AARHUS
UNIVERSITY
DEPARTMENT OF
MOLECULAR BIOLOGY AND GENETICS

FAUPE Kick Off Meeting

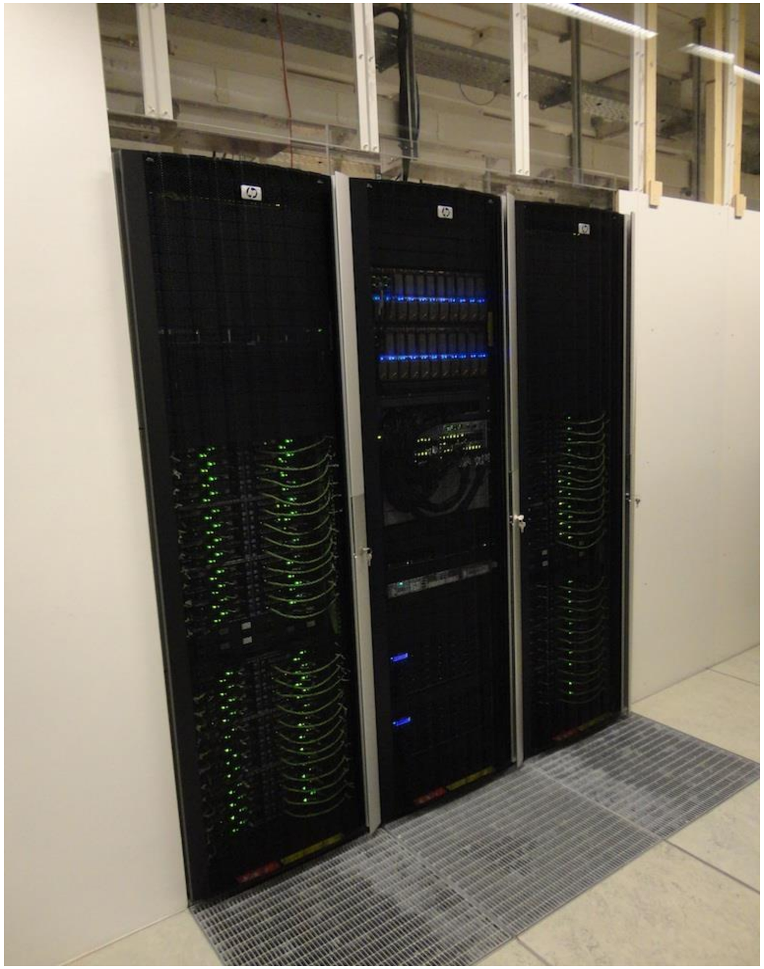
WP 2 Aims

- Establishment of a high performance computing cluster
- Establishment of a genomics breeding decision support system

WP 2 Aims

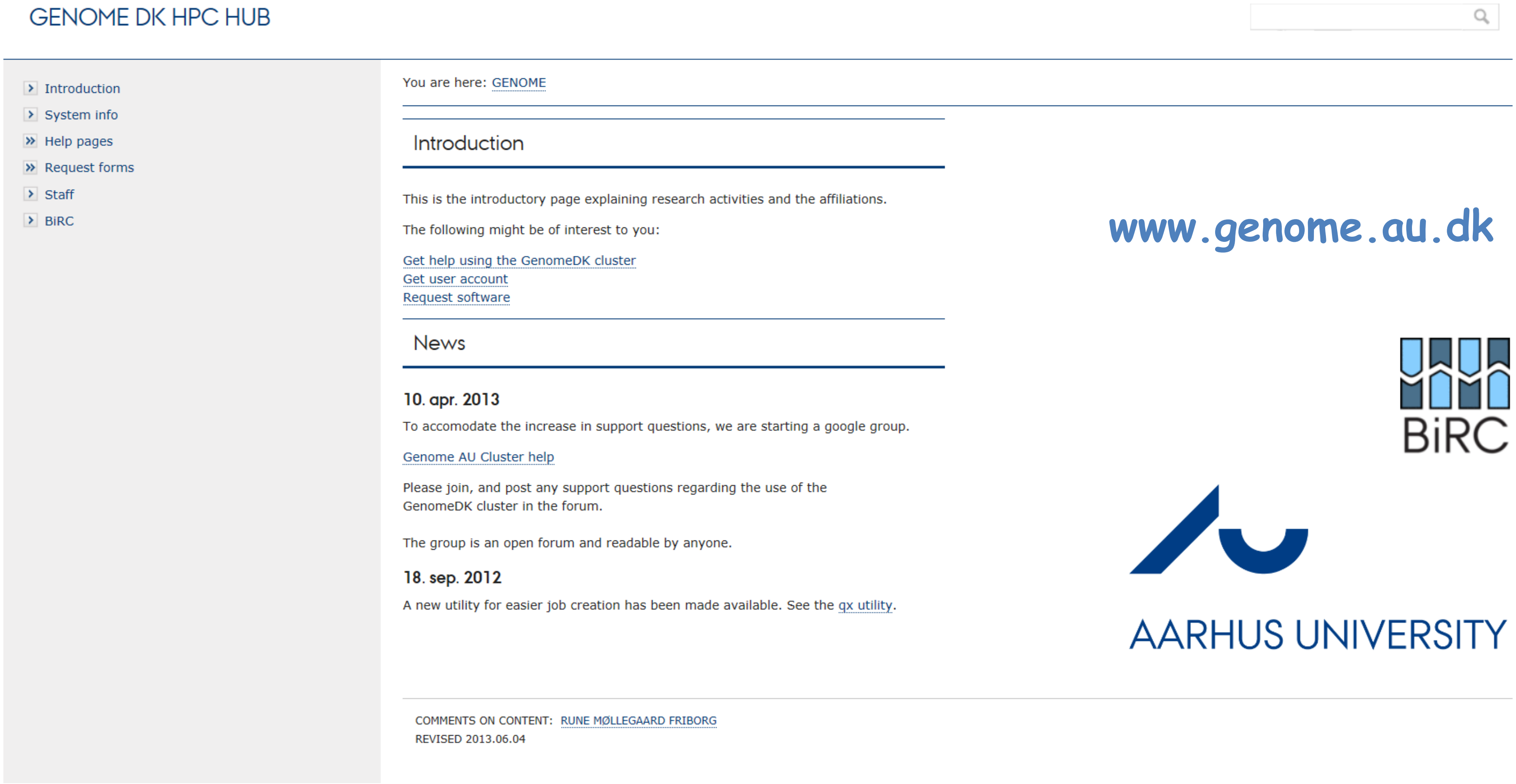
- Establishment of a high performance computing cluster

The GenomeDK Cluster



AARHUS
UNIVERSITY
DEPARTMENT OF
MOLECULAR BIOLOGY AND GENETICS

The GenomeDK Cluster Home Page



GENOME DK HPC HUB

- Introduction
- System info
- Help pages
- Request forms
- Staff
- BiRC

You are here: GENOME

Introduction

This is the introductory page explaining research activities and the affiliations.

The following might be of interest to you:

Get help using the GenomeDK cluster
Get user account
Request software

News

10. apr. 2013

To accomodate the increase in support questions, we are starting a google group.

Genome AU Cluster help

Please join, and post any support questions regarding the use of the GenomeDK cluster in the forum.

The group is an open forum and readable by anyone.

18. sep. 2012

A new utility for easier job creation has been made available. See the qx utility.

COMMENTS ON CONTENT: RUNE MØLLEGAARD FRIBORG
REVISED 2013.06.04

www.genome.au.dk

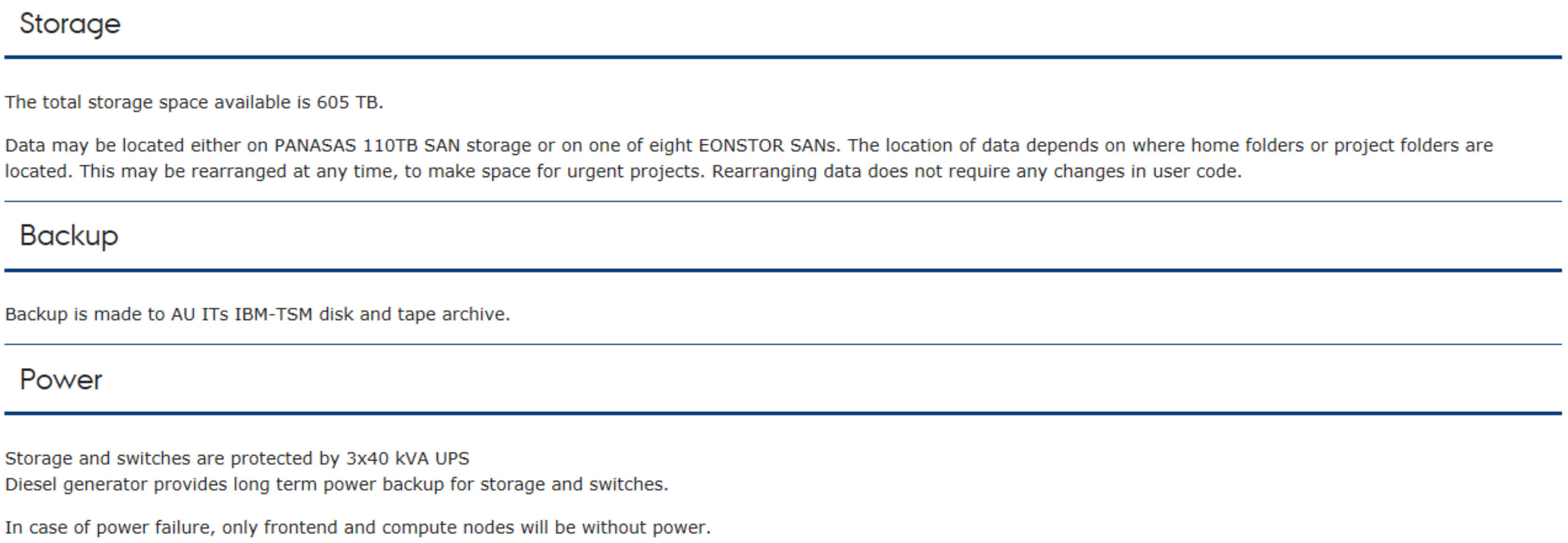
BiRC

AARHUS UNIVERSITY

GenomeDK System Info - Nodes

Queue name	Nodes / cores	Node description and info	Remarks
normal	95 / 1520	› Two Intel/"Sandy Bridge" E5-2670 CPUs @ 2.67 GHz, 8 cores/CPU › 64 GB memory @ 1600 MHz › 2 TB SATA disks. Raid 0: ~280MB/s › 10 GigE and 1 GigE NIC's.	Default walltime 740 hours
normal	56 / 896	› Two Intel/"Sandy Bridge" E5-2670 CPUs @ 2.67 GHz, 8 cores/CPU › 128 GB memory @ 1600 MHz › 2 TB SATA disks. Raid 0: ~280MB/s › InfiniBand 4X QDR and 1 GigE NIC's	Default walltime 740 hours
qfat1	1 / 32	› Four AMD/Opteron 6212 CPUs @ 2.67 GHz, 8 cores/CPU › 512 GB memory @ 800 MHz › 2 TB SAS disk: ~200MB/s › 10 GigE and 1 GigE NIC's.	Default walltime 740 hours, CPU/memory performance 25% of nodes in the normal queue.
qfat2	3 / 72	› Four Intel/"Westmere" E7-4807 CPUs @ 1.87 Ghz, 6 cores/CPU › 1024 GB memory @ 800 MHz › 2 TB SAS disk: ~200MB/s › 10 GigE and 1GigE NIC's.	Default walltime 740 hours, CPU/memory performance 50% of nodes in the normal queue.

GenomeDK System Info – Storage and Backup



Storage

The total storage space available is 605 TB.

Data may be located either on PANASAS 110TB SAN storage or on one of eight EONSTOR SANs. The location of data depends on where home folders or project folders are located. This may be rearranged at any time, to make space for urgent projects. Rearranging data does not require any changes in user code.

Backup

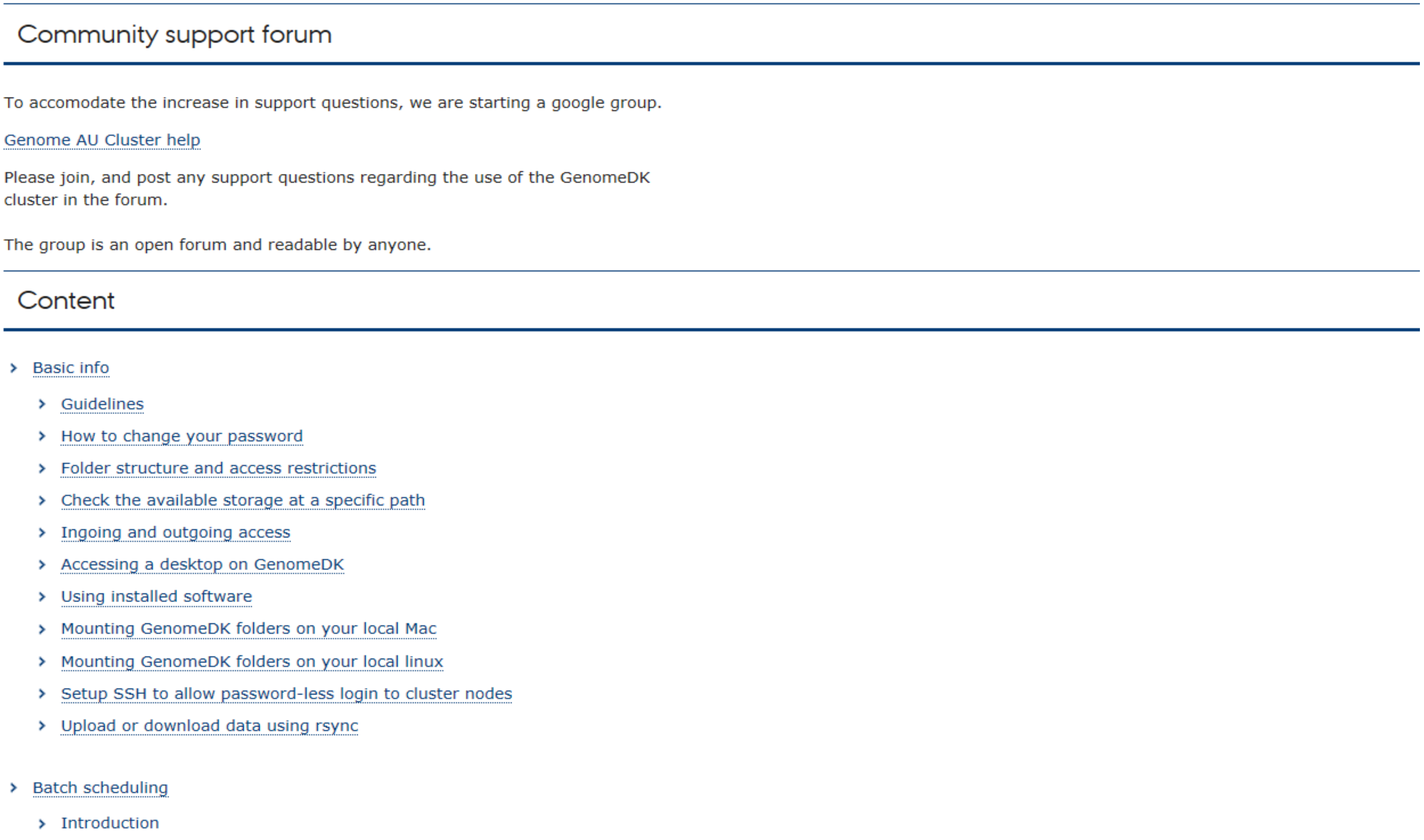
Backup is made to AU ITs IBM-TSM disk and tape archive.

