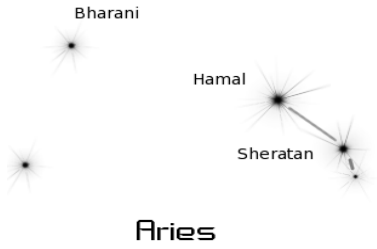
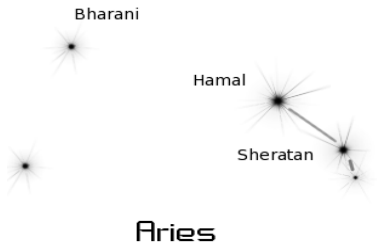


a(ccessory)**g**(enome) **MLST**

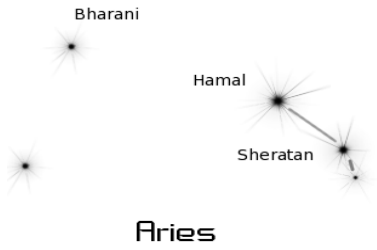
from loci to alleles



Before Starting

Why (ML) Sequence Typing? (instead of sequences)

data portability
data sharing (string profiles)
independent by platform
simple comparison algorythms



Current Solutions

MLST (7 loci)

adv: large db, few data, universally fed

con: could not be enough discriminative

does not provide particular functional information

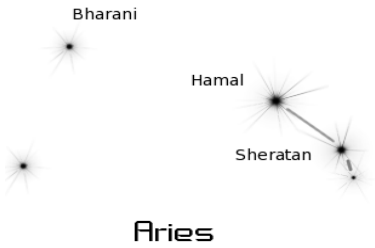
cgMLST (core genome)

adv: discriminative, can give functional information

con: dependent by platform and assembly

static schema or dynamic?

not adequate for organisms with large or variable accessory genome



Our Approach, agMLST

Focusing on ACCESSORY genome

before in Hrevap

it can give insights in close related strains (discriminative)

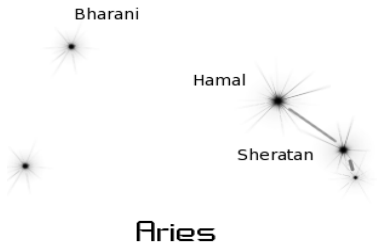
it can provide functional information

it could be more curated since it receives attention from scientific community

agMLST

a couple of tools to invesigate it

and to convert it into alleles



The agMLST core: the database

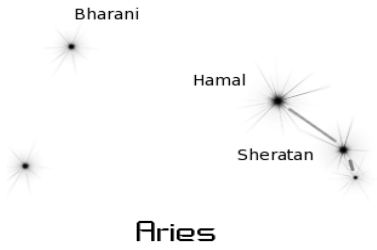
MLST approaches work well only if they have a good DB behind…
It represents the core of the method and its value

Our DB contains 79 loci over a total of 879 alleles

1-144 alleles per locus (Avg: 11)
alleles len 117-9672 (Avg: 1716)
Manually Curated!!!

;-) Thanks Valeria!!!

```
>astA:1:AF161000
ATGCCATCAACACAGTATATCCGGAGACCCACATCCAGTTATGCATCGTGCATATGGTGCGCAACAGTCTGC
GCTTCGTGTCATGGAAGGACTACAAAGCCGTCACTCGCGACCTGA
>astA:2:AF161001
ATGCCATCAACACAGTATATCCGGAGACCCACATCCAGTTATGCATCGTGCATATGGTGCGCAACAGCGTGC
GCTTCGTGTCATGGAAGGACTACAAAGCCGTCACTCGCGACCTGA
>astA:3:AB042005
ATGCCATCAACACAGTATATCCGGAGGCCCGCATCCAGTTATGCATCGTGCATATGGTGCGCAACAGCCTGC
GCTTCGTGTCATGGAAGGACTACAAAGCCGTCACTCGCGACCTGA
>astA:4:AB042002
ATGCCATCAACACAGTATATCCGAAGGCCCGCATCCAGTTATGCATCGTGCATATGGTGCGCAACAGCCTGC
GCTTCGTGTCATGGAAGGACTACAAAGCCGTCACTCGCGACCTGA
>astA:5:AF160998
ATGCCATCAACACAGTATATCCGGAGGCCAGCATCCAGTTATGCATCGTGCATATGGTGCGCAACAGCCTGC
GCTTCGTGTCATGGAAGGACTACAAAGCCGTCACTCGCGACCTGA
>astA:6:AY545598
ATGCCATCAACACAGTATATCCGAAGGCCCGCATCCAGTTATGCATCGTGCATATGGTGCGCAACAGCCTGC
GCTTCGTGTCATGGAAGGACTACAAGGCCGTCACTCGCGACCTGA
>astA:7:AF411067
ATGCCATCAACGCAGTATATCCGAAGGCCCGCATCCAGTTATGCATCGTGCATATGGTGCGCAACAGCCTGC
GCTTCGTGTCATGGAAGGACTACAAAGCCGTCACTCGCGACCTGA
```



Alleles Conversion Strategies

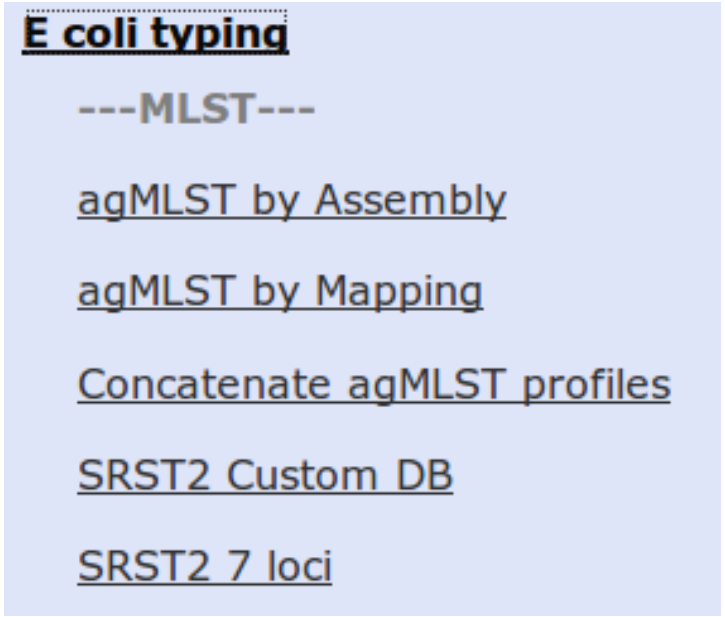
How to convert Sequences to strings…

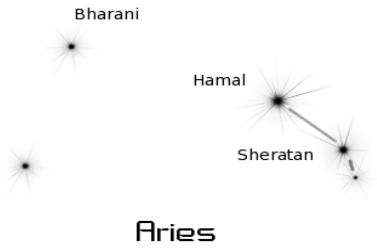
Already Existing tools:

- on line tools (pubmlst)
- **SRST2**
- *Commercial solutions*

Our solutions:

- agMLST by reads
- agMLST by assembly





SRST2

It derives *traditional* MLST from reads

It also accepts custom DBs

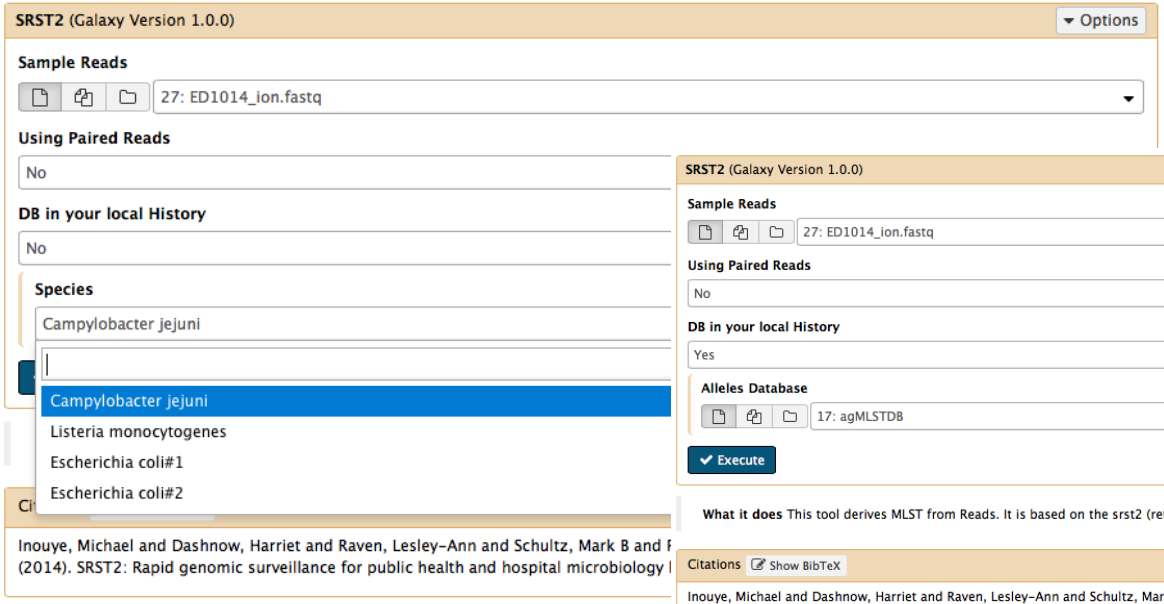
Inouye *et al. Genome Medicine* 2014, **6**:90
http://genomemedicine.com/content/6/11/90

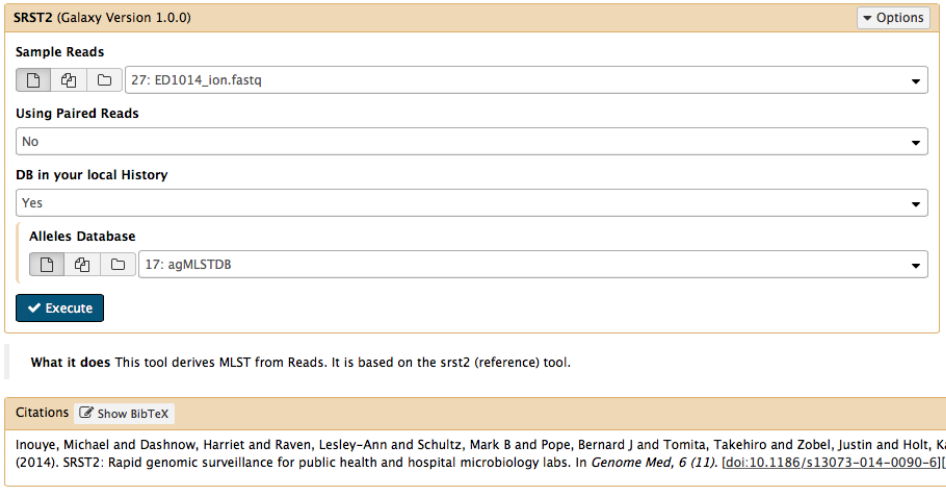
Genome Medicine

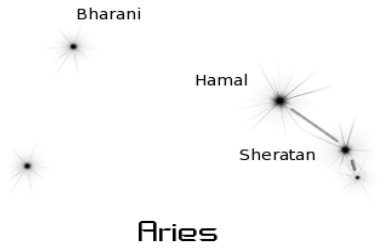
SOFTWARE Open Access

SRST2: Rapid genomic surveillance for public health and hospital microbiology labs

Michael Inouye[1,2], Harriet Dashnow[3,4], Lesley-Ann Raven[1], Mark B Schultz[3], Bernard J Pope[4,5], Takehiro Tomita[2,6], Justin Zobel[5] and Kathryn E Holt[3*]



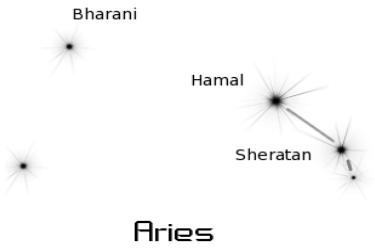




SRST2

It returns an array of allels (*the classical 7 loci schema*)
It includes: snps, indels, coverage, alternative allele frequency

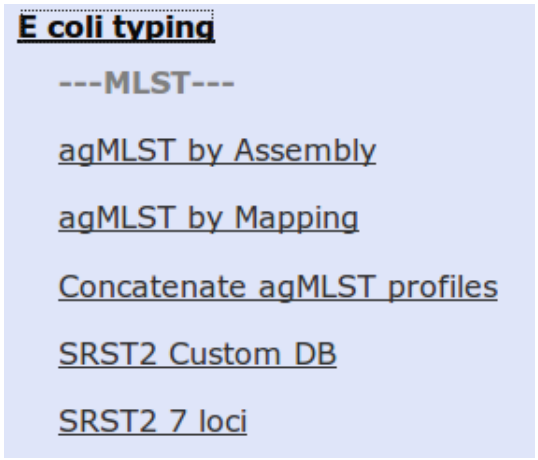
1	2	3	4	5	6	7	8	9	10	11	12	13	14
Sample	DB	gene	allele	coverage	depth	diffs	uncertainty	divergence	length	maxMAF	clusterid	seqid	annotation
dataset_50782	dataset_50786	stx2A	70_683	100.0	66.923			0.0	960	0.136	67	683	AY143336 a
dataset_50782	dataset_50786	stx2B	27_754	100.0	48.402			0.0	270	0.038	68	754	AE005174 a
dataset_50782	dataset_50786	iha	4_334	90.326	4.381	137snp6indel199holes	edge0.0	7.253	2088	0.5	34	334	AF399919
dataset_50782	dataset_50786	espJ	1_210	100.0	52.527			0.0	654	0.196	21	210	AB303060
dataset_50782	dataset_50786	espP	3_214	100.0	122.951			0.0	3903	0.232	22	214	GQ259888
dataset_50782	dataset_50786	lpfA	3_413	100.0	49.876			0.0	573	0.39	43	413	AP010953
dataset_50782	dataset_50786	katP	1_408	100.0	124.519			0.0	2211	0.153	41	408	AB011549
dataset_50782	dataset_50786	cif	2_65	99.882	64.122	1snp1indel		0.118	849	0.143	8	65	AY128535
dataset_50782	dataset_50786	toxB	3_849	100.0	116.057			0.0	9501	0.429	73	849	AB456530
dataset_50782	dataset_50786	tir	34_844	99.876	52.459	1snp1indel1holes		0.062	1617	0.222	72	844	AB426060
dataset_50782	dataset_50786	iss	8_391	100.0	104.414			0.0	294	0.248	39	391	CP001665
dataset_50782	dataset_50786	eae	45_139	100.0	65.663	2snp		0.071	2820	0.148	12	139	ECU59503
dataset_50782	dataset_50786	efa1	11_153	99.979	61.531	2indel		0.0	9672	0.438	14	153	AJ277443
dataset_50782	dataset_50786	espB	13_190	100.0	69.044			0.0	945	0.195	17	190	AF054421
dataset_50782	dataset_50786	espA	22_176	100.0	66.603			0.0	579	0.233	16	176	FM201463
dataset_50782	dataset_50786	espF	2_196	100.0	42.051	4snp		0.641	624	0.5	19	196	AF116900
dataset_50782	dataset_50786	ehxA	7_324	100.0	119.967			0.0	2997	0.116	32	324	HM138194
dataset_50782	dataset_50786	gad	20_267	100.0	93.909	2snp		0.143	1401	0.494	31	267	AP010953
dataset_50782	dataset_50786	prfB	13_547	100.0	45.279			0.0	882	0.222	56	547	CP002970
dataset_50782	dataset_50786	nleB	11_499	100.0	69.592			0.0	990	0.219	51	499	AF453441
dataset_50782	dataset_50786	nleA	12_482	100.0	64.067			0.0	1323	0.3	50	482	AM422003
dataset_50782	dataset_50786	nleC	3_505	100.0	51.593	1snp		0.101	987	0.341	52	505	AP010953

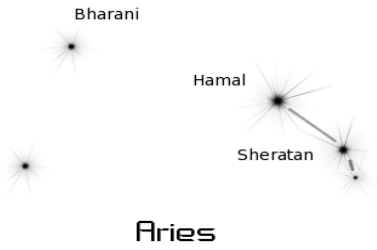


Our agMLST implementation

Two different strategies: by mapping, by assembling

Both of them based on three steps: 1) format DB
2) aligning / assembly
3) scoring





agMLST format DB

To guarantee compatibily with SRST2, headers of custom fasta files in the format Locus/allele are converted in:

familyID__Locus__allele__counter eg:

>astA:3 AB000123

>3__astA__3__123 AB000123

space

Annotation (opt)

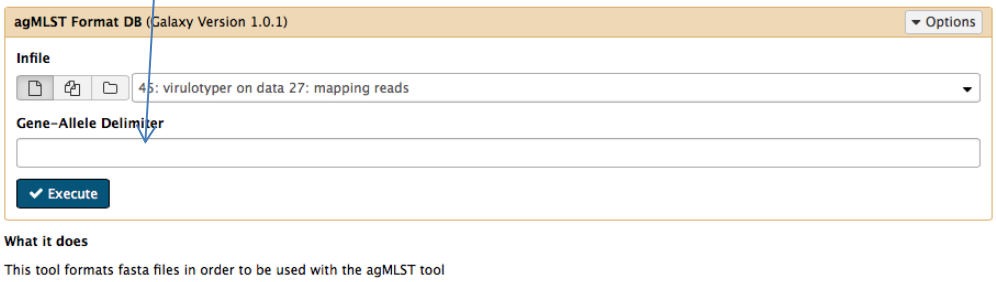
Family ID

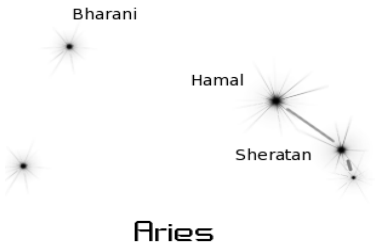
Locus ID

Allele ID

Unique Identifier

Double underline as separator





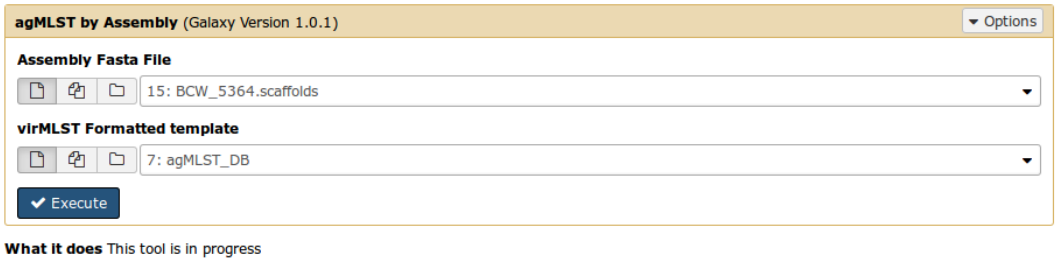
Why Two agMLST tools?

