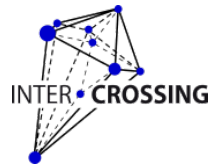


NGS and cloud computing

Raquel Tobes

2013-08-26

Era7 bioinformatics

ohnosequences!

NGS data analysis

The emergence of the Bioinformatics bottleneck

Bioinformatics analysis, before, during and after experimental work

NGS data analysis

NGS is perfectly suited for the cloud

A lot of standard NGS data analysis processes:

- imply storage needs near to terabytes
- are inherently parallel
- require high computational power
- are peaks over the baseline of computational needs

NGS data analysis

INTER • CROSSING

NGS is inherently parallel

Next Generation Sequencing =
Massively Parallel Sequencing

Short reads → High Coverage

Era7 bioinformatics
ohnosequences!

NGS data analysis in the cloud

NGS is inherently parallel

Tasks to be automated and/or parallelized

Management of reads:

- Quality analysis
- Pre-processing:
  - De-multiplexing
  - Filtering
  - Trimming
  - Indexing

Era7 bioinformatics

ohnosequences!

NGS data analysis in the cloud

INTER • CROSSING

NGS is inherently parallel

Tasks to be automated and/or parallelized

Functional Annotation:

- gene-centric annotation
- protein-centric annotation
- transcript-centric annotation

Era7 bioinformatics

ohnosequences!

NGS data analysis in the cloud

NGS is inherently parallel

Tasks to be automated and/or parallelized:

- Taxonomic assignment
- Motif search
- Ortholog protein analysis

NGS data analysis

NGS demands high computational power

- Assembly
- Comparative genomics:
  - Massive similarity analysis: BLAST, MUMmer,...
  - Massive alignment
  - Variant detection: SNPs, indels, rearrangements
- Protein networks, regulatory networks, pathways
- Analysis of data with hierarchical structures
- Visualization

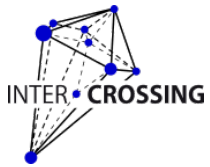
de novo assembly

'genome assembly is one of the most fundamental problems to address. Before any kind of genomic analysis can commence we need to assemble the reads'

'Accurate genome assembly requires sequencing at high depth, and assembling millions of these short reads into a full-length genome is computationally difficult as for each read, contiguous sequences need to be identified from a large unstructured pool of short reads.'

Computational solutions for omics data.
Berger B *et al.*, Nat Rev Genet. 2013

what is assembly?



tion to under

Bioinf

to understand Biology

atics is the science of us

tion to understand Biology

stand Biology

using informa

the science of

using information for

Bioinform

matics is the science

Bioinformatics is

unders

ormatics is the science

using information

Bioinfor

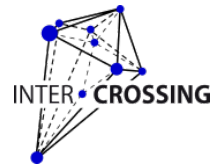
nce of using informa

tand Biology

Era7 bioinformatics

ohnosequences!

what is assembly?



de novo assembly

Bioinf

Bioinformatics is

atics is the science of us

Bioinfor

the science of

using informa

matics is the science

using information

ormatics is the science

of using information to

Bioinform

nce of using informa

tion to understand Biology

to understand Biology

unders

tand Biology

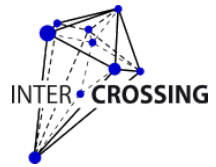
tion to under

stand Biology

Era7 bioinformatics

ohnosequences!

what is assembly?

