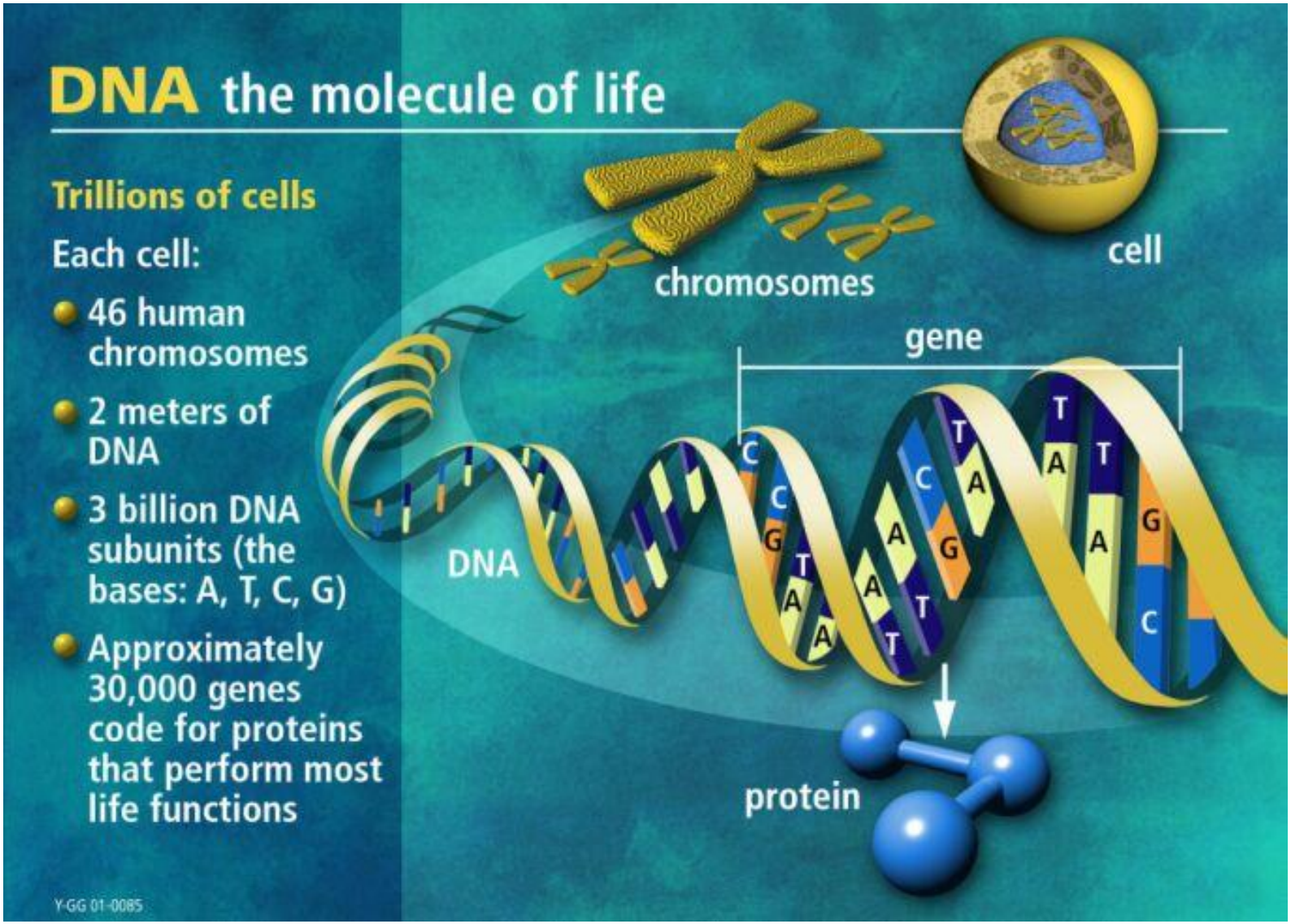
Bioinformatics pipelines for environmental genomics sequencing data.

Julien Tremblay, PhD
julien.tremblay@nrc.ca

National Research Council Canada | Conseil national de recherches Canada
NRC · CNRC

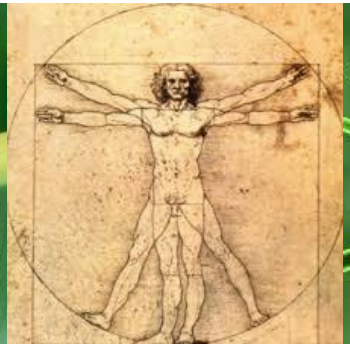
Genome to life



www.flyfishingdevon.co.uk

Genomics





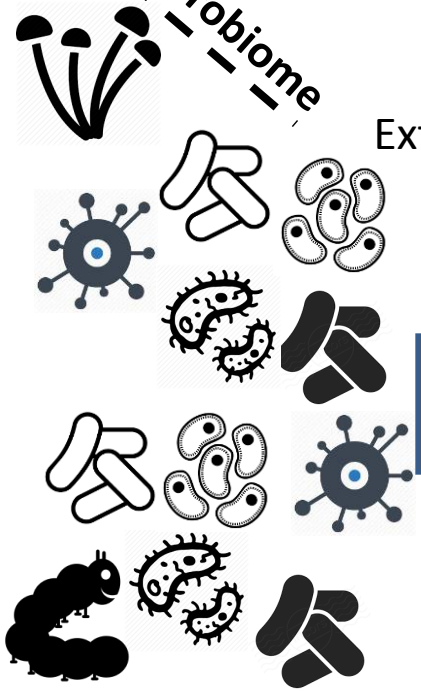

Analyze: find mutations, investigate diseases, metabolic pathways, functional domains etc.

Extract DNA

Sequence

Construct full genome from Sequencing data.

Environmental genomics (aka metagenomics): sequence everything



Extract DNA

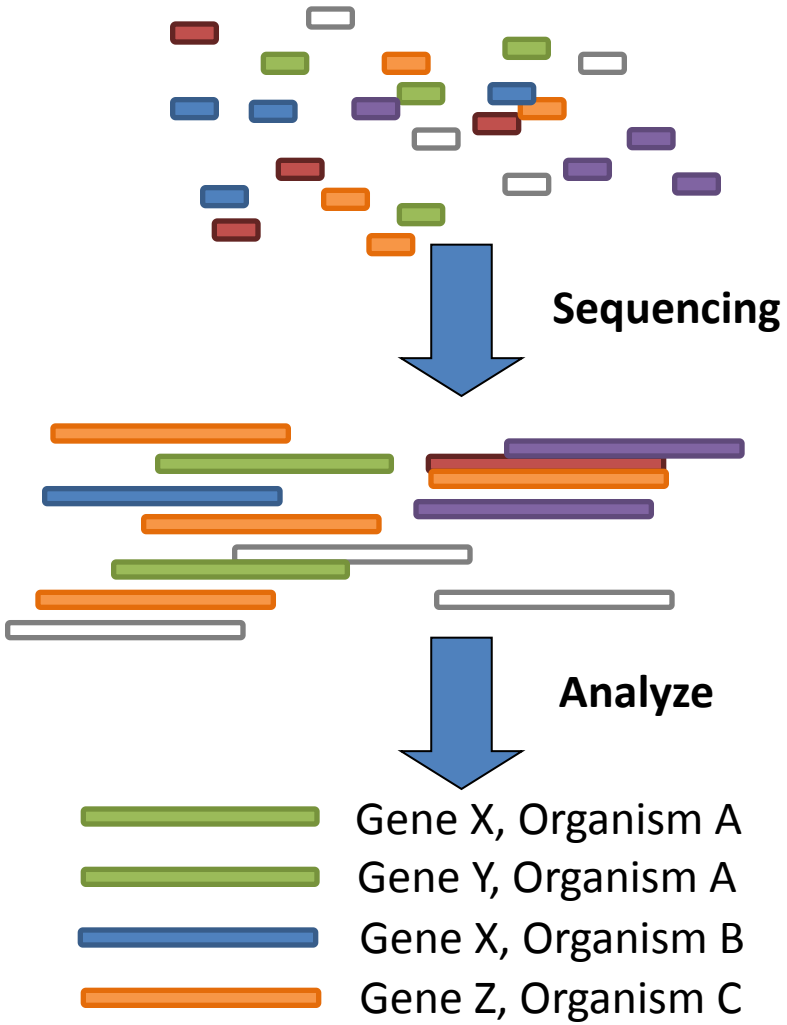
Sequence

Who`s there?
What are they doing?
Who`s doing what?

