



Genome Informatics Unit
Biomedicum Helsinki
<http://www.giu.fi>

FIMM Solexa IT

-Current Status and Future Plans

Juha Knuuttila, Juha Muilu,
Timo Miettinen, Kyösti Sutinen

FIMM Genome and Technology Center
Institute for Molecular Medicine Finland

Topics

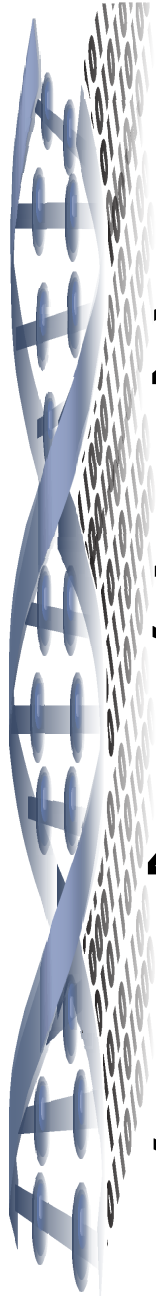
1. About the Solexa Data & Pipeline

2. Hardware solutions

3. Software solutions

4. Future plans

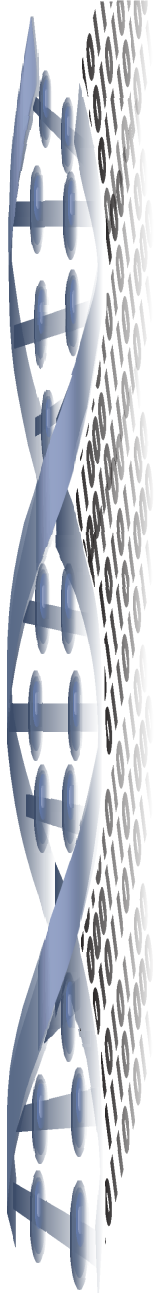
5. Discussion



Volume of Solexa Data: Raw Data

I. Images (also called "Raw Data")

- Typical case / run:
 - single-end run with 36 cycles x 2 rows x 50 columns x 4 colors x 8 lanes x 7.76KB/image = ~ 0,9TB
 - same with 72 cycles (in a paired-end run) ~1,8TB
- In more extreme cases:
 - up to 2 TB / run (single-end)
 - up to 4 TB / run (paired-end)
- 1 run / week => in 3 weeks: 3-12TB of raw data
- This varies with the amount of cycles run
- Lossless JPEG compression to about 50% of the size can be achieved

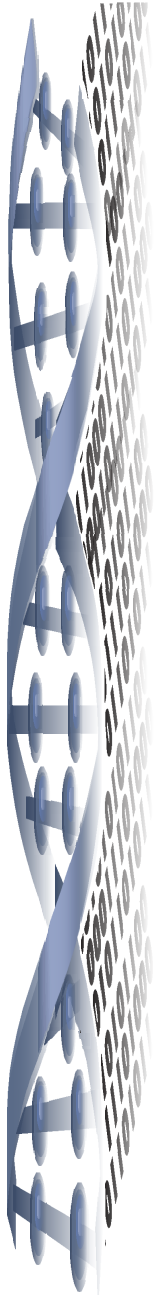


Volume of Solexa Data: Primary Data

II. Traces, intermediate files, visualization files, sequence fragments, assembled sequences, annotations... (also called primary data, or analysis data, or the "end-product")

The volume of this (250 – 600GB) varies with:

1. the pipeline software used
 2. sample quality
 3. reagent quality
 4. amount of cycles run
 5. what will be stored (in FIMM, initially everything but images)
 6. ...
- These are mostly text files, which compress well if necessary



Example: Illumina's Solexa Pipeline Software (GOAT)

