
bioportal
www.bioportal.uio.no

Contents

1. **Bioportal**

   A) Registration.
   B) Managing projects, files, and jobs.
   C) Submitting / checking jobs.

2. **AIR (Appender, Identifier, and Remover)**

3. **PhyloSity**

If you do not have Bioportal account.

Step 1: Feide –Login using your UIO email address (Instant access / Registration)

Step 2: Email: **bioportal-drift@usit.uio.no**

Subject: bpcourse access.

If you already have Bioportal account:

Proceed with Step 2 only.

Bioportal

Bioportal is a web-based bioinformatics service at University of Oslo (http://www.bioportal.uio.no/).

- 590 CPU total connected to the Bioportal
- Additional access to over 4000 CPUs on TITAN cluster
- Continual software upgrades.
- Total number of jobs in 2009 = 17 515 (>1 500 000 CPU hours).
- Till today = 15332 jobs already.

Applications available on Bioportal

Phylogenetic analysis

- MrBayes
- PAUP
- PhyML
- Phylobayes
- Garli
- PAML
- Modeltest/Protest
- RAxML
- PHASE
- POY
- Treefinder
- BEAST
- AIR

Population genetics

- FAMHAP
- LAMARC
- NPMLE
- STRUCTURE
- PHASE
- UNPHASED
- SIMWALK2
- PSCL

Bioinformatics applications

- Blast
- MAFFT
- PhyloSity
- Newbler
- Pfam
- PhredPhrap
- AUTODOCK4
- Adscreening
- Preassemble
- Transeq

Chemistry / Statistical application

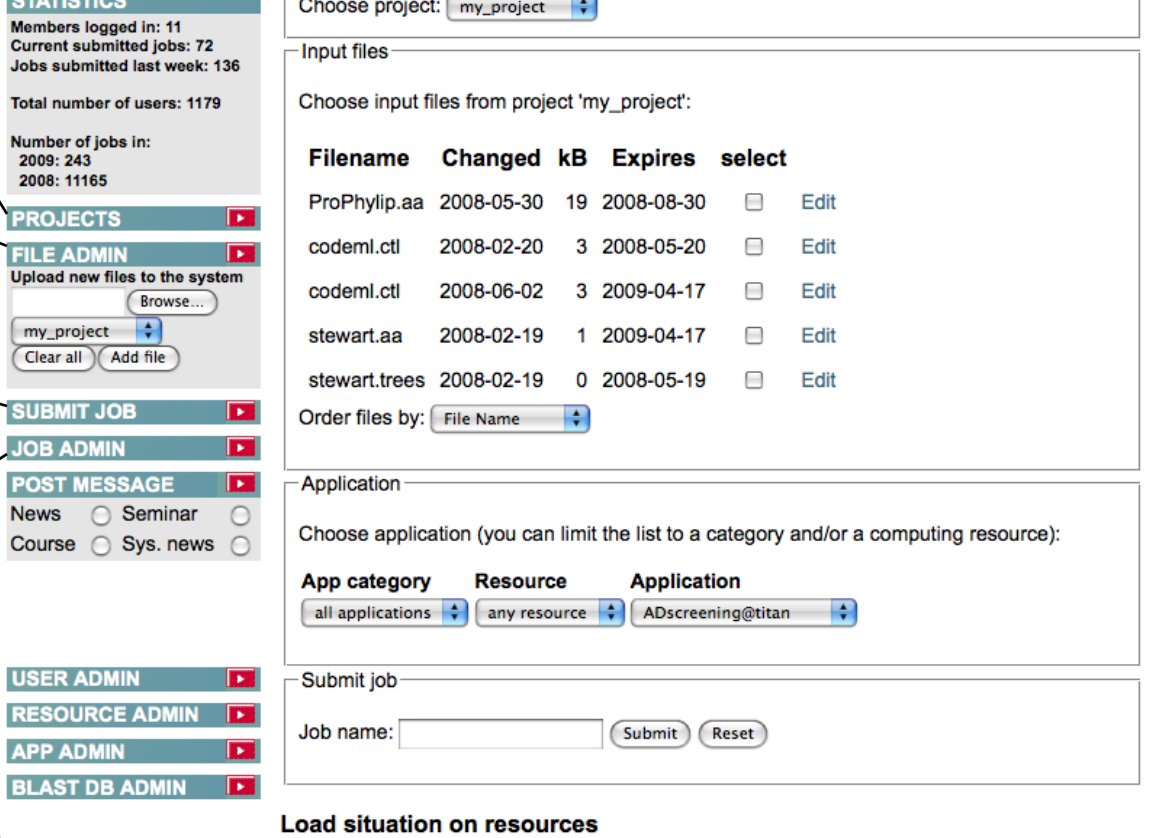
- DALTON
- DIRAC
- GAUSSIAN
- Meltprofile

