Advancing the Frontiers of Metagenomic Science

Daniel Falush, Wally Gilks,
Susan Holmes, David Kolsicki,
Christopher Quince,
Alexander Sczyrba, Daniel Huson

Open for Business
Isaac Newton Institute, Cambridge, UK
14 April 2014

“Mathematical, Statistical and Computational Aspects of the New Science of Metagenomics”

24 March – 17 April, 2014

Organisers

Wally Gilks	University of Leeds
Daniel Huson	University of Tübingen
Elisa Loza	National Health Service Blood Transfusion
Simon Tavaré	University of Cambridge
Gabriel Valiente	Technical University of Catalonia
Tandy Warnow	University of Illinois at Urbana-Champaign

Advisors

Vincent Moulton	University of East Anglia
Mihai Pop	University of Maryland

Agenda

Week 1: Workshop

Week 2: Forming research themes

Week 3: Developing research themes

Week 4: *Open for Business*

Consolidating collaborations

Research

Theme	*Convener*
• Taxonomic profiling	Daniel Falush
• Ecological modelling	Christopher Quince
• Functional modelling	Rodrigo Mendes
• Design and analysis	Susan Holmes
• Reference-free analysis	David Koslicki, Gabriel Valiente
• CAMI	Alice McHardy, Alexander Sczyrba
• Fourth domain	Wally Gilks

Taxonomic Profiling

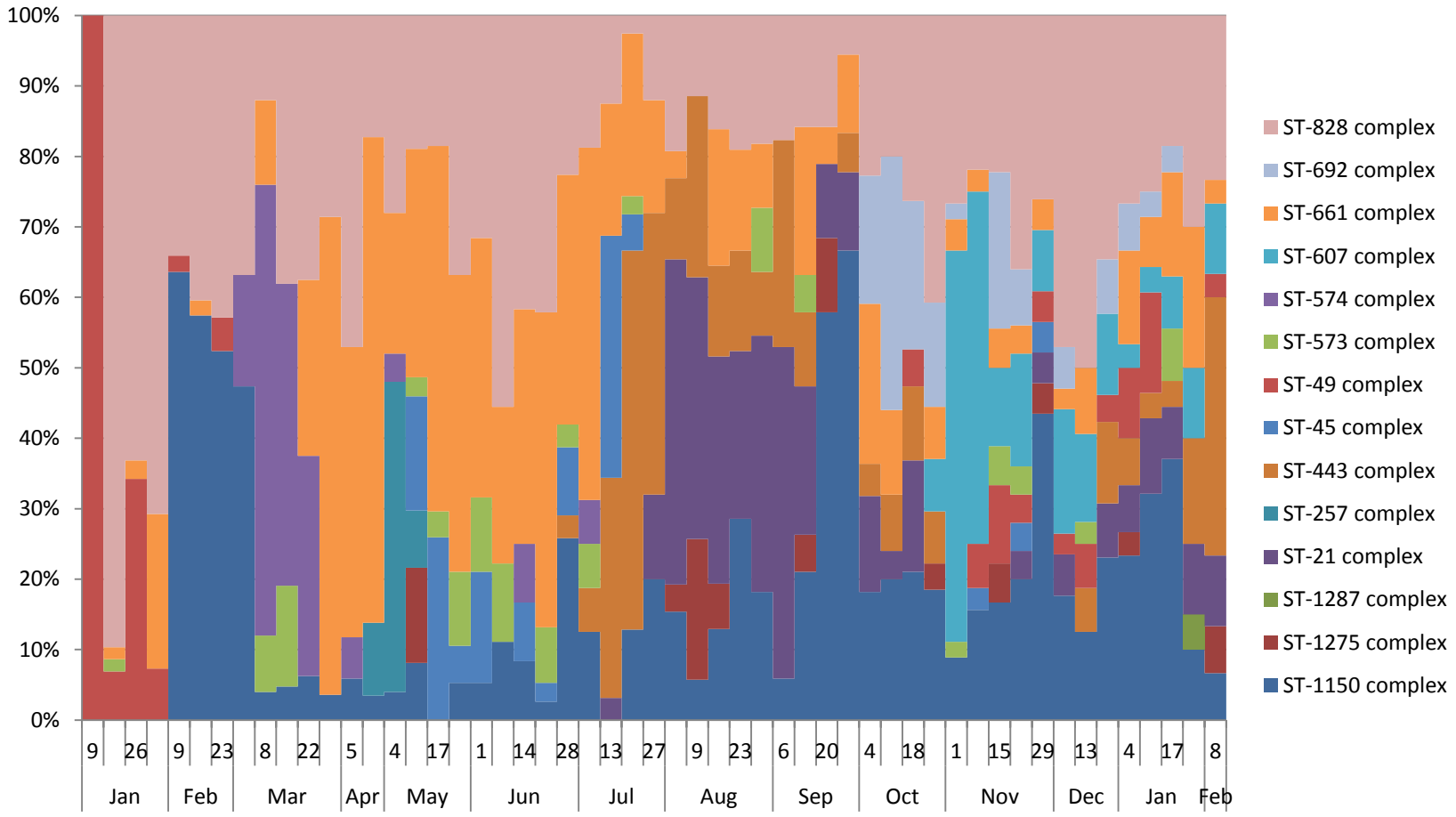
Presented by Daniel Falush

Max-Planck Institute for Evolutionary Anthropology

Strain level profiling of metagenomic communities using chromosome painting

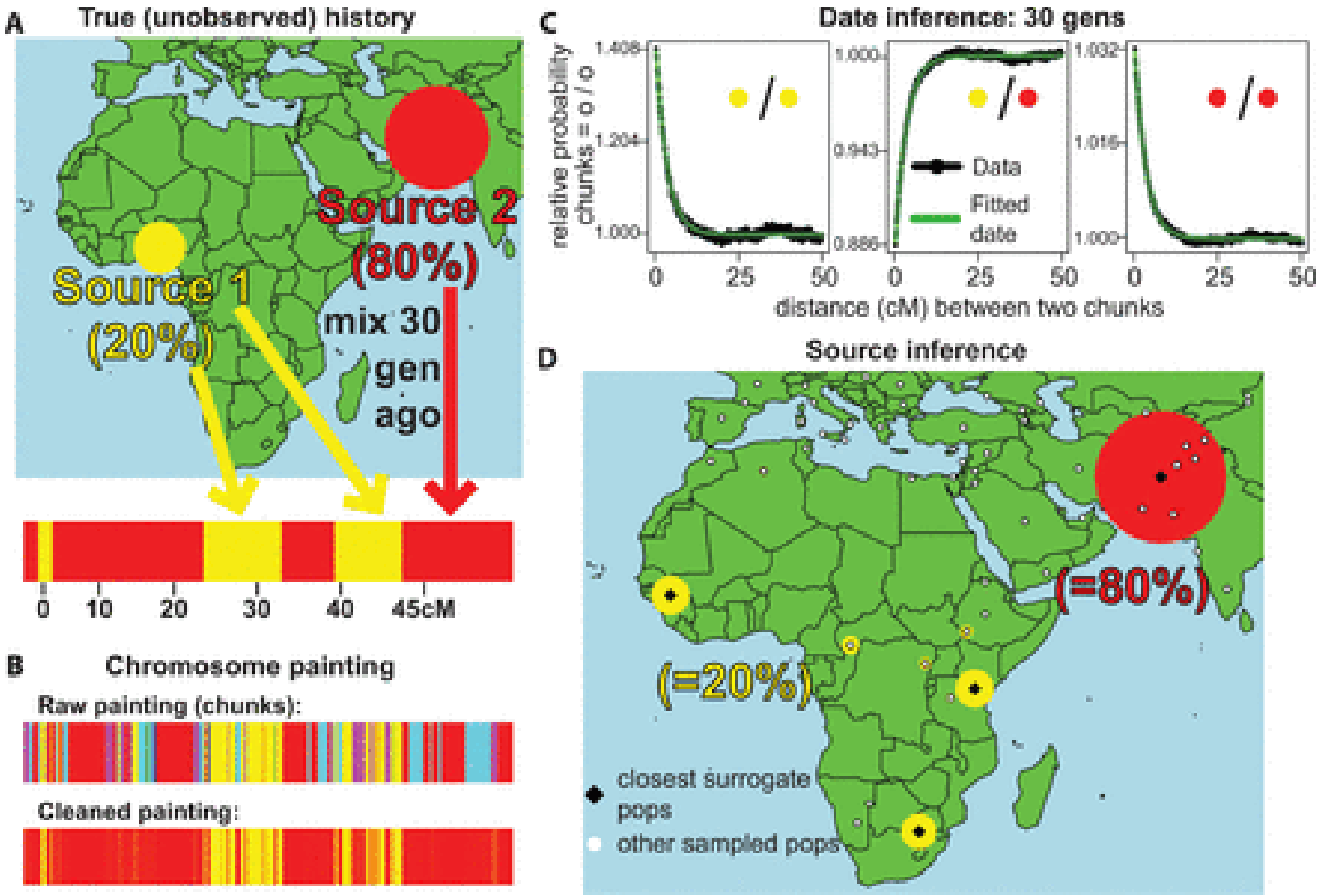
David Kosliki,
Nam Nguyen
Daniel Alemany
Daniel Falush

Strain level variation tells its own story Campylobacter Clonal complexes isolated from a broiler breeder flock over time



Colles *et al,* Unpublished

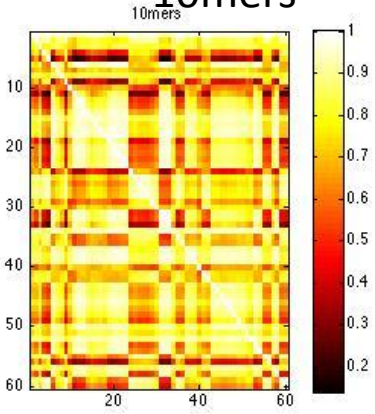
Chromosome painting: powerful data reduction and modelling technique from human genetics



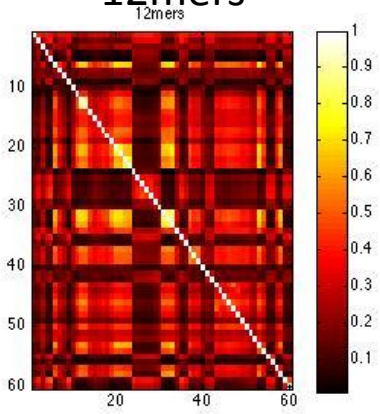
Chromopainter/FineSTRUCTURE/Globetrotter

Painting bacterial genomes based on Kmers of different lengths

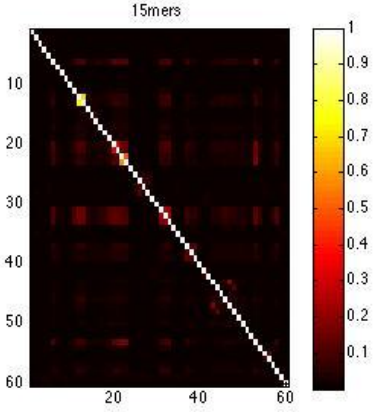
10mers



12mers



15mers



Mixture modelling

Basic Idea:

Count occurrence of common kmers

Add columns to *A* corresponding to internal nodes.

$$Ax = y$$

$$A_{i,j} = \sum_{k \in \text{kmers}(\text{seq}_i, \text{seq}_j)} \text{occ}_k(\text{seq}_j)$$

$$y_i = \sum_{k \in \text{kmers}(\text{seq}_i, \text{sample})} \text{occ}_k(\text{sample})$$

$$x_i = \text{freq. of seq}_i \text{ in sample}$$

$$\text{minimize } \| x \|_1^2 + \lambda \| Ax - y \|_2^2$$

Our approach

- Uses a large fraction of the information in the data
- Should work on wide variety of datasets, including 16S and metagenomes.
- Should provide strain resolution when the data supports it or classify at species or genus level when it does not.

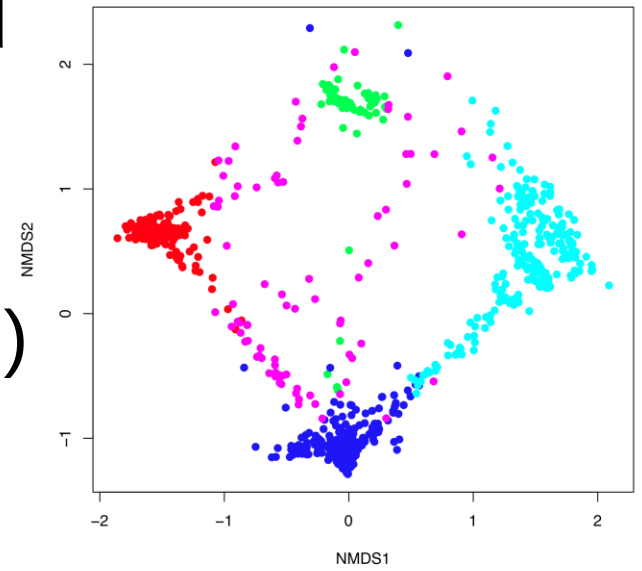
Ecological Modelling

Presented by Christopher Quince

University of Glasgow

Ecological Modelling

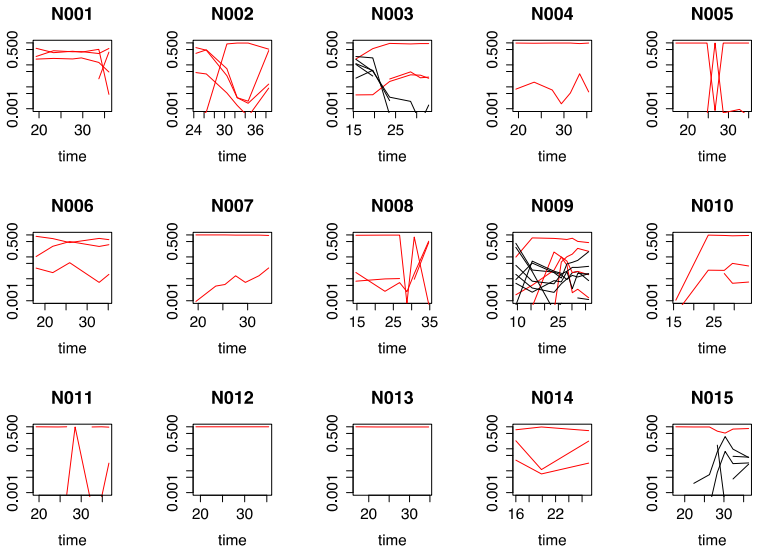
- Develop ecologically inspired approaches for modelling microbiomics data:
  - Mixture models (Daniel Falush)
  - Niche-neutral theory
  - Communities and phylogeny (Susan Holmes)
  - Analysis of vaginal microbiome time series data (Stephen Cornell)



Stephen Cornell

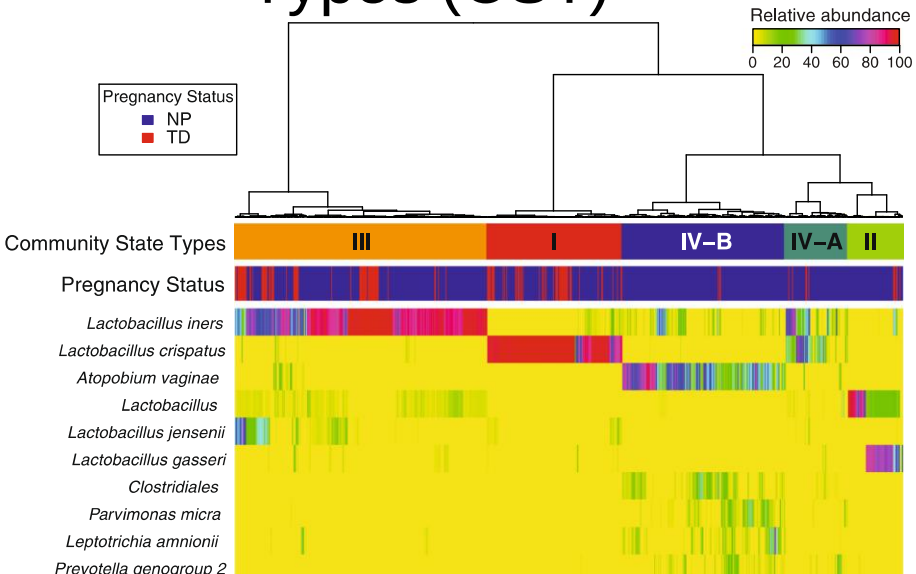
Data from Romera et al. *Microbiome* (2014)

Modelling dynamics of Vaginal Bacterial communities

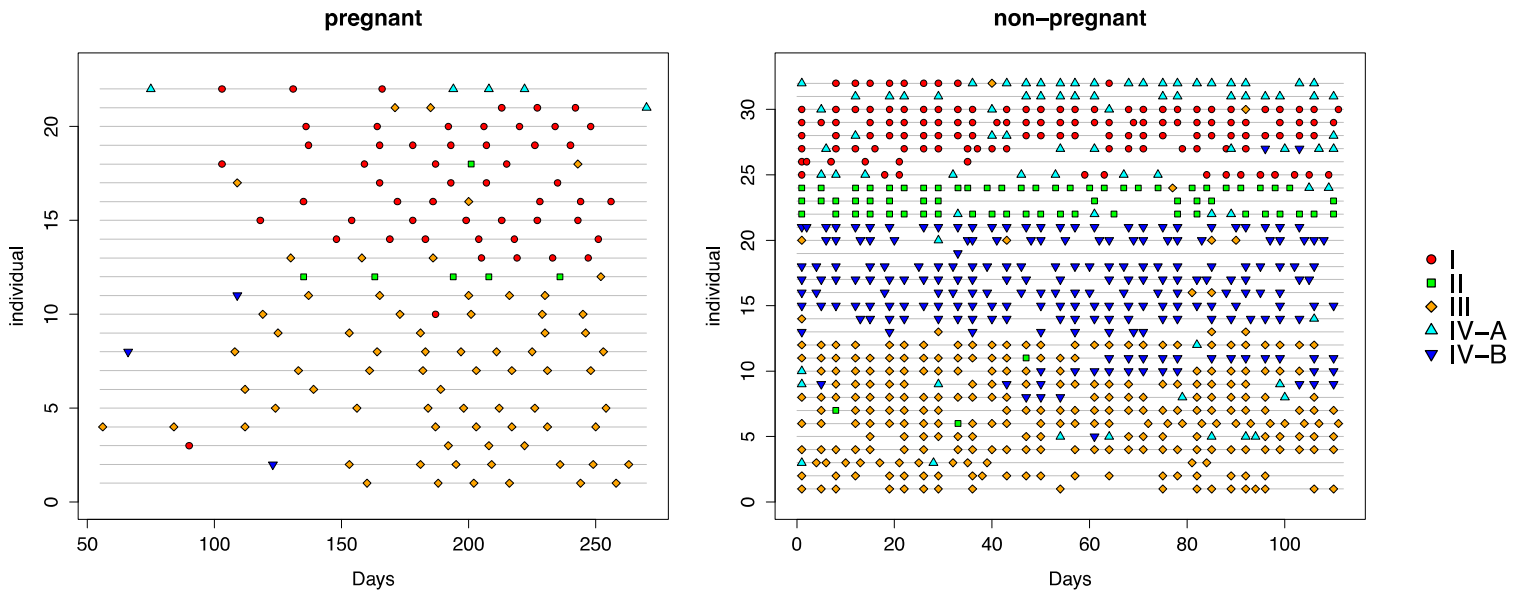


- How do the dynamics differ between 22 pregnant and 32 non-pregnant women?
- 143 bacterial species, strong fluctuations

- Simplified description: clustering by community relative abundances
  - identifies 5 Community State Types (CST)



Stephen Cornell



- Dynamic model (Markov process) accounts for differences in sampling frequency
- Underlying dynamics of CST differs between pregnant/non-pregnant
- Pregnant communities more stable (time constant: 143 days (pregnant) vs. 45 days (non-pregnant))
- Pregnant communities much less likely to switch to IV-A (a state correlated with bacterial vaginosis)
- Transition probability depends on both incumbent and invading CST
  - Invasion is not just a “lottery”

Design and Analysis

Presented by Susan Holmes

Stanford University

Challenges in Statistical Design and Analyses of Metagenomic Data

Susan Holmes

http://www-stat.stanford.edu/~susan/

Bio-X and Statistics, Stanford

Isaac Newton Institute Meeting

April,14, 2014

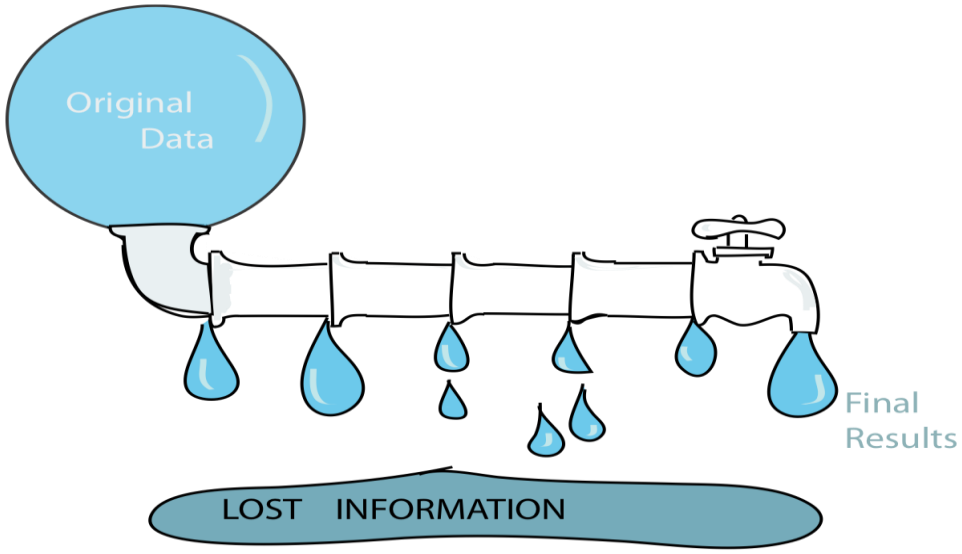
Challenges for the Design of Meta Genomic Data Experiments

- Heterogeneity.
- Lack of calibration.
- Iteration, multiplicity of choices.
- Graph or Tree integration.
- Reproducibility.
- Data Dredging of high throughput data.
- Statistical Validation (p-values?).

Heterogeneity

- Status : response/ explanatory.
- Hidden (latent)/measured.
- Different Types :
- Continuous
  - Binary, categorical
  - Graphs/ Trees
  - Images/Maps/ Spatial Information
- Amounts of dependency: independent/time series/spatial.
- Different technologies used (454, Illumina, MassSpec, RNA-seq, Images).
- Heteroscedasticiy (different numbers of reads, GC context, binding, lab/operator)..

Losing information and power



Statistical Sufficiency, data transformations.

Mixture Models.

Documentation and Record Keeping

P-values are overrated

- Many significant findings today are not reproducible (see JPA Ioannidis - 2005).
- Why?
- Data dredging?

P-values are overrated

JANUARY 6, 2011

When Science Goes Psychic

DEBATERS

No Sacred Mantle
LAWRENCE KRAUSS, THEORETICAL PHYSICIST

A Cutoff for Craziness
DOUGLAS HOFSTADTER, AUTHOR, "GÖDEL, ESCHER, BACH"

A Buyer-Beware Market
BEN GOLDACRE, AUTHOR, "BAD SCIENCE"

No Need to Panic
RICHARD WISEMAN, PROFESSOR OF PSYCHOLOGY

Addressing the Bias Problem
JONATHAN SCHOOLER, PROFESSOR OF PSYCHOLOGY

An Assault on Rationality
DAVID HELFAND, COLUMBIA ASTROPHYSICS LABORATORY

How Open Inquiry Works
STANLEY W. TRIMBLE, PROFESSOR OF GEOGRAPHY

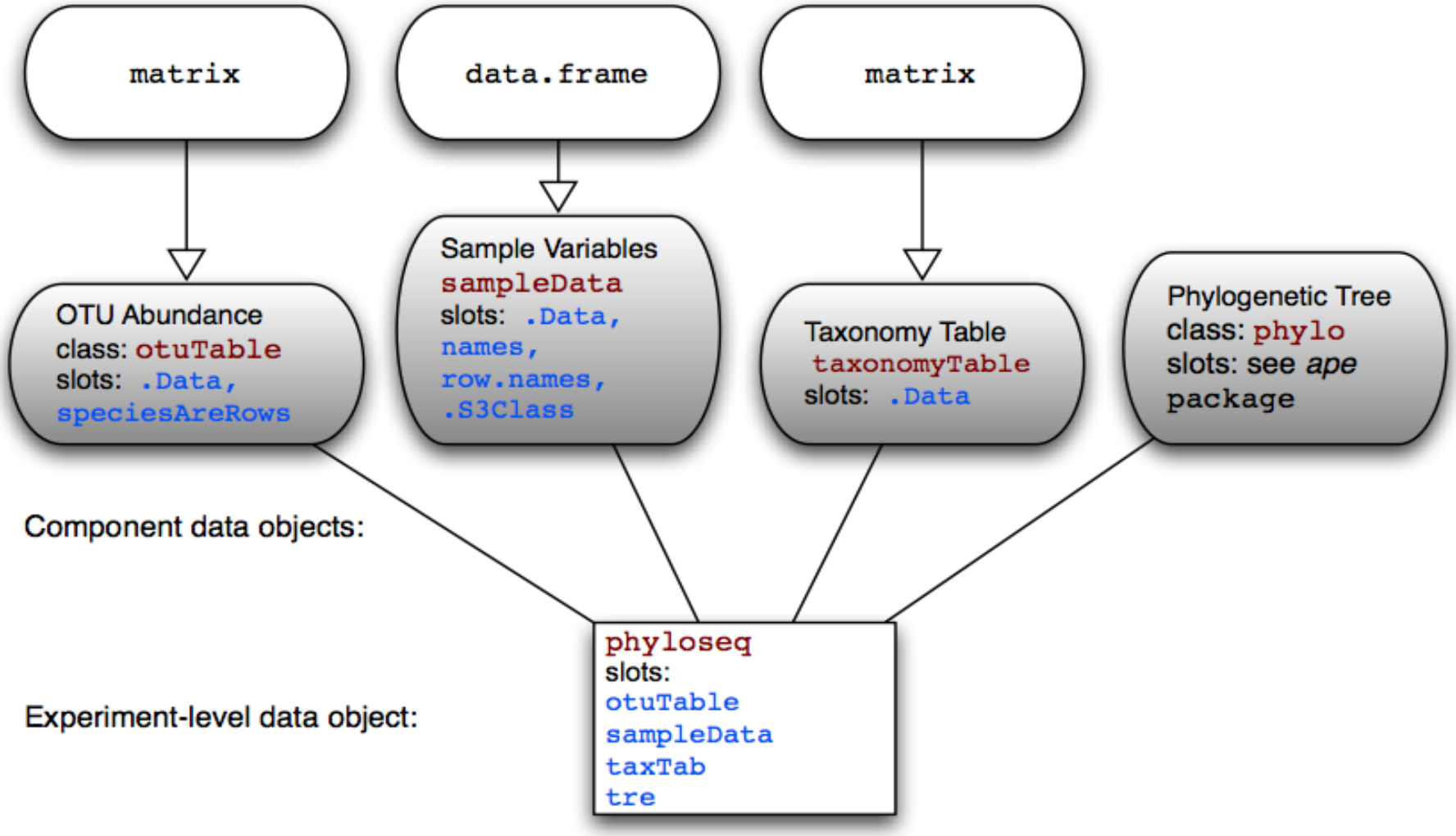
INTRODUCTION

A respected psychology journal has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events. Though the paper was peer reviewed, there are many critics who say the research is nonsense. A group of psychologists published a rebuttal paper in the same journal.

© Kts/Dreamstime.com

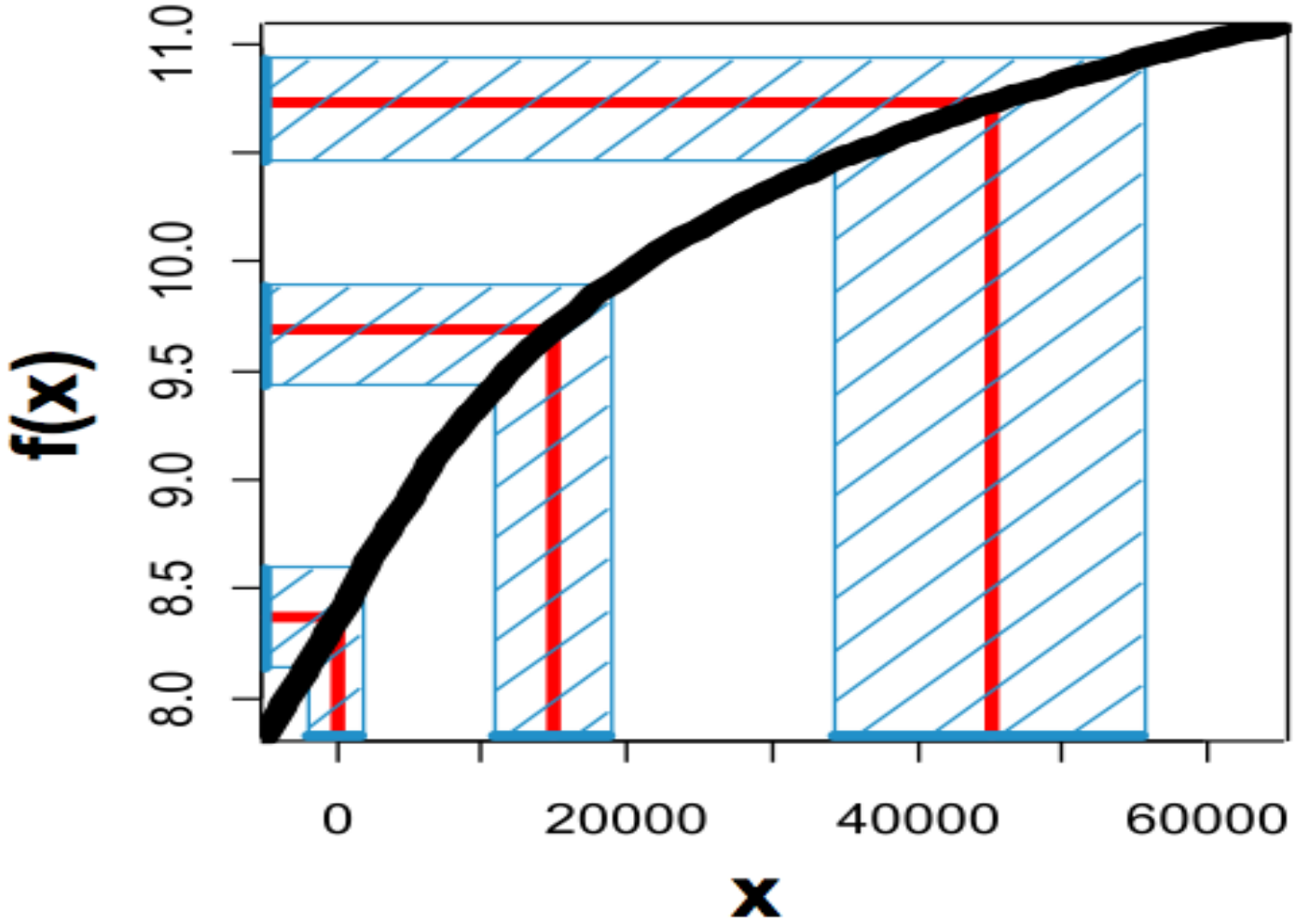
How does the peer review process ensure good quality research? Are there factors that the standard process cannot take account of? Or is ESP simply a claim that should not be entertained as a subject of scientific inquiry?

Keeping all the information



Normalization

- variance stabilizing transformation



Optimality Criteria Chosen at the time of the experiment's design

Optimality Criteria:

- Sensitivity or Power: True Positive Rate.
- Specificity: True Negative Rate.
- Detection of Rare variants

- We have to control for many sources of error (blocking, modeling, etc..)

- Use of available resources for depth, technical replicates or biological replicates?

Conclusions:

- Error structure, mixture models, noise decompositions.
- Power simulations.
- Data integration phyloseq, use all the data together.
- Reproducibility: open source standards, publication of source code and data. (R) knitr and RStudio.

Needed:

Better calibration, conservation of all the relevant information, ie number of reads, variability, quality control results.

Reference-free Analysis

Presented by David Koslicki

Oregon State University

Reference-free analysis

Zam Iqbal, David Koslicki, Gabriel Valie

What can be said about metagenomic samples in the absence of (good) references?

Global analysis: How diverse is the sample?

How does one sample differ from another?

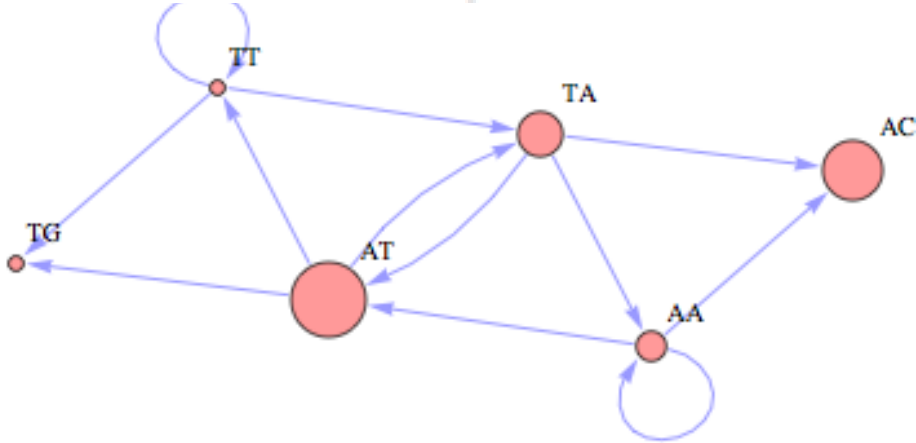
K-mer approach: Can multiple k-mer lengths be used to obtain a multi-scale view of a sample?

What is the "right" way to compare k-mer counts across samples?

Tools: Complexity function

$$P_S(k) = \#\,\text{unique}\; k-\text{mers in metagenome}\; S$$

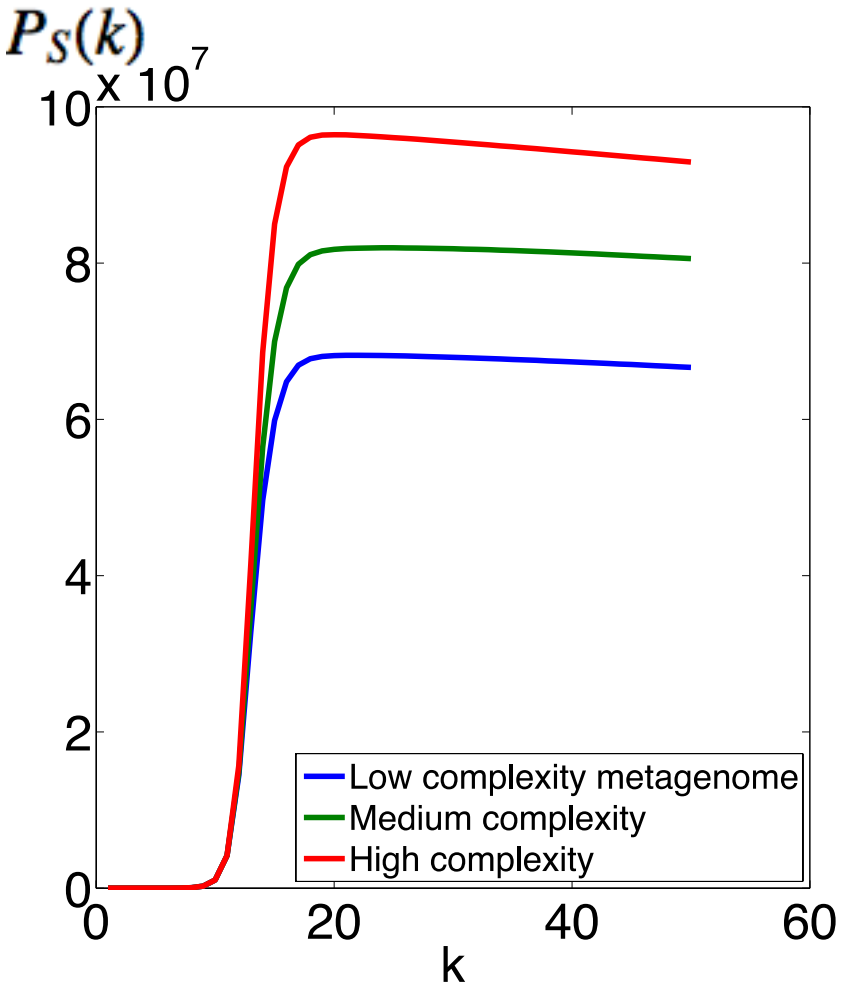
De Bruijn graph

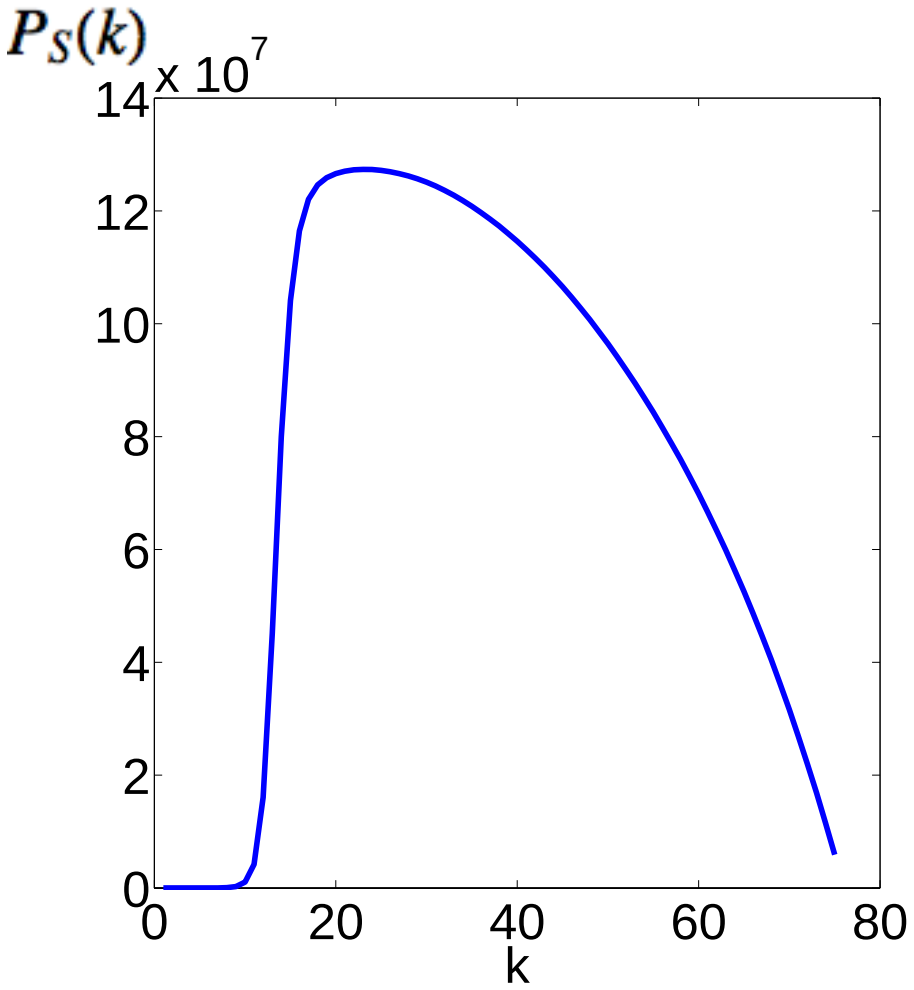


(K-mer) Size Matters

How diverse is the sample?

$$P_S(k) = \#\,\text{unique } k-\text{mers in metagenome } S$$

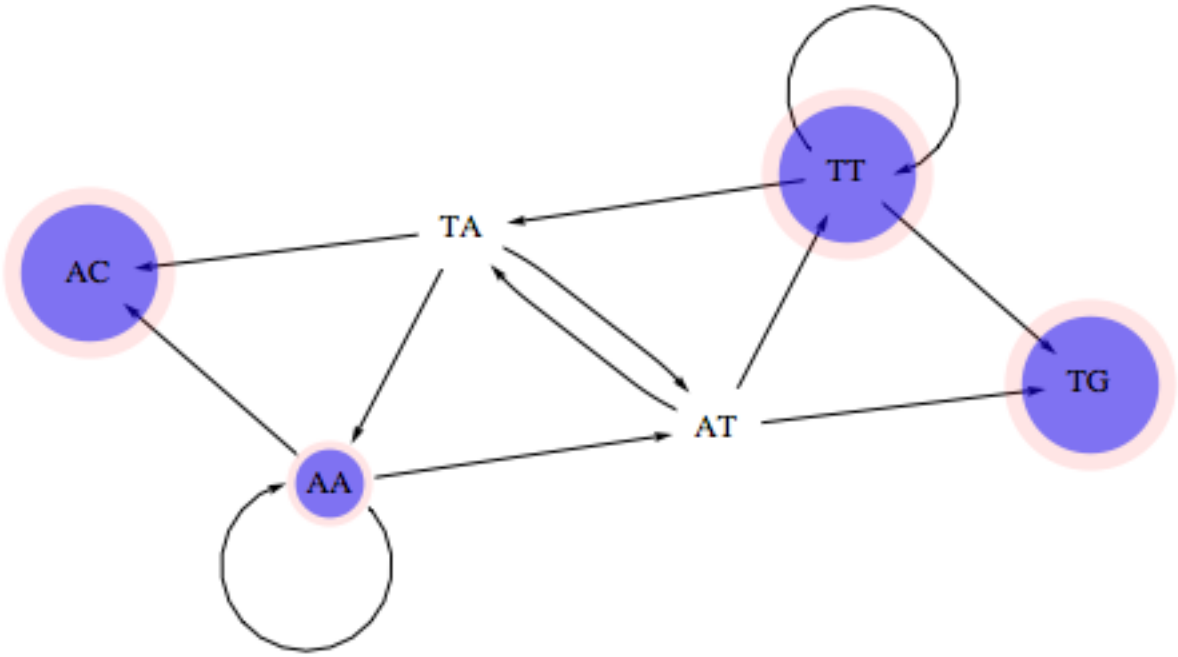




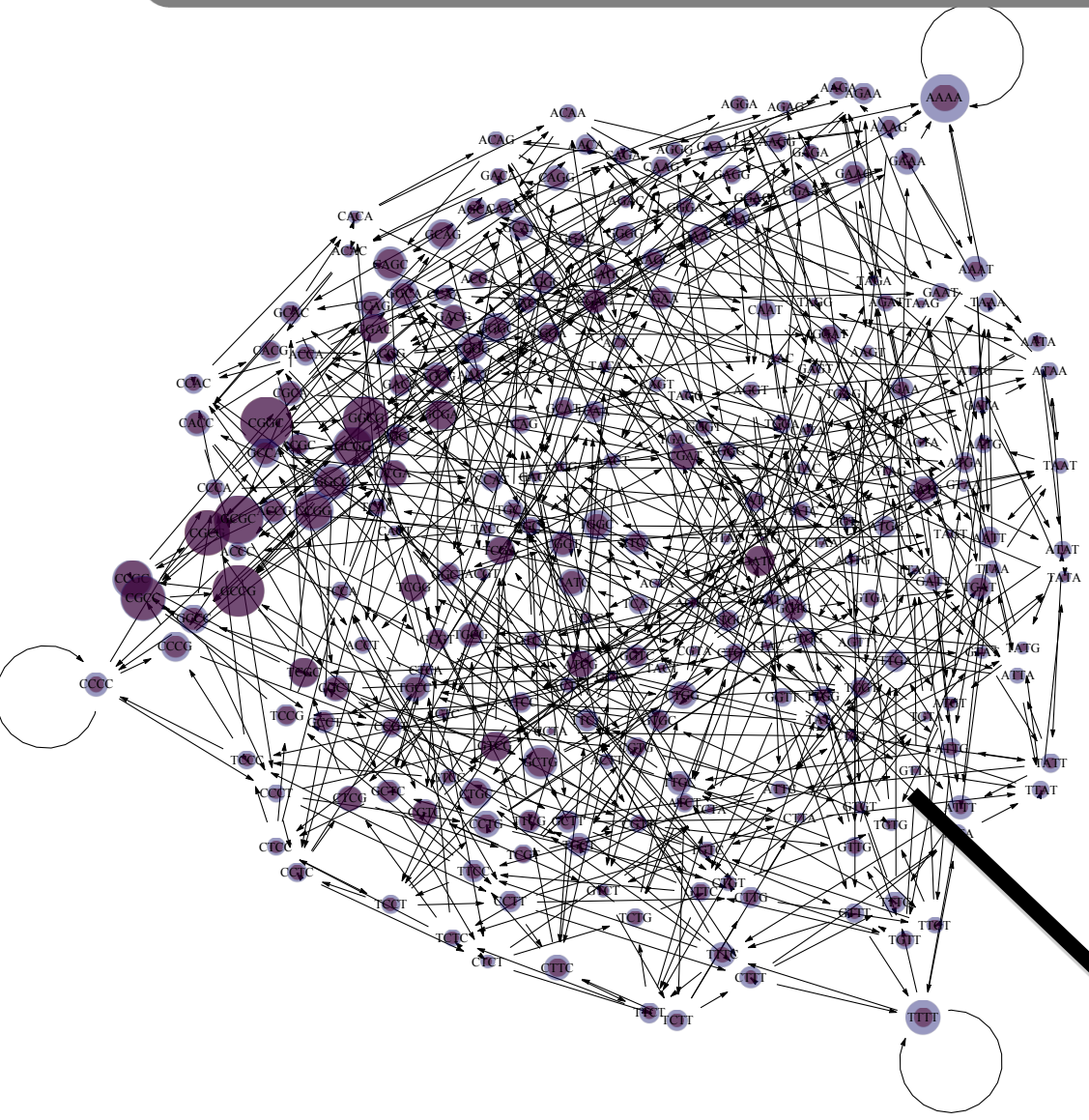
De Bruijn-based metrics

How does one sample differ from another?

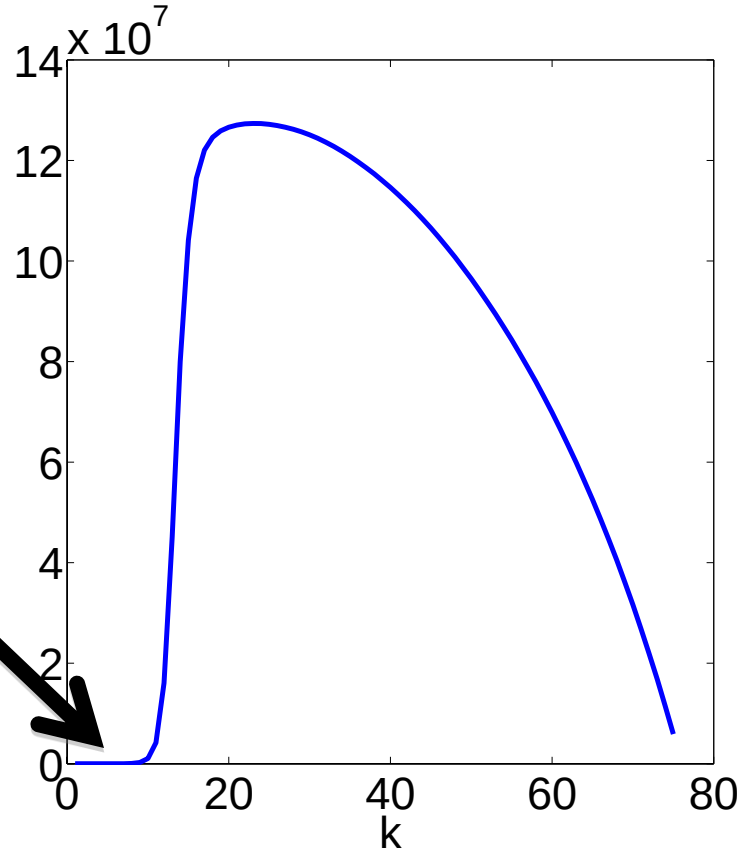
Keep track of *how much* mass needs to be moved *how far.*



De Bruijn-based metrics

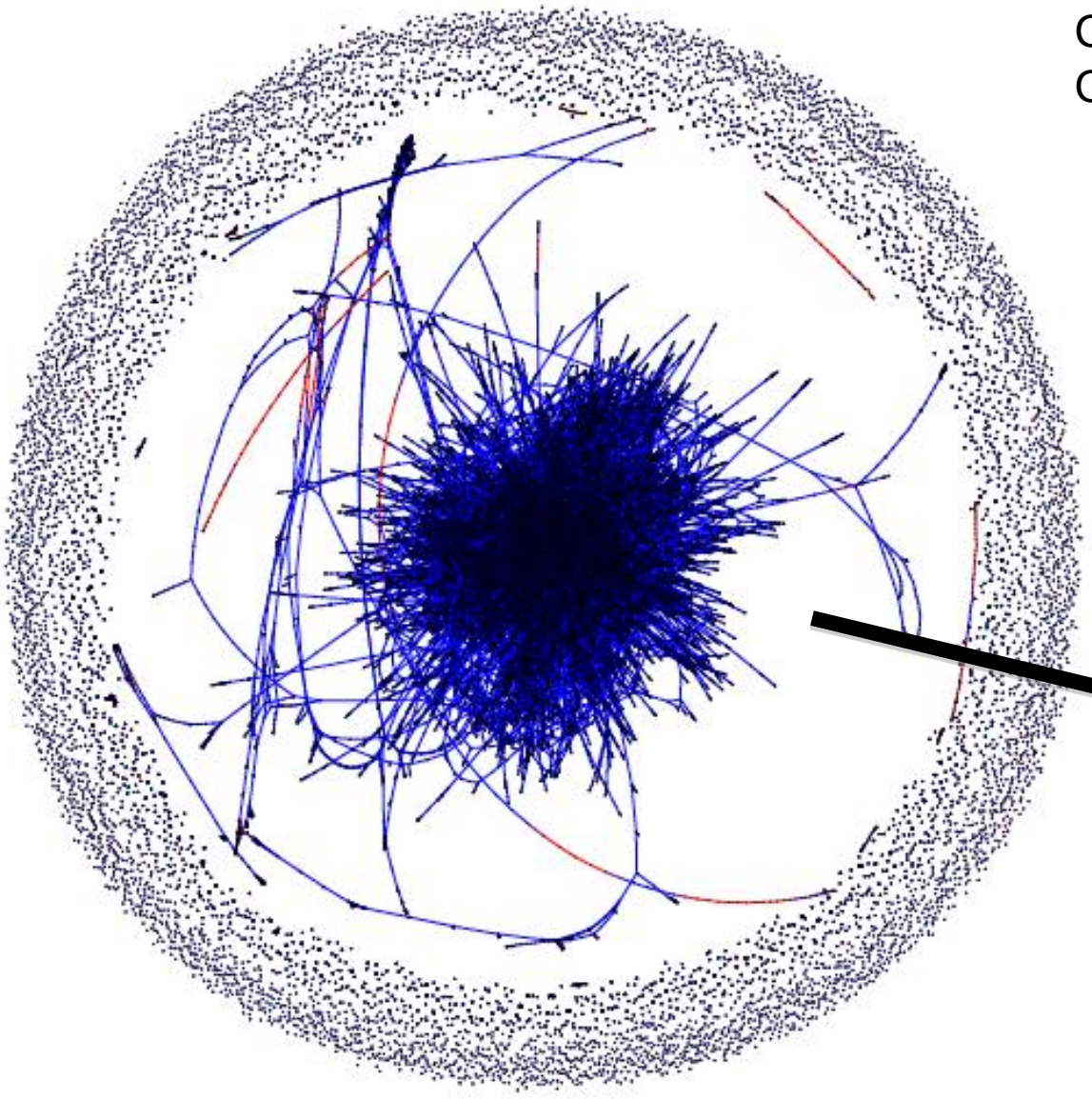


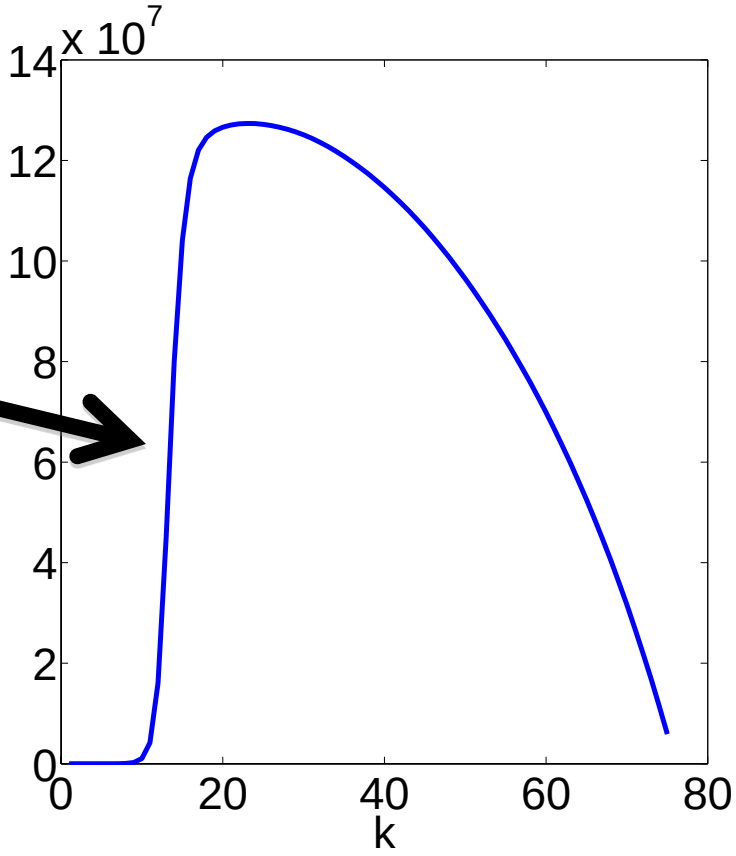
Connections to de Bruijn Graphs



De Bruijn-based metrics

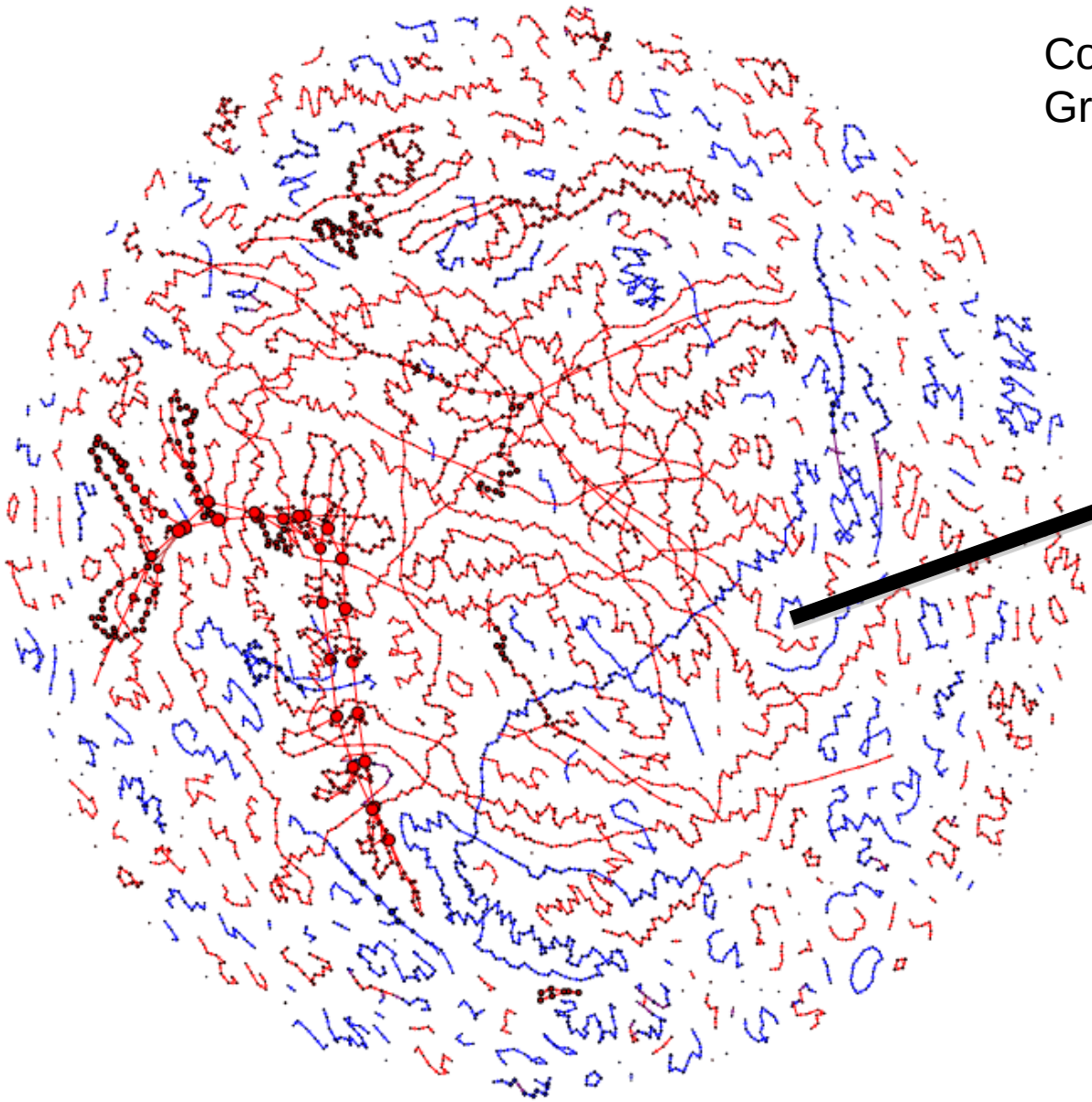
Connections to de Bruijn Graphs

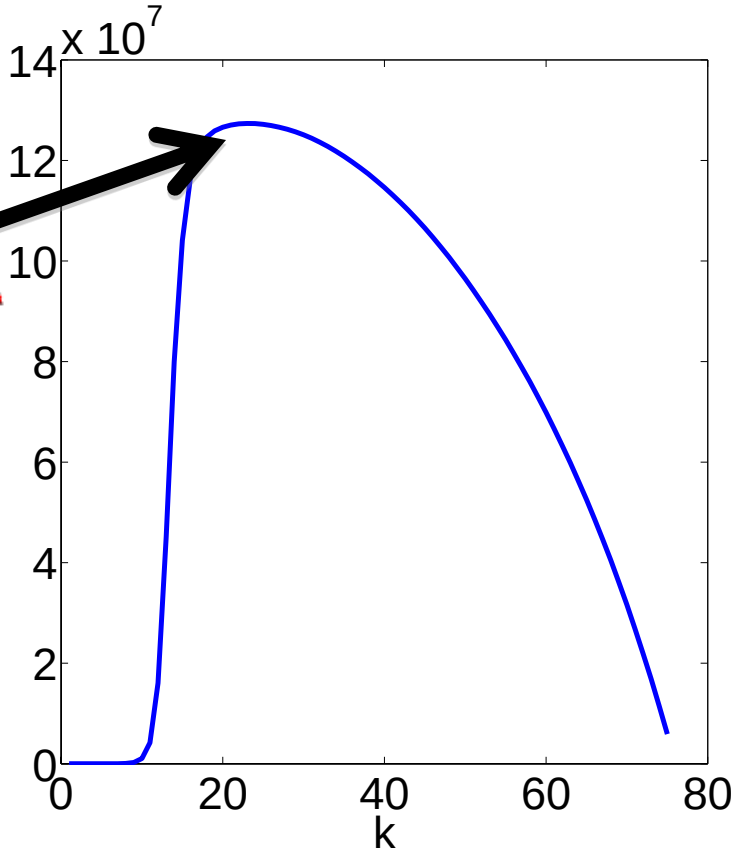




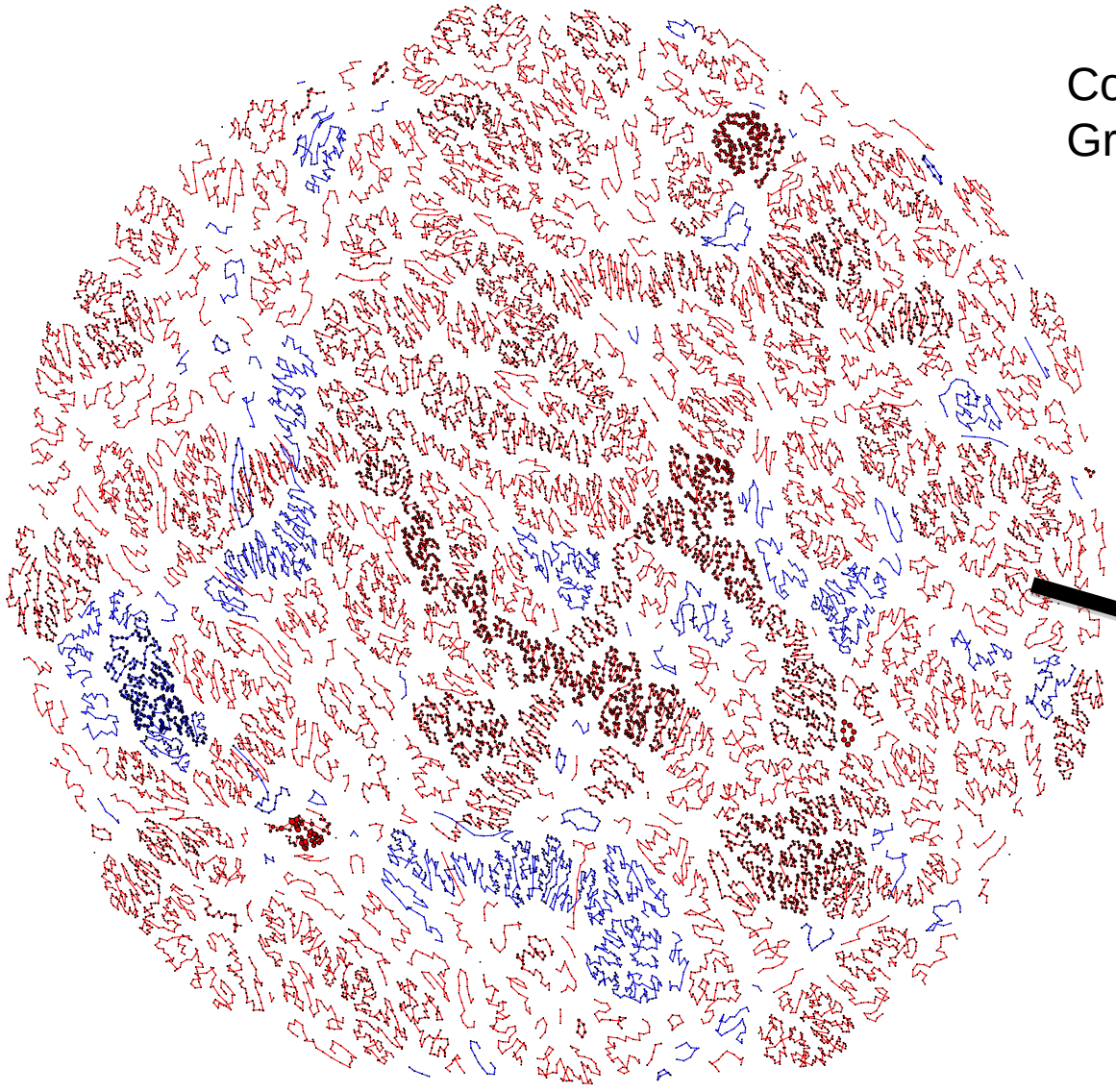
De Bruijn-based metrics

Connections to de Bruijn Graphs

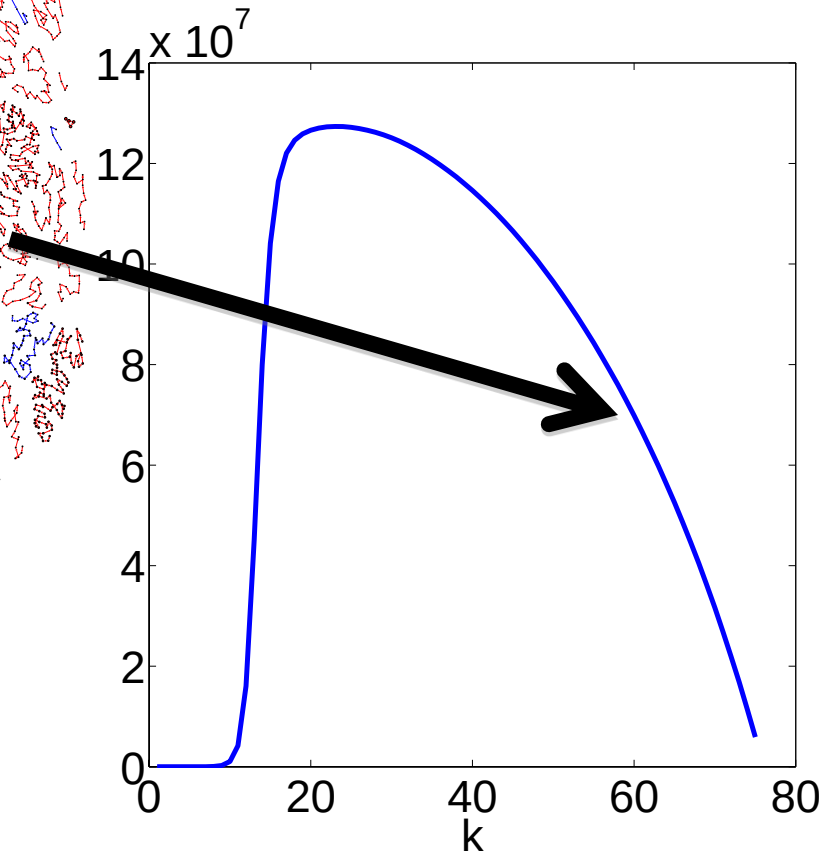




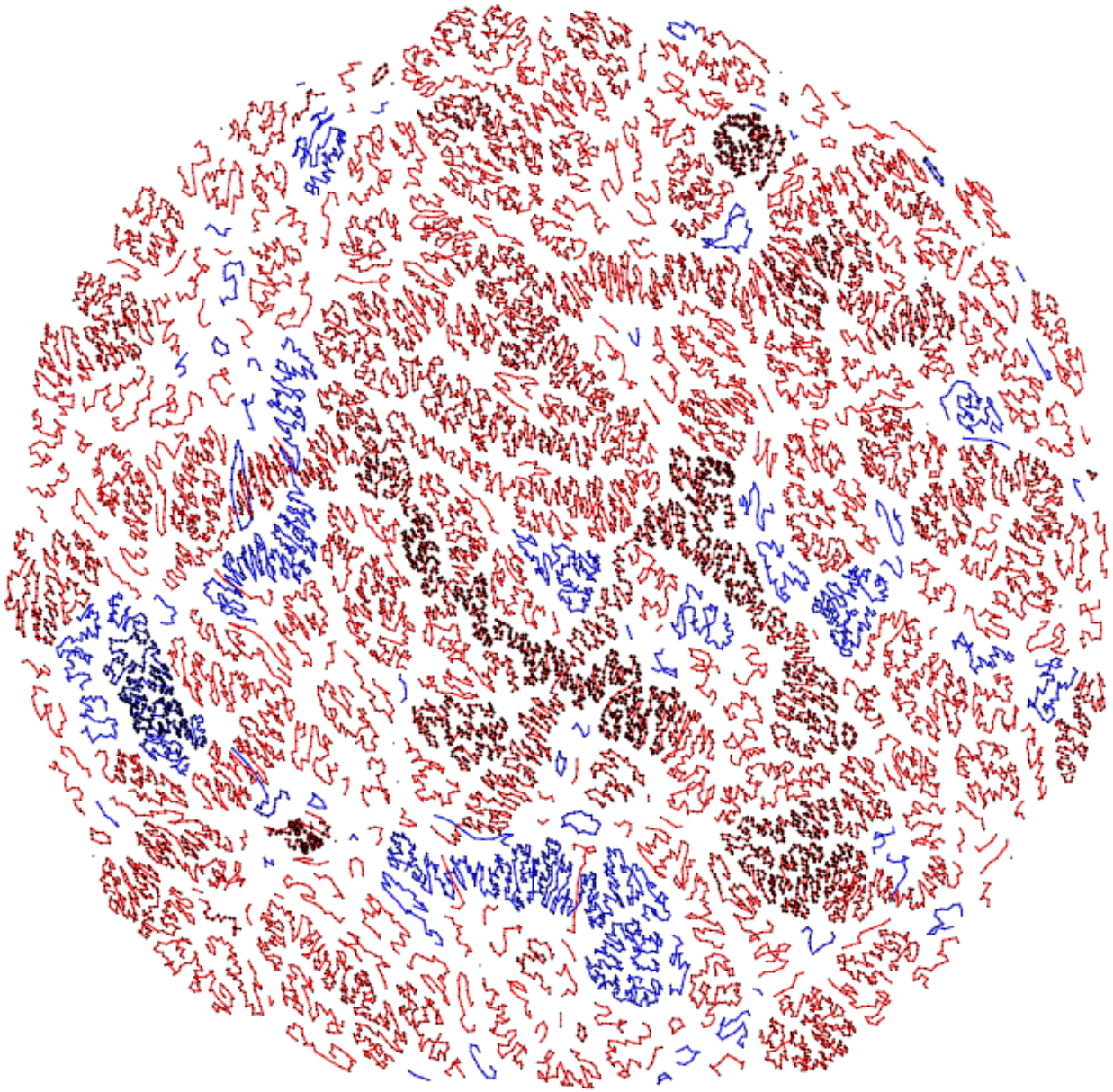
Connection to complexity



Connections to de Bruijn Graphs



De Bruijn-based metrics



CAMI: Critical Assessment of Metagenomic Interpretation

Presented by Alexander Sczyrba

University of Bielefeld

CAMI
Critical Assessment of Metagenomic Interpretation

Organisers:
Alice McHardy (U. Düsseldorf), Thomas Rattei (U. Vienna), Alex Sczyrba (U. Bielefeld)

Outline

- Assessment of computational methods for metagenome analysis
  - WGS assembly
  - binning methods
- Set of simulated benchmark data sets
  - generated from unpublished genomes
- Decide on set of performance measures
- Participants download data und submit assignments via web
- Joint publication of results for all tools and data contributors

Benchmark data sets

- High Complexity, Medium Complexity samples with replicates
- Include strain level variations, include species at different taxonomic distances to reference data
- Simulate Illumina and PacBio reads from unpublished assembled genomes
- Distribute unassembled simulated metagenome samples for assembly and binning

Assessment

Assembly measures

- Reference-dependent measures
  (NG50, COMPASS, REAPR, Feature Response Curves, etc.)
- Reference-independent measures
  (ALE, LAP, ?)

(Taxonomic) binning measures

- (macro-) precision and –recall accuracy,
- taxonomy-based measures

(earth movers distance, i.e. UniFrac, etc.)

- bin consistency (taxonomy-aware, or not)

Main Goals

- comparison of available assemblers and binning tools
- best practice for metagenomic assembly and binning
- develop a set of guidelines
- develop better assembly metrics

Contributors

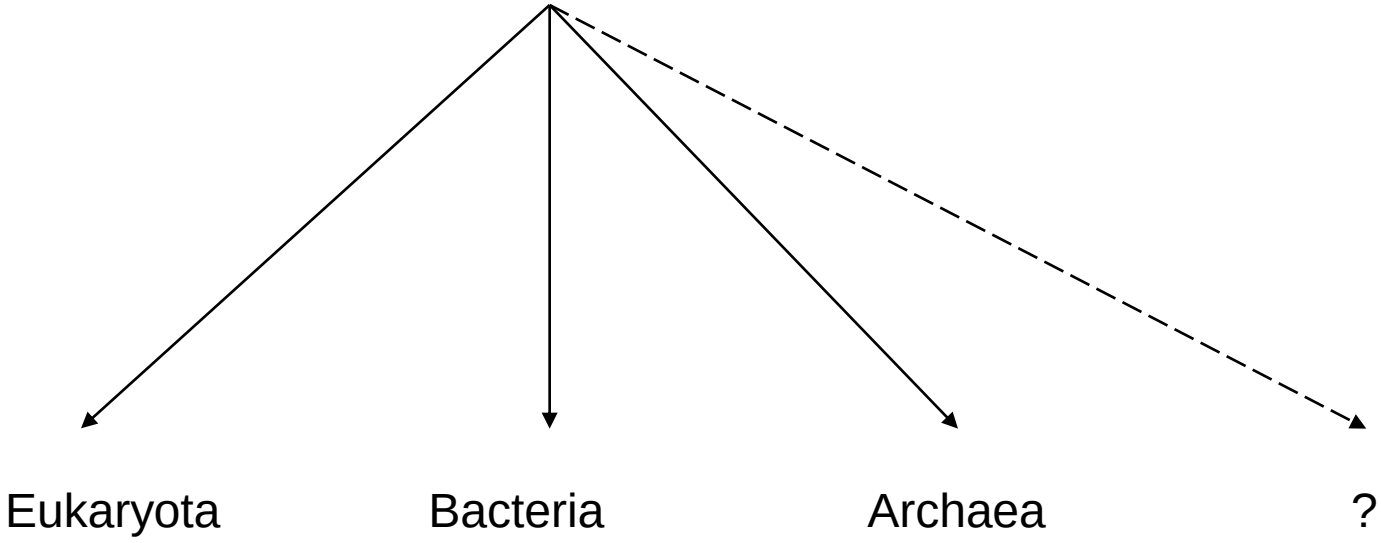
- Daniel Huson
- Richard Leggett
- Folker Meyer
- Mihai Pop
- Eddy Rubin
- Monica Santamaria
- Gabriel Valiente
- Tandy Warnow
- …?

Fourth Domain

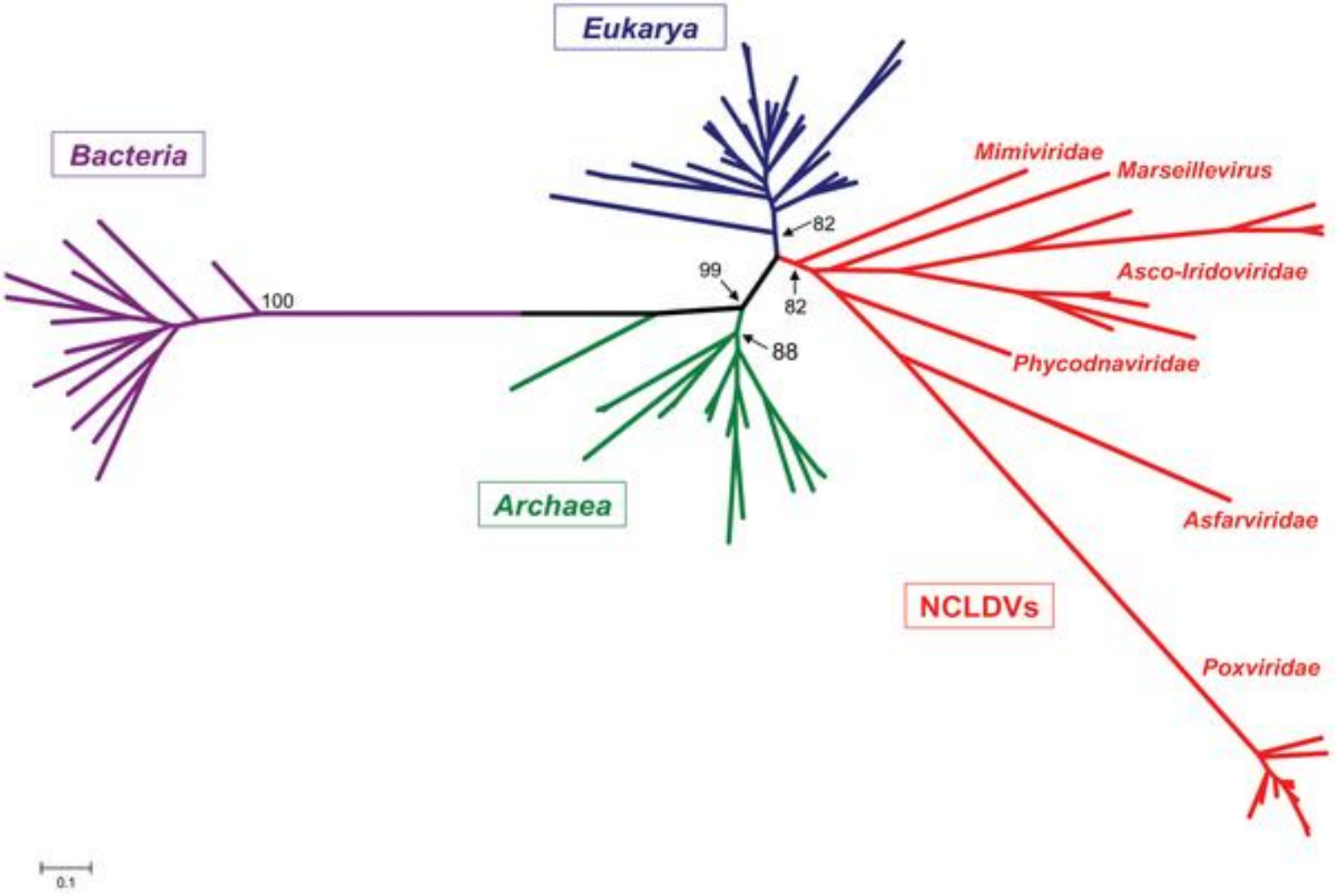
Presented by Wally Gilks

University of Leeds

Fourth Domain



Phylogeny of Giant RNA Mimivirus ribosomal genes



Boyer M, Madoui M-A, Gimenez G, La Scola B, et al. (2010) Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses. PLoS ONE 5(12): e15530. doi:10.1371/journal.pone.0015530
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0015530

PLOS | ONE

Questions?