Status of Work Package 2

Work Package 2 Deliverables

- Phase 2.1 - High Performance Computing (HPC) cluster hardware
- Phase 2.2 – Data security and confidentiality
- Phase 2.3 – Data standards and software
- Phase 2.4 – Database development
- Phase 2.5 – GWAS analysis and integration of phenotype and genotype data

Status of Work Package 2

- Phase 2.1 - High Performance Computing (HPC) cluster hardware ✓
- Phase 2.2 – Data security and confidentiality ✓
- Start of post doc delayed – will start January 4th

The GenomeDK Cluster – Before Upgrade

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.

www.genome.au.dk

BiRC

AARHUS UNIVERSITY

COMMENTS ON CONTENT: RUNE MØLLEGAARD FRIBORG
REVISED 2013.06.04

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 1.6 PB.

Data may be located either on PANASAS 110TB SAN storage (panfs), on one of eight EONSTOR SANs (nfs) or on our 1 PB FhGFS distributed file system (fast storage).

Home and project folders are located on panfs and nfs storage by default. The fast storage is reserved for large data files involved in I/O intensive computations.

Panfs and nfs can deliver read/write performance of up to 700MB/s, while the fast storage can reach an aggregated read/write performance of up to 25GB/s. Fast storage is a 32 node distributed file system running FhGFS.

Backup

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

AARHUS
UNIVERSITY
DEPARTMENT OF
MOLECULAR BIOLOGY AND GENETICS

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

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

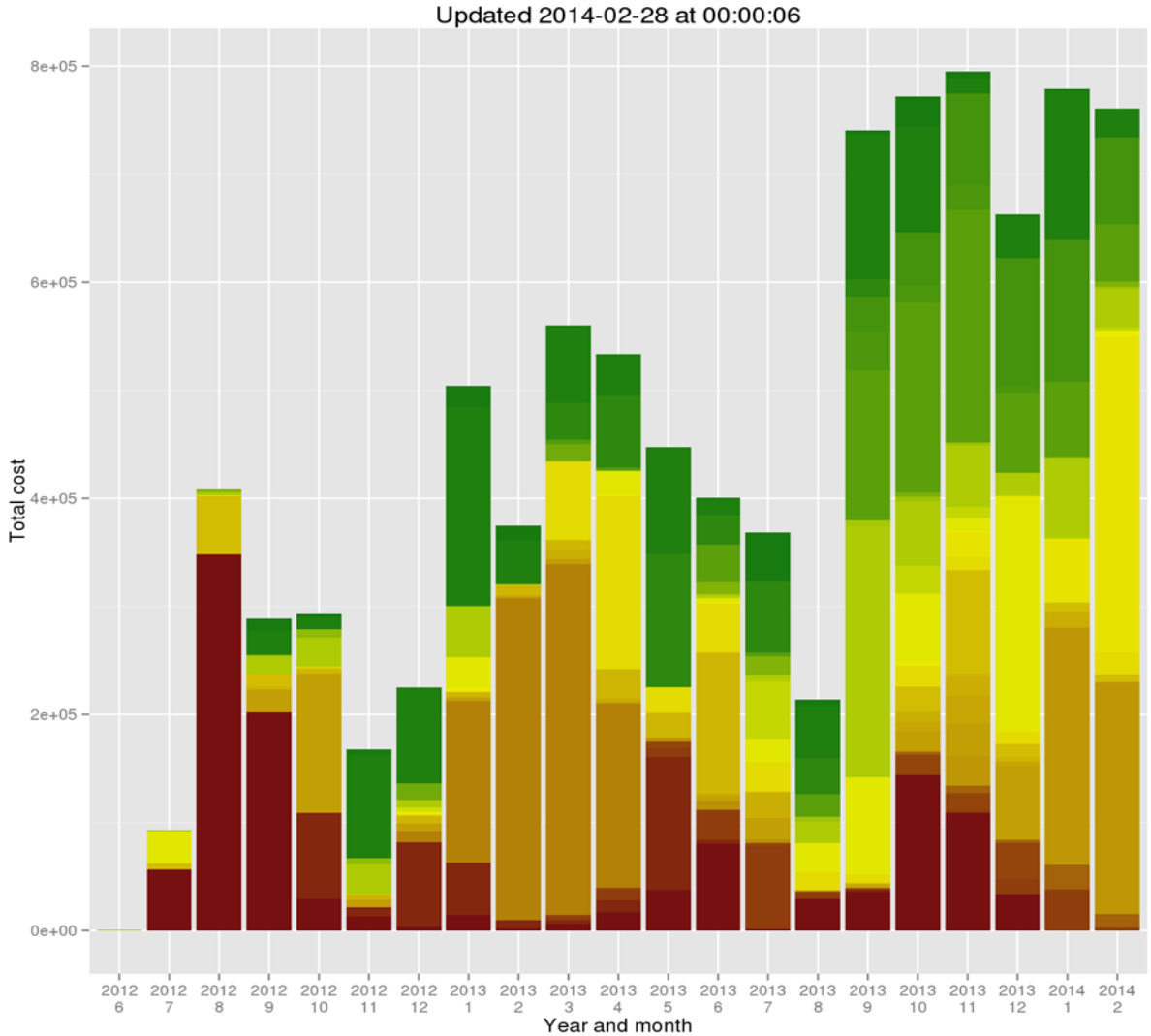
```
[user@fe1]$ vncserver -list
TigerVNC server sessions:

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).

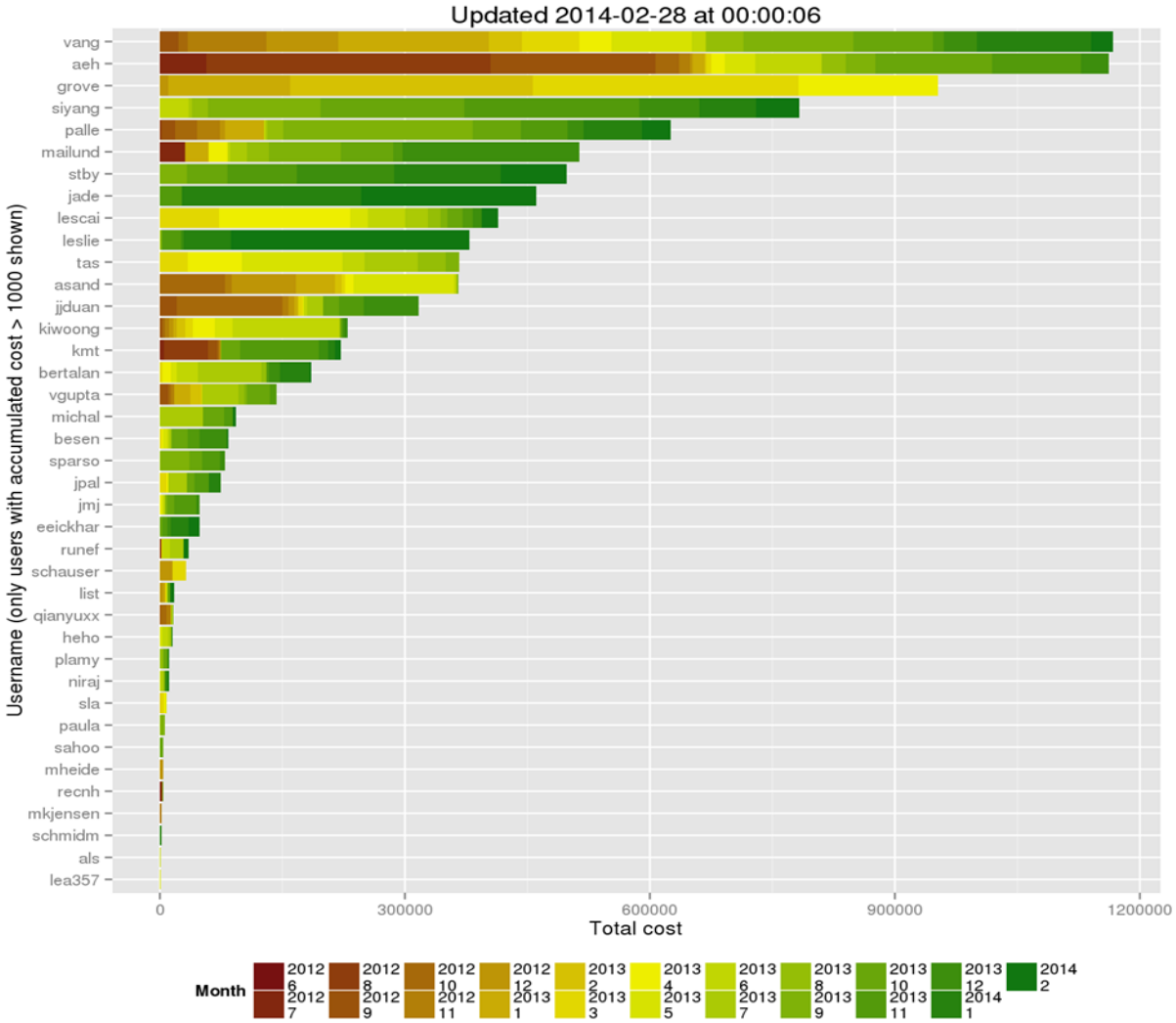
AARHUS
UNIVERSITY
DEPARTMENT OF
MOLECULAR BIOLOGY AND GENETICS

Load on the GenomeDK Cluster



AARHUS UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

Accumulated Costs



FAUPE contribution to GenomeDK

- 1.4 mill DKK investment in hardware as part of the FAUPE project