Microbiome therapy gains market traction, *Nature*, 2014, 509, 269-270

Metagenomics = Modern jigsaw puzzle

Environmental samples =
1000s of microorganisms
~5Mbp each, 5,000 genes



Millions of reads of… 100 bp

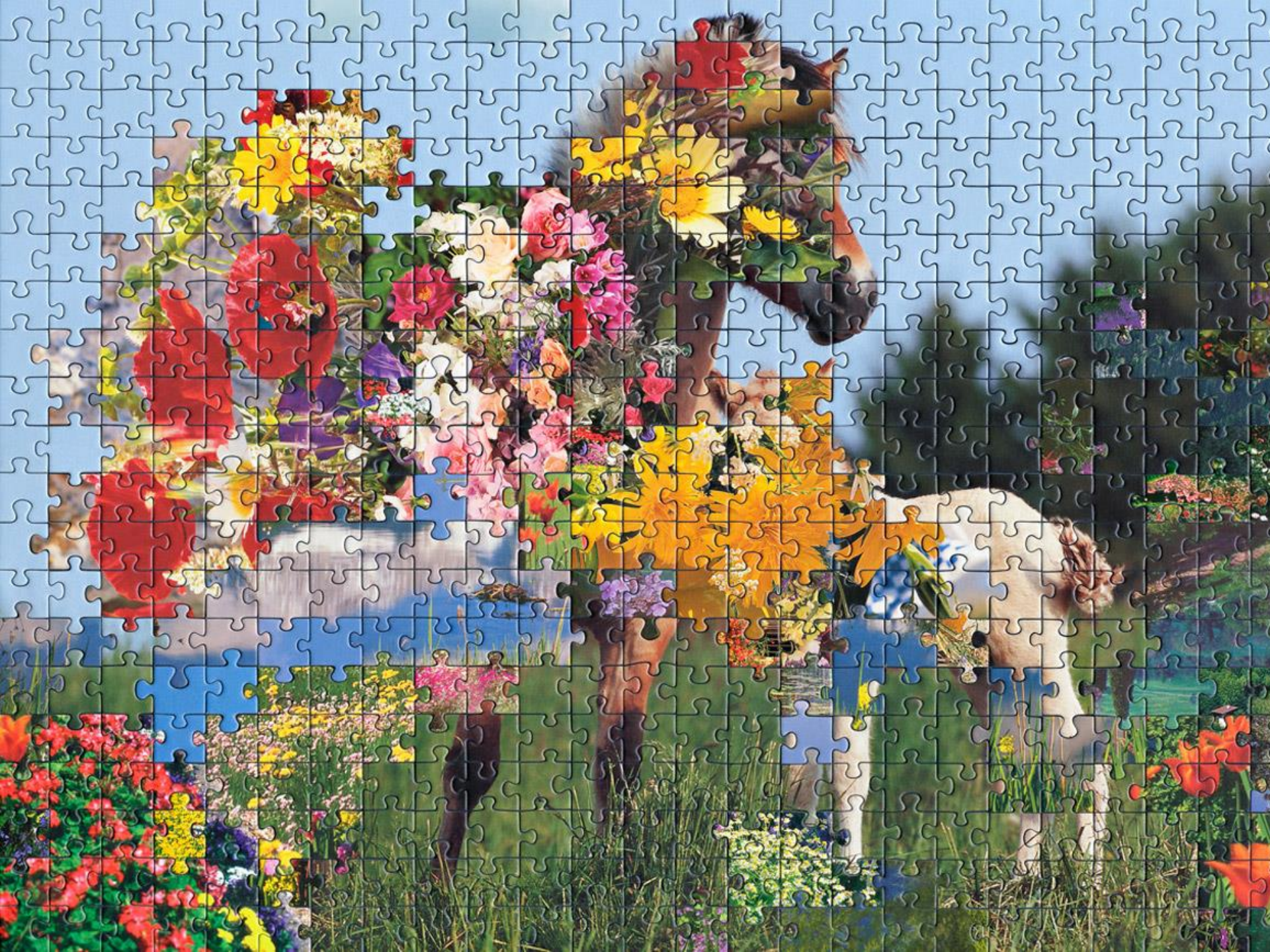
Assembly

Comparison to **known** genes / organisms

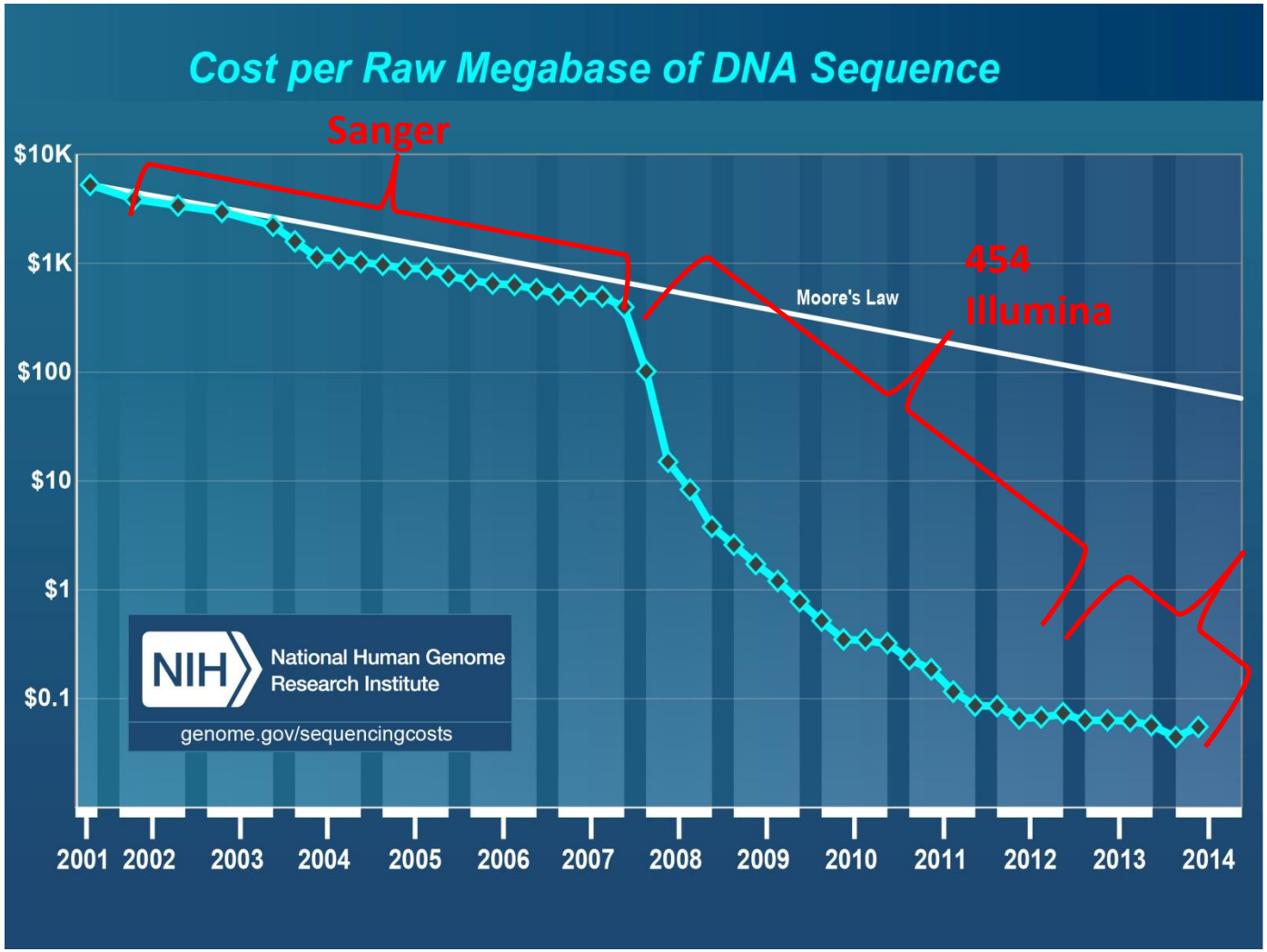
What is bioinformatics?

Field that develops methods and software tools for understanding biological data. Bioinformatics combines computer
science, statistics, mathematics and engineering to study and process biological data.