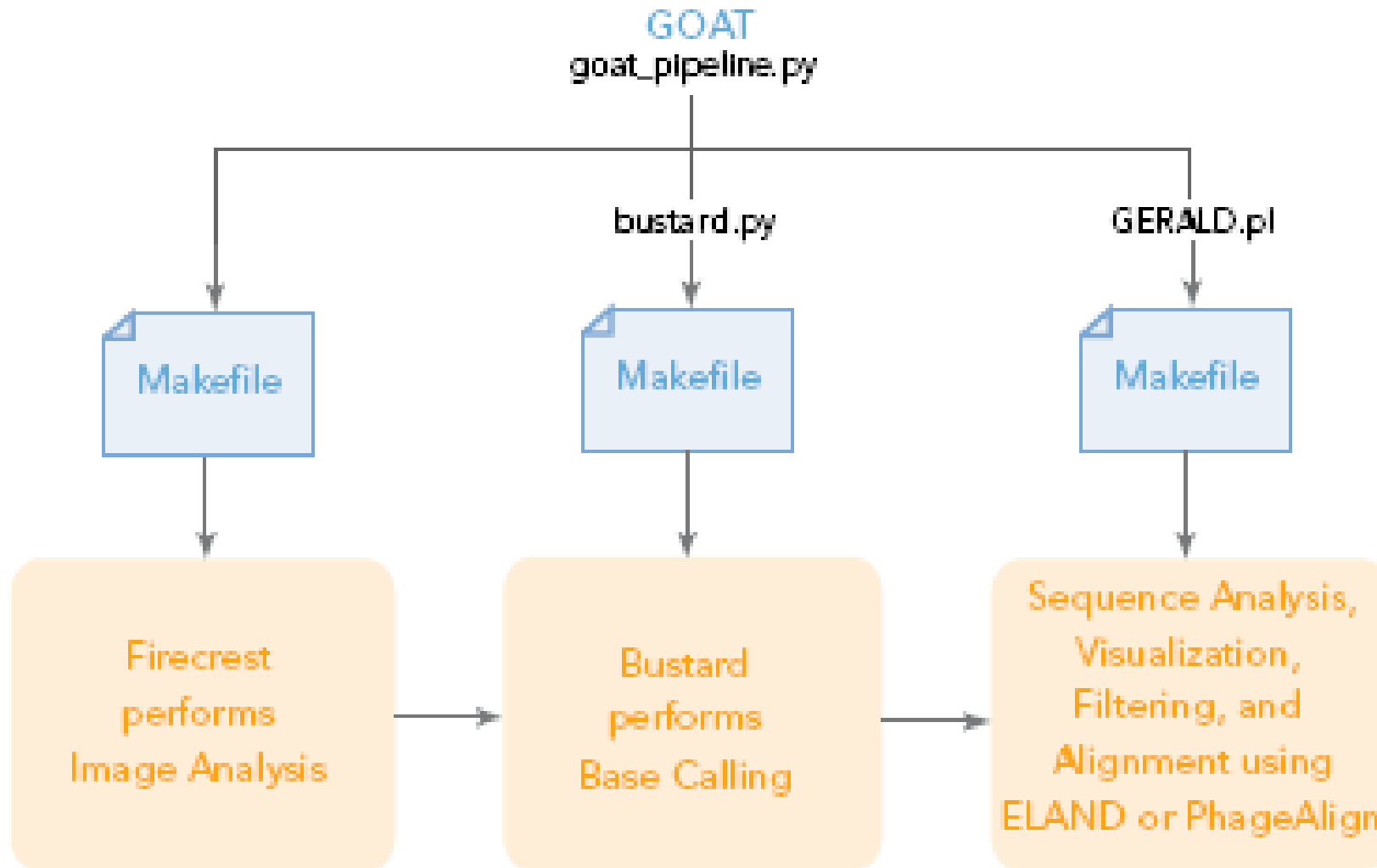


Figure 2 Pipeline Modules

(from Illumina Manual 1003881_Pipeline_User_Guide.pdf)