Create/Manage project

Upload files

Select files and Application

Check status of submitted jobs



BMC Bioinformatics

BioMed Central

Software

Open Access

AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses

Surendra Kumar[1], Åsmund Skjæveland[1], Russell JS Orr[1], Pål Enger[1,2], Torgeir Ruden[2], Bjørn-Helge Mevik[2], Fabien Burki[3], Andreas Botnen[2] and Kamran Shalchian-Tabrizi*[1]

Address: [1]Microbial Evolution Research Group (MERG), Department of Biology, University of Oslo, Norway, [2]Centre of Information Technology, University of Oslo, Norway and [3]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

Email: Surendra Kumar - surendra.kumar@bio.uio.no; Åsmund Skjæveland - asmund.skjaveland@bio.uio.no; Russell JS Orr - russell.orr@bio.uio.no; Pål Enger - pal.enger@usit.uio.no; Torgeir Ruden - t.a.ruden@usit.uio.no; Bjørn-Helge Mevik - b.h.mevik@usit.uio.no; Fabien Burki - burkif@interchange.ubc.ca; Andreas Botnen - andreas.botnen@gmail.com; Kamran Shalchian-Tabrizi* - Kamran@bio.uio.no

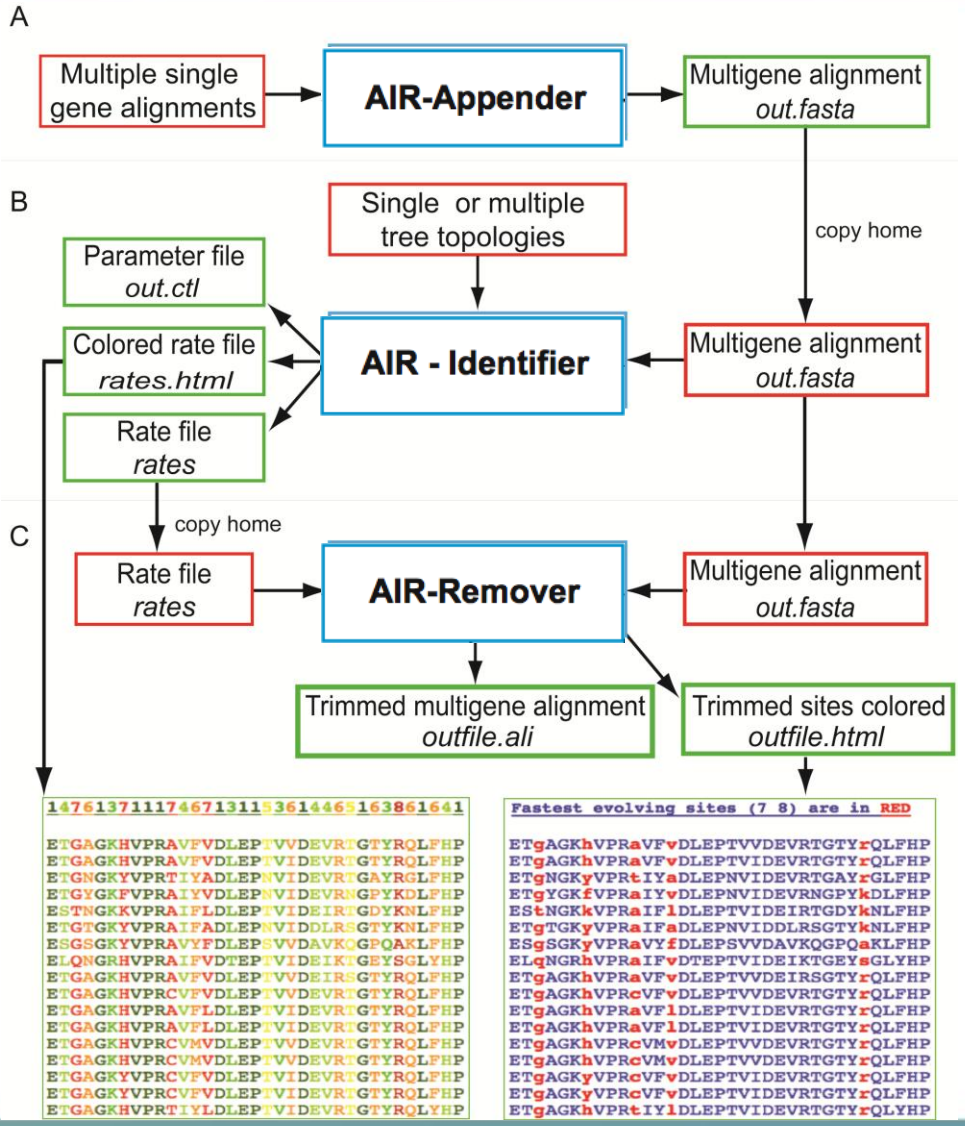
* Corresponding author

Published: 28 October 2009

BMC Bioinformatics 2009, **10**:357 doi:10.1186/1471-2105-10-357

Received: 21 April 2009
Accepted: 28 October 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/357



Merge several single gene alignment into one multi gene alignment

Identifying fast evolving sites

Removing fast evolving sites

A

File 1

>Human
atgcatgcatgcatgc
> Rat
atgcatgcatgcatgc
>Cow
atgcatgcatgcatgc
>Horse
atgcatgcatgcatgc
>Dog
atgcatgcatgcatgc

File 2

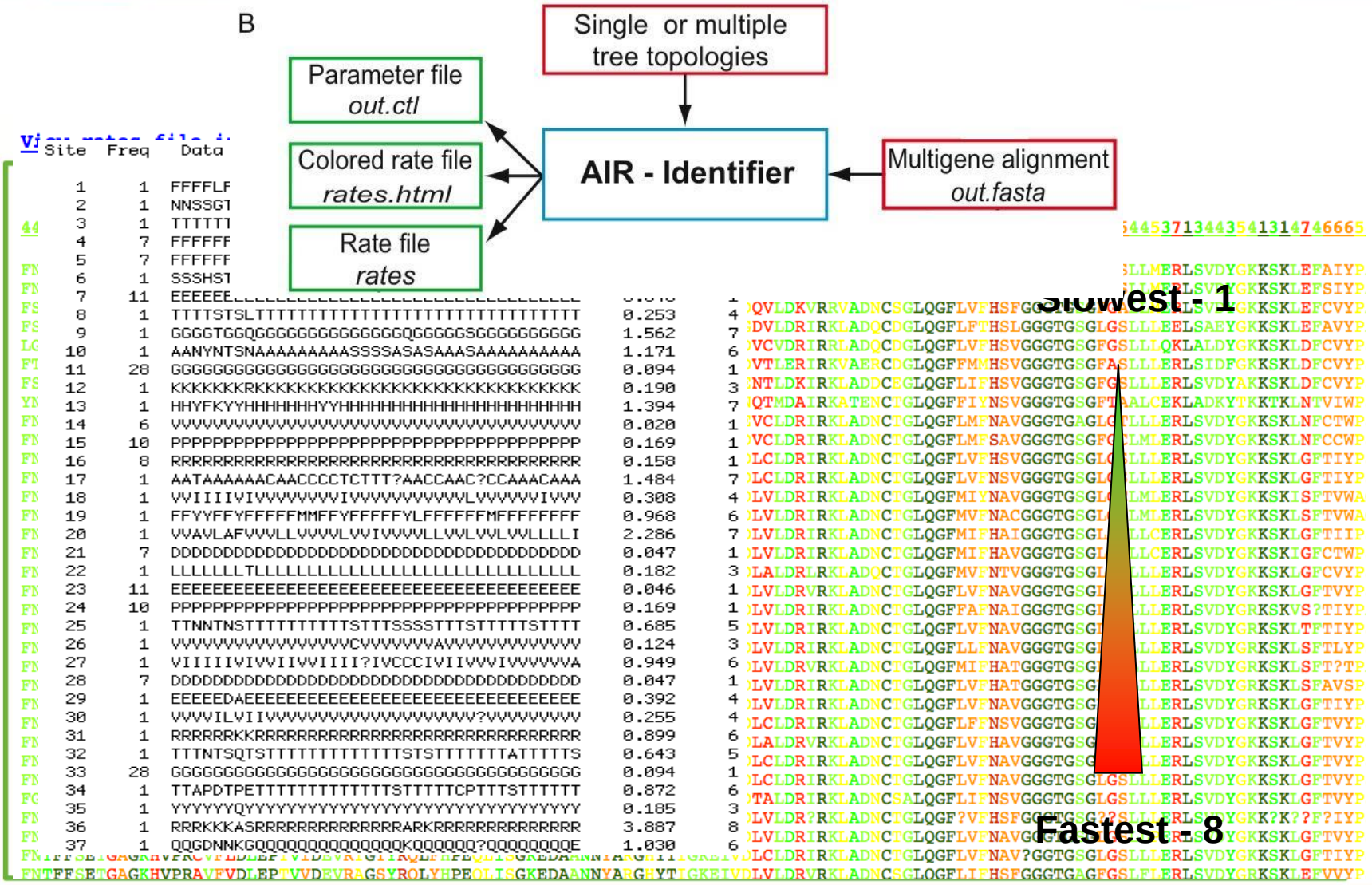
>Human
ATGCATGCATGCATGC
>Rat
ATGCATGCATGCATGC
>Cow
ATGCATGCATGCATGC
>Horse
ATGCATGCATGCATGC
>Dog
ATGCATGCATGCATGC

