ngasp

Pipeline Development

Research Program: ***Plant and Animal Genomics***
Research Group: ***Statistical and Population Genomics***

Software Engineer: J. Jené

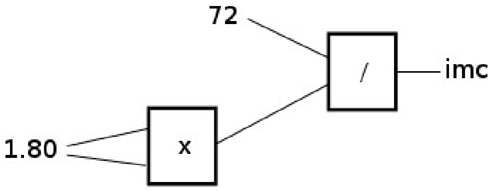
Principal Investigator: S. Ramos
Technical Supervisor: G. Vera

Workshop – Session 1 - June 8, 2017

Experiment vs Pipeline

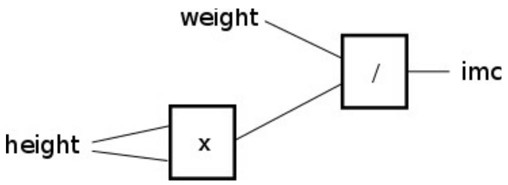
- **Experiment:** It is a workflow with real input / output data.

  Experiment Example:

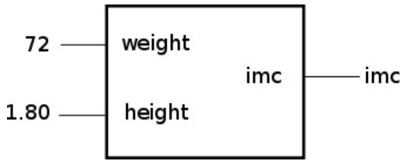


- **Pipeline:** It is a generic workflow. It requires an experiment that executes it.

  Pipeline Example:



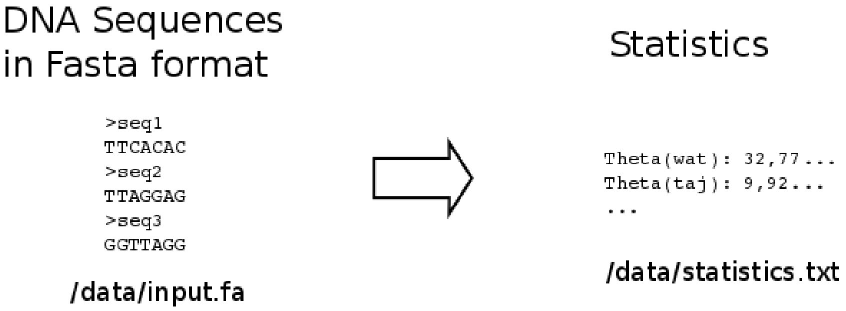
Experiment Example that calls the previous pipeline:



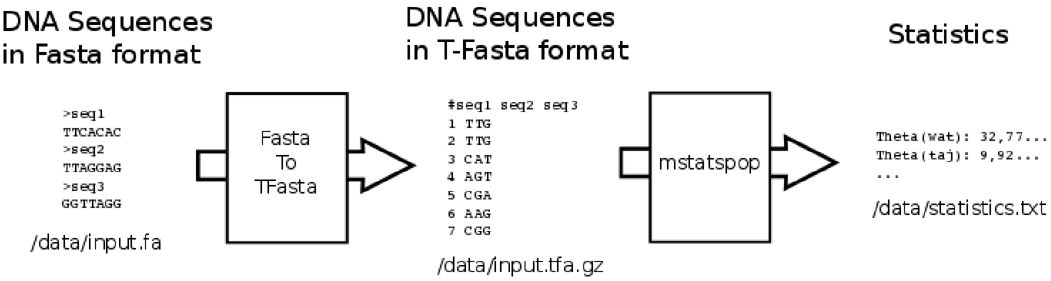
Note: Another difference between *experiments* and *pipelines* is that pipelines can be looped, experiments cannot.

1) Creation of an Experiment

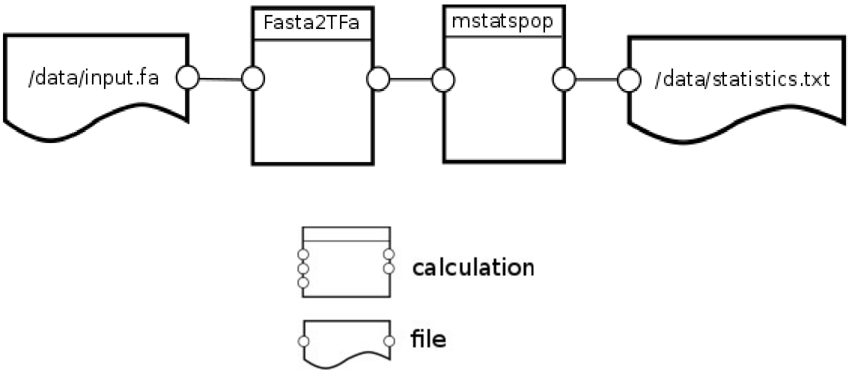
- Objective: Get statistics from DNA sequences in fasta format.

