



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**



## *Sequencers: HiSeq*



### Characteristics:

- High throughput.
- Paired end.
- High accuracy.
- Read length  $2 \times 150\text{bp}$ .
- Relatively long run time.
- Relatively expensive.

Figure 1: HiSeq 2000.

*Sequencers: Ion Torrent*

Figure 2: Ion torrent.

## Characteristics:

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

*Sequencers: Pacific Biosciences SMRT*



Figure 3: PacBio RS.

*Pacific Biosciences Single Molecule, Real-Time*

## Characteristics:

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

*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 ( $\pm 250$ bp).

## Base calling

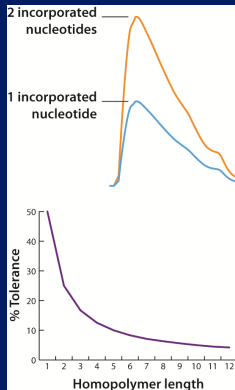


Figure 4: Ion torrent

Yahya Anvar

## Base calling

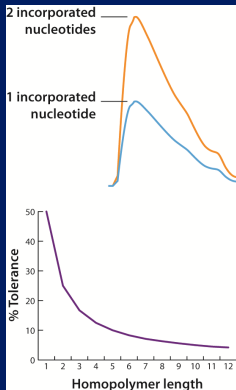


Figure 4: Ion torrent

Yahya Anvar

Some difficulties for the Ion Torrent:

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



## Base calling

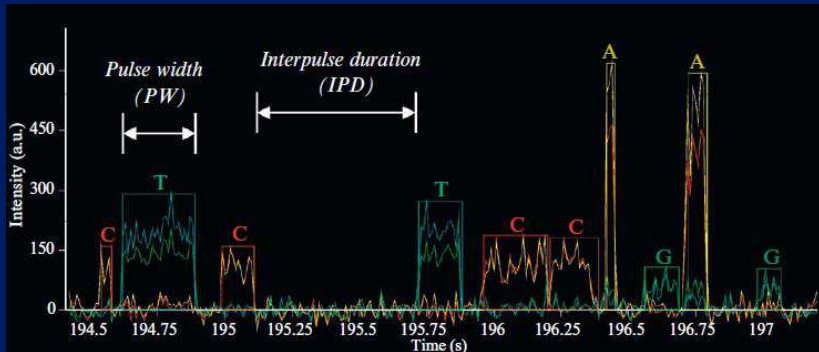


Figure 5: PacBio

*Secondary data analysis*

Alignment issues.

*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.

## *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.

*Secondary data analysis*

De novo assembly:

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

**Jeroen Frank, Yahya Anvar, Ken Kraaijeveld**

*Emphasis on reproducibility and reliability*

<https://humgenprojects.lumc.nl/>

**Michel Villerius, Martijn Vermaat, Zuotian Tatum**

### *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**

### *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**



### *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**

### *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**

## *NGS LIMS*

Technical details:

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

**Fedde Schaeffer, Zuotian Tatum, Martijn Vermaat**

## *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.

**Fedde Schaeffer, Zuotian Tatum, Martijn Vermaat**

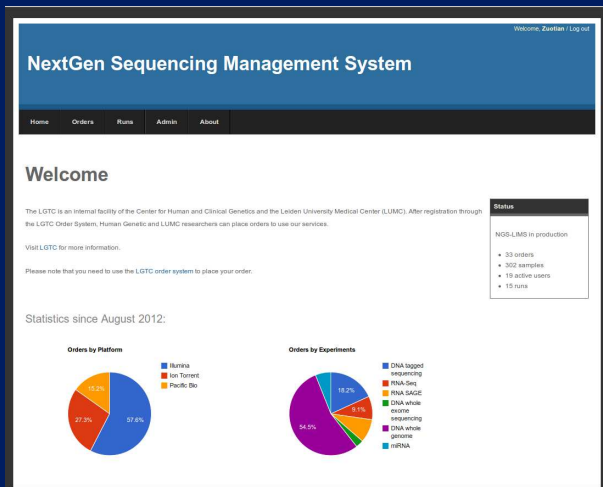


Figure 6: LIMS welcome page.

# Computing facilities

HOME - Orders

## Select order to change

Search

Filter

All dates: August 2012 September 2012 October 2012

Action:   0 of 30 selected

<input type="checkbox"/>	ID	Sequencing platform	State	Created by	LGCT order reference	Tube count
<input type="checkbox"/>	46	Illumina	Submitted			1
<input type="checkbox"/>	45	Illumina	Submitted			1
<input type="checkbox"/>	47	Illumina	Submitted			1
<input type="checkbox"/>	26	Illumina	Approved			1
<input type="checkbox"/>	44	Ion Torrent	Pending			1
<input type="checkbox"/>	43	Ion Torrent	Pending			1
<input type="checkbox"/>	41	Ion Torrent	Sequencing			1
<input type="checkbox"/>	42	Ion Torrent	Sequencing			1
<input type="checkbox"/>	38	Illumina	Approved			1
<input type="checkbox"/>	40	Illumina	Approved			1
<input type="checkbox"/>	39	Illumina	Approved			1
<input type="checkbox"/>	37	Ion Torrent	Submitted			1
<input type="checkbox"/>	18	Illumina	Sequencing			4
<input type="checkbox"/>	25	Illumina	Sequencing			1
<input type="checkbox"/>	23	Illumina	Sequencing			2
<input type="checkbox"/>	22	Illumina	Sequencing			1
<input type="checkbox"/>	21	Pacific Bio	Sequencing			4
<input type="checkbox"/>	17	Ion Torrent	In preparation			1
<input type="checkbox"/>	20	Ion Torrent	Pending			0
<input type="checkbox"/>	19	Illumina	Sequencing			8
<input type="checkbox"/>	16	Illumina	Sequencing			8

By State

- All
- Approved
- Failed
- In preparation
- Pending
- Sequencing
- Submitted
- Succeeded

By Sequencing platform

- All
- Illumina
- Ion Torrent
- Pacific Bio

By experiment type

- All
- DNA tagged sequencing
- RNA-Seq
- RNA CAGE
- DNA ChIPSeq
- RNA SAGE
- DNA whole exome sequencing
- DNA whole genome
- miRNA
- DNA custom capture sequencing
- RNA custom capture sequencing

**Recent Actions**

- 23 | Illumina | RNA tagged SAGE
- Run design requirement for Pacific Bio
- Quality control only
- 1 | Illumina | RNA-Seq
- 318 chip + 200 cycles

Figure 7: LIMS order panel.

The screenshot shows the 'NextGen Sequencing Management System' administration interface. At the top, there is a navigation bar with 'Home', 'Orders', 'Runs', 'Admin', and 'About'. Below this is a 'Systems administration' section. On the left, a table lists various system components, each with 'Add' and 'Change' options. On the right, a 'Recent Actions' panel shows a list of recent system events, including orders and run design requirements.

Systems	
Barcode groups	<a href="#">Add</a> <a href="#">Change</a>
Barcodes	<a href="#">Add</a> <a href="#">Change</a>
Cluster stations	<a href="#">Add</a> <a href="#">Change</a>
Cycles or time choices	<a href="#">Add</a> <a href="#">Change</a>
Data analysis choices	<a href="#">Add</a> <a href="#">Change</a>
Experiment types	<a href="#">Add</a> <a href="#">Change</a>
Fragment size choices	<a href="#">Add</a> <a href="#">Change</a>
Platforms	<a href="#">Add</a> <a href="#">Change</a>
Run design requirements	<a href="#">Add</a> <a href="#">Change</a>
Sample preparation choices	<a href="#">Add</a> <a href="#">Change</a>
Sample requirements	<a href="#">Add</a> <a href="#">Change</a>
Sample species choices	<a href="#">Add</a> <a href="#">Change</a>
Sample type choices	<a href="#">Add</a> <a href="#">Change</a>
Units	<a href="#">Add</a> <a href="#">Change</a>

Recent Actions	
My Actions	
<a href="#">23   Illumina   RNA tagged SAGE</a>	Order
<a href="#">Run design requirement for Pacific Bio</a>	Run design requirement
<a href="#">User</a>	User
<a href="#">[Baros]</a>	User
<a href="#">Quality control only</a>	Data analysis choice
<a href="#">1   Illumina   RNA-Seq</a>	Order
<a href="#">User</a>	User
<a href="#">310 chip + 200 cycles</a>	Cycles or time choice

At the bottom of the page, there are links for 'User Profile', 'Feedback', and 'Links'.

Figure 8: LIMS administration.

***DVD***

The *Diagnostic Variant Database*.

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

**Martijn Vermaat, David van Enckevort, Hailiang Mei**



## *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**

*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**

## *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**

## *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**

## *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**

## *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**

## *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**

## *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**



## *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**

## *NGS Introduction Course*

Full day to get people acquainted with NGS data analysis.

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

**M. van Galen**

## *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**

## *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**

## *Clusters*

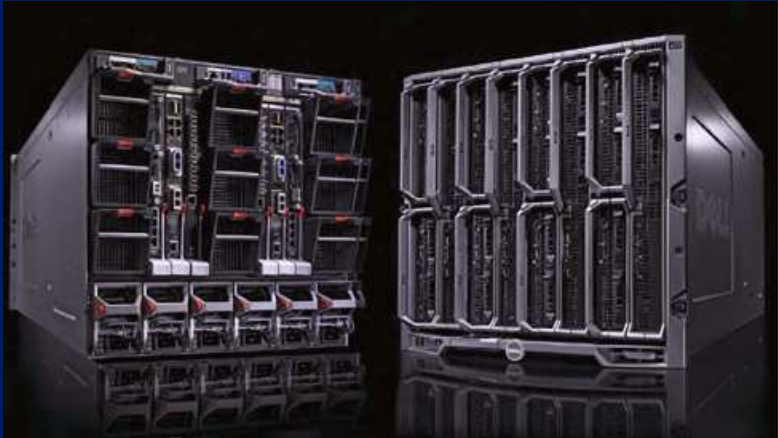


Figure 9: Dell M610 blade server

## *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.

## *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

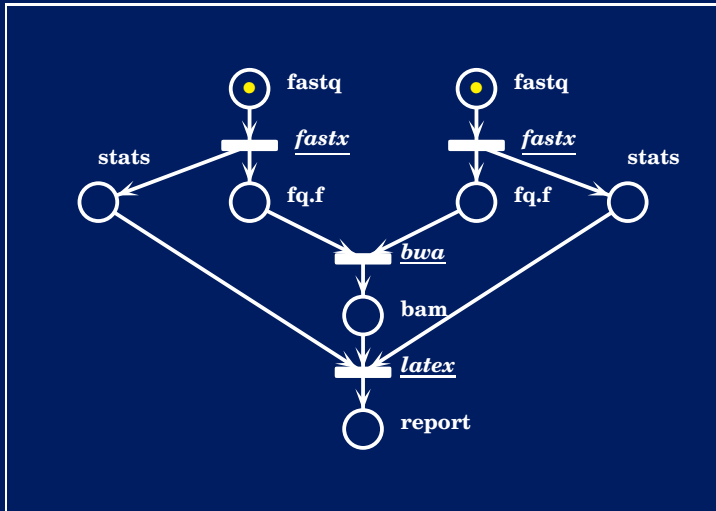


Figure 10: A parallel workflow



# Full data analysis

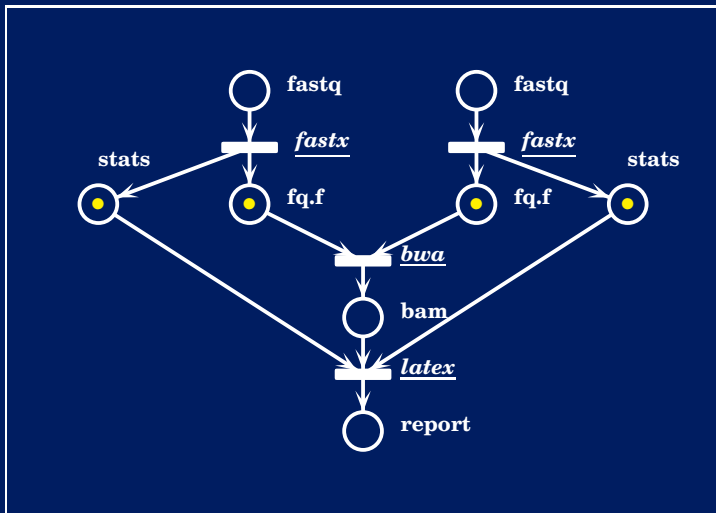


Figure 10: A parallel workflow

# Full data analysis

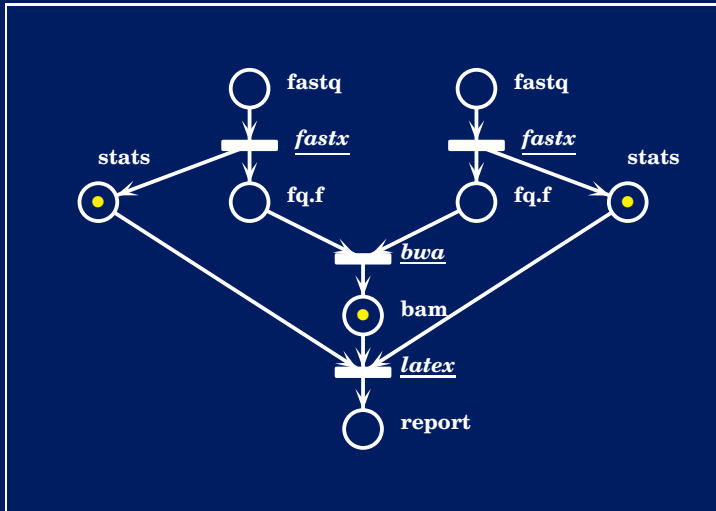


Figure 10: A parallel workflow

# Full data analysis

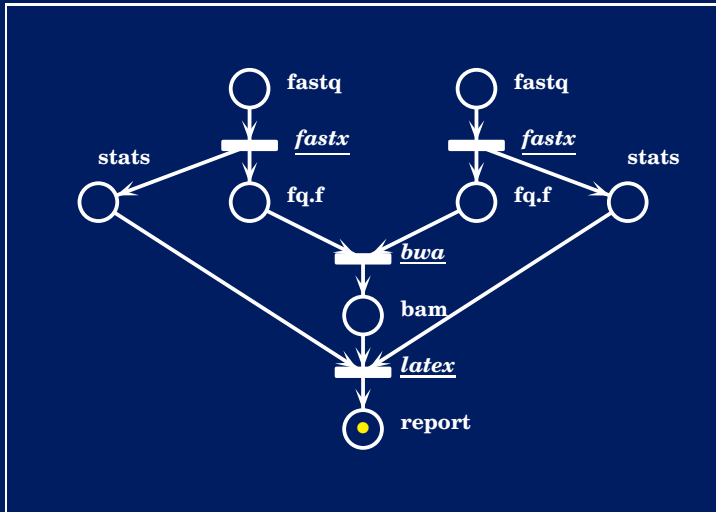


Figure 10: A parallel workflow

## The GAPSS3 workflow

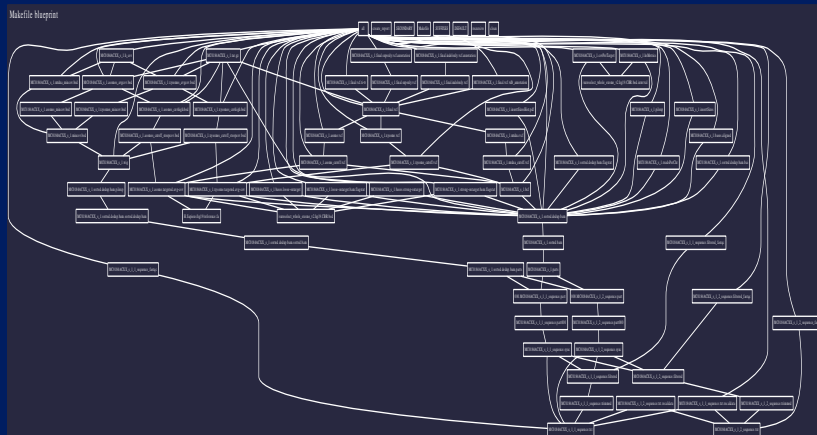


Figure 11: GAPSS3

## The GAPSS3 workflow

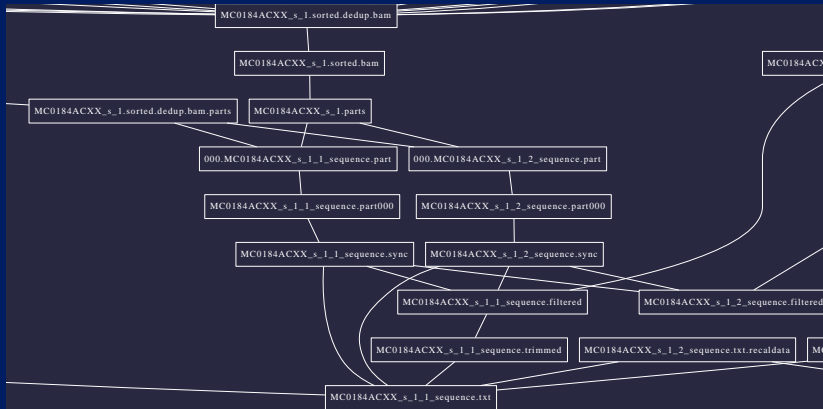


Figure 12: Part of GAPSS3

***GAPSS report***

18 pages with details about the 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.

*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.
  - ...



## *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

## *Metrics for NGS data files*

Distance between wiggle files.

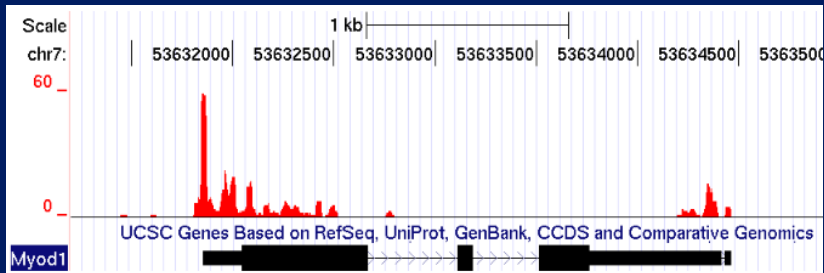


Figure 13: Wiggle file.

Pairwise comparison based on the *multiset* distance measure.

## Metrics for NGS data files

Comparing  $k$ -mer profiles.

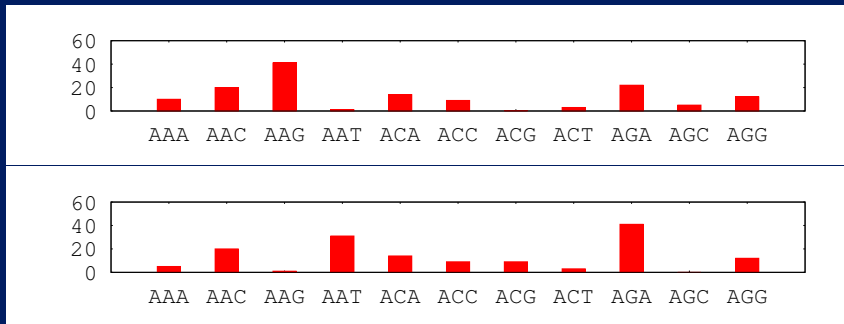


Figure 14: Two  $k$ -mer profiles.

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

***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.

## *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**

***SASC: free support***

Initiative by:

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

***SASC: free support***

Initiative by:

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

Set up:

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



## *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.

### *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.

### *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.

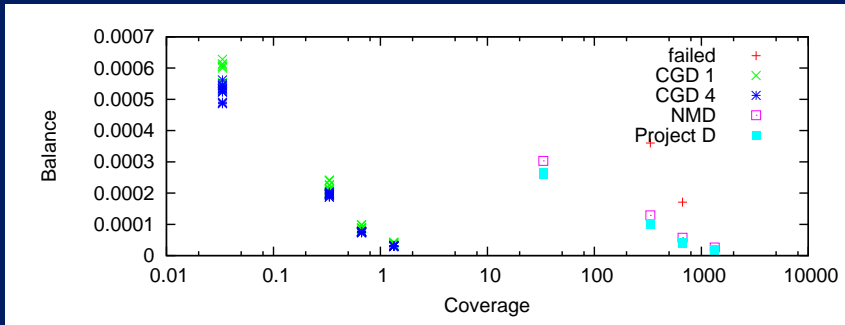


Figure 15: Coverage vs. 9-mer balance.



## 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