LEIDEN UNIVERSITY MEDICAL CENTER

Bioinformatics at the LGTC

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Sequencers: HiSeq

Figure 1: HiSeq 2000.

Characteristics:

- High throughput.
- Paired end.
- High accuracy.
- Read length 2×150bp.
- Relatively long run time.
- Relatively expensive.

Introduction

Sequencers: Ion Torrent

Figure 2: Ion torrent.

Characteristics:

- Moderate throughput.
- Single end (for now).
- High accuracy.
- Read length $\pm$200bp.
- Short run time.
- Cheap runs.

Introduction

Sequencers: Pacific Biosciences SMRT

Figure 3: PacBio RS.

Introduction

Pacific Biosciences Single Molecule, Real-Time

Characteristics:

- Long reads (several kilobases).
- Relatively high error rate.
- Relatively high throughput (comparable with the Roche 454).

Introduction

Pacific Biosciences Single Molecule, Real-Time

Characteristics:

- Long reads (several kilobases).
- Relatively high error rate.
- Relatively high throughput (comparable with the Roche 454).

Circular consensus sequencing.

- Sequence the same molecule several times.
- Extremely high accuracy.
- Acceptable read length ($\pm$250bp).

Core analysis

Base calling

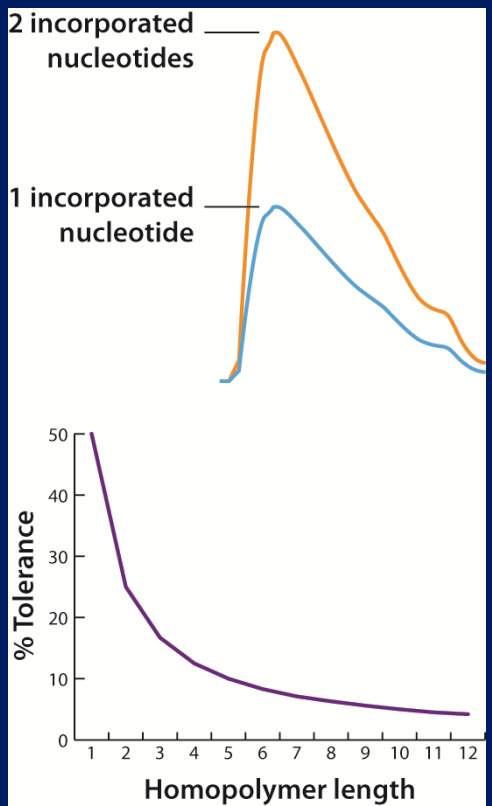

Figure 4: Ion torrent

Yahya Anvar

Core analysis

Base calling

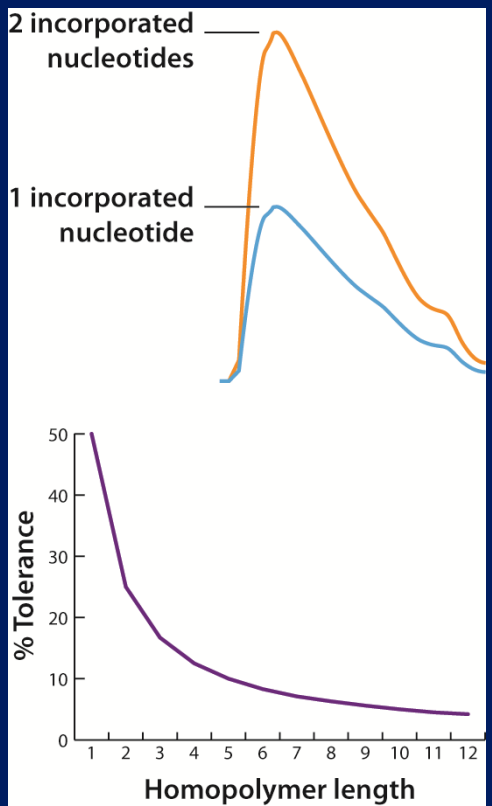

Figure 4: Ion torrent

Some difficulties for the Ion Torrent:

- Mononucleotide stretches.
- Strand specific insertions.
- Context dependent biases.

Yahya Anvar

Core analysis

Base calling

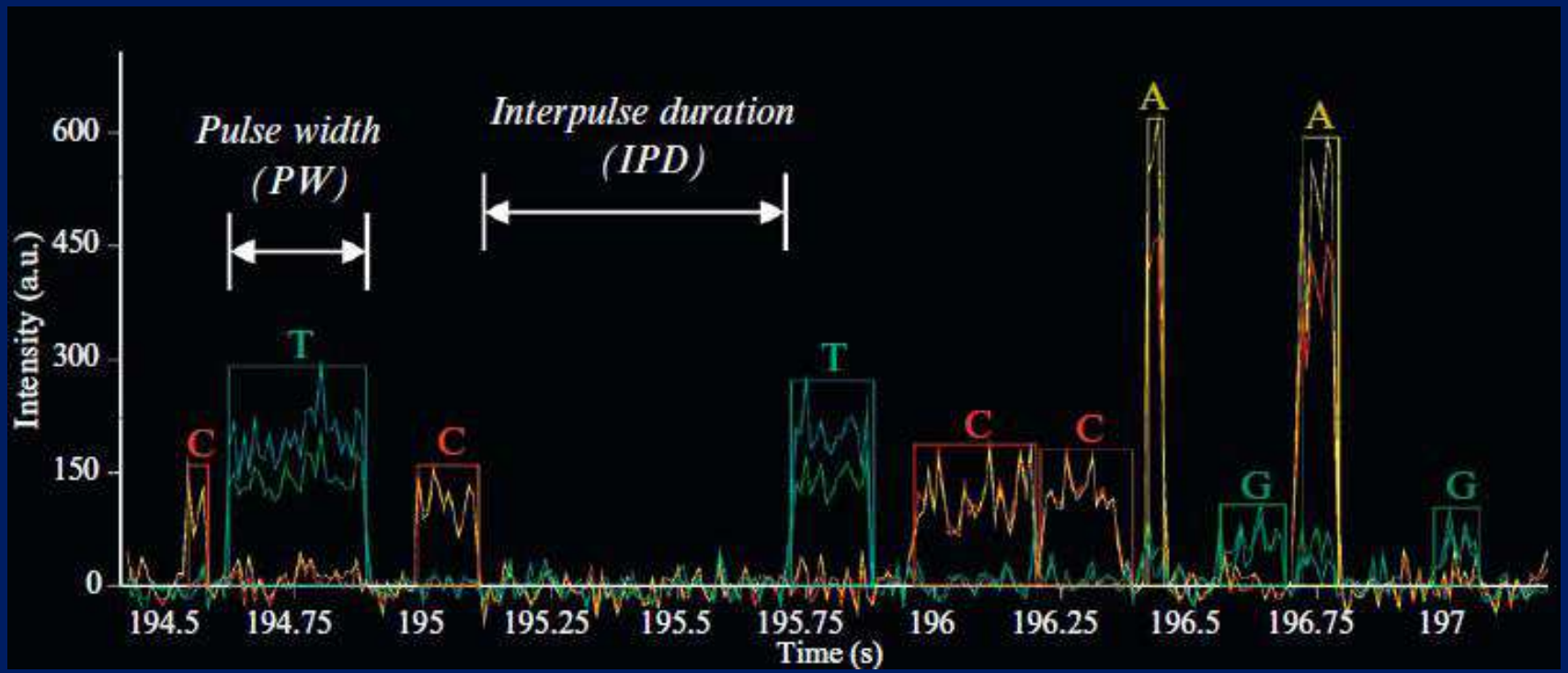

Figure 5: PacBio

Yahya Anvar

Core analysis

Secondary data analysis

Alignment issues.

Michiel van Galen, Yahya Anvar, Martijn Vermaat

Core analysis

Secondary data analysis

Alignment issues.

Different platforms:

- HiSeq — Dropping accuracy at the end of the reads.
- Ion Torrent — Mononucleotide stretches.
- PacBio — Uniform error, slight bias to insertions.