Biology ⟷ Mathematics
Advanced algorithmic



Rise of high throughput DNA sequencing



Illumina
PacBio
Oxford nanopores
Ion Torrent

Sequencing data deluge ahead

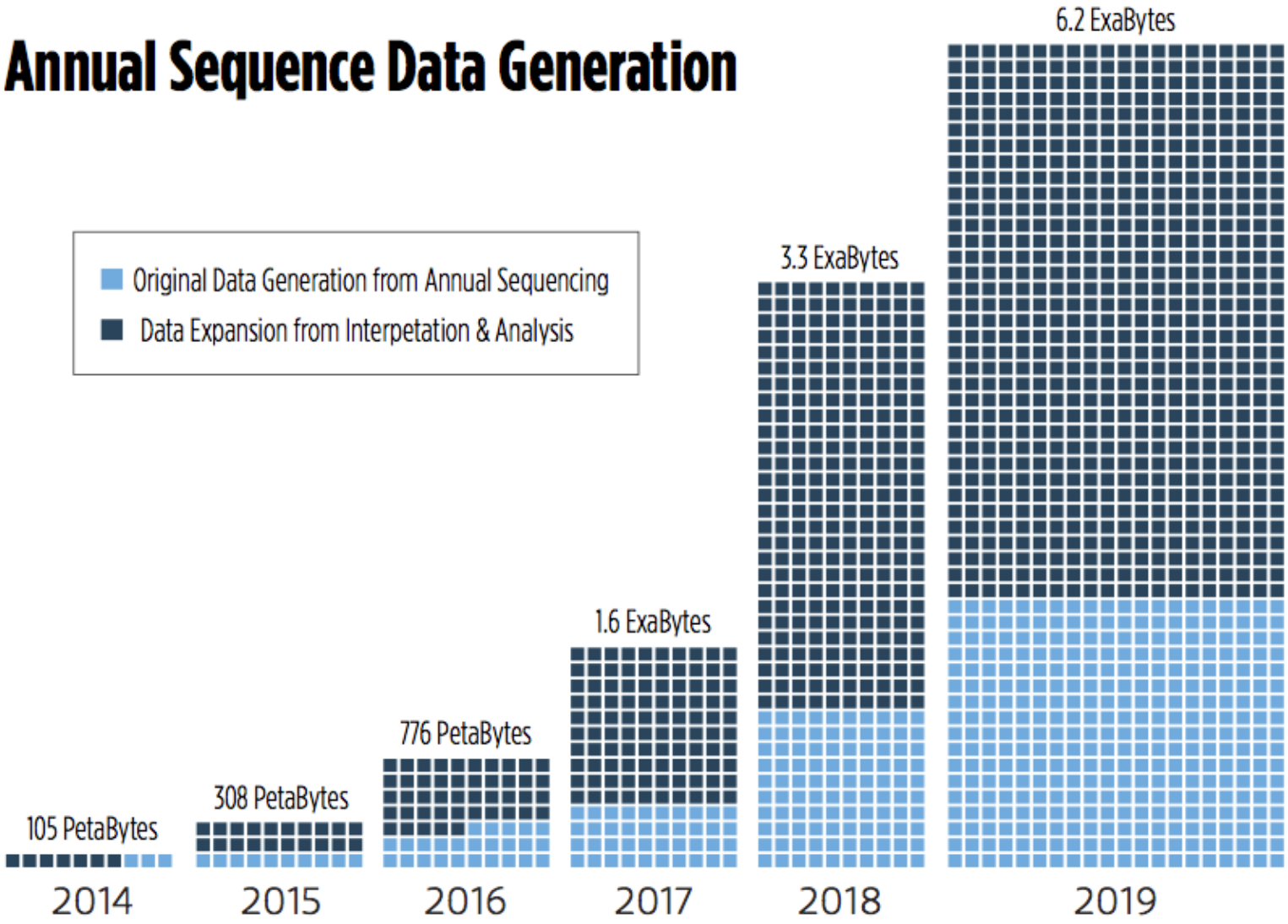


Image: www.onrampbioinformatics.com

1000$ genome... ?

Illumina Hiseq4000

Modern sequencers

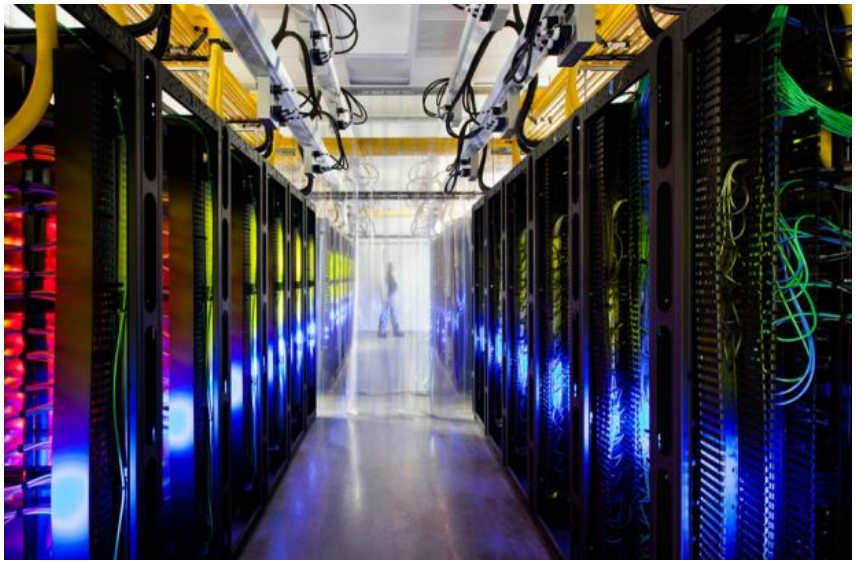
Sequencing cost = cheap!

TBytes of raw data

Data analysis is expansive!

