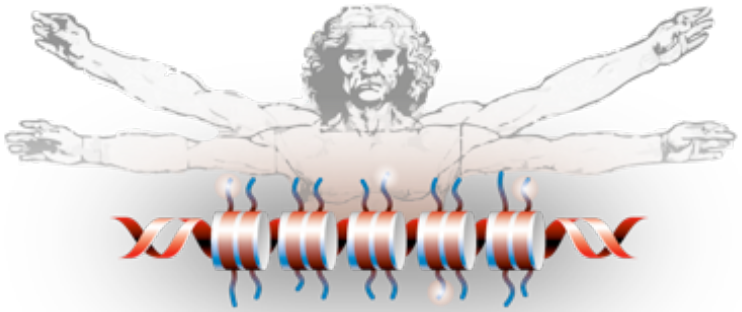
Max Planck Institute of
Immunobiology & Epigenetics
Max Planck Institut für Immunbiologie & Epigenetik

UNI FREIBURG

# Galaxy Course

## Sep 21-25 2015

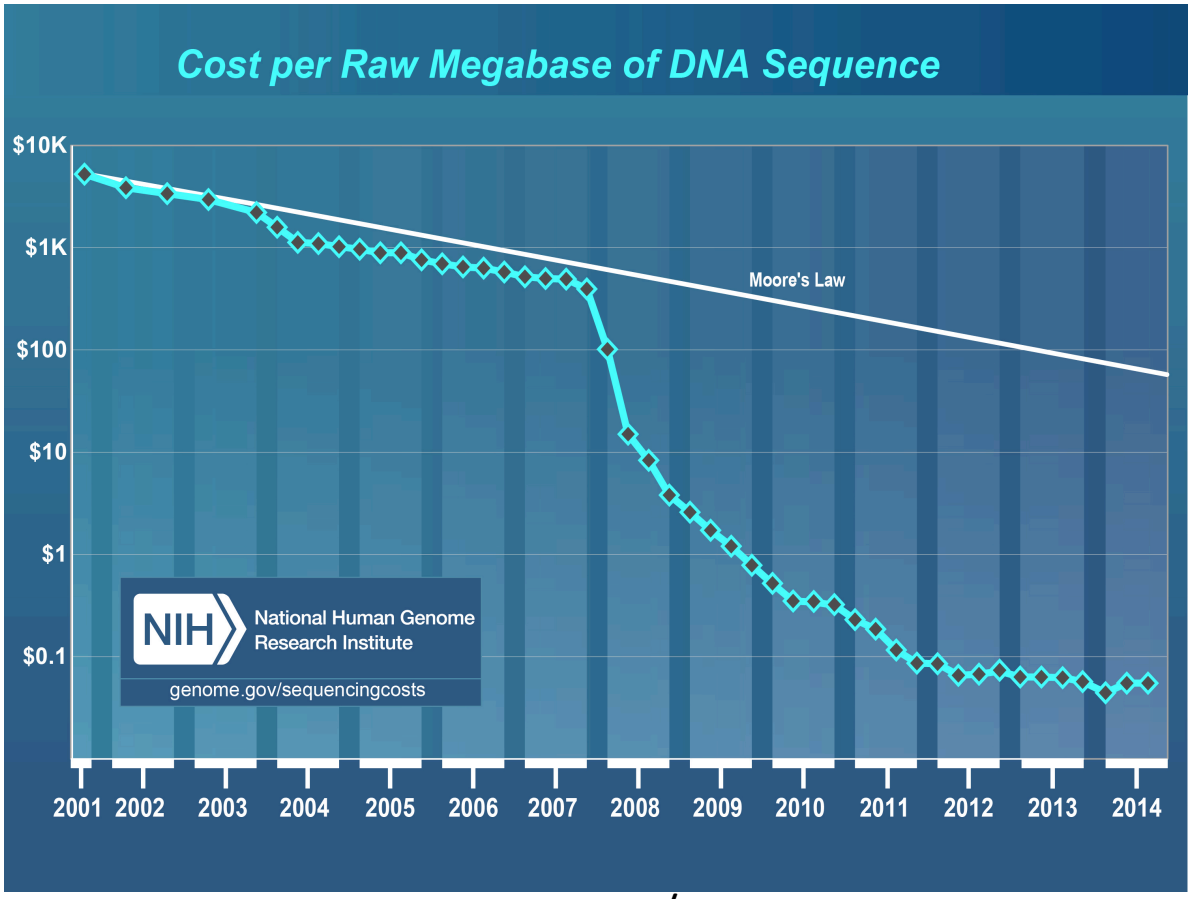
Thomas Manke



# Outline

- Prolog
- Pipelines
- Problems

# Cheap Data & Expensive Analysis



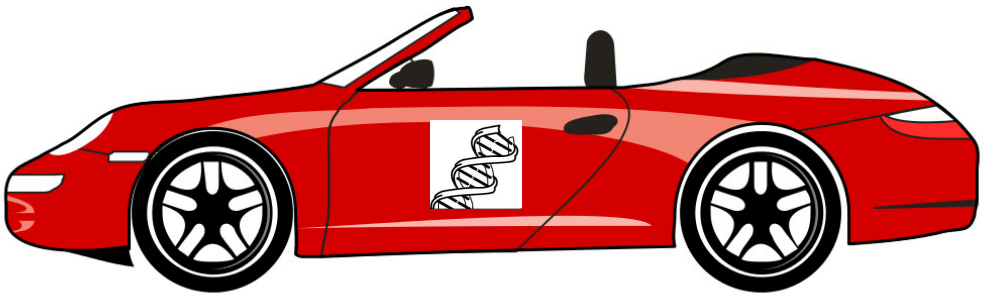

http://www.genome.gov/sequencingcosts/

Weekly output
@ MPI-IE

- $>10^{11}$ bp
- 30 genomes
- 1-2 TB

How to analyze sequencing data at production rate?

# What if …



… the car industry had developed as rapidly as sequencing technology ?

| | Price | Speed |
|---|---|---|
| 2000 | 100,000 Euro | 280 km/h |
| 2015 | 10 cents | 21 Millionen km/h |

# Why do we sequence?

## Reference


| | |
|---|---|
| time | • 1990-2003 |
| bp | • 3,000,000,000 |
| $ | • 3,000,000,000 |

## Variation

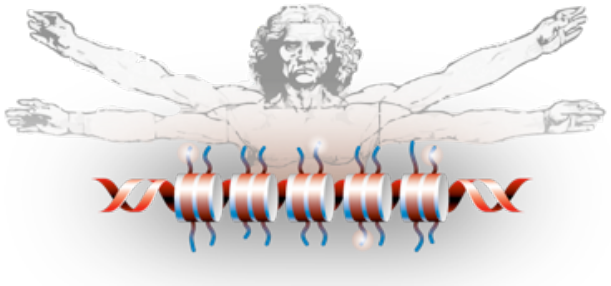

~2500 genomes

| | |
|---|---|
| time | • 2008-2012 |
| bp | • 3,000,000,000,000 |
| $ | • 120,000,000 |

# Why do WE sequence?

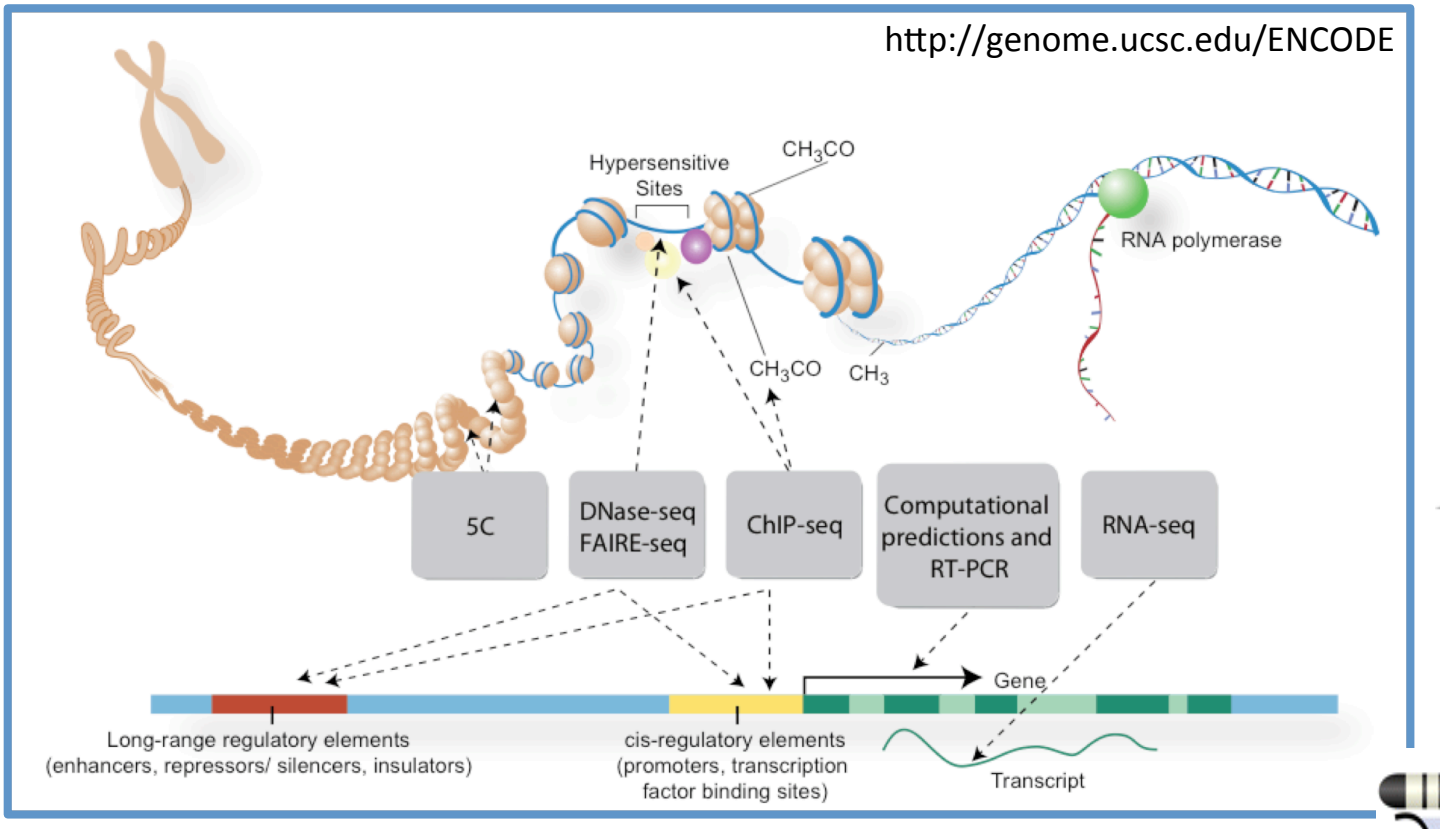
Deutsches Epigenom Programm

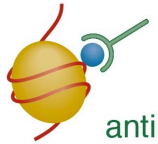
CRC 992: MEDEP



# Genome Architecture

Rules for functional organization and gene expression ?




ChIP

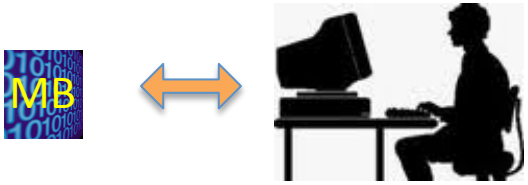
Sequencing

Mapping

Location → sequence counts → enrichments?

# New Demands and Workloads

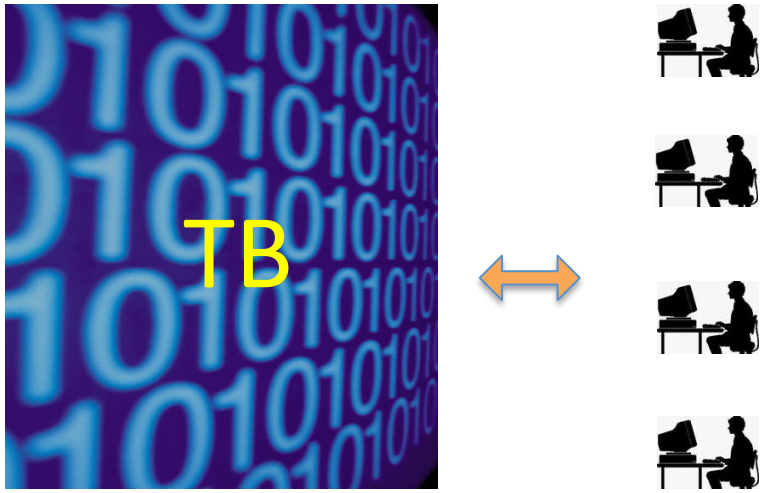
Expensive Data




Method Development

Programming & Algorithmic Comparisons

Expensive Analysis




Data Management

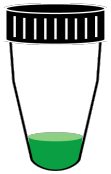
Software Management

Data Reduction & Pipelines

# Pipelines

Data Production &
Data Reduction

# Data Production Pipeline



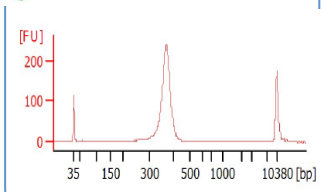
DNA

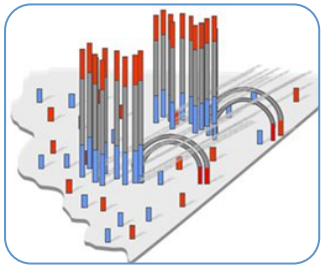
Enzymes
Adapters

Adapter-ligated DNA

Library Preparation

QC: fragment size



Cluster Generation

Sequencing

32 TB

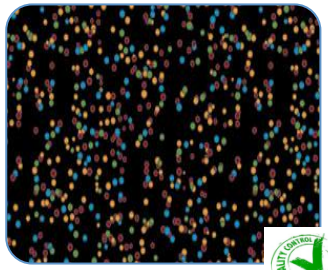