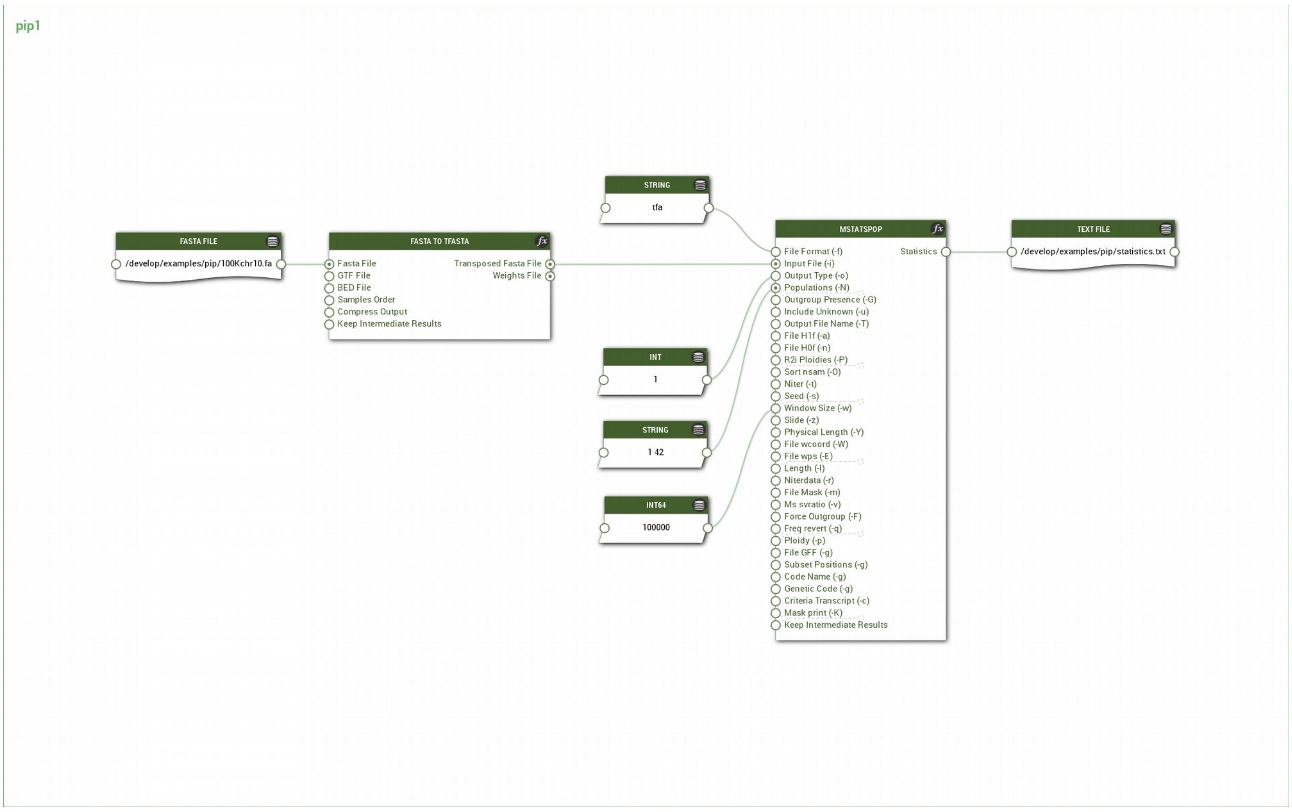


- Draft Workflow:



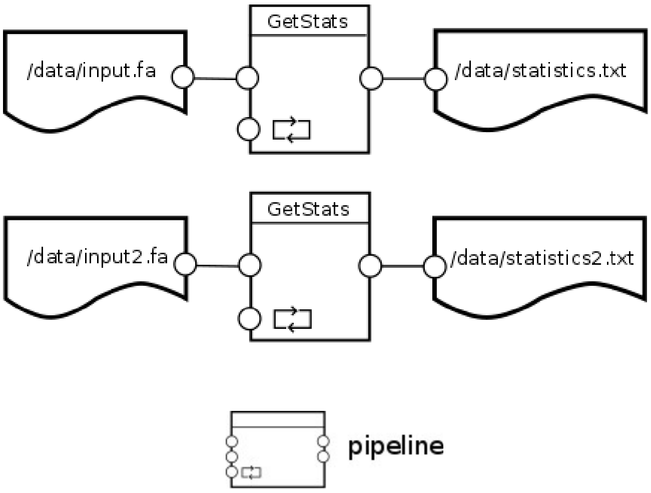
- ngasp Experiment:



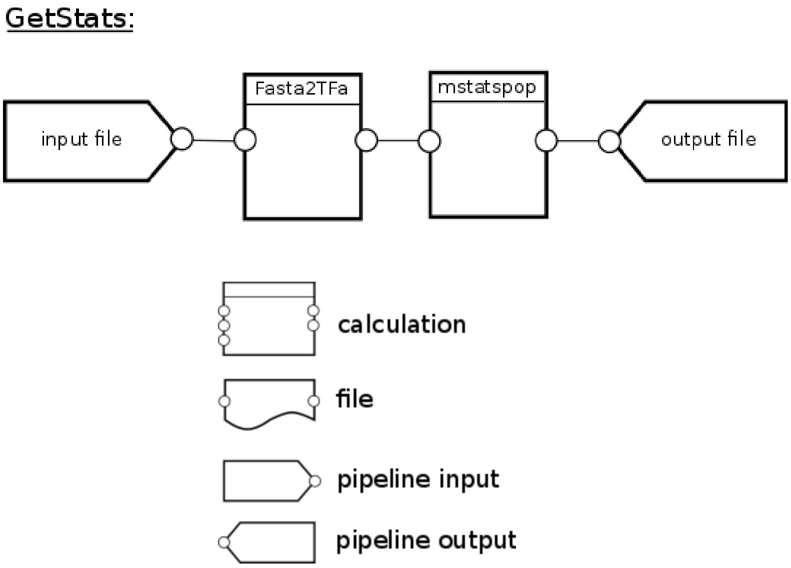

pip1
FASTA FILE
/develop/examples/pip/100Kchr10.fa
FASTA TO TFASTA
Fasta File
GTF File
BED File
Samples Order
Compress Output
Keep Intermediate Results
Transposed Fasta File
Weights File
STRING
tfa
INT
1
STRING
1 42
INT64
100000
MSTATSPOP
File Format (-f)
Input File (-i)
Output Type (-o)
Populations (-N)
Outgroup Presence (-G)
Include Unknown (-u)
Output File Name (-T)
File H1f (-a)
File H0f (-n)
R2i Ploidies (-P)
Sort nsam (-O)
Niter (-t)
Seed (-s)
Window Size (-w)
Slide (-z)
Physical Length (-Y)
File wcoord (-W)
File wps (-E)
Length (-l)
Niterdata (-r)
File Mask (-m)
Ms svratio (-v)
Force Outgroup (-F)
Freq revert (-q)
Ploidy (-p)
File GFF (-g)
Subset Positions (-g)
Code Name (-g)
Genetic Code (-g)
Criteria Transcript (-c)
Mask print (-K)
Keep Intermediate Results
Statistics
TEXT FILE
/develop/examples/pip/statistics.txt

2) Creation of a Generic and Reusable Pipeline

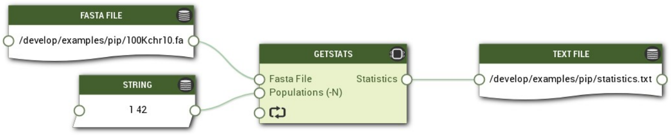
- Desired New Pipeline Usage:



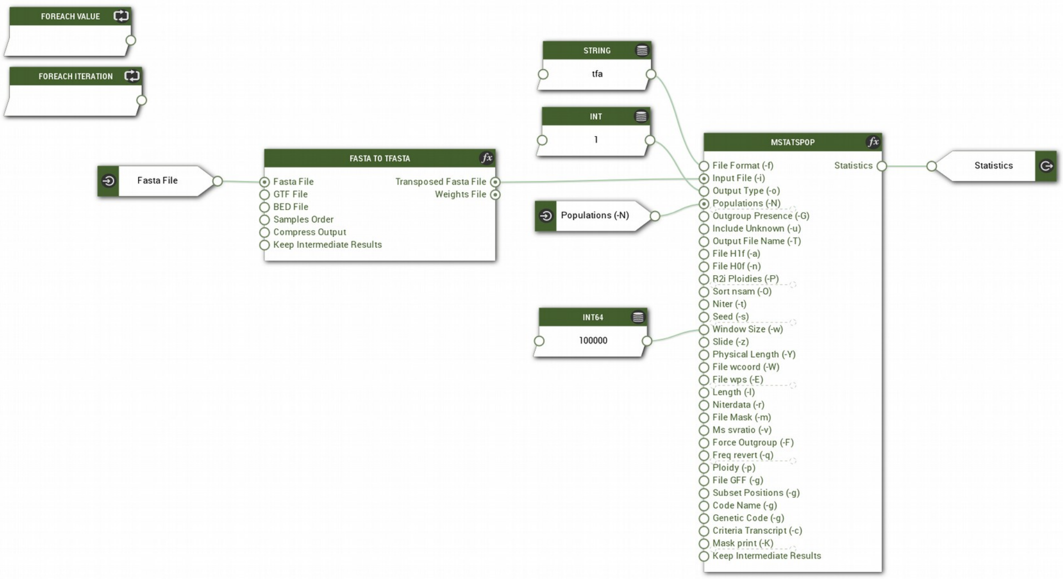
- Create a Pipeline replacing input / otput data by pipeline inputs and outputs:



pip2A

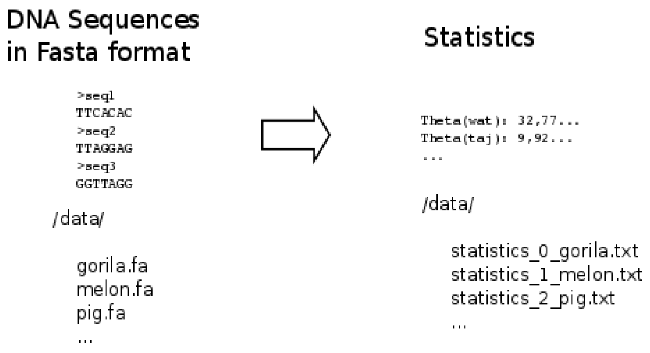


GetStats

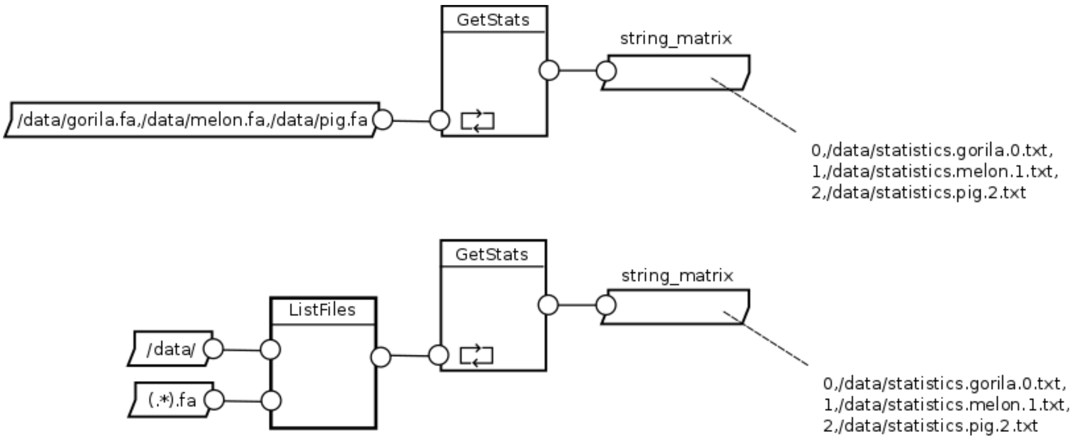


3) Creation of a Generic and Reusable Pipeline for Multiple Input and output Files

- Objective:

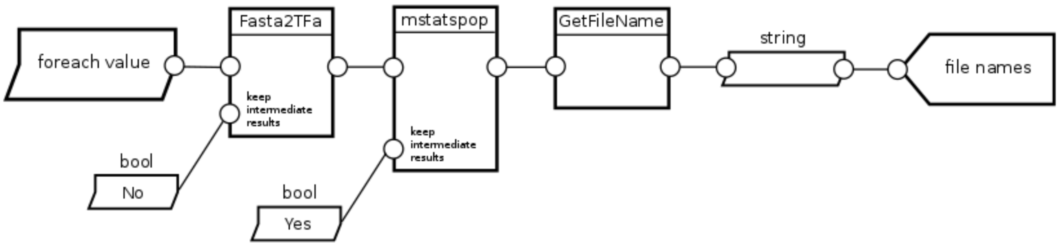


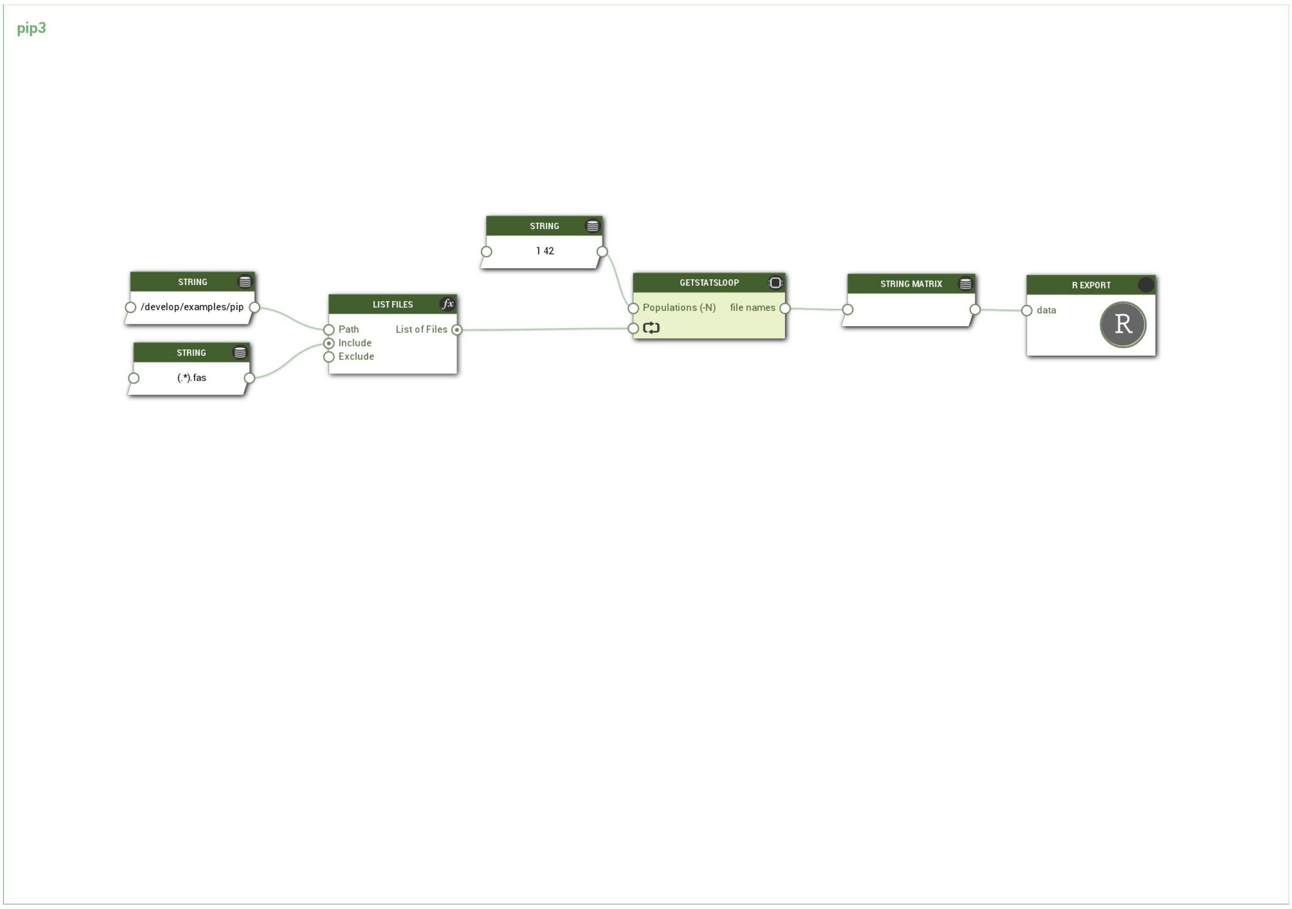
- Pipeline Desired Usage:

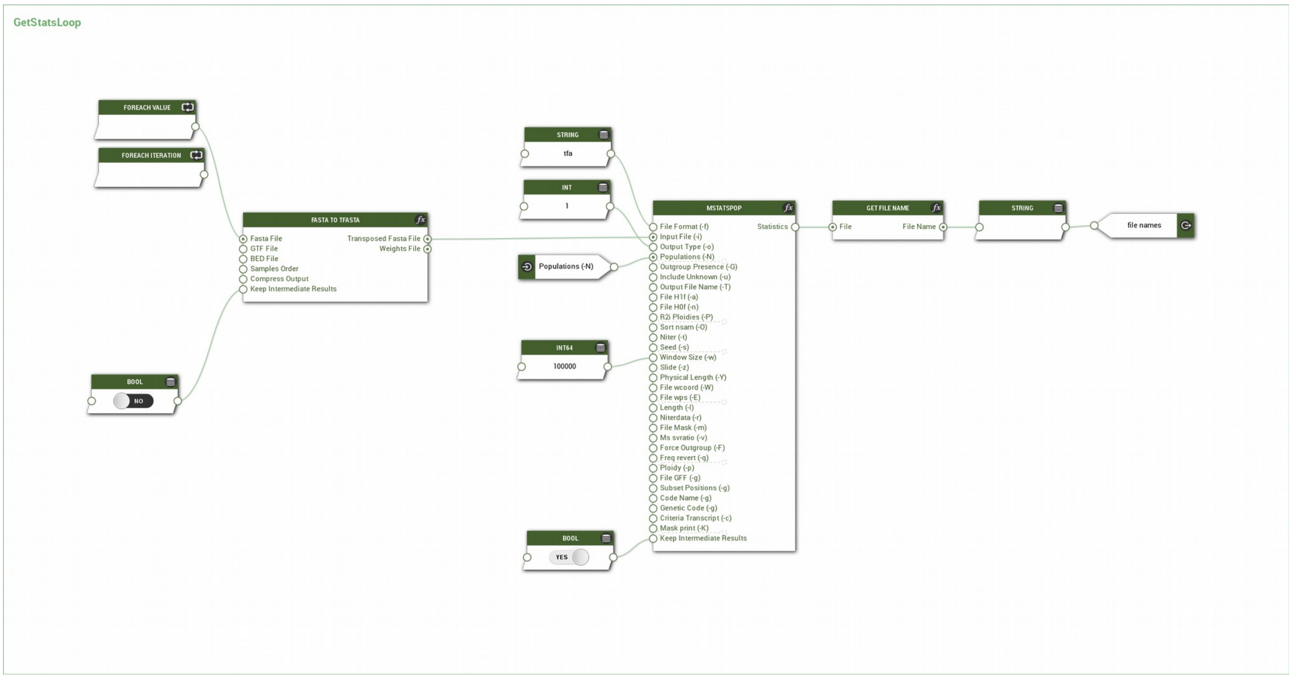

GetStats
string_matrix
/data/gorila.fa,/data/melon.fa,/data/pig.fa
0,/data/statistics.gorila.0.txt,
1,/data/statistics.melon.1.txt,
2,/data/statistics.pig.2.txt
GetStats
string_matrix
ListFiles
/data/
(.*).fa
0,/data/statistics.gorila.0.txt,
1,/data/statistics.melon.1.txt,
2,/data/statistics.pig.2.txt

