Parco Tecnologico Padano

Entrepreneurial research in ag-biotech

Bioinformatics for metagenomics analysis

John Williams

Parco Tecnologico Padano

Human MicroBiome project

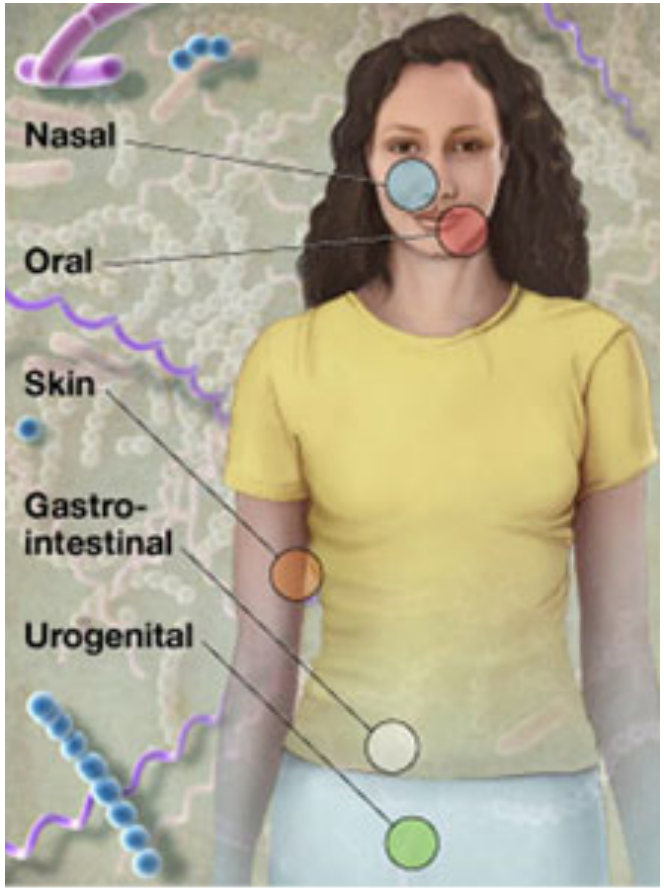
Parco Tecnologico Padano

Entrepreneurial research in ag-biotech

- Sampling 5 sites
- Reference Microbial genomes
- Resource Repositories

- New Computational technologies
- New Computational tools
- Data Analysis center
- Evaluation of data ve disease

- Legal Aspects



HMP

RUMINOMICS

MetaGenome analysis within Ruminomics

Parco Tecnologico Padano

Entrepreneurial research in ag-biotech

- Samples
  - Extreme 50 samples among 1000 cows from Italy, Sweden and UK.
  - from rumen swap between reindeer and cow
- Explore differences in
  - microbial composition
  - gene composition
- Impacts of a change in the microflora on
  - biological pathways
  - degradation of specific nutrient sources

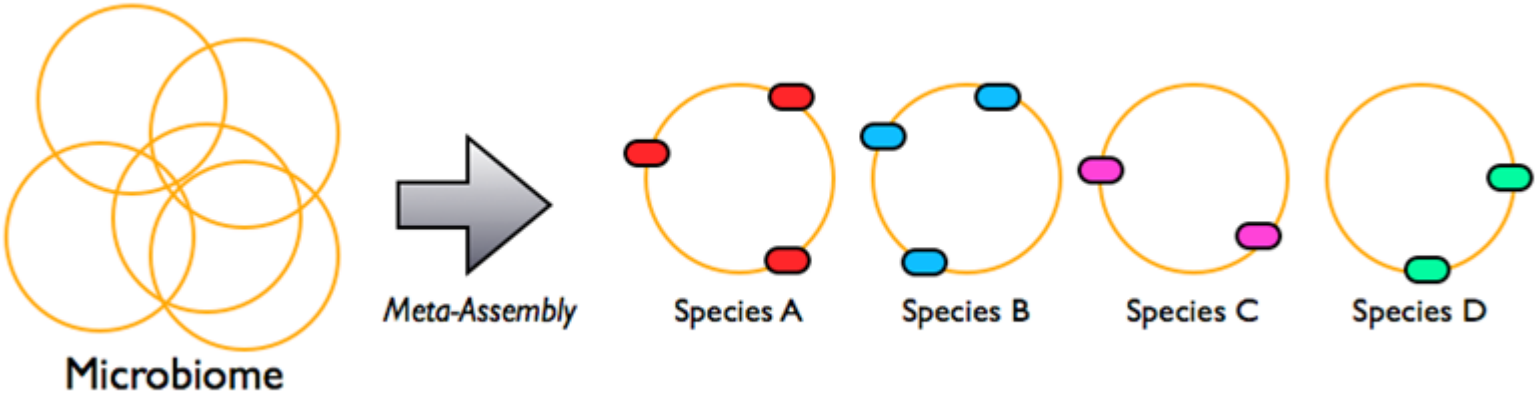
RUMINOMICS

Meta-biome challenge

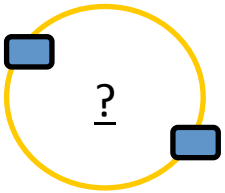
Parco Tecnologico Padano

Entrepreneurial research in ag-biotech

- Resolving micro-organisms present by
  - content
  - proportions
  - Identification of unknowns



