
Introduction to STRaM Analysis Workflow in Galaxy

(STRaM Loci Ver. 18)

Li Binglin et al. 2025.05

Part One: STRaM Analysis Workflow Fast Run Protocol

1. Apply for a web or lab Galaxy platform account for login (Figure 1).

Galaxy online: <https://usegalaxy.org/>.

The screenshot shows the Galaxy web platform interface. The top navigation bar includes the Galaxy logo, a search bar, and a 'Login or Register' button. The main content area is divided into two sections. The left section, titled 'Welcome to Galaxy, please log in', contains a login form with fields for 'Public Name or Email Address' and 'Password', and a 'Login' button. A red box labeled 'Login' is around the login form. A red arrow points from the 'Login' button to the 'Click to register' text. The right section, titled 'Create a Galaxy account', contains a registration form with fields for 'Email address', 'Password', 'Confirm password', and 'Public name'. A red box is around the registration form. The page also includes a 'Forgot password? Click here to reset your password.' link and a 'Don't have an account? Register here.' link.

[Galaxy](https://usegalaxy.org/) of Web Platform (<https://usegalaxy.org/>)

Figure 1: Schematic Flow

2. Import STRaM analysis workflow.
 - 2.1 Click "Workflow" in the left menu bar.
 - 2.2 Click "Import" to upload a workflow file.
 - 2.3 Browse and select the latest STRaM workflow file and click "Import workflow".
 - 2.4 The workflow file will be uploaded successfully and displayed.
- Note: Current workflow version is "STRaM analysis workflow_v26.ga".
- See workflow in Figure 2.

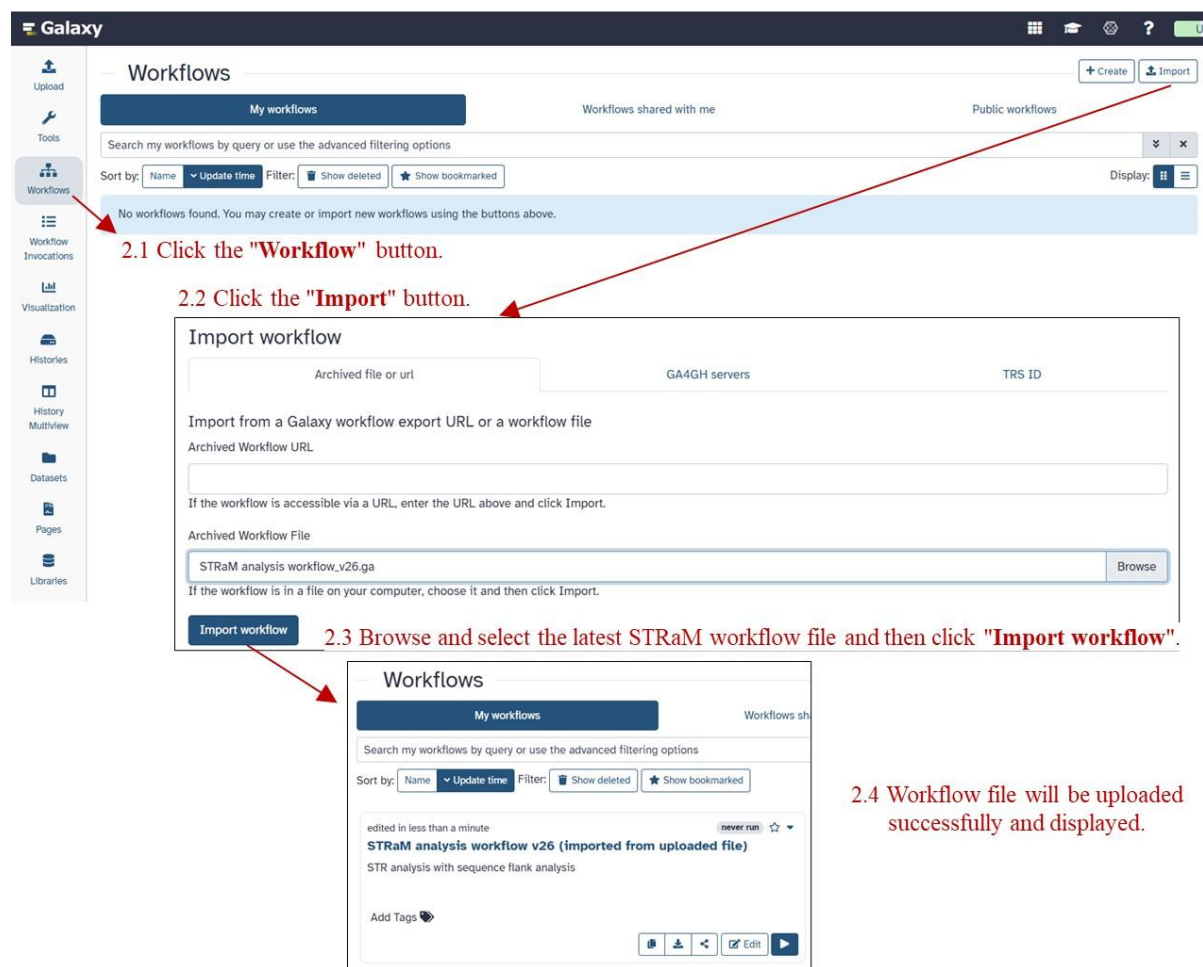


Figure 2: Workflows

3. Upload all datasets into a new history in Galaxy Server

- 3.1 Create a new analysis history and assign a name, such as "STRaM analysis".
- 3.2 Click "Upload" in the left menu bar.
- 3.3 Choose "Local files" to upload datasets from your local drives file.
- 3.4 Click "Start" to begin the upload.
- 3.5 Files will upload successfully and appear in your "STRaM analysis" history.

See schematic flow in Figure 3

Uploaded files: Raw data Paired-end read 1, Raw data Paired-end read 2, Reference human genome hg38 (e.g., GRCh38_no_alt_analysis_set), 3' reference flanking sequences of loