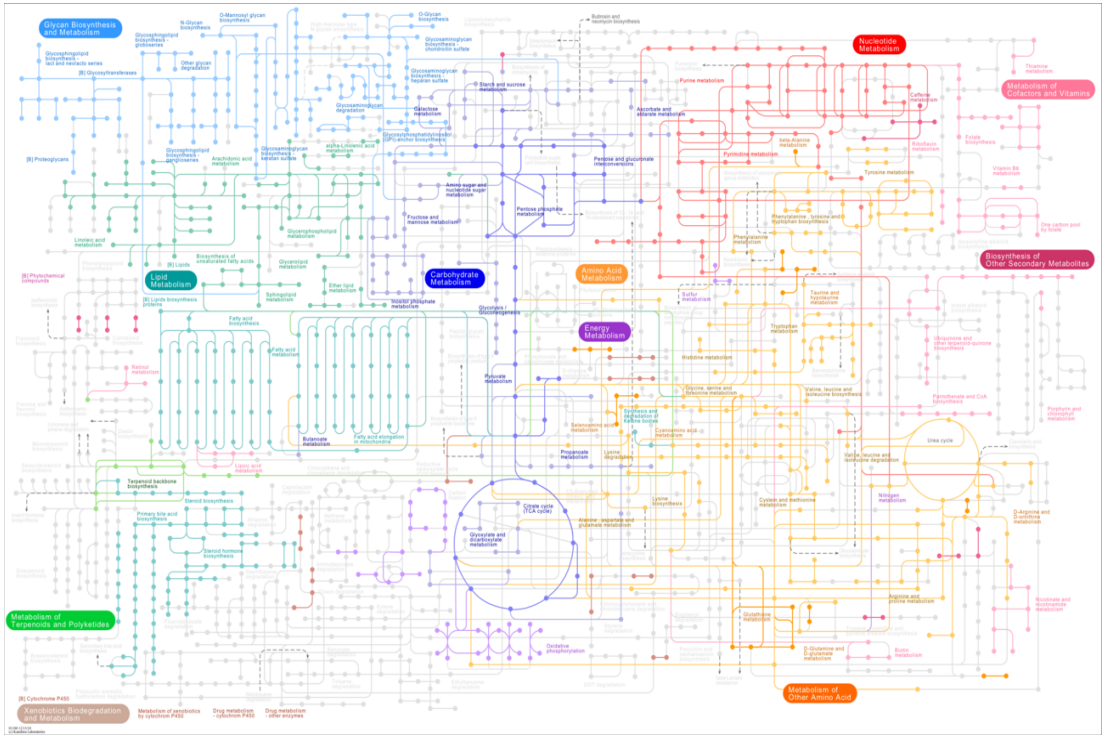
Data integration

"combining data from different sources for a unified view"

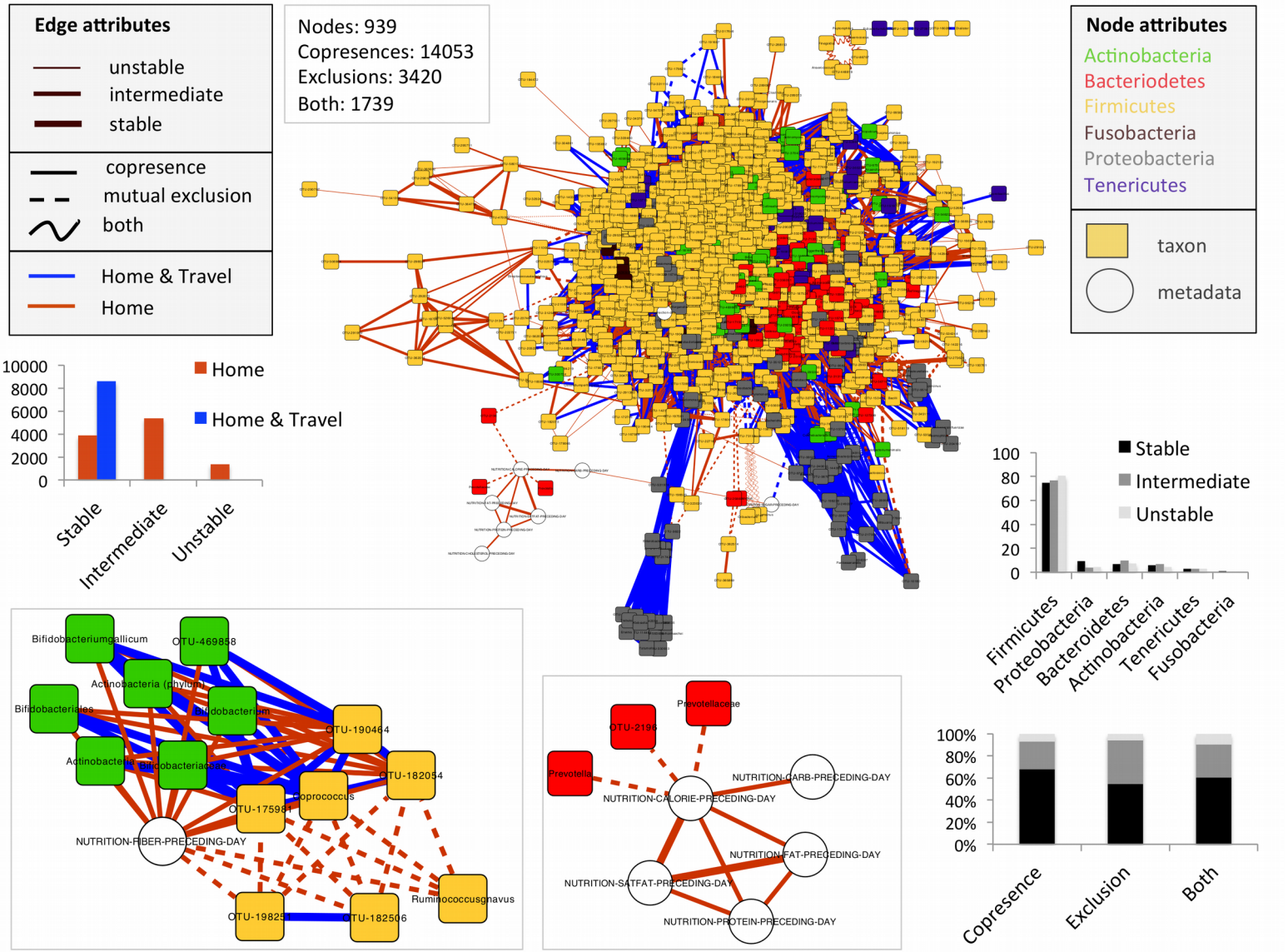


Leo Lahti | www.iki.fi/Leo.Lahti | leo.lahti@iki.fi | @antagomir
D.Sc. / Adj. Prof. University of Turku, Finland & VIB/KU Leuven

VIB Community Meeting, Brussels Feb 27, 2017

Complex systems, versatile data

sample size; metadata; supporting data; complementary views



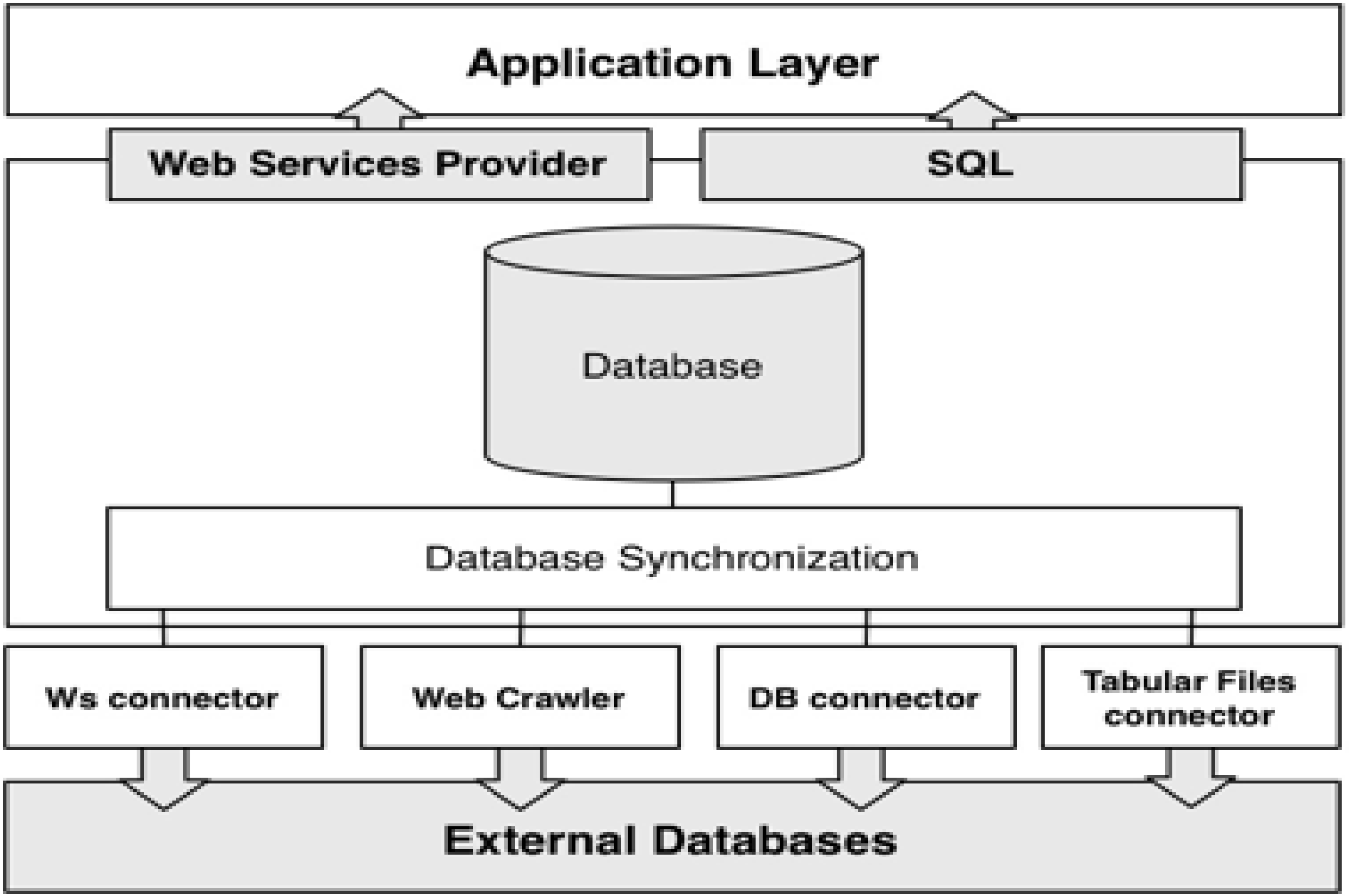
Faust et al. Curr. Op. Microbiol. 2015

Varieties of data integration

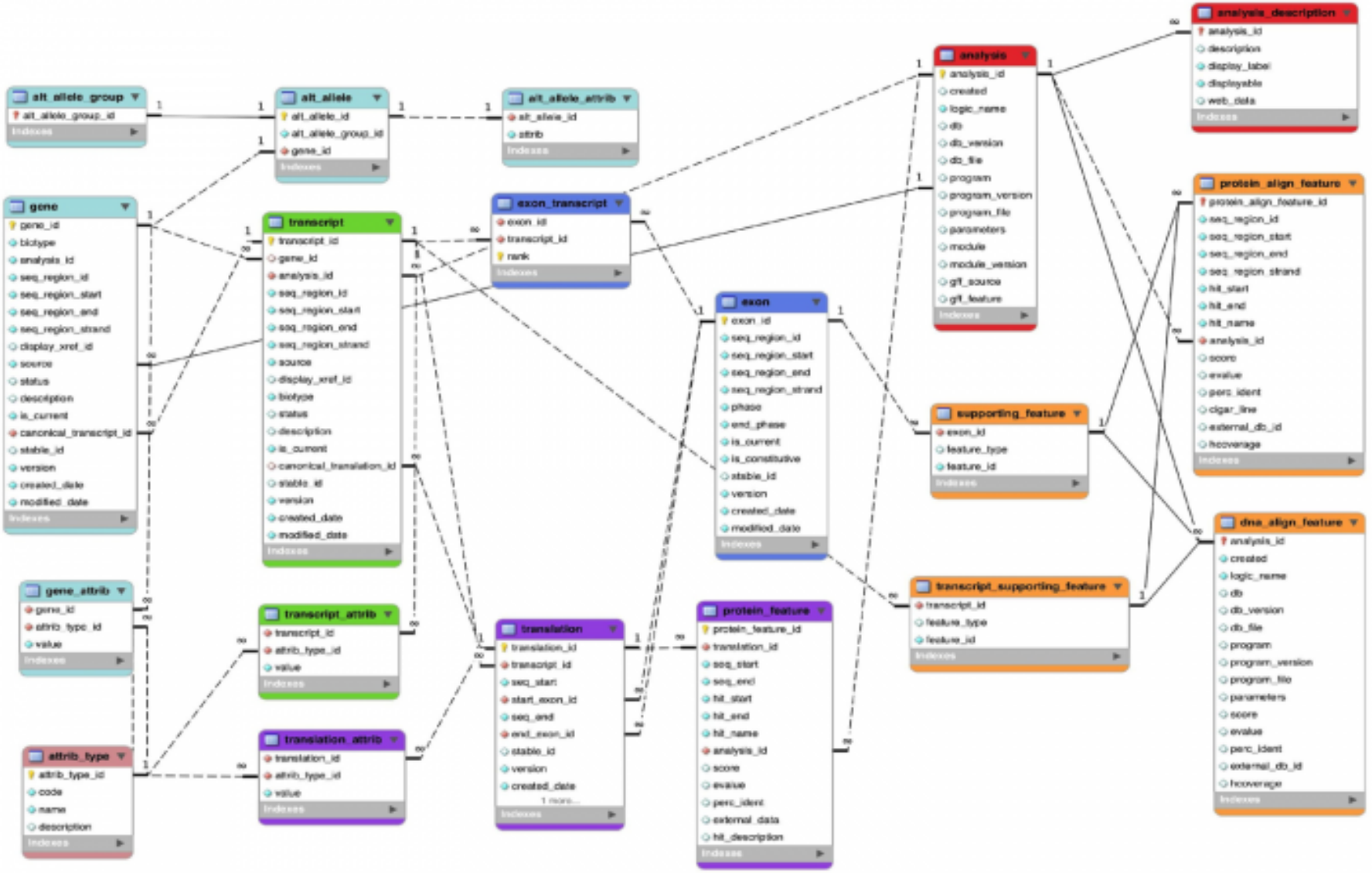
1. Infrastructure view → workflows & compatibility
    - Data access & curation; Linked data; Semantic web
2. Meta-analysis view (single source)→ sample size, context
    - Pool data/results to increase sample size
3. Supervised view (single source, supported by others)
    - Support single source analysis
4. Multi-view learning (multiple sources)
    - Joint analysis of multiple data sources
    - Higher-level mechanisms

1. Infrastructure view

"Information silo" occurs whenever a data system is incompatible or not integrated with other data systems.



Interlinking in Ensembl



Linked data enables association analyses

drug interactions with gut microbiome

Falony et al. Science 352, 2016.

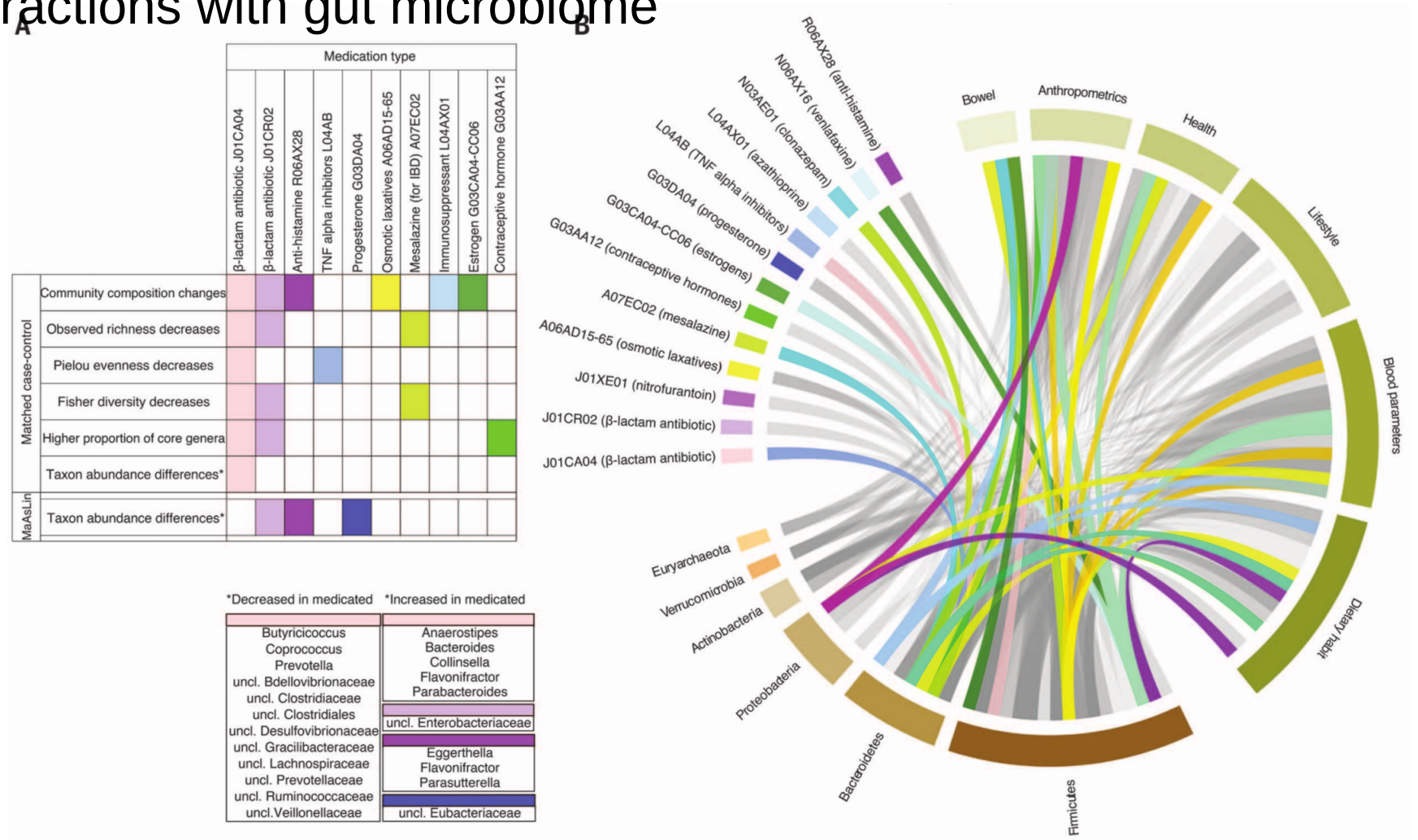
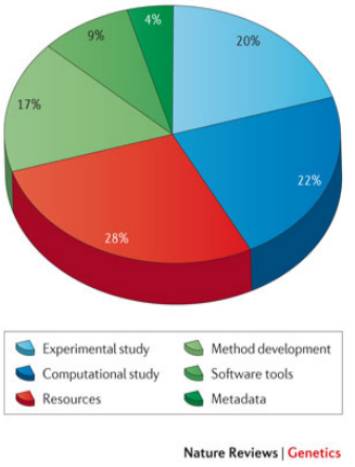