PacBio RSII

Software



Storage

Compute nodes

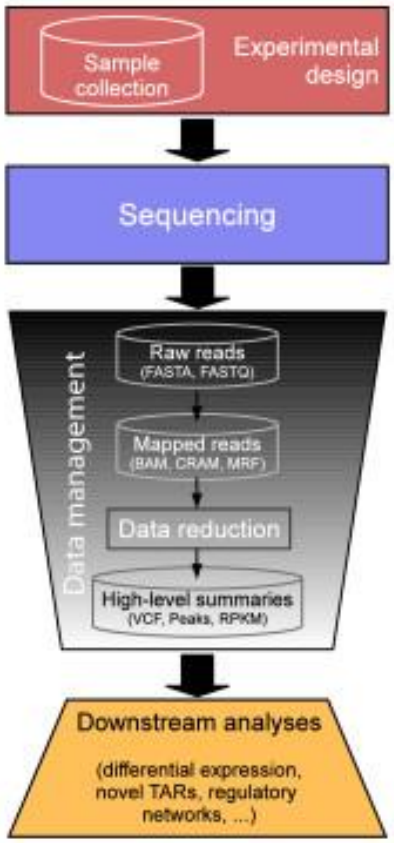
Hardware maintenance

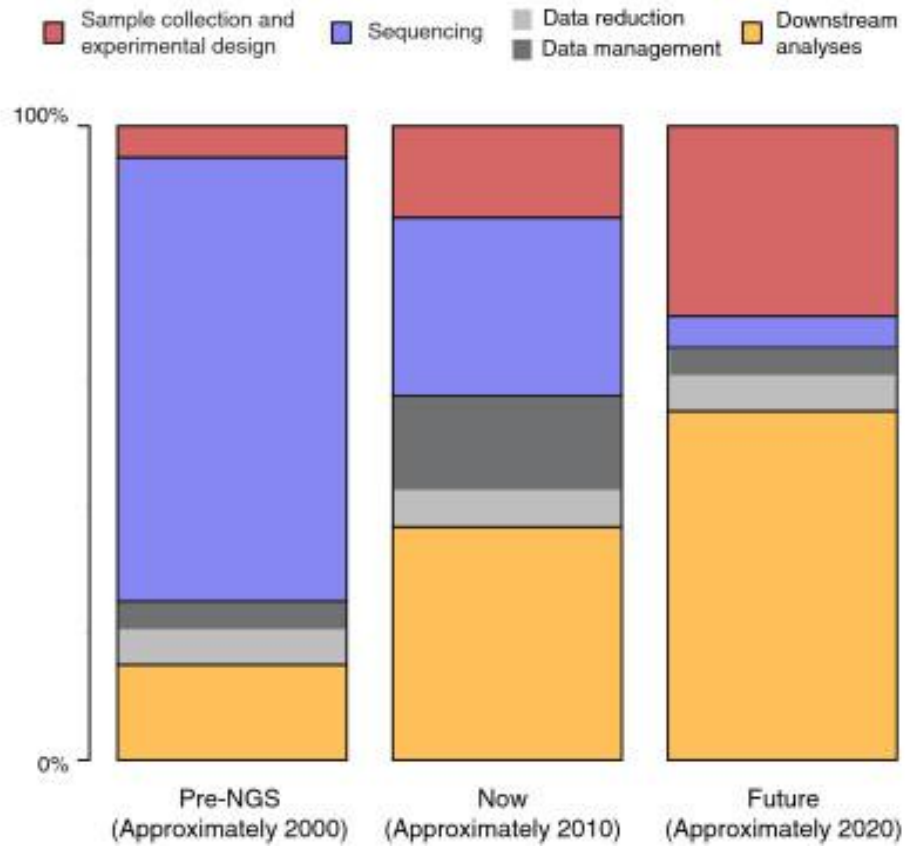
Electricity

Network infrastructure

The 1,000$ genome, the 100,000$ analysis?, Genome Medicine, 2010, 2:84

Bottleneck





Sequencing DNA is now the easy part

Computational pipelines are increasingly complex

Contain many steps

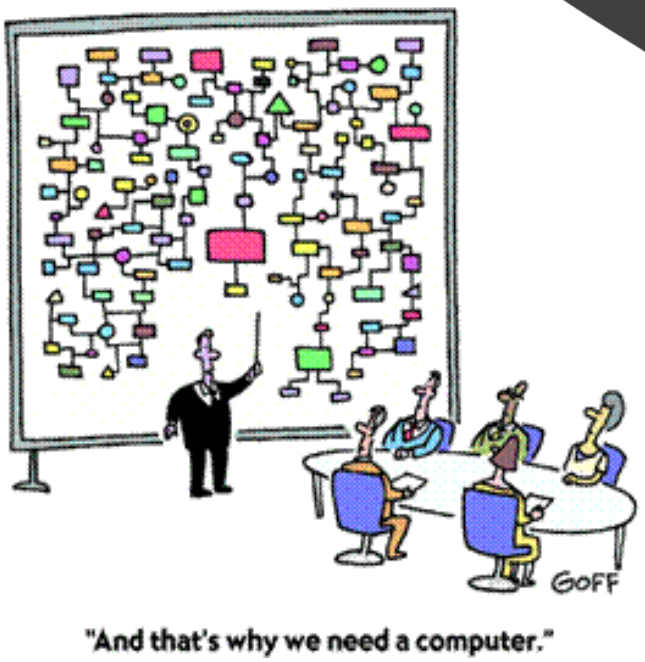
Need for automation

The real cost of sequencing: Higher than you think!, Genome Biology, 2011,12:125

Unintelligible information

Metadata,
chemistry data

Hypotheses
+ experimental design



"And that's why we need a computer."

- What genes are differentially expressed in my data?
- By what organism?
- Correlates with metadata?
- Which genes are co-expressed in variable x vs y? and why?

