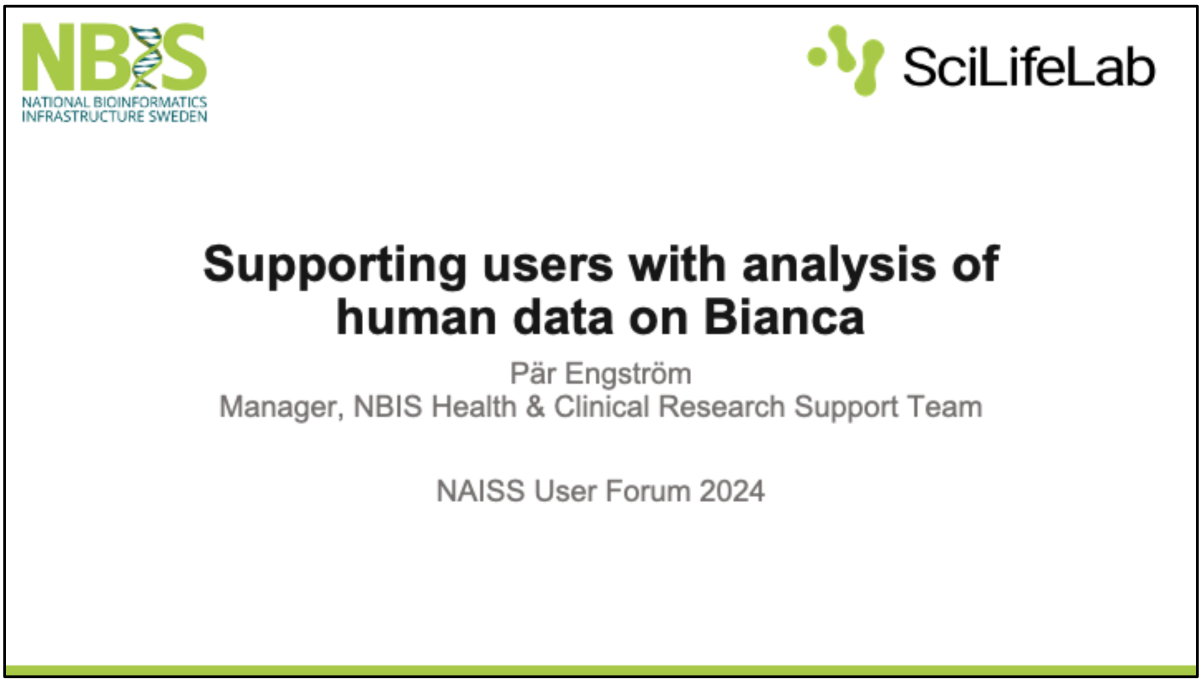
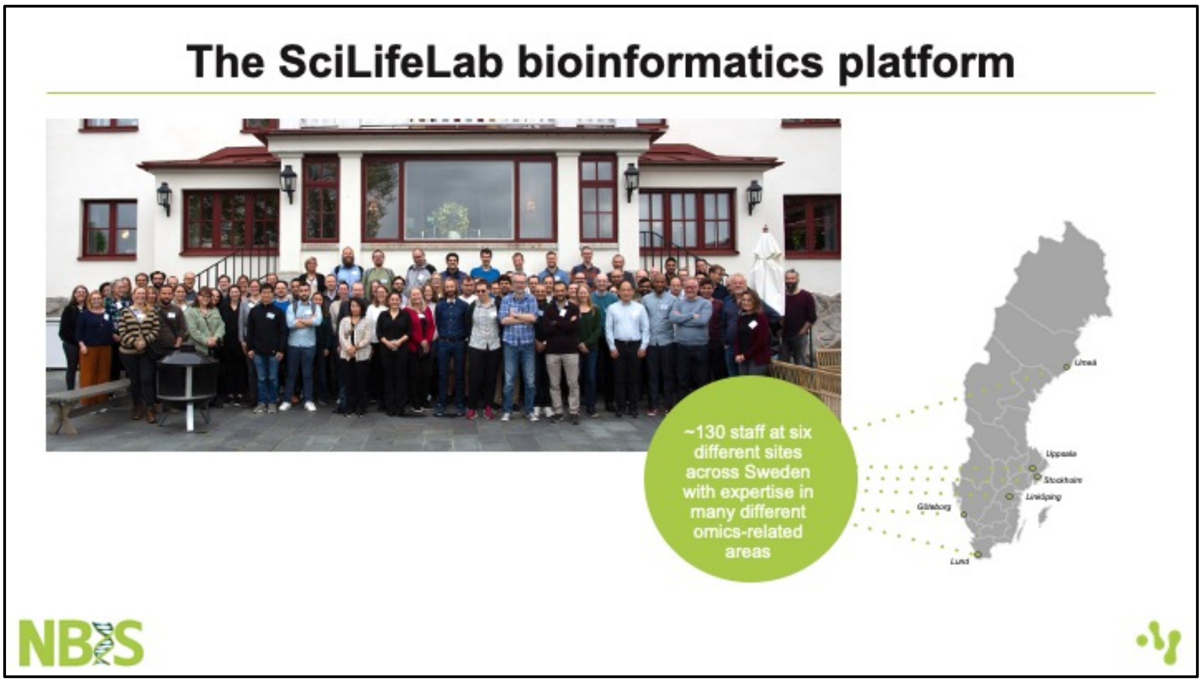


Supporting users with analysis of human data on Bianca

Pär Engström
Manager, NBIS Health & Clinical Research Support Team

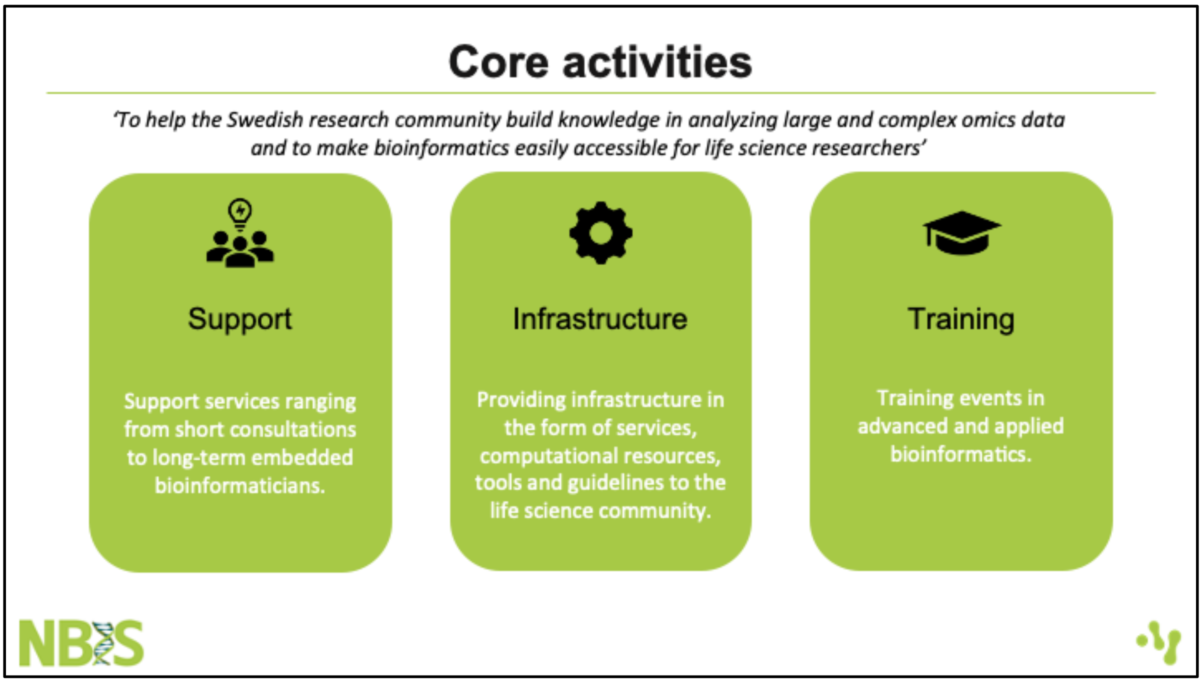
NAISS User Forum 2024



The SciLifeLab bioinformatics platform

~130 staff at six different sites across Sweden with expertise in many different omics-related areas

NBIS is the SciLifeLab bioinformatics platform, a distributed national infrastructure with about 130 staff across Sweden.



Core activities

'To help the Swedish research community build knowledge in analyzing large and complex omics data and to make bioinformatics easily accessible for life science researchers'

Support

Support services ranging from short consultations to long-term embedded bioinformaticians.

Infrastructure

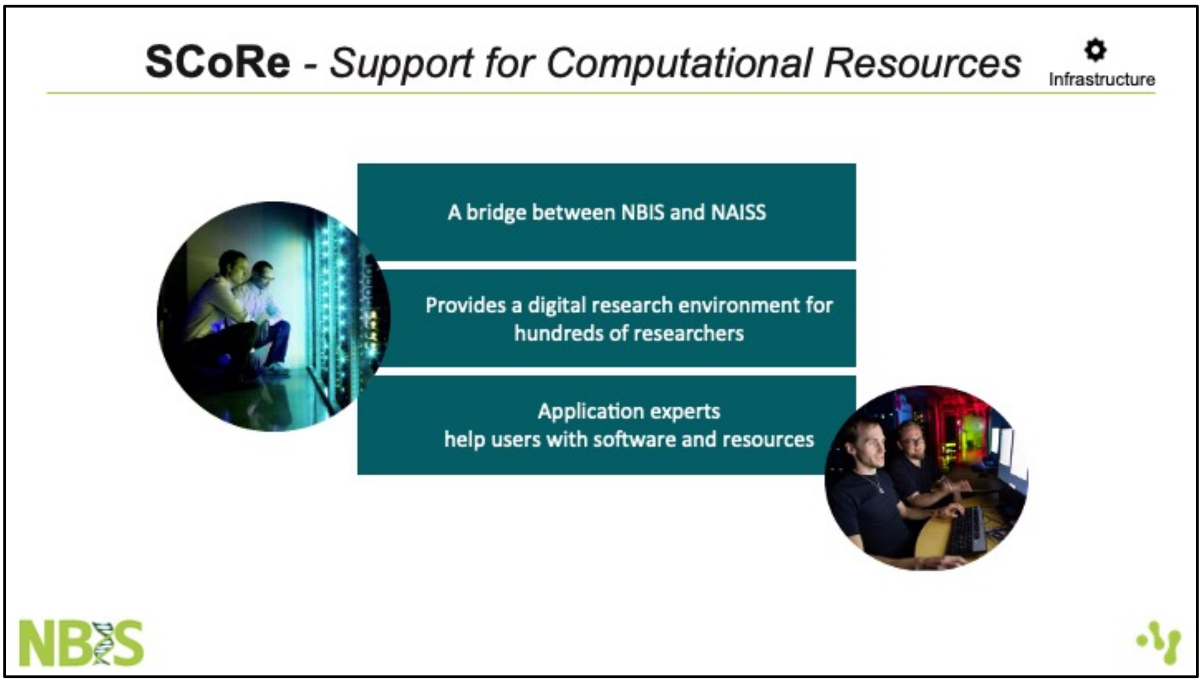
Providing infrastructure in the form of services, computational resources, tools and guidelines to the life science community.

Training

Training events in advanced and applied bioinformatics.

NBIS

NBIS has a number of activities in the areas of Support, Infrastructure and Training – to ensure that bioinformatics is easily accessible for life science researchers in Sweden.



SCoRe - *Support for Computational Resources*

Infrastructure

A bridge between NBIS and NAISS

Provides a digital research environment for hundreds of researchers

Application experts
help users with software and resources

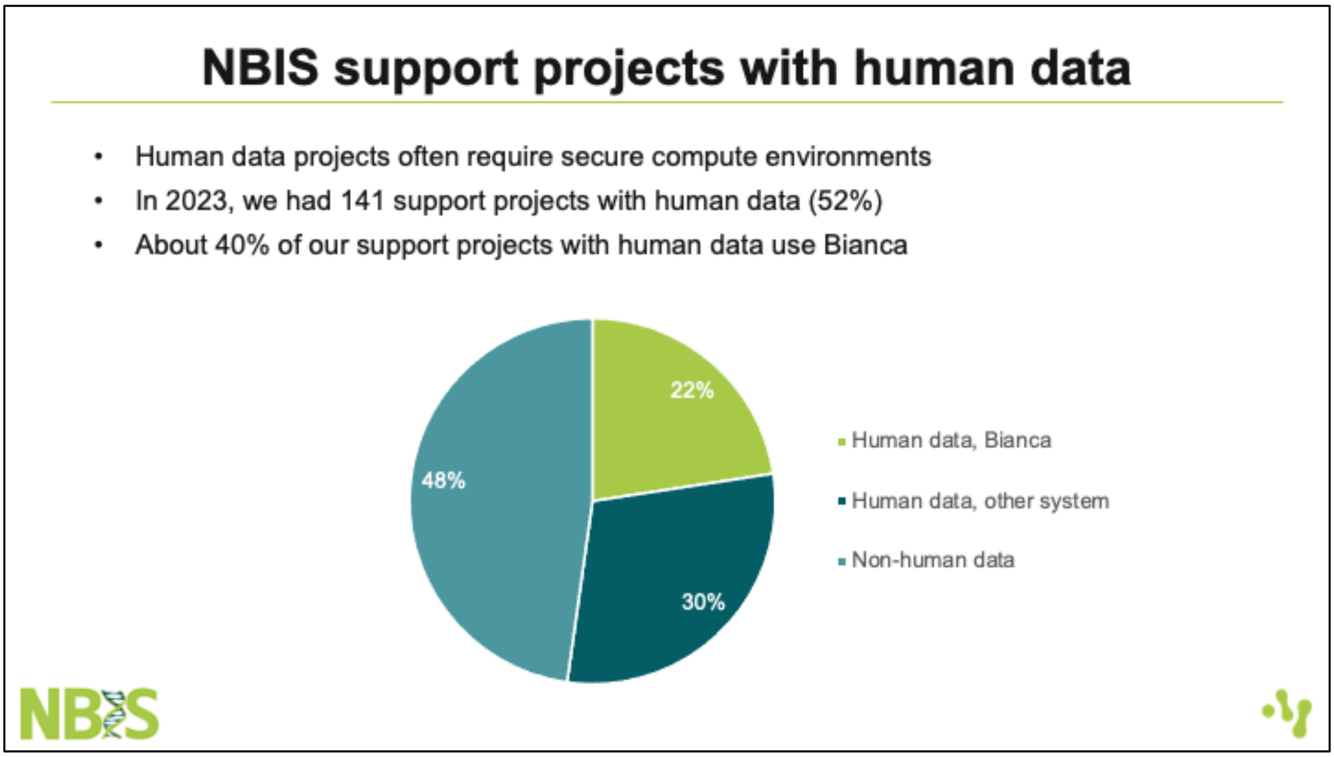
NBIS has a team called SCoRe – Support for Computational Resources.
They serve as a bridge between NBIS and the NAISS compute resources, for example maintaining the rich set of bioinformatics software installations on Rackham and Bianca.

Support with data processing & analysis

Support

nbis.se/get-support

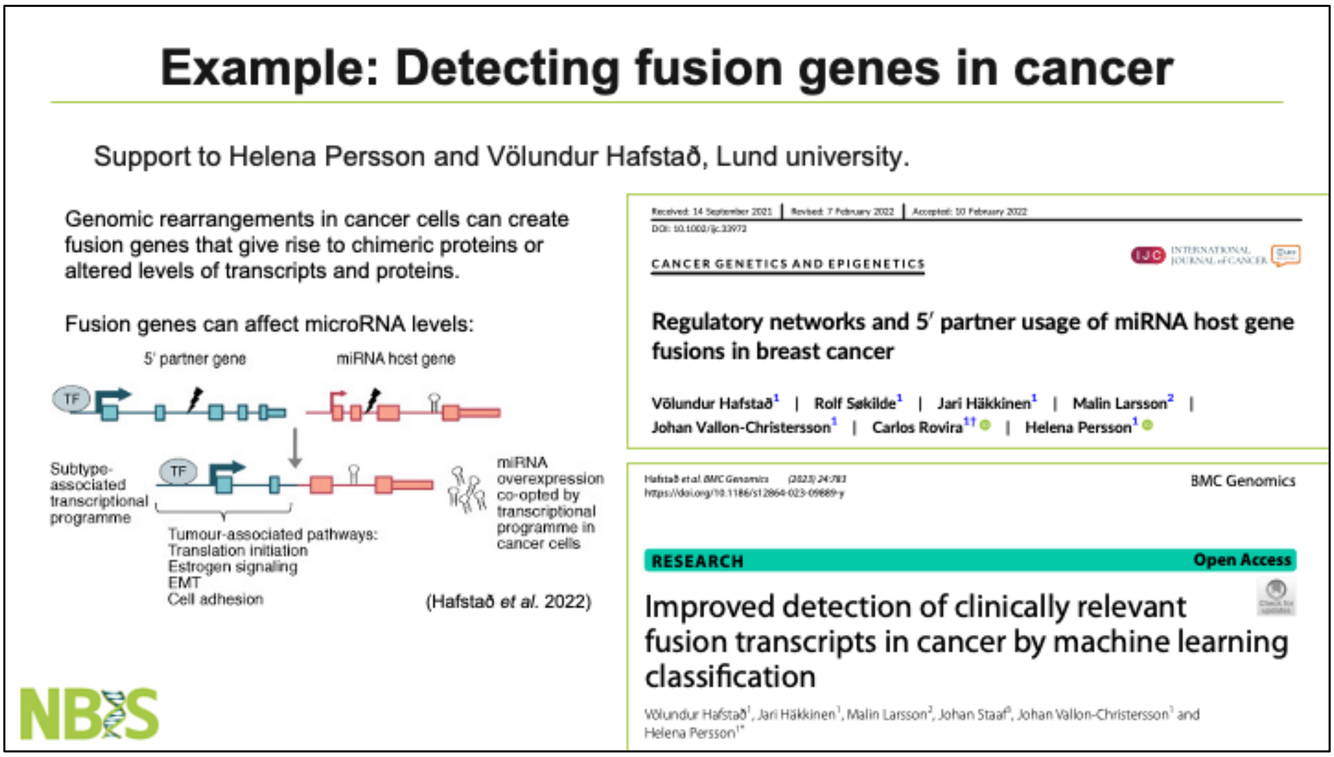
Today I will focus on our support for data processing and analysis, which is the largest activity in NBIS.
Here we typically work hands-on in research projects to help research groups with their bioinformatics needs.



NBIS support projects with human data

- Human data projects often require secure compute environments
- In 2023, we had 141 support projects with human data (52%)
- About 40% of our support projects with human data use Bianca

About half of all projects we support have human data. Last year, we had about 140 projects with human data.
When we work with human data, we need to make sure we use a secure enough compute environment.
Bianca is very secure. It’s not always necessary to work there, but often it is the best choice, in particular if there is genetic data in the project, as genetics can be used to identify individuals.



Example: Detecting fusion genes in cancer

Support to Helena Persson and Völundur Hafstað, Lund university.

Genomic rearrangements in cancer cells can create fusion genes that give rise to chimeric proteins or altered levels of transcripts and proteins.

Fusion genes can affect microRNA levels:

5′ partner gene
miRNA host gene
TF
Subtype-associated transcriptional programme
TF
miRNA overexpression co-opted by transcriptional programme in cancer cells
Tumour-associated pathways:
Translation initiation
Estrogen signaling
EMT
Cell adhesion

(Hafstað *et al.* 2022)

Received: 14 September 2021 | Revised: 7 February 2022 | Accepted: 10 February 2022
DOI: 10.1002/ijc.33973

CANCER GENETICS AND EPIGENETICS

IJC INTERNATIONAL JOURNAL of CANCER

Regulatory networks and 5′ partner usage of miRNA host gene fusions in breast cancer

Völundur Hafstað[1] | Rolf Søkilde[1] | Jari Häkkinen[1] | Malin Larsson[2] | Johan Vallon-Christersson[1] | Carlos Rovira[1†] | Helena Persson[1]

Hafstað et al. *BMC Genomics* (2023) 24:783
https://doi.org/10.1186/s12864-023-09889-y