Fig. 5. Drug interactions in the FGFP. (**A**) Overview of the association between different types of medication and microbiome composition. Colored boxes (color coding according to medication) represent a significant result in the matched case-control (FDR<5%) or boosted additive general linear modeling (FDR<10%, table S11) analyses. The effect (decrease/increase) of medication on genera abundances is specified. (**B**) Circos plot showing correlations between covariates and genus abundances (FDR<10%) interacting with drugs. Genera are grouped at phylum level; ribbons represent genus-phenotype associations and are colored according to the confounding medication (gray indicates nonconfounded).

2. Meta-analysis (single source)



Review

Nature Reviews Genetics **14**, 89-99 (February 2013) | doi:10.1038/nrg3394

Reuse of public genome-wide gene expression data

Johan Rung[1] & Alvis Brazma[1] About the authors

top

Our understanding of gene expression has changed dramatically over the past decade, largely catalysed by technological developments. High-throughput experiments — microarrays and large amounts of genome-wide g archives. Added-value databases to make them accessible to every the gene expression data that ar making use of these data. Reuse many obstacles in data preparati results. We will discuss these ch believe can improve the utility of

Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas

Larsson Omberg, Kyle Ellrott, Yuan Yuan, Cyriac Kandoth, Chris Wong, Michael R Kellen, Stephen H Friend, Josh Stuart, Han Liang & Adam A Margolin

Affiliations | Contributions | Corresponding authors

Nature Genetics **45**, 1121–1126 (2013) | doi:10.1038/ng.2761

Published online 26 September 2013

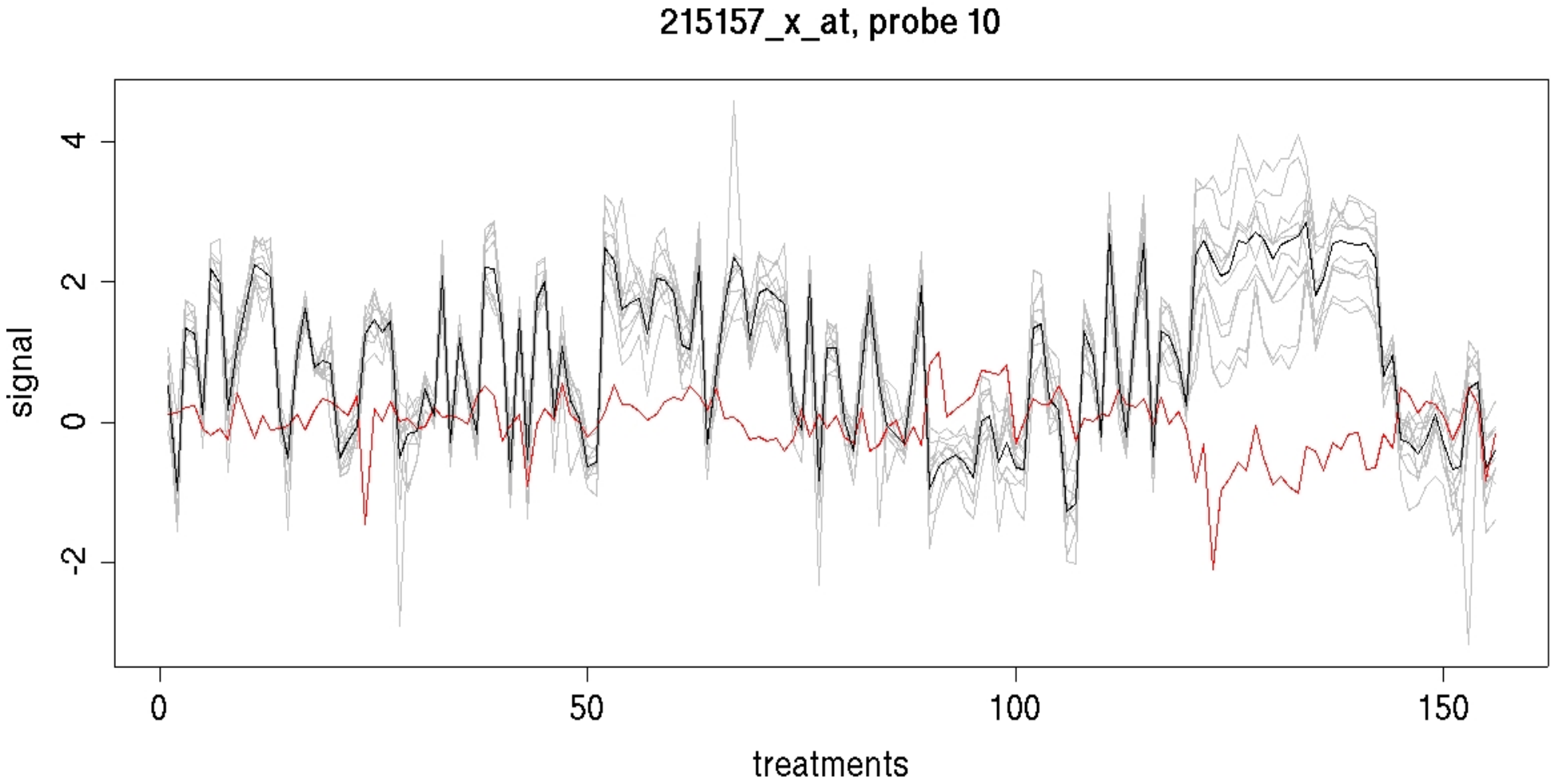
Automated multidimensional phenotypic profiling using large public microarray repositories

Min Xu[a,1], Wenyuan Li[a,1], Gareth M. James[b], Michael R. Mehan[a], and Xianghong Jasmine Zhou[a,2]

[a]Molecular and Computational Biology, Department of Biological Sciences, and [b]Marshall School of Business, University of Southern California, Los Angeles, CA 90089