Knowledge (high level summarized analyses)

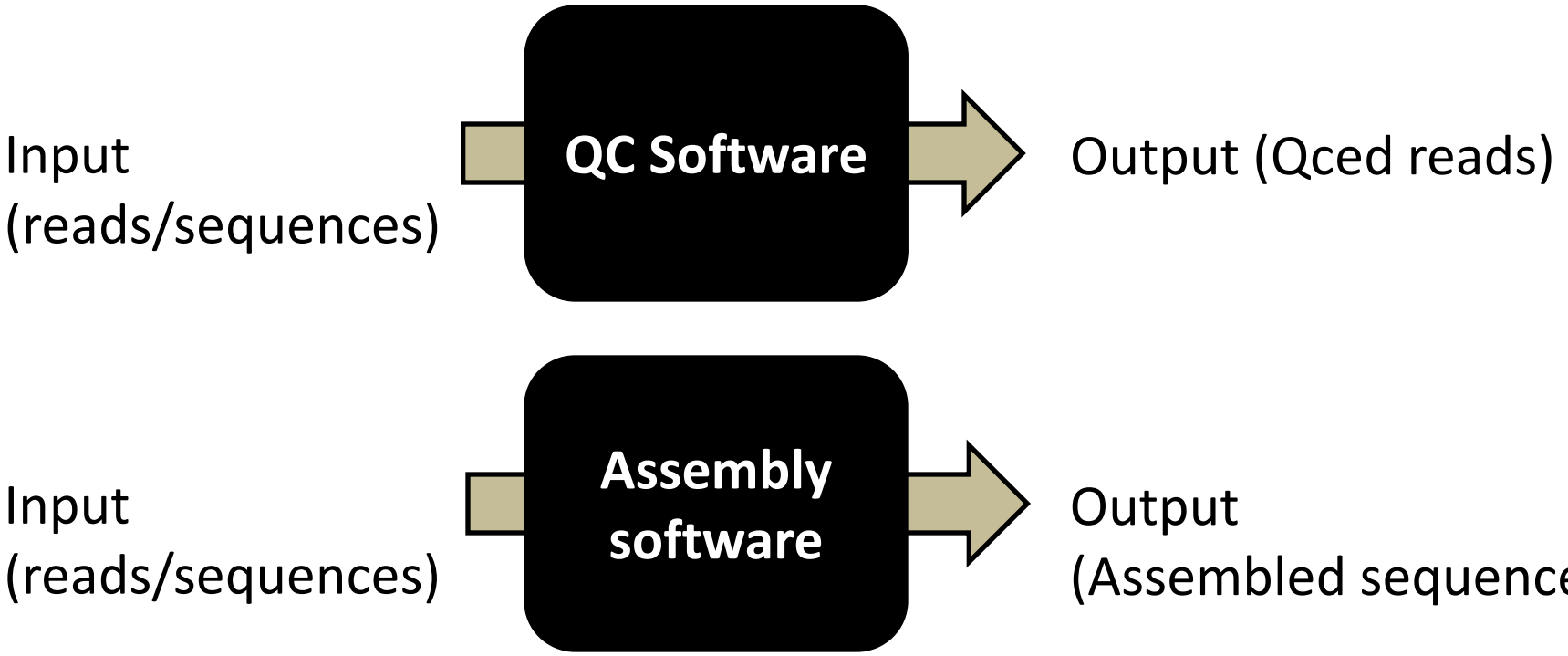
Challenges in bioinformatics

- Bioinformatics software varies greatly in quality.
- A few established core packages are stable
  - Raw reads quality control
  - Reads alignment
  - Reads assembly
  - Clustering
  - Reads sorting
- Software Installation/maintenance = challenging.

Ten recommendations for creating usable bioinformatics command line software. 2013, *GigaScience* 2013, 2:15

