Using Galaxy for the analysis of NGS-derived pathogen genomes in clinical microbiology

Anthony Underwood*, Paul-Michael Agapow, Michel Doumith and Jonathan Green.

Bioinformatics Unit, Health Protection Agency, Colindale

The Health Protection Agency

- **The Health Protection Agency's role is to provide an integrated approach to protecting UK public health**
- **The role of the Microbiology Services Division is to provide specialist and reference microbiology to assist with**
  - infectious disease surveillance
  - microbial epidemiology
  - co-ordination of the investigation and cause of national and uncommon outbreaks
- **Equivalent to the CDC in the USA**

The Health Protection Agency: Activities

- **The specialist and reference microbiology activities are comprised of two primary functions**
  - Identification
    - Determining the **species** of an infectious agent
    - Is the microbe responsible for the disease symptoms described for the patient?
  - Typing
    - Determining the **strain** of the an infectious agent
    - Does the microbe have the same type as others seen in an outbreak or that seen in environmental or food samples

The changing face of microbiology

- **Public health microbiology was based for a long time on phenotypic testing**
  - Selective growth media
  - Colony morphology
  - Gram straining and cell morphology
  - Serotyping
  - Biochemical tests
- **Over the last 2 decades some of the functions have become replaced with molecular tests**
  - Identification
    - 16S rRNA gene sequencing
    - Other genes for difficult groups such as *Bacillus* species

The changing face of microbiology 2

- **Typing microbes has seen the biggest revolution with many molecular tests now commonly used**

  - **Multi Locus Sequence Typing** (MLST)
    Sequencing of 7 house keeping genes resulting in an allelic profile where a single base change results in a new allele

Locus/ ST	*adk*	*fumC*	*gyrB*	*icd*	*mdh*	*purA*	*recA*
10	10	11	4	8	8	8	2

  - Some bacteria require **additional loci** to provide sufficient discrimination

    - For example *por*A and *fetB* sequencing in Neisseria

The changing face of microbiology 3

- **Other molecular typing techniques**
  - Some organisms are typed using the sequence from a **single gene**
    - For example sequencing of the *emm* gene that codes for the M protein can replace the Lancefield serotyping scheme for Group A Streptococci
  - **Drug resistance** determination
    - e.g mutations in *rpo*B and *gyr*A causing resistance to rifampicin and fluoroquinolones respectively in *Mycobacterium tuberculosis*
  - Multi Locus **VNTR** Analysis (MLVA)
    - The copy number at several repeat loci are concatenated to produce a digital barcode/profile e.g 2-5-4-2-1

      These profiles are compared to identify types

Next Generation Sequencing and Microbiology

- **Next Generation sequencing may change the way we do pubic health microbiology**
  - The average microbial genome is relatively small
  - By multiplexing samples using molecular tags and the amount of data generated by the Illumina HiSeq machines high coverage paired end data can be generated for £100 (€115)
  - This will probably fall to approx £40 (€45) by end of 2011
  - These prices are close to or cheaper than that required for MLST

Next Generation Sequencing: The future?

Looking ahead it is not too crazy to suggest that every pathogen isolated from a patient will have its entire genome sequenced

Next Generation Sequencing and Microbiology 2

- **There is already the potential to genome sequence an infectious agent and perform ‘typing +’**
  - The MLST type can be determined
  - But so can the presence/sequence of other genes
    - Virulence gene profiles
    - Resistance genes
    - Point mutations in genes involved in the infectious process
    - Any other gene that at a later time may be of interest – great for retrospective studies

Next Generation Sequencing and Microbiology 3

- The current ‘Next Generation’ technologies have limitations for real time results since library prep and sequencing times take days/weeks
- New technologies such as Ion Torrent or new machines such as the MiSeq promise much faster sequence delivery in under 24 hours
- For the moment the utility of NGS is confined to medium term projects
- However it is capturing the imagination of public health microbiologists
- The problem is in the analysis

Next Generation Sequencing Analysis

Health Protection Agency

- Over 50 NGS projects underway
- Very few bioinformaticians attached to projects
- The burden of analysis falls on a core team of 3 or 4 bioinformaticians

Galaxy and microbial genome analysis

Health Protection Agency

- Enter Galaxy

- **Assessment of Galaxy led us to believe that it might provide a solution and kill 2 birds with 1 stone**
  - Provide a means for laboratory scientists with little/no command line or bioinformatics analysis to analyse NGS data
  - Relieve the burden on bioinformaticians of having to perform processing steps enabling them to concentrate on more complex downstream comparative analyses

Galaxy use within the HPA
Warning

Galaxy and microbial genome analysis 2

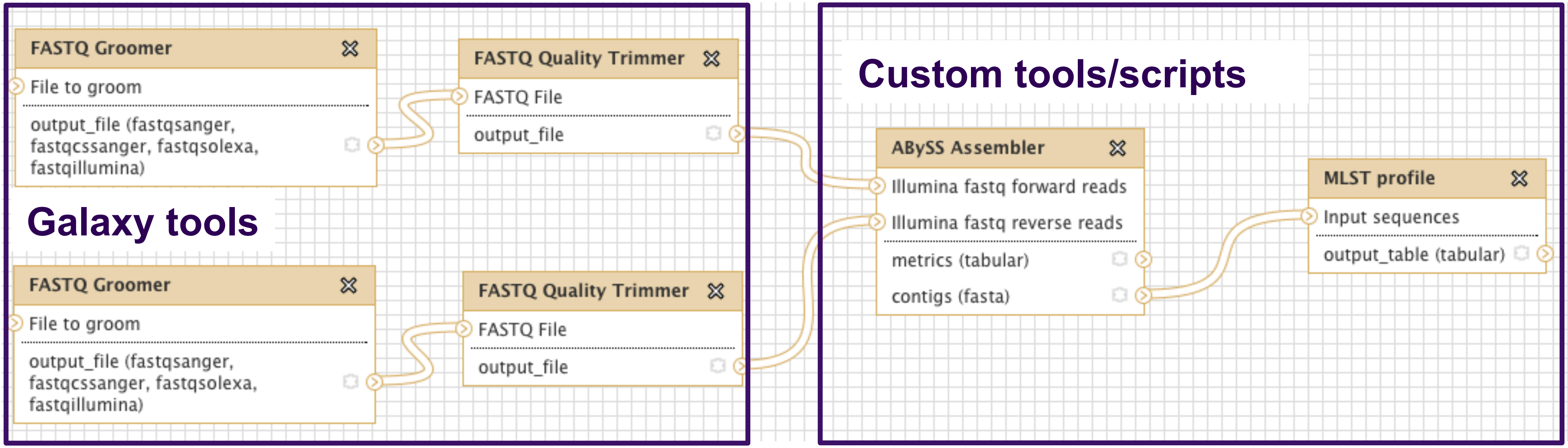
- **What kind of simple analyses might clinical microbiologists want to perform?**
  - QC assessment of samples before further processing
  - Mapping of reads to a reference
    - SNP calling and filtering of ‘interesting SNPs’
  - *De novo* assembly with QC ‘gateways’
    - Assigning MLST type
    - Determine genotype e.g *emm* type
    - Produce virulence profile

Galaxy
MLST determination

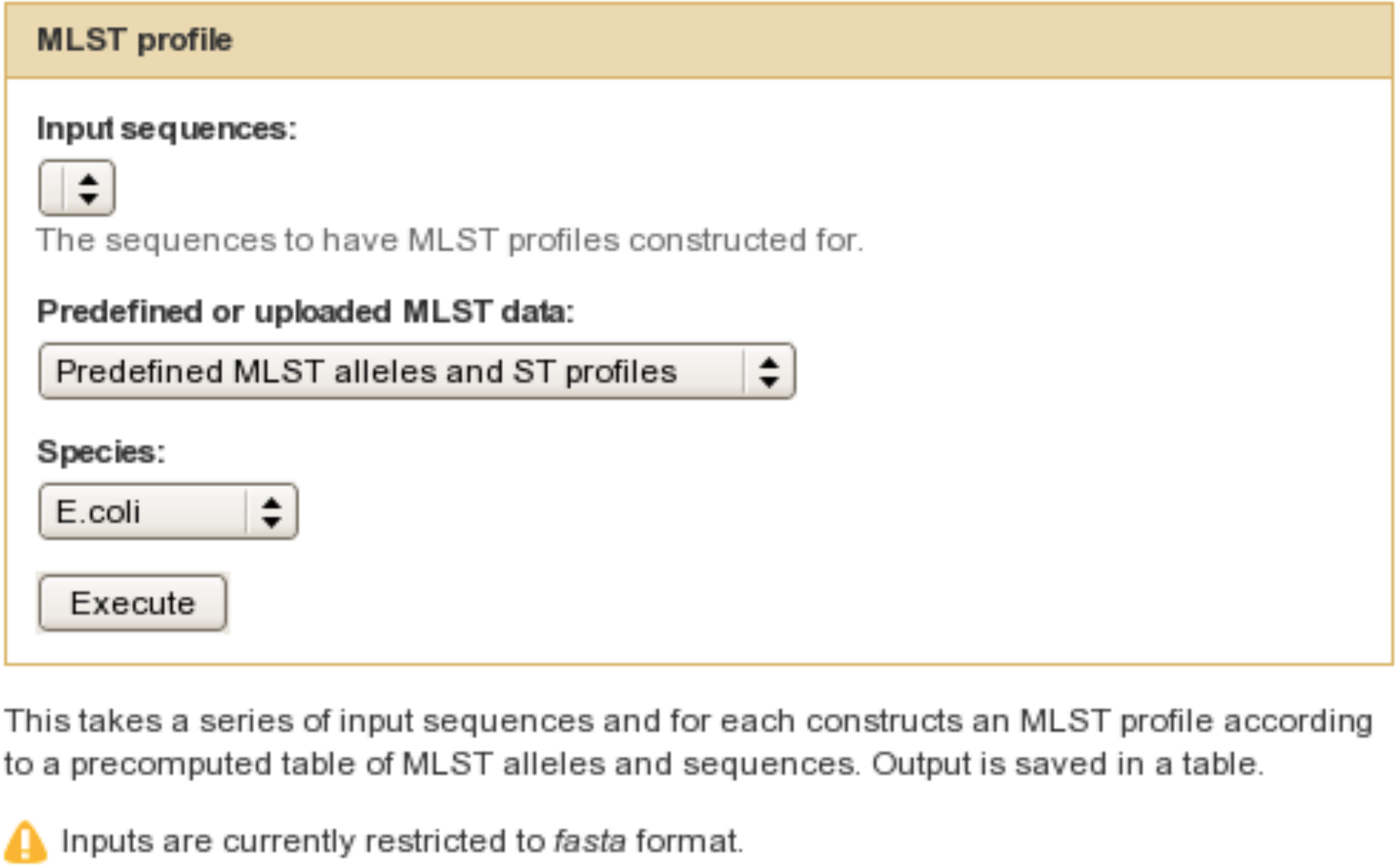
- Scripts already existed within our group that could extract MLST and virulence profiles
- Scripts written within the group are in a range of languages – python, ruby, perl, C++
- The ability to use existing Galaxy NGS tools in combination with ‘in house’ scripts provided the flexibility to deliver bespoke solutions
- The fact that Galaxy is language agnostic makes it an appealing solution to our polyglot group

Galaxy
MLST pipeline

- **Galaxy tools: FASTQ Groomer → Trimmer**
- **Custom scripts: ABySS assembly → MLST profile**
- **MLST profile**
  - Make a blast database from *de novo* assembly contigs
  - Extract sequence of 7 loci by blast from contigs
  - Compare each locus sequence with MLST database to discover an exact match (existing allele) or inexact match (new allele)



Galaxy
MLST input



MLST profile

Input sequences:

The sequences to have MLST profiles constructed for.