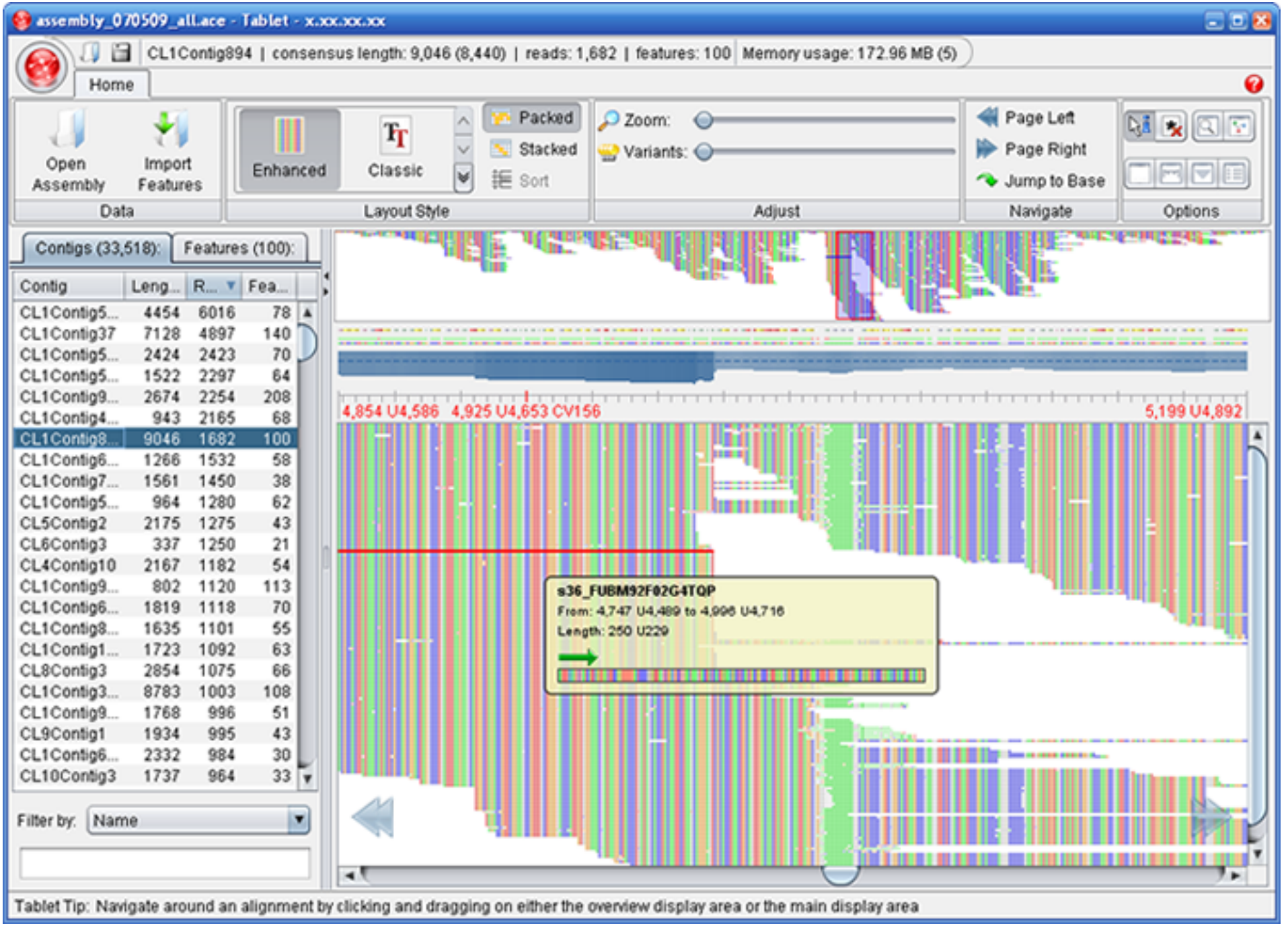


mapping to a reference

Bioinformatics is the science of using information to understand Biology				
Bioinf	ormatics is the science	of using informa	tion to under	stand Biology
Bioinformatics is	the science of	using informa	tion to understand Biology	
Bioinfor	matics is the science	of using information to	unders	tand Biology
Bioinform	atics is the science of us	ing information	to understand Biology	

4x coverage

Era7 bioinformatics
ohnosequences!


assembly_070509_all.ace - Tablet - x.xx.xx.xx
CL1Contig894 | consensus length: 9,046 (8,440) | reads: 1,682 | features: 100
Memory usage: 172.96 MB (5)
Home
Open Assembly
Import Features
Data
Enhanced
Classic
Layout Style
Packed
Stacked
Sort
Zoom:
Variants:
Adjust
Page Left
Page Right
Jump to Base
Navigate
Options
Contigs (33,518):
Features (100):
Contig Leng... R... Fea...
CL1Contig5... 4454 6016 78
CL1Contig37 7128 4897 140
CL1Contig5... 2424 2423 70
CL1Contig5... 1522 2297 64
CL1Contig9... 2674 2254 208
CL1Contig4... 943 2165 68
CL1Contig8... 9046 1682 100
CL1Contig6... 1266 1532 58
CL1Contig7... 1561 1450 38
CL1Contig5... 964 1280 62
CL5Contig2 2175 1275 43
CL6Contig3 337 1250 21
CL4Contig10 2167 1182 54
CL1Contig9... 802 1120 113
CL1Contig6... 1819 1118 70
CL1Contig8... 1635 1101 55
CL1Contig1... 1723 1092 63
CL8Contig3 2854 1075 66
CL1Contig3... 8783 1003 108
CL1Contig9... 1768 996 51
CL9Contig1 1934 995 43
CL1Contig6... 2332 984 30
CL10Contig3 1737 964 33
Filter by: Name
4,854 U4,586 4,925 U4,653 CV156
5,199 U4,892
s36_FUBM92F02G4TQP
From: 4,747 U4,489 to 4,996 U4,716
Length: 250 U229
Tablet Tip: Navigate around an alignment by clicking and dragging on either the overview display area or the main display area

What is assembly?

Two totally different assembly methods

- *De novo* assembly:

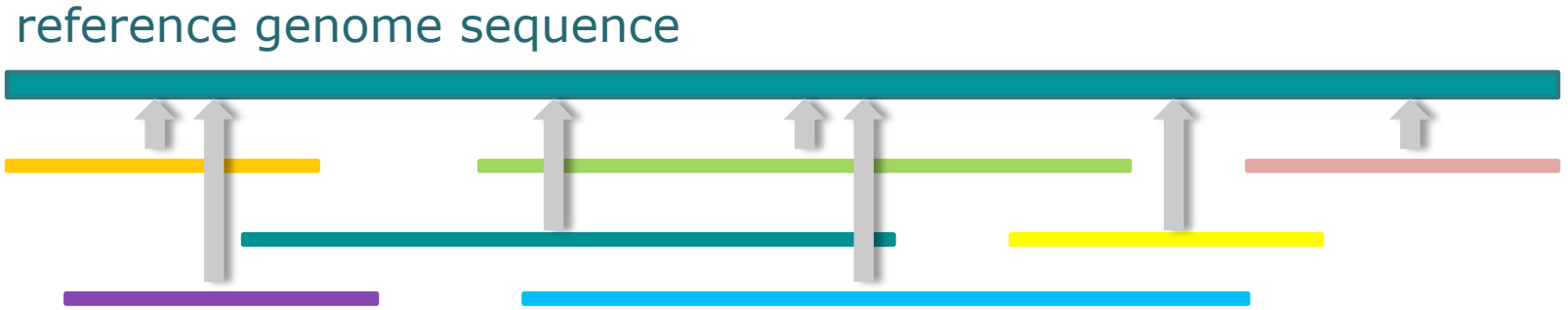
- Assembly mapping to a reference sequence

Era7 bioinformatics

ohnosequences!

What is assembly?

Assembly mapping to a reference



- The determinant point is the alignment of each read to the reference sequence
- In contrast to *de novo* assembly the rest of the reads and the overlapping are not crucial

What is assembly?

Velvet- *de novo* assembly

Algorithms for de novo short read assembly using de Bruijn graphs:

- A de Bruijn graph is a compact representation based on short words (k-mers)
- Velvet is ideal for high coverage, very short read data sets and also assembles and handles paired-end reads
- Velvet produces contigs of up to 50-kb N50 length in simulations of prokaryotic data and 3-kb N50 on simulated mammalian BACs

What is assembly?

MIRA – sequence assembler

Whole Genome Shotgun and EST Sequence Assembler for Sanger, 454 and Solexa / Illumina:

- **Hybrid de-novo assemblies** with Sanger, 454 and Illumina / Solexa
- **Mapping against a reference:** mapping assemblies and automatic tagging of difference site (SNPs, insertions or deletions) of mutant strains against a reference sequence.

What is assembly?

