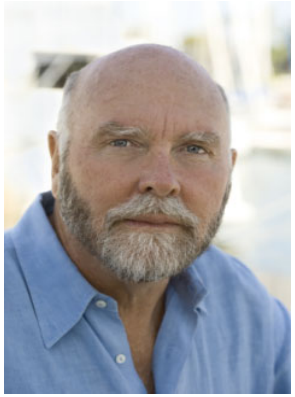
Introduction to Galaxy

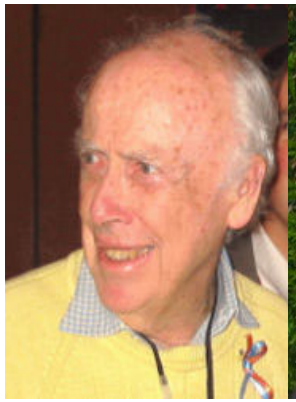
Sequencing of the human genoma, 2001

- Many years of hard work
- BAC based approaches towards WGS
- Amplification in bacterial culture
- More than 20,000 BAC clones
- Each containing about 100kb fragment
- Together provided a tiling path through each human chromosome
- Isolation, select pieces about 2-3 kb
- Subcloned into plasmid vectors, amplification, isolation
- recreate contigs
- Refinement, gap closure, sequence quality improvement
- (less 1 error/ 40,000 bases)
- 2.7 billion dollars

Sequencing of individual genomes, 2007-



$100 million

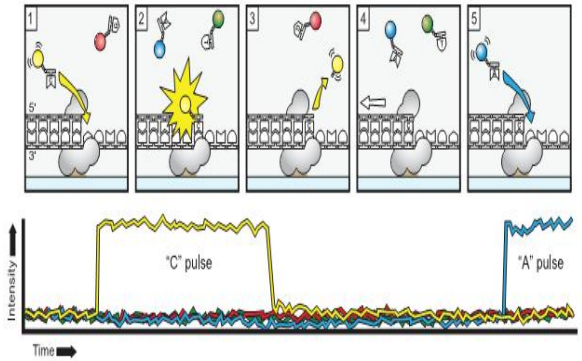


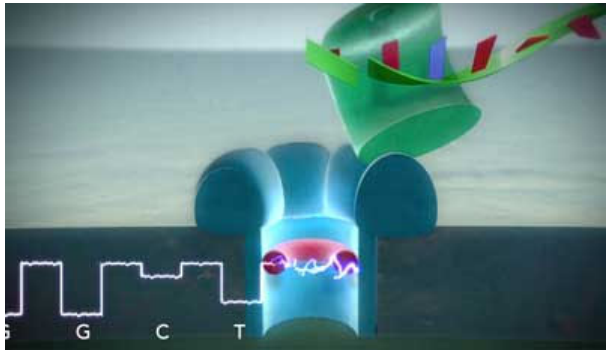
$1 million

$1000

2nd and 3rd generation sequencing

- Illumina/Solexa Genome Analyzer: 2006
- Pacific Biosciencies SMRT: launching 2010
- Oxford Nanopore MinION 2014

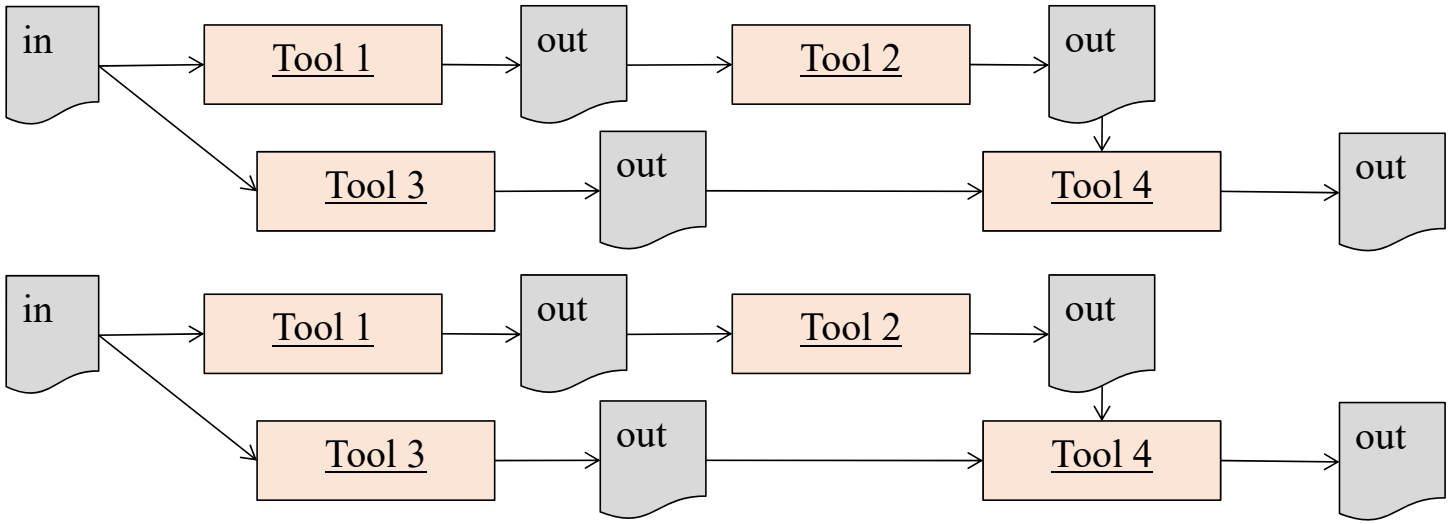




What is Galaxy?

- Galaxy is a free web-based scientific analysis platform that can be used by scientists across the world
- Galaxy's tools assist in the analysis of large biological datasets that are common in fields such as genomics or bioinformatics

Motivation of Galaxy



Sequencers throughput require parallel processing of multiple samples

$\Rightarrow$ how do you efficiently monitor all these workflow executions ?

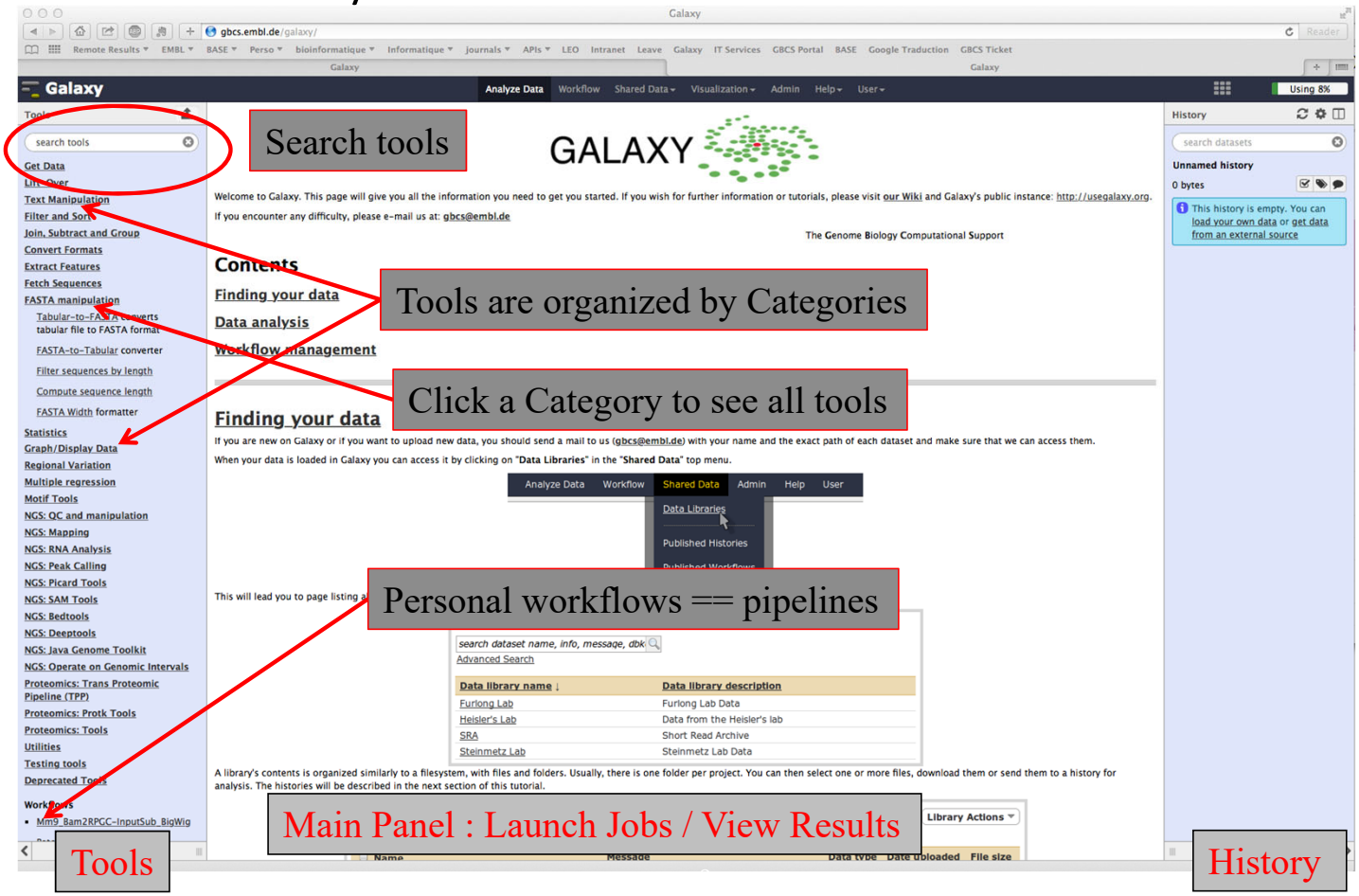
What makes Galaxy special?