Michiel van Galen, Yahya Anvar, Martijn Vermaat

Core analysis

Secondary data analysis

Alignment issues.

Different platforms:

- HiSeq — Dropping accuracy at the end of the reads.
- Ion Torrent — Mononucleotide stretches.
- PacBio — Uniform error, slight bias to insertions.

Different experiments:

- tags — Short reads, no allowance for errors.
- RNASeq — Exon-exon spanning reads.
- mtDNA — Circular.

Michiel van Galen, Yahya Anvar, Martijn Vermaat

Core analysis

Secondary data analysis

De novo assembly:

- Springtail.
- Cobra.
- Clavipes.
- Wolbachia.
- Acinetobacter plasmids.
- . . .

Jeroen Frank, Yahya Anvar, Ken Kraaijeveld

Computing facilities

Emphasis on reproducibility and reliability

https://humgenprojects.lumc.nl/

Michel Villerius, Martijn Vermaat, Zuotian Tatum

Computing facilities

Emphasis on reproducibility and reliability

Facilities:

- A compute cluster (29 nodes, 368 CPUs).
- Redundant and fast storage (576T).
- A convenient way to export (share / view) data.

https://humgenprojects.lumc.nl/

Michel Villerius, Martijn Vermaat, Zuotian Tatum

Computing facilities

Emphasis on reproducibility and reliability

Facilities:

- A compute cluster (29 nodes, 368 CPUs).
- Redundant and fast storage (576T).
- A convenient way to export (share / view) data.

Support for programmers and clients of our programs:

- Version Control Systems for code.
- Bugtracking systems.

https://humgenprojects.lumc.nl/

Michel Villerius, Martijn Vermaat, Zuotian Tatum

Computing facilities

Emphasis on reproducibility and reliability

Facilities:

- A compute cluster (29 nodes, 368 CPUs).
- Redundant and fast storage (576T).
- A convenient way to export (share / view) data.

Support for programmers and clients of our programs:

- Version Control Systems for code.
- Bugtracking systems.

Tracking and automation:

- LIMS system.

https://humgenprojects.lumc.nl/

Michel Villerius, Martijn Vermaat, Zuotian Tatum

Computing facilities

Emphasis on reproducibility and reliability

Facilities:

- A compute cluster (29 nodes, 368 CPUs).
- Redundant and fast storage (576T).
- A convenient way to export (share / view) data.

Support for programmers and clients of our programs:

- Version Control Systems for code.
- Bugtracking systems.

Tracking and automation:

- LIMS system.

Reuse of previous analysis and quality control:

- Variant database.

https://humgenprojects.lumc.nl/

Michel Villerius, Martijn Vermaat, Zuotian Tatum

Computing facilities

NGS LIMS

Technical details:

- Support for all sequencing platforms.
- An API to communicate with a scheduler.
- Based on Django for rapid development.

Feddde Schaeffer, Zuotian Tatum, Martijn Vermaat

Computing facilities

NGS LIMS

Technical details:

- Support for all sequencing platforms.
- An API to communicate with a scheduler.
- Based on Django for rapid development.

Bringing pipelines together:

- Scheduler communicates with:
  - LIMS.
  - Sequencers.
  - Cluster / Storage.

Feddde Schaeffer, Zuotian Tatum, Martijn Vermaat

Computing facilities

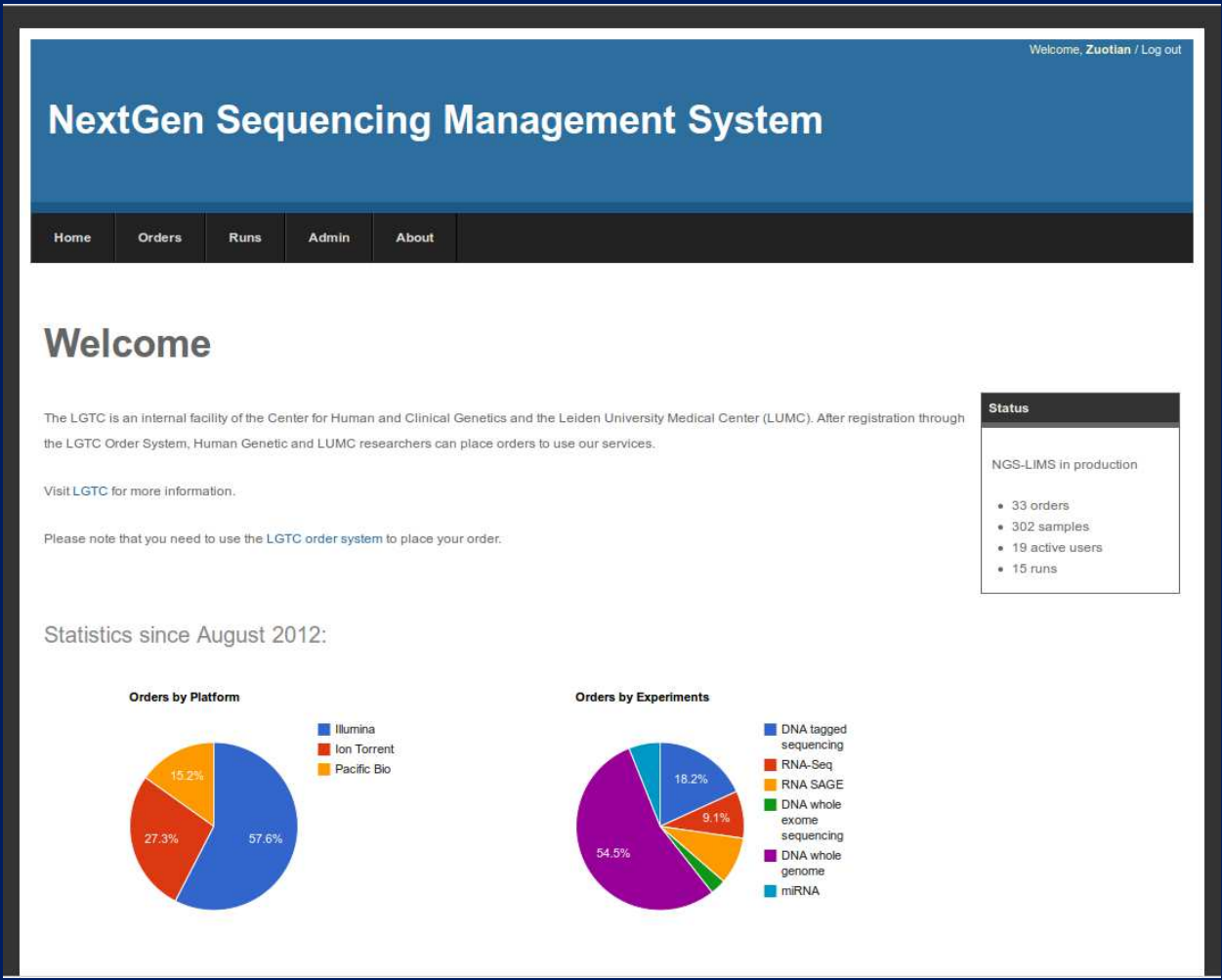

Figure 6: LIMS welcome page.

Computing facilities

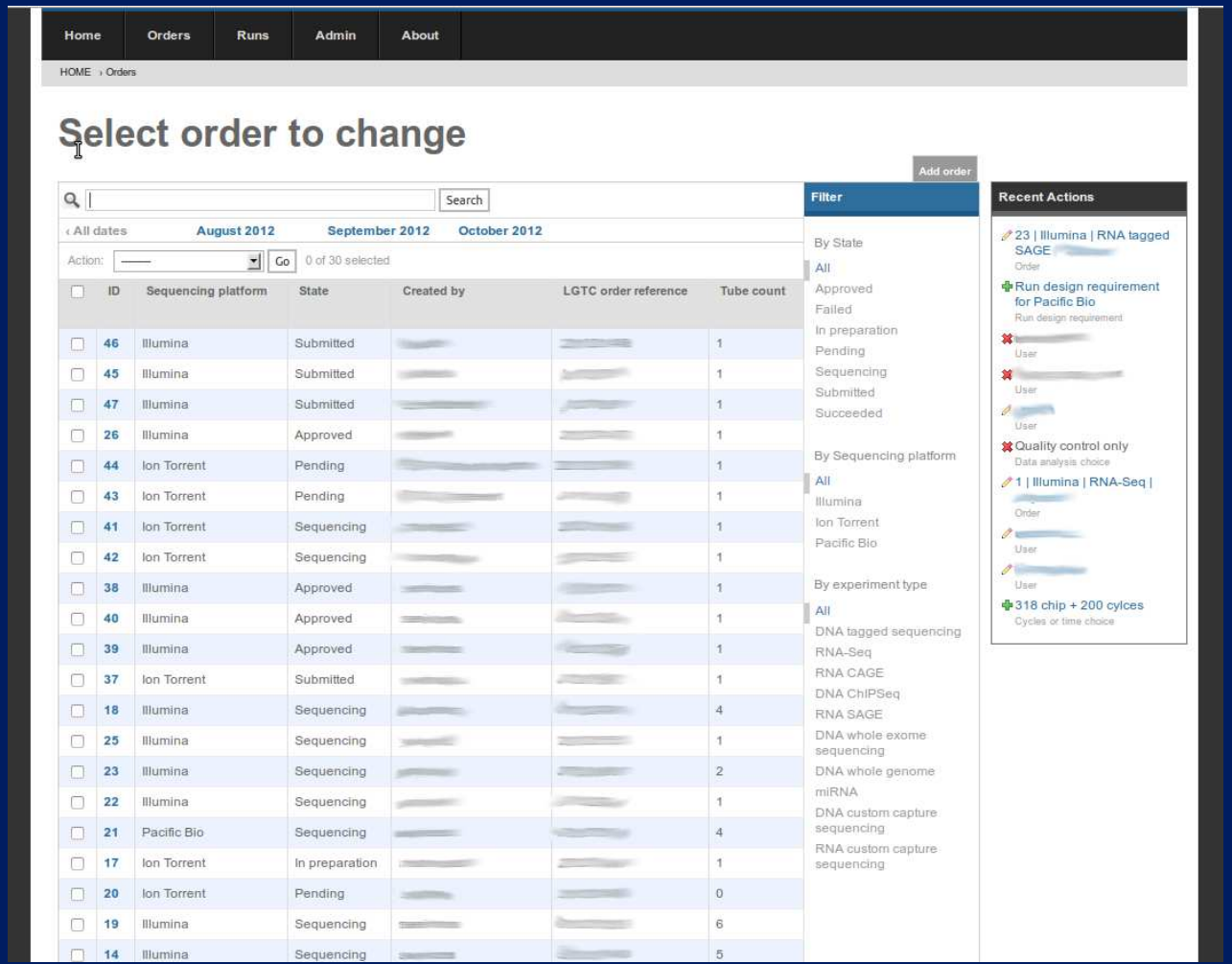

Figure 7: LIMS order panel.

Computing facilities

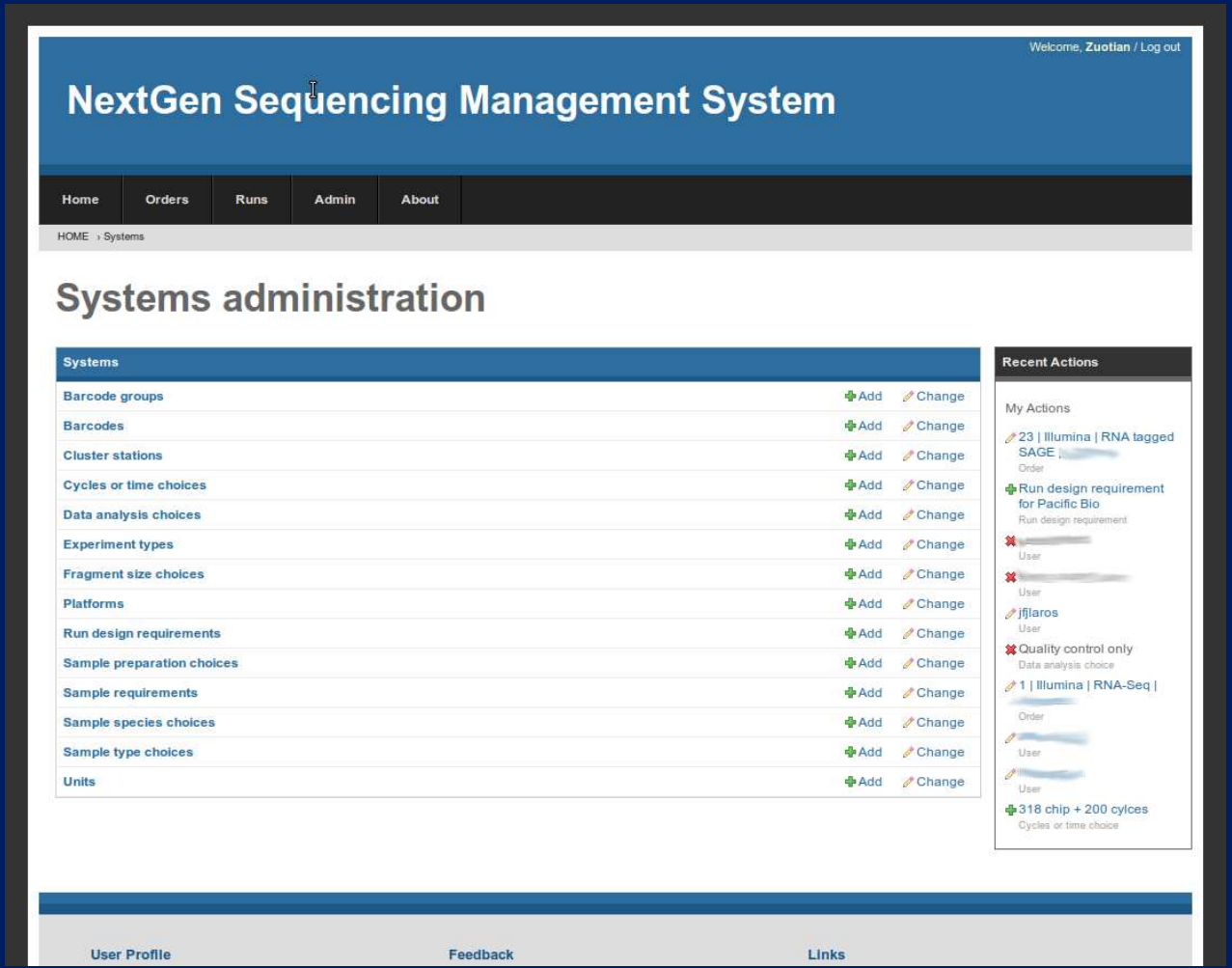

Figure 8: LIMS administration.

Computing facilities

DVD

The *Diagnostic Variant Database*.

- Share variants found in exome sequencing experiments.
- Find functionally relevant variants.

Martijn Vermaat, David van Enckevort, Hailiang Mei

Computing facilities

DVD

The *Diagnostic Variant Database*.

- Share variants found in exome sequencing experiments.
- Find functionally relevant variants.

Technical details:

- Store coverage information to determine reference calls.
- Disambiguation of variant descriptions.
- Pooling without loss of information.
- Duplicate sample detection.
  - Allows for re-annotation without polluting the database.
- Encrypted connection with authentication.

Martijn Vermaat, David van Enckevort, Hailiang Mei

Education

Next Generation Sequencing data analysis

11-13 September 2012

- PhD students, postdocs, senior researchers.

J.M. Boer, J.T. den Dunnen, W. van IJcken, C. van Gelder

Education

Next Generation Sequencing data analysis

11-13 September 2012

- PhD students, postdocs, senior researchers.
- Discussion of different platforms and produced data.
    - Illumina, Roche, ABI, Ion Torrent, etc.