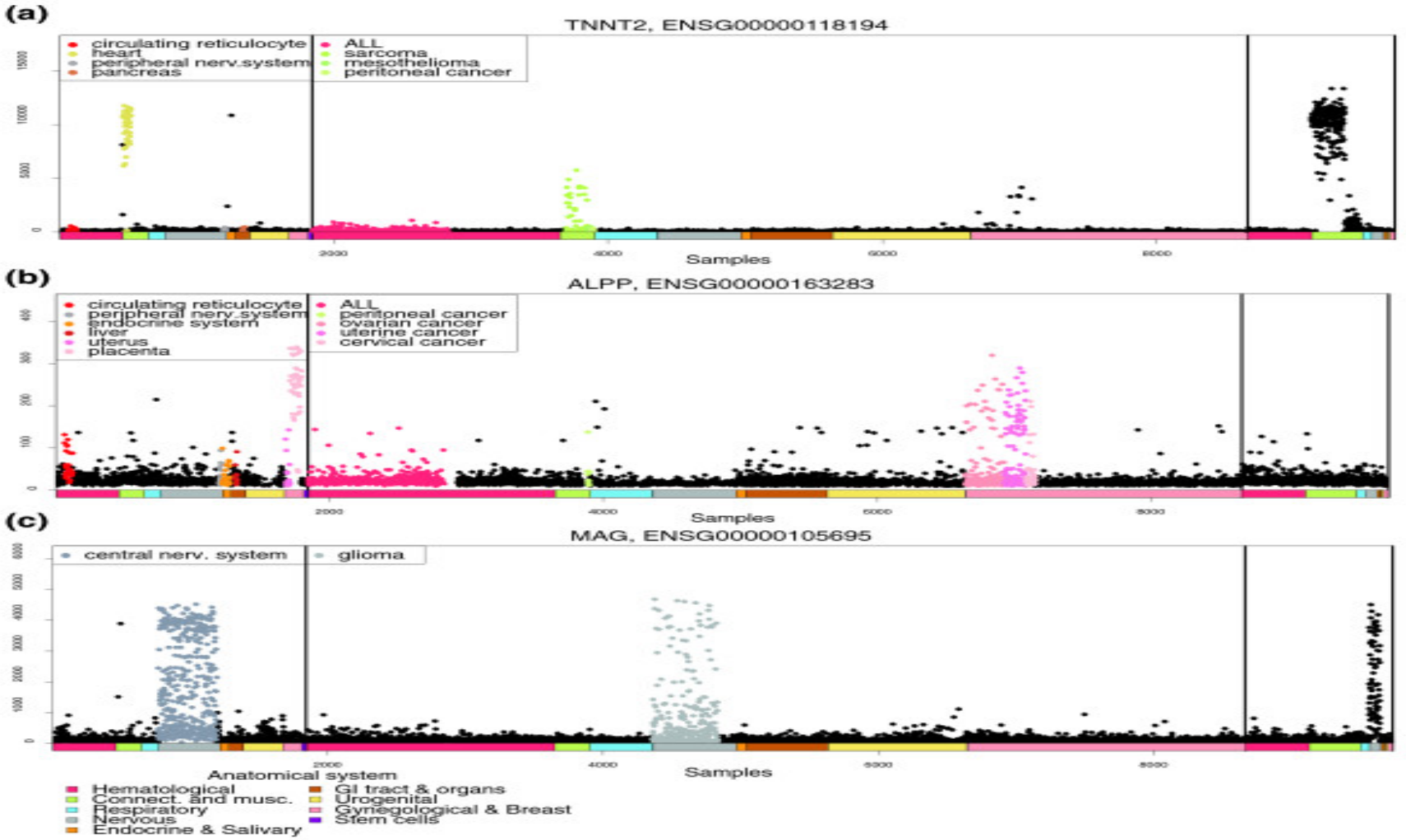
More reliable results by integrating independent views

detect noisy probes by analysing joint performance across many samples



Lahti et al., TCBB/IEEE 2011 & NAR 2013

Pooled analysis of ~10,000 samples reveals tissue-specific RNA activity (multiple measurement platforms!):

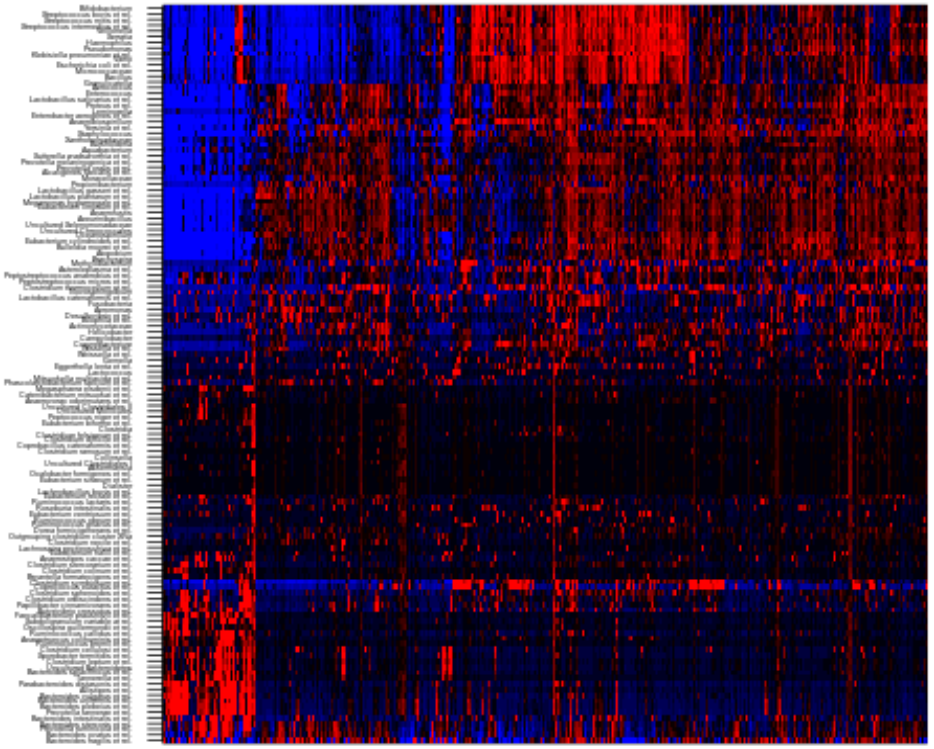


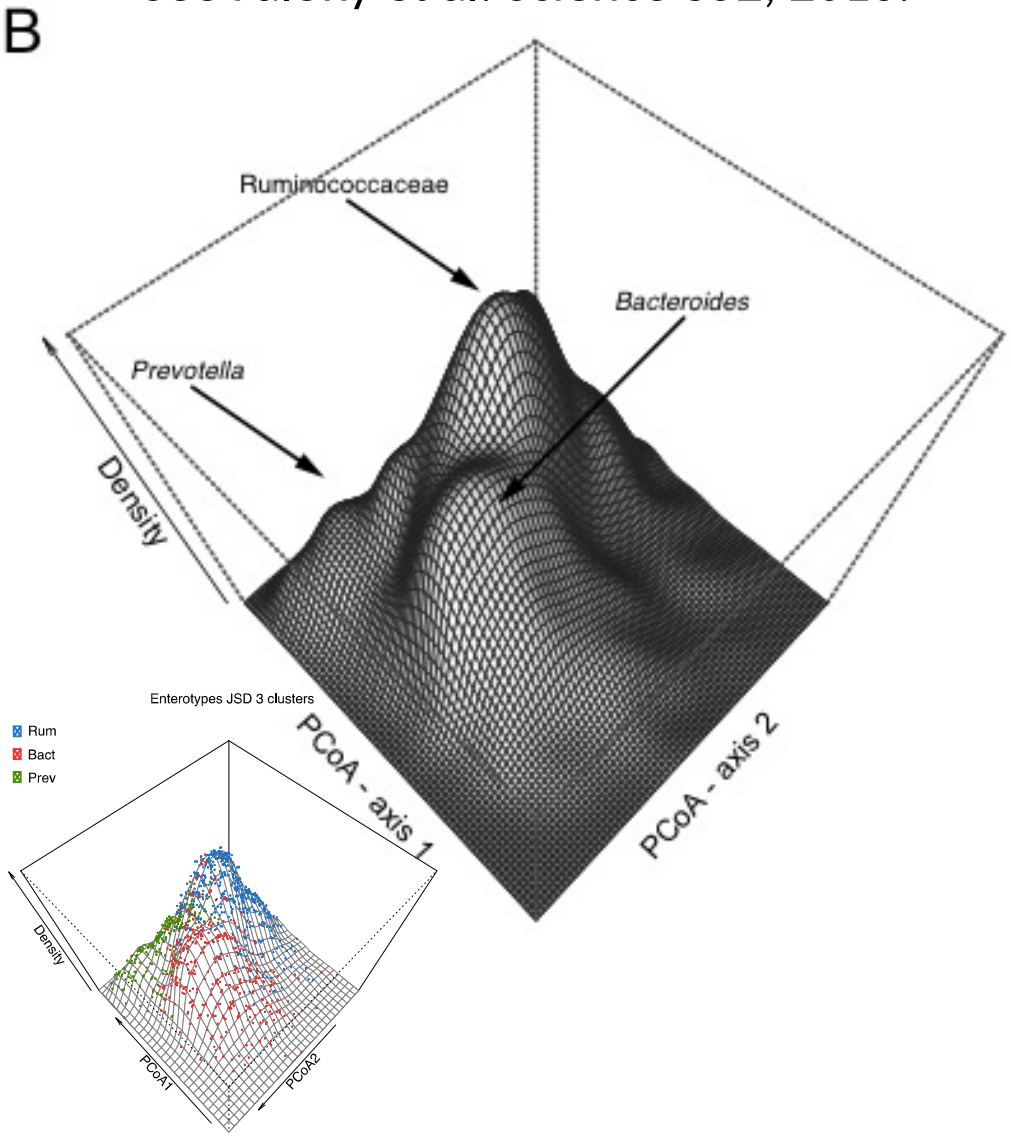
Kilpinen et al. *Genome Biology* 2008, GeneSapiens database

Enterotype landscape of the Flemish Population shows density peaks of ‘preferred’ ecosystem constellations

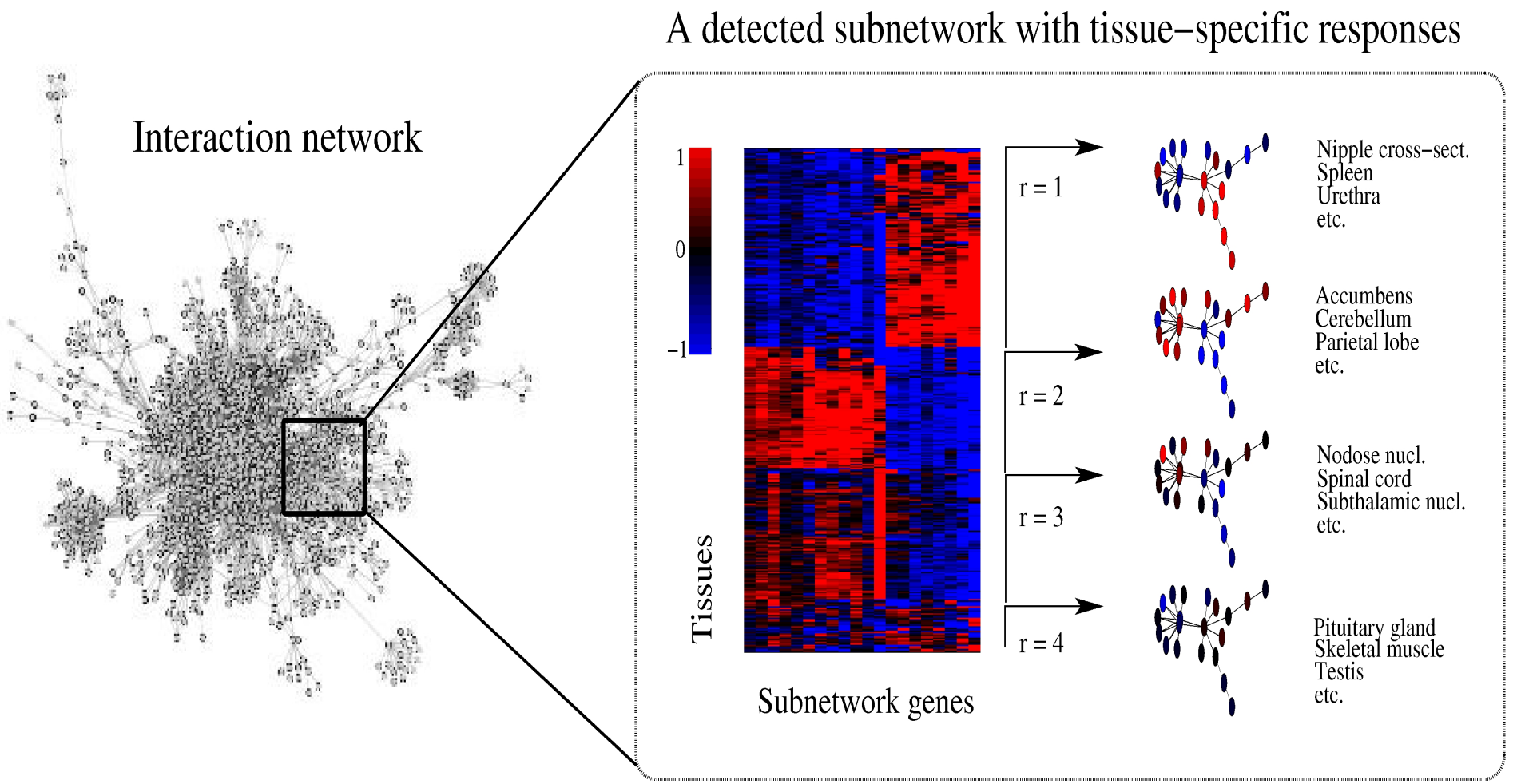
See Falony et al. Science 352, 2016.

Similar evidence with HITChip Atlas: 10,000 gut microbiota samples; see Lahti et al. Nature Comm. 2014





3. Supervised analysis: transcriptome activity with known signaling pathways



Improved contig clustering by integrating sample coverage & sequence composition

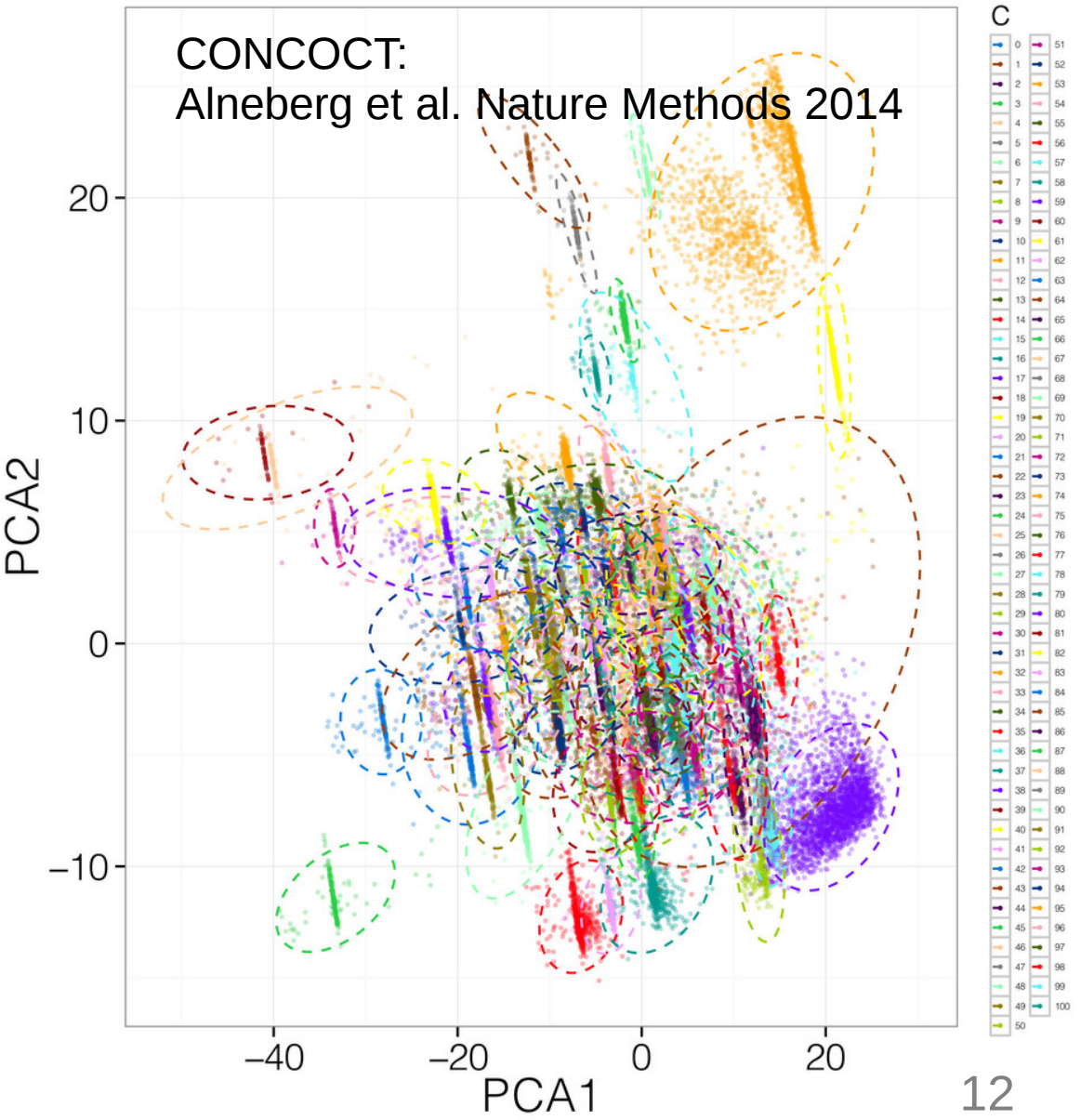
Coverage: contig abundance across samples (N=64)

Composition: frequency of k-mers (N=136 4-mers)

Joint data: 100,000 contigs x 201 features (abundances + k-mers)

Dimension reduction: with PCA from 201 → 21 (>90% information retained)

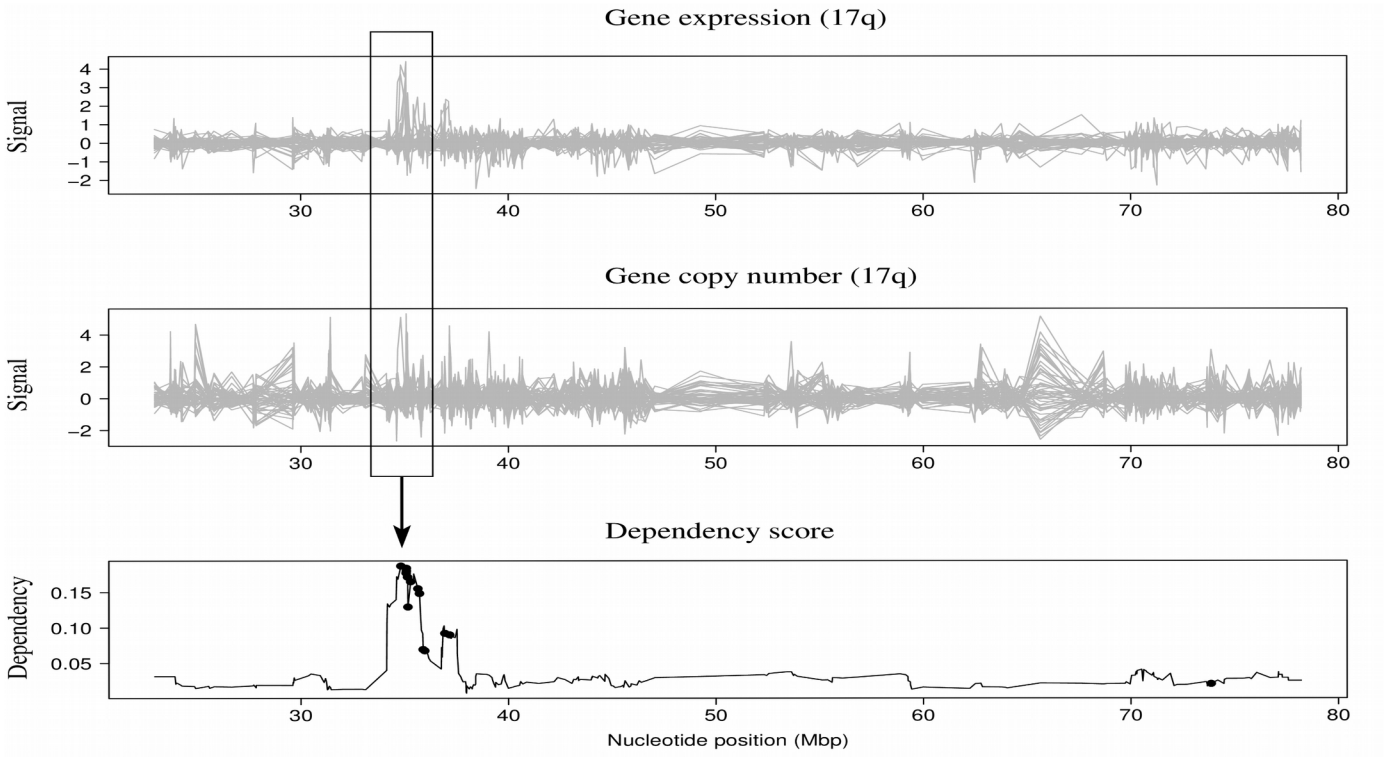
Clustering: variational Gaussian mixtures

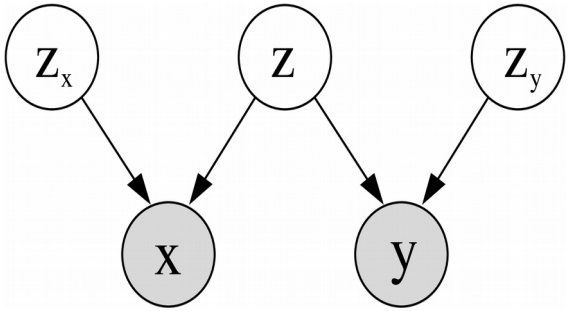


4. Multi-view analysis & Data fusion

Cancer gene discovery with latent variable models

Integrating multiple views on the same object for a unified and improved representation





$$\begin{cases} X = W_x \mathbf{z} + \varepsilon_x \\ Y = W_y \mathbf{z} + \varepsilon_y \end{cases}$$

R/Bioconductor: pint. Lahti et al. MLSP 2009

Summary

Varieties

1. Infrastructure
2. Meta-analysis
3. Supervised analysis
4. Multi-view learning

Advantages

→ data access and reuse potential

→ sample size & statistical power

→ simplify & guide analysis

→ reliability; independent evidence

→ higher-level mechanisms

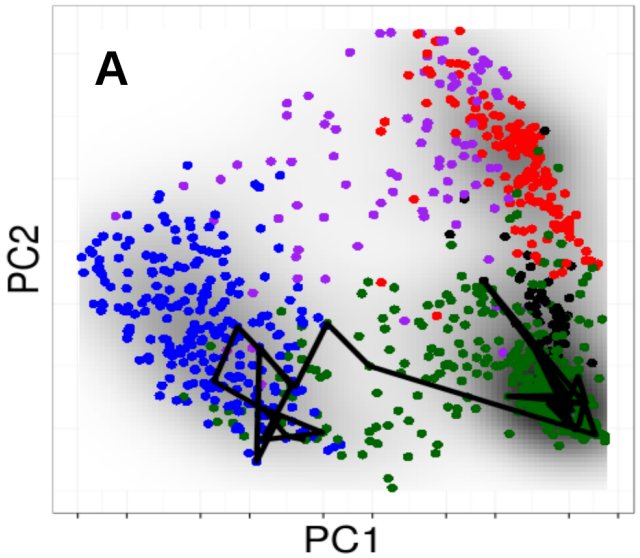
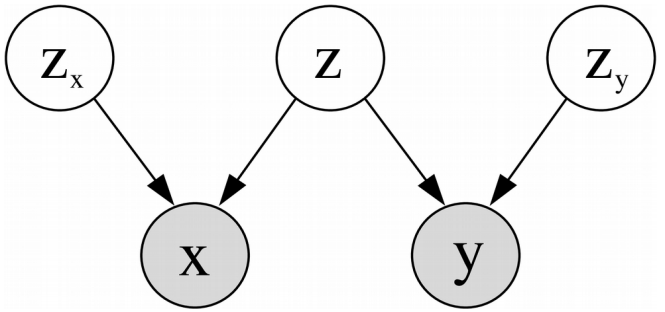


Figure: Faust et al. Curr. Op. Microbiol. 2015
Data: Gajer et al. 2012

Thank You !



Jeroen Raes

Karoline Faust

Christopher Quince

Willem M de Vos

Contact:
http://www.iki.fi/Leo.Lahti

VIB

KU LEUVEN

Suomen Akatemia
Finlands Akademi • Academy of Finland

Turun yliopisto
University of Turku