- Low cost
- Repeatable

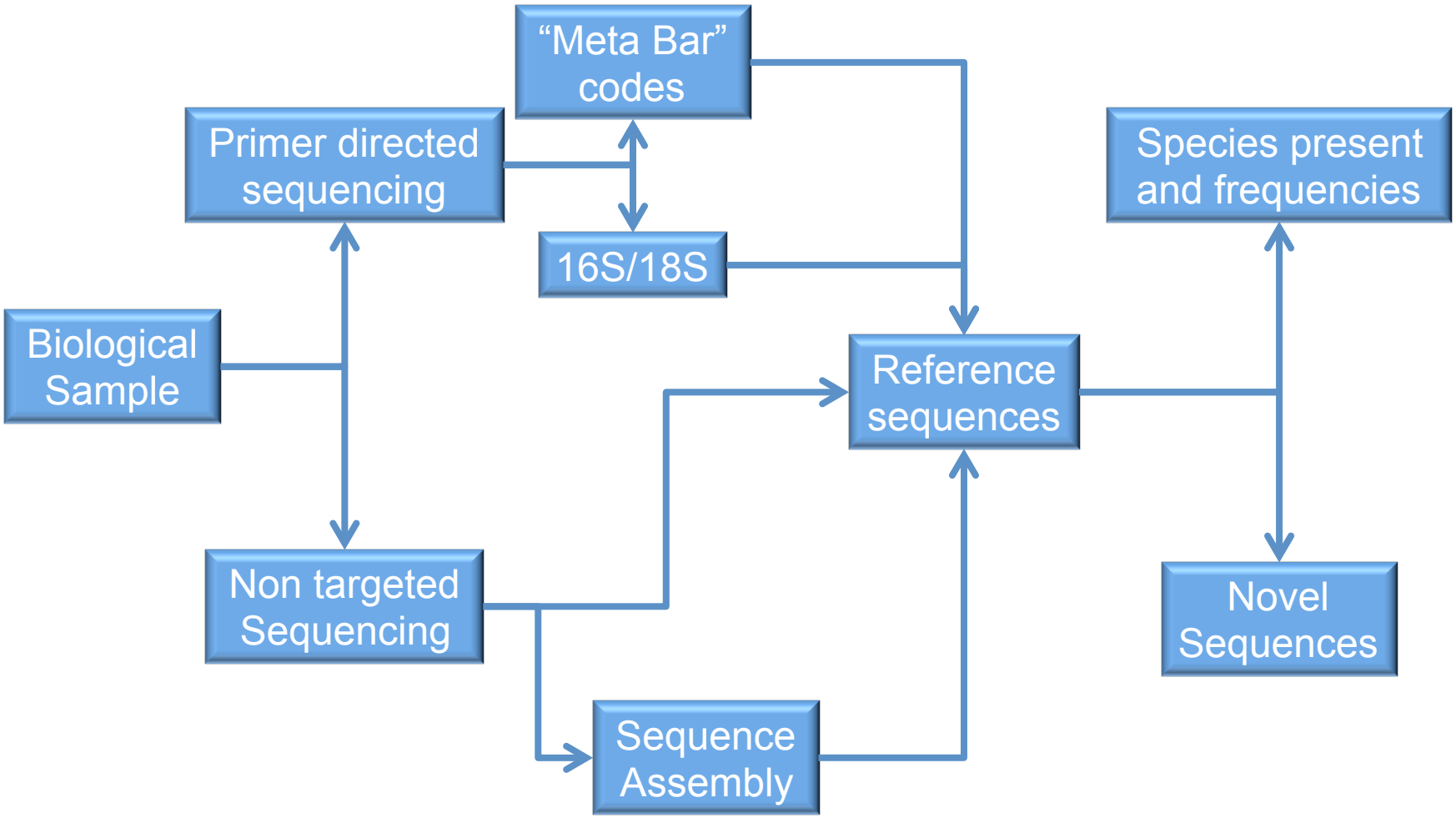


RUMINOMICS

Meta-biome analyses strategies

Parco Tecnologico Padano

Entrepreneurial research in ag-biotech



Considerations

- Culture based methods
  - May miss a large proportion of the species…
- Primer based methods
  - Bias as clades missed
  - Preferential amplification
- Sequence based methods
  - Large data sets
  - Complex analysis

Advantages of metagenomics

- Analysis of the whole consortium of microorganisms
  - Sequencing without targeted amplification
  - Determination of microbial composition and abundances
    - With or without a reference set of genomes
- Expression capabilitity
  - Mata-transcriptome
  - Meta-proteome
  - Impact of environment on the transcription
- sequence to gene and function

Sequence information

- Classical sequencing (Sanger)
  - *complete genomes*
  - annotation to locate genes, operons etc
  - contained data sets per species
- Metagenome sequencing
  - fragmented and incomplete sequences
  - reference required to identify individual or species
  - need to reconstruct genomes
  - large data sets
  - **computational challenges**

Challenges

- The rate of of genome sequencing has exceeded the bioinformatic capability to analyse the data
  - This is now the bottleneck
- Challenges
  - More complex than the assembly of single genomes
  - Variable represention of different species
  - Genetic diversity within strains
  - Homology among strains

PTP meta-genome pipeline

Parco Tecnologico Padano

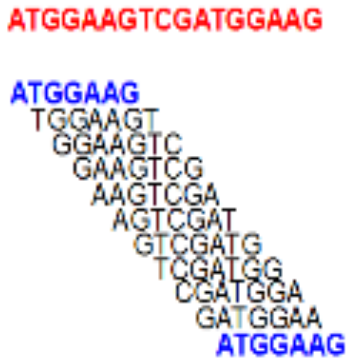
Entrepreneurial research in ag-biotech

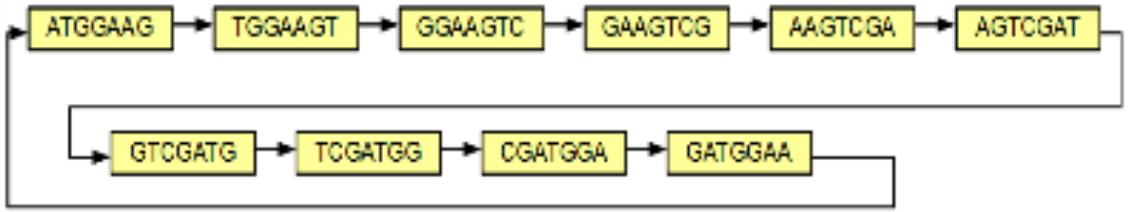
- Sampling
- Illumina sequencing
- Sequence QC
  - Removal of low quality sequence
  - Trimming of low quality ends
  - Removal of duplicates
- Assembly vs references
  - Choice of meta-assembler
  - Optimization of parameters
- Assessment of representation
  - standarization of sequence length vs sequence counts
  - Assigment and counting of assemblies
- Analysis of unassembled sequences

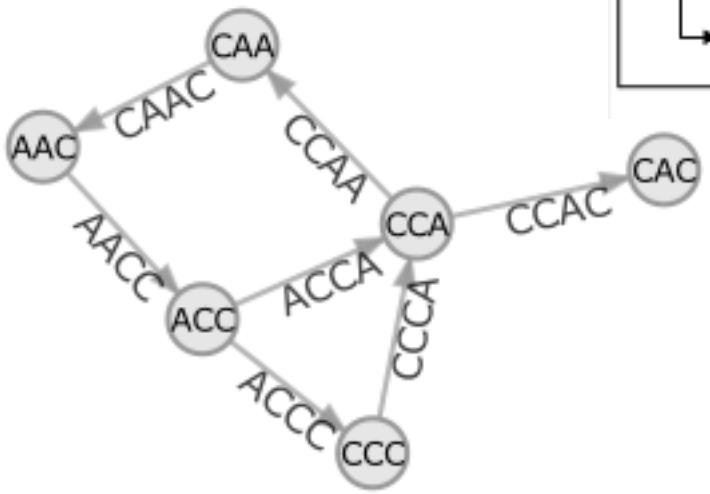
RuminOmics

Meta-assemblers single genomes

- Mostly based on Debruijn graphs
  - Split sequence into K-mers
  - Identify overlaps
  - Construct Debruijn graphs
  - Deduce contigs
  - Verify using pair-end data



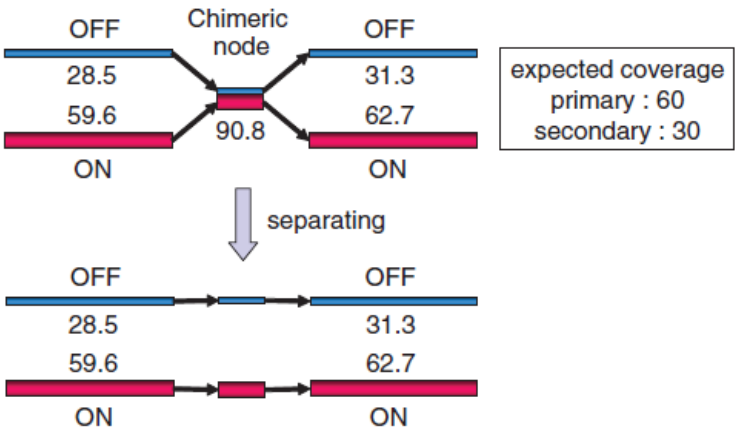




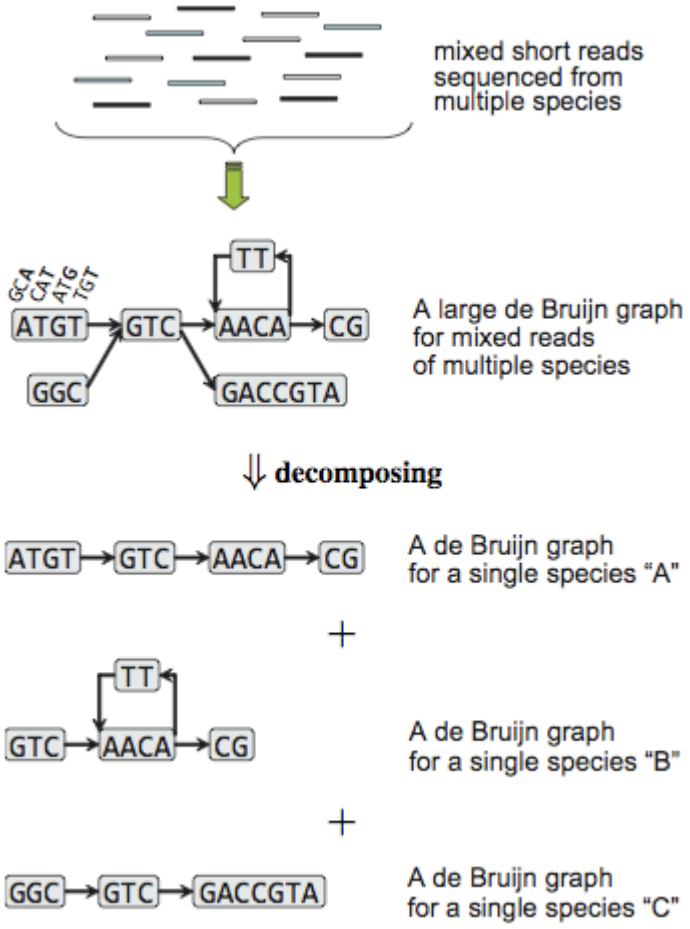
RUMINOMICS

Resolving the graphs, multiple species

- MetaVelvet
  - Velvet (single genomes)
  - To meta-genome analysis
  - Resolved Debruijn graphs
  - But on a single computer



expected coverage
primary : 60
secondary : 30



- Use of paired end data

RUMINOMICS

MetaRay vs MetaVelvet

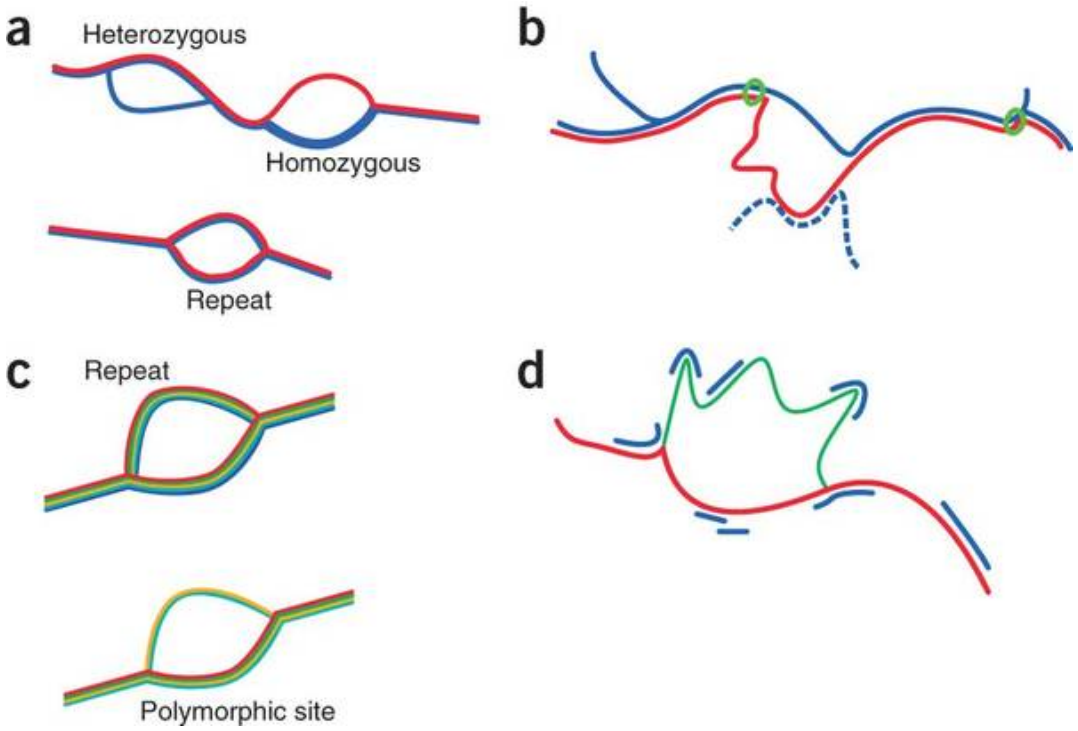
- Solutions to computing complexity
  - parallel array computing. eg ABySS ,RAY… for single genomes
- MetaRay
  - distributed array computing
    - increased data handling time
    - but decreased real time processing
  - uses reference genomes to assign K-mers
- Better results on the longer contigs
  - With contigs greater than 500bp

MataRay: Boisvert et al. Genome Biology 2012, 13:R122
http://genomebiology.com/2012/13/12/R122c

RUMINOMICS

MetaRay

- Construction of a de Bruijn graph from the raw sequence
- de novo assembly
- colouring with finished bacterial genomes

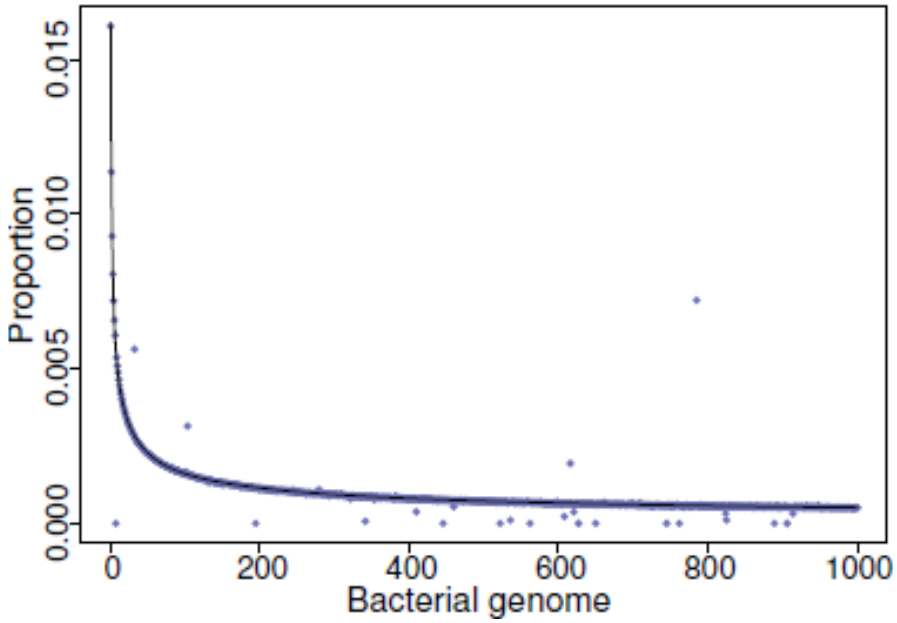


Iqbal et al. Nat. Genetics 2012

RUMINOMICS

Simulated performance

- Data set
  - 1000 bacterial genomes
  - 1% human
  - 3 X 10^9 (Illumina HighSeq Flow Cell)
- Errors
  - 2.6K miss-asemblies (1.3%)
  - 4 genomes overestimated
  - 20 underestimated
- Analysis
  - 15 hours
  - 1,024 cores 1.5GB/core
- Output
  - 974K contigs,
  - N50 76K
  - 974 and 2,894K per genome



Boisvert et al. Genome Biology 2012, 13:R122
http://genomebiology.com/2012/13/12/R122

RUMINOMICS

Ruminomics data PTP GenHome

Parco Tecnologico Padano

Entrepreneurial research in ag-biotech

- Data
  - 1 sample
  - HighSeq 32Gb

	MetaRay (k = 31)	MetaRay (k = 63)	MetaVelvet (k = 51)	MetaVelvet (k = 80)
# of contigs	3666701	113241	8871806	594313
N50	228	1069	382	839
Largest contig	190086	57744	51187	43580

- Analysis
  - MetaRay
  - 512 Gb 48 cores
  - memory usage 200Gb
  - 12 hours

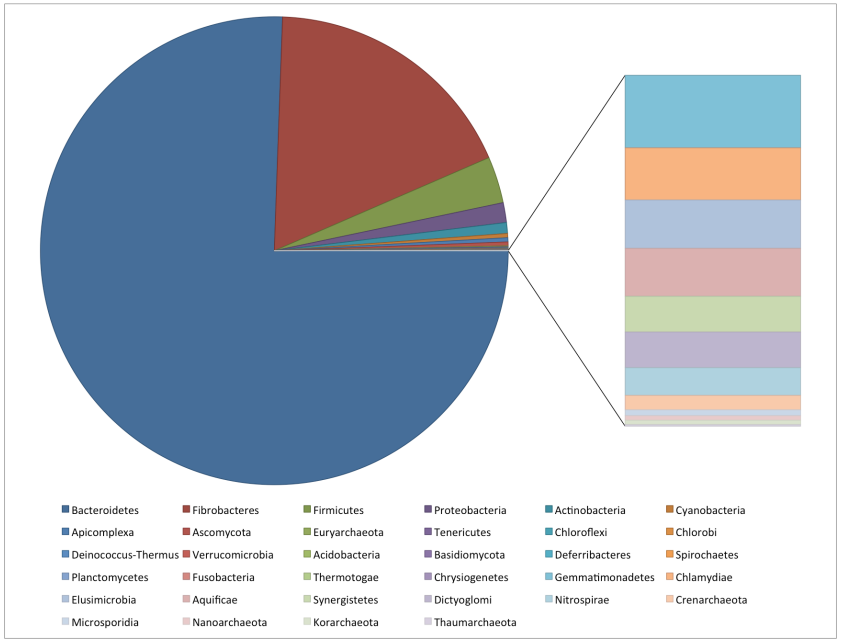
RUMINOMICS

Taxonomic results

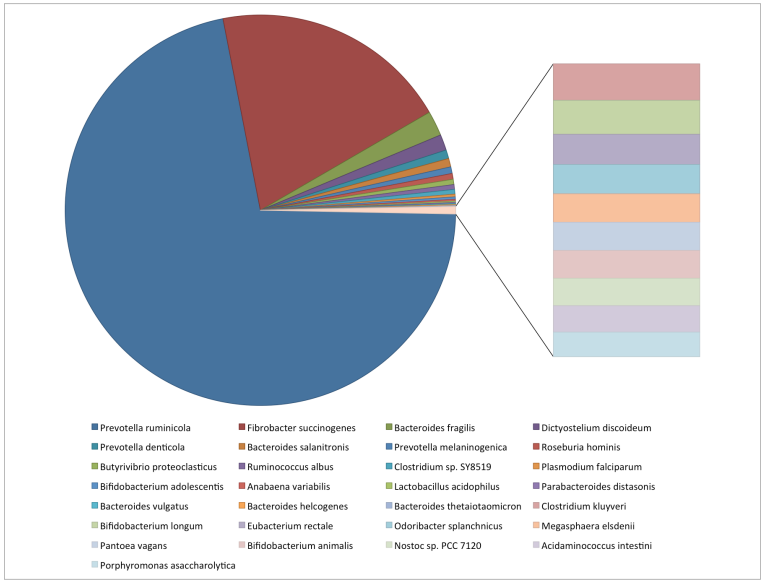
Parco Tecnologico Padano

Entrepreneurial research in ag-biotech

Phylum



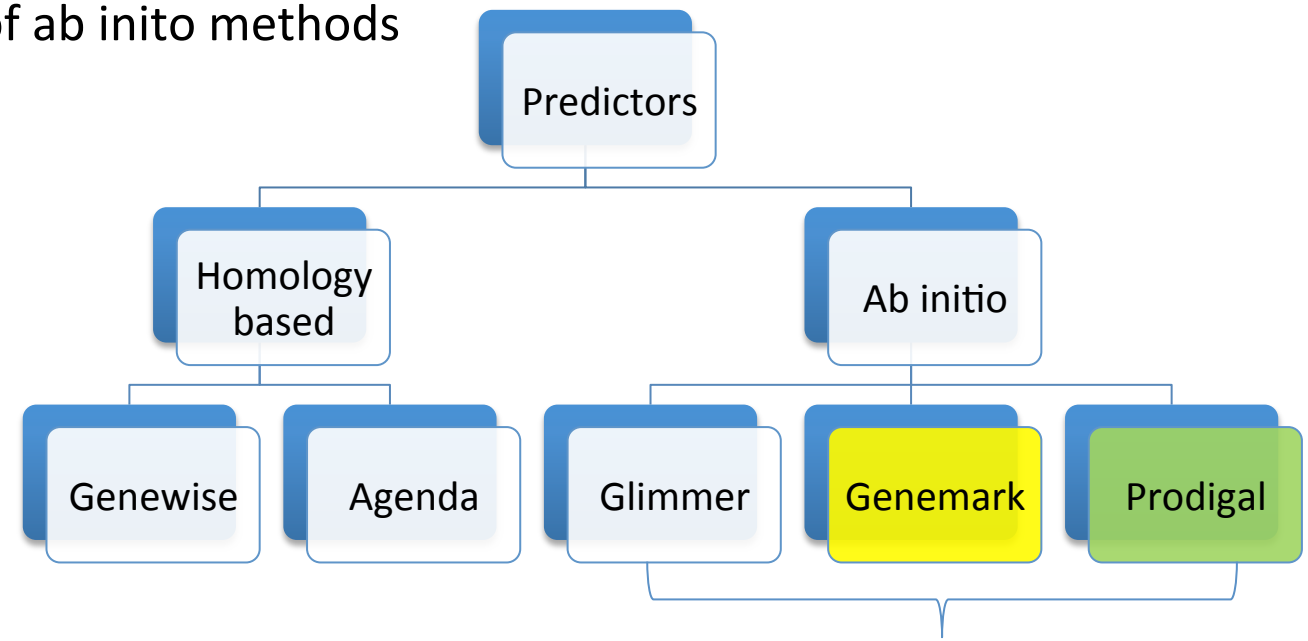
Species



RUMINOMICS

Gene prediction

- When assemble meta-genomics data will have :
  - high fraction of unknown sequences
  - Use of ab inito methods

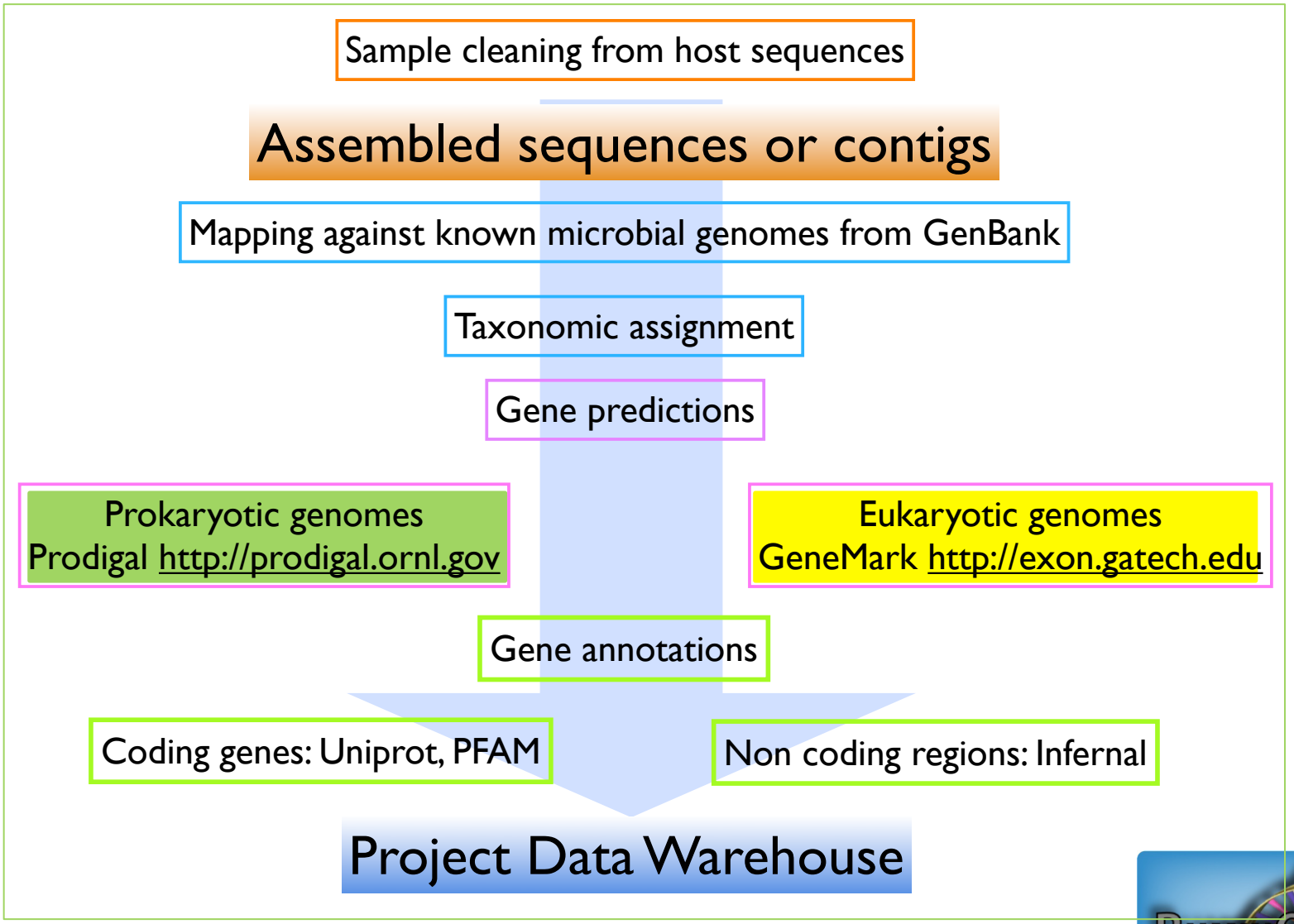


Tools within the NCBI annotation pipeline

From sequence to genes

Parco Tecnologico Padano

Entrepreneurial research in ag-biotech



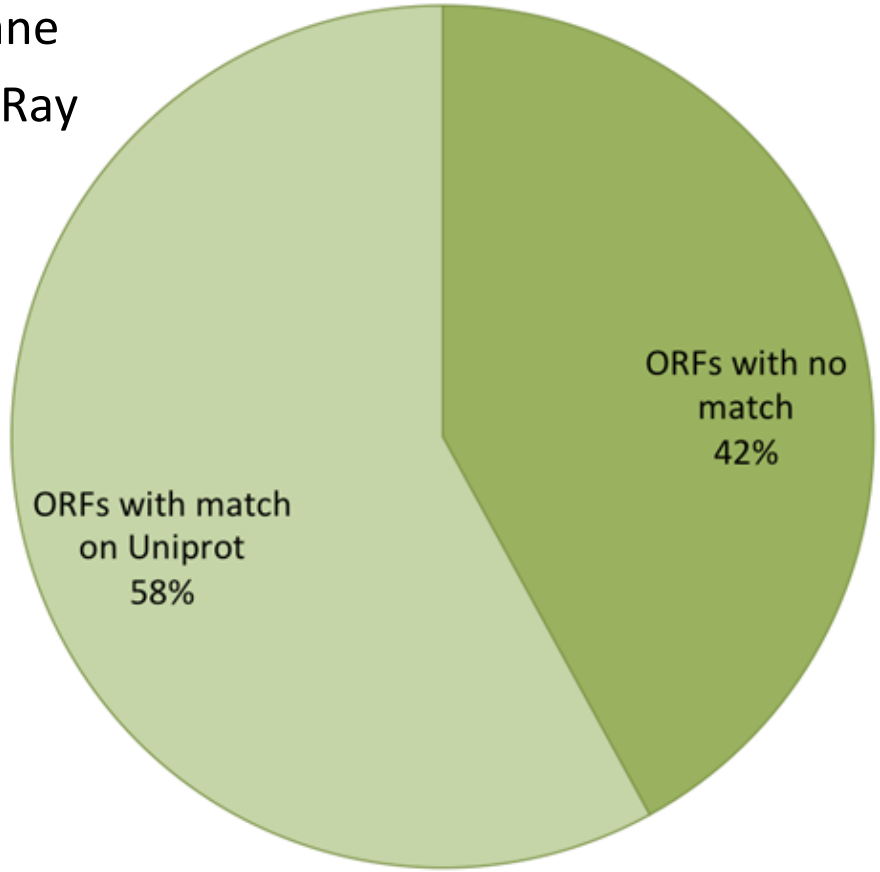
RUMINOMICS

First Ruminomics Meta-data

Parco Tecnologico Padano

Entrepreneurial research in ag-biotech

- Total ORFs predicted : 206,054
  - 1 sample on 1 HiSeq lane
  - assembled with Meta-Ray