By Contigs/Draft:

you can have draft but not reads…

assemblies are easy to share

independent by platform

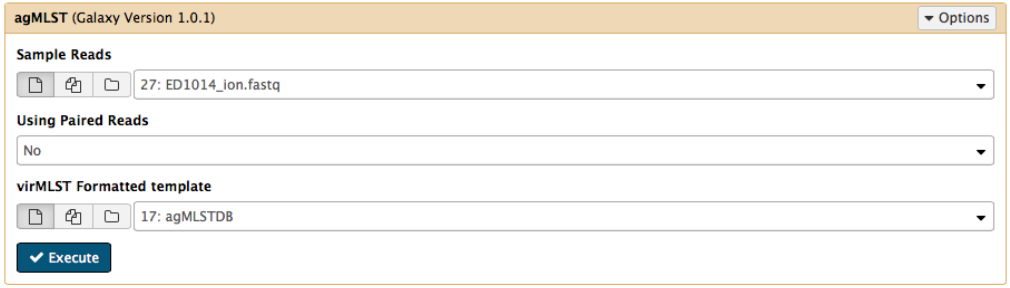


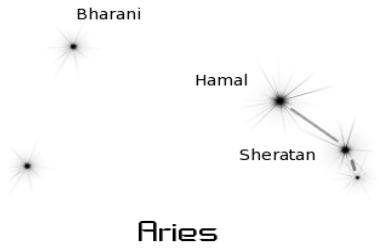
By Reads

if you have reads you can assembly them, but you wouldn't

comparison of performances with exhisting tools

partially assembled genes?

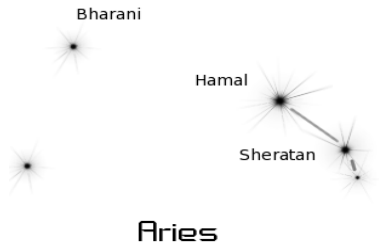




agMLST, how does it work?

By Contigs/Draft:
DB blasted against assembly
hits are ranked
best matches are compared to the DB for new alleles discovery

By Reads
reads are aligned against all alleles
consensus are generated and ranked
best consensus is compared to the DB for new alleles discovery



agMLST, what does it return?

alleles array

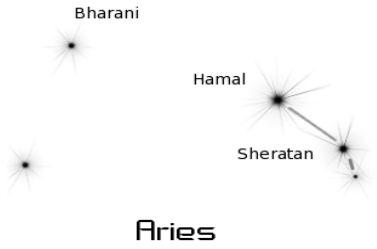
new alleles fasta file

extended log file

Same Format of SRST2

New Alleles to DB Curators
(Valeria, obviously...)

By Contigs/Draft: raw blast results
By Reads: raw align file (pileup/bam)



agMLST, what does it return?

Profiles, they can be easily concatenated and clustered

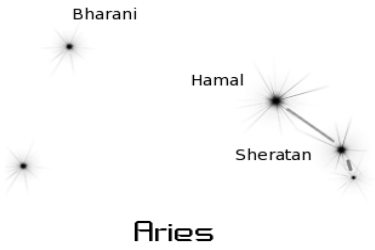
astA	mchB	mchC	fasA	ehxA	tccP	sfaS	tir	ipaH9.8	cofA	picU	iha	sat	sigA	espI	espJ
NM	3	NM	NM	3	NM	NM	NM	NM	NM	NM	NM	7	NM	NM	34*

New Alleles

```
# agMLST Alleles Pipeline on /clusterfs/galaxy/galaxy-dist/database/files/050/dataset_50925.dat Contigs File
# ------------------------------------------------------------------------------
# Running blast...
Locus :  nleC   New Allele! closest: 3   Aligned:  987   Identity:  99.59 Mismatches:  4   Gaps:  0
Locus :  nleB   Allele:  11
Locus :  cif   New Allele! closest: 2   Aligned:  849   Identity:  99.88 Mismatches:  0   Gaps:  1
Locus :  prfB   Allele:  13
Locus :  espJ   Allele:  1
Locus :  iss   Allele:  8
Locus :  stx2A   Allele:  70
Locus :  stx2B   Allele:  27
Locus :  lpfA   Allele:  3
Locus :  eae   New Allele! closest: 45   Aligned:  2820   Identity:  99.93 Mismatches:  2   Gaps:  0
Locus :  ehxA   Allele:  7
Locus :  tir   New Allele! closest: 34   Aligned:  1617   Identity:  99.88 Mismatches:  2   Gaps:  0
Locus :  katP   Allele:  1
Locus :  espF   Allele:  2
Locus :  espA   Allele:  22
Locus :  espP   Allele:  1
Locus :  efa1   Allele:  11
Locus :  espB   Allele:  13
Locus :  toxB   Allele:  3
# ------------------------------------------------------------------------------
```

```
>tir__NewAllele_closeTo_34
ATGCCTATTGGTAATCTTGGCCACAATCCCAATGTGAGAGCTTTAATTCCACCTGCACCGCCATTACCTTCACAAACCGACGGTGCAGGAGGTGCCCGTA
ATCAGCTCATTAACTCAAATGGCCCGATGGGGTCTCGTTTGCTATTTACGCCTATAAGGAATTCTGTTGCTGATGCTGCTGATTCTCGTGCCAGTGATAT
TCCCGGACTTCCTACAAATCCACTGCGCTTTGCTGCGTCCGAGGTATCTTTGCATGGTGCGCTTGAAGTTCTTCATGATAAAGGGGGGCTTGATACTCTT
AACTCTGCTATTGGATCTTCGTTATTCCGTGTTGAAACTCGGGATGATGGCAGCCATGTTGCTATCGGGCAAAAAAATGGCCTCGAGACCACTGTTGTTT
TAAGTGAGCAAGAGTTTTCTAGCTTACAGTCCCTTGATCCTGAAGGTAAAAACAAATTTGTATTTACTGGAGGCCGCGGTGGCGCAGGGCATGCTATGGT
CACGGTTGCTTCAGATATCGCCGAAGCCCGTCAGAGGATAATAGATAAATTAGAACCAAAGGATACAAAGGAGACGAAGGAGCCAGGGGATCCAAATAGT
GGCGAGGGAAAAATCATTGAAATTCATACCTCAACCTCAACTTCTAGCCTCCGTGCAGATCCTAAACTTTGGTTGTCATTGGGGACTATTGCTGCAGGTC
TGATAGGGATGGCTGCGACGGGGATTGCACAGGCTGTTGCGTTGACTCCAGAGCCGGATGACCCAATCACTACCGACCTGATGCTGCAGCAAACACAGC
TGAAGCAGCGGCAAAAGATCAGTTAACGAAAGAAGCATTCCAGAACCCAGATAACCAGAAAGTTAATATCGATGAGAACGGAAATGCAATTCCGTCCGGG
GAACTAAAAGATGATGTTGTTGCGCAAATAGCAGAACAAGCTAAAGCGGCGGGTGAACAGGCCAGACAGGAAGCTATTGAAAGTAATTCTCAGGCGCAGC
AAAAATATGATGAACAGCATGCTAAACGCGAACAGGAAATGTCTCTTTCATCGGGGGTTGGCTACGGTATTAGTGGTGCGCTGATTCTTGGCGGGGGAAT
TGGTGCCGGTGTTACTGCTGCTCTTCATCGGAAAAACCAACCGGCAGAACAAACAATCACTACACGTACGGTAGTCGATAATCAGCCTACGAATAACGCA
TCTGCGCAGGGCAATACTGACACAAGTGGGCCAGAAGAGTCCCCGGCGAGCAGACGTAATTCGAATGCCAGCCTCGCATCGAACGGGTCTGACACCTCCA
GCACGGGCACGGTAGAGAATCCGTATGCTGACGTTGGAATGCCCAGAAATGATTCACTGGCTCGCATTTCAGAGGAACCTATTTATGATGAGGTCGCTGC
AGATCCTAATTATAGCGTCATTCAACATTTTTCAGGGAACAGCCCAGTTACCGGAAGGTTAGTGGGAACCCCAGGGCAAGGTATCCAAAGTACTTATGCG
CTTCTGGCAAGCAGCGGCGGATTGCGTTTAGGTATGGGAGGATTAACGGGGGGGGGCGAGAGCGCAGTAAGTACTGCCAATGCCGCACCAACGCCGGGAC
CCGCACGTTTCGTTTAA
>eae__NewAllele_closeTo_45
ATGATTACTCATGGTTTTATGCCCGGACCCGGCACAAGCATAAGCTAAAAAAAACATTTATTATGCTTAGTGCTGGTTTAGGATTGTTTTTTTATGTTA
ACCAGAATTCATTTGCAAATGGTGAAAATTATTTTAAATTGAGTTCAGATTCAAAACTGTTAACTCAAAATGCCGCTCAGGATCGCCTTTTTTATACGTT
AAAAACAGGTGAAACTGTTGCCAATATTTCTAAATCACAGGGTATCAGTTTATCGGTAATTTGGTCACTGAATAAACATTTATACAGTTCCGAAAGCGAA
ATGATGAAGGCTGGACCTGGTCAGCAGATCATTTTGCCACTCAAAAAACTGTCTGTTGAATATAGTGCCTTACCTGTCTTAGGTTCGGCACCTGTTGTTG
CTGCAGGTGGTGTCGCTGGTCATACGAATAAAATGACTAAAATGTCCCCGGACGCGACTAAAAGCAACACGACCGATGACAAGGCTCTAAATTATGCGGC
ACAACAGGCCGCGAGCCTTGGTAGCCAGCTCCAGTCGCGCTCACTGAACGGCGATTACGCGAAAGATACCGCTCTTGGTATGGCCAGCAGCCAGGCTTCG
TCACAGTTGCAGGCCTGGTTACAACATTATGGAACGGCAGAGGTTAATCTGCAGAGTGGTAATAACTTTGACGGTAGTTCACTGGACTTCTTATTACCGT
TCTATGATTCCGAAAACATGCTGGCATTTGGTCAGGTCGGTGCGCGTTACATTGACTCCCGCTTTACGGCAAATTTAGGTGCTGGCCAGCGTTTTTCCT
TCCTGAAAATATGTTGGGCTATAACGTCTTCATTGATCAGGATTTTTCTGGTGATAATACCCGTTTAGGTATTGGTGGCGAATACTGGCGAGACTATTTC
AAAAGTAGCGTTAACGGCTATTTCCGCATGAGCGGCTGGCATGAGTCATACAATAAGAAAGACTATGATGAGCGCCCGGCAAATGGTTTTGATATCCGCT
TTAATGGCTATTTACCATCATATCCGGCATTAGGCGCCAAACTGATGTACGAACAGTATTATGGTGATAATGTTGCTTTGTTTAATTCCGATAAGTTGCA
GTCGAATCCTGGCGCGGCGACCGTTGGTGTAAACTACACTCCGATTCCTCTGGTGACGATGGGGATCGATTACCGTCATGGTACGGGTAATGAAAATGAT
CTCCTTTACTCAATGCAGTTCCGTTATCAGTTTGATAAACCGTGGTCTCAGCAAATCGAGCCACAGTATGTTAACGAGTTAAGAACATTATCGGGCAGCC
GTTACGATCTGGTTCAGCGTAATAACAATATTATTCTGGAGTACAAAAAGCAGGATATTCTTTCTCTGAATATTCCGCATGATATTAATGGTACTGAACA
CAGTACGCAGAAGATTCAATTGATCGTTAAGAGCAAATACGGTCTGGATCGTATCGTCTGGGATGATAGCGCATTACGCAGTCAGGGCGGTCAGATTCAG
CATGGCGGAAGCCAAAGCGCACAAGACTACCAGGCTATTTTGCCTGCTTATGTGCAAGGCGGCAGCAATATTTATAAAGTGACCGCTCGCGCCTATGACC
GAAATGGTAATAGTTCTAATAATGTACAGCTCACTATTACCGTTTTACCGAATGGGCAGGTTGTGGACCAGGTTGGGGTAACGGACTTTACGGCTGATAA
AACATCGGCTAAAGCGGATGGCATAGAAGCTATTACCTATACCGCGACGGTTAAAAAGAATGGTGTAGCTCAGGCTAATGTCCCTGTAACATTTAGTATT
GTATCCGGGACTGCAACTCTTGGGGCAAATAGTGCCAGAACGGATGGTAACGGTAAGGCGACCGTAACGCTGAAGTCGGCTACGCCAGGACAGGTCGTCG
TGTCTGCTAAAACCGCGGAGATGACTTCGCCACTTAATGCCAGCGCGGTTATATTTGTTGATCAAACCAAGGCCAGTATTACTGAGATTAAGGCTGATAA
AACAACAGCGAAGGCAGATGGTTCTGATGCGATTACCTATACTGTCAGAGTGATGAAGGAGGGGGCACCCGTAGTAGATCAGAAAGTGACCTTTTCTAAG
GATTTTGGGACCCTGAATAAGACTGAAGCAACAACCGATCAGAATGGTTATGCTACTGTAAAATTATCATCCAATACTCCTGGCAAGGCCATTGTTAGTG
CAAAAGTGAGTGGAGTAGGTACAGAAGTTAAGGCTACTACCGTTGAGTTTTTTGCCCCGTTGAGTATTGATGGTGATAAAGTGACCGTAATTGGTACTGG
TATCACGGGGGCTCTGCCAAAGAACTGGTTACAGTATGGTCAGGTTAAGCTACAGGCAACAGGGGGCAATGGAAAATACACATGGAAATCCAGTAATACT
AAAATTGCTTCTGTTGATAACTCGGGAGTGATAACCTTAAATGAAAAAGGGAGTGCCACAATTACTGTAGTATCTGGTGATAATCAGAGTGCGACATACA
CAATTAATGCACCGGGTAGTATTGTAATTGCTGTGGATAAAAATACTCGAGTTACGTATTTTGATGCCGAAAACAAATGTAAGACAATAGCGCAAATTT
AGCACAGTCAAAAGAACTATTGGCCAATATCTATTCAACATGGGGTGCTGCAAATAAATATCCTTACTATTCTGGTTCTAAATCATTGACTGCTTGGATT
AAACAATCCTCTTCTGAACAGTCATCAGGTGTATCAAGCACATATGATTTGGTTACGAAGAACCAGTTGATCAATGTTGGAGTAAACAATAAGAATGCTT
TTTCTGTTTGTGTAAAATAA
```