- Free
- Available world-wide
- Massive toolbox for data analysis — More than 5500 tools available
- Free training materials — Useful for learning the user interface and several tools
- Helpful community and forums
- Reproducibility — “Workflows” can be saved to show step-by-step uses of tools on data

How to Start

- Go to usegalaxy.org and click the ”Login or Register” tab at the top
- Click on the “Register here” link below the login to begin creating an account
  - Once done filling in the required information, click ”Create”
  - You will be required to verify your account’s email address before being able to use any of the tools or data available
- After creating and verifying your account, you are ready to begin your introduction into Galaxy

Welcome to Galaxy



Noticeable Tool Categories

Upload personal files

Fetch public data from UCSC, ENSEMBL, mouseMINE…

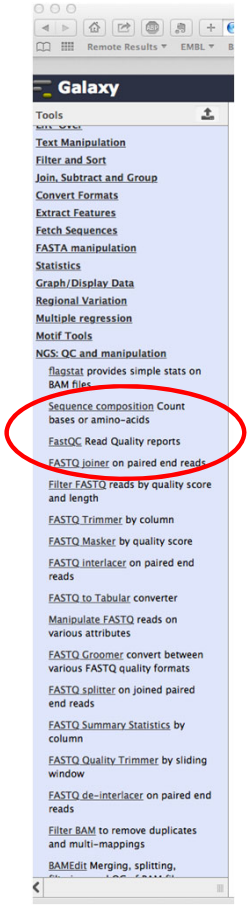
General sequence and text format manipulation Tools

NGS Tools organized by:
- application (QC, mapping, peak calling, RNA, …)
- package name (bedtools, picard)

