CSIRO tools presentation

Dealing with Big data

Philip Moncuquet | Bioinformatician

24 March 2014

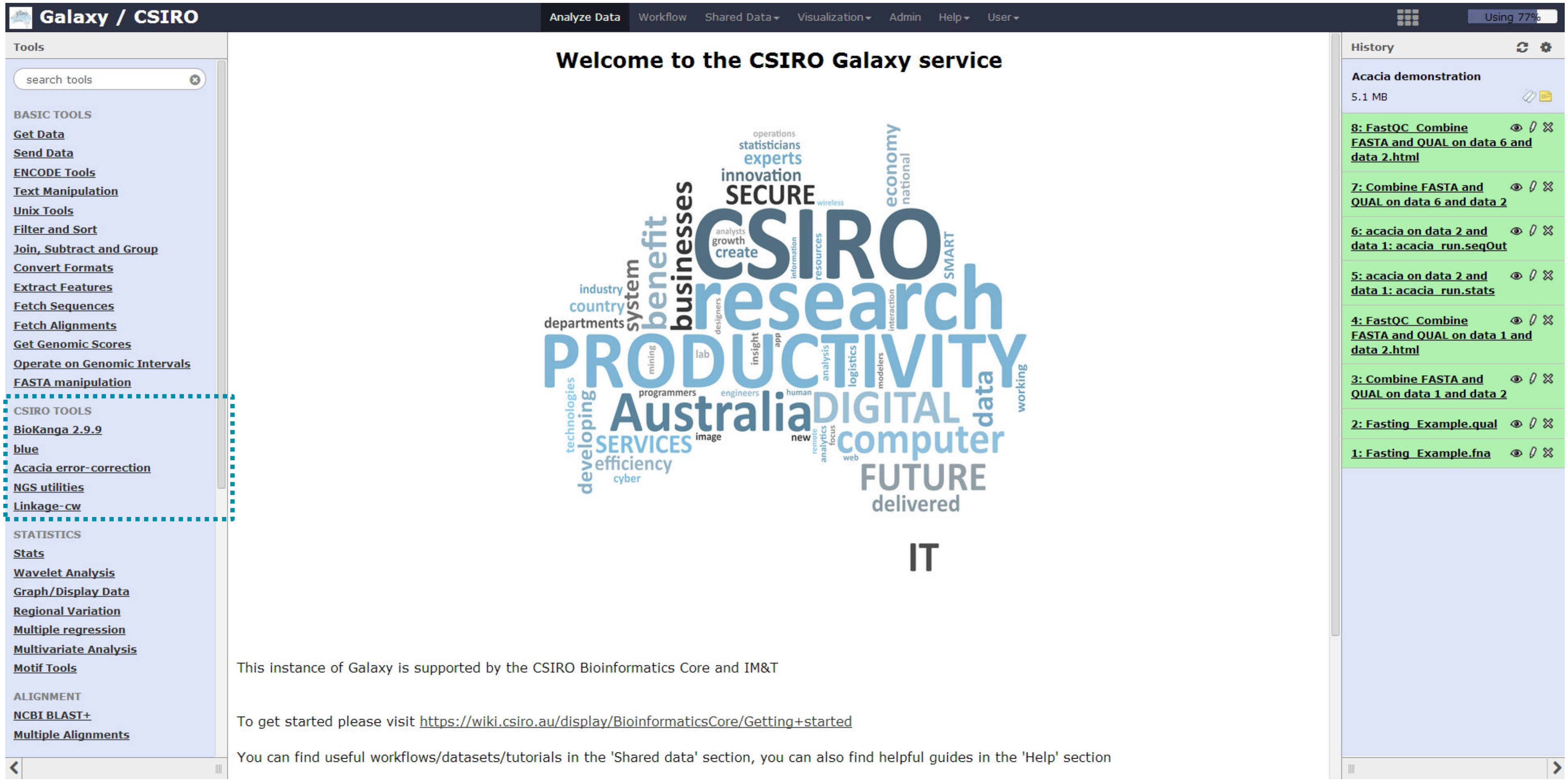
Galaxy Australasia Workshop 2014

CSIRO

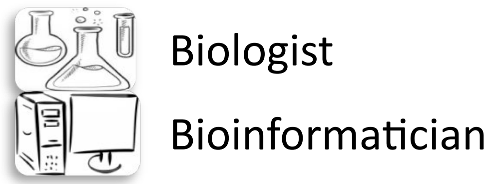
1 – CSIRO in house tools

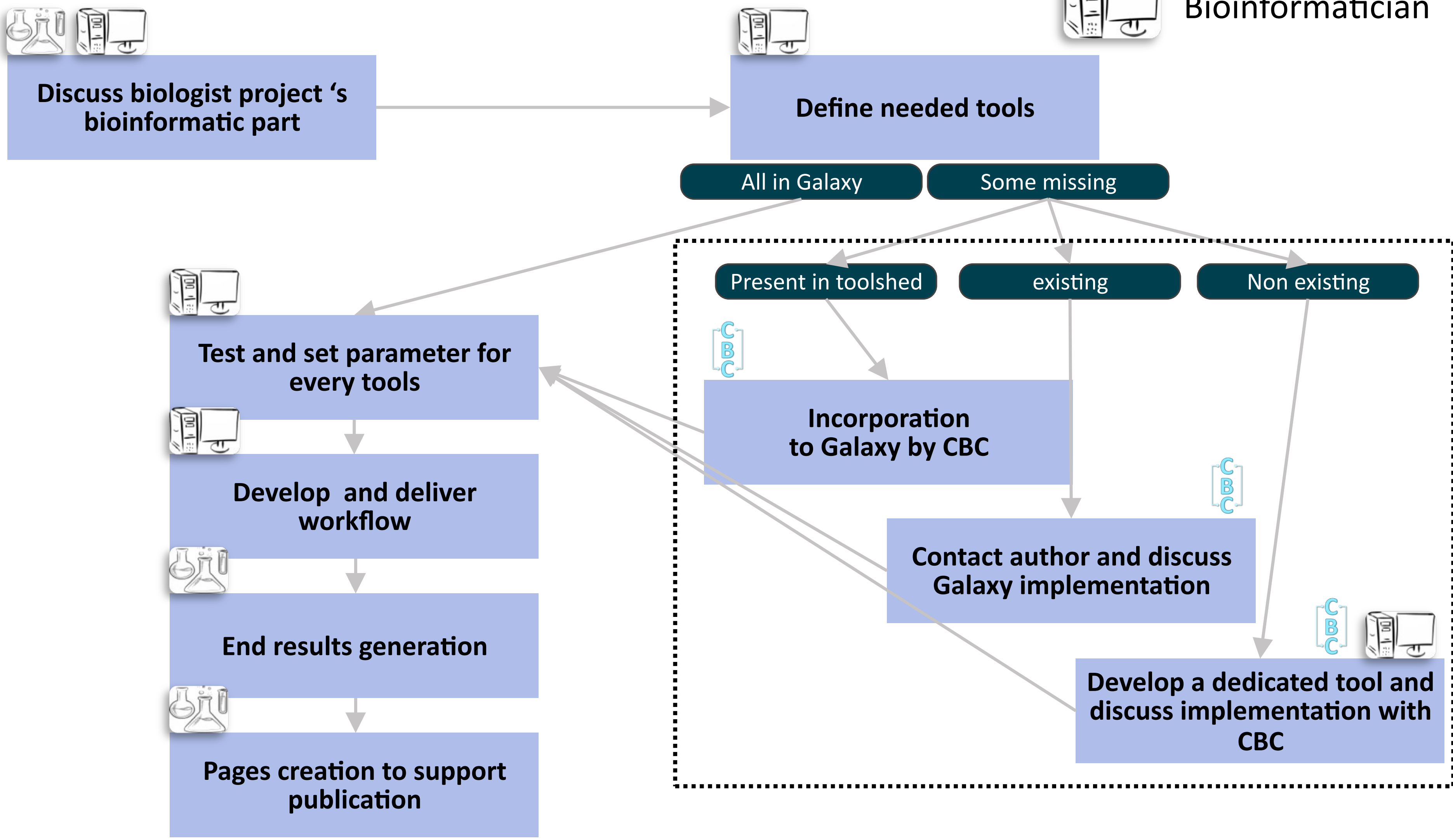
2 – Realistically sized project

Tools presentation - Interface



Tools presentation – Integration process





Tools presentation – Integration process

⟹ Steps for a successful integration

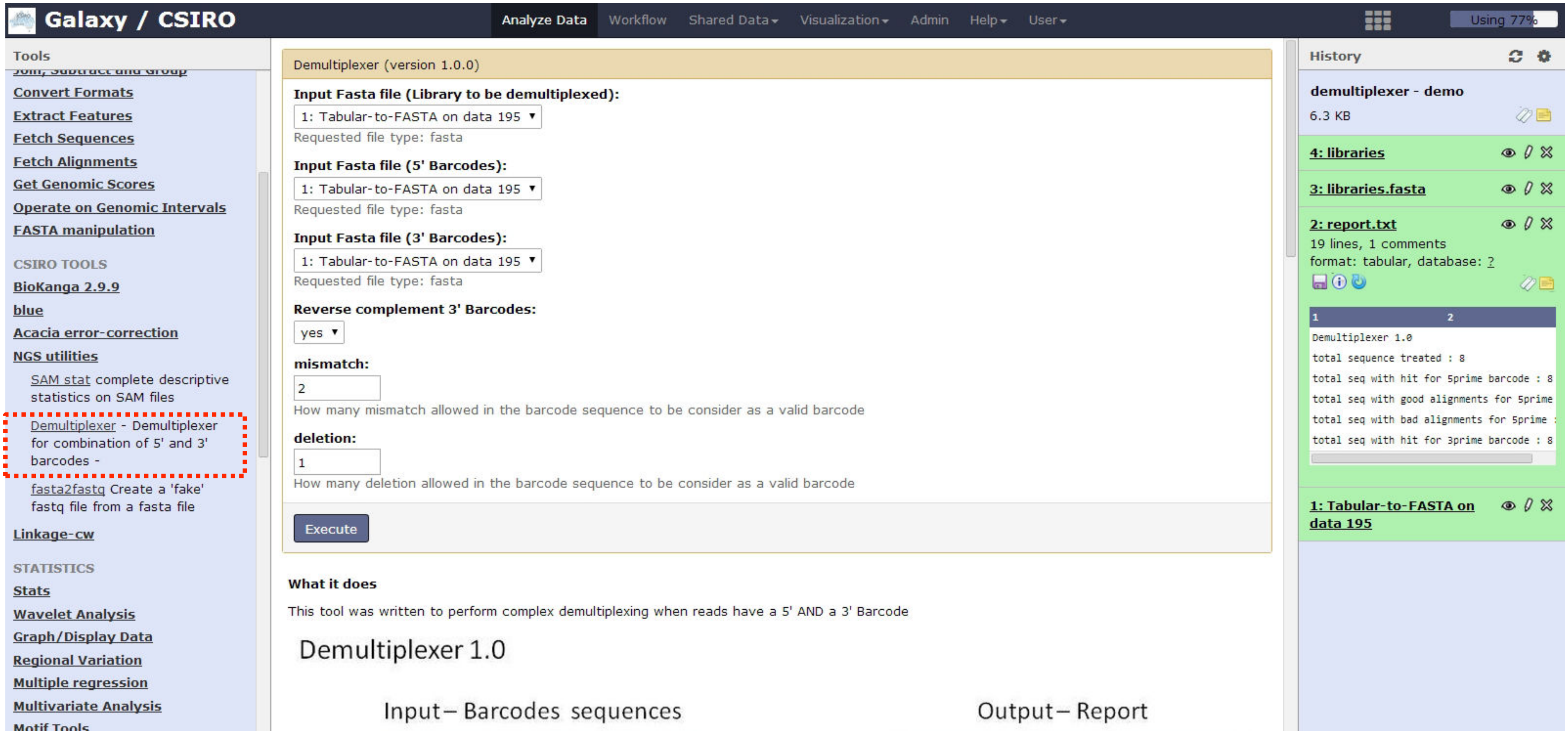
- Evaluation
  - Is the tool fitted for Galaxy ?
  - Is Galaxy fitted for the tool ?

- Coding
  - Definition of available parameters
  - Definition of inputs and outputs
  - Development of the wrapper
  - Testing on the development instance

- Integration to the production Instance
  - Make documentation available
  - Create and publish History/Workflows
  - Create and publish 'Tool Page'

⟹ All theses steps require back and forth communication between the author of the tool and the person integrating it

Tools presentation – NGS utilities - demultiplexer



Tools presentation – NGS utilities - demultiplexer

Galaxy / CSIRO   Analyze Data   Workflow   Shared Data   Visualization   Admin   Help   User

Private Page | demultiplexer

Demultiplexer

What it does

This tool was written to perform complex demultiplexing when reads have a 5' AND a 3' Barcode

Input

Library you want to demultuiplex (fasta format)

sequences of 5' barcode (fasta format)

sequences of 3' barcode (fasta format)

Parameter list

mismatch : number of mismatch allowed in the barcode alignment

deletion : number of deletion allowed in the barcode alignment

Output

There are two outputs.

The first is a text file that reports some figures about the process of barcode alignments.

The second one is a zipped files that contains all the different fasta files that were demultiplexed.

Galaxy History | demultiplexer - demo

Galaxy Workflow | demultiplexer - demo

Tools presentation – Acacia

nature methods
Techniques for life scientists and chemists

nature.com ▸ journal home ▸ archive ▸ issue ▸ correspondence ▸ abstract

NATURE METHODS | CORRESPONDENCE

Fast, accurate error-correction of amplicon pyrosequences using Acacia

Lauren Bragg, Glenn Stone, Michael Imelfort, Philip Hugenholtz & Gene W Tyson

Affiliations | Corresponding author

Nature Methods **9**, 425–426 (2012) | doi:10.1038/nmeth.1990
Published online 27 April 2012

Tools presentation – Blue

Bioinformatics

Blue: correcting sequencing errors using consensus and context

Paul Greenfield[1,2*], Konsta Duesing[3], Alexie Papanicolaou[4] and Denis C. Bauer[1]
[1]CSIRO Computational Informatics, Sydney, Australia.
[2]School of IT, University of Sydney, Sydney, Australia
[3]CSIRO Animal, Food and Health Sciences, Sydney, Australia
[4]CSIRO Ecosystem Sciences, Canberra, Australia
Received on XXXXX; revised on XXXXX; accepted on XXXXX

http://www.bioinformatics.csiro.au/blue/

CSIRO

CSIRO Bioinformatics

Blue

Blue is a fast, accurate short-read error-correction tool based on k-mer consensus and context. It will correct both Illumina and 454-like data, and accepts sequence data files in both FASTQ and FASTA formats. Blue is made available under the General Public License and comes with absolutely no warranty. Blue is written in C# and runs natively on Windows, and with mono on Linux.

An article describing Blue is in preparation.

LINKS

Home
Contact
Publications

SOFTWARE

CSIRO

Biokanga – an integrated toolset for basic bioinformatics NGS tasks
Release 2.95.0

Overview

Biokanga is a suite of tools developed by Stephen Stuart to cope with NGS data analysis.

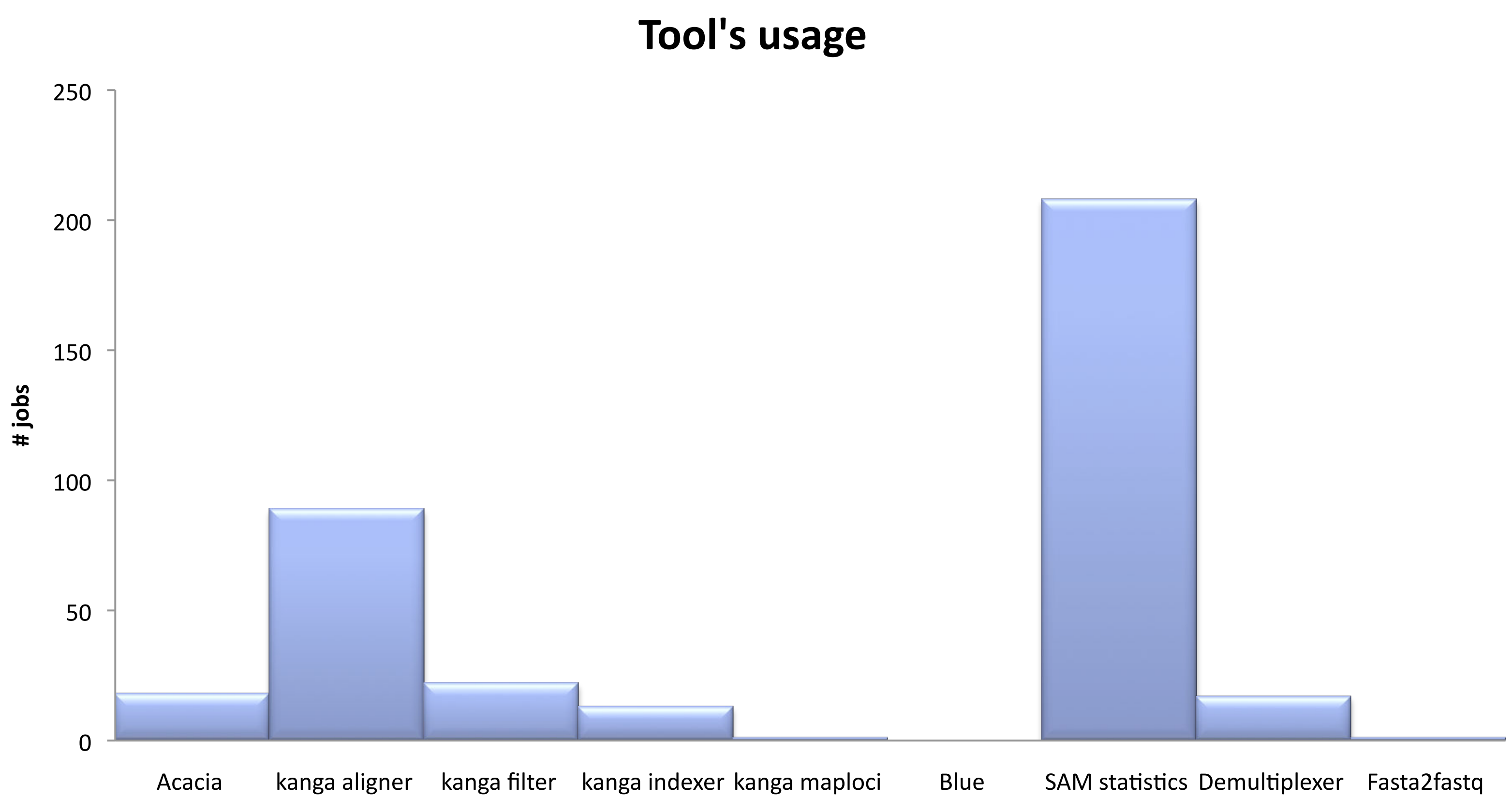
It is composed of several modules that can be arranged in workflow to carry different analysis step :

- index : Generate index over genome assembly or sequences
- aligner : Align NGS reads to indexed genome assembly or sequences
- filter : Filter NGS reads for sequencer errors and/or exact duplicates
- maploci : Map aligned reads loci to known features
- pseudogenome : Concatenate sequences to create pseudo-genome assembly
- rnade : RNA-seq differential expression analyser with optional Pearsons generation
- gene2seq : Generate tab delimited counts file for input to DESeq or EdgeR

Tools presentation - NGS utilities – Dark matter script

⟹ Tool Factory !