BMC Genomics

RESEARCH **Open Access**

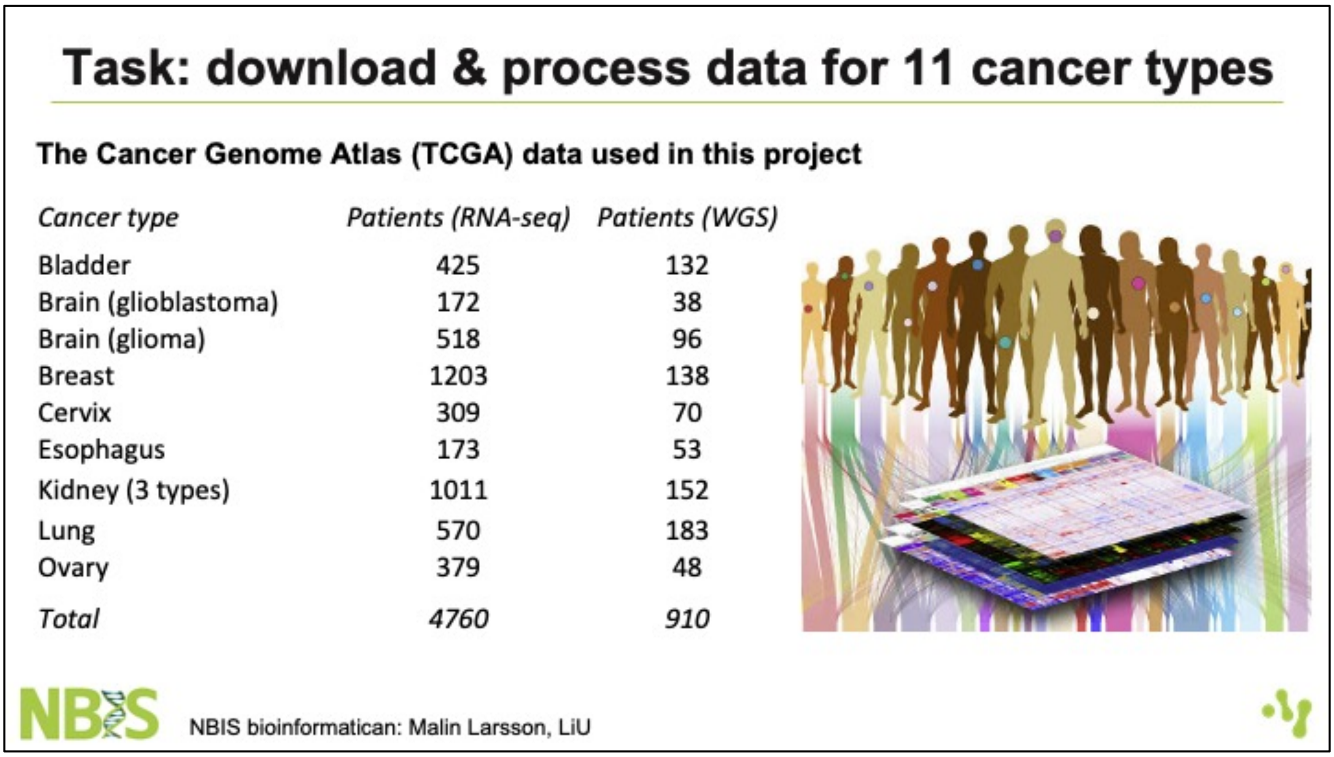
Improved detection of clinically relevant fusion transcripts in cancer by machine learning classification

Völundur Hafstað[1], Jari Häkkinen[1], Malin Larsson[2], Johan Staaf[3], Johan Vallon-Christersson[1] and Helena Persson[1*]

Here is an example of one of our recent support projects where we used Bianca.
In this project, we helped Helena Persson and her PhD student Völundur Hafstað at Lund university.
They studied a type of cancer-causing mutation called gene fusions.
Such mutations cause two different genes to be combined resulting in a chimeric gene with unwanted effects.

A programming analogy:
Gene fusions are as if you would cut and paste in a computer program so that part of a subroutine gets replaced with part of a different subroutine.
That would result in a buggy program with undesirable effects - perhaps spinning out of control and threatening to take over the system like a tumour in the body.

If you are interested in the science, here are a couple of papers that resulted from the project.
One is about an improved method for detecting gene fusions, and the other presents novel biological discoveries about how fusion genes can cause cancer.



