Nordic Big Biomedical Data for Action

Davit Bzhalava, PhD
Dept. of Laboratory Medicine,
Karolinska Institutet,
Sweden

Nordic Center of Excellence in Health-Related e-Sciences

(NIASC)

Main goals for NIASC 2014-2018

- Overall aim: develop eScience tools to improve health and societal preconditions for health
- Develop and enhance risk prediction algorithms for risk of breast, prostate, colorectal and cervical cancer
- Develop eScience tools to enable implementation of risk-stratified cancer screening polices

→ Establish the use of at least one eScience-based predictive algorithm within a national screening program by program end in 2018

→ Establish a significant legacy of eScience tools as outlined below and later during today's discussion

The need for a Nordic concerted effort

- Will achieve improved eScience and improved translation to health policies through increased researcher mobility between Nordic countries, training and synergy effects
- Joint development and sharing of eScience tools, adapted to unique Nordic infrastructures (registers, biobanks and organised screening programs)

NIASC IT/Bioinformaticians/DBA team

- 4 IT/bioinformatician/DBAs were recruited in Norway, Sweden, Denmark and Finland.

- In March 2015, the group had a meeting focusing on to get to know each other, to identify the most critical aspects for collaboration and to plan future activities.

Nordic Workshop on Best Practices for Big Data Handling in Biobanking

- Workshop brought together Nordic national biobanks, national biobank infrastructures, e-infrastructure service providers for sensitive data processing

- Three working groups were formed that will:
  - Formulate a joint use case on IT needs for the Nordic National Biobanks & Infrastructures.
  - Develop and optimize useful bioinformatics softwares/algorithms that are scalable and able to efficiently utilize parallel computing
  - Develop IT solutions that will enable mapping of multiple outcomes from registries to different biological entities (Reverse data mining or “registromics”).

Expected outcomes:

- Design a use case for on Nordic Big Biomedical Data for Action
- The NordForsk Nordic Colorectal Cancer consortium – as an expansion towards a clinical use case exploiting the infrastructure.
- To learn from each other (Where we are, what are the needs, how should we proceed)
- Prepare for big data calls.

Agenda – day 1

- **Session 1:**
- **Data privacy regulation, ethics and legal issues**
  - 13:00 – 13:40 The GDPR, balanced national legislation and exchange of data for research while still respecting national legislation.
  - 13:40 – 14:00 Discussions
- **Session 2: Nordic Colorectal Cancer consortium – a use case for Synergy of NORDIC Big Biomedical Data Sources**
  - 14:00-14:25 Joakim Dillner– Overview of the Nordic CRC Project
  - 14:25-15:50 Karin Sundström – Sweden Available resources and scientific interests of CRC Project
  - 15:50-15:15 Frederik Trier Møller – Acquisition of consumer data, storage, content and possibilities for CRC studies.
- **15:15 - 15:30 Coffee Break**
  - 15:30 - 15:50 Miriam Elfström – NKCx, storage, content and possibilities for CRC studies.
  - 15:50 - 16:15 Anu Jalanko – Finnish Biobanks – Available resources for CRC studies
  - 16:15 - 16:40 Tõnu Esko – Estonian Genome Center, Large scale population genomics and possibilities for CRC studies
  - 16:40 - 17:05 Laufey Tryggvadóttir - Infrastructure and Potential Participation in Nordic CRC Poject.
  - 17:05 - 17:30 E. Christina M. Wennerström - LABMED registry, storage, content and possibilities for CRC studies'
- **18:00 Dinner**

Agenda – day 2

Infrastructures and emerging technologies for processing big biomedical data

- 09:00 - 09:15 Davit Bzhalava Welcome, summary of day 1 and goals of the day 2

- **Session 1: Data centers that could handle sensitive biomedical data in NIASC consortium**
  - 09:15 – 09:45 Antti Pursula - Tryggve: Nordic services for sensitive data.
  - 09:45 – 10:15 Davit Bzhalava – Biomedical Computing Center at KI
  - 10:15 – 10:45 Jaakko Leinonen - ePouta infrastructure
  - 10:45 – 11:15 Joel Hedlund – Tryggve2 planning: scopes and roadmap.
  - 11:15 – 11:45 Discussions
- **12:00 - 13:00 Lunch**
- **Session 2: Emerging technologies for storing and processing big biomedical data**
  - 13:00 – 13:30 Jim Dowling – Hops Hadoop
  - 13:30 – 14:00 Maarala Ilari – Big Data processing for genomics.
  - 14:00 – 14:30 Petri Klemelä -Delivering bioinformatics software as virtual machine image
  - 14:30 – 15:00 Discussions
- **15:00 - 15:15 Coffee Break**

Agenda – day 2

<u>Session 3: Session algorithms for processing big biomedical data:</u>

15:15 – 15:35 Piotr Bala – PCJ BLAST
15:35 – 15:55 Raul Vicente – Deep learning for computational biology
15:55 – 16:15 Davit Bzhalava genome assembly on distributed file system
16:15 – 16:35 Bartlomiej Wilkowski – The Biobank Register: A search system for big biomedical and biobank data
16:35 – 17:00 Discussions

18:00 Dinner

Agenda – day 3

- **Day 3 (09:00 - 12:00)**
  - 09:00 - 09:30 Eero Pukkala – The Nordic Occupational Cancer Study (NOCCA), storage, content and possibilities for CRC studies.
- **Session 1: Software sustainability**
  - 09:30 - 10:00 Mike Croucher – Is your research software correct?
  - 10:00 - 10:30 Radovan Bast – CodeRefinery
  - 10:30 - 11:00 Discussions
- **Session 2: Plans for the future**
  - 11:00 – 11:30 Joakim Dillner – Future work and upcoming big data calls
  - 11:30 – 12:00 Discussions and planning for the future.
- **12:00 – 13:00 Lunch**