J.M. Boer, J.T. den Dunnen, W. van IJcken, C. van Gelder

Education

Next Generation Sequencing data analysis

11-13 September 2012

- PhD students, postdocs, senior researchers.
- Discussion of different platforms and produced data.
  - Illumina, Roche, ABI, Ion Torrent, etc.
- General applications.
  - Resequencing, structural variation, de novo assembly, visualisation, pipelines, etc.

J.M. Boer, J.T. den Dunnen, W. van IJcken, C. van Gelder

Education

Next Generation Sequencing data analysis

11-13 September 2012

- PhD students, postdocs, senior researchers.
- Discussion of different platforms and produced data.
  - Illumina, Roche, ABI, Ion Torrent, etc.
- General applications.
  - Resequencing, structural variation, de novo assembly, visualisation, pipelines, etc.
- Specific applications.
  - QC, statistics, expression, ChIP-seq, metagenomics, etc.

J.M. Boer, J.T. den Dunnen, W. van IJcken, C. van Gelder

Education

Next Generation Sequencing data analysis

11-13 September 2012

- PhD students, postdocs, senior researchers.
- Discussion of different platforms and produced data.
  - Illumina, Roche, ABI, Ion Torrent, etc.
- General applications.
  - Resequencing, structural variation, de novo assembly, visualisation, pipelines, etc.
- Specific applications.
  - QC, statistics, expression, ChIP-seq, metagenomics, etc.
- Practical sessions.
  - Galaxy, NextGENe, CLCbio.

Room for 60 people, always full.

J.M. Boer, J.T. den Dunnen, W. van IJcken, C. van Gelder

Education

Basic Linux Course

Full day to get people acquainted with Linux.

- Available on request (given four times now).

https://humgenprojects.lumc.nl/trac/humgenprojects/wiki/LinuxCourse

M. van Galen, M. van Iterson

Education

Basic Linux Course

Full day to get people acquainted with Linux.

- Available on request (given four times now).
- Covered topics:
  - Installation practical.
  - Command line, files, directories, ownership, etc.
  - Installing software.
  - Regular expressions.
  - Connecting to remote machines.

https://humgenprojects.lumc.nl/trac/humgenprojects/wiki/LinuxCourse

M. van Galen, M. van Iterson

Education

Basic Linux Course

Full day to get people acquainted with Linux.

- Available on request (given four times now).
- Covered topics:
  - Installation practical.
  - Command line, files, directories, ownership, etc.
  - Installing software.
  - Regular expressions.
  - Connecting to remote machines.
- All topics have a lecture and a practical.

https://humgenprojects.lumc.nl/trac/humgenprojects/wiki/LinuxCourse

M. van Galen, M. van Iterson

Education

NGS Introduction Course

Full day to get people acquainted with NGS data analysis.

https://humgenprojects.lumc.nl/trac/GAPSS3/wiki/course

M. van Galen

Education

NGS Introduction Course

Full day to get people acquainted with NGS data analysis.

- Sort version of the Linux course.
- Command line and NGS tools.
- Connecting to remote machines (clusters).
- NGS pipelines.
- Using Galaxy.

https://humgenprojects.lumc.nl/trac/GAPSS3/wiki/course

M. van Galen

Education

NGS Introduction Course

Full day to get people acquainted with NGS data analysis.

- Sort version of the Linux course.
- Command line and NGS tools.
- Connecting to remote machines (clusters).
- NGS pipelines.
- Using Galaxy.

Given twice in the last year.

- University of Leuven, Belgium.
- Avans Hogeschool Breda.
    - Will be repeated this year.

https://humgenprojects.lumc.nl/trac/GAPSS3/wiki/course

M. van Galen

Full data analysis

Clusters

Figure 9: Dell M610 blade server

Full data analysis

A different look at pipelines

What if. . .

- All commands are atomic.
  - We describe input and output.
- We build a *dependency graph*.
- Trace a path in this graph to find a workflow.

Full data analysis

A different look at pipelines

What if. . .

- All commands are atomic.
  - We describe input and output.
- We build a *dependency graph*.
- Trace a path in this graph to find a workflow.

This way we do not need to:

- Design a workflow.
- Figure out which parts can be run in parallel.

