The Challenge of Big Data

Keeping track of thousands of samples and many terabytes of data.

Experience from SciLifeLab Stockholm.
Science for Life Laboratory (SciLifeLab)

Vision

"To create a internationally leading center for high-throughput research in molecular biosciences and medicine, with special focus on disease biology and translational medicine"

- A national resource center
- Technology for translational medicine
- Close collaboration with Swedish hospitals
- Efficient contact with industry and society
- Strong part in European research infrastructures
Main Technology Platforms, SciLifeLab Stockholm

- Genomics, Massive Parallel Sequencing
- Proteomics and Protein Profiling
- Bioimaging and Functional Biology
- Bioinformatics and Systems Biology
Antibody-based Protein profiling

- The world's largest resource of antibodies (HPA-project funded by KAW)
- Four SciLifeLab facilities for translational research based on this resource

Tissue profiles (IHC) human tissues (SciLifeLab Uppsala)
Subcellular profiles (IF) in cell lines (SciLifeLab Sthlm)
Biobank profiles in patient cohorts (SciLifeLab Sthlm)
Tissue profiles in rodent models (SciLifeLab Sthlm)
The ALM and CLICK imaging facilities at SciLifeLab Stockholm

“Our mission is to become one of the leading competence centers for light microscopy in Europe”

**ALM**

*Advanced Light Microscopy*

Key application: Super resolution light microscopy

Director: Hjalmar Brismar
Manager: Hans Blom

**CLICK**

*Center for Live Imaging of Cells at KI*

Key application: Live cell microscopy

Director: Per Uhlén
Manager: Göran Månsson

Additional competence and resources:
1 application specialist, 2 postdocs, 3 postgrad students and 1 BMA.
Mass spectrometry facility (2011)

Facility Manager: Janne Lehtiö (KS)

Responsible Platform Director: Roman Zubarev (KI)

Personnel 2011: 15 FTEs (other funding include SLL, VR, EU, Cancerfonden)

Scientific focus: Clinical proteomics (analysis of clinical samples using mass spectrometry)

Ambition: To be internationally competitive in the field of clinical proteomics, focus on proteomics for Individualized medicin.
• NGI; formerly SNISS
• Service organisation
  • Sweden, Nordic countries, EU
  • Project-based
  • Fee-for-service
    • Academics: Cost of consumables
    • Commercial: Full cost
• Technology platform: Massively parallel sequencing
• From DNA/RNA to sequence
  • Project planning, experiment design
  • “Best practice” bioinformatics analysis
  • Deeper project involvement: WABI
This is the customer portal for National Genomics Infrastructure hosted by SciLifeLab (NGI SciLifeLab).

NGI SciLifeLab primarily serves projects originating in Swedish research groups, academic as well as commercial. Projects from the other Nordic countries may be admitted under certain conditions, such as collaborations. Other European projects will be served only if they are part of a EU-sponsored project such as ESGi. Non-European customers cannot, unfortunately, use NGI SciLifeLab.

An brief description of the available services can be found in Overview. See also the FAQ (Frequently Asked Questions).

Starting January 1st 2013, the National Genomics Infrastructure (NGI) was launched, originating from the Swedish Research Council RFI infrastructure SNiSS. NGI follows the VR RFI guidelines for national infrastructures and is supported by the Swedish Research Council (VR), host universities (KTH, UU), and SciLifeLab. NGI will include facilities established by Profs Ulf Gyllensten, Ann-Christine Svanen and Joakim Lundeberg. In addition, the Knut and Alice Wallenberg Foundation (KAW) has provided critical capital support for new instruments and computational infrastructure.
Genomics Platform, SciLifeLab Stockholm

- One of three NGI nodes
  - Two nodes in Uppsala
- Current machinery
  - 5 Illumina HiSeq 2500
  - 2 Roche Genome Sequencer 454 FLX+
  - 1 Life Technologies SOLiD 550 XL
  - 2 Illumina MiSeq
  - 1 Argus Optical Mapper
Which applications?
Distribution of projects, from 2012-01-01
# projects and samples started
### Achievements 2012

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projects completed</td>
<td>202</td>
</tr>
<tr>
<td>Samples processed</td>
<td>4539</td>
</tr>
</tbody>
</table>
# Sequencing applications 2012

<table>
<thead>
<tr>
<th>Sequencing Application</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole genome</td>
<td>de novo</td>
</tr>
<tr>
<td>Whole genome</td>
<td>Re-sequencing</td>
</tr>
<tr>
<td>Targeted genome</td>
<td>Exome re-sequencing</td>
</tr>
<tr>
<td>RNA sequencing</td>
<td>Digital profiles</td>
</tr>
<tr>
<td>Metagenomics/transcriptomics</td>
<td>Shot-gun</td>
</tr>
<tr>
<td>Signature sequencing</td>
<td>ChIP-Seq, epigenomics</td>
</tr>
</tbody>
</table>
### Genomics: Infrastructure Experimental

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Read length</th>
<th>Data/run</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina HiSeq 2500 rapid mode</td>
<td>2*150 bp</td>
<td>180 Gb</td>
<td>2 days</td>
</tr>
<tr>
<td>Illumina HiSeq 2000 hi output mode</td>
<td>2*100 bp</td>
<td>200-300 Gb</td>
<td>11 days</td>
</tr>
</tbody>
</table>

Agilent Bravo workstation, Caliper GX and XT, REM, EZ bead system, Robotics, PCRs, Bioanalyzer, Covaris etc

**Illumina’s new HiSeq 2000. The company plans to start shipping the instrument next month.**
New instruments introduced 2012...

January 2012 – Announcements

Life Technologies - Ion Proton
Illumina – HiSeq2500

Short reads
Human genome for 1 000 USD
Human genome in a day
Launch Q3, Q4 2012

February 2012 – Announcements

Oxford Nanopore

Long reads
USB sequencing - MiniION
Human genome less than 1 000 USD – GridION
Human genome in 15 minutes – GridION
Launch Q3, Q4 2012
Informatics challenge

LIMS (lab data)

Data management (sequence data)

Data analysis (research data)
Genomics staff and resource persons

Sequence group

Anders Jemt
Christian
Natanaelsson
Desirée von Tell
Elsie Castro
Emma Sernstad
Joel Gruselius

Application specialists

Genome sequencing (re-sequencing) - Tiina Robins
Seq cap (re-sequencing) - Anna Sköllermo
De novo sequencing (mate-pairs) - Elsie Castro
RNA sequencing - Emma Sernstad
ChIP Seq/Signature - Mattias Ormestad
Automation - Joel Gruselius
Quality Assurance (ISO) system - Kicki Holmberg
Emerging technologies/protocols - Kicki Holmberg
Sample Coordinator - Mattias Ormestad

PhD/postdoc specialist support (bioinfo)

Genome sequencing (re-sequencing) – Henrik, Pelin, Lilia
Seq cap (re-sequencing) – Magnus, Olof
De novo sequencing (mate-pairs) – Lars
RNA sequencing – Beata, Henrik
ChIP Seq/Signature – Patrik S., Olof
Metagenomics/Amplicon – Anders A., Daniel L.

PhD/postdoc specialist support (wet lab)

Genome sequencing (re-sequencing) – Henrik, Pelin,…
Seq cap (re-sequencing) – Magnus, Erik B…
De novo sequencing (mate-pairs) – Sverker, …
RNA sequencing – Beata, Henrik
ChIP Seq/Signature – Patrik S.,
ChIP Seq/Amplicon – Anders A…

Analysis group

Genome sequencing (re-sequencing) – Per Unneberg
Seq cap (re-sequencing) – Per Unneberg
De novo sequencing – Ellen Sherwood
De novo assembly – Björn Nystedt
Spruce project – Andrej Alekseenko
Small genome assembly – Alexej Kutsenko
RNA sequencing – Mikael Huss
ChIP Seq/Signature – Mikael Huss

Informatics group

Systems Architect – Per Kraulis
DevOp – Roma Valls Guimera
Bioinfo Systems Op – Pontus Larsson
Systems programmer – Valentine Svensson
Bioinfo analyst – Paul Costea
Basic "best practice" bioinfo analysis