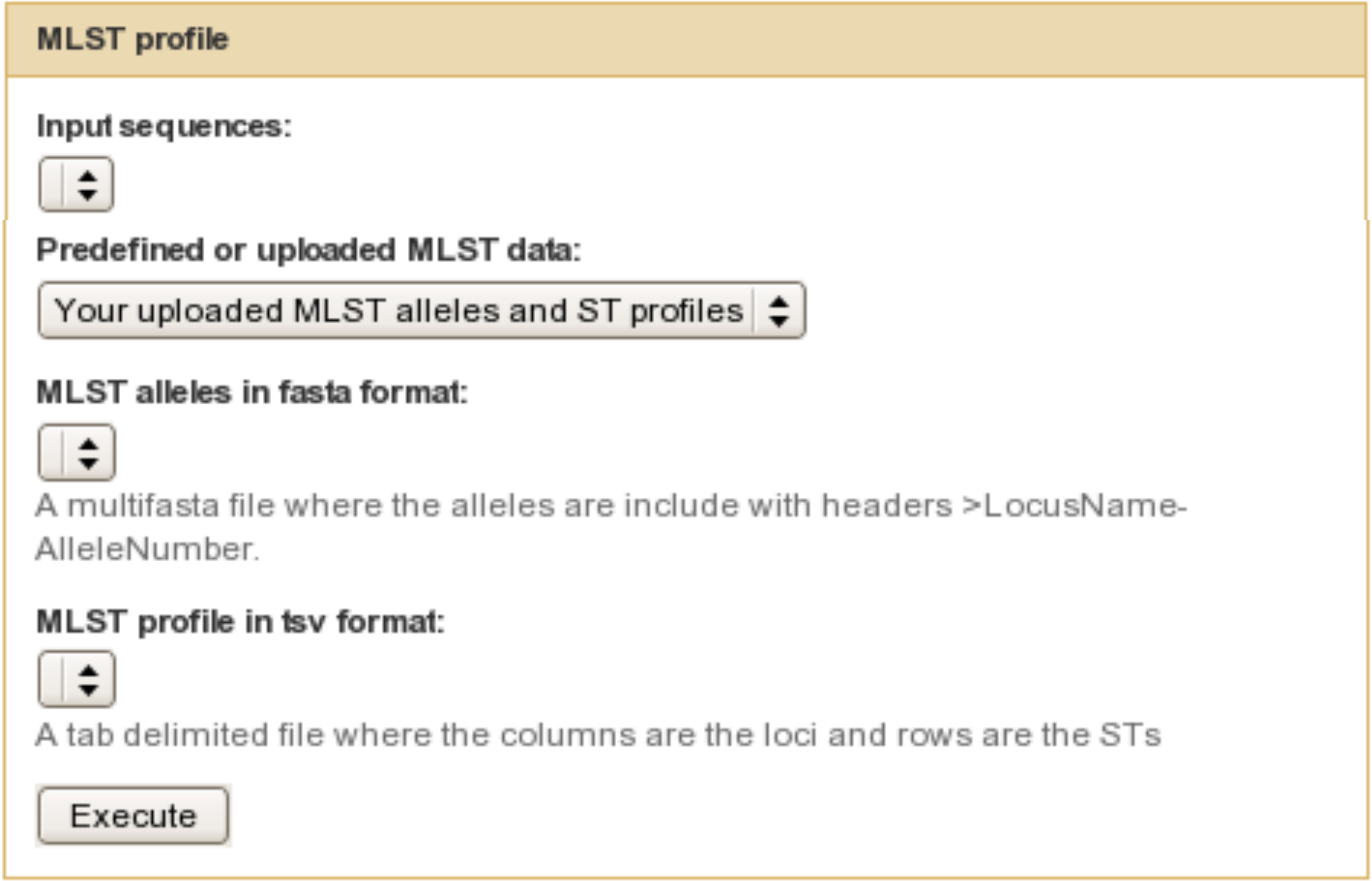
Predefined or uploaded MLST data: Predefined MLST alleles and ST profiles

Species: E.coli

This takes a series of input sequences and for each constructs an MLST profile according to a precomputed table of MLST alleles and sequences. Output is saved in a table.

⚠ Inputs are currently restricted to *fasta* format.