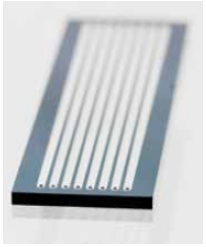
Image Analysis

2 TB

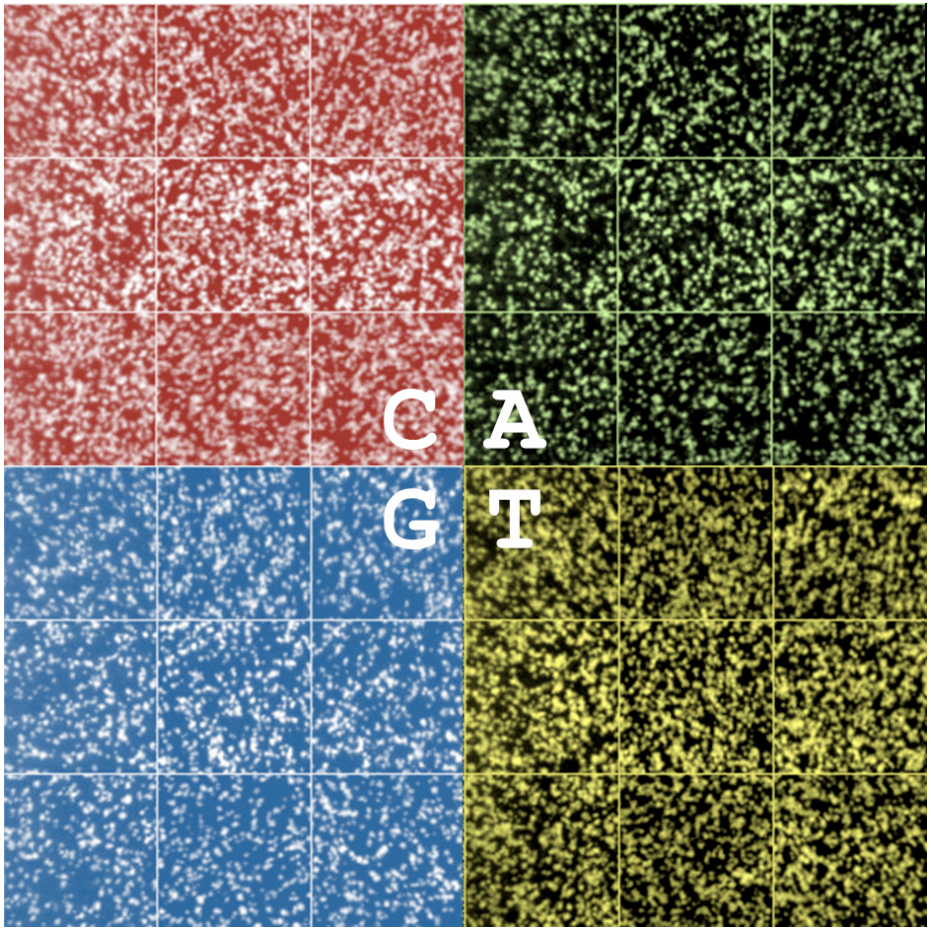
For each sample

Multiplexing: combine samples

8 lanes for m x 8 samples



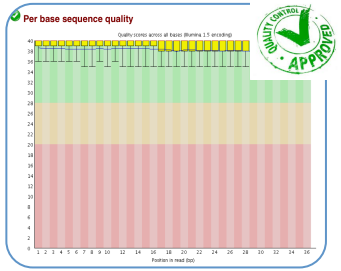
# Data Reduction



# Data Reduction Pipeline

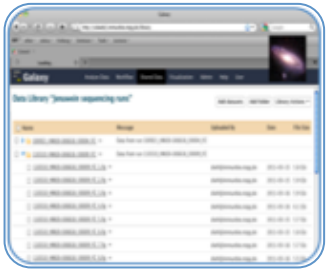
2 TB

Transfer

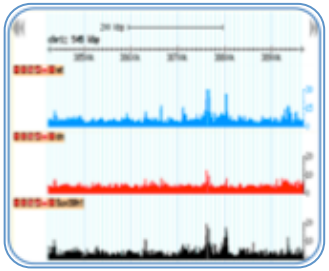



Quality Control

Mapping



Data Repository



Web services

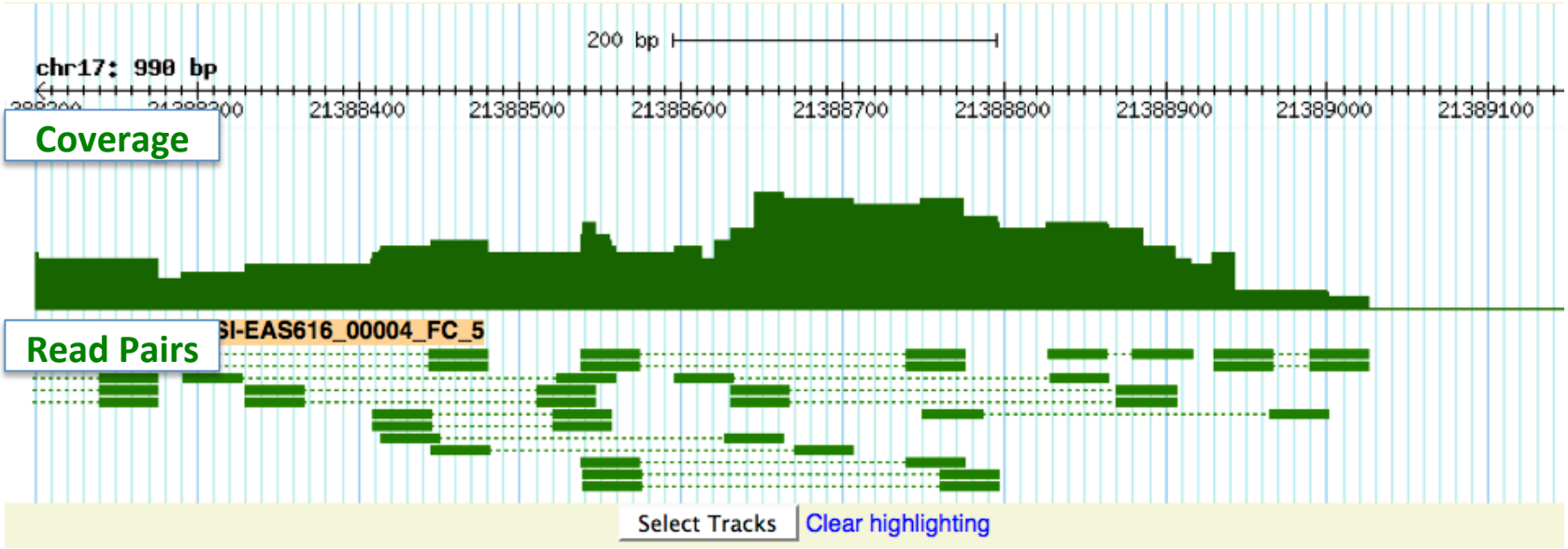
Sequences & Qualities

- Required for publication (fastq file: 25 GB / sample)

Coordinates

- Required for further analysis (bam file: 5 GB / sample)

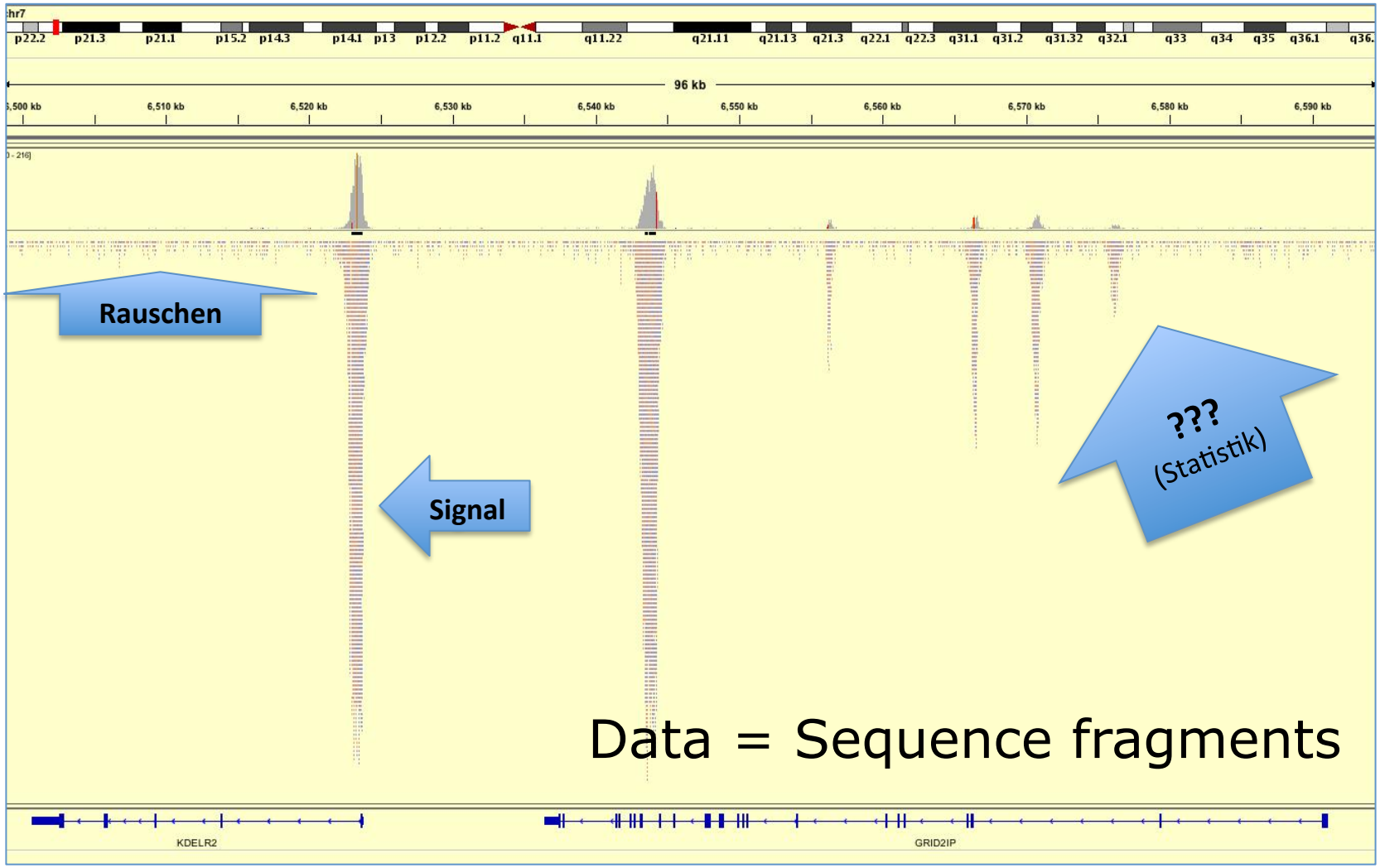
# Visualization of Reads & Coverage




Genome Browsers: **IGV**, UCSC, ModEncode, …

→ Data handling, data formats, inspection/visualization of loci

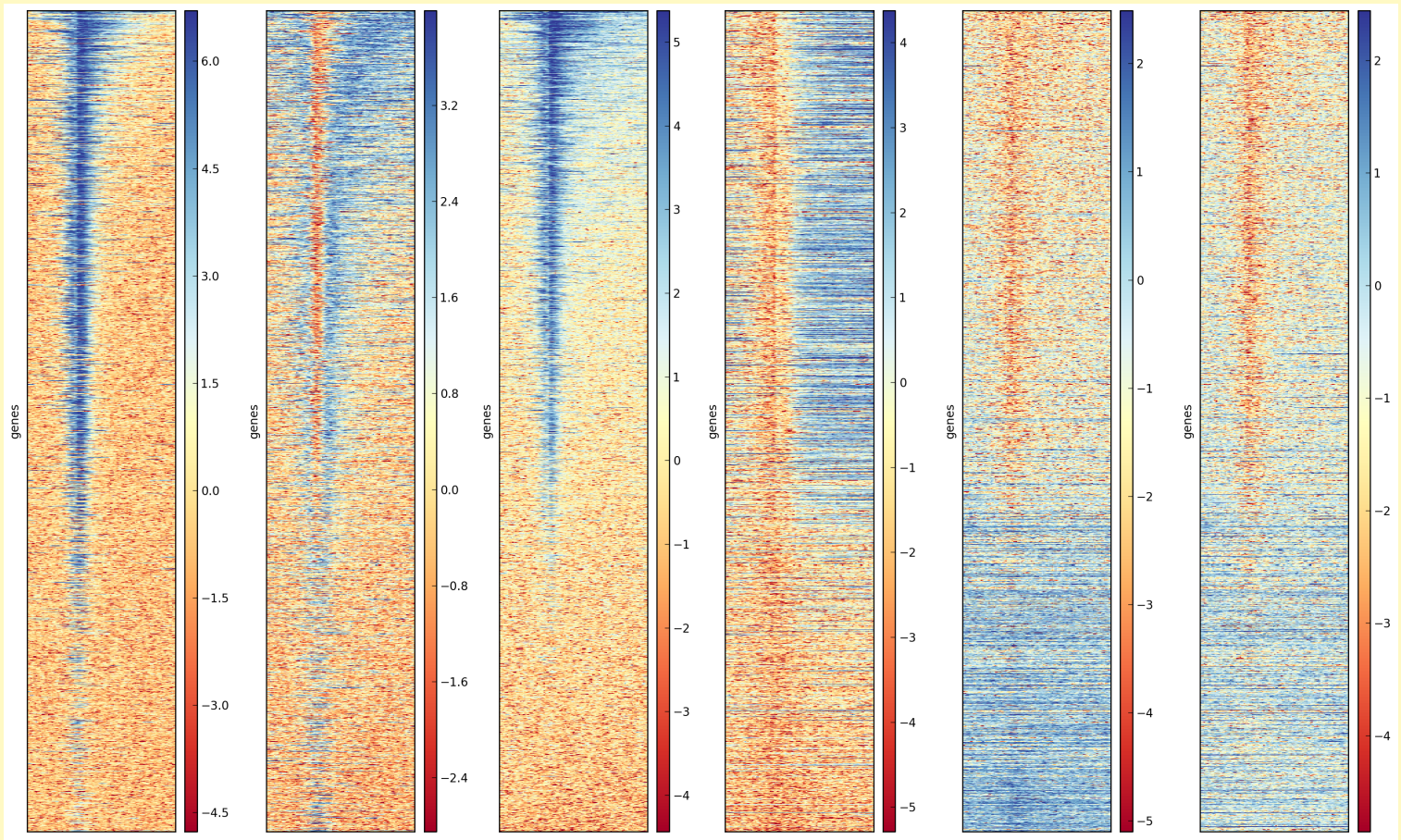
# Signal Detection




More Data (depth,replicates) → more reliable predictions

# Meta-Analysis

deepTools




→ Data integration, genome-wide analysis, *interpretation*

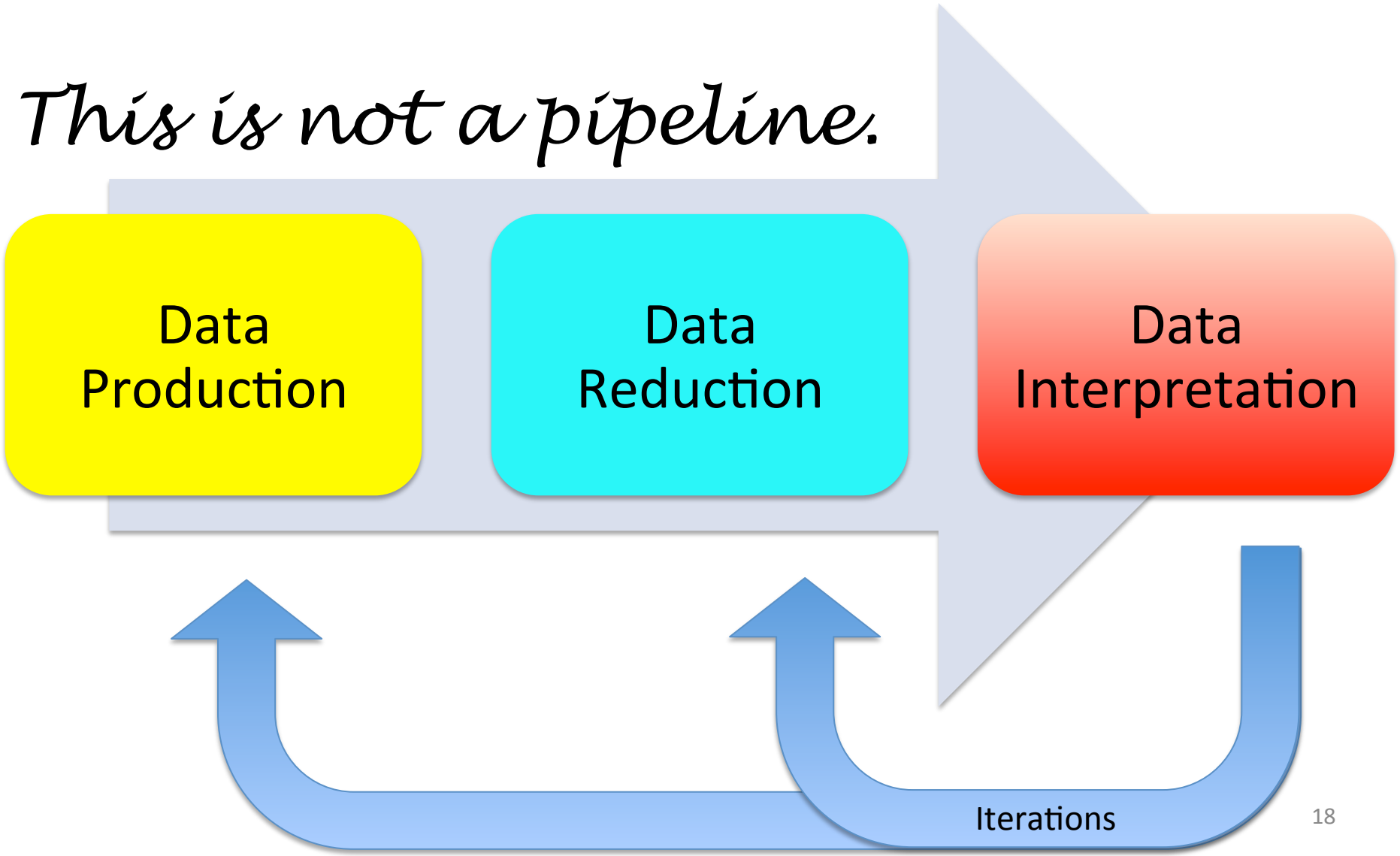
# Project timeline

| | |
|---|---|
| Data Production | • 1-2 weeks |
| Data Reduction | • 1-2 weeks |
| Secondary analysis | • Months!!! |

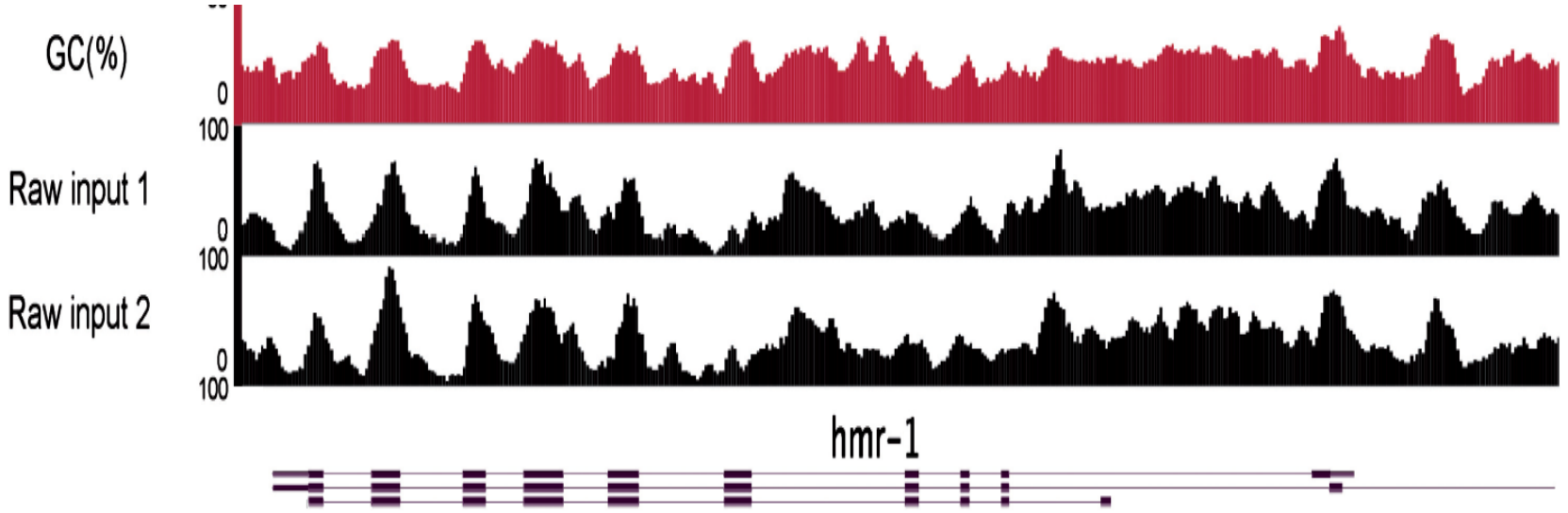
# Problems

# Challenge I

*This is not a pipeline.*



# Challenge II: Biases

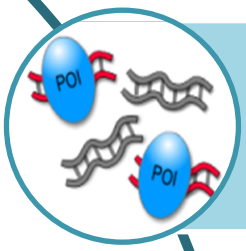



Cheung et al NAR (2011)

need appropriate background models & controls

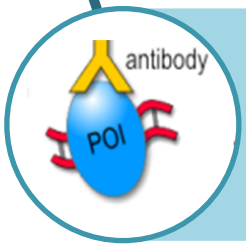
# Biases in ChIP-seq

**Sample heterogeneity?**



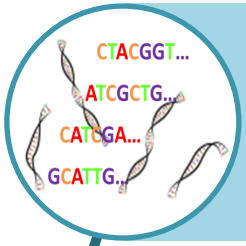

**FRAGMENTATION OF PROTEIN-CHROMATIN COMPLEXES**

- inappropriate sonication conditions
- **open chromatin is favoured**



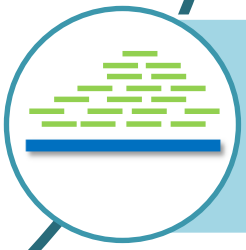

**CHROMATIN-IMMUNOPRECIPITATION**

- insufficient antibody specificity and/or sensitivity
- cross-reactivity or context-dependency




**AMPLIFICATION AND SEQUENCING**

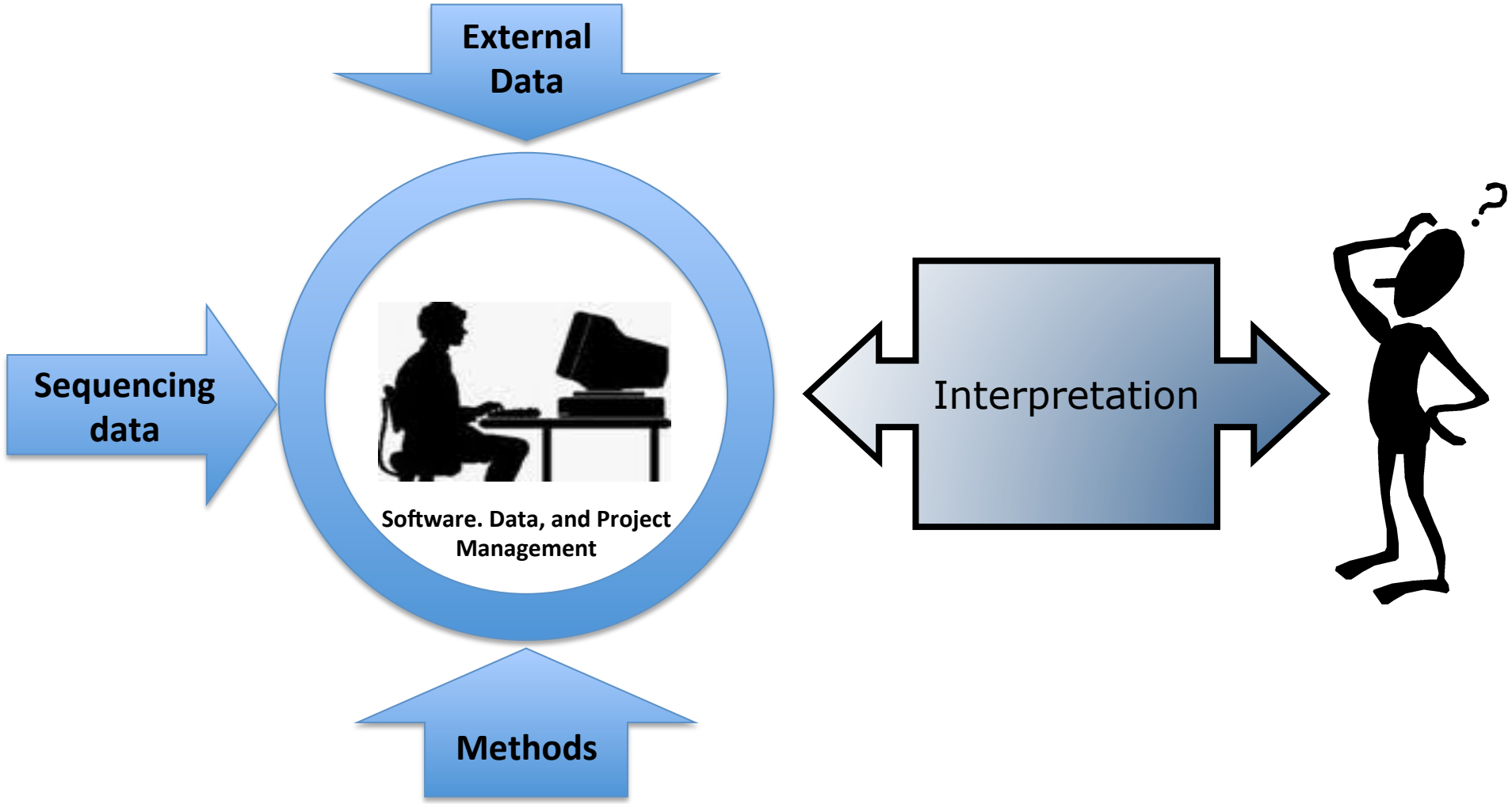
- PCR artefacts (e.g. duplicates)
- **varying amplification efficiency (GC-bias)**
- erroneous base calling