Power

Storage and switches are protected by 3x40 kVA UPS
Diesel generator provides long term power backup for storage and switches.

In case of power failure, only frontend and compute nodes will be without power.

GenomeDK – Help Pages



Community support forum

To accomodate the increase in support questions, we are starting a google group.

Genome AU Cluster help

Please join, and post any support questions regarding the use of the GenomeDK cluster in the forum.

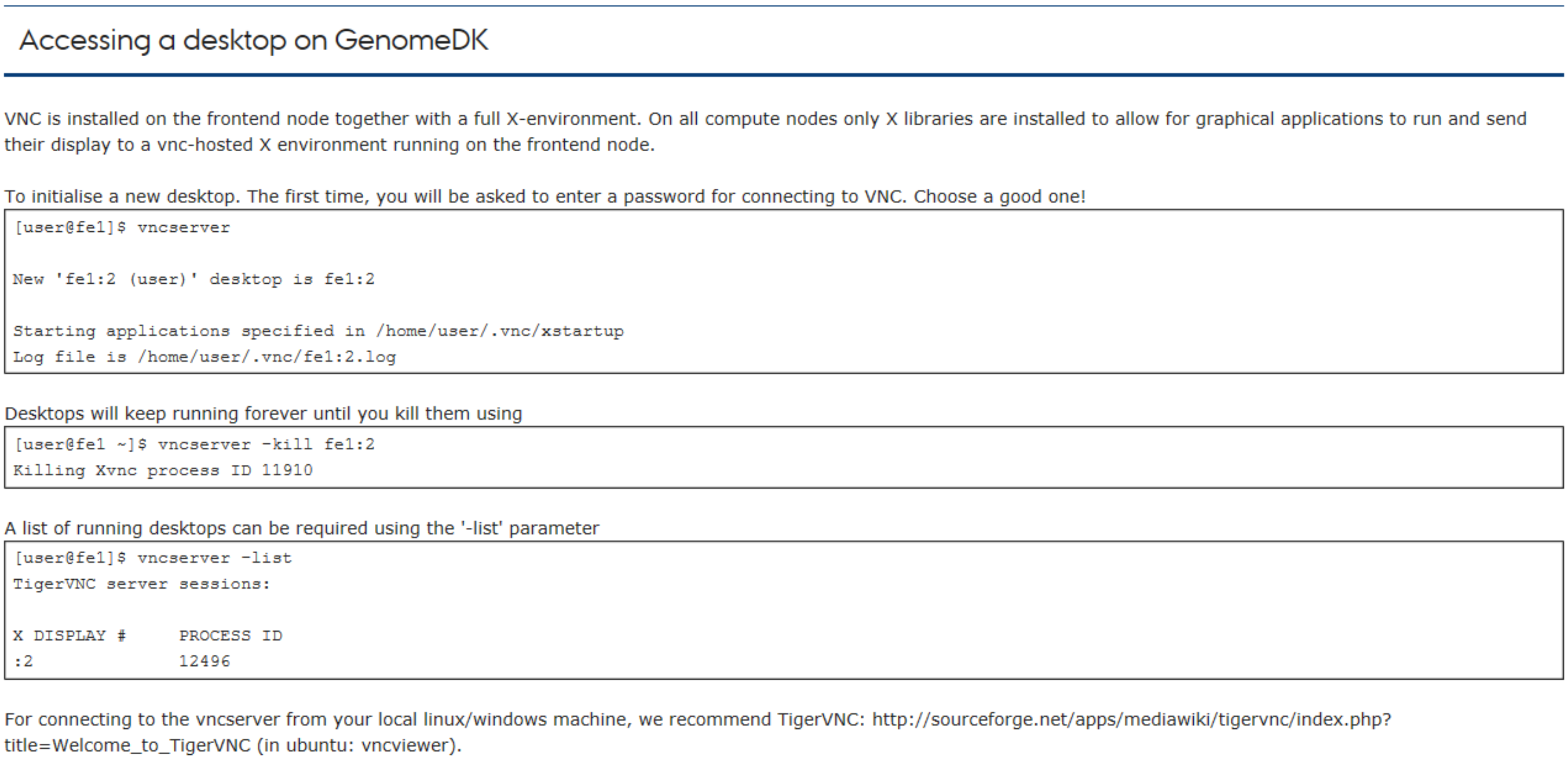
The group is an open forum and readable by anyone.

Content

- Basic info
    - Guidelines
    - How to change your password
    - Folder structure and access restrictions
    - Check the available storage at a specific path
    - Ingoing and outgoing access
    - Accessing a desktop on GenomeDK
    - Using installed software
    - Mounting GenomeDK folders on your local Mac
    - Mounting GenomeDK folders on your local linux
    - Setup SSH to allow password-less login to cluster nodes
    - Upload or download data using rsync
- Batch scheduling
    - Introduction

AARHUS
UNIVERSITY
DEPARTMENT OF
MOLECULAR BIOLOGY AND GENETICS

GenomeDK – Help Page Example



Accessing a desktop on GenomeDK

VNC is installed on the frontend node together with a full X-environment. On all compute nodes only X libraries are installed to allow for graphical applications to run and send their display to a vnc-hosted X environment running on the frontend node.