Galaxy
MLST input



MLST profile

Input sequences:

Predefined or uploaded MLST data:

Your uploaded MLST alleles and ST profiles

MLST alleles in fasta format:

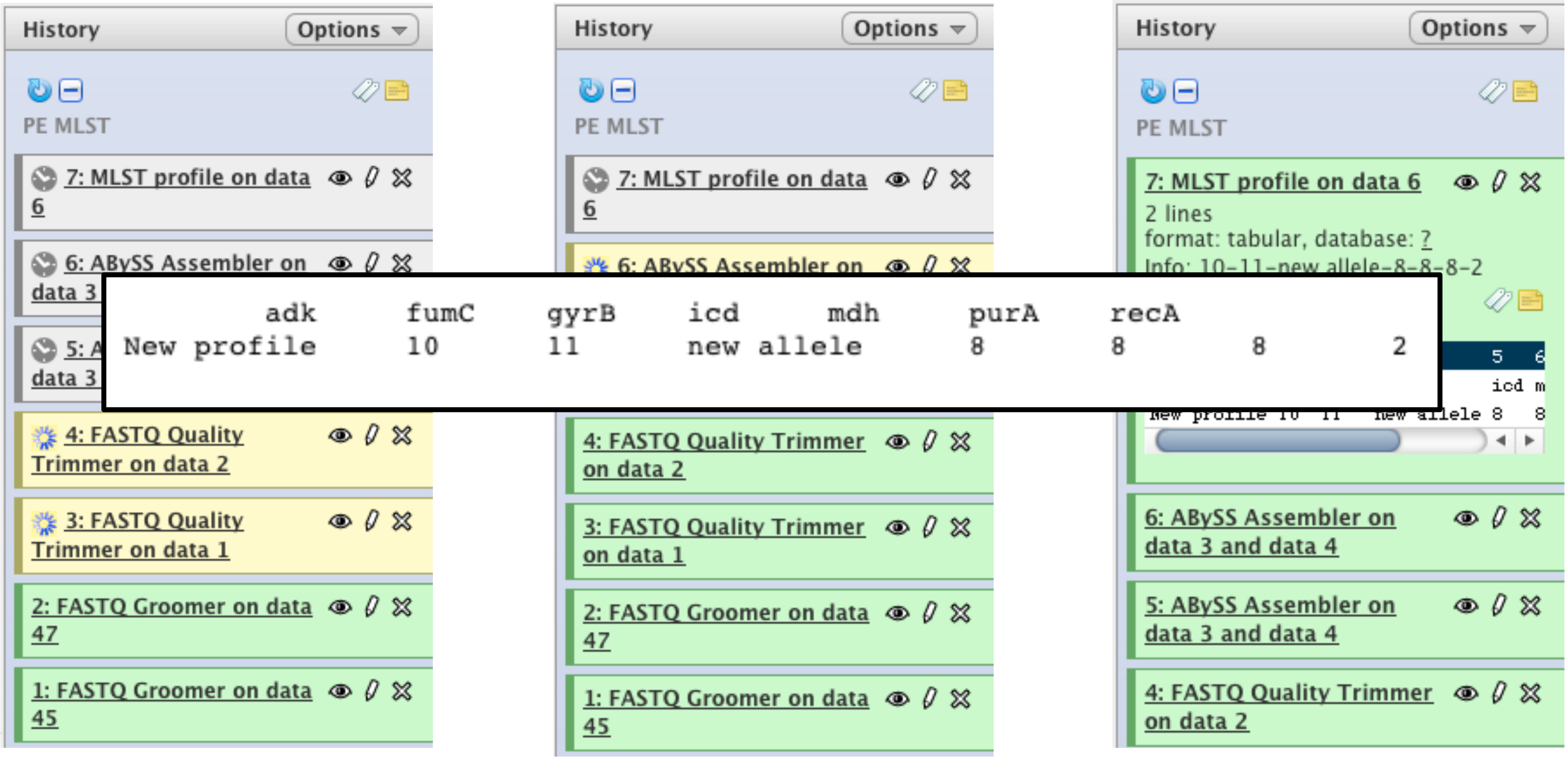
A multifasta file where the alleles are include with headers >LocusName-AlleleNumber.