**MAPPING TO REFERENCE GENOME**

- Sequencing depth. Saturation ?
- variable mappability of the genome (esp. repeat regions)

# Challenge III: Communication




400++ programs, 100++ file formats,
20+ databases, lack of standardization

External Data

Sequencing data

# Two goals

Improve Communication

Increase Efficiency

Methods

# Galaxy:
## a central repository for data access, analysis & sharing

### Data Management

- Minimize data transfer
- Format conversion tools
- Access Control

### Analysis Management

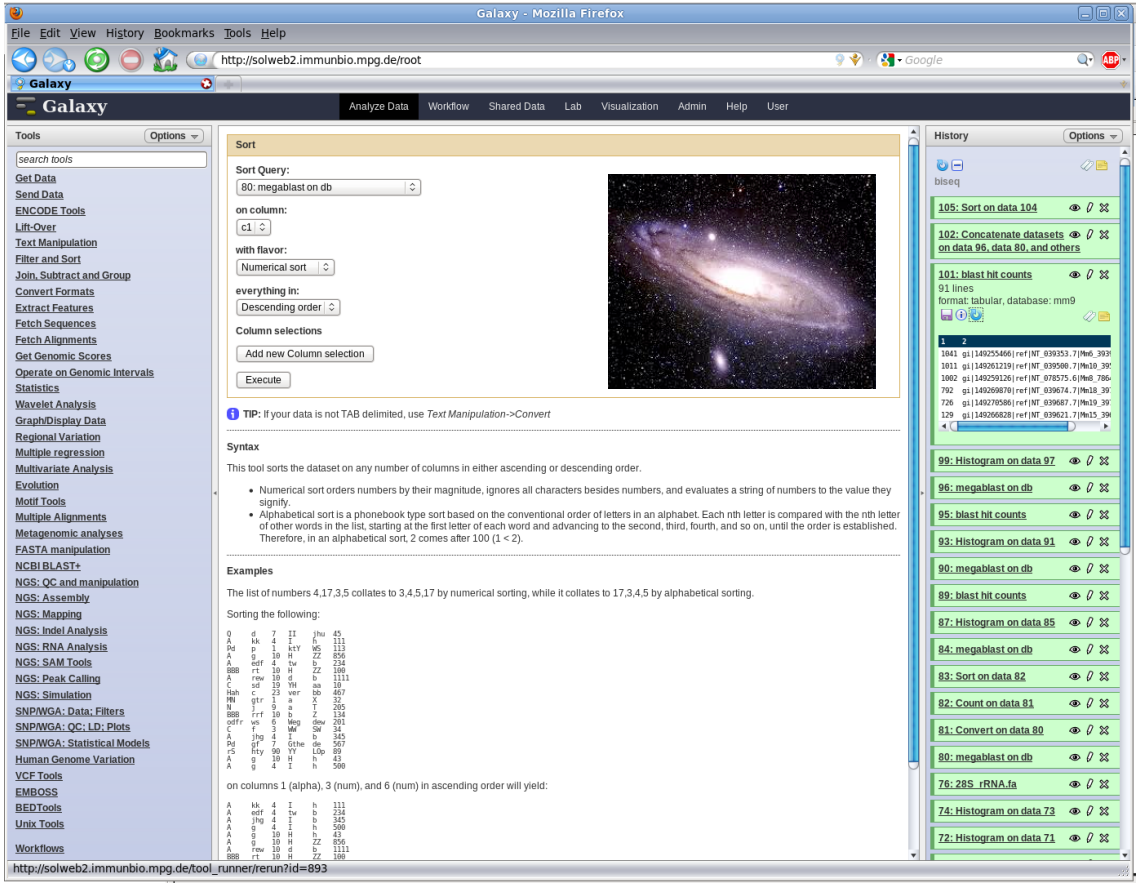
- Flexible and customizable plugins
- Reproducible workflows & histories
- Develop and integrate custom tools

### Project Management

- Share data, histories & workflows
- Easy-to-use web interface

### Community Effort

- rapidly expanding, free, standard
- Exchange Protocols (UCSC,BioMart)



# Programme

| | | |
|---|---|---|
| Mon | • Galaxy Introduction | Björn  |
| Tue | • ChIP-seq | Andreas   Fidel |
| Wed | • Exome-seq | Björn  |
| Thu | • RNA-seq | Fabian   Pavan |
| Fri | • Case study |  You |