Enabling Translational Medicine with e-Science

Ola Spjuth
ola.spjuth@ki.se
ola.spjuth@farmbio.uu.se
Cancer

Normal cells accumulate mutations

- Chromosomes
- 1 mutation
- 2 mutations
- 3 mutations
- 4 mutations

Normal cell → Malignant cell

Cancer cells divide excessively and form tumors

Malignant cancer cells can invade other tissues
Cancer cells have altered genomes

- Each cancer has its own genome
  → Calls for individualized/stratified diagnostics and treatment

Source: www.path.cam.ac.uk/~pawefish/BreastCellLineDescriptions/HCC38.html
Today: We have access to high-throughput technologies to study biological phenomena
New challenges: Data management and analysis

- Storage
- Analysis methods, pipelines
- Scaling
- Automation
- Data integration, security
- etc.
Translational medicine

- **Aim:** *Turn Basic Research into Medicines and Treatments*
  - “from bench to bedside”
  - “from laboratory to clinic”

- Traditionally a slow process
  - May take 10 -20 years for original research to translate to routine medical practice
e-Science in translational medicine

- e-Science solutions are desperately needed in order to translate high-throughput technologies into clinical settings (diagnostics, treatment, …)
  - e-infrastructure (computers, storage, networks, frameworks)
  - Methods, workflows/pipelines
  - Operations, experience and expertise

- e-Science can aid translational medical research
  - Simulation and prediction models
  - Large-scale collaborative, integrative research
2012-2014: e-Science for Cancer Prevention and Control
- a SeRC flagship project

Use statistical modeling and data integration in cancer research:

→ Individualized prevention strategies
→ Individualized treatments

PIs: Juni Palmgren & Jan-Eric Litton
Cancer Risk Prediction Centre, CRiSP
VR Linnaeus 2008-2018

Breast cancer is the most common cancer among women in Sweden with almost 8,000 new cases annually. In Sweden, 1,500 women die from breast cancer yearly but there is a remarkable difference between outcomes of localized vs advanced disease. Prostate cancer is the most common cancer among men in Sweden today and yearly almost 10,000 new cases are diagnosed. Despite the old age of onset, the morbidity and mortality of this cancer is substantial with more than 2,500 deaths annually.

We know that cancer mortality can be reduced if cases are detected and treated early, but there is a problem with over-diagnosis and over-treatment. What if we instead could predict the risk for aggressive cancers? Our research focuses on understanding cancer risk and how to design individualized prevention strategies.

PI: Per Hall
PI: Henrik Grönberg

Personalized Cancer Prevention!
Personalized screening; Why?

- Early detection is important.
- Efficiency of current practice for early detection of breast and prostate cancer is questioned!

We have:
- Organized mammographic screening
- Widespread opportunistic PSA testing

- Rates of detection of slow growing cancers increase
  - Beware of over-diagnosis, overtreatment, increasing cost and increase in side effects!
- Aggressive cancers and mortality do not decrease enough
e-Science components in research on personalized screening

highlight 5 eCPC case studies

Data Availability
Breast density as risk factor
Biomarkers for breast and prostate cancer
Designing screening programs
Predicting risk for prostate cancer
Data Integration
SAIL – Sample Availability System

- Overview of data – what is available
- Plan studies, investigate data available for subset of patients/samples across data archives such as biobanks

Breast density as risk factor
Imaging technology and computational techniques

- Applying e-Science methods to image analysis data can feed into prediction modeling and help guide screening and preventative strategies.

- eCPC developed new measures of mammographic density, with the aim of providing stronger risk factors for breast cancer.

Keith Humphreys
Biomarker discovery and validation

- Validated genetic and protein markers conveying prostate cancer risk
- Studied genomic profile of breast cancers

→ Data-intensive bioinformatics making use of national HPC e-Infrastructures (SNIC)
eCPC Microsimulation Model

- Simulate individual event histories
- Aggregate to population level
- Assess the model fit (calibration)
- Evaluate effects of intervention

Example: Reduction in Cervical Cancer Incidence

(Bodhager Diebert et al. 2007)

Vaccination only

Screening every 2 years only

Screening every 2 years with vaccine
eScience and eCPC microsimulation

- Used OpenMP/MPI or PGAS for Bayesian calibration of the microsimulation on HPC
- Compared HPC with MapReduce for the microsimulation
- Trans-compile from C++ to WebAssembly/JavaScript for microsimulation on the browser
- Compared client-server and client-only web interfaces for the microsimulation
Microsimulation and STHLM-2

- **Stockholm 2 study**: Find out how heredity, environment and genetic changes can be used to predict the risk of developing prostate cancer.
  → 50,000 participants

- Microsimulation was developed and used in Sthlm 2 for evaluating and planning screening strategies
  → Resulted in the design of Stockholm 3 study: Develop strategies to improve risk predictions
Sthlm 3 study: Predicting risk for prostate cancer based on PSA + biomarkers

- Constructed and validated Stockholm 3 model (S3M)
  - Prediction algorithm for predicting a man’s risk for having prostate cancer, which shows improved specificity (given constant sensitivity) over prostate specific antigen (PSA) testing alone in the population-based STHLM3 diagnostic trial
  - 50,000 men in Study Q1 2015
  - Thermo Fisher customized chip

- The developed microsimulation model will be used for health economics calculations
2015: eCPC sharpens the focus on translational medicine

- Adding new prioritized field: **Clinical sequencing**
  - How can new high-throughput sequencing technologies aid cancer diagnostics and treatments?