Topics

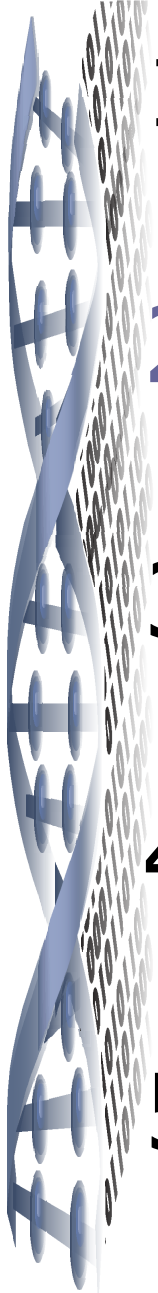
1. About the Solexa Data & Pipeline

2. Hardware solutions

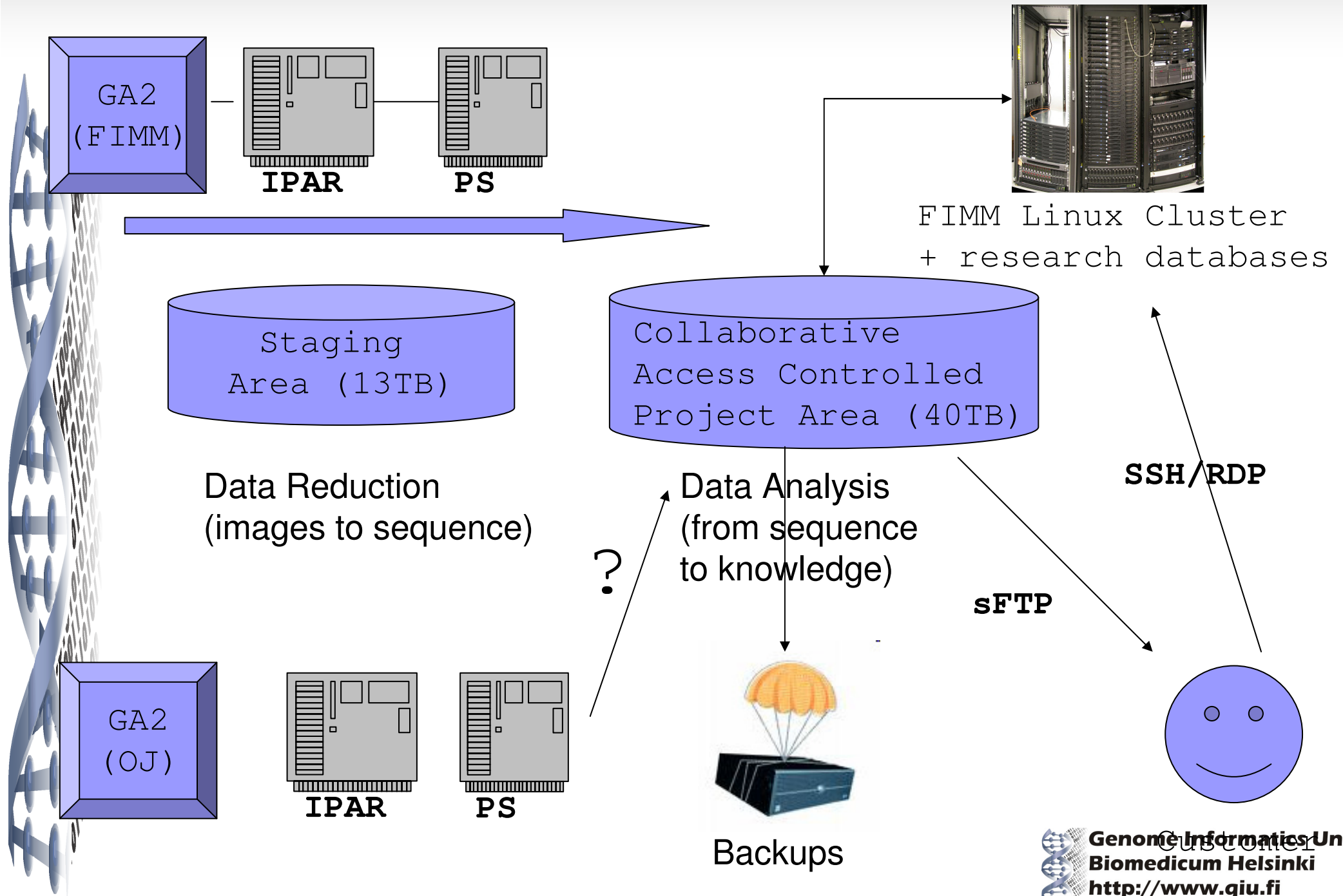
3. Software solutions

4. Future plans

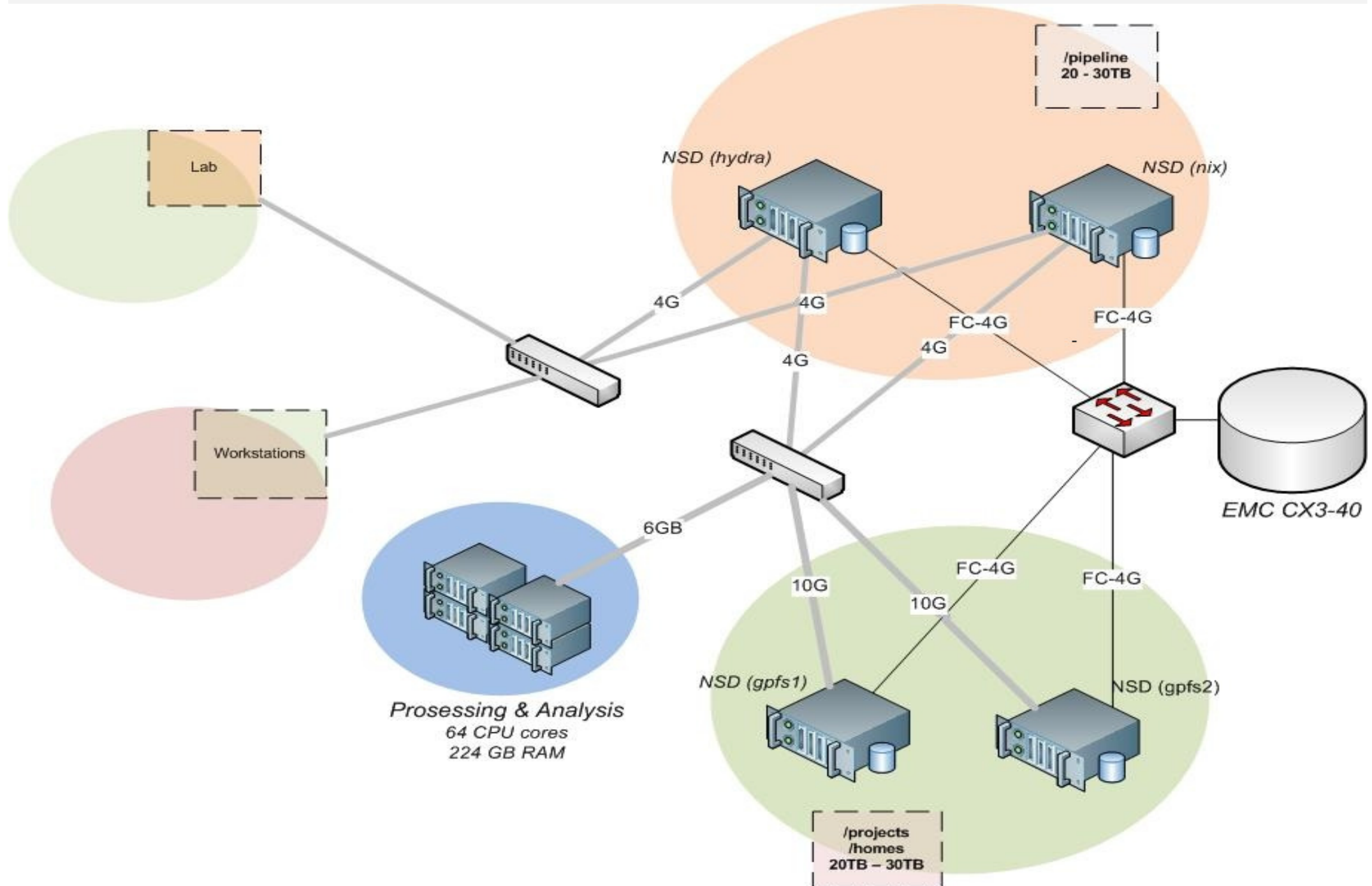
5. Discussion



Hardware



Technical System Overview



IT Hardware Included in the Solexa Order

2 x GA Instrument PC

- Two Dual core 2.66 Ghz
- **1 TB storage**
- Windows XP professional

2 x Integrated Primary Analysis and Reporting (IPAR)

- Two XEON 5450 quad core @ 3 Ghz
- **3TB storage.**
- Windows XP 64bit professional

1 x Genome Analyzer Pipeline Server

- *Four Dual-Core 64-bit Intel® Xeon® Processors 7140M*
- (3.40GHz)
- 32GB memory
- **7TB storage**

FIMM Hardware

Longer Term Primary Data Storage

- 6x DAE 15x 750GB SATA (6x **9TB** = **54TB** effective w/ raid6)

Pipeline Data Processing

- 2x Blade servers 16GB 2x QuadCore for *primary data reduction*
- 2x Blade servers 32GB 2x QuadCore for *assembly and analysis*
- 1x Blade server 32GB 2x QuadCore for *testing*

Pipeline Configuration

Tier 1 Storage - Staging area

Mountpoint: */pipeline*

Capacity: **13TB** ?

Tier 2 Storage - Project folders

Mountpoint: */gpfs/projects*

Capacity: **~40TB** ?

Backups

LTO4 Ultrium tape technology

Capacity: **80TB**

Topics

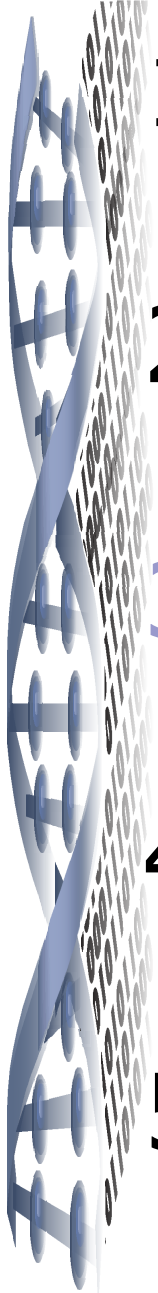
1. About the Solexa Data & Pipeline

2. Hardware solutions

3. Software solutions

4. Future plans

5. Discussion



Data Management and Analysis Pipelines

Data Management (for sample and run tracking)

- GLIMS (extend current functionality to Solexa)
 - sample tracking
 - run tracking
 - project mgmt