- Hardware integrated with existing GenomeDK cluster hardware
- Funding used for buying extra storage capacity and nodes to optimise the performance of GenomeDK
- CID users will because of the large investment be prioritized users
- Advantages:
  - Platform for high-performance computing in a secured system with backup
  - Easy sharing of large datastes between partners without data transfer
  - Platform for data visualization
    - Genome browers for genomes, transcriptoms and markers with password protected access will be established

Orthologs vs Paralogs

- **Homolog:** A gene related to a second gene by descent from a common ancestral DNA sequence. The term, homolog, may apply to the relationship between genes separated by the event of speciation (see ortholog) or to the relationship between genes separated by the event of duplication (see paralog).

- **Orthologs** are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution.

- **Paralogs** are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

AARHUS UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

Identification of Ortholog Groups by OrthoMCL

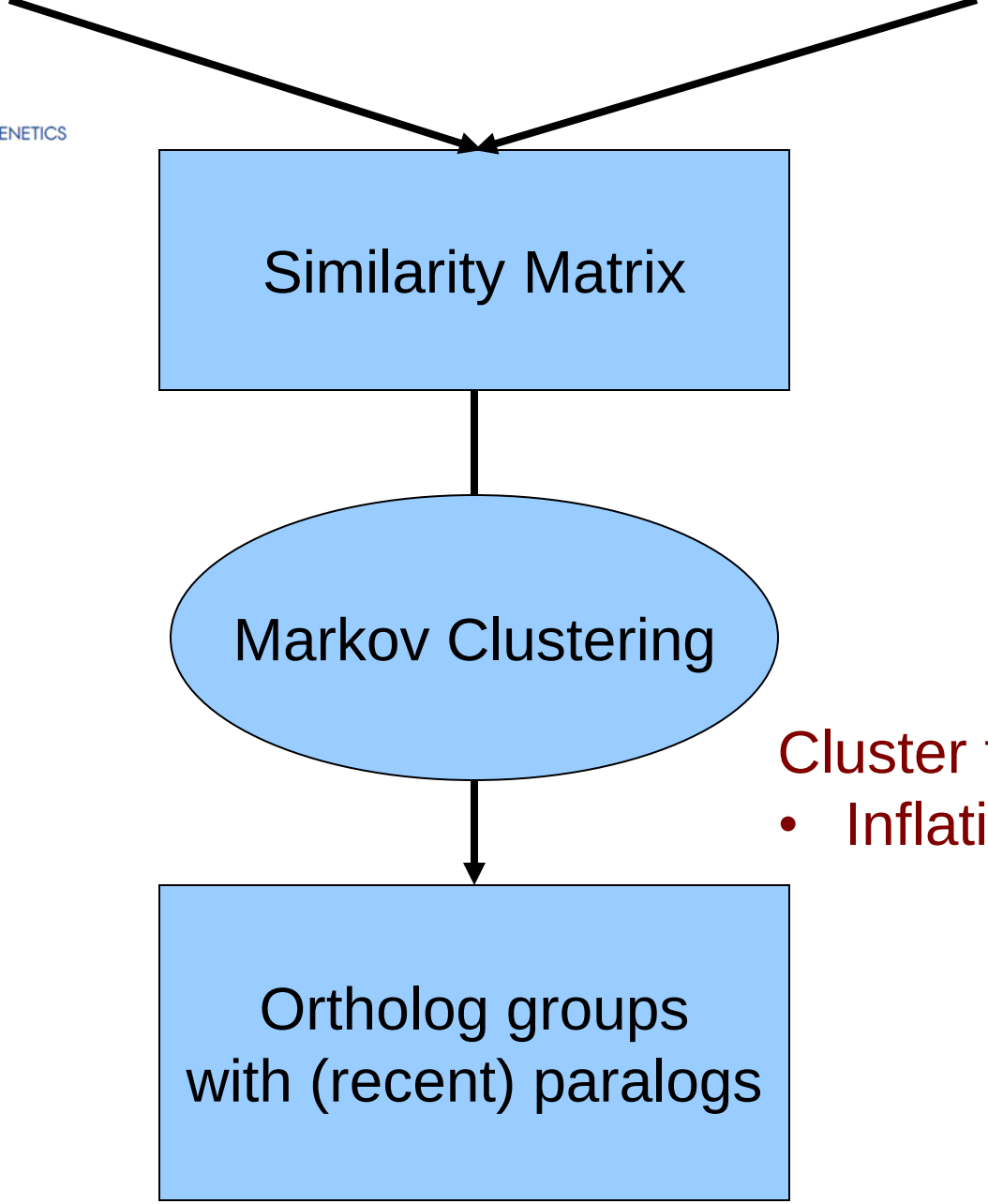
Protein sequences from organisms of interest

All-against-all BLASTP

Similarity cutoff:
- P-value
- % overlap

<u>Between Species:</u>
Reciprocal best similarity pairs
Putative orthologs

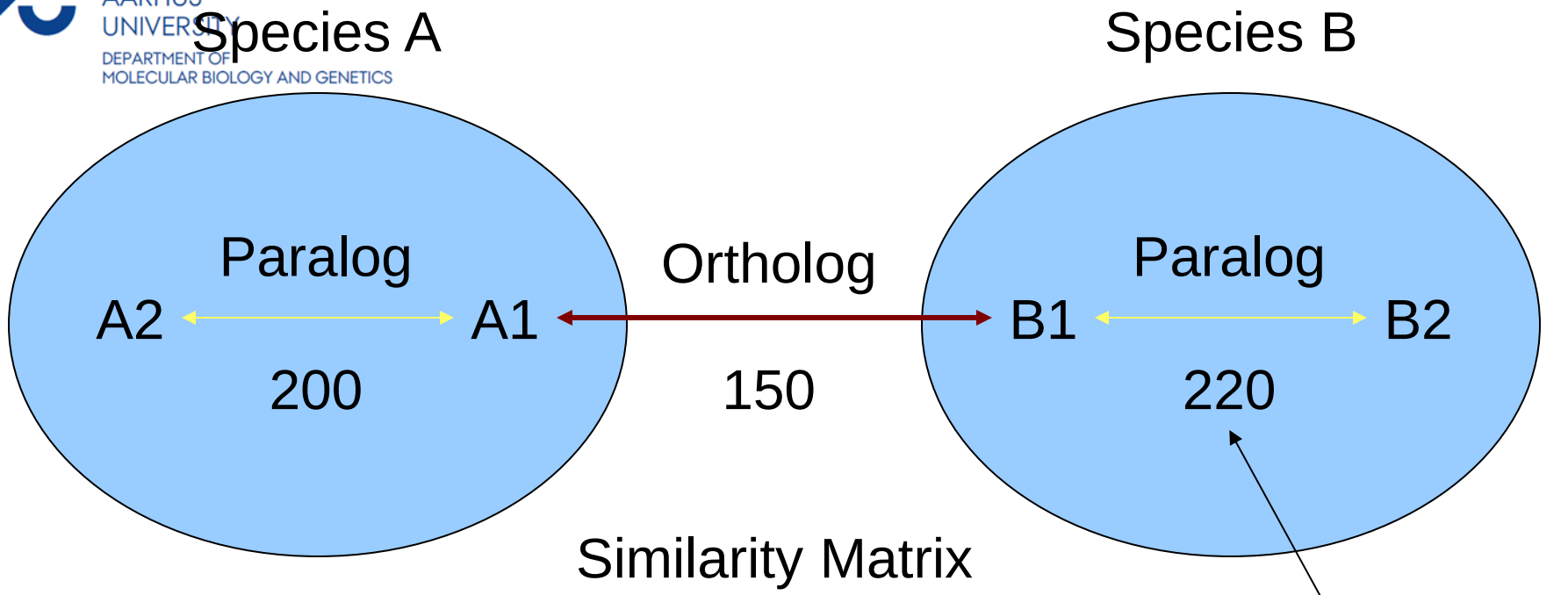
<u>Within Species:</u>