Full data analysis

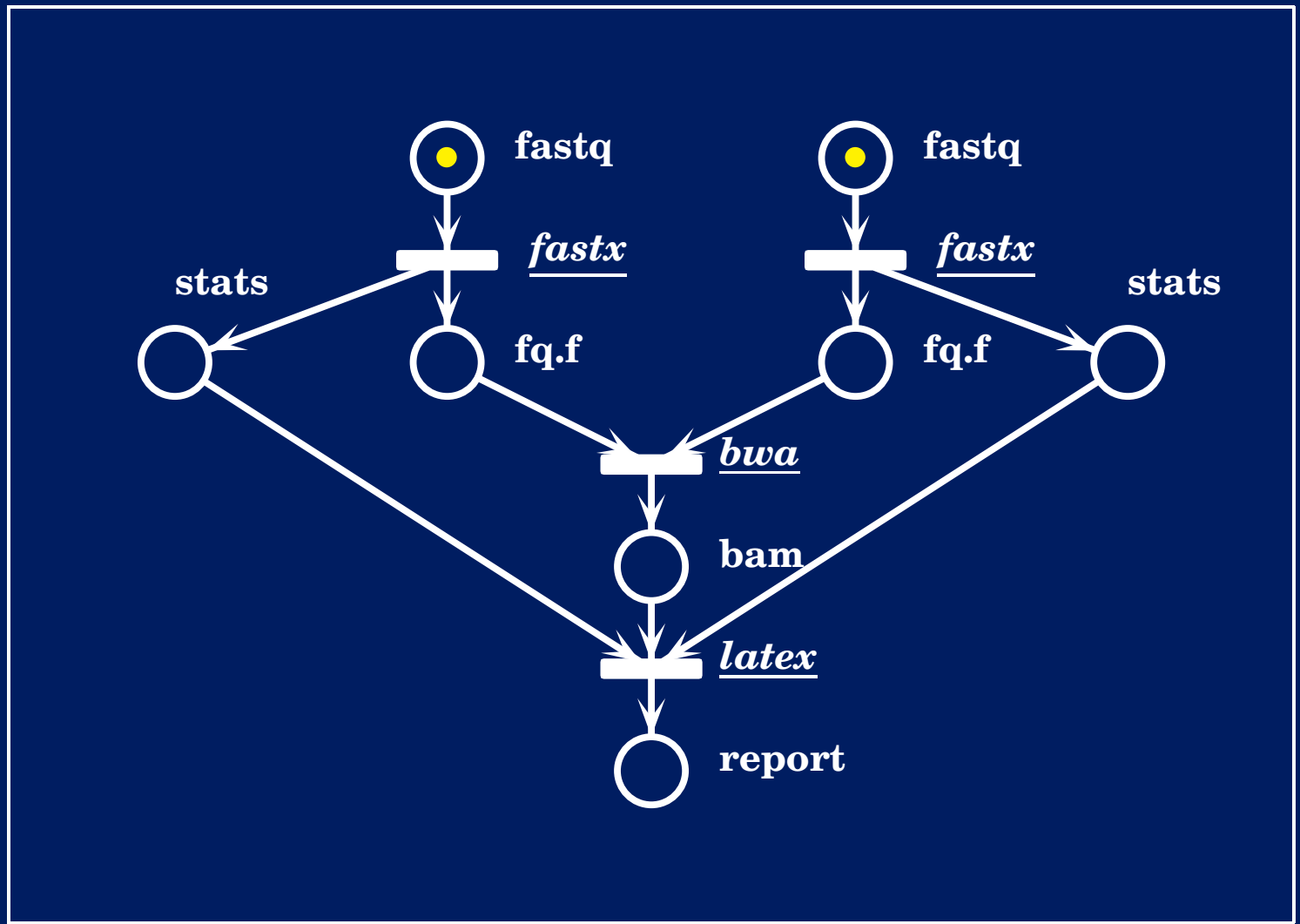

Figure 10: A parallel workflow

Full data analysis

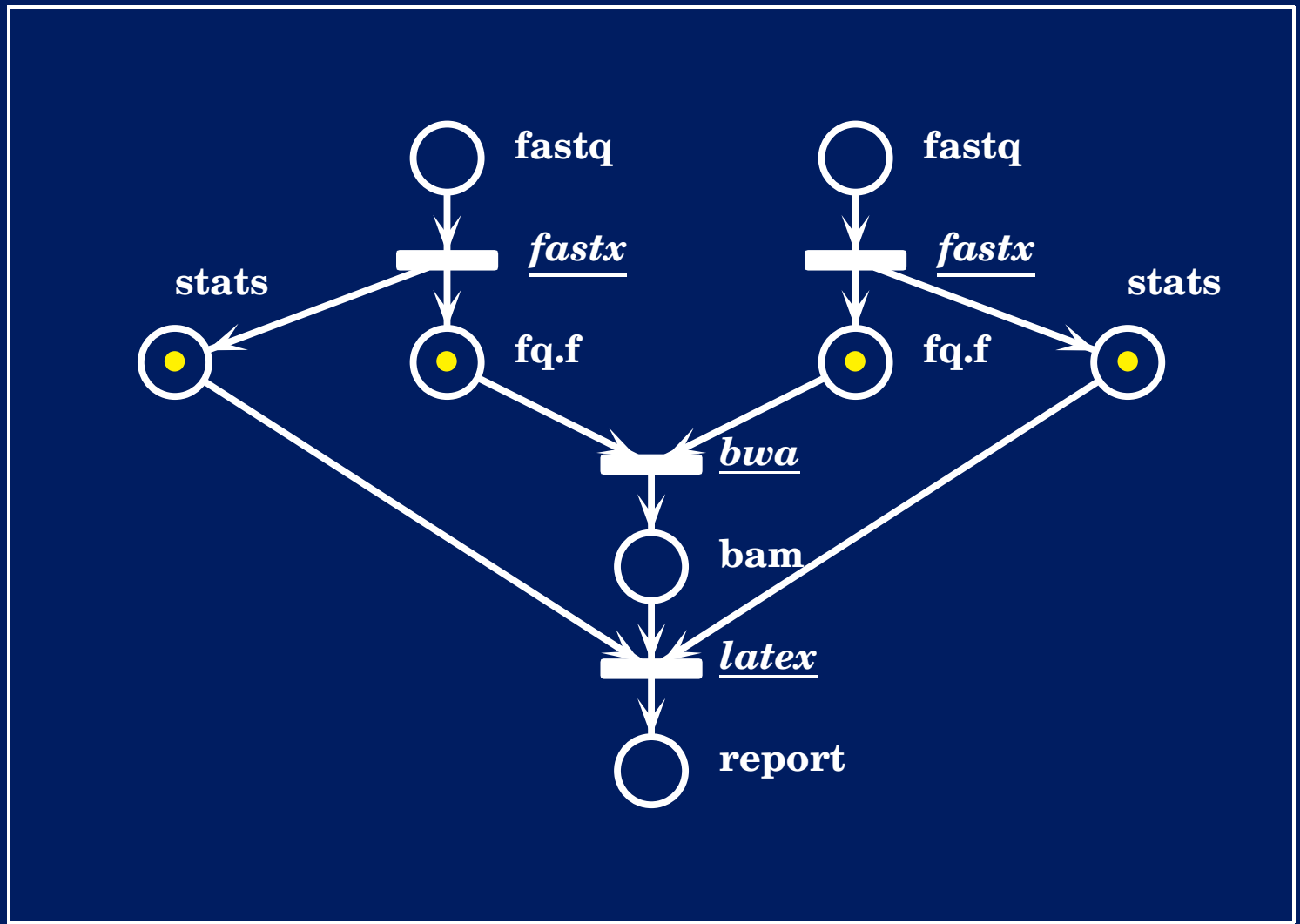

Figure 10: A parallel workflow

Full data analysis

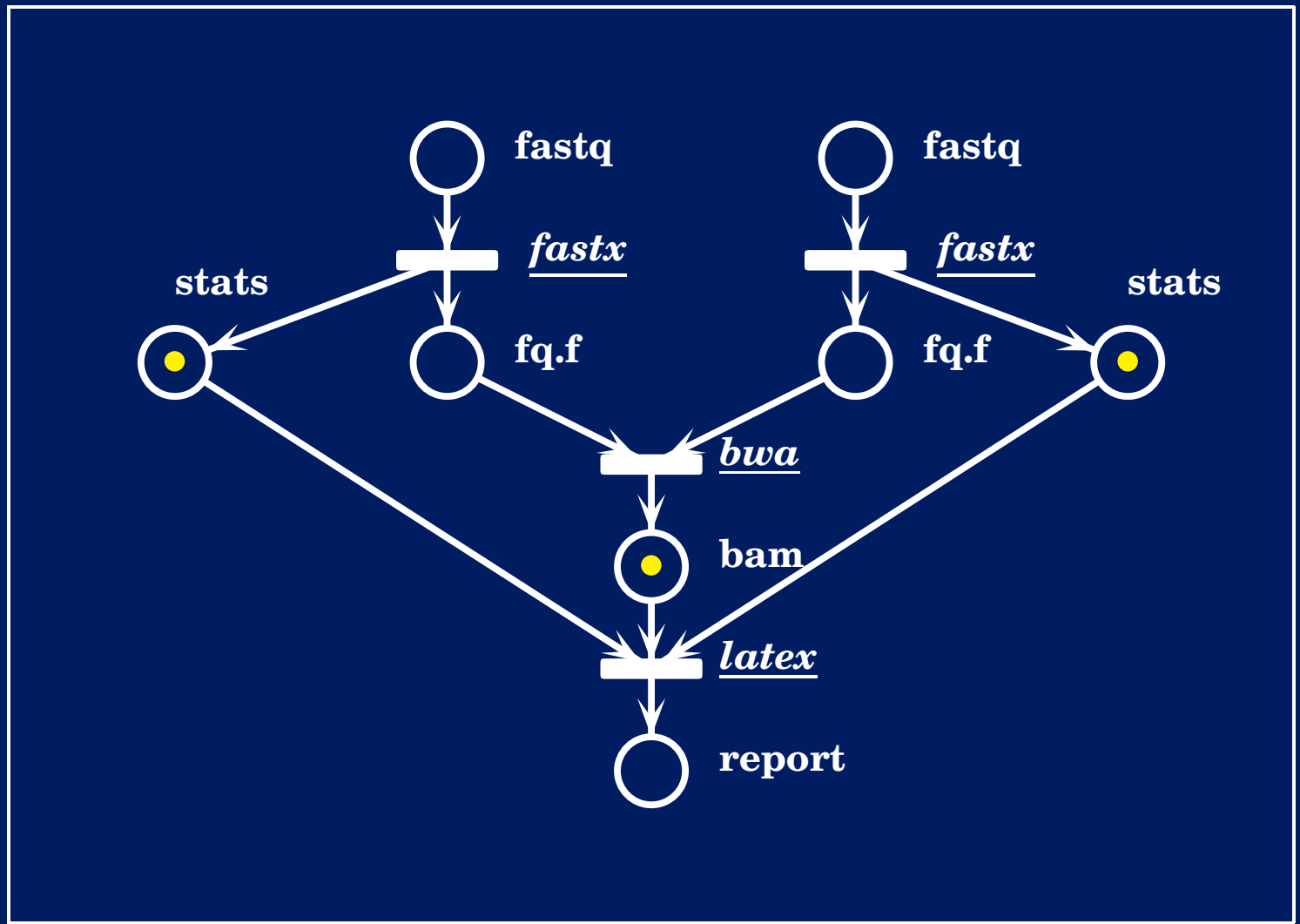

Figure 10: A parallel workflow

Full data analysis

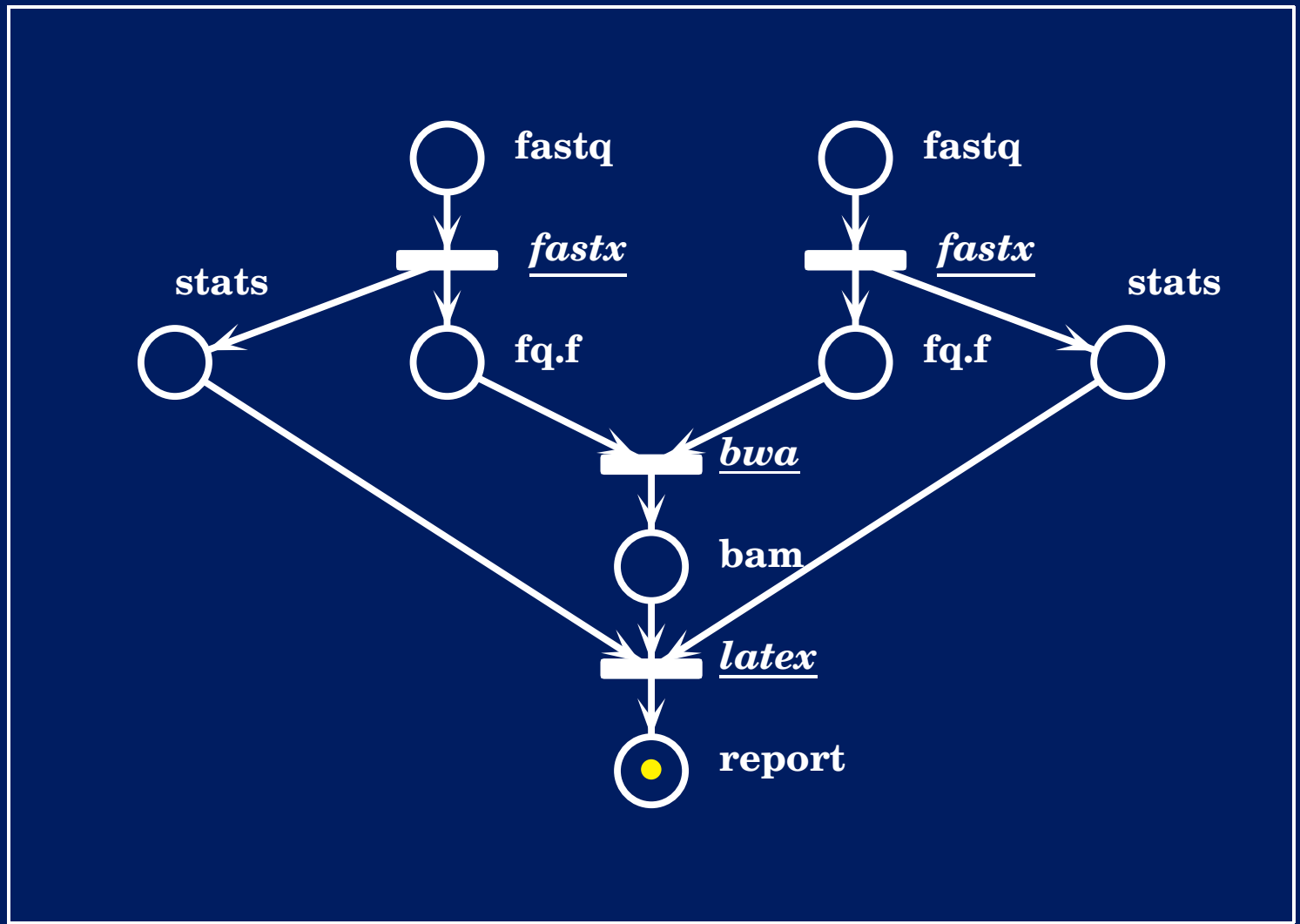

Figure 10: A parallel workflow

Full data analysis

The GAPSS3 workflow

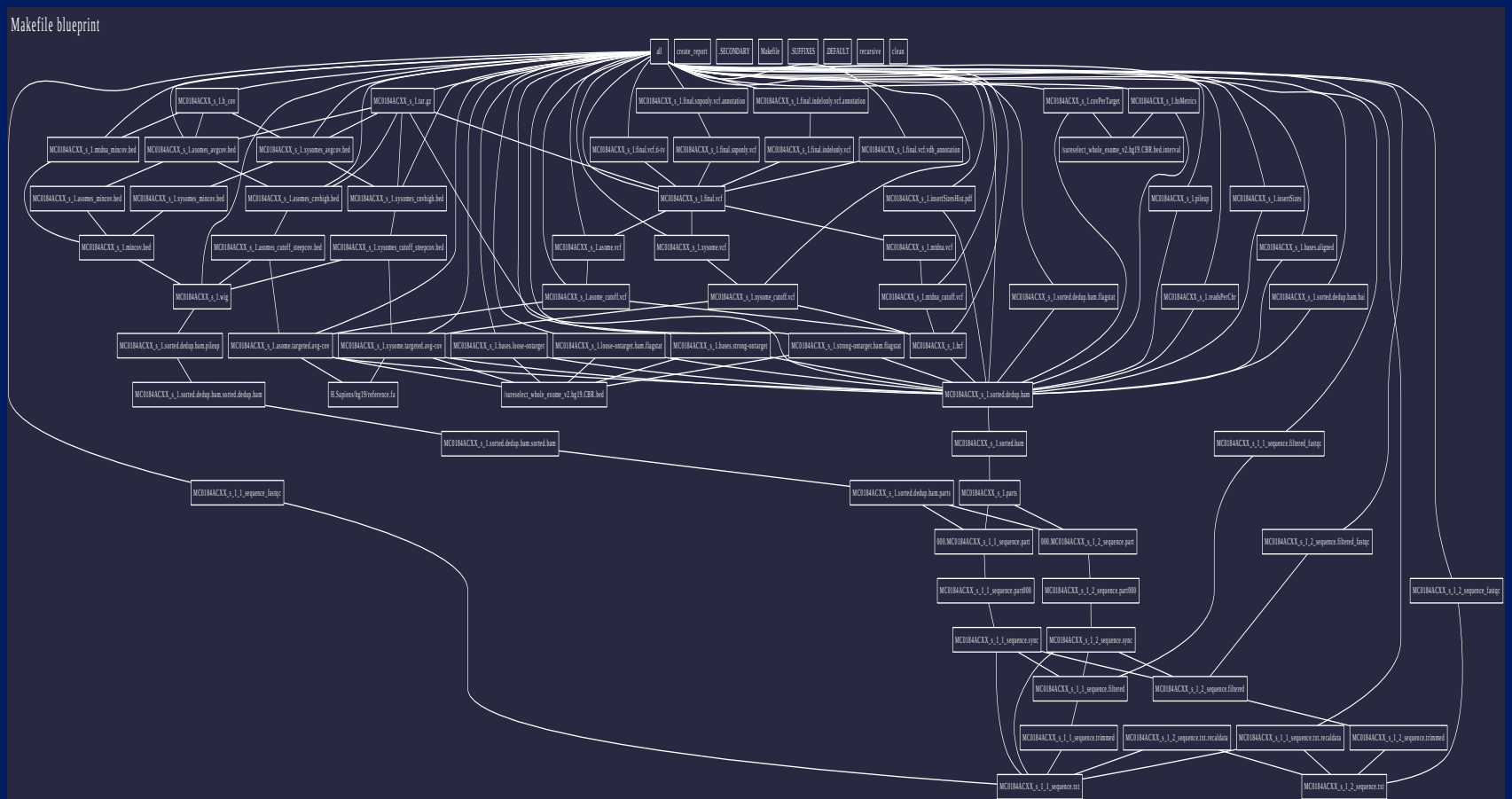

Figure 11: GAPSS3

Full data analysis

The GAPSS3 workflow

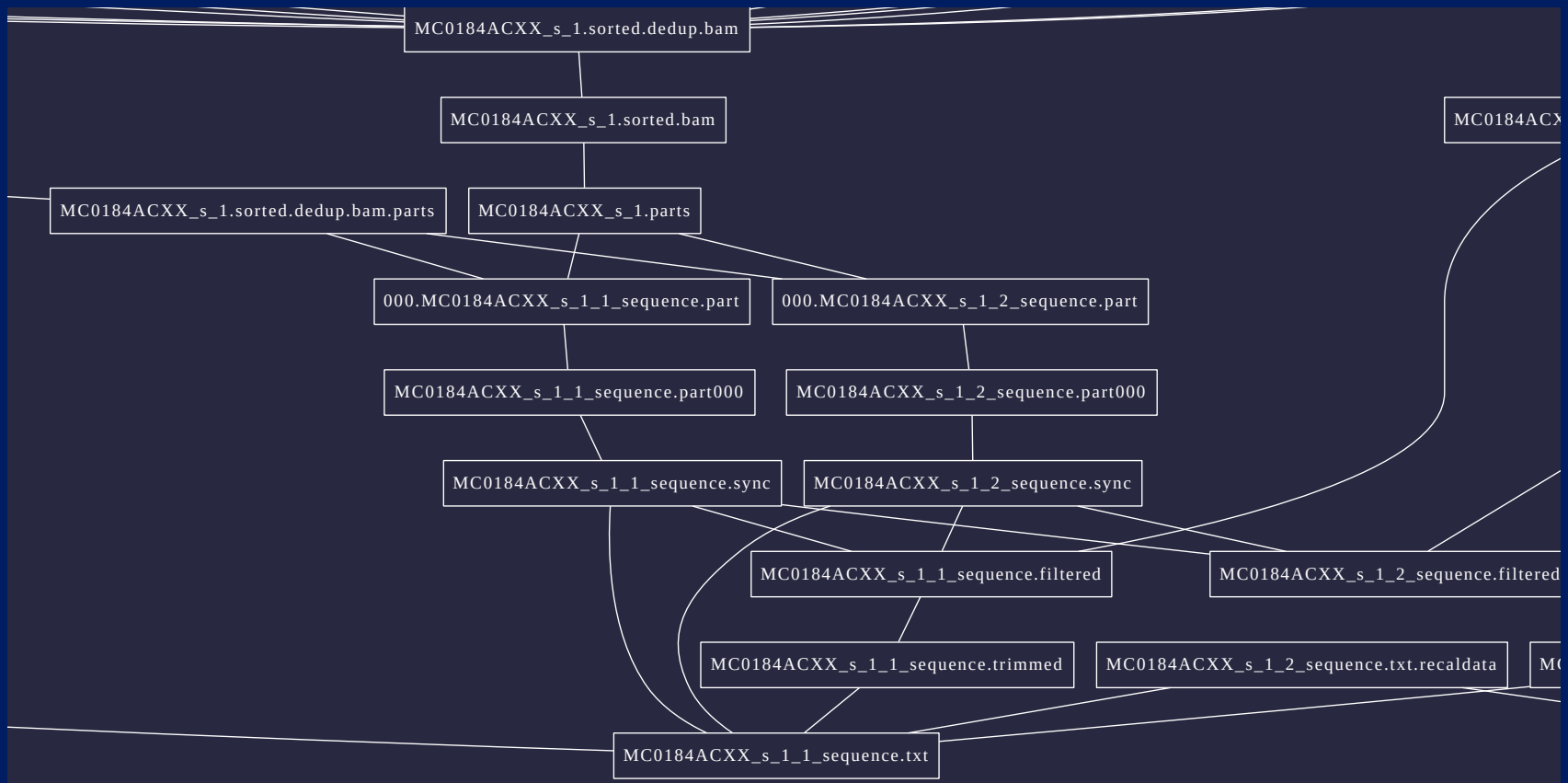

Figure 12: Part of GAPSS3

Full data analysis

GAPSS report