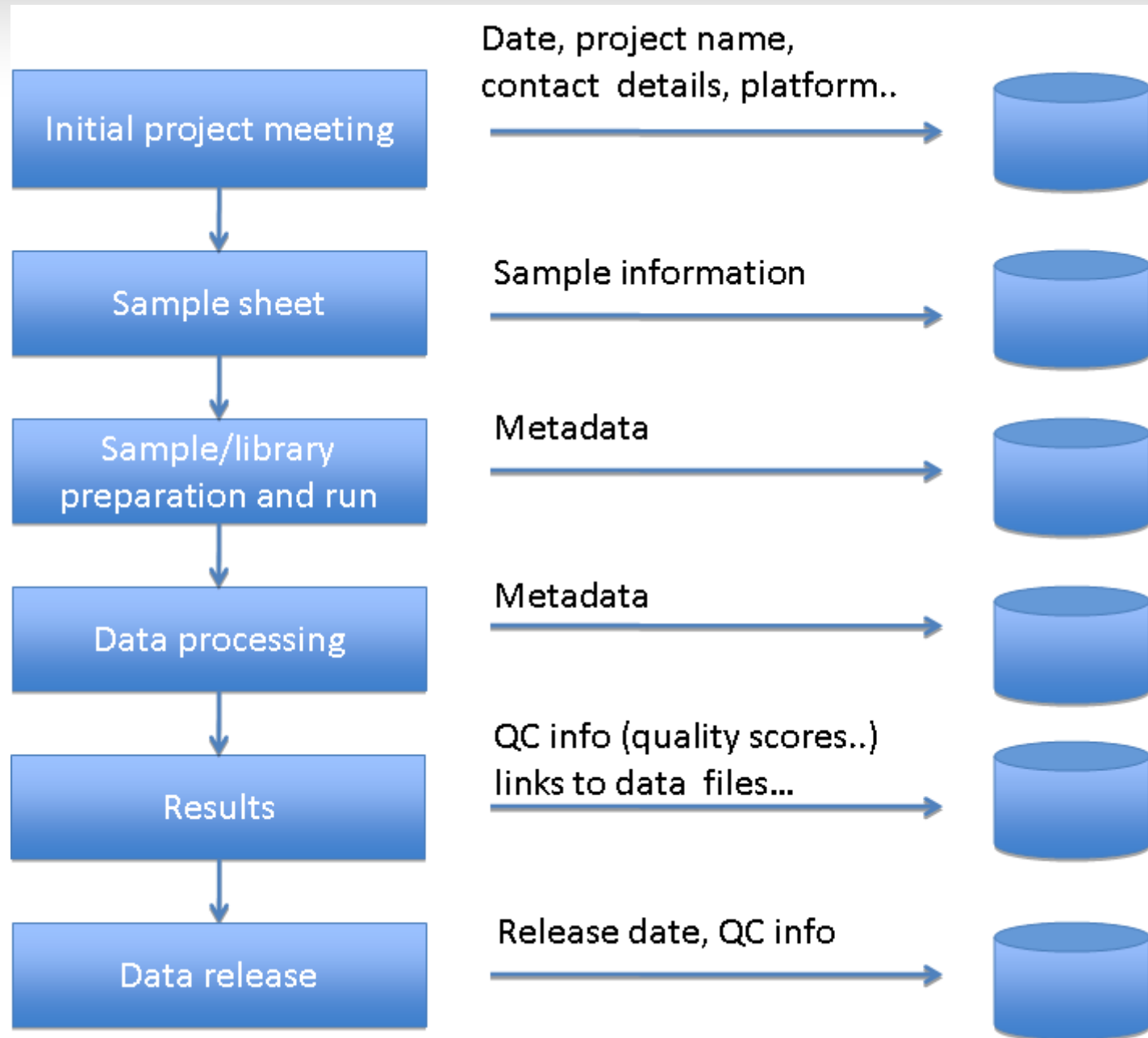
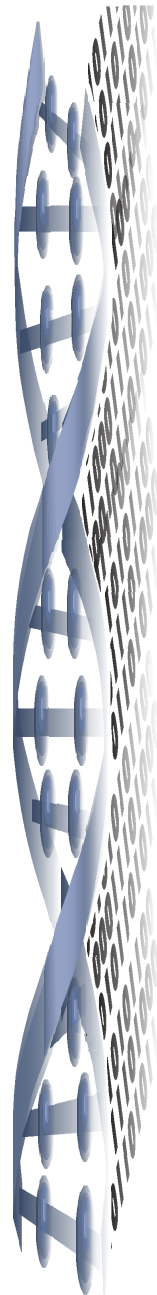
Basic Analysis (from images to sequence)

- GOAT (installed)
- MaQ (to be tested soon)
- Swift (to be tested later)

Downstream Analysis (from sequence to knowledge)

- Software XYZ
- ... More can be installed as needed
- jobs can be assigned to high memory queues