Proteomics Tools

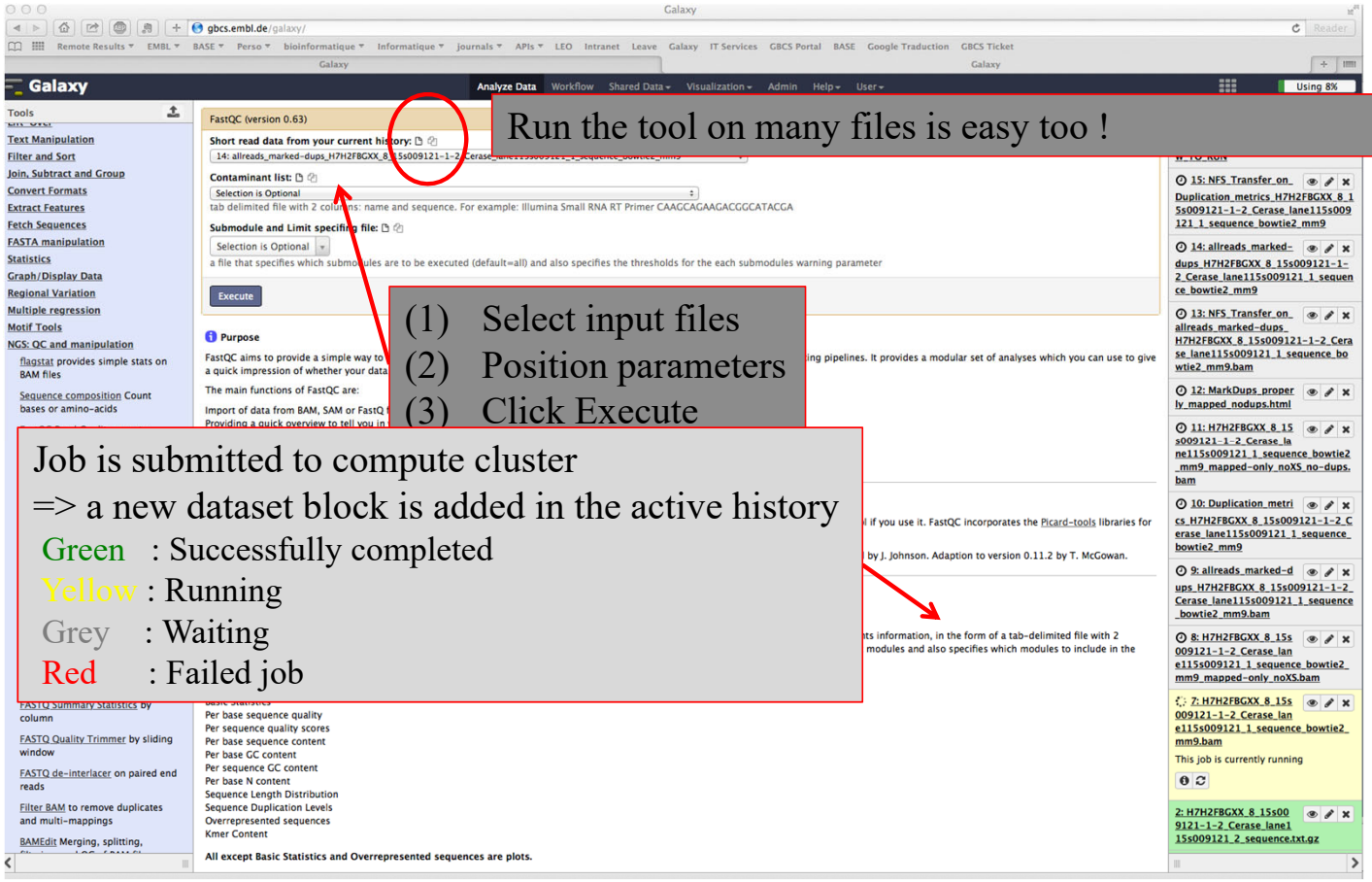
Utility (local), Test (beta version) and Deprecated (for backward comp.)

Running a Tool is easy



Click a tool to bring it up in the middle panel eg FastQC

Running a Tool is easy



Tool summary

More than >1000 tools available !