Reciprocal better similarity pairs
(Recent) paralogs



Cluster tightness:

- Inflation values (I)

AARHUS UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

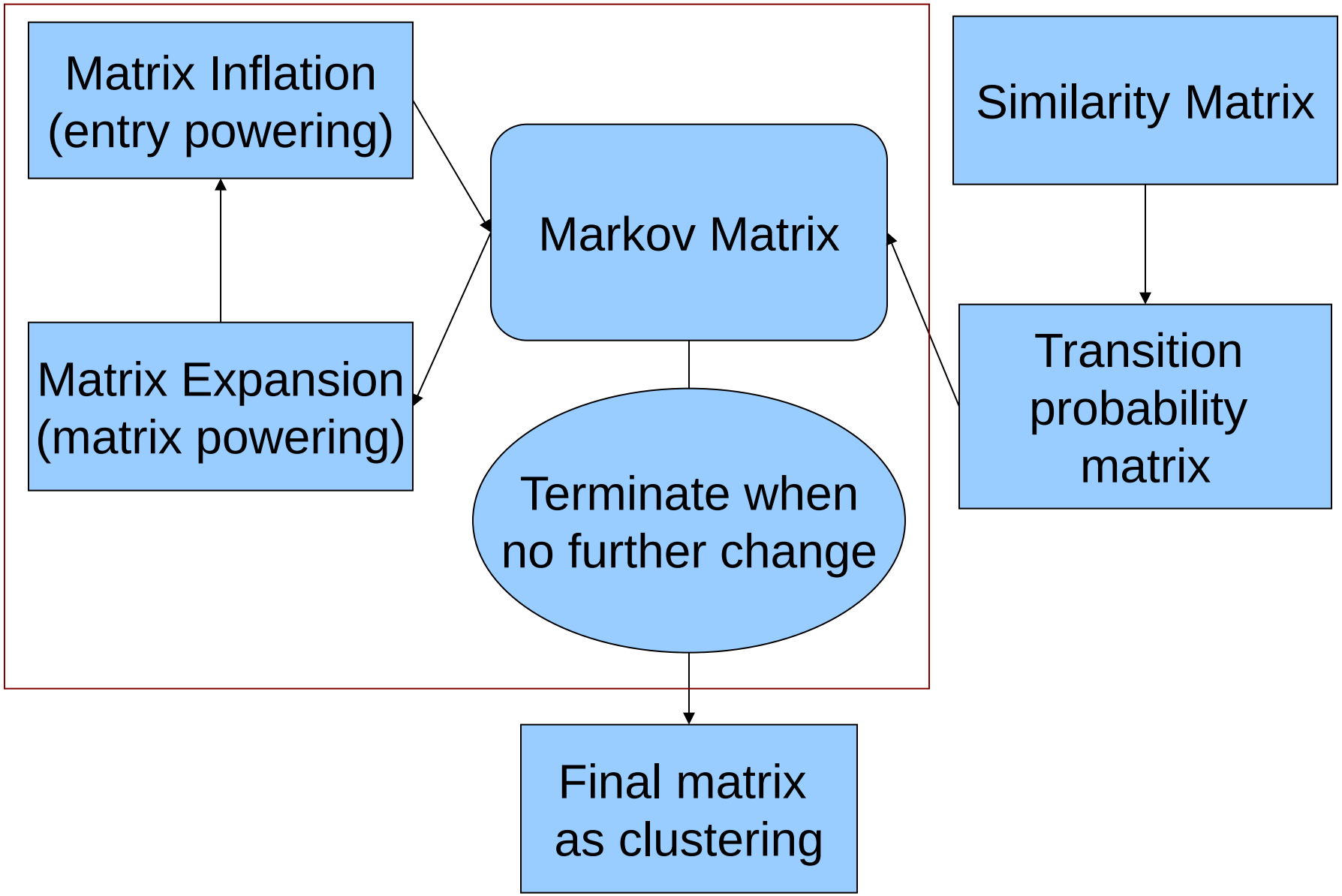


Similarity Matrix

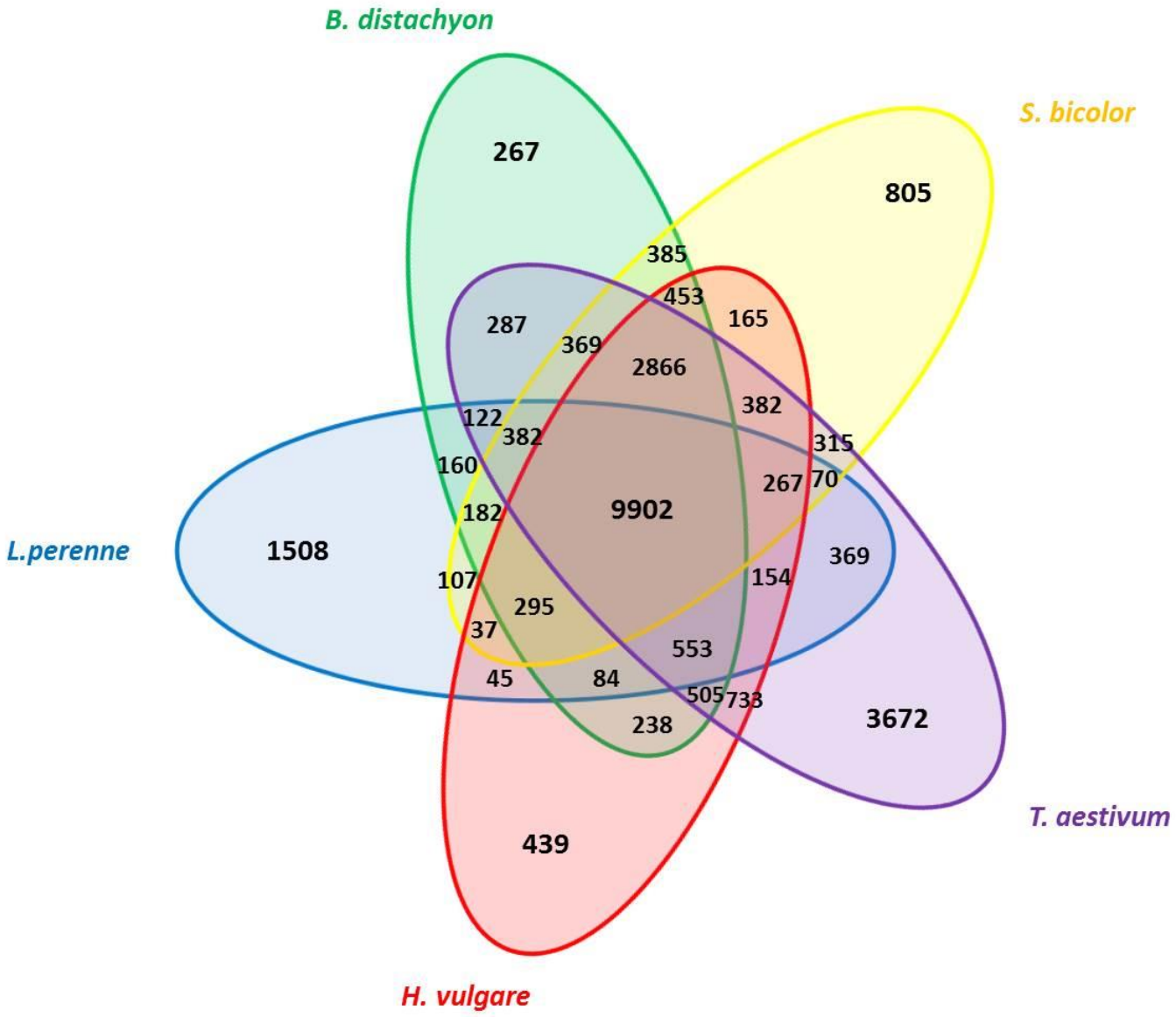
	A1	A2	B1	B2
A1	–	200	150	0
A2	200	–	0	0
B1	150	0	–	220
B2	0	0	220	–

Similarity score

Markov Clustering (MCL) Algorithm

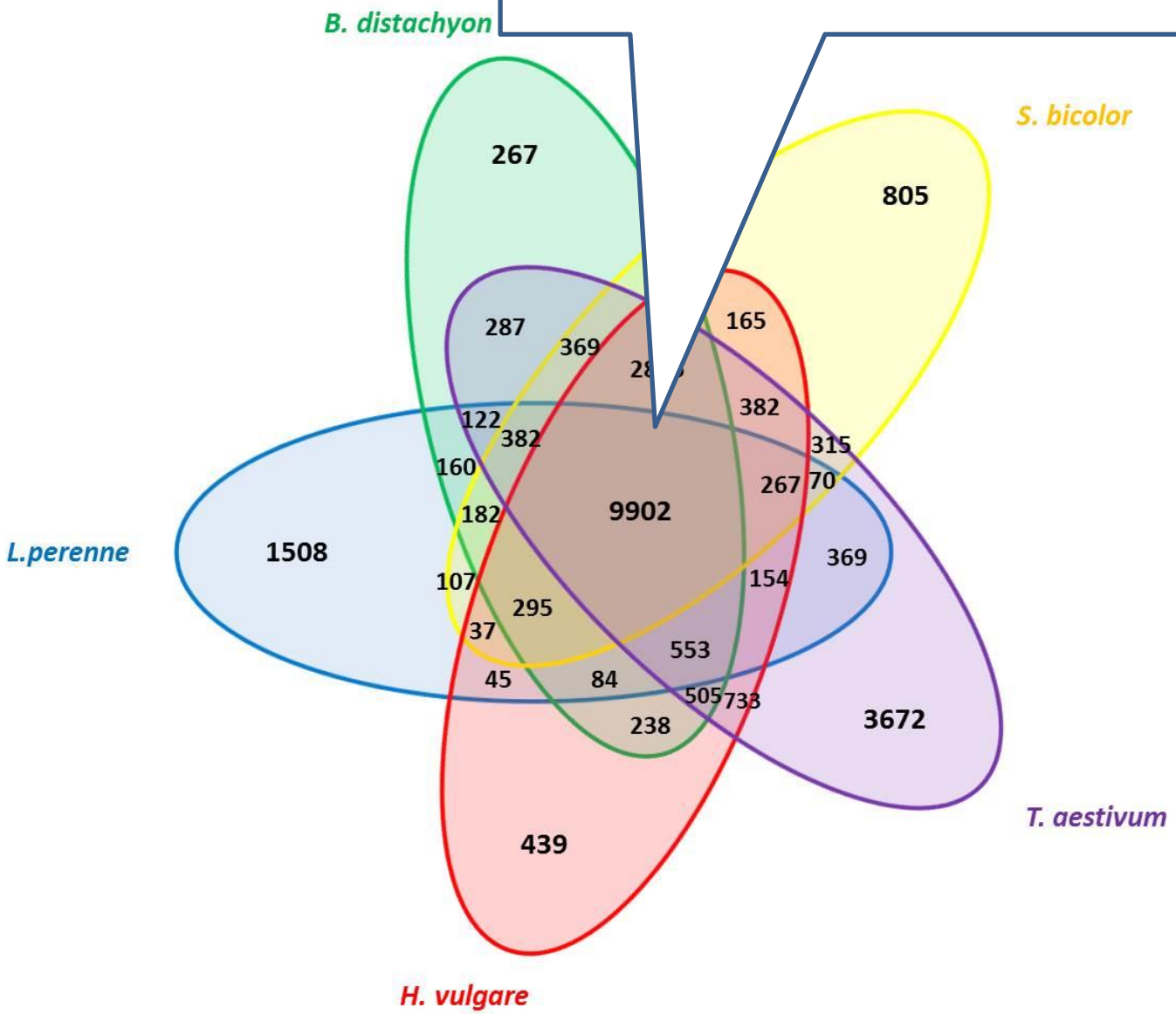


AARHUS
UNIVERSITY
DEPARTMENT OF
MOLECULAR BIOLOGY AND GENETICS



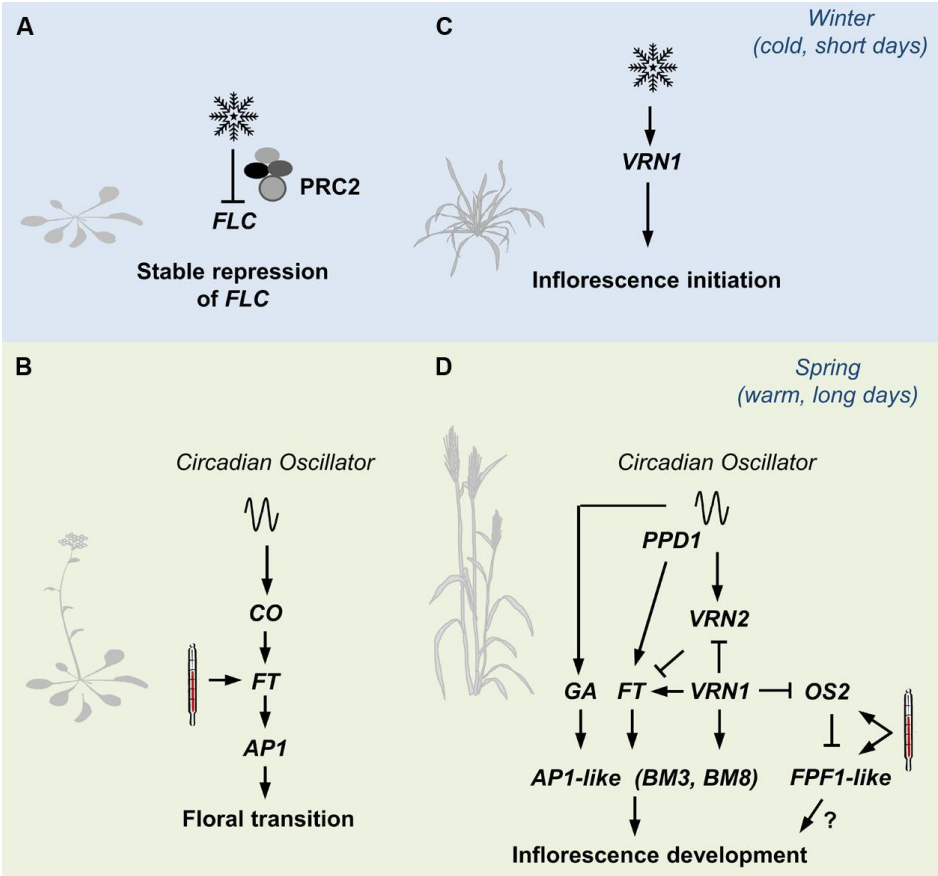
AARHUS UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

Potato will be included as the next species



AARHUS UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

Seasonal Flowering Responses in Grasses and Arabidopsis



(A) The prolonged cold of winter triggers lasting repression of FLOWERING LOCUS C (FLC) in Arabidopsis, via the PolycombRepressor Complex2(PRC2).

(B) The long days of spring activate expression of FLOWERING LOCUS T, a process mediated by the circadian oscillator via CONSTANS (CO). FT activates expression of genes such as a APETALA1 that trigger floral development. High-temperatures can also activate expression of FT to accelerate flowering.

(C) Winter cold activates expression of VERNALIZATION1 (VRN1) in cereals and related grasses. VRN1 promotes inflorescence initiation at the shoot.

(D) VRN1 remains active after winter and down-regulates VRN2, which would otherwise repress the long-day flowering response in leaves. As day length increases after winter, expression of FT-like1 is activated by the circadian oscillator,via PHOTOPERIOD1. The long-day flowering response activates expression of genes at the shoot apex that promote the development of floral organs.

*Fjellheim et al., 2014 (Front. Plant Sci., doi: 10.3389/fpls.2014.00431)

AARHUS UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

VRN2 Orthologous Groups

Decreasing degree of perenniallity, percistency, and cold tolerance

VRN2	Group16799	Group11720
F. pratensis	1	1
L. perenne	1	1
L. multiflorum	0	0
L. m. westerwoldicum	0	0
L. temulentum	0	0

AARHUS
UNIVERSITY
DEPARTMENT OF
MOLECULAR BIOLOGY AND GENETICS

PPD-H1 Orthologous Groups

Ppd-H1 Pseudo-response regulator	Group2472	Group11735	Group37150
F. pratensis	4	1	0
L. perenne	1	1	0
L. multiflorum	2	2	0
L. m. westerwoldicum	2	2	1
L. temulentum	2	0	1

Working with Gene Orthologs Allows to:

- Transfer of information from one species to another
  - Create a genome-wide link at the gene level across species
  - Candidate genes, or markers, for a trait can be transferred across species
  - QTL
- Prediction across-species for genomic selection; i.e. RadiMax
- Identification of genes unique in a species (rapidly evolving genes)
  - Disease resistance genes
  - Unique genes for a trait of interest in one species; i.e. endophytes in grasses
  - Plant organs; i.e. potato tubers