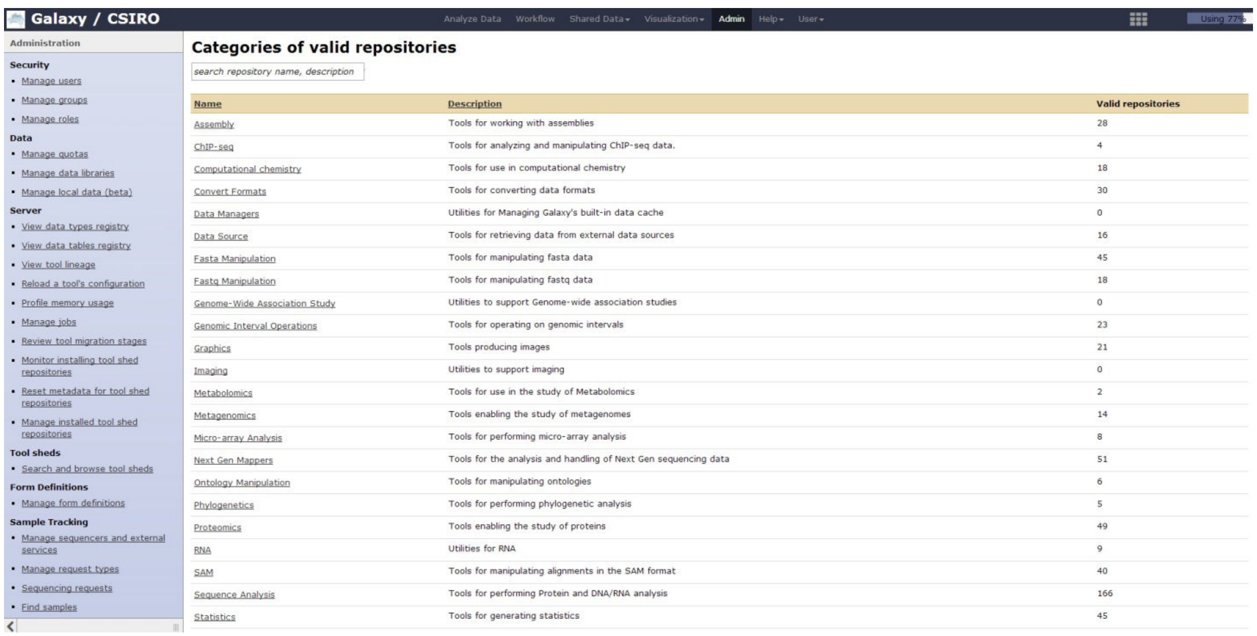
Tools presentation – Usage



Tools integration – Outlooks

Extending the tool availability to the scientific community

- CSIRO local toolshed
- Transfer to the test toolshed
- Transfer to the main toolshed
- Support publication process



Name	Description	Valid repositories
Assembly	Tools for working with assemblies	28
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	4
Computational chemistry	Tools for use in computational chemistry	18
Convert Formats	Tools for converting data formats	30
Data Managers	Utilities for Managing Galaxy's built-in data cache	0
Data Source	Tools for retrieving data from external data sources	16
Fasta Manipulation	Tools for manipulating fasta data	45
Fastq Manipulation	Tools for manipulating fastq data	18
Genome-Wide Association Study	Utilities to support Genome-wide association studies	0
Genomic Interval Operations	Tools for operating on genomic intervals	23
Graphics	Tools producing images	21
Imaging	Utilities to support imaging	0
Metabolomics	Tools for use in the study of Metabolomics	2
Metagenomics	Tools enabling the study of metagenomes	14
Micro-array Analysis	Tools for performing micro-array analysis	8
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	51
Ontology Manipulation	Tools for manipulating ontologies	6
Phylogenetics	Tools for performing phylogenetic analysis	5
Proteomics	Tools enabling the study of proteins	49
RNA	Utilities for RNA	9
SAM	Tools for manipulating alignments in the SAM format	40
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	166
Statistics	Tools for generating statistics	45

Big Data Project

→ What is a big data project ?

- Raw data over 100 Gb to no limit
- Whole project data likely to reach over 1Tb

→ What issues?