MLST profile in tsv format:

A tab delimited file where the columns are the loci and rows are the STs

Galaxy
MLST results

- **A paired end data set consisting of 14 million reads took 1 hour to convert, trim, assemble and call the MLST profile. Hands on time 1 minute!**



7: MLST profile on data 6

6: ABySS Assembler on data 3

5: A data 3

4: FASTQ Quality Trimmer on data 2

3: FASTQ Quality Trimmer on data 1

2: FASTQ Groomer on data 47

1: FASTQ Groomer on data 45

7: MLST profile on data 6

6: ABySS Assembler on

4: FASTQ Quality Trimmer on data 2

3: FASTQ Quality Trimmer on data 1

2: FASTQ Groomer on data 47

1: FASTQ Groomer on data 45

7: MLST profile on data 6
2 lines
format: tabular, database: ?
Info: 10-11-new allele-8-8-8-2

```
5   6
        icd  m
New profile 10  11   new allele 8   8
```

6: ABySS Assembler on data 3 and data 4

5: ABySS Assembler on data 3 and data 4

4: FASTQ Quality Trimmer on data 2

```
            adk        fumC       gyrB       icd        mdh        purA       recA
New profile            10         11         new allele            8          8          8          2
```

Galaxy
Typing by reference genes

We have:

- A set of reads from an unknown (untyped) microbe(s)
- Already characterised sets of reference (usually virulence) genes
- Typing scheme(s) based on the presence and absence of given reference genes

We want to know:

- Whether any genes of interest are present
- Based on presence/absence what types are present

Galaxy
Generic genotyping

A simple, generic, extensible, updatable approach:

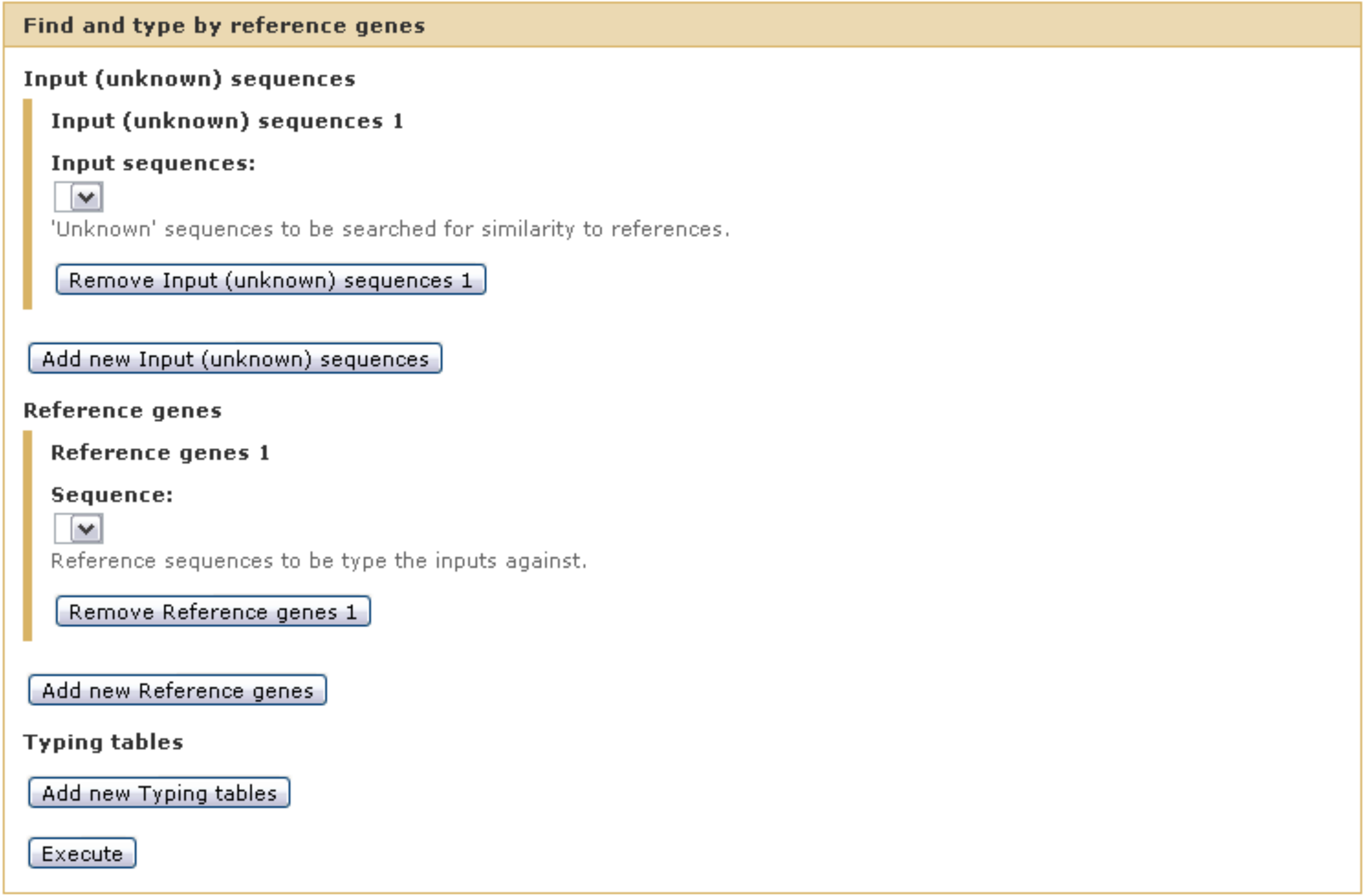
- Inputs microbial genomes are just Fasta files
- References, likewise
- Typing schemes are just a table