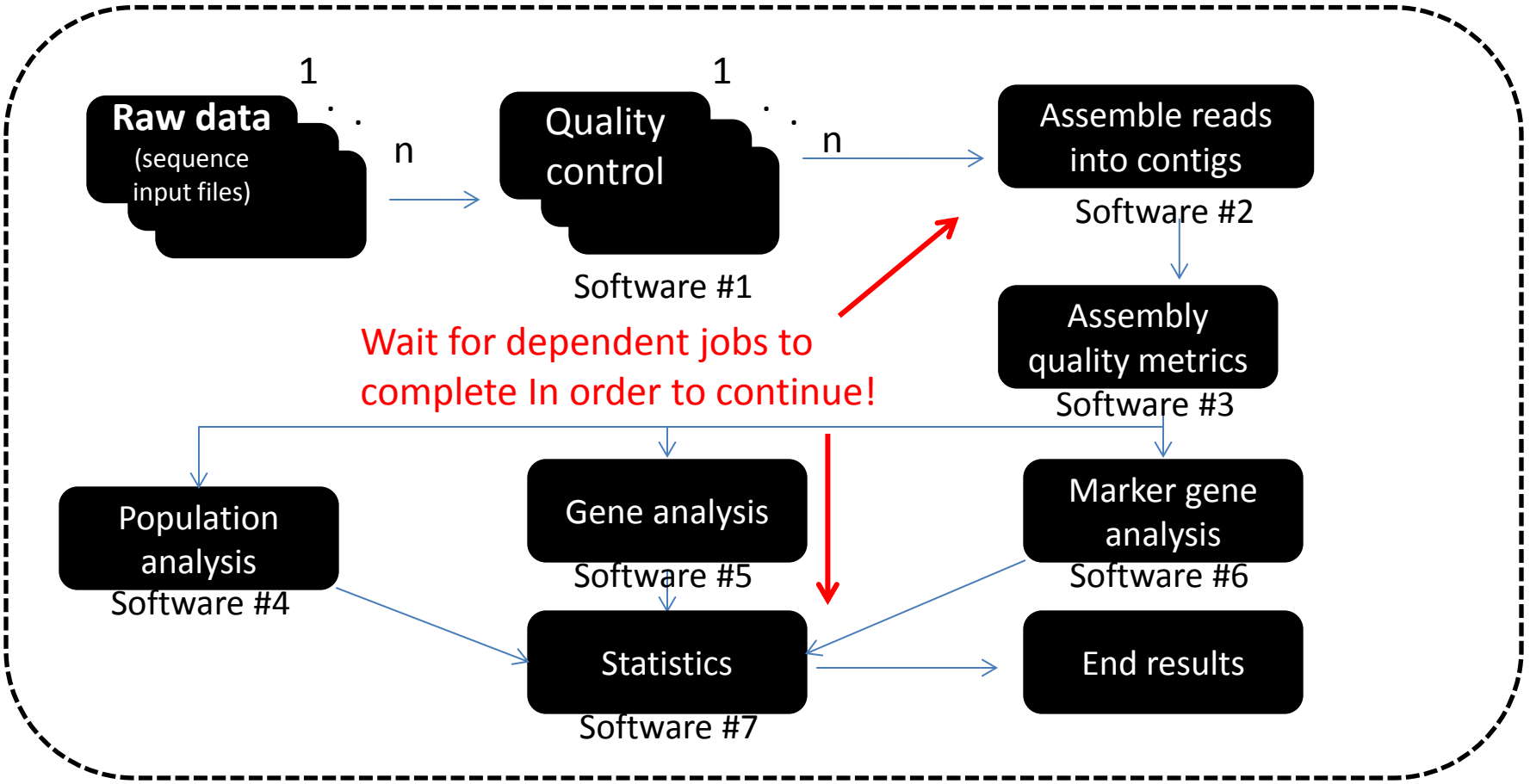
Challenges in bioinformatics

- Using a bioformatics package is easy…



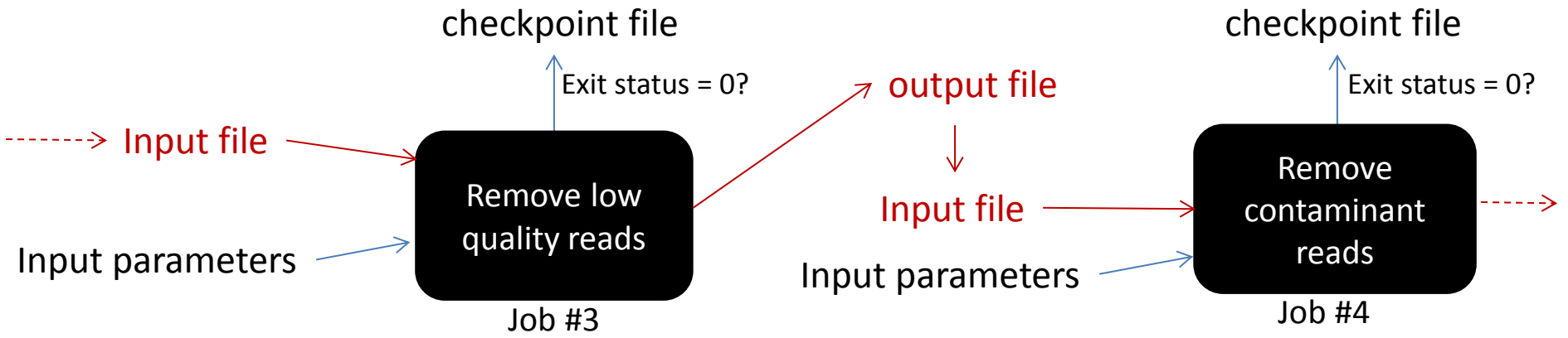
Challenges in bioinformatics

- Executing bioformatics packages in a specific order is a little harder…



Pipeline wrapper module

- McGill University and Genome Quebec Innovation Centre's pipeline module. https://bitbucket.org/mugqic/mugqic_pipelines
- Generates PBS jobs + manage their dependencies + Smart restart mechanism in case of job failure.



Stats on production pipelines

	Marker genes pipeline	Shotgun assembly + annotation pipeline. (i.e. metagenomics)	Shotgun metatranscriptomics pipeline
Input data	1-10 GB	0.5 TB – 2 TB	0.5 TB – 2TB
Number of jobs	**115**	**~9,000**	**~2,000 jobs**
Size intermediate files	4 to 20 GB	3 TB to 10 TB	1.5 TB to 5 TB
CPU time (core hours)(cummulative)	~70 hrs	~25,000 hrs	~5,000 hrs
RAM (cummulative)	~30 GB	~25 TB	~20 TB
Number of third party packages	22	43	35

Generate highly summarized data

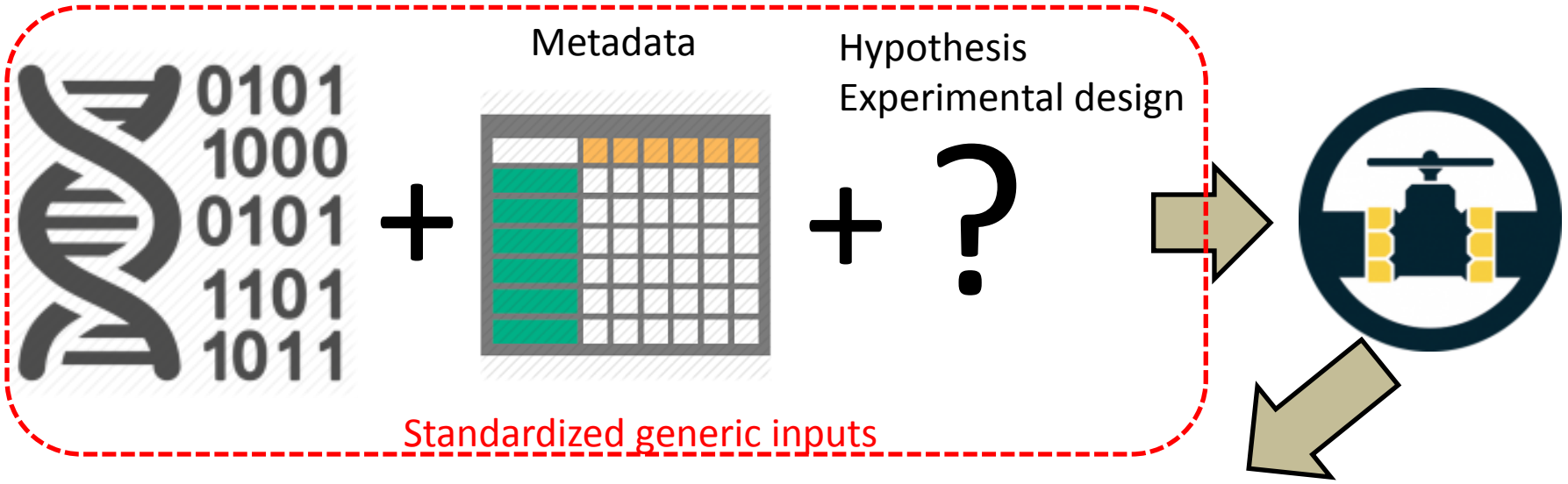
- Assembling reads into contigs is great, but ≠ end results…
- <u>Need metadata for each sequenced sample!</u>

Metadata example for a project investigating Wheat metagenome.