- Transfer
- Computational requirement
- Storage

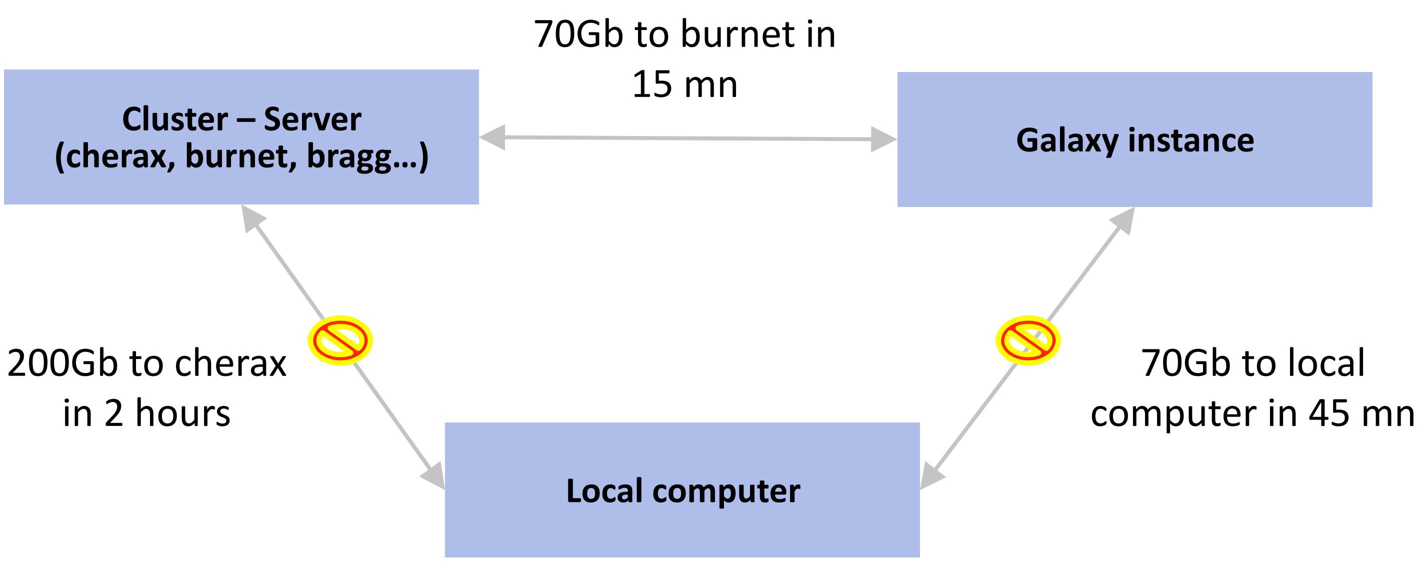
→ Assessing requirements

- End user fluency with Galaxy
- Project feasibility
- Time line

Big Data Project - Transfer

⟹ Data needs to be in and out of Galaxy in reliable and fast manner

- FTP – Filezilla
- File_to_FTP tool by Geert Vandeweyer



Big Data Project – Storage – disk space

⟹ Dealing with group quotas

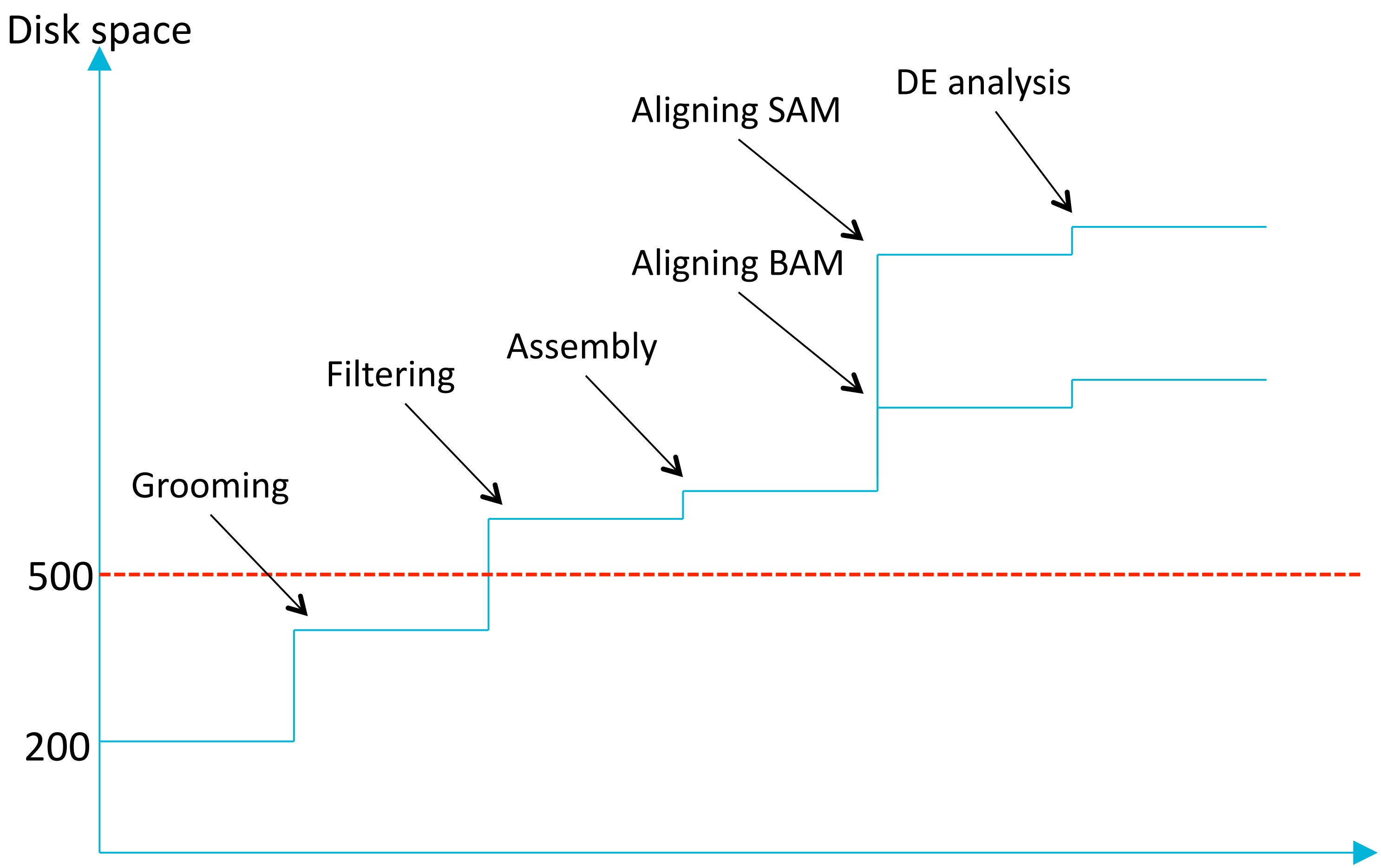
- Default : 200 Gb
- Project : 500 Gb
- Admin : 1000 Gb

⟹ Storage strategy

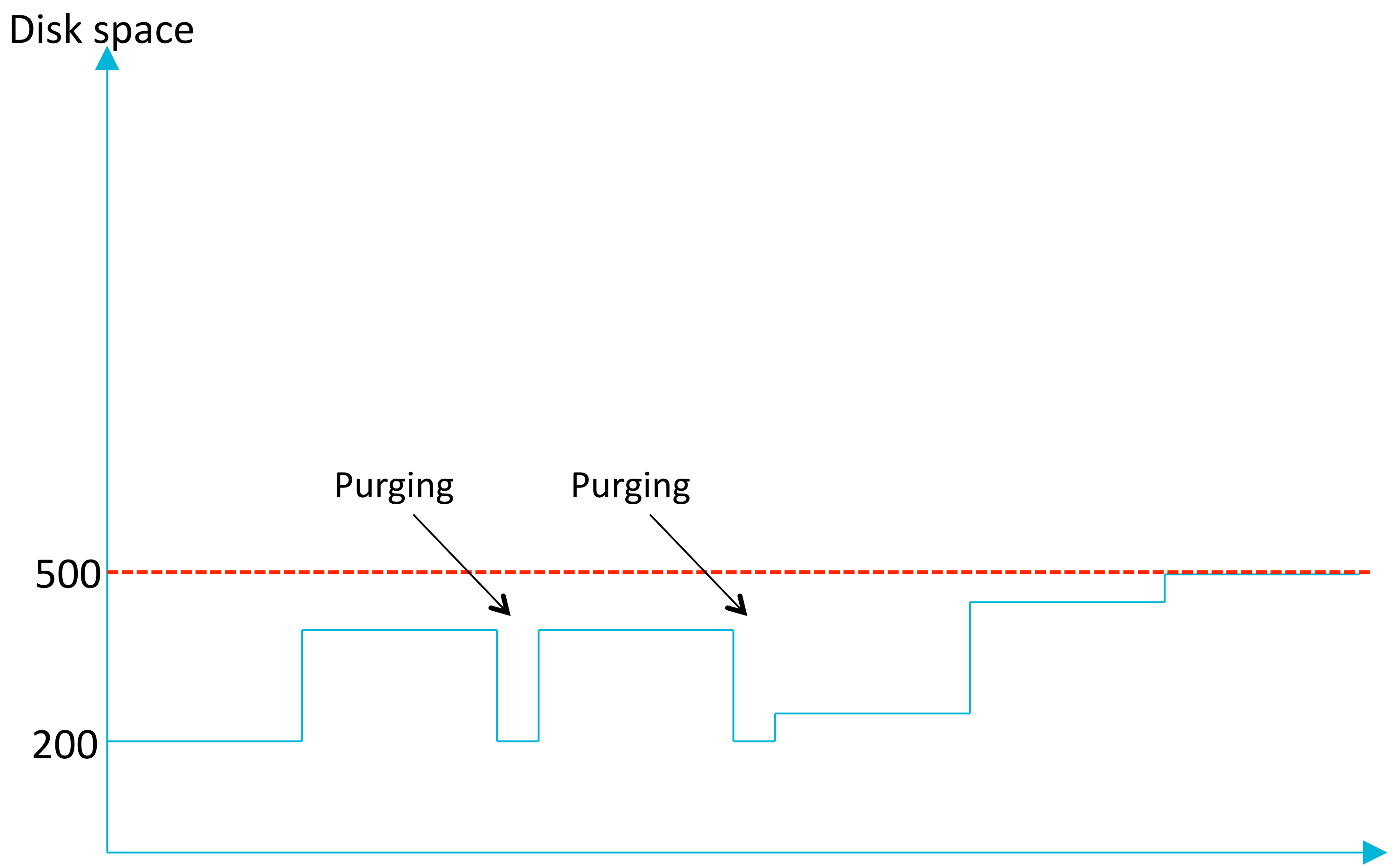
- Galaxy is not meant to be a storage solution
- Raw data and workflows should be sufficient to reproduce the analysis
- ‘Copy datasets’ and purge approach

Big Data Project – Storage – disk space

RNASeq experiment – 900 millions read



Big Data Project – Storage – disk space



Big Data Project – Managing computational needs

- Managing tools
  - Job_conf.xml

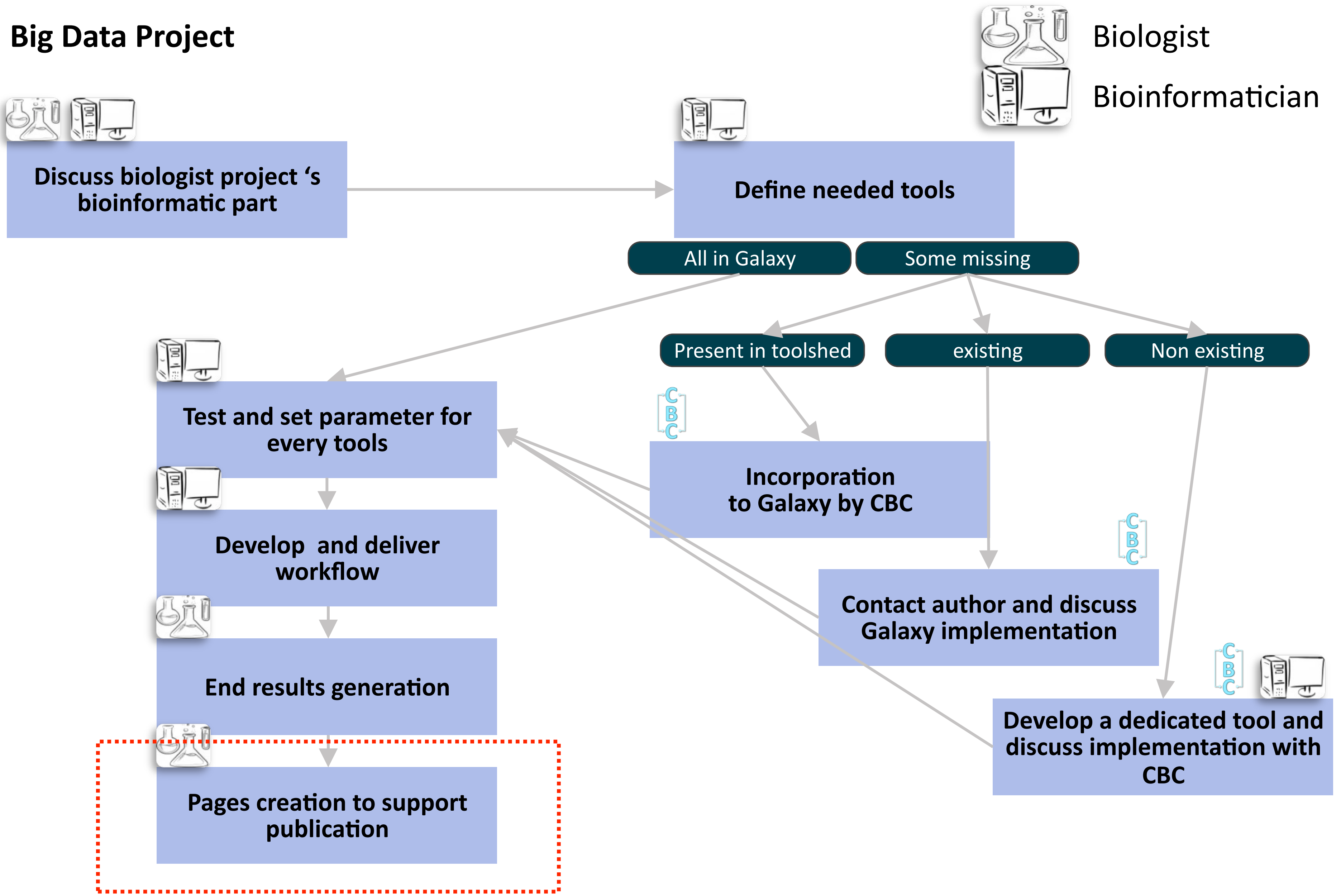
```
<tools>
  <!-- following tools were defined to run local in previous setup.  All other tools will run through PBS by default -->
  <tool id="biomart" destination="local"/>
  <tool id="encode_db1" destination="local"/>
  <tool id="hbvar" destination="local"/>
  <tool id="microbial_import1" destination="local"/>
  <tool id="ucsc_table_direct1" destination="local"/>
  <tool id="ucsc_table_direct_archaea1" destination="local"/>
  <tool id="ucsc_table_direct_test1" destination="local"/>
  <tool id="upload1" destination="local"/>
  <tool id="ncbi_blastn_wrapper" destination="big_jobs"/>

  <!-- tools that are going to bragg can be put here -->
  <tool id="remote_sort_tool" handler="remote" destination="bragg_1cpu_5min"/>
      <tool id="bragg_blastp_cpu" handler="remote" destination="bragg_1cpu_1hr_8gb"/>
      <tool id="bragg_blastp_gpu" handler="remote" destination="bragg_1cpu_1gpu_1hr_8gb"/>
</tools>
```

- Managing user
  - User job limit and walltime
  - No perfect solution

- Dynamic management of execution (from Nate Coraor)

Big Data Project



Big Data Project

Published Pages | mon13m | hookamphipod1

About this Page

454 pyrosequencing-based analysis of gene expression profiles in the amphipod Melita plumulosa: transcriptome assembly and toxicant induced changes

Sharon E. Hook1*, Natalie A. Twine2, Stuart L. Simpson1, David A. Spadaro1, Philippe Moncuquet3, Marc R. Wilkins2

1. CSIRO Land and Water, Locked Bag 2007, Kirrawee, NSW 2232 Australia
2. NSW Systems Biology Initiative, School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia
3. CSIRO Mathematics, Informatics, and Statistics, Acton, ACT, Australia

* Corresponding author

Abstract

Background: Next generation sequencing using Roche's 454 pyrosequencing platform can be used to generate genomic information for non-model organisms, although there are bioinformatic challenges associated with these studies. These challenges are compounded by a lack of a standardized protocol to either assemble data or to evaluate the quality of a de novo transcriptome. This study presents an assembly of the control and toxicant responsive transcriptome of Melita plumulosa, an Australian amphipod commonly used in ecotoxicological studies.

Results: Sequencing generated 1.3 million reads from control, juvenile, metal-exposed and diesel-exposed amphipods. Different read filtering and assembly protocols were evaluated to generate an assembly that i) had an optimal number of contigs; ii) had long contigs; iii) contained a suitable representation of conserved genes; and iv) had long othrolog alignment lengths relative to the length of each contig. A final assembly, generated using fixed-length trimming based on the sequence quality scores, followed by assembly using the MIRA algorithm, produced the best results. The 26,625 contigs generated via this approach were annotated using Blast2GO, and the differential expression between treatments and control was determined by mapping with BWA followed by DESeq. Although the mapping generated low coverage, many differentially expressed contigs, including some with known developmental or toxicological function, were identified.

Conclusions: This study demonstrated that 454 pyrosequencing is an effective means of generating reference transcriptome information for organisms, such as the amphipod Melita plumulosa, that have no genomic information available in databases or in closely related sequenced species. It also demonstrated how optimization of read filtering protocols and assembly approaches changes the utility of results obtained from next generation sequencing studies, and establishes criteria to determine the quality of a de novo assembly in species lacking a reference genome. This new transcriptomic knowledge provides the genomic foundation for the creation of microarray and qPCR assays, serving as a reference transcriptome in future RNAseq studies, and allowing both the biology and ecotoxicology of this organism to be better understood. This approach will allow genomics-based methodology to be applied to a wider range of environmentally relevant species.

Keywords: de novo assembly, transcriptome assembly, RNA Seq, amphipod, toxicogenomics

Datatsets

- Galaxy Dataset | 2.GAC.454Reads.qual
- Galaxy Dataset | 1.GAC.454Reads.qual
- Galaxy Dataset | 2.GAC.454Reads.fna
- Galaxy Dataset | 1.GAC.454Reads.fna

WORKFLOW

- Galaxy Workflow | 454 reads filtering and assembly - amphipod version

Author

mon13m

Related Pages

All published pages
Published pages by mon13m

Rating

Community
(0 ratings, 0.0 average)

Yours

Tags

Community: none

Yours:

Thank you

Bioinformatics Core :

Annette McGrath, Sean Li, Moncuquet Philip, Ondrej Hlinka, Sean McWilliams