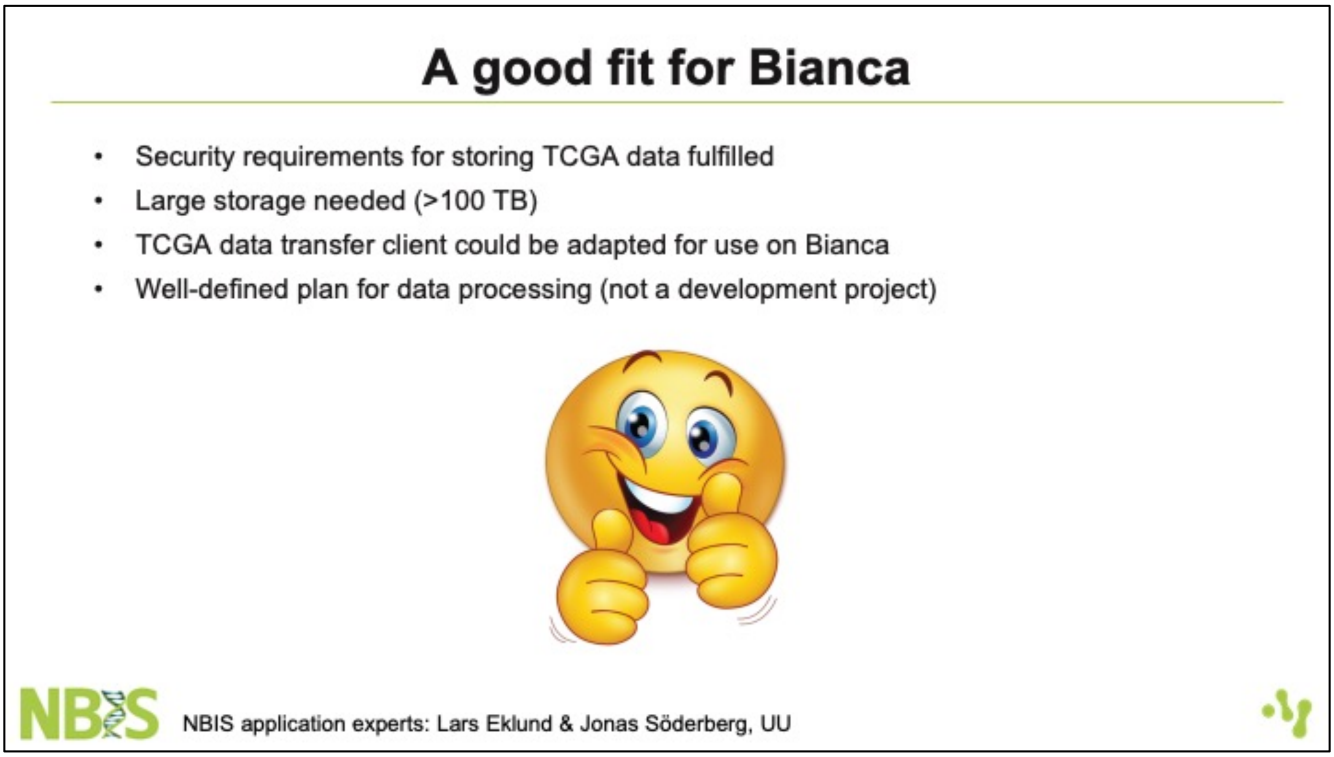
Task: download & process data for 11 cancer types

The Cancer Genome Atlas (TCGA) data used in this project

Cancer type	*Patients (RNA-seq)*	*Patients (WGS)*
Bladder	425	132
Brain (glioblastoma)	172	38
Brain (glioma)	518	96
Breast	1203	138
Cervix	309	70
Esophagus	173	53
Kidney (3 types)	1011	152
Lung	570	183
Ovary	379	48
Total	*4760*	*910*

NBIS bioinformatican: Malin Larsson, LiU

To make this research possible, the group wanted our help with identifying fusion genes in data from cancer patients.
And they wanted to to this on a large scale, so they asked us for help in downloading and processing part of a large American database, The Cancer Genome Atlas.
The database contains genetic data collected from cancer samples, so called RNA sequencing and whole genome sequencing data.
We used data for 11 different cancer types, and several thousand patients in total.
We ran different programs and scripts on the data to identify and study fusion genes.