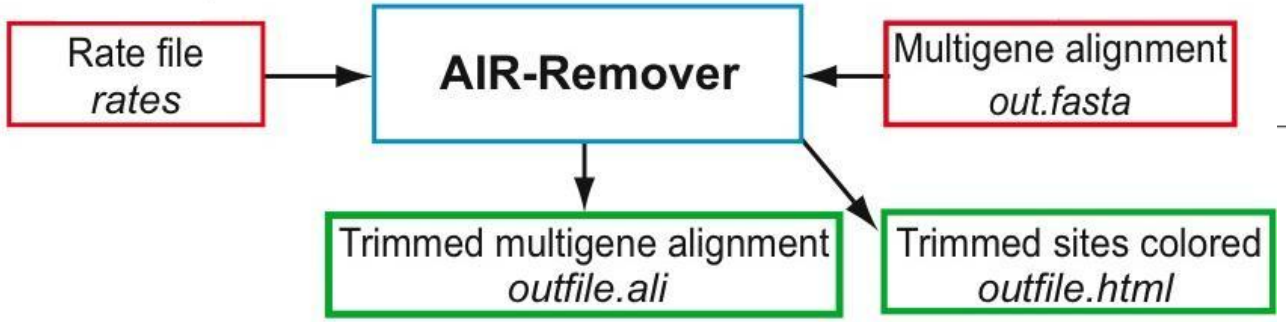
File n

>Human
atgcatgcatgcatgc
>Rat
atgcatgcatgcatgc
>Cow
atgcatgcatgcatgc
>Dog
atgcatgcatgcatgc

>Human
atgcatgcatgcatgc--ATGCATGCATGCATGC--atgcatgcatgcatgc
> Rat
atgcatgcatgcatgc--ATGCATGCATGCATGC--atgcatgcatgcatgc
>Cow
atgcatgcatgcatgc--ATGCATGCATGCATGC--atgcatgcatgcatgc
>Horse
atgcatgcatgcatgc--ATGCATGCATGCATGC--????????????????
>Dog
atgcatgcatgcatgc--ATGCATGCATGCATGC--atgcatgcatgcatgc

PAML (by ZihengYang) - http://abacus.gene.ucl.ac.uk/software/





- **Generates a ready to use alignment file after removing the fast evolving sites.**
- **Generates output file with statistics about the sites in rates files.**
- **Generates colored alignment file displaying the removed sites in RED.**