Utilizing Existing FIMM Infrastructures



Topics

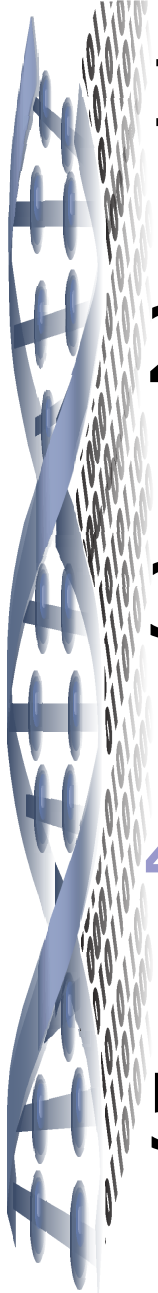
1. About the Solexa Data & Pipeline

2. Hardware solutions

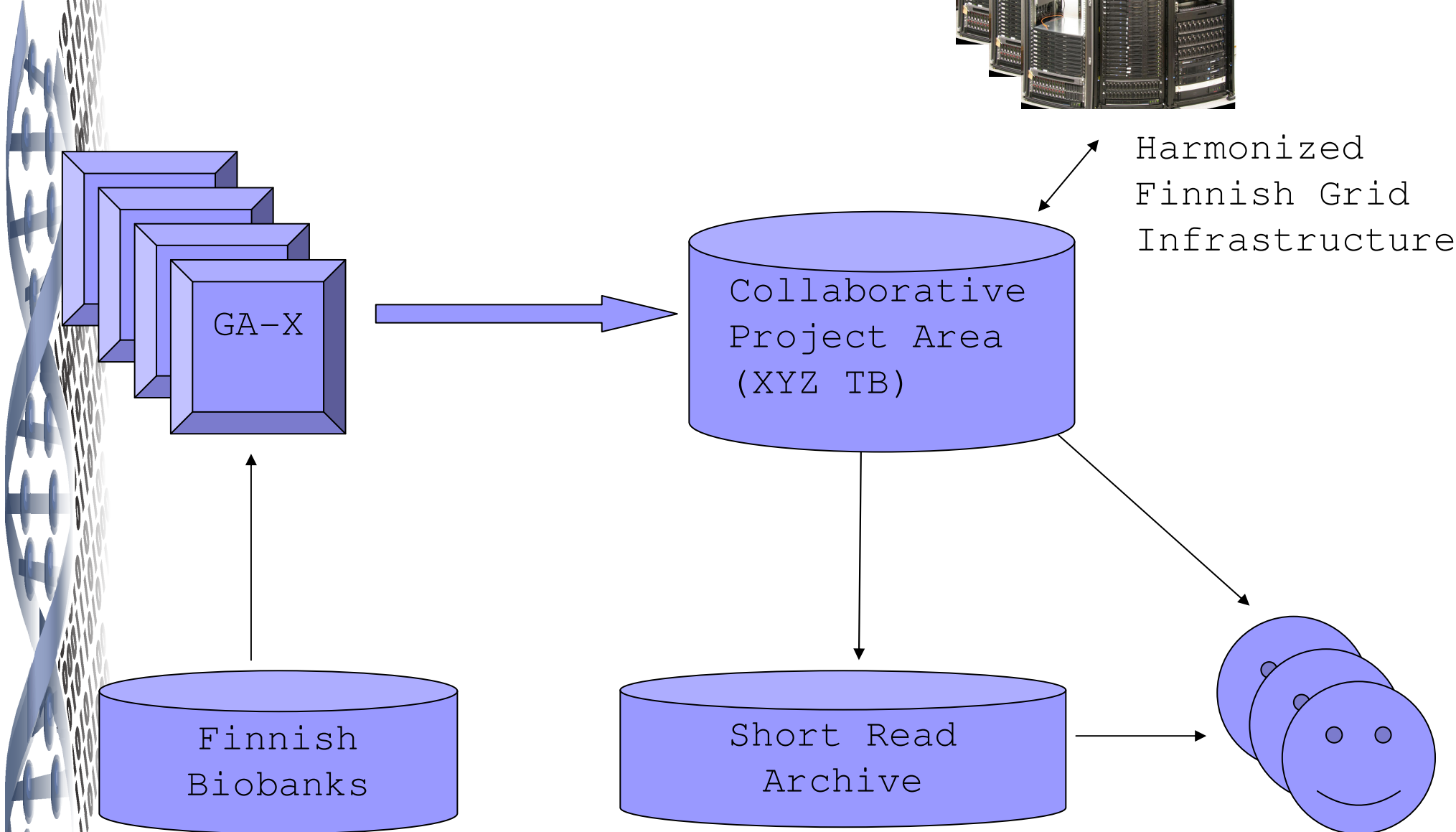
3. Software solutions

4. Future plans

5. Discussion



Possible Future Scenario



Topics