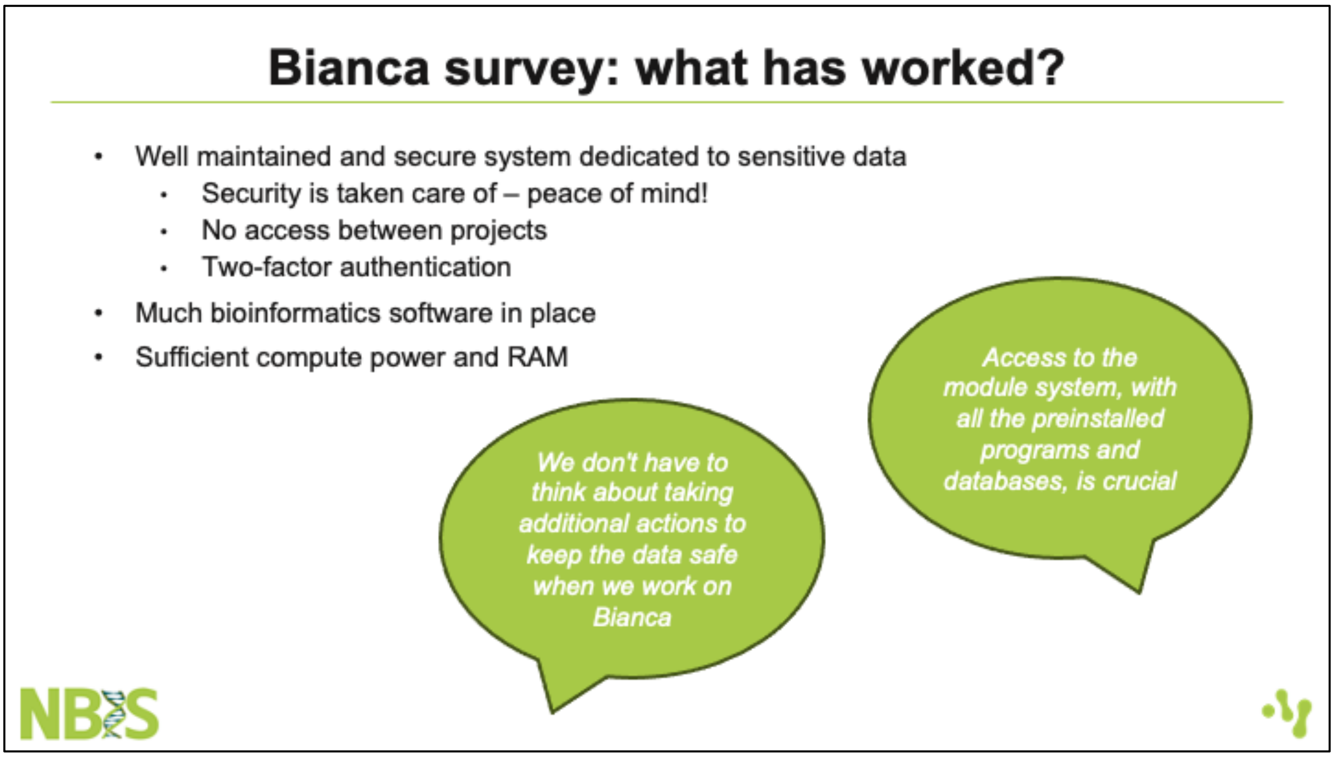


A good fit for Bianca

- Security requirements for storing TCGA data fulfilled
- Large storage needed (>100 TB)
- TCGA data transfer client could be adapted for use on Bianca
- Well-defined plan for data processing (not a development project)

NBIS application experts: Lars Eklund & Jonas Söderberg, UU

For this project, Bianca was a very good fit.
Having a resource like Bianca is essential for projects like this.



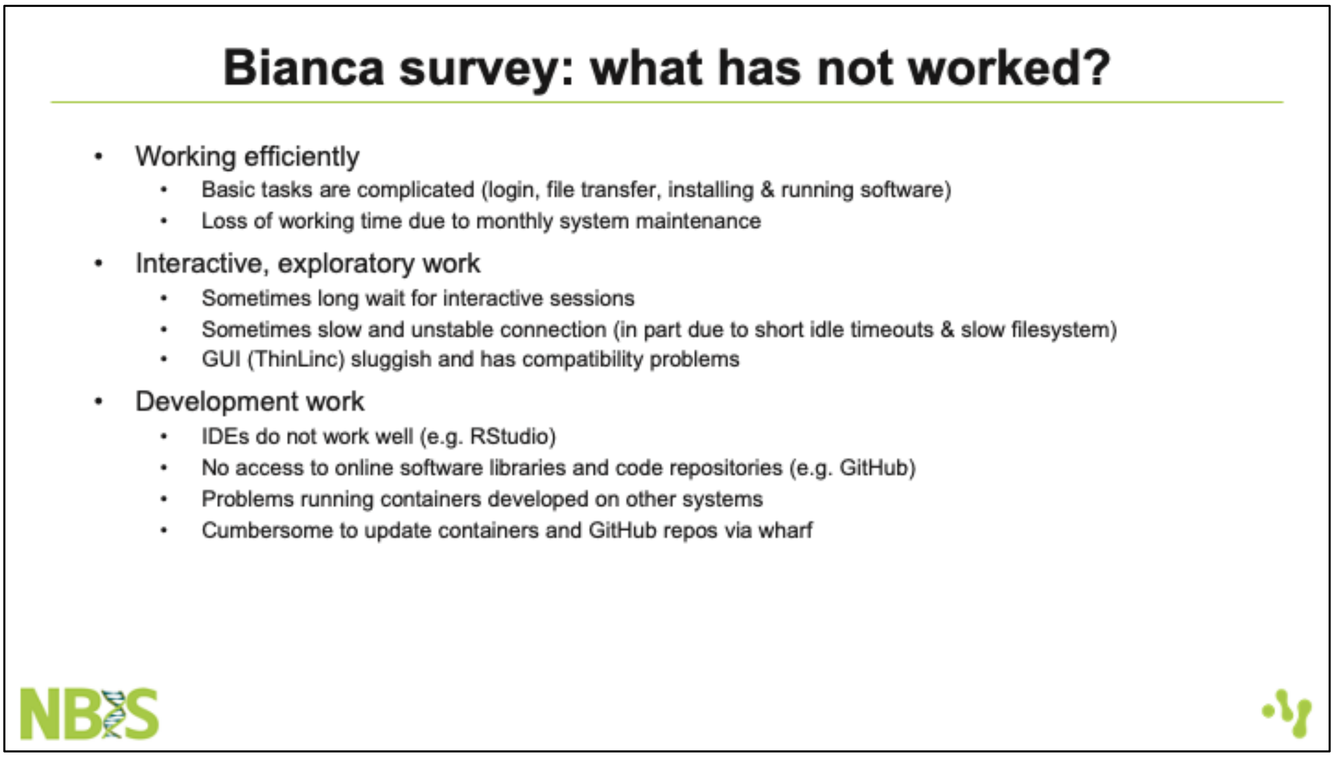
Bianca survey: what has worked?

- Well maintained and secure system dedicated to sensitive data
  - Security is taken care of – peace of mind!
  - No access between projects
  - Two-factor authentication
- Much bioinformatics software in place
- Sufficient compute power and RAM

Access to the module system, with all the preinstalled programs and databases, is crucial

We don't have to think about taking additional actions to keep the data safe when we work on Bianca

We asked NBIS bioinformaticians about their experiences of working on Bianca. Here is what they said about what has worked well.



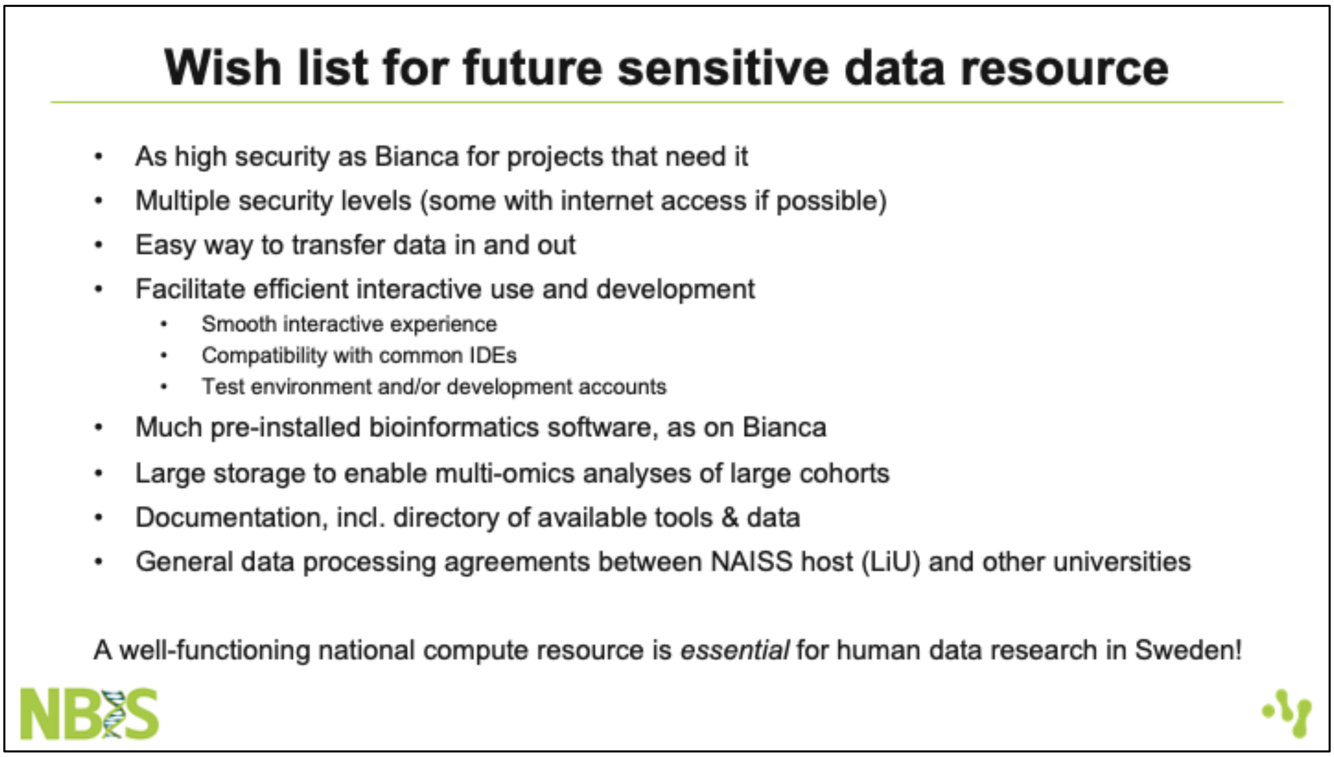
Bianca survey: what has not worked?

- Working efficiently
  - Basic tasks are complicated (login, file transfer, installing & running software)
  - Loss of working time due to monthly system maintenance
- Interactive, exploratory work
  - Sometimes long wait for interactive sessions
  - Sometimes slow and unstable connection (in part due to short idle timeouts & slow filesystem)
  - GUI (ThinLinc) sluggish and has compatibility problems
- Development work
  - IDEs do not work well (e.g. RStudio)
  - No access to online software libraries and code repositories (e.g. GitHub)
  - Problems running containers developed on other systems
  - Cumbersome to update containers and GitHub repos via wharf

And here is what they said about what has not worked.

The obstacles for working efficiently are a concern.
The added overhead with a system like Bianca is a problem for life science researchers who work in competitive fields. Also, it affects our ability to deliver support efficiently.

In addition, it is challenging to carry out interactive, exploratory and development work on Bianca.
For non-sensitive data, one can often do exploratory analysis and development on a local machine, but for sensitive data, one needs a secure system also for such tasks that don’t require much compute.



Wish list for future sensitive data resource

- As high security as Bianca for projects that need it
- Multiple security levels (some with internet access if possible)
- Easy way to transfer data in and out
- Facilitate efficient interactive use and development
  - Smooth interactive experience
  - Compatibility with common IDEs
  - Test environment and/or development accounts
- Much pre-installed bioinformatics software, as on Bianca
- Large storage to enable multi-omics analyses of large cohorts
- Documentation, incl. directory of available tools & data
- General data processing agreements between NAISS host (LiU) and other universities

A well-functioning national compute resource is *essential* for human data research in Sweden!

Based on the survey, here is a wish list for future sensitive data compute resources.