Log File - Description



Briefly, an example

AGMLST

SRST2 7 loci

SRST2 Custom DB

agMLST DB reformat

agMLST by Mapping

agMLST by Assembly

Concatenate agMLST profiles

Assembled

Trimmed Reads

Formatted DB

DB

Trimmed Reads Ion

46: BCW_5364.scaffold.fasta

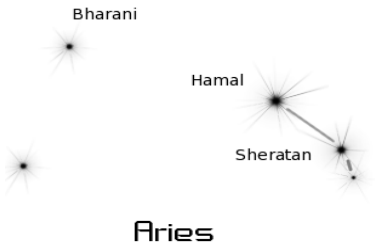
11: BCW_5364_2.fastq

10: BCW_5364_1.fastq

3: agMLST_DB

2: virulence_ecoli.fa

1: ED1014_ion.fastq



Briefly, an example

AGMLST

SRST2 7 loci

SRST2 Custom DB

agMLST DB reformat

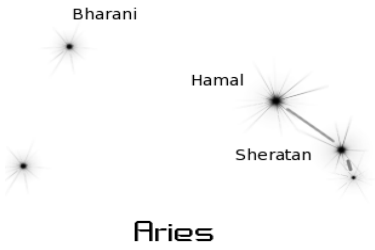
agMLST by Mapping

agMLST by Assembly

Concatenate agMLST profiles

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help

Sample	ST	adk	fumC	gyrB	icd	mdh	purA	recA	mismatches	uncertainty	depth	maxMAF
dataset_1340	21	16	4	12	16	9	7	7	0	-	13.5141428571	0.125



Briefly, an example

AGMLST

SRST2 7 loci

SRST2 Custom DB

agMLST DB reformat

agMLST by Mapping

agMLST by Assembly

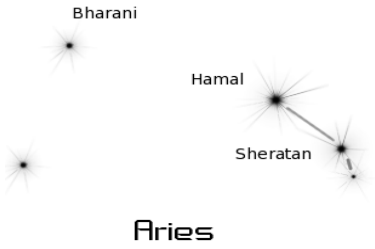
Concatenate agMLST profiles

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User

Sample	cif	eae	efa1	ehxA	espA	espB	espF	espJ	espP	gad	iha	iss	katP
dataset_1340	2_65*	45_139*	11_153	7_324	22_176	13_190	2_196*	1_210	1_864	20_267*	14_344*	8_391	1_408

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help

Sample	DB	gene	allele	coverage	depth	diffs	uncertainty	divergence	length	maxMAF
dataset_1340	dataset_1332	stx2A	70_683	100.0	17.424			0.0	960	0.111
dataset_1340	dataset_1332	stx2B	27_754	100.0	20.546			0.0	270	0.037
dataset_1340	dataset_1332	iha	14_344	100.0	29.439	10snp		0.478	2091	0.5
dataset_1340	dataset_1332	espJ	1_210	100.0	16.4			0.0	654	0.083
dataset_1340	dataset_1332	espP	1_864	100.0	20.448			0.0	774	0.091
dataset_1340	dataset_1332	lpfA	3_413	100.0	14.908			0.0	573	0.125
dataset_1340	dataset_1332	katP	1_408	100.0	20.82			0.0	2211	0.105
dataset_1340	dataset_1332	cif	2_65	99.882	15.822	1indel		0.0	849	0.091
dataset_1340	dataset_1332	toxB	3_849	100.0	23.012			0.0	9501	0.2
dataset_1340	dataset_1332	tir	34_844	100.0	15.558	2snp		0.124	1617	0.083
dataset_1340	dataset_1332	iss	8_391	100.0	31.831			0.0	294	0.048
dataset_1340	dataset_1332	eae	45_139	100.0	16.454	2snp		0.071	2820	0.111
dataset_1340	dataset_1332	efa1	11_153	100.0	18.118			0.0	9672	0.364
dataset_1340	dataset_1332	espB	13_190	100.0	15.368			0.0	945	0.118
dataset_1340	dataset_1332	espA	22_176	100.0	19.769			0.0	579	0.053
dataset_1340	dataset_1332	espF	2_196	100.0	18.55	1snp		0.16	624	0.478
dataset_1340	dataset_1332	ehxA	7_324	100.0	21.649			0.0	2997	0.083
dataset_1340	dataset_1332	gad	20_267	100.0	26.141	1snp		0.071	1401	0.457
dataset_1340	dataset_1332	prfB	13_547	100.0	13.978			0.0	882	0.333
dataset_1340	dataset_1332	nleB	11_499	100.0	18.507			0.0	990	0.095
dataset_1340	dataset_1332	nleA	13_483	100.0	26.326	1snp		0.076	1323	0.5
dataset_1340	dataset_1332	nleC	6_508	100.0	14.332	4snp		0.405	987	0.091



Briefly, an example

AGMLST

SRST2 7 loci

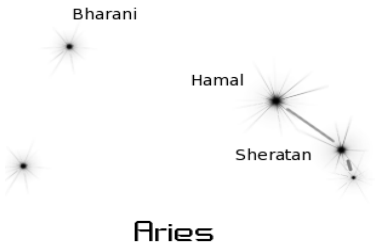
SRST2 Custom DB

agMLST DB reformat

agMLST by Mapping

agMLST by Assembly

Concatenate agMLST profiles



Briefly, an example

AGMLST

SRST2 7 loci

SRST2 Custom DB

agMLST DB reformat

agMLST by Mapping

agMLST by Assembly

Concatenate agMLST profiles

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
sampleID		astA		cnf1		cofA		eae		eatA		efa1		epeA		espA		espB
t	pic		prfB		rpeA		sat		senB		celb		sepA		sfaS		sigA		sta1
test9	NM	NM	NM	NM	NM	NM	NM	22	NM	NM	2	NM	NM	1	NM			

85: agMLST New Alleles File

4 sequences

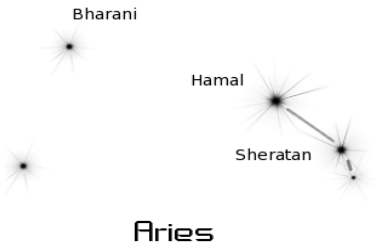
format: fasta, database: ?

```
>tir__NewAllele_closeTo_34
ATGCCTATTGGTAATCTTGGCCACAATCCCAATGTGAGAGCTTTAATTCCACCTGCACCGCCATTACCTTC
ATCAGCTCATTAACTCAAATGGCCCGATGGGGTCTCGTTTGCTATTTACGCCTATAAGGAATTCTGTTGCTC
TCCCGGACTTCCTACAAATCCACTGCGCTTTGCTGCGTCCGAGGTATCTTTGCATGGTGCGCTTGAAGTTC
AACTCTGCTATTGGATCTTCGTTATTCCGTGTTGAAACTCGGGATGATGGCAGCCATGTTGCTATCGGGCA
TAAGTGAGCAAGAGTTTTCTAGCTTACAGTCCCTTGATCCTGAAGGTAAAAACAAATTTGTATTTACTGGAC
```

Galaxy — Analyze Data | Workflow | Shared Data | Visualization | Help | User

```
# agMLST Alleles Pipeline on /Users/massimilianoorsini/Sftw/galaxy/database/files/001/dataset_1403.dat Reads File
# ------------------------------------------------------------
# DB Statistics. Loci : 78. Alleles: 879
Error: Read 6/14/2016 21:56:14 program started has more read characters than quality values.
libc++abi.dylib: terminating with uncaught exception of type int
(ERR): bowtie2-align died with signal 6 (ABRT)
[samopen] SAM header is present: 879 sequences.
[mpileup] 1 samples in 1 input files
<mpileup> Set max per-file depth to 8000
```

# summary.	locus	HorCov	VertCov	Assembled		Indels	Mismatches	Ns	Lenght	Expected
#passing	39__iss__5__388	91.2	7	291	0	0	23	291	294	
#passing	52__nleC__6__508	98.9	13	986	0	2	10	986	987	
#passing	34__iha__13__343	99.0	5	2082	0	0	12	2082	2091	
#passing	32__ehxA__7__324	100.0	10	2996	0	1	0	2996	2997	
#passing	32__ehxA__5__322	97.6	9	2995	0	1	69	2995	2997	
#passing	16__espA__22__176	99.7	14	577	0	0	0	577	579	
#passing	31__gad__9__256	92.0	2	1399	0	0	110	1399	1401	
# notpassing	17__espB__13__190	99.9	5	944	1	0	0	944	945	
#passing	56__prfB__13__547	99.9	5	881	0	2	0	881	882	
#passing	17__espB__9__186	92.4	3	944	0	4	71	944	945	
#passing	39__iss__13__396	95.0	13	341	0	0	16	341	342	
#passing	34__iha__12__342	95.9	4	2089	0	0	83	2089	2091	
# notpassing	14__efa1__10__152	96.8	6	9670	4	5	310	9670	9672	
# notpassing	17__espB__10__187	91.9	5	943	1	0	75	943	945	
#passing	16__espA__23__177	92.1	10	577	0	0	44	577	579	
# notpassing	12__eae__45__139	99.9	17	2819	1	2	1	2819	2820	
#passing	72__tir__32__842	94.6	8	1616	0	0	87	1616	1617	
#passing	14__efa1__11__153	97.6	7	9669	0	6	231	9669	9672	
# notpassing	22__espP__3__214	100.0	13	3902	5	1	0	3902	3903	
#passing	72__tir__34__844	98.2	8	1615	0	0	27	1615	1617	
#passing	39__iss__6__389	92.2	2	292	0	0	21	292	294	
#passing	39__iss__8__391	98.3	11	293	0	0	4	293	294	
#passing	52__nleC__3__505	98.0	13	986	0	3	19	986	987	
# notpassing	73__toxB__4__850	100.0	15	9500	8	0	2	9500	9501	



Briefly, an example

AGMLST

SRST2 7 loci

SRST2 Custom DB

agMLST DB reformat

agMLST by Mapping

agMLST by Assembly

Concatenate agMLST profiles

76: agMLST Profiling

2 lines

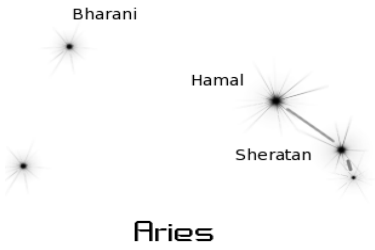
format: tabular, database: ?

[H[2J# agMLST Alleles Pipeline on /Users/massimilianoorsini/Sftw/galaxy/database/files/001/dataset_1403.dat Reads File # ---------- # DB Statistics. Loci : 78. Alleles: 879 # summary. locus HorCov VertCo

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
sampleID		astA	cnf1	cofA	eae	eatA	efa1	epeA	espA	espB	espC	espF	bfpA	espI	espJ								
t	pic	prfB	rpeA	sat	senB	celb	sepA	sfaS	sigA	sta1	stb	stx1A	stx1B	stx2A	stx2B								
test9	NM	NM	NM	NM	NM	NM	NM	22	NM	NM	2	NM	NM	1	NM	NM	NM	NM	NM	NM	NM	NM	NM

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help

```
# agMLST Alleles Pipeline on /Users/massimilianoorsini/Sftw/galaxy/database/files/001/dataset_1375.dat Contigs File
# ------------------------------------------------------------------------
# Running blast...
# Locus :  nleC   New Allele! closest: 3   Aligned:  987   Identity:  99.59 Mismatches:  4   Gaps:  0
# Locus :  nleB   Allele:  11
# Locus :  cif   New Allele! closest: 2   Aligned:  849   Identity:  99.88 Mismatches:  0   Gaps:  1
# Locus :  prfB   Allele:  13
# Locus :  espJ   Allele:  1
# Locus :  iss   Allele:  8
# Locus :  stx2A   Allele:  70
# Locus :  stx2B   Allele:  27
# Locus :  lpfA   Allele:  3
# Locus :  eae   New Allele! closest: 45   Aligned:  2820   Identity:  99.93 Mismatches:  2   Gaps:  0
# Locus :  ehxA   Allele:  7
# Locus :  tir   New Allele! closest: 34   Aligned:  1617   Identity:  99.88 Mismatches:  2   Gaps:  0
# Locus :  katP   Allele:  1
# Locus :  espF   Allele:  2
# Locus :  espA   Allele:  22
# Locus :  espP   Allele:  1
# Locus :  efa1   Allele:  11
# Locus :  espB   Allele:  13
# Locus :  toxB   Allele:  3
# ------------------------------------------------------------------------
```