SOAPdenovo assembly method

- SOAPdenovo is a short-read assembly method that can build a *de novo* draft assembly for the human-sized genomes.
- It is specially designed to assemble Illumina GA short reads.
- It uses de Bruijn graph for assembly

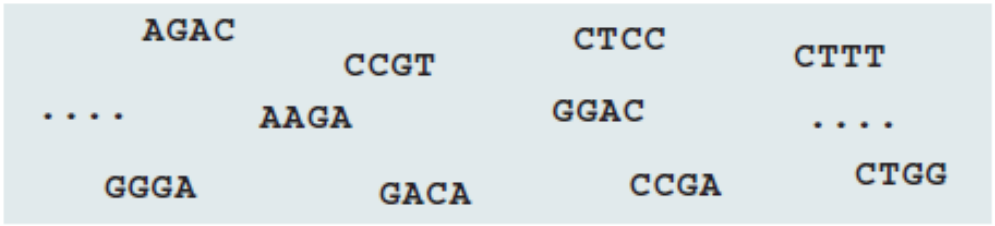
What is assembly?

ALLPATHS assembly method

ALLPATHS-LG is an algorithm (de Bruijn) for genome assembly able to manage massively parallel DNA sequence data from the human and mouse genomes, generated on the Illumina platform.

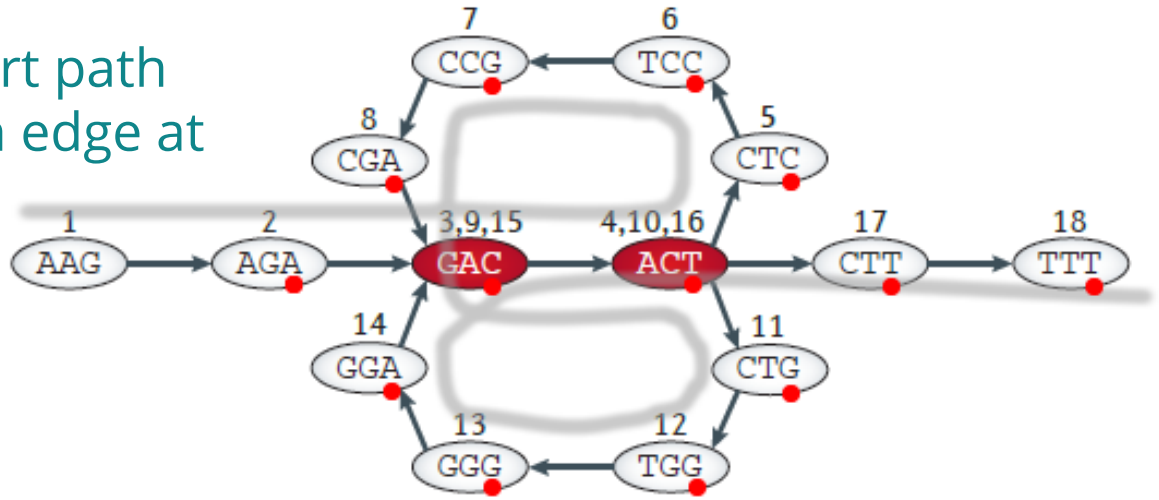
It generates draft genome assemblies with good accuracy (≥99.95%) , short-range contiguity N50 size = 11.5 Mb for human and 7.2 Mb for mouse), long-range connectivity, and coverage

De Bruijn graphs



to find the short path that visits each edge at least once

Read lenght=4
k=3



Modification of Figure 1 of 'Computational solutions for omics data. Berger B *et al.*, Nat Rev Genet. 2013'

assembly in the cloud

Tasks that could be parallelized:

- indexing of prefixes for *de Bruijn* graph building

Tasks that require high computational power:

- to find a path that visits each edge exactly once

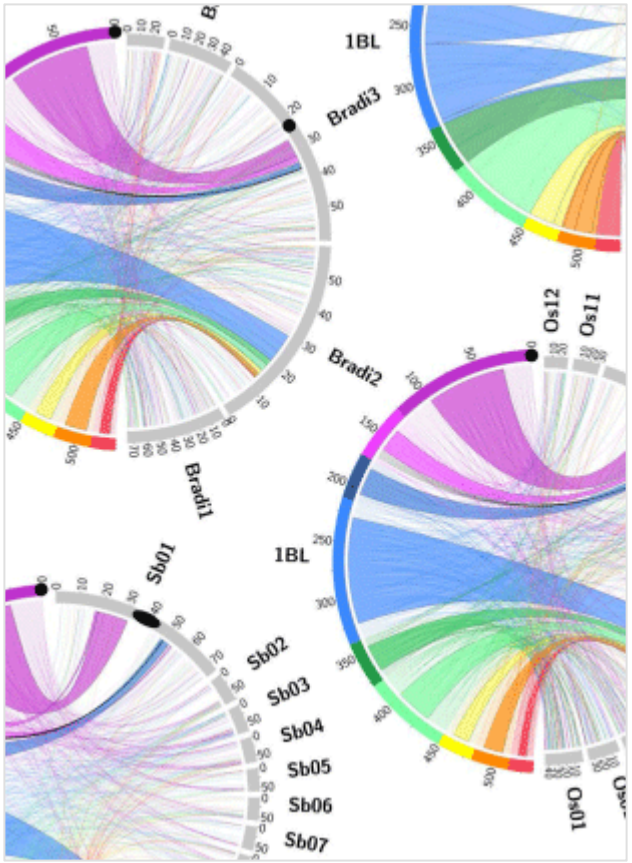
Comparative genomics

Interpreting results to extract biological insights

- Phylogenetic profiles
- Functional profiles
- GO annotation analysis
- Variant analysis
- Evolutionary studies
- Population genetics analysis
- Differential expression analysis
- Taxonomic diversity analysis