The script builds a database from the inputs, blast the references against it, and looks up the results in the typing scheme table

```
>aidA
ATGAATAAGGCCTACAG
TATCATATGGAGCCACT
CCAGACAGGCCTGGAT
TGTGGCCTCAGAGTTA
GCCAGAGGACATGGTT
TTGTCCTTGCAAAAAT
ACACTGCTGGTATTGGC
GGTTGTTTCCACAATC
```

```
   , chuA, yja2, TSPE4_C2
B2, +,     +,     ~
D,  +,     -,     ~
B1, -,     ~,     +
A,  -,     ~,     -
```

Galaxy
Generic genotyping 2



Find and type by reference genes

Input (unknown) sequences

Input (unknown) sequences 1

Input sequences:

'Unknown' sequences to be searched for similarity to references.

Remove Input (unknown) sequences 1
Add new Input (unknown) sequences

Reference genes

Reference genes 1

Sequence:

Reference sequences to be type the inputs against.

Typing tables

Galaxy
Generic genotyping 3

Make a virtue of laziness

- Use standard, simple types
- User can select as many input, references and typing tables as needed
- Use metadata of Fasta headers to usefully label output
- Output is saved as YAML

```
----
Datetime:
2011-05-23T16:16:20+01:00
Hits:
  -
    Name: unknown-12
      -
        Name: aah et al.
        Matches:
          Full: [aah]
          Partial: []
        Phylo_matches: [B2]
      -
        Name: aidA and iroN
        Matches:
          Full: []
          Partial: [iroN, ompT]
        Phylo_matches: [D1, B2]
```

Galaxy
Galgen

Being even more virtuous …

- There’s a lot of repetition in Galaxy tool construction
- Can we save effort in making a new tool?
- Can we prevent errors by automating tool generation?

Yes …

Label-seqs-by-data.rb –in-table epidates.csv uk.fasta

To

label-seqs-by-date tool dir, template and conf entry

Galaxy
Galgen 2

Galgen:

- Sniffs a command-line and infers tool and executable name, options, input datasets and outputs, etc.
- Checks these with the user
- Generates necessary basic tool config and template files
- Uses hints on command-line (bracket options, file extensions, etc.)

Label-seqs-by-data.rb (–in-table epidates.csv) input_uk.fasta

Can’t guess everything, but aim for all simple cases and provide skeleton for more complex.

Coming … “soon” (a month)

Galaxy
Future Direction

- To process genomes and call SNPs
- To filter SNPs for those in genes of interest
- To report SNPs that may result in drug resistance

- To develop a generic genotyper that can extract the sequence used in genotyping from a draft genome and call the type

- For longer read (454) data report copy number for repeats that have a short enough repeat length

Galaxy
Additional functionality

- **Tasks we need to complete**
    - With 50 projects anticipated we need to find an efficient way of storing and organising data using Galaxy datasets
    - To fully integrate the Galaxy instance with our Condor cluster to be able to perform jobs more efficiently in parallel
- **Desirables**
    - To process multiple samples with one workflow and organise the final results that makes it easy to link samples to results
    - To organise data sources so scientists can easily select which of 100s of samples to process
    - To organise results so scientists other than those performing the analyses can quickly navigate and view them.

Acknowledgments

- **Galaxy Team**
- **Laboratory Scientists from Health Protection Agency**
    - ARMRL
    - LHI
    - APU