- Joint collaboration, SeRC and eSSENCE

PI: Juni Palmgren
eCPC research on clinical sequencing

Applied e-Science research

e-Science methods development

Individualized diagnostics

Prediction models, machine learning

e-Infrastructure development

Automation, Big Data

Posters: 18 + 28

Posters: 41 + 46
Clinical sequencing supported by eCPC

Stockholm: ClinSeq
- Short read technology
- Pan-cancer approach
- www.clinseq.se

Uppsala
- Long read technology
- Targeted approaches
ClinSeq pipeline

- Surgery
- Pathology and DNA/RNA extraction
- Sample prep
- Actionable Panel Low-pass WGS RNAseq
- Data analysis
- Deliver report to treating physician

Patient value
Research opportunities

Pan-cancer approach - the same pipeline for all cancers
E-science in ClinSeq

- **HPC / distributed computing**
  - Bioinformatic pre-processing in CLINSEQ: 15 Gb raw data per patient (~100 CPU hours / patient)
  - Automated bioinformatic pipeline that run in cluster environment (e.g. UPPMAX)

- **Machine learning and computational statistics**: Methods development and application
  - Supervised and unsupervised learning across multiple types of high-dimensional data
  - Biomarker discovery (robust methods for variable selection): definition of robust sets of clinically relevant biomarkers

- Integration of in-house data with **public DBs and data sets**
  - The Cancer Genome Atlas (TCGA), clinically relevant mutations (COSMIC, ClinVar)
ClinSeq in breast cancer

- Results suggest that routine clinical biomarkers could be replaced with DNA and RNA sequencing-based diagnostics.

- The ClinSeq profile adds value by providing more detailed diagnostic information (subtype, mutations, transcriptomic grade).

- Prospective validation study is planned (N=500).
Supporting clinical sequencing at Uppsala Academic Hospital

1. Sample registration
2. Sequencing request
3. Sequence delivery
4. Analysis pipelines on high-performance computers
5. Data archiving
6. Results delivery
7. Decision support
8. Treatments and outcomes
9. Data sharing

Sample transfer

Clinician

Knowledge base

Other instances
Chronic Myeloid Leukemia (CML)

<table>
<thead>
<tr>
<th>Sample</th>
<th>T315I</th>
<th>F359C</th>
<th>F359I</th>
<th>H396R</th>
<th>D276G</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>R3740(infection)</td>
<td>88.9</td>
<td>4.2</td>
<td></td>
<td></td>
<td></td>
<td>2013-09-17</td>
</tr>
<tr>
<td>R7394(1m)</td>
<td>94.8</td>
<td>2.2</td>
<td></td>
<td>1</td>
<td>1.8</td>
<td>2013-10-20</td>
</tr>
<tr>
<td>R7840(2m)</td>
<td>13.5</td>
<td>83.1</td>
<td></td>
<td></td>
<td></td>
<td>2013-12-01</td>
</tr>
</tbody>
</table>

Patient 1

External ID: QRX00027
Age: 68
Sex: F
e-Science in eCPC clinical sequencing in Uppsala

- Construct and automate pipelines on HPC and Cloud
- Transfer results into database system
  - Structured data (variants, tables, visualizations, data files etc.)
  - Inside hospital domain with clinical phenotypes (future)
- Security – sensitive personal information
- Encrypt and archive raw data
- Build up knowledge base for future predictive modeling

Now: In pilot production at UAH
Future: Extend to other cancers and genomic regions
eCPC involvements

- NIASC – The Nordic Information for Action e-Science Center of Excellence (http://nordicehealth.se/, 2014-1018)
  - Integrated Nordic research in population-based cancer screening
  - generic eScience infrastructure
  - eScience-based predictive algorithm within a national screening program

- PhenoMeNaI H2020 project (start Sept. 1st 2015)
  - clinical metabolomics
Sweden and Nordic countries
Enormous potential for eScience in medical research

- Reliable demographics and healthcare registers
- Clinical and population cohorts
- National biobanks
- Biotechnology and Information technology
- High quality epidemiology and clinical research
- Bioinformatics, computational biology, biostatistics
- Thank you -

www.ecpc.e-science.se