Comparative genomics

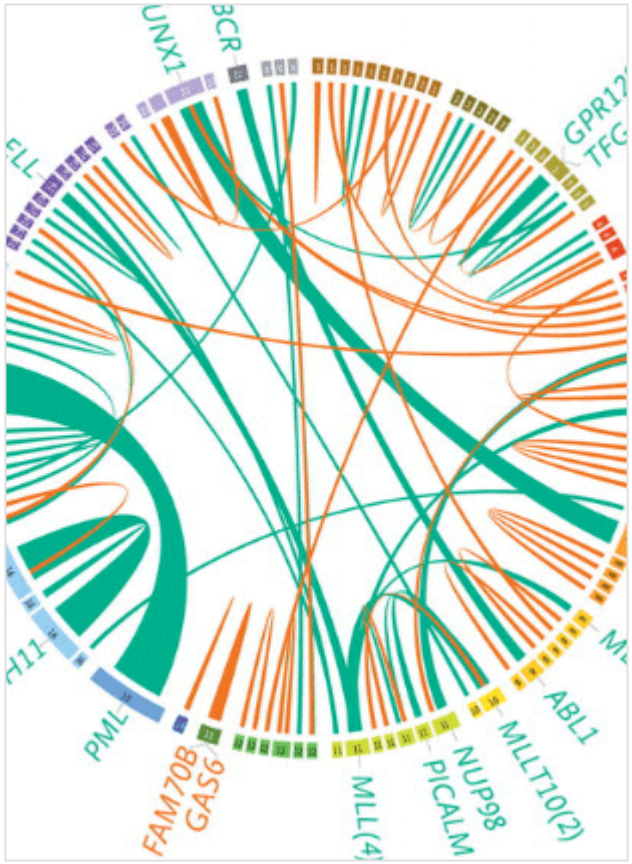
Comparative genomics



Evolutionary studies

25 June 2013 | Philippe R, Paux E, Bertin I *et al.*2013. A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat *Genome Biol* **14**:R64.