Briefly, an example

AGMLST

SRST2 7 loci

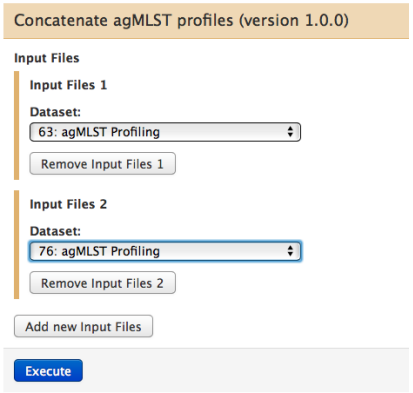
SRST2 Custom DB

agMLST DB reformat

agMLST by Mapping

agMLST by Assembly

Concatenate agMLST profiles



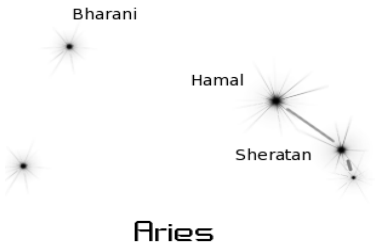
89: SRST2 Alleles Multi SampleTable

4 lines

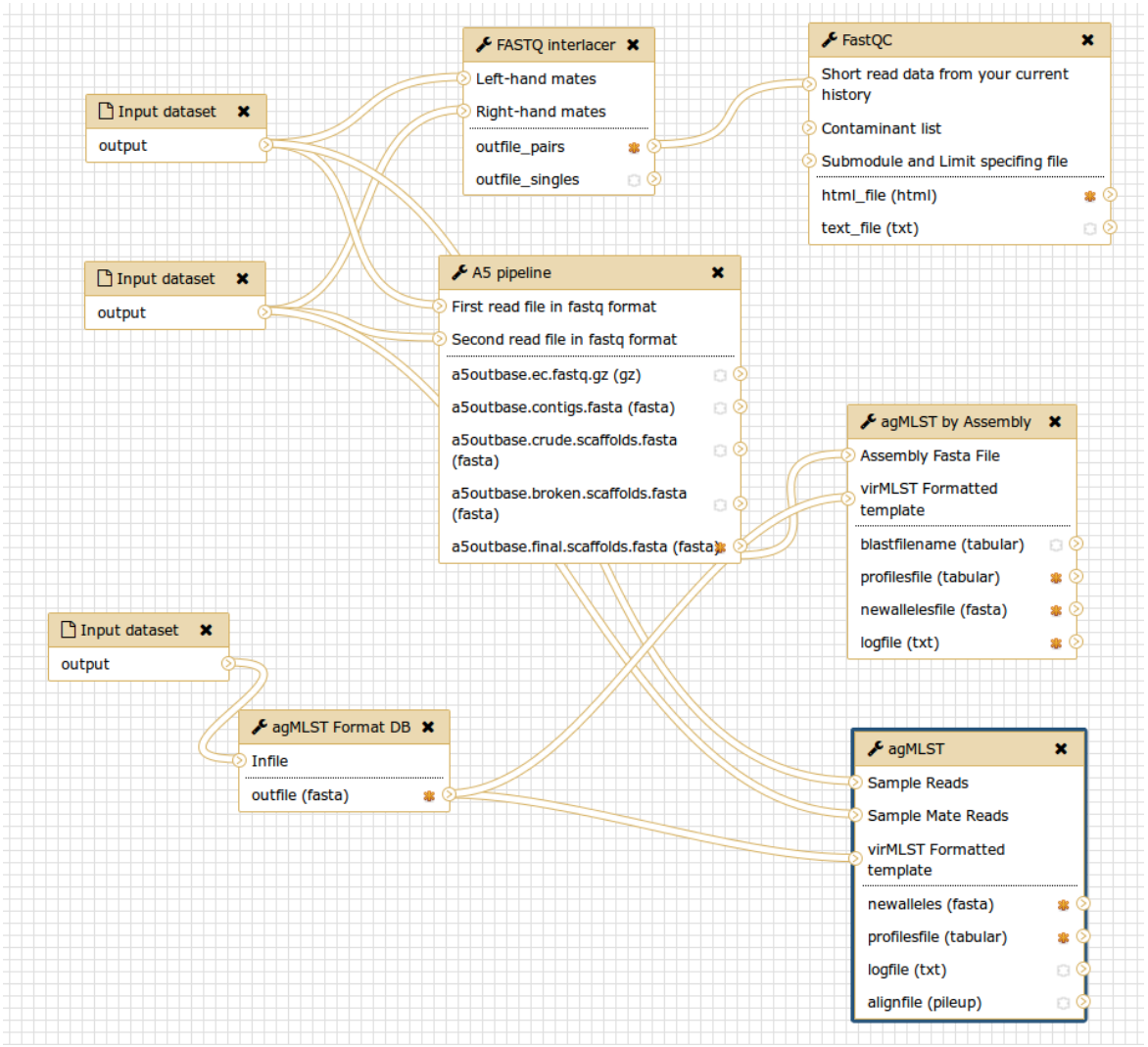
format: tabular, database: ?

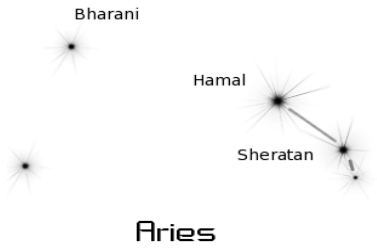
1	2	3	4	5	6	7	8	9	10
Sample	cif	eae	efa1	ehxA	espA	espB	espF	espJ	espP
dataset_1340	2_65*	45_139*	11_153	7_324	22_176	13_190	2_196*	1_210	1_86
dataset_1340	2_65*	45_139*	11_153	7_324	22_176	13_190	2_196*	1_210	1_86
dataset_1339	2_65*	45_139*	11_153	7_324	22_176	13_190	2_196	1_210	3_21

Yes, I picked same sample more than one time... ☹



Why, do not concatenate all?



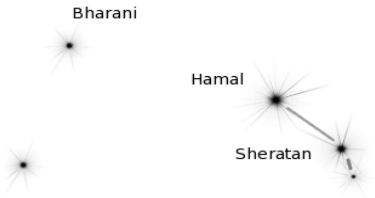


Future Implementation

To improve Scoring Algorhitms

To improve new alleles discovery (working on protein sequence?)
To set parameters for different platforms

agMLST it’s just a starting point, help us to help you…



Let's try it…