18 pages with details about the analysis.

Full data analysis

GAPSS report

18 pages with details about the analysis.

- QC of raw data.
- QC of cleaned data (trimmed, clipped).
- Alignment statistics.
- Capture statistics.
- Variant calling statistics.

Full data analysis

GAPSS report

18 pages with details about the analysis.

- QC of raw data.
- QC of cleaned data (trimmed, clipped).
- Alignment statistics.
- Capture statistics.
- Variant calling statistics.
  - Ti/Tv rate.
  - Number of hits in LOVD.
  - ...

Full data analysis

Pipeline development

A selection of pipelines under development:

• Y-STR haplotype diversity	Thirsa Kraaijenbrink
• Trisomy detection	Hailiang Mei
• TALENS analysis	Yahya Anvar, Marcel Veltrop, Cor Breukel, Sjef Verbeek
• Strand specific RNAseq	Willeke van Roon
• CNV pipeline	Michiel van Galen
• Bacterial strain identification	Sunita Paltansing, Sandra Bernards
• Antibiotic resistance gene identification	Sunita Paltansing, Sandra Bernards

Full data analysis

Metrics for NGS data files

Distance between wiggle files.

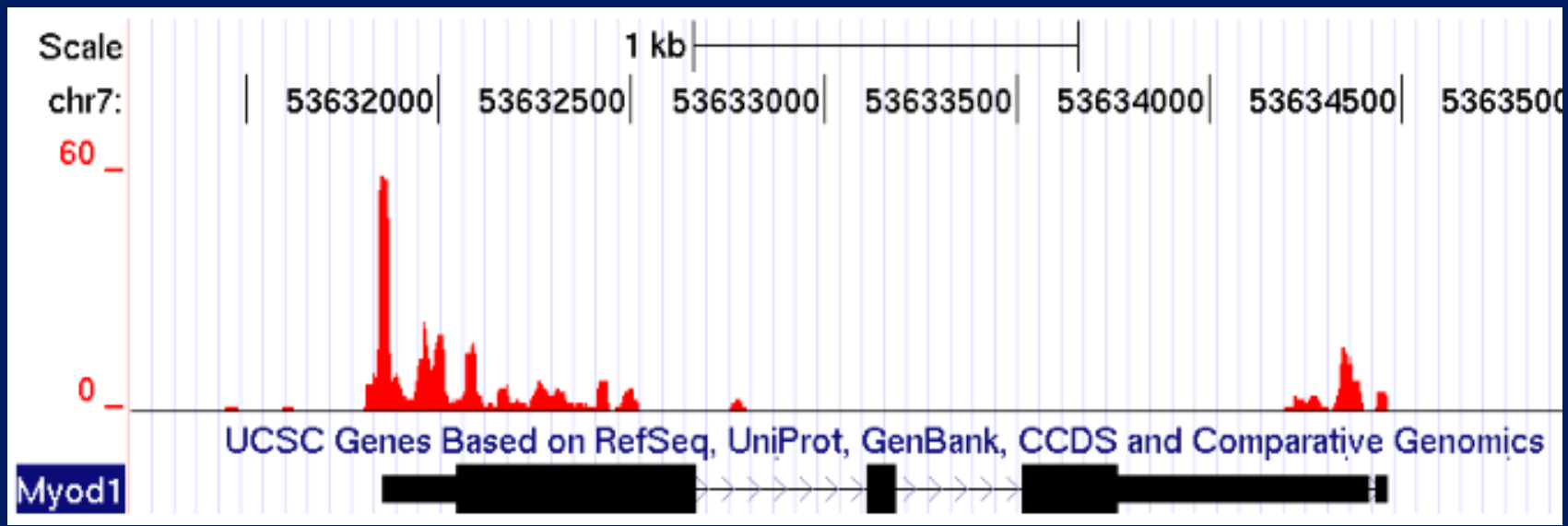

Figure 13: Wiggle file.

Pairwise comparison based on the *multiset* distance measure.

Full data analysis

Metrics for NGS data files

Comparing k-mer profiles.