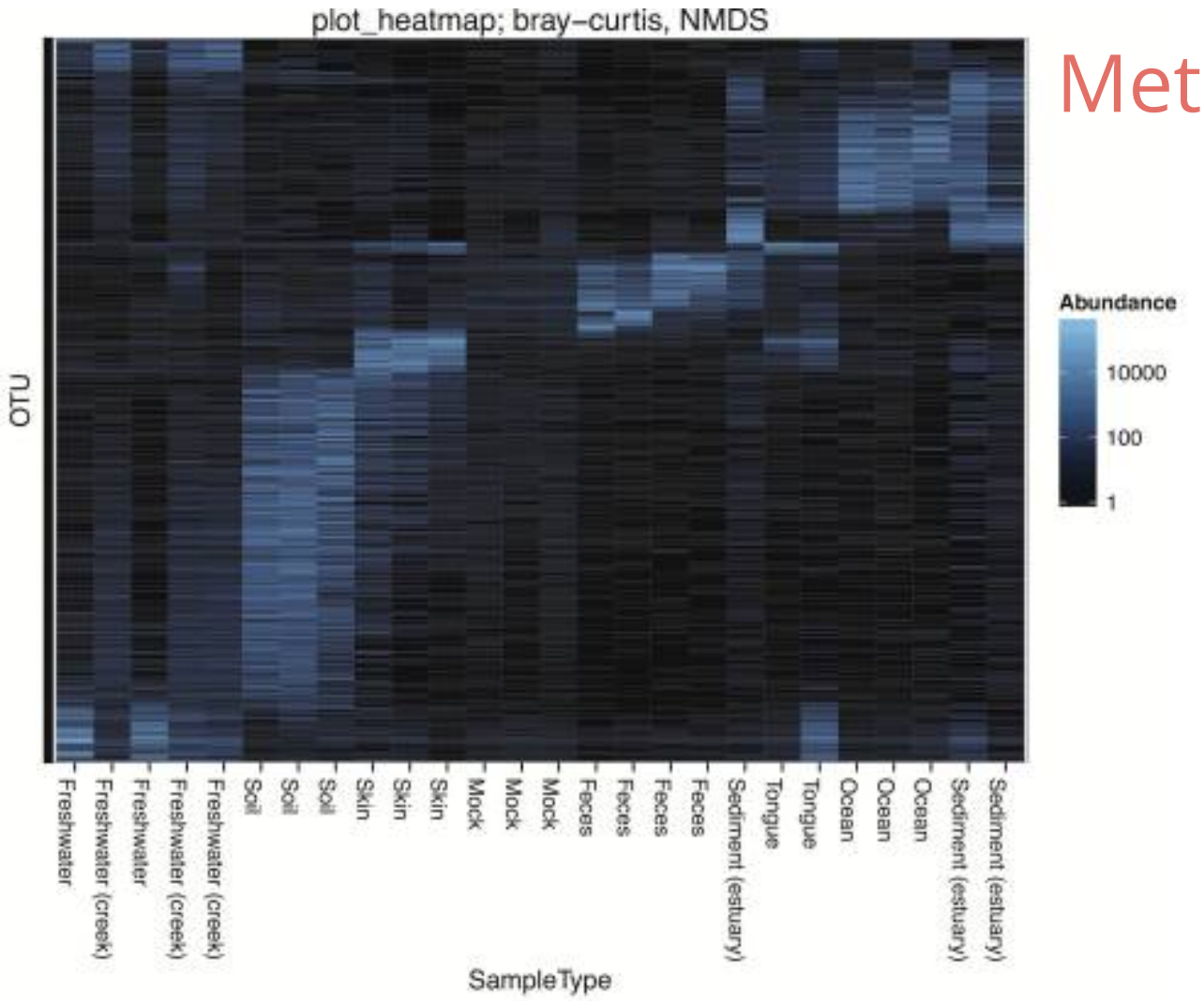
Comparative genomics



Epigenomics studies

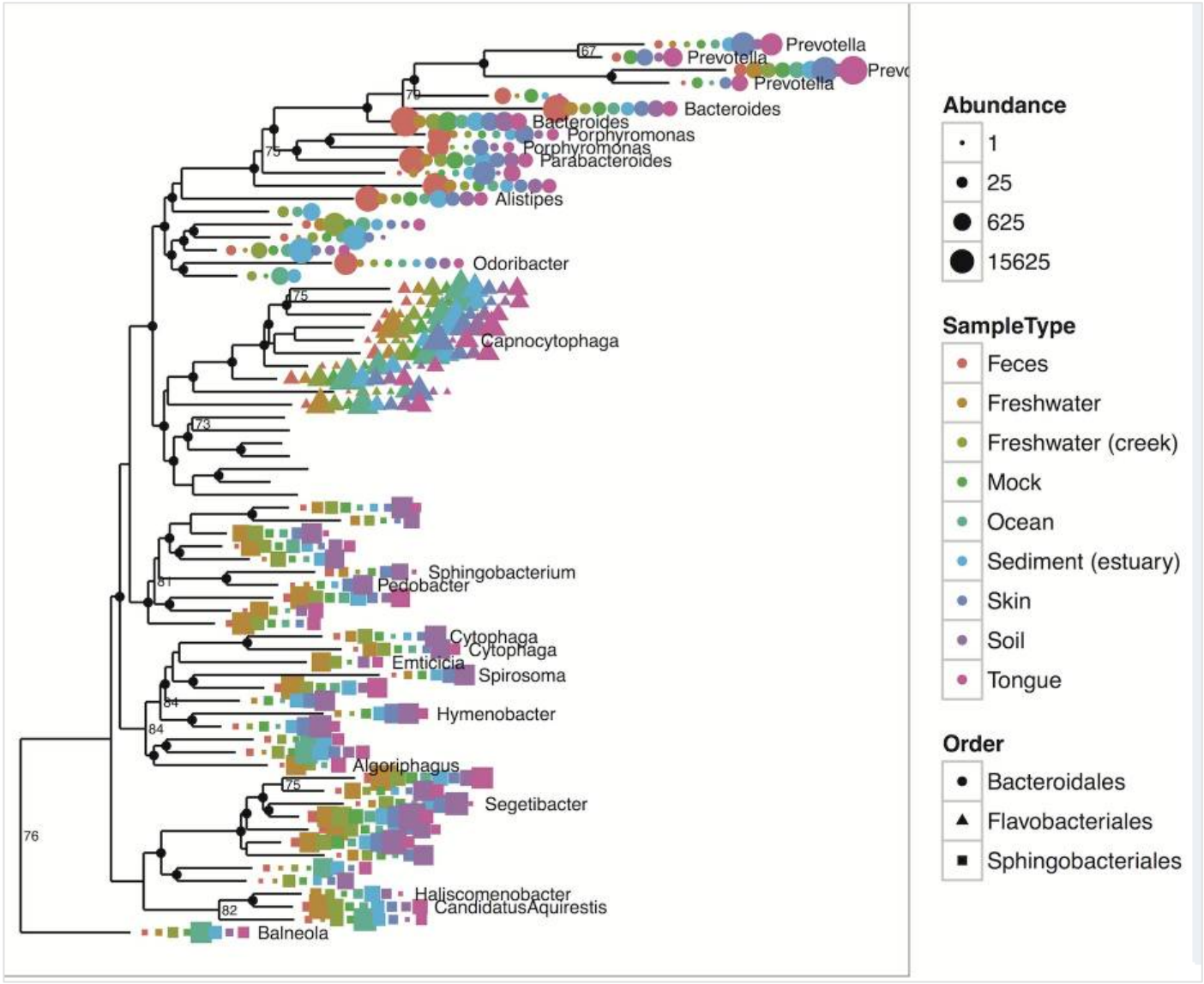
30 May 2013 | Network TCGAR2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia *The New England journal of medicine* **368**:2059-2074.

Comparative genomics



Metagenomics studies

phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.
McMurdie PJ, Holmes S. PLoS One. 2013 Apr 22;8(4):e61217. doi: 10.1371/journal.pone.0061217

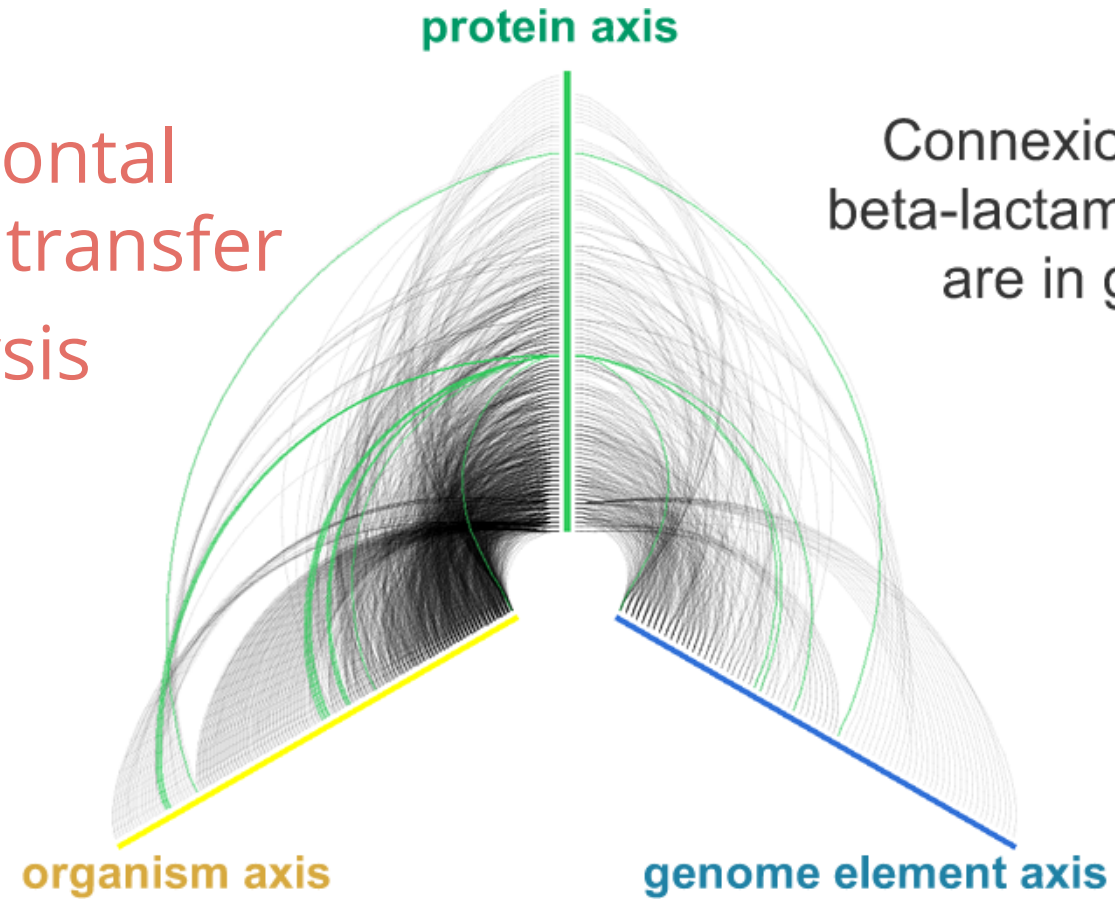


Microbiome studies

[phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.](#) McMurdie PJ, Holmes S. PLoS One. 2013 Apr 22;8(4):e61217