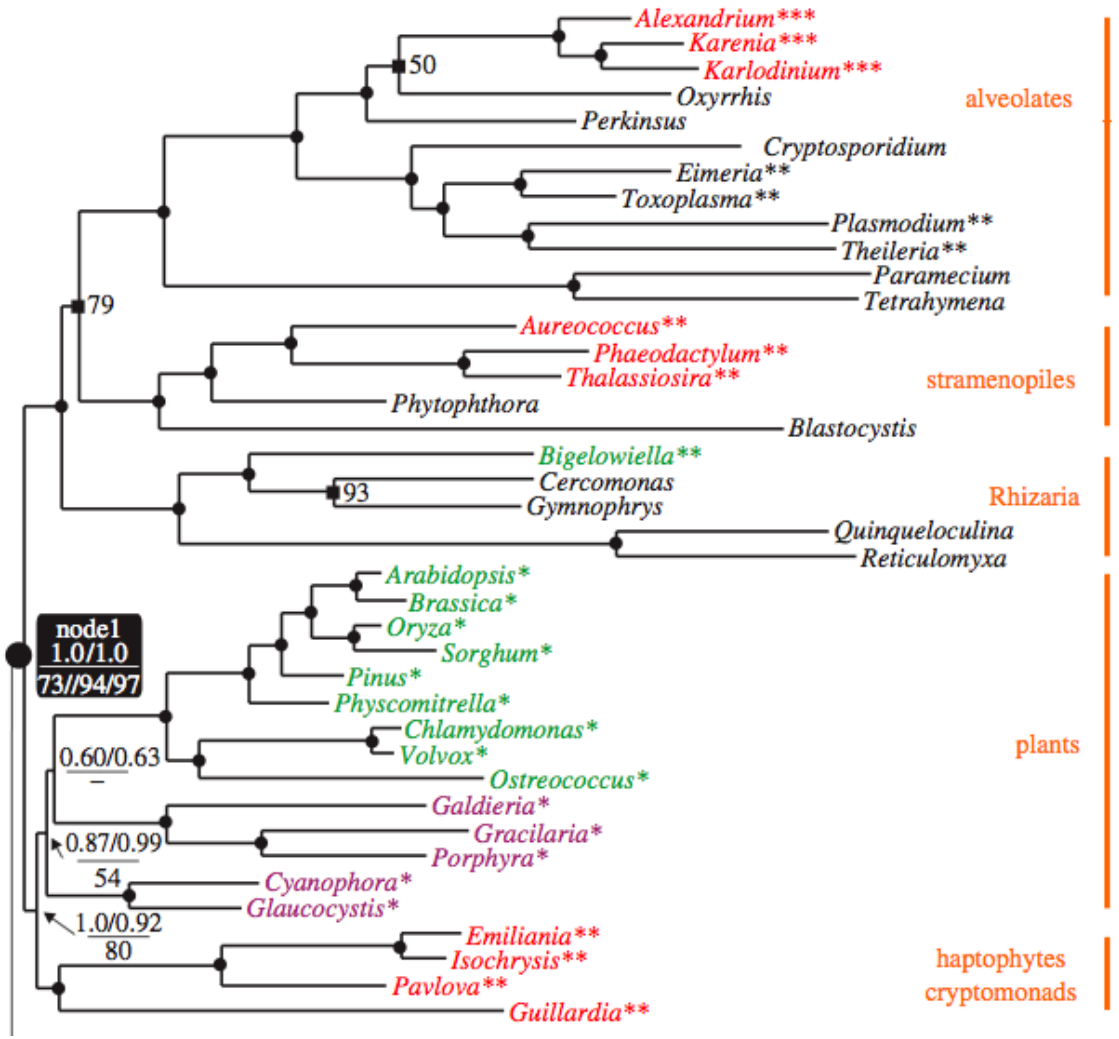
biology letters
Evolutionary biology
Biol. Lett. (2008) 4, 366–369
doi:10.1098/rsbl.2008.0224
Published online 3 June 2008

Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes

Fabien Burki[1,*], Kamran Shalchian-Tabrizi[2] and Jan Pawlowski[1]

[1]*Department of Zoology and Animal Biology, University of Geneva, 1211 Geneva 4, Switzerland*
[2]*Microbial Evolution Reseach Group, Department of Biology, University of Oslo, 1066 Blindern, 0316 Oslo, Norway*
**Author for correspondence (fabien.burki@zoo.unige.ch).*

