: 1500 classification trees
Predictor 1500

Prediction level



Using the a-priori trained forest:

- **Each time we want to classify**
- **From the local computer**

Data

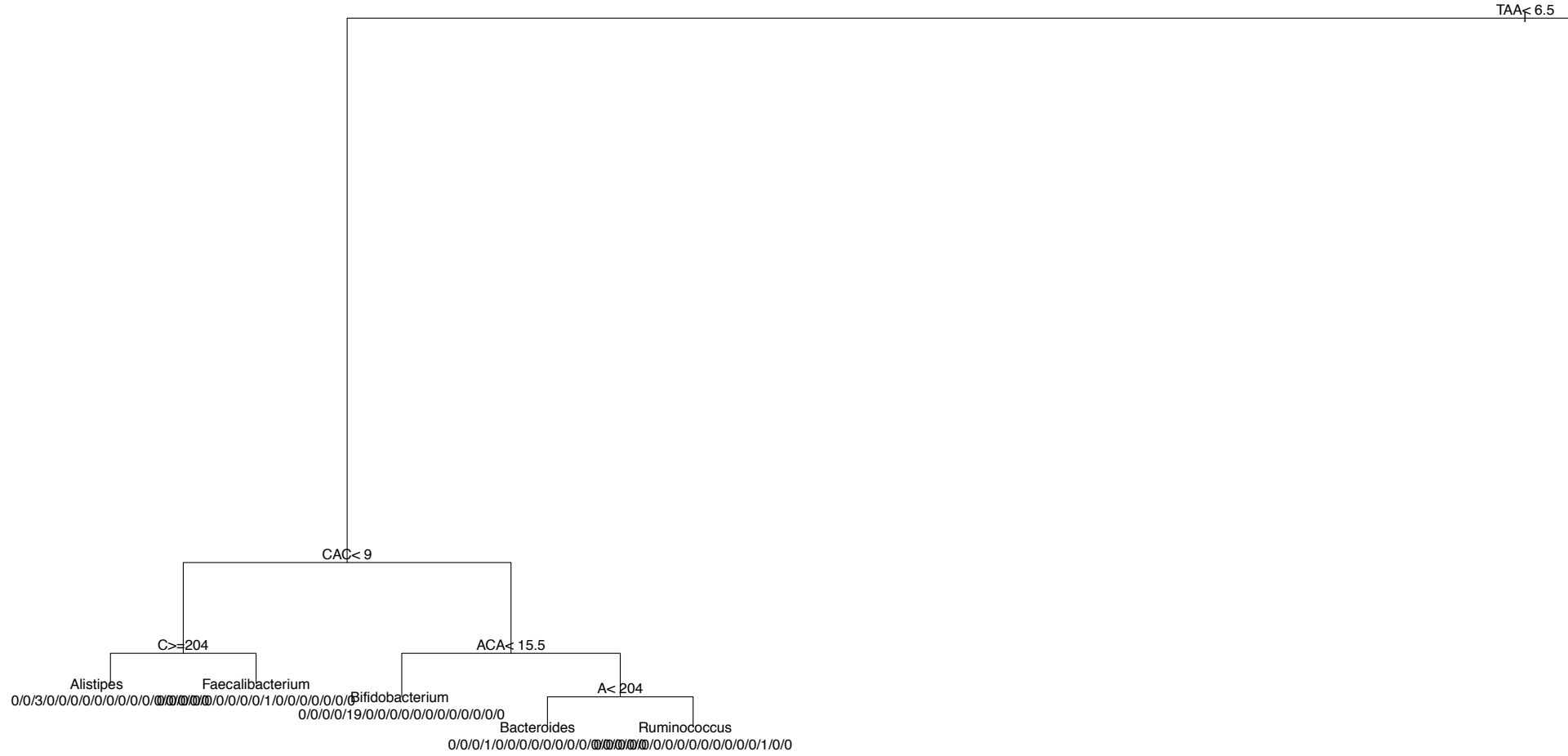
- Microorg. Databases + MetaSim 454-read simulations (Richter et al. PLoS ONE (2008), 3(10): e3373):
- Reads (150,000) per genera (17):
  - 245 Acidaminococcus, 256 Akkermansia, 2406 Alistipes, 40981 Bacteroides, 31404 Bifidobacterium, 1277 Coprococcus, 331 Eggerthella, 29118 Escherichia, 2378 Eubacterium, 4468 Faecalibacterium, 220 Megasphaera, 427 Parabacteroides, 1693 Prevotella, 4401 Roseburia, 2731 Ruminococcus, 4340 Shigella, 23324 Streptococcus
- Explanatory variables:
  - 1-,2-,3-,4-mer frequencies,
  - GC ratio, GC skew.