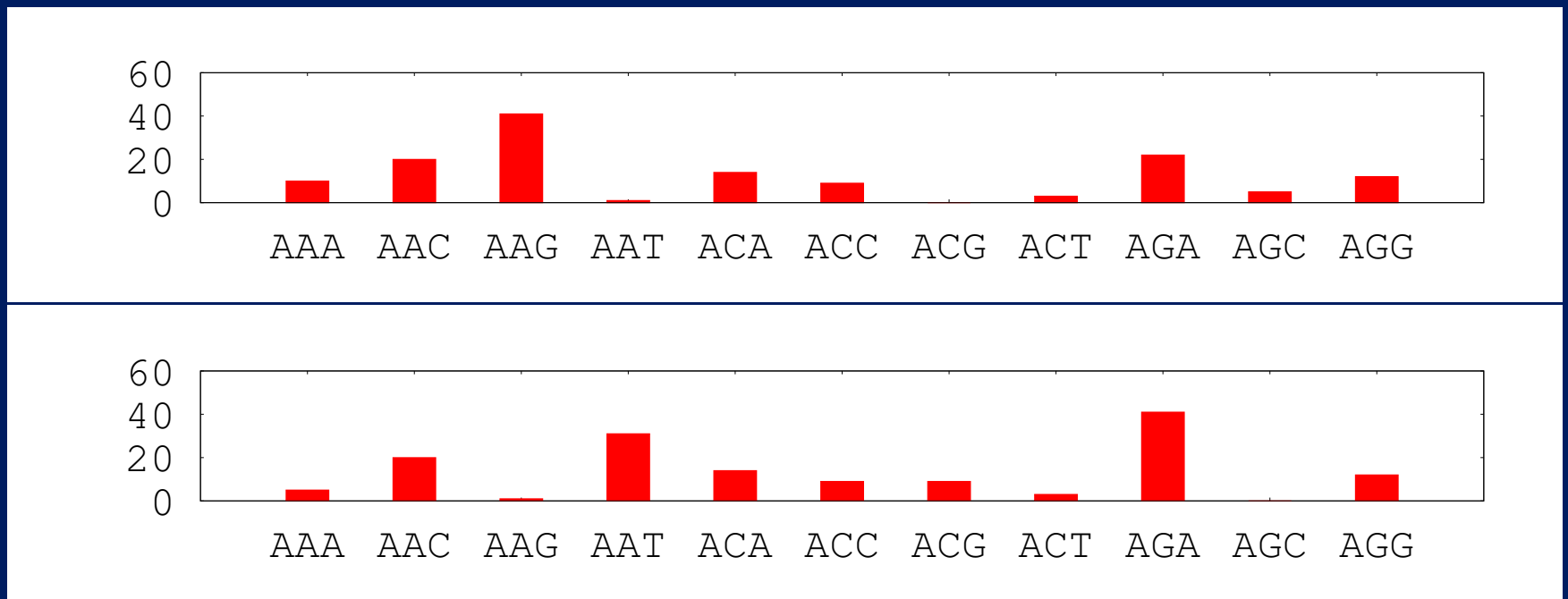

Figure 14: Two k-mer profiles.

Jaap van der Heijden, Yahya Anvar

Full data analysis

Forensics

Jaap van der Heijden, Kris van der Gaag, Peter de Knijff

Full data analysis

Forensics

STR profiling:

- Look deeper into STRs by using sequencing.
- Semi-global alignment of flanking sequences.
- Regular expressions for known alleles.
- Classification of new alleles.

Jaap van der Heijden, Kris van der Gaag, Peter de Knijff

Full data analysis

Forensics

STR profiling:

- Look deeper into STRs by using sequencing.
- Semi-global alignment of flanking sequences.
- Regular expressions for known alleles.
- Classification of new alleles.

SNP profiling:

- Highly variable regions in a certain population.
- Easier to work with than with STRs.

Jaap van der Heijden, Kris van der Gaag, Peter de Knijff

Sequencing Analysis Support Core

SASC: free support

Initiative by:

- Medical Statistics.
- Clinical Genetics.
- Human Genetics.

Sequencing Analysis Support Core

SASC: free support

Initiative by:

- Medical Statistics.
- Clinical Genetics.
- Human Genetics.

Set up:

- Three research assistants.
- Will facilitate anyone in the LUMC in principle.

Sequencing Analysis Support Core

SASC: free support

Initiative by:

- Medical Statistics.
- Clinical Genetics.
- Human Genetics.

Set up:

- Three research assistants.
- Will facilitate anyone in the LUMC in principle.

Tasks:

- Assist in data interpretation.
- Set up a diagnostic pipeline.
- Identify recurring questions and automate.

In principle no primary or secondary data analysis.

Future plans

Improve quality

Quality control:

- Better integration of QC pipelines via the LIMS.
- Further automation for improved reproducibility.
- Better integration with the wet lab via the LIMS.
    - E.g., if a capture is done, we can call certain QC pipelines.

Future plans

Improve quality

Quality control:

- Better integration of QC pipelines via the LIMS.
- Further automation for improved reproducibility.
- Better integration with the wet lab via the LIMS.
    - E.g., if a capture is done, we can call certain QC pipelines.

Experimental pipelines.

Future plans

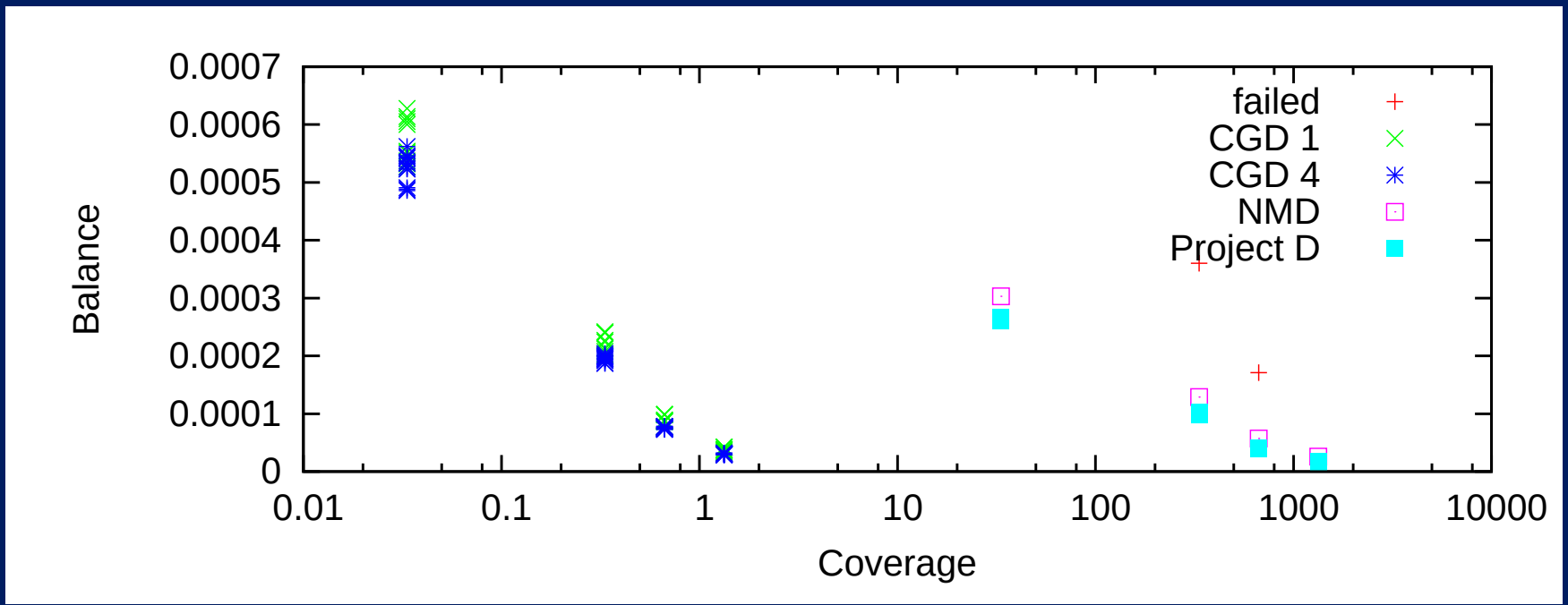

Figure 15: Coverage vs. 9-mer balance.

Questions?

Acknowledgements:

Michiel van Galen
Martijn Vermaat
Michel Villerius
Yayha Anvar
Zuotian Tatum
Jaap van der Heijden
Fedde Schaeffer
Hailiang Mei
Ken Kraaijeveld
Johan den Dunnen