- Pipeline Design:

GetStats:

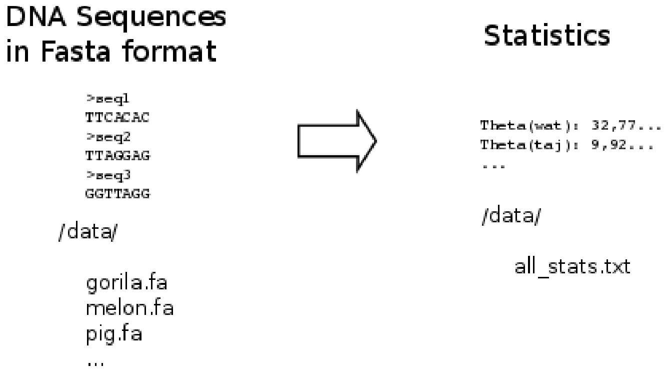



pip3
STRING
/develop/examples/pip
STRING
(.*).fas
LIST FILES
Path
Include
Exclude
List of Files
STRING
1 42
GETSTATSLOOP
Populations (-N)
file names
STRING MATRIX
R EXPORT
data
R

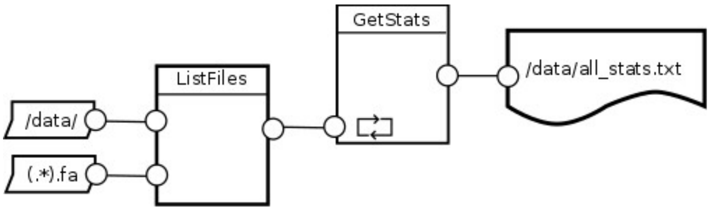

GetStatsLoop
FOREACH VALUE
FOREACH ITERATION
FASTA TO TFASTA
Fasta File
GTF File
BED File
Samples Order
Compress Output
Keep Intermediate Results
Transposed Fasta File
Weights File
BOOL
NO
STRING
tfa
INT
1
Populations (-N)
INT64
100000
BOOL
YES
MSTATSPOP
File Format (-f)
Input File (-i)
Output Type (-o)
Populations (-N)
Outgroup Presence (-G)
Include Unknown (-u)
Output File Name (-T)
File H1f (-a)
File H0f (-n)
R2i Ploidies (-P)
Sort nsam (-O)
Niter (-t)
Seed (-s)
Window Size (-w)
Slide (-z)
Physical Length (-Y)
File wcoord (-W)
File wps (-E)
Length (-l)
Niterdata (-r)
File Mask (-m)
Ms svratio (-v)
Force Outgroup (-F)
Freq revert (-q)
Ploidy (-p)
File GFF (-g)
Subset Positions (-g)
Code Name (-g)
Genetic Code (-g)
Criteria Transcript (-c)
Mask print (-K)
Keep Intermediate Results
Statistics
GET FILE NAME
File
File Name
STRING
file names

4) Creation a Generic and Reusable Pipeline for Multiple Input Files that Generate Only One Output File

- Objective:

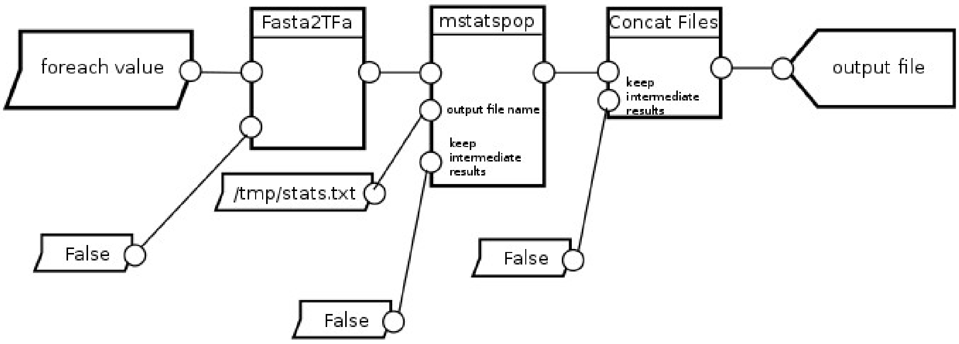


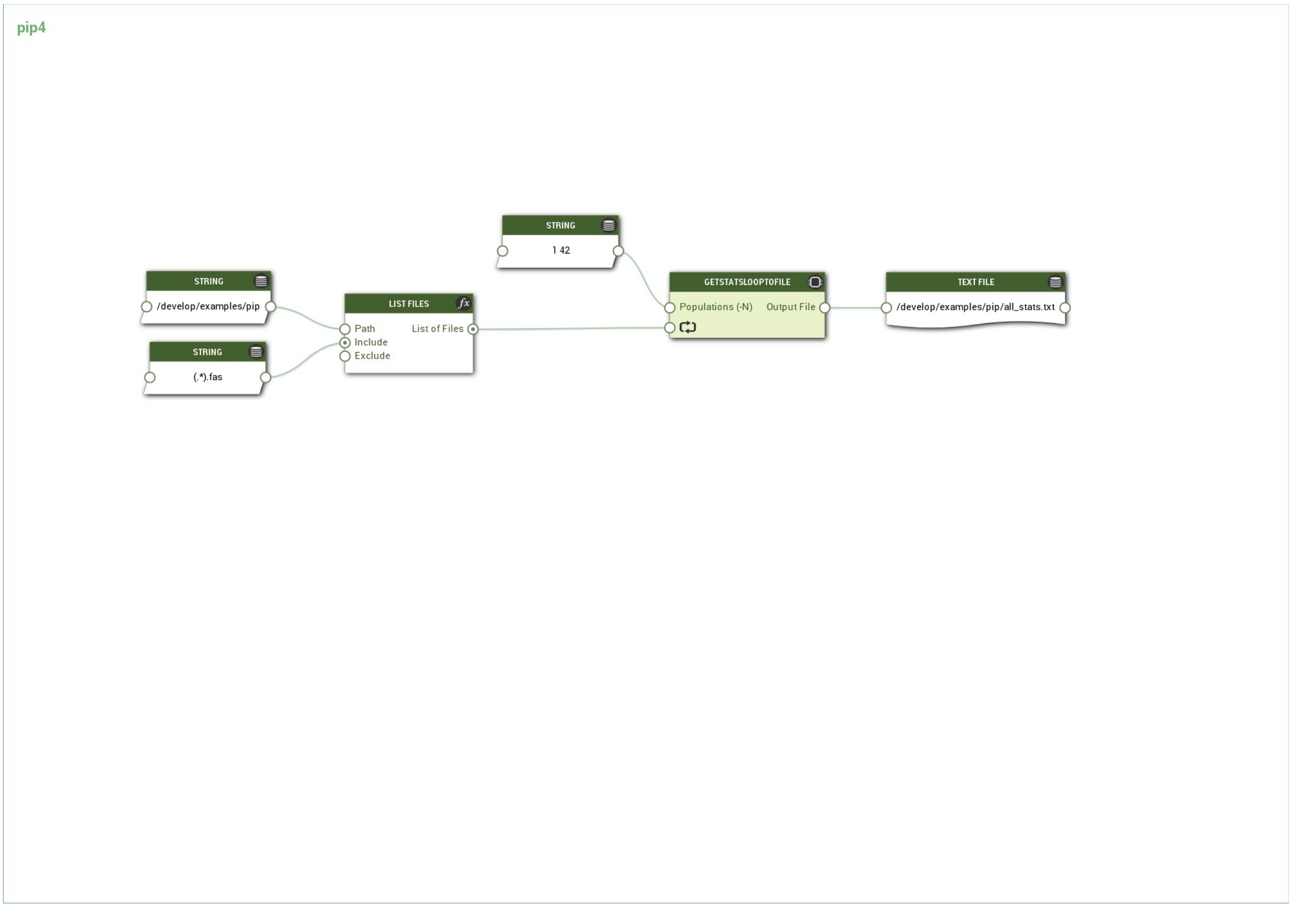
- Desired Pipeline Usage:

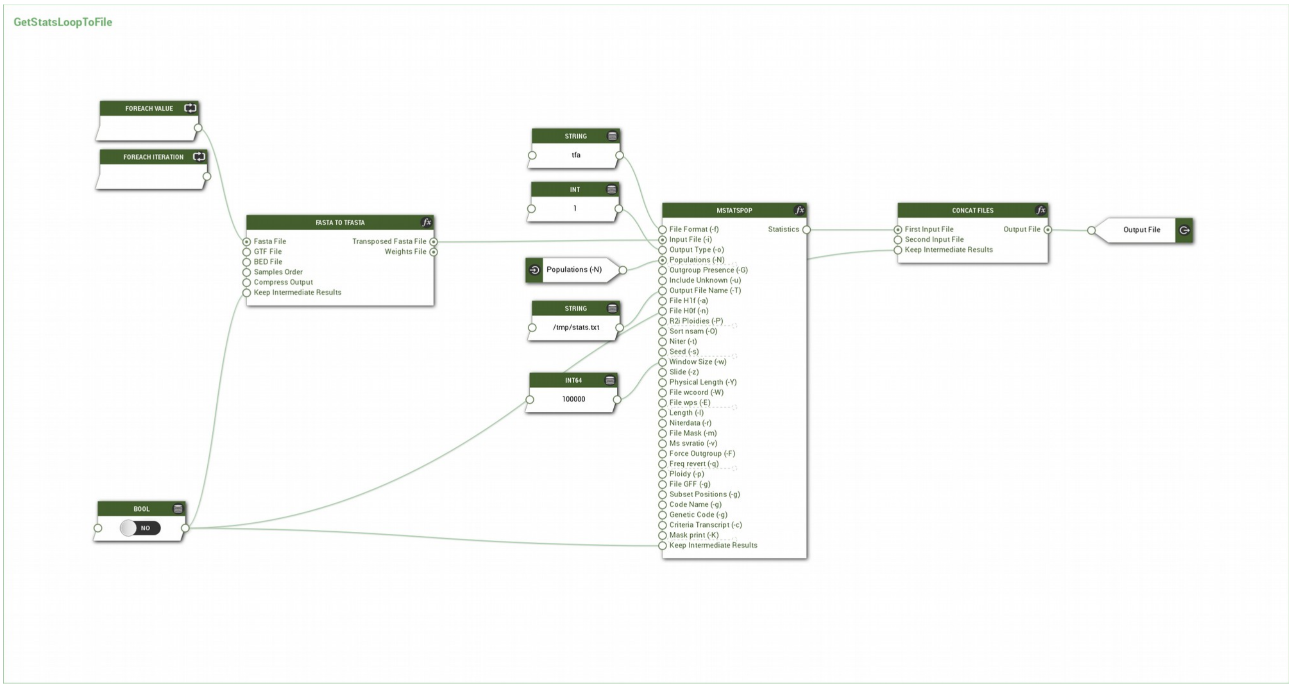


- Pipeline Design:

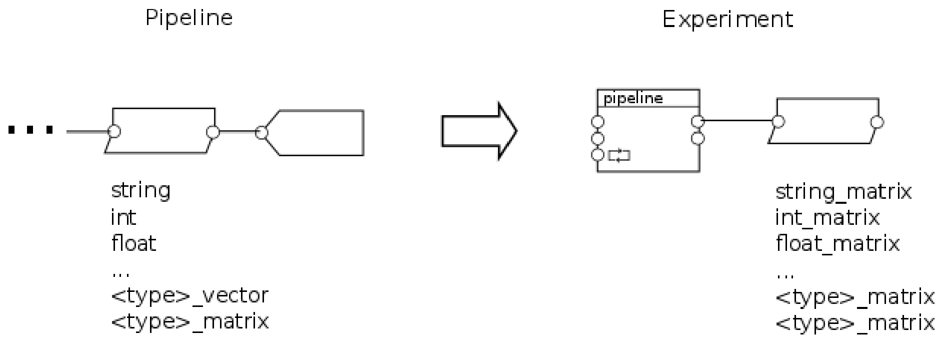
GetStats:

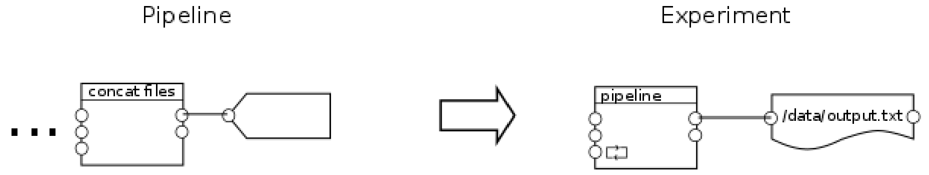



pip4
STRING
/develop/examples/pip
STRING
(.*).fas
LIST FILES
Path
Include
Exclude
List of Files
STRING
1 42
GETSTATSLOOPTOFILE
Populations (-N)
Output File
TEXT FILE
/develop/examples/pip/all_stats.txt


GetStatsLoopToFile
FOREACH VALUE
FOREACH ITERATION
FASTA TO TFASTA
Fasta File
GTF File
BED File
Samples Order
Compress Output
Keep Intermediate Results
Transposed Fasta File
Weights File
STRING
tfa
INT
1
Populations (-N)
STRING
/tmp/stats.txt
INT64
100000
BOOL
NO
MSTATSPOP
File Format (-f)
Input File (-i)
Output Type (-o)
Populations (-N)
Outgroup Presence (-G)
Include Unknown (-u)
Output File Name (-T)
File H1f (-a)
File H0f (-n)
R2i Ploidies (-P)
Sort nsam (-O)
Niter (-t)
Seed (-s)
Window Size (-w)
Slide (-z)
Physical Length (-Y)
File wcoord (-W)
File wps (-E)
Length (-l)
Niterdata (-r)
File Mask (-m)
Ms svratio (-v)
Force Outgroup (-F)
Freq revert (-q)
Ploidy (-p)
File GFF (-g)
Subset Positions (-g)
Code Name (-g)
Genetic Code (-g)
Criteria Transcript (-c)
Mask print (-K)
Keep Intermediate Results
Statistics
CONCAT FILES
First Input File
Second Input File
Keep Intermediate Results
Output File
Output File

5) Pipeline Outputs Sumary


Pipeline
Experiment
...
pipeline
string
int
float
...
<type>_vector
<type>_matrix
string_matrix
int_matrix
float_matrix
...
<type>_matrix
<type>_matrix


Pipeline
Experiment
...
concat files
pipeline
/data/output.txt