SampleID	Nitrogen conc.	Phosphore conc.	Treatment	Field
Wheat.1	33.5	78.9	1-year rotation	A
Wheat.2	21.7	78.8	1-year rotation	A
Wheat.3	44.8	77.4	2-year rotation	A
Wheat.4	12.3	56.7	2-year rotation	A
Wheat.5	11.3	43.6	1-year rotation	B
Wheat.6	13.5	43.5	1-year rotation	B
Wheat.7	13.5	43.5	2-year rotation	B

Clear answers

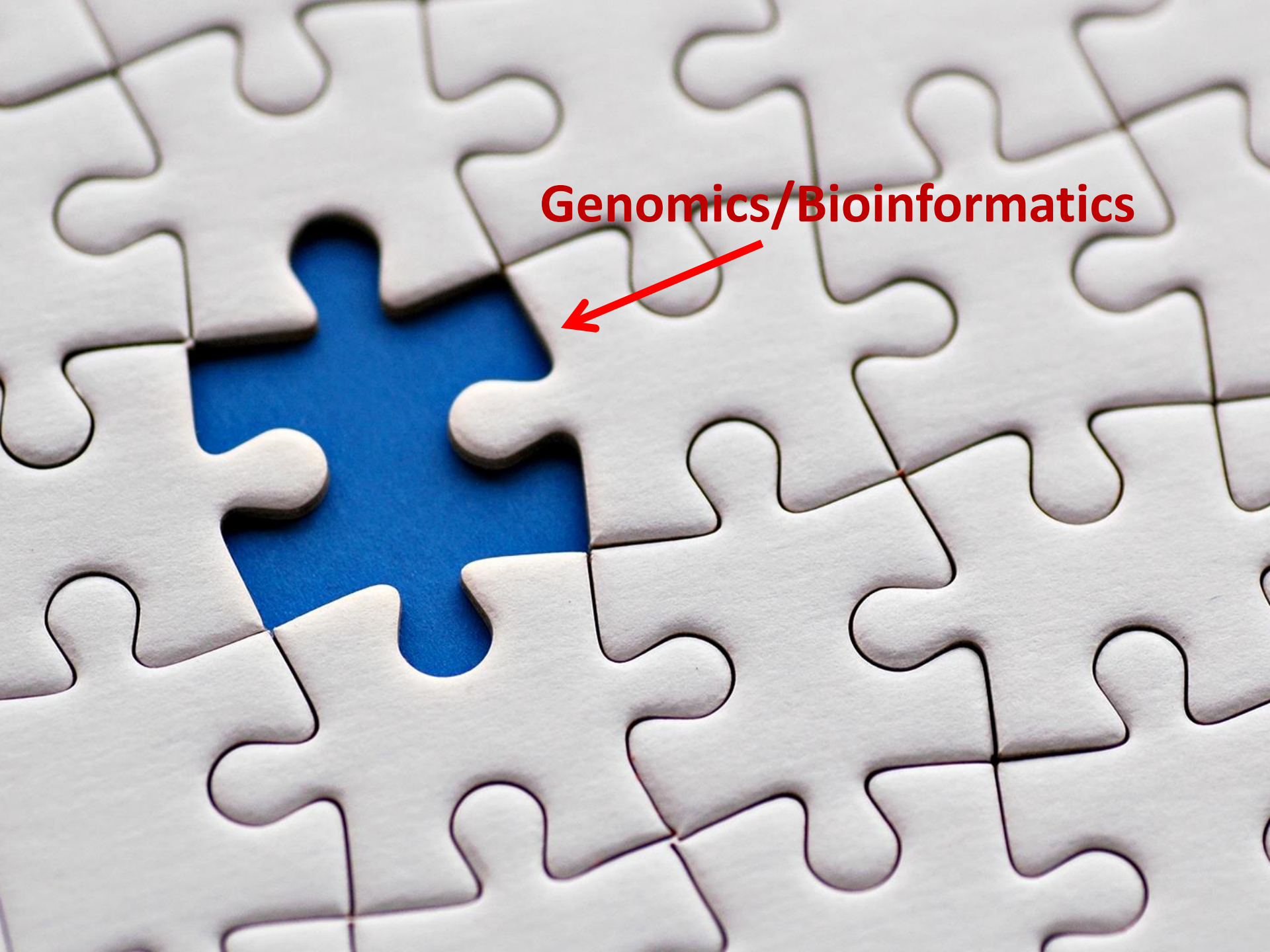
- Assembling reads into contigs + annotation is great, but ≠ end results…
- <u>Need metadata for each sequenced sample!</u>



Treatment #1 significantly enhanced Wheat crop yields.

Treatment #1 was correlated with upregulation of gene x which is known to enhance uptake of nitrogen…

The sequence of gene x shows unusual domain structure…


Genomics/Bioinformatics

Future direction

- Improve integration of metadata with genomics data.
- Develop visualization methods/tools highly dimensional end results. R, D3.js…
- Get ready for next generation sequencers
  - Oxford Nanopores will generate 10 TBytes of data / day XD → Insatiable thirst for compute power and storage!

Acknowledgments

- National Research Council's Biomonitoring group
  - Étienne Yergeau
  - Charles Greer
- McGill University and Genome Quebec Innovation Centre
  - Joel Fillon
  - Louis Letourneau
- McGill HPC