135 genes
65 taxa



Phylosity - An online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation

Test Dataset download link:

https://www.bioportal.uio.no/onlinemat/online_material.php

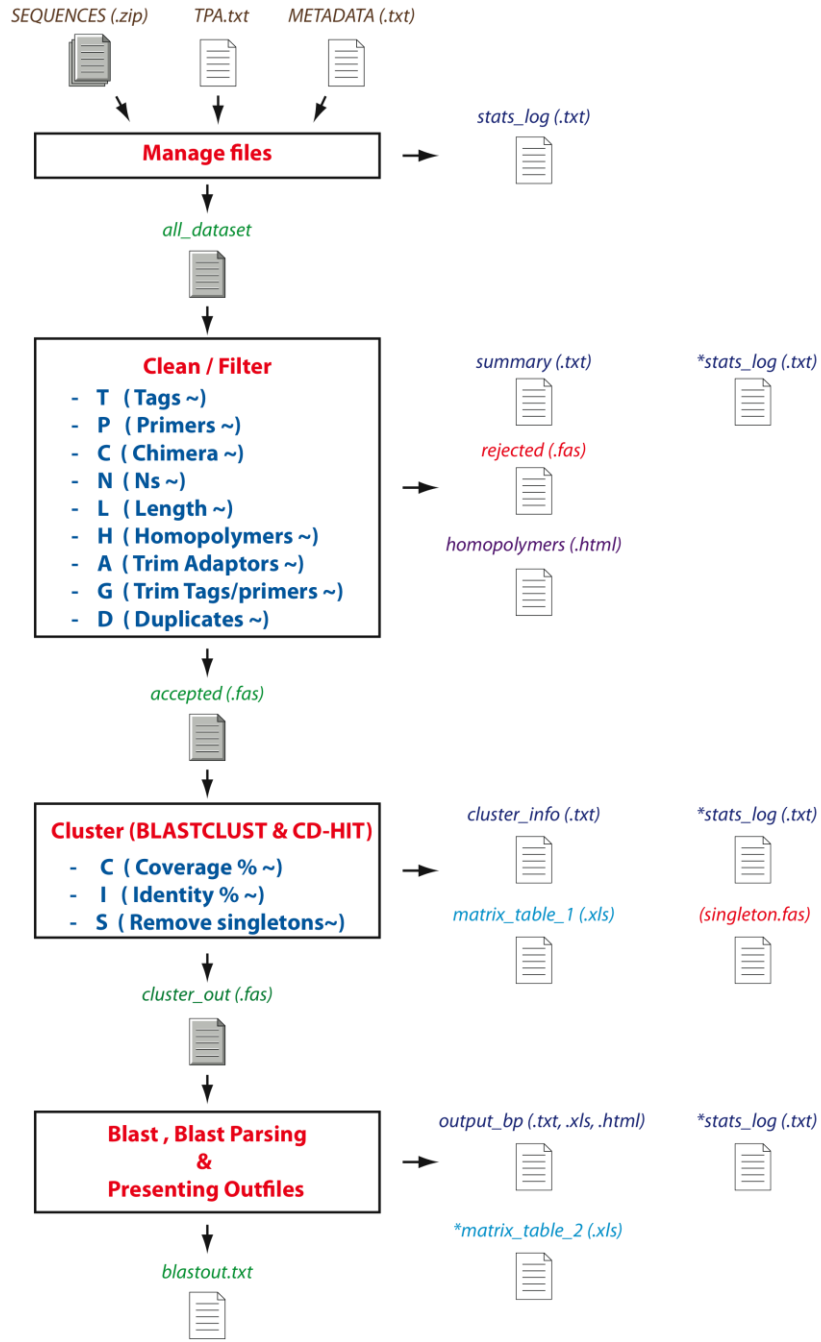
The Pipeline includes :-

Three steps:

1. Filtering low quality sequence reads followed by trimming the undesired segment.

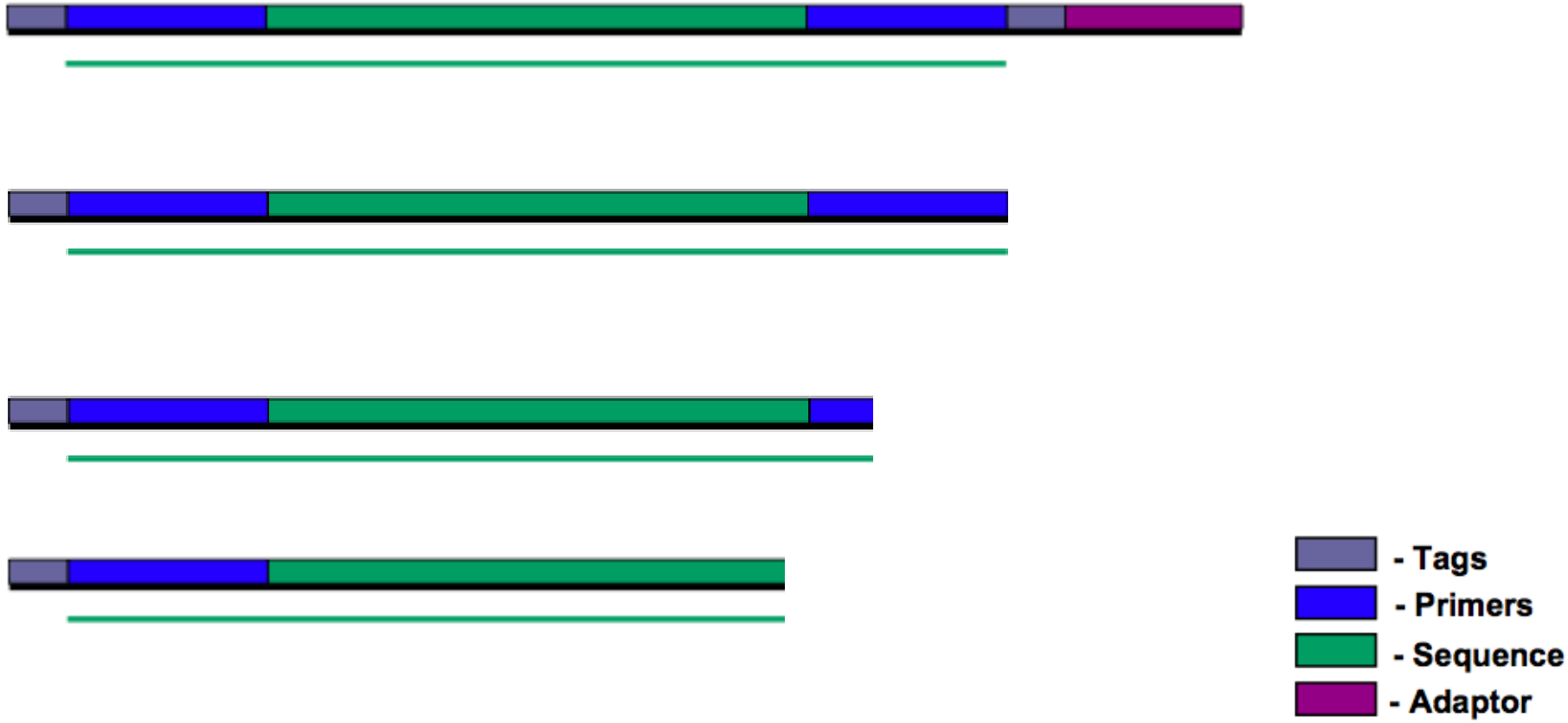
2. Clustering sequence reads in Operational Taxonomic Unit (OTUs).

3. Taxonomic annotation of sequences / OTUs using BLAST.


SEQUENCES (.zip)
TPA.txt
METADATA (.txt)
Manage files
stats_log (.txt)
all_dataset
Clean / Filter
- T (Tags ~)
- P (Primers ~)
- C (Chimera ~)
- N (Ns ~)
- L (Length ~)
- H (Homopolymers ~)
- A (Trim Adaptors ~)
- G (Trim Tags/primers ~)
- D (Duplicates ~)
summary (.txt)
*stats_log (.txt)
rejected (.fas)
homopolymers (.html)
accepted (.fas)
Cluster (BLASTCLUST & CD-HIT)
- C (Coverage % ~)
- I (Identity % ~)
- S (Remove singletons~)
cluster_info (.txt)
*stats_log (.txt)
matrix_table_1 (.xls)
(singleton.fas)
cluster_out (.fas)
Blast , Blast Parsing
&
Presenting Outfiles
output_bp (.txt, .xls, .html)
*stats_log (.txt)
*matrix_table_2 (.xls)
blastout.txt

Input files

1. Raw or Processed 454 Sequence data (.zip)



2. METADATA list (.txt)

```
FVCIMFP01.fna  S1
FVCIMFP02.fna  S2
FVCIMFP03.fna  S3
```

FVCIMFP01.fna

```
>FVCIMFP01AS1H7 length=251
AACAACGC……..
>FVCIMFP01ATTT6 length=201
AACAACGC……..
>FVCIMFP01APYQN length=227
TCACTCGC……..
```

FVCIMFP02.fna

```
>FVCIMFP01AS1R7 length=281
AACAACGC……..
>FVCIMFP01ATTS6 length=281
AACAACGC……..
>FVCIMFP01APYAN length=247
TCACTCGC……..
```

FVCIMFP03.fna

```
>FVCIMFP01AS15I length=281
AACAACGC……..
>FVCIMFP01ATTG2 length=281
AACAACGC……..
>FVCIMFP01APHGR length=247
TCACTCGC……..
```

```
>S1|FVCIMFP01AS1H7 length=251
AACAACGC……..
>S1|FVCIMFP01ATTT6 length=201
AACAACGC……..
>S1|FVCIMFP01APYQN length=227
TCACTCGC……..
>S2|FVCIMFP01AS1R7 length=281
AACAACGC……..
>S2|FVCIMFP01ATTS6 length=281
AACAACGC……..
>S2|FVCIMFP01APYAN length=247
TCACTCGC……..
>S3|FVCIMFP01AS15I length=281
AACAACGC……..
>S3|FVCIMFP01ATTG2 length=281
AACAACGC……..
>S3|FVCIMFP01APHGR length=247
TCACTCGC……..
```

3. TPA file(.txt)

<Tags>

AACAAC

AACCGA

GGCTAC

TTCTCG

<Forward primer>

GCTGCGTTCTTCATCGATGC

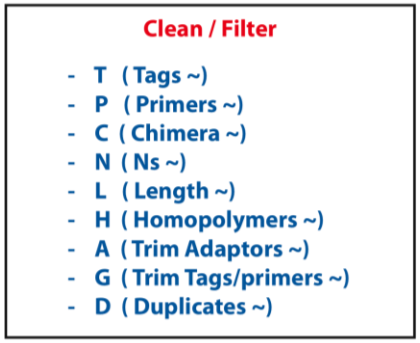
<Reverse Primer>

CCTTGTTACGACTTTTACTTCC

<Adaptor>

CTGATGGCGCGAGGGAGGC

<EOF>



Filtering and Trimming

- T - Sequences with incorrect tags
- P - Sequences with non-matching primers
- C – Sequences with non-compatible tags
- N - Sequences with Ns
- L - Sequences with length < user specified value (e.g. 150)
- H - Collapse homopolymers
- D - Identical sequences
- G - Trim tags or/and primers from the sequences
- A - Trim Adaptor sequences

Accepted

Rejected

>S3|FVCIMFP10F76XJ|308
>S3|FVCIMFP10F76XJ|308|T-TTCTCG
>S3|FVCIMFP10F76XJ|308|T-TTCTCG|FPY
>S3|FVCIMFP10F76XJ|308|T-TTCTCG|FPY|RPY
>S3|FVCIMFP10F76XJ|308|T-TTCTCG|FPY|RPY|rTY-CGAGAA
>S3|FVCIMFP10F76XJ|308|T-TTCTCG|FPY|RPY|rTY-CGAGAA|AR

Trim tags

Length

Duplicates

Homopolymers

>**S3**|FVCIMFP10F76XJ|308|**T**-TTCTCG|**FPY**|**RPY**|**rTY**-CGAGAA|**AR**|**GB**|**L**=287|**D**_2|**HP**_1_287:286

Clustering - BLASTCLUST

- Clustering by a single-linkage method.
- The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster.
- BLASTCLUST used megablast algorithm for DNA sequences and blastp for protein sequences.
- Longest sequence is the representative sequences of each cluster.

ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.24/

Clustering - CDHIT

- Fast greedy incremental clustering process.
- Sequences are first sorted in order of decreasing length.
- The longest one becomes the representative of the first cluster
- Then, each remaining sequence is compared to the representatives of existing cluster.

Clustering - CDHIT

- If the similarity with any representative is above a given threshold, it is grouped into that cluster.

- Otherwise, a new cluster is defined with that sequence as representative.

Download link:

http://www.bioinformatics.org/cd-hit/

Blast search

- BLASTN
- Blast search parameters
- NCBI-nr / custom database
- Custom databases is not automated but can be made available within the pipeline on Bioportal
- Blast parsing options (Overlapping% & Identity%)

Test Dataset download link:

https://www.bioportal.uio.no/onlinemat/online_material.php

QUESTIONS ?