To initialise a new desktop. The first time, you will be asked to enter a password for connecting to VNC. Choose a good one!

```
[user@fe1]$ vncserver

New 'fe1:2 (user)' desktop is fe1:2

Starting applications specified in /home/user/.vnc/xstartup
Log file is /home/user/.vnc/fe1:2.log
```

Desktops will keep running forever until you kill them using

```
[user@fe1 ~]$ vncserver -kill fe1:2
Killing Xvnc process ID 11910
```

A list of running desktops can be required using the '-list' parameter

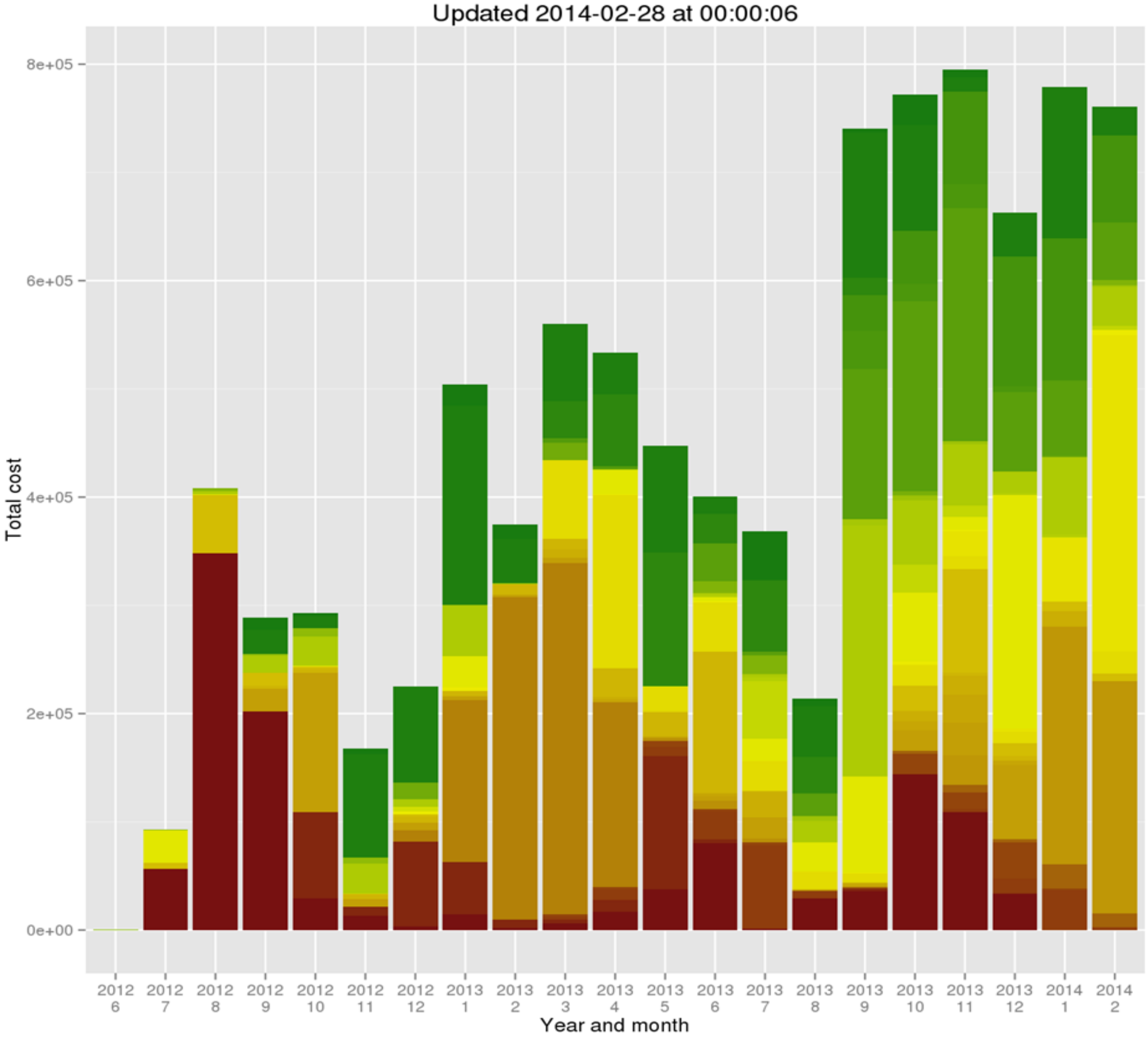
```
[user@fe1]$ vncserver -list
TigerVNC server sessions:

X DISPLAY #     PROCESS ID
:2              12496
```

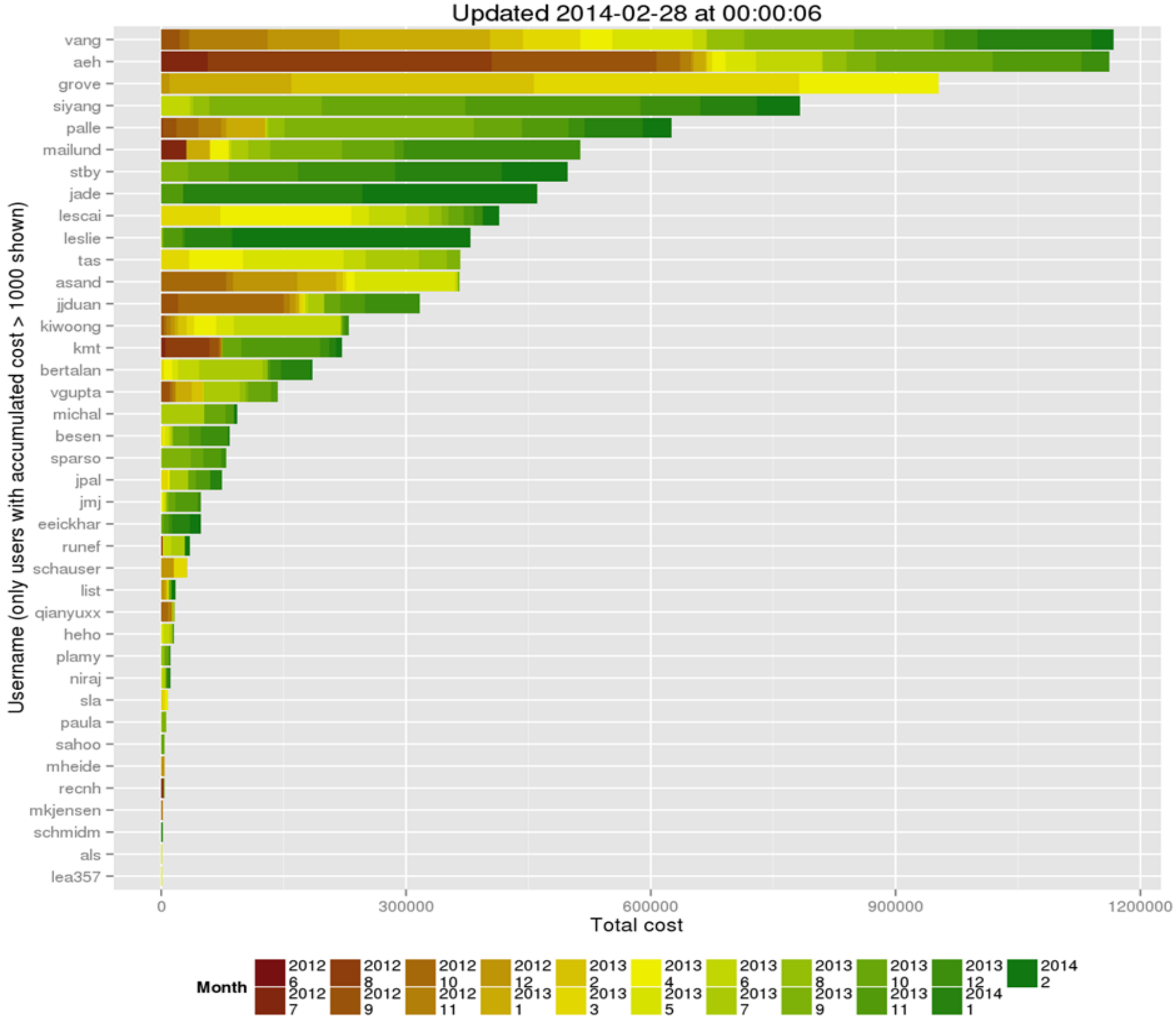
For connecting to the vncserver from your local linux/windows machine, we recommend TigerVNC: http://sourceforge.net/apps/mediawiki/tigervnc/index.php?title=Welcome_to_TigerVNC (in ubuntu: vncviewer).

AARHUS
UNIVERSITY
DEPARTMENT OF
MOLECULAR BIOLOGY AND GENETICS

Load on the GenomeDK Cluster



Accumulated Costs



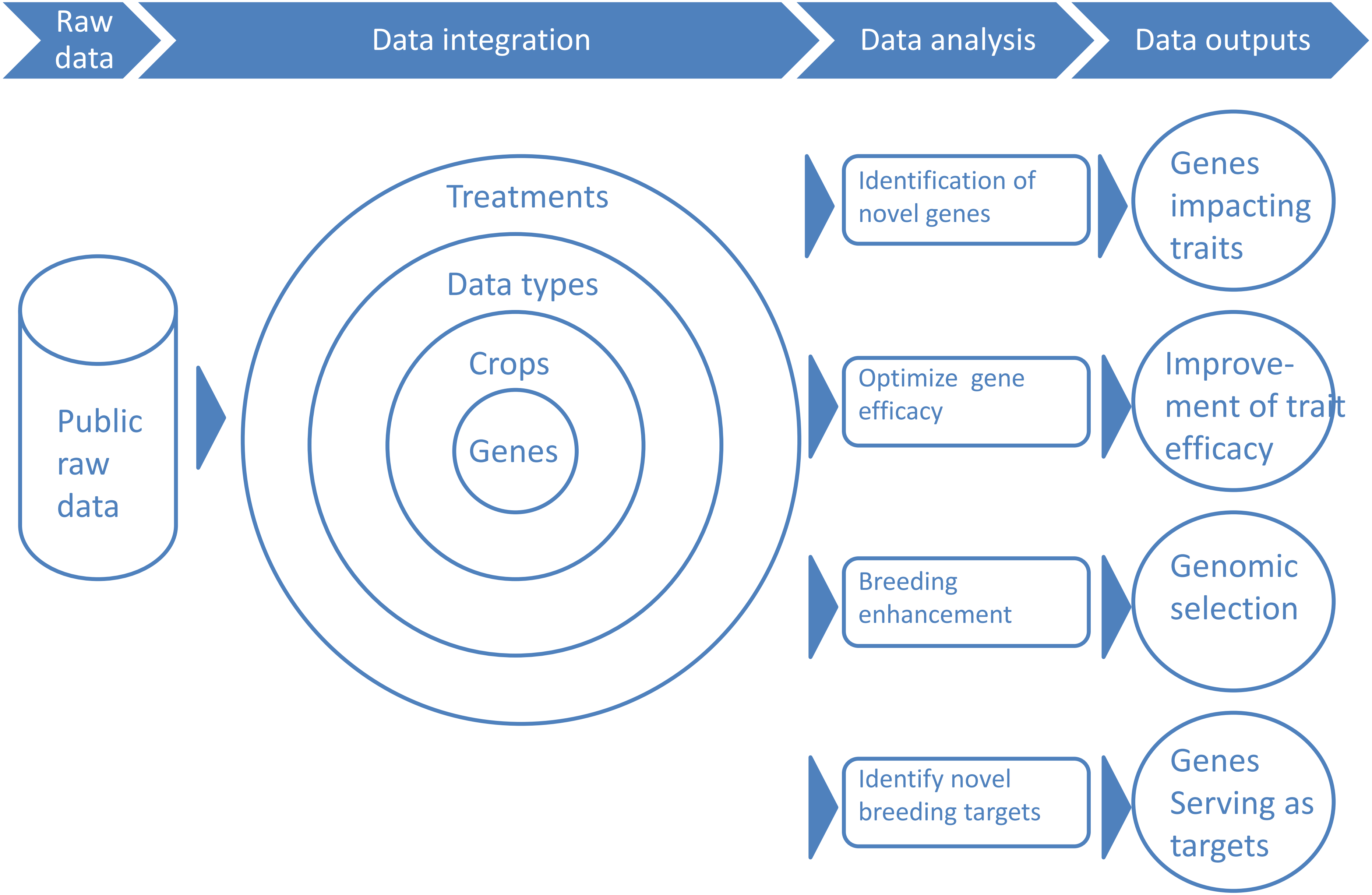
CID Cluster Specifications (preliminary)

- Integrated with existing GenomeDK cluster at Aarhus University
- Two types of nodes
  - 1 node with 24 cores and 1 TB RAM (RAM demanding data analysis)
  - 40 nodes with 16 cores and 64GB RAM per node (CPU-demanding data analysis).
  - 100TB data storage with backup
  - Continous upgrades of any hardware can be performed, if needed

WP 2 Aims

- Establishment of a CID genomics breeding decision support system
- A resource that will complement existing breeding activities (traditional and genomic selection breeding programs)
- Build on:
  - Sequence data available through public databases
    - Genomes, transcriptomes, SNPs……
  - Molecular pathway information
  - Published trait information
- Proprietary data is not a requirement - can optionally be included
- A strategy for integration of phenotyping and genotyping data