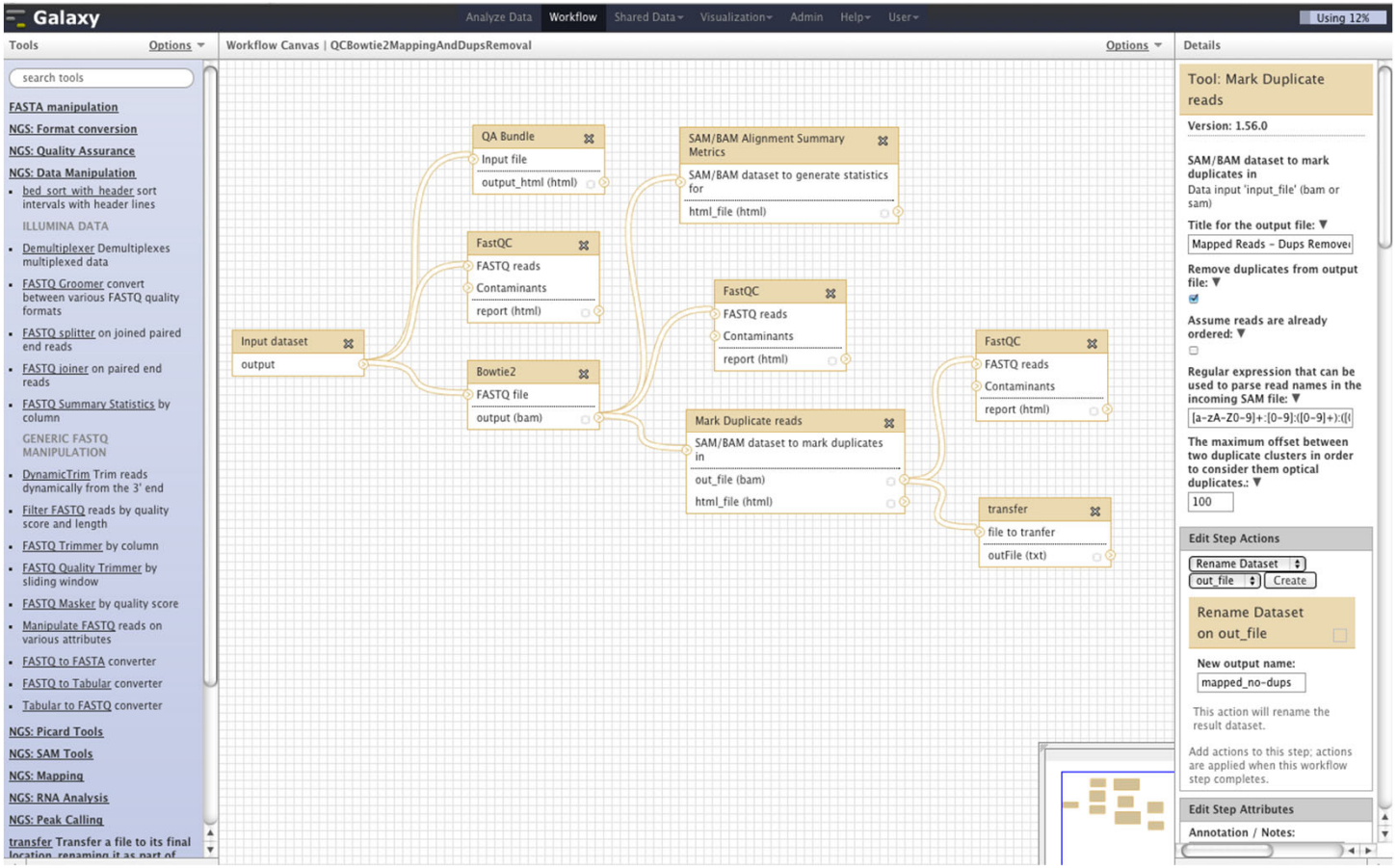
✓ All results can be downloaded or directly transferred to your project folder

✓ Missing Tools can be easily integrated

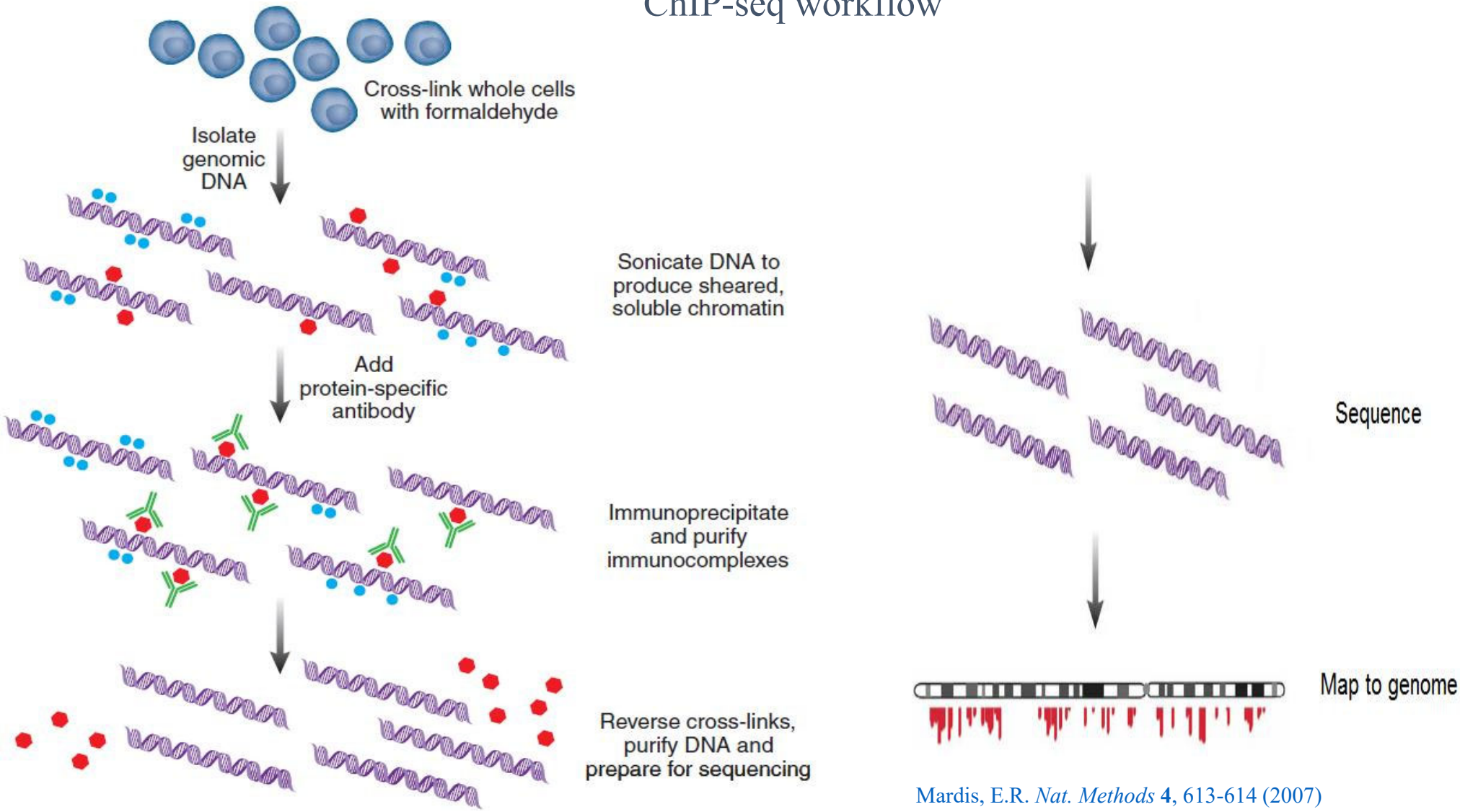
✓ Easy way to add a GUI to your own script

✓ Parameters for cluster submission can be adjusted for each tool (and even be dynamically computed)