Tree example

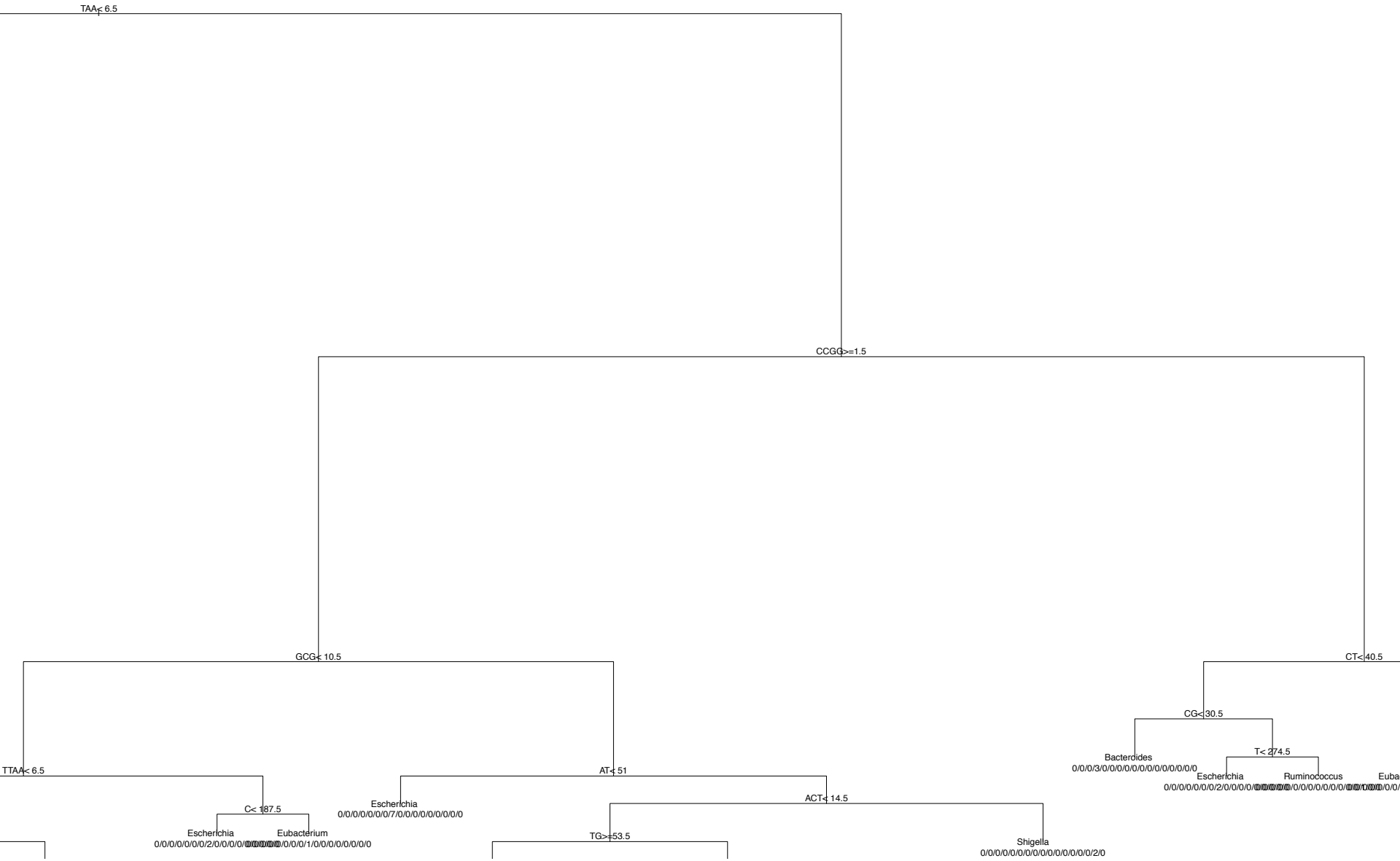
```
node), split, n, loss, yval, (yprob)
      * denotes terminal node

   1) root 100 71 Bacteroides (0 0 0.04 0.29 0.19 0.02 0 0.12 0.03 0.01 0.01 0.01 0 0.02 0.05 0.02 0.19)
     2) TAA< 6.5 25  6 Bifidobacterium (0 0 0.12 0.04 0.76 0 0 0 0 0.04 0 0 0 0 0.04 0 0)
       4) CAC< 9 4  1 Alistipes (0 0 0.75 0 0 0 0 0 0 0.25 0 0 0 0 0 0 0)
         8) C>=204 3  0 Alistipes (0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0) *
         9) C< 204 1  0 Faecalibacterium (0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0) *
       5) CAC>=9 21  2 Bifidobacterium (0 0 0 0.048 0.9 0 0 0 0 0 0 0 0 0 0.048 0 0)
        10) ACA< 15.5 19  0 Bifidobacterium (0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0) *
        11) ACA>=15.5 2  1 Bacteroides (0 0 0 0.5 0 0 0 0 0 0 0 0 0 0 0.5 0 0)
          22) A< 204 1  0 Bacteroides (0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0) *
          23) A>=204 1  0 Ruminococcus (0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0) *
     3) TAA>=6.5 75 47 Bacteroides (0 0 0.013 0.37 0 0.027 0 0.16 0.04 0 0.013 0.013 0 0.027 0.053 0.027 0.25)
       6) CCGG>=1.5 47 22 Bacteroides (0 0 0.021 0.53 0 0.043 0 0.21 0.043 0 0.021 0.021 0 0.043 0.021 0.043 0)
        12) GCG< 10.5 31  8 Bacteroides (0 0 0 0.74 0 0.032 0 0.065 0.032 0 0 0.032 0 0.065 0.032 0 0)
          24) TTAA< 6.5 28  5 Bacteroides (0 0 0 0.82 0 0.036 0 0 0 0 0 0.036 0 0.071 0.036 0 0)
            48) GGAT< 4.5 20  0 Bacteroides (0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0) *
            49) GGAT>=4.5 8  5 Bacteroides (0 0 0 0.37 0 0.12 0 0 0 0 0 0.12 0 0.25 0.12 0 0)
              98) A< 201 3  0 Bacteroides (0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0) *
              99) A>=201 5  3 Roseburia (0 0 0 0 0 0.2 0 0 0 0 0 0.2 0 0.4 0.2 0 0)
               198) ATA>=15.5 3  2 Coprococcus (0 0 0 0 0 0.33 0 0 0 0 0 0.33 0 0 0.33 0 0)
                 396) A>=218.5 2  1 Coprococcus (0 0 0 0 0 0.5 0 0 0 0 0 0.5 0 0 0 0 0)
                   792) A< 245.5 1  0 Coprococcus (0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0) *
```