Data Integration – a Gene Centric Appoach



Raw data

Data integration

Data analysis

Data outputs

Public raw data

Treatments

Data types

Crops

Genes

Identification of novel genes

Genes impacting traits

Optimize gene efficacy

Improve-ment of trait efficacy

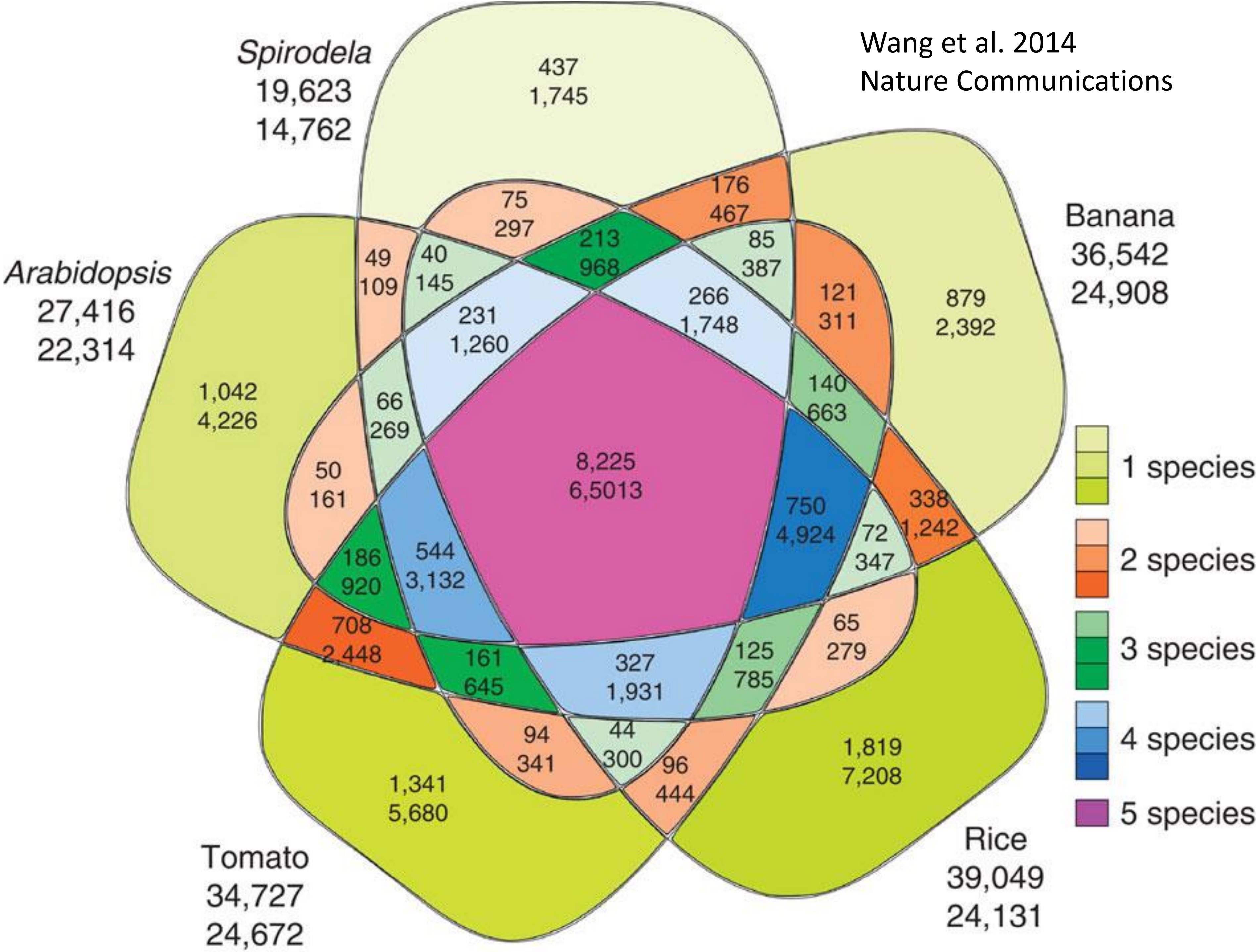
Breeding enhancement

Genomic selection

Identify novel breeding targets

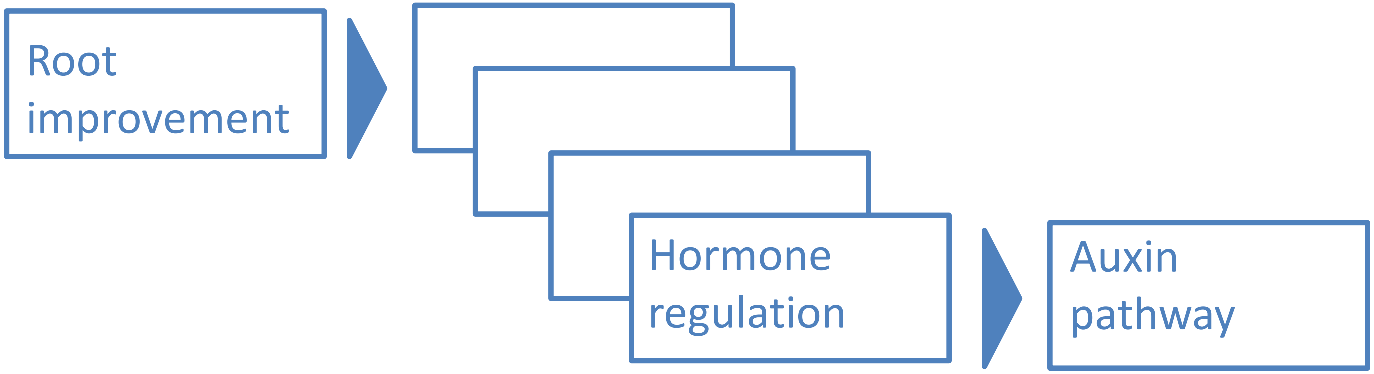
Genes Serving as targets

Wang et al. 2014
Nature Communications



Trait understanding

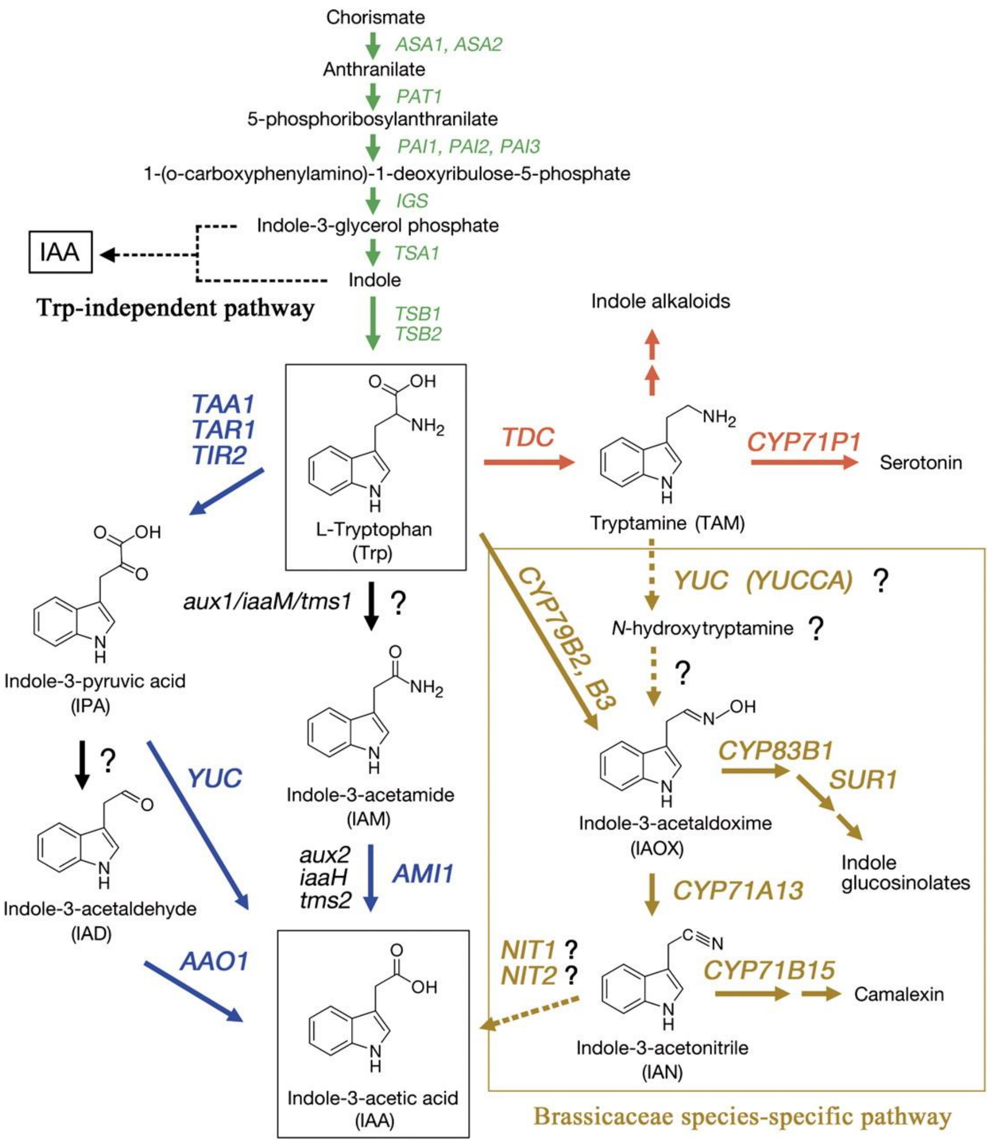
- **Trait will be broken down into precise biological and physiological terms in order to focus on relevant processes and facilitate gene discovery for trait improvement**



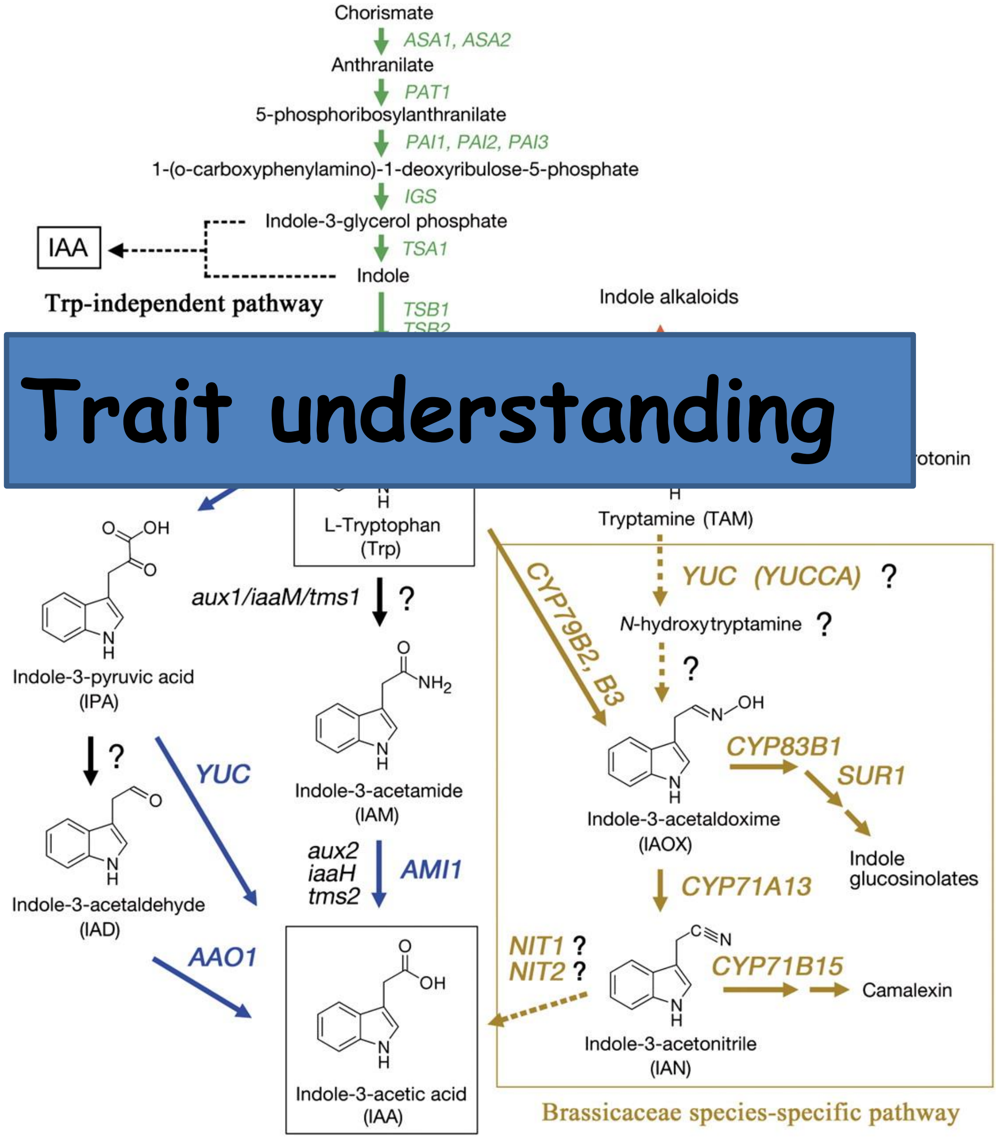
Root improvement

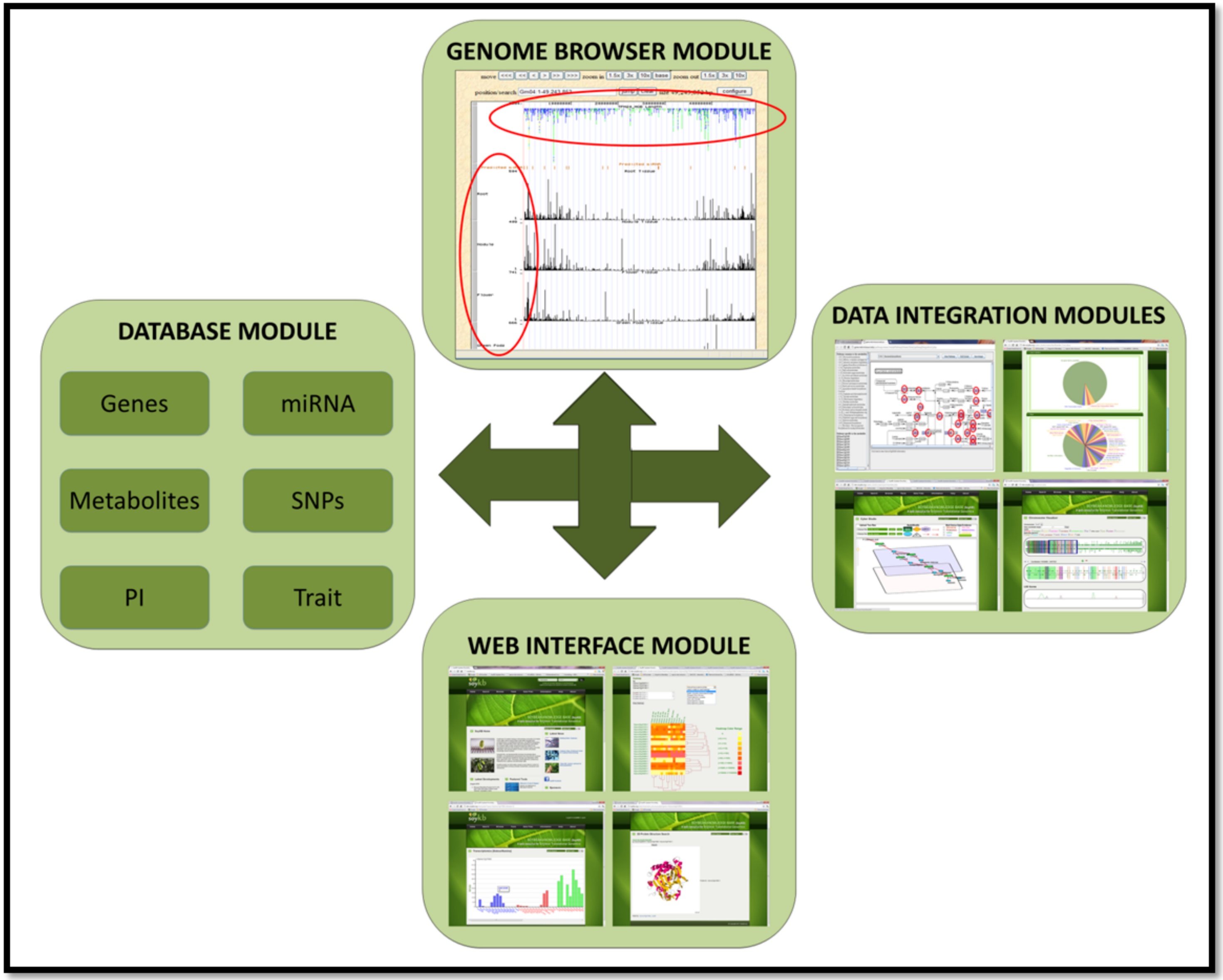
Hormone regulation

Auxin pathway


Chorismate
ASA1, ASA2
Anthranilate
PAT1
5-phosphoribosylanthranilate
PAI1, PAI2, PAI3
1-(o-carboxyphenylamino)-1-deoxyribulose-5-phosphate
IGS
Indole-3-glycerol phosphate
TSA1
Indole
IAA
Trp-independent pathway
TSB1
TSB2
Indole alkaloids
TAA1
TAR1
TIR2
L-Tryptophan
(Trp)
TDC
Tryptamine (TAM)
CYP71P1
Serotonin
Indole-3-pyruvic acid
(IPA)
aux1/iaaM/tms1 ?
CYP79B2, B3
YUC (YUCCA) ?
N-hydroxytryptamine ?
?
?
Indole-3-acetamide
(IAM)
Indole-3-acetaldoxime
(IAOX)
CYP83B1
SUR1