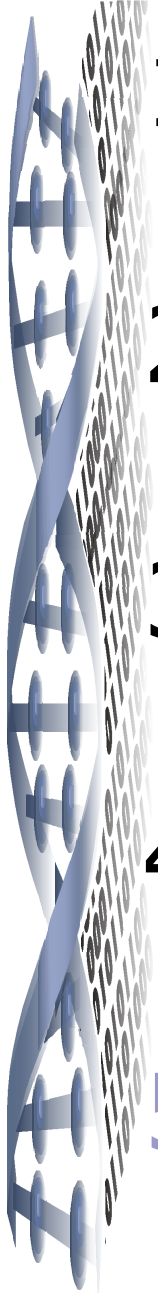
1. About the Solexa Data & Pipeline

2. Hardware solutions

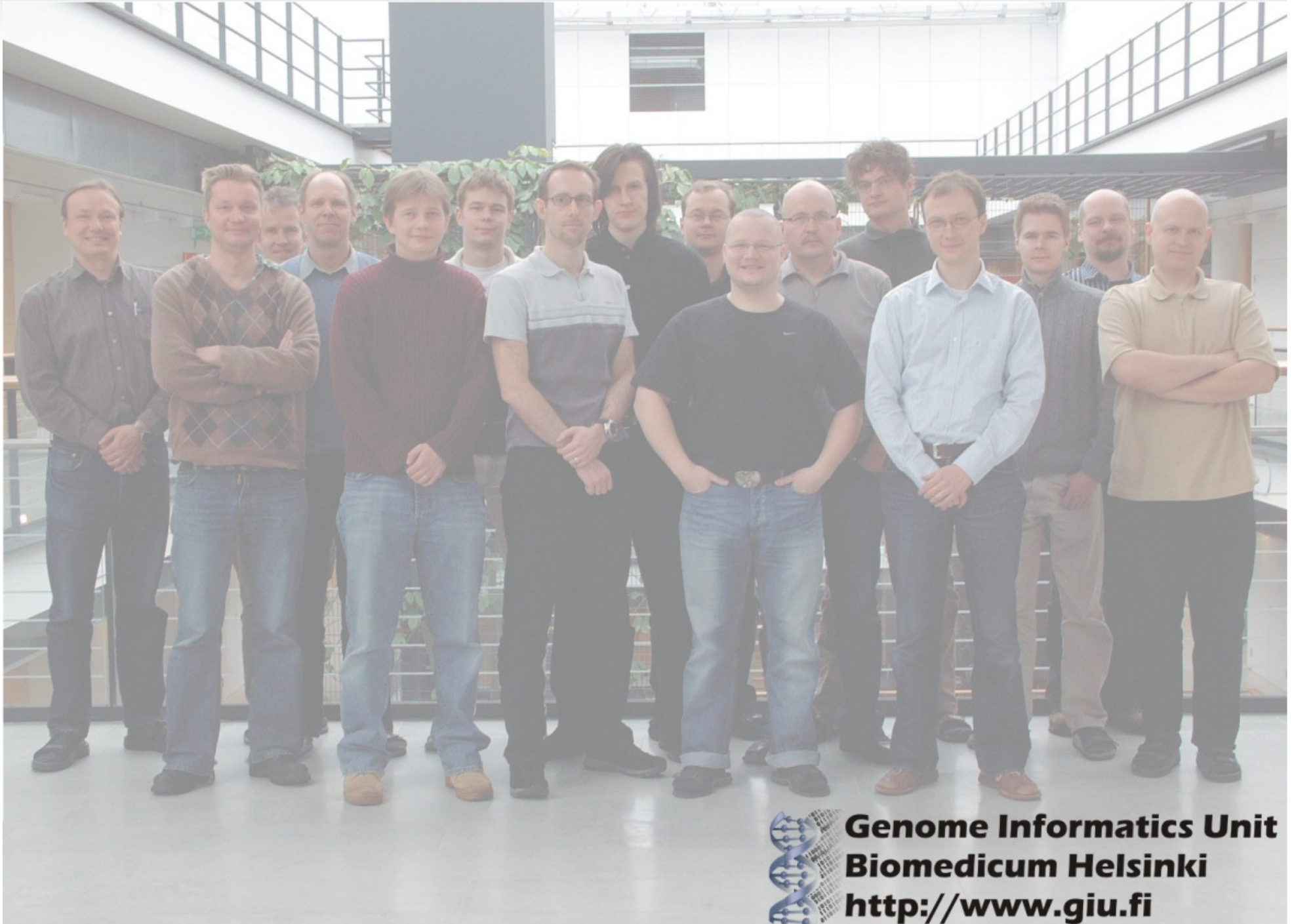
3. Software solutions

4. Future plans

5. Discussion



Acknowledgements



Genome Informatics Unit
Biomedicum Helsinki
<http://www.giu.fi>