...

Tree example



Tree example (cont.)



Results

- Performance:
  - Predicting over data not included in construction.
  - Accuracy estimation=77%
  - Example:

Pred/True	Acidaminococcus	Akkermansia	**Alistipes**	**Bacteroides**	Bifidobacter	Coprococcus
Alistipes	0%	0%	**100%**	0%	0%	0%
Bacteroides	0.34%	0.13%	0.76%	**69.14%**	1.62%	2.01%

- Classification-importance of each variable:
  - Increasing Accuracy: GCG, GC, CGC, %GC, TTA, CG, TA, CCGG, TAA, AAA, TTT, CGG, TTTA, CCG, TAAA, AA, CGA, GCGC, TTAA, TT.
  - Decreasing Gini Impurity: %GC, GCG, CGC, GC, CCGG, CG, TA, TTA, CGG, TAA, CCG, TTT, GCGC, AAA, TTTA, TAAA, TT, AA.

Conclusions

- Multiple trees approach show high accuracy.
- Improving estimations of genus concentrations:
  - Genus prediction on non-assembled or unknown species,
  - No needing blast.
- Next:
  - Measuring the impact on the total pipeline,
  - Focusing on target microorganisms.

www.assar-lab.cl

THANKS

GRAZIE

www.assar-lab.cl