- Contigs homology
  - BlastN on Uniprot
  - E-value cut-off 1e-10

RUMINOMICS

Functional analyses

- Gene classification by
  - structure
  - function
  - Definition of differences among microbial species
- Coding vs non-coding genes
- Assessment of coding capacity
  - metabolic pathways
  - newly identify strains

Statistical analyses

- Correlation between functional genomics information and phenotypic information
- Model for analysis includes
  - Factors for environmental effects on phenotype
  - Covariances among complex phenotypes
  - Presence/absence of single species /phila
  - Groups defined by clustering
- Use non parametric methods to solve

In summary

- A “Ruminomics” pipeline is available
  - Assembly
  - Composition
  - Function
- Approaches are computationally demanding
  - Tested on a small data set
  - Yet to be tested on a large data set

RUMINOMICS

Acknowledgements

- Bioinformatics group
- Francesco Strozzi

- Statistical Genetics
- Alessandra Stella

- GenHome Unit
- Chiara Ferandi

Thank you !

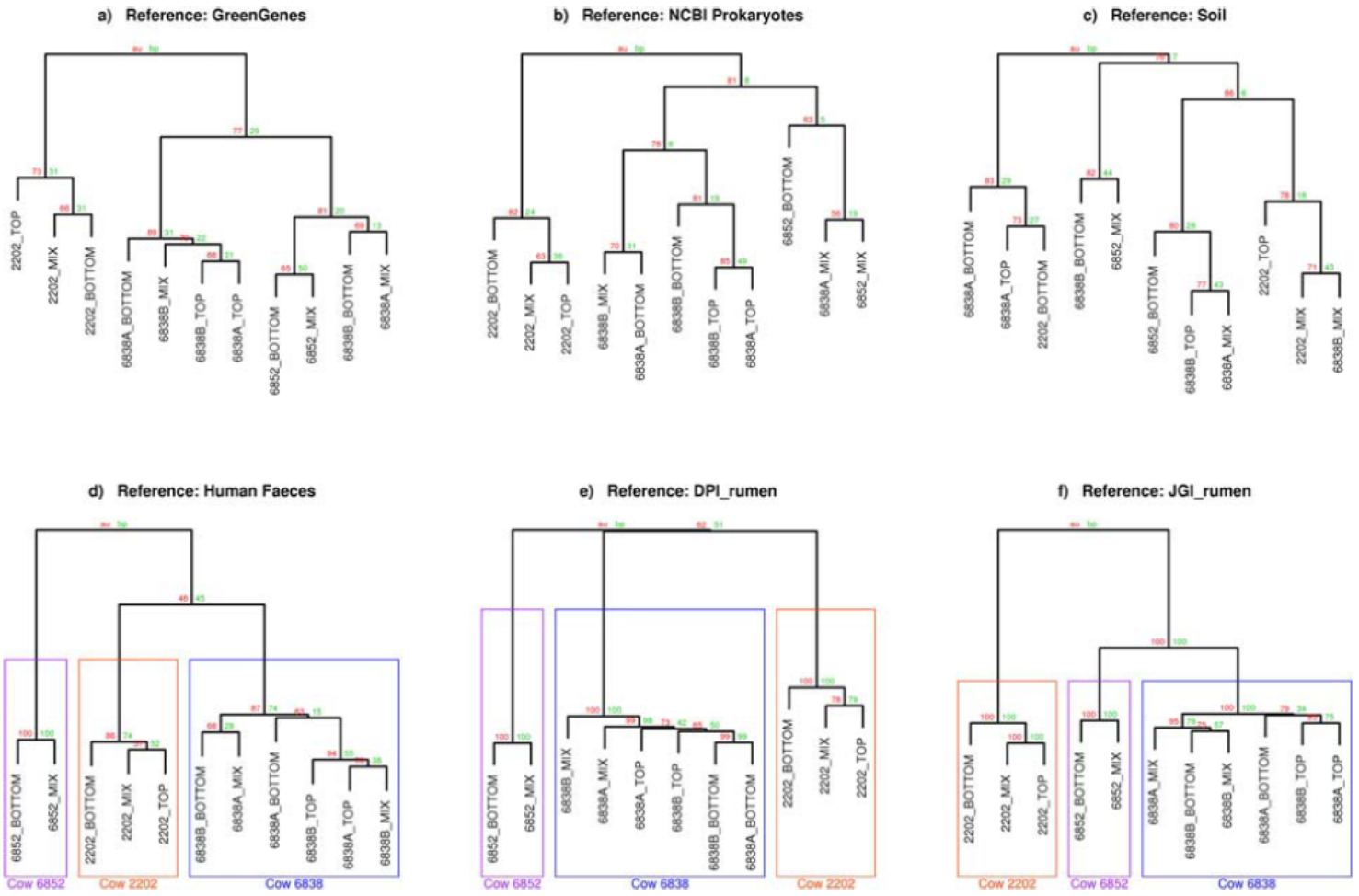
Parco Tecnologico Padano

Entrepreneurial research in ag-biotech

Impact of Reference Data

Parco Tecnologico Padano

Entrepreneurial research in ag-biotech



Ross et al. BMC Genetics 2012, 13:53
http://www.biomedcentral.com/1471-2156/13/53

RUMINOMICS