Tools can be assembled into workflows



ChIP-seq workflow

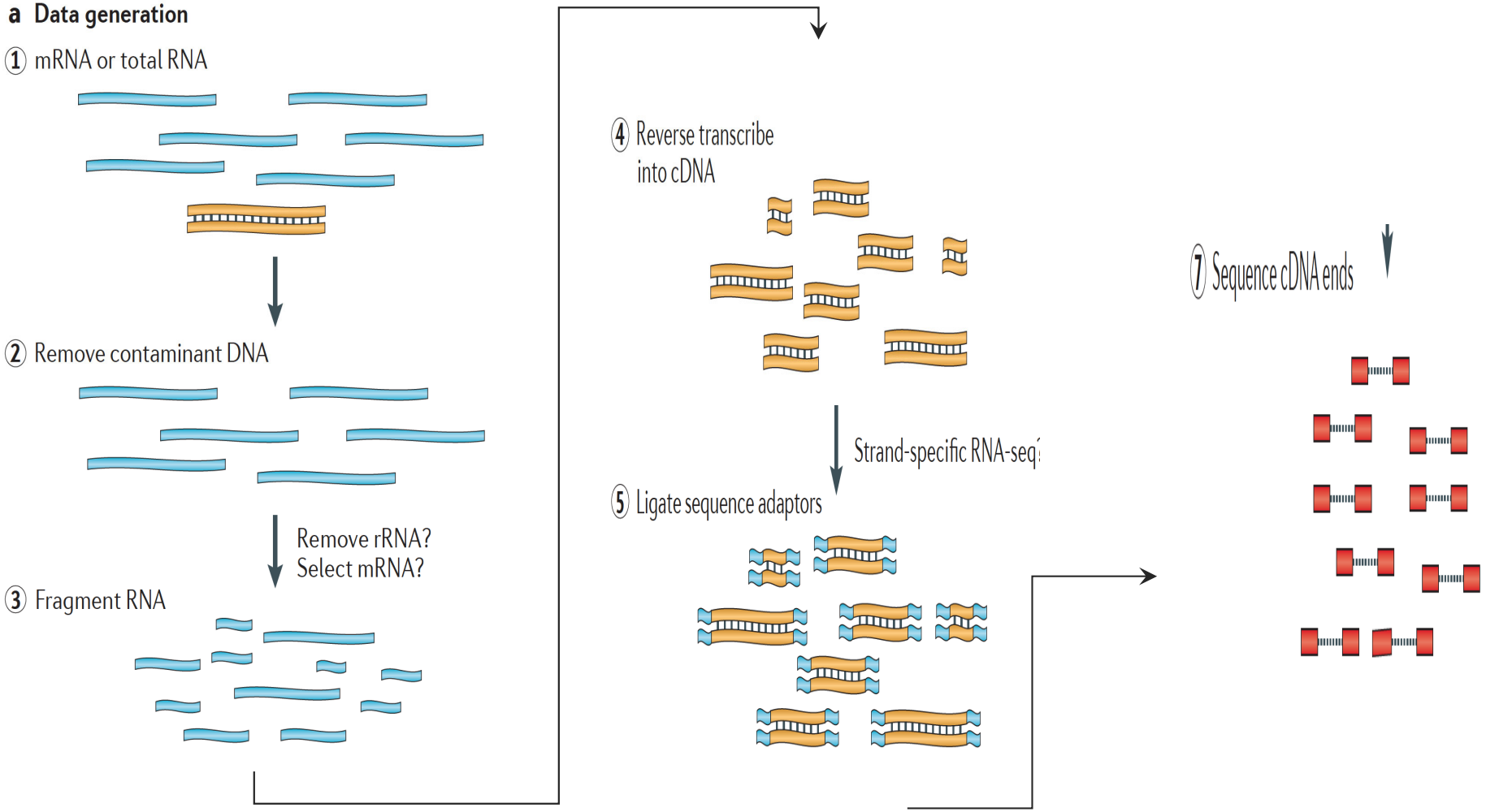


Mardis, E.R. *Nat. Methods* **4**, 613-614 (2007)

Major steps in ChIP-seq data analysis

- Trimmomatic or Cutadapt to remove adaptor sequences and filter low quality reads
- Fastqc to evaluate the quality
- Bowtie 2 to map reads to references
- MACS2 to define peaks