Comparative genomics

Horizontal gene transfer analysis

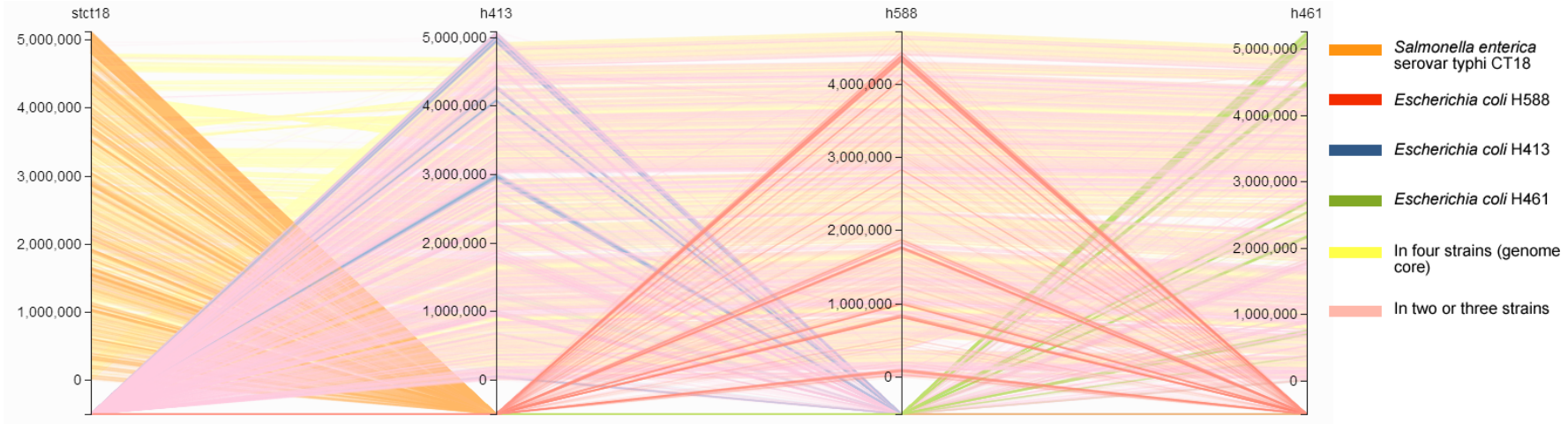


Connexions of beta-lactamases are in green

Era7 bioinformatics
ohnosequences!