Indole
glucosinolates
Indole-3-acetaldehyde
(IAD)
YUC
aux2
iaaH
tms2
AMI1
CYP71A13
AAO1
NIT1 ?
NIT2 ?
Indole-3-acetonitrile
(IAN)
CYP71B15
Camalexin
Indole-3-acetic acid
(IAA)
Brassicaceae species-specific pathway

Trait understanding





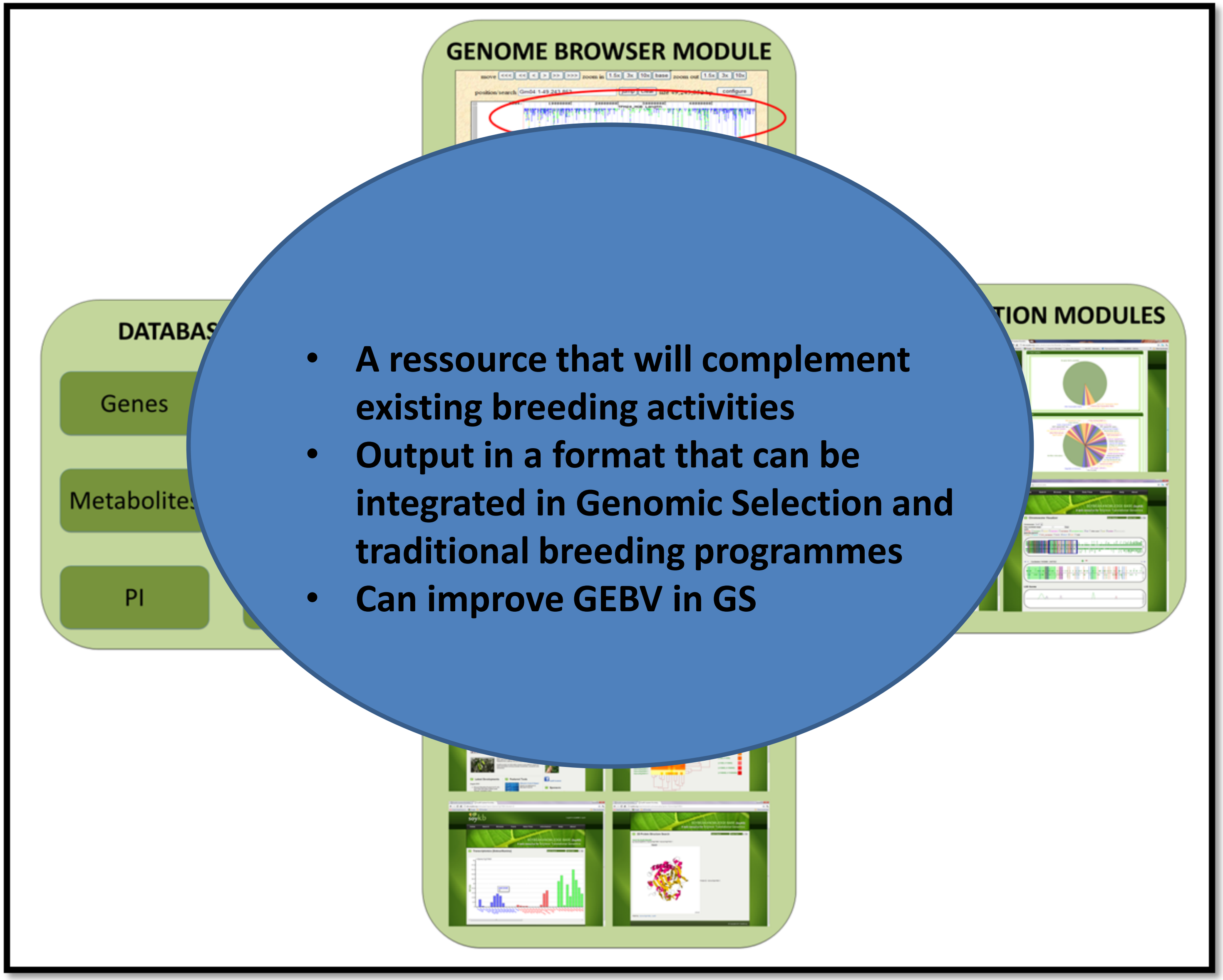
GENOME BROWSER MODULE

DATABASE MODULE

- Genes
- miRNA
- Metabolites
- SNPs
- PI
- Trait

DATA INTEGRATION MODULES

WEB INTERFACE MODULE



- A ressource that will complement existing breeding activities
- Output in a format that can be integrated in Genomic Selection and traditional breeding programmes
- Can improve GEBV in GS