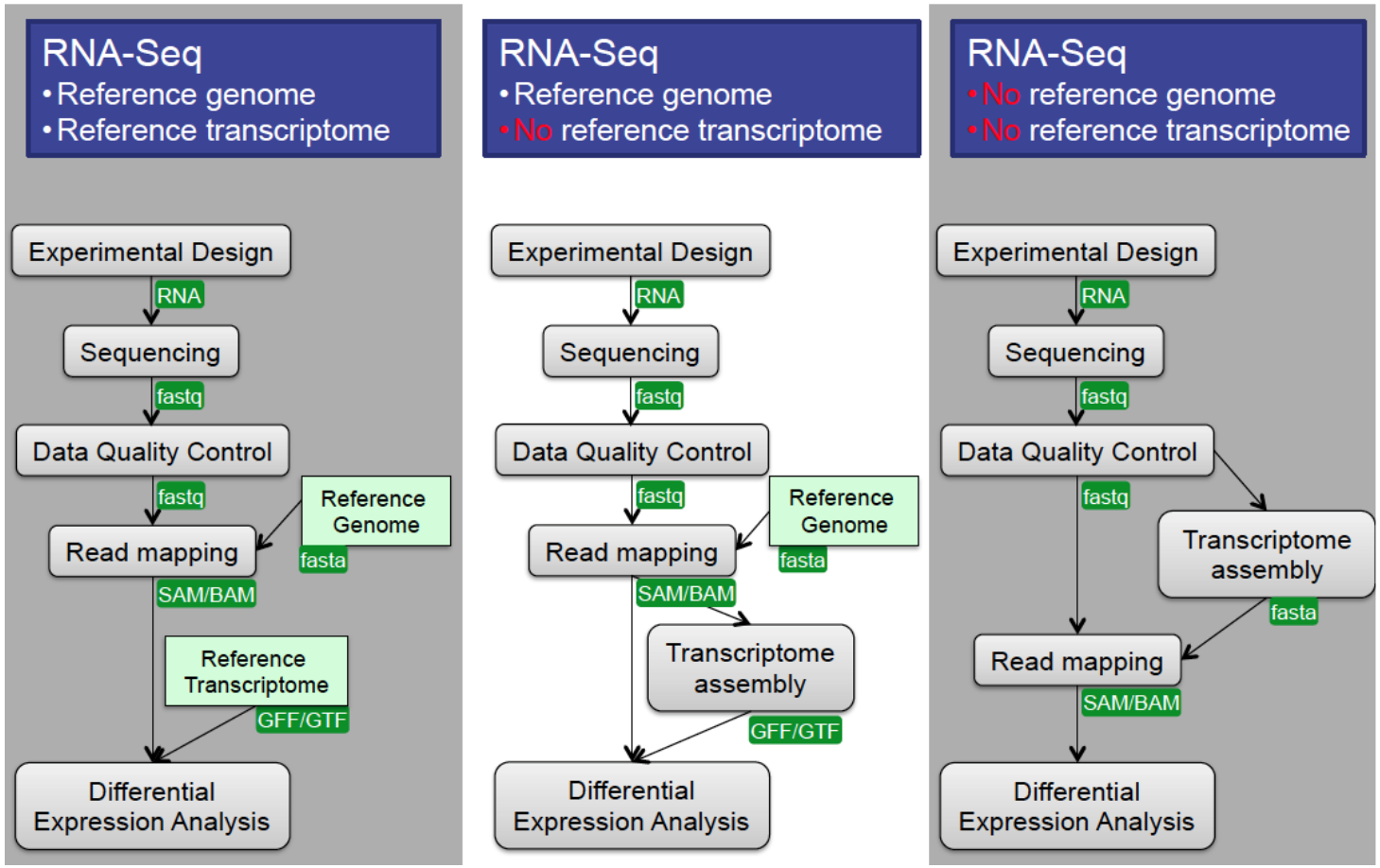
RNA-seq workflow



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Differential Gene Expression

Options for DGE analysis



How does one pick the right tools?

1. Quality Check - **FASTQC**
2. Trimming – **Trimmomatic, cutadator**
3. Splice-aware alignment – **STAR, hisat2**

   Bacterial alignment - **BWA** or **Novoalign**
4. Counting reads per gene – **featureCounts, stringtie**

- Counting reads per isoform **- Salmon**

5. DGE Analysis – **edgeR, DESeq limma**

De novo transcriptome assembly - **Trinity**

References cont.

- https://galaxyproject.github.io/
- https://galaxyproject.org/usegalaxy/
- usegalaxy.org