NGS Bioinformatics

- collection of reads
- basic draft assembly

De Novo

- mapped reads
- defined target model
- list of variation (SNP’s etc)

Variation Analysis

- mapped reads
- defined gene model
- quantified transcript levels (RPKM)

Functional Genomics

- quality control
- mechanisms for data delivery
- tools and methods for further analysis
- access to know-how

Informatics infrastructure
The way we work

**Fee-for-service**
- best practice
- quality
- attributes of sequence data
- variation
- quantitation
- draft assemblies

**Collaborations**
- project group
- involvement in design
- assigned bioinformatician
- iterative analyses
- evaluation and conclusion
- suggestions for validation

Descriptive analyses

Explorative analyses with context
Circos Plot of U2OS genome
Current whole genome sequencing (de novo) projects

- Skin fungus (8 Mbp)
- Surirella Diatom (100 Mb)
- C. tentans (insect) (150 Mbp)
- Herring (1 Gbp)
- Spruce (20 Gbp)
Barnacle de novo sequence project
Balanus improvisus
Anders Blomberg, GU
LIMS: Concepts

Entities
- Project
- Sample
- Collection
  - Plate
  - Workset
- Result
  - Data file(s)
  - QC/QA
- Analysis

Processes
- Workflow
- Task
- Process

Tracking
- Current state
- History: Who did what when?
- To-do lists
LIMS: The problem

- The ontology is fundamental
  - Which entities?
  - What relations?
  - Decisions required! Constraints!
  - Must be explicit!

- For example: What is a sample?
  - Original sample (customer)
  - Derived sample (in lab)
  - Dilutions, aliquots, transformations?
  - Sample material vs. sample tube?
  - Identifier scheme
LIMS: Moving target

- Instruments
  - New versions
  - New types

- Procedures
  - Workflows
  - Protocols

- Personnel
  - Turnover
  - Changing roles

- Constraints
  - Protocol change during workflow?

- Trackability
  - Exactly which protocol?
  - Relation to older
LIMS: Integration

- Customers
- Management
- Instruments
- QA system
- Status displays
- Invoices
- Downstream processing
- Result delivery
Interfaces required for people and programs!

Web interface, obviously... (but custom clients are still common)

API: few LIMS have it!

Ideal: Web and API the same; RESTful!

Human interface: Do user interaction testing!
“Customization” is a positive term...

But: More customization
    == more work
    == more maintenance

And: Customization can never undo fundamental ontology!

(and don't believe anyone who says: “Our LIMS can be customized to do anything!”)
LIMS: Introduction into the lab

Two alternative strategies:

1) Customize LIMS to fit current work patterns
2) Modify work patterns to fit LIMS

In practice, a compromise must be found.

Get acceptance ("buy-in") from lab personnel
- Essential
- LIMS must clearly simplify the work
  - If not now, then in foreseeable future...
- If the lab doesn't "get" the LIMS, it is dead...
LIMS: Google Docs

- Current system at SciLifeLab Stockholm
- Excel-ish
- Access control
- Central storage
- Web interface
- Auto-backup

But clearly suboptimal!
- No proper database
- Trackability problems
- Cut-and-paste mistakes
LIMS: GeneSifter (Geospiza)

- Attempt 2010 at SciLifeLab Stockholm
- Web interface
- Nice licensing model
- For NextGen Sequencing

Abandoned early 2011
- Ontology problems
  – Muddled
  – Terminology issues, or not?
- User interaction problems
  – “Where am I?”
  – ”Where did that value go?”
LIMS: slog

• Project during 2011
• Home-grown system
• Python, CouchDB
• Web interface
• RESTful API

• Data part of ontology finished
• Problem: Workflow part
• Problem: Interface for plate handling
• Abandoned late 2011
  • GenoLogics
LIMS: GenoLogics

- Attempt 2012 at SciLifeLab Stockholm
- Java client (!)
- Web interface in development
- For NextGen Sequencing
- Clear Ontology
- Proper RESTful API
- Expensive
Contact information

http://www.scilifelab.se

http://portal.scilifelab.se/genomics

Genomics Stockholm
Per Kraulis (Genomics IT) per.kraulis@scilifelab.se