Comparative genomics

Study of orthologs using parallel coordinates graphs

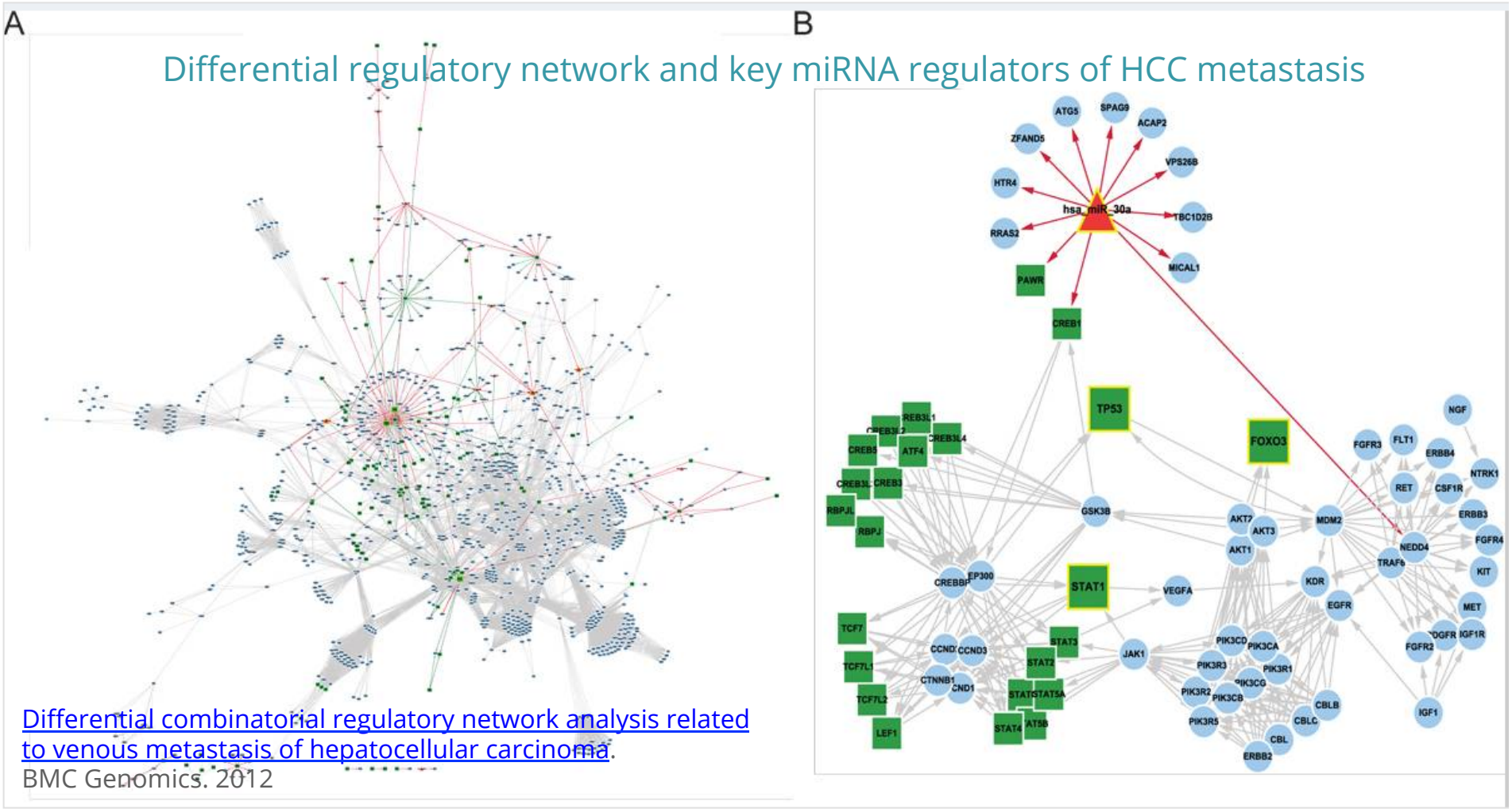


owing all 7133 rows

e	h588	h413	h461	stct18	GO	function	GO id	Pathway	Protein f...	Keywords	Subcellul...	Organism	EC numb...	InterPro	Length	PubMed ...
0294...	5690	2621	38007	296011	cytoplasm	0	GO:0005...	0	UPF0294...	Cytoplasm	Cytoplas...	Escheric...	0	IPR0051...	274	0
bra...	8769	5698	41084	299115	carbon-...	FUNCTI...	GO:0016...	0	Transgly...	Cell mem...	Cell mem...	Escheric...	4.2.2.n1	IPR0233...	452	12471157
oxy...	9596	6525	41911	299942	glutathio...	FUNCTI...	GO:0006...	Seconda...	Metallo-...	Complet...	0	Shigella ...	3.1.2.6	IPR0012...	251	16275786
hyltr...	9630	6559	41945	299977	methyltr...	0	GO:0008...	0	0	Methyltr...	0	Shigella ...	0	IPR0132...	240	0
nucl...	10816	7745	43131	301163	cytoplas...	FUNCTI...	GO:0005...	0	RNase H...	Complet...	Cytoplas...	Klebsiell...	3.1.26.4	IPR0228...	155	0
pol...	10881	7810	43196	301227	DNA bin...	FUNCTI...	GO:0003...	0	0	Complet...	0	Salmonel...	2.7.7.7	IPR0060...	243	1167760.
ein ...	23637	2748880	2918313	309016	ATP bind...	FUNCTI...	GO:0005...	0	ClpA/clp...	ATP-bin...	Cytoplas...	Pseudom...	0	IPR0035...	902	1098404.
0012...	44873	67787	74908	361834	hydrolas...	0	GO:0016...	0	UPF0012...	Complet...	0	Escheric...	3.5.-.-	IPR0030...	256	9278503.
-co...	47987	70901	78022	364387	acyl-Co...	FUNCTI...	GO:0003...	Lipid me...	Acyl-Co...	Complet...	0	Escheric...	1.3.99.-	IPR0060...	814	1120655.
pho...	48227	71141	78262	364627	D-glycer...	FUNCTI...	GO:2001...	Carbohy...	SIS famil...	Carbohy...	Cytoplas...	Escheric...	5.3.1.28	IPR0045...	192	18676672

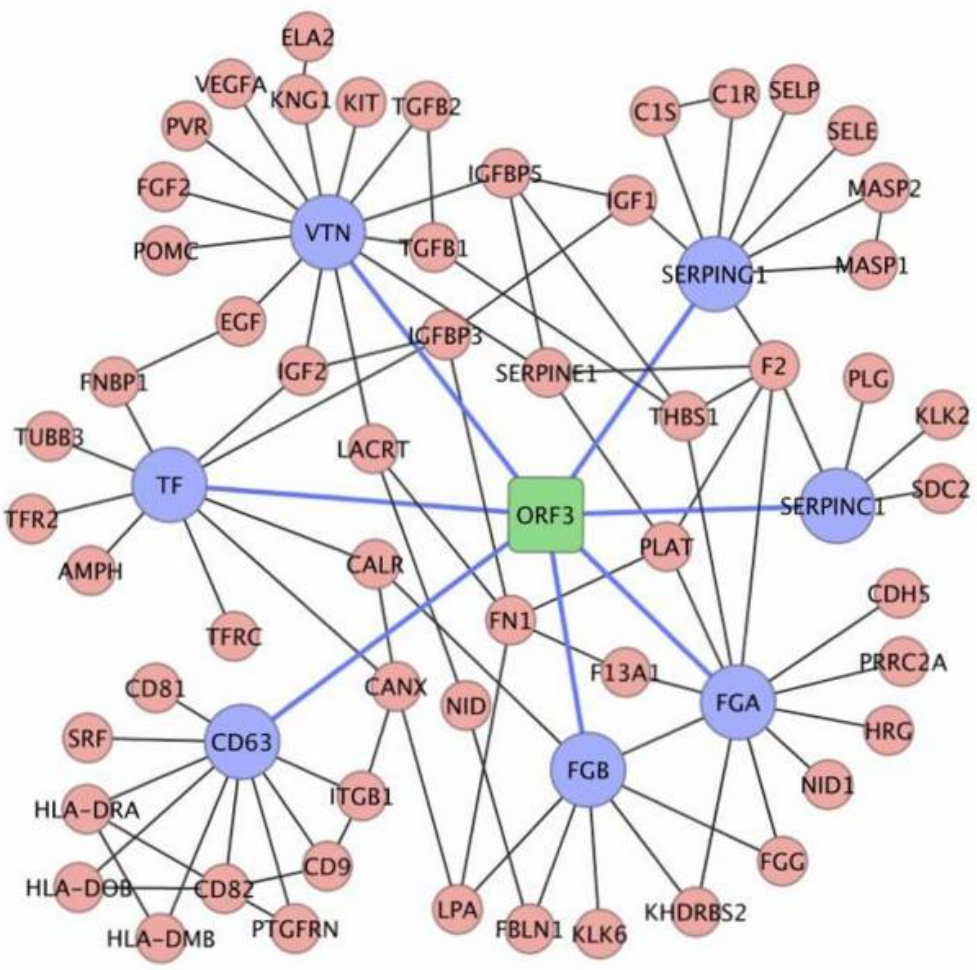
Network analysis

Regulatory networks: analyzing complex relationships



[Differential combinatorial regulatory network analysis related to venous metastasis of hepatocellular carcinoma](). BMC Genomics. 2012

Network analysis

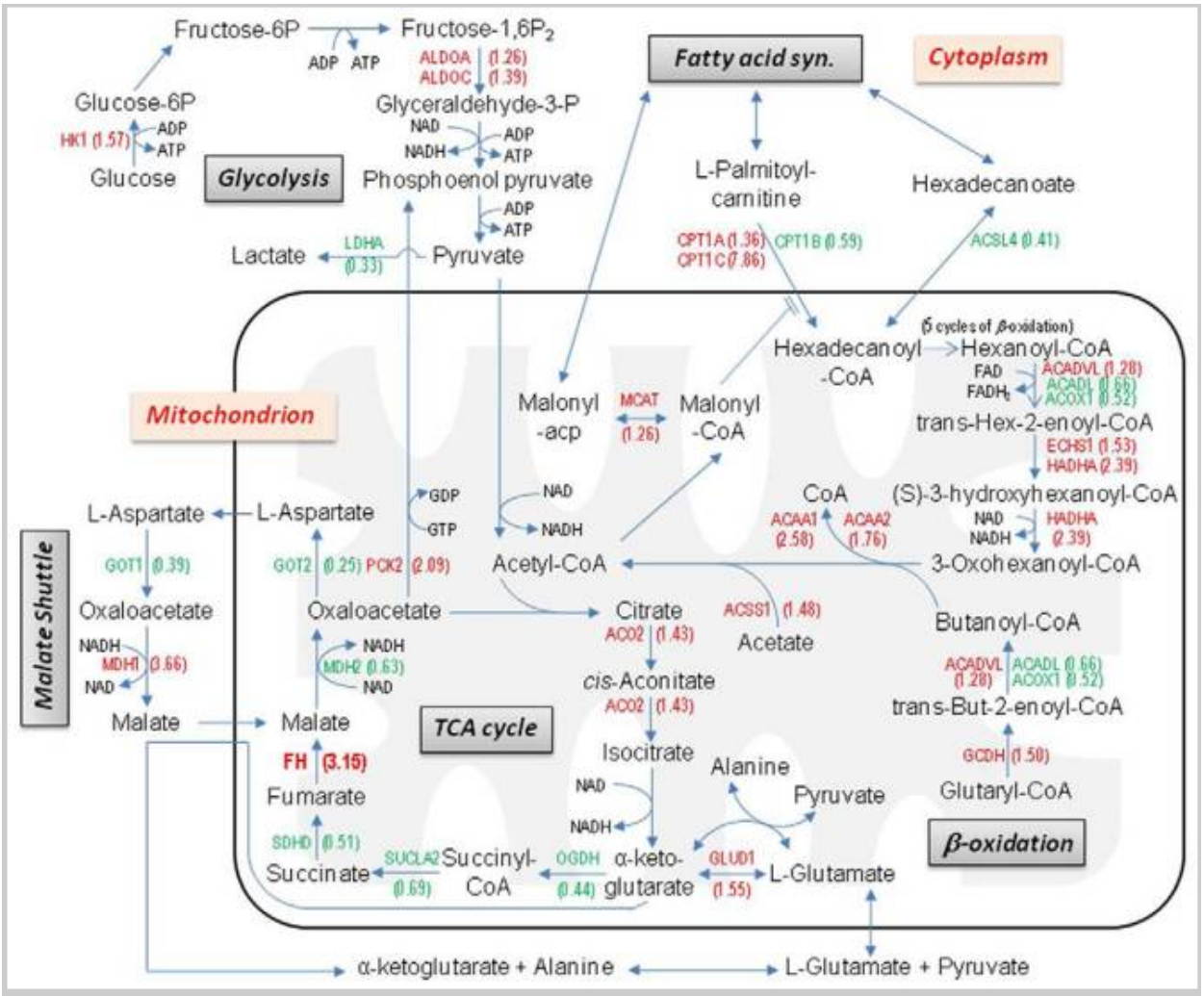


Protein interaction networks

Interactions between HEV ORF3 interacting proteins and host proteins associated to “Hemostasis”

Virus host protein interaction network analysis reveals that the HEV ORF3 protein may interrupt the blood coagulation process.
Geng Y, Yang J, Huang W, Harrison TJ, Zhou Y, Wen Z, Wang Y.
PLoS One. 2013;8(2):e56320

Network analysis



Metabolic pathways

Metabolic pathways with dysregulated genes in renal cancer cell line UOK268 compared to the normal renal epithelial cell line HK-2

[A novel fumarate hydratase-deficient HLRCC kidney cancer cell line, UOK268: a model of the Warburg effect in cancer.](#)
Yang Y et al.
Cancer Genet. 2012 Jul-Aug; 205(7-8):377-90

cloud for NGS data analysis

"Moore's law says that computing power and storage capacity doubles every 18 months, whereas the volume of new sequence data has grown tenfold every year since 2002"

Cloud computing can help to avoid the widening gap between sequence data generation and computing power

cloud for NGS data analysis

"Efficient means for storing, searching and retrieving data are of foremost concern as they are necessary for any analysis to proceed"

"Efficient processing, storage and retrieval of large scale sequencing data sets are crucially important for modern 'big-data-driven' life science"

Cloud computing is especially well-suited for the development of the new 'big-data-driven' life science

Computational solutions for omics data.
Berger B *et al.,* Nat Rev Genet. 2013

cloud for NGS data analysis

"Biological data are exploding, both in size and complexity. High-throughput instruments are now routinely used in individual laboratories around the world in basic science applications as well as in efforts to understand and treat human disease. This trend towards the democratization of genome-scale technologies means that large data sets are being generated and used by individual bench biologists."

For anyone to extract biological insights from these data sets, familiarity with increasingly sophisticated computational techniques is required

Computational solutions for omics data.
Berger B *et al.,* Nat Rev Genet. 2013

INTER • CROSSING

High Throughput Technologies

cloud computing

High quality research
New biological insights

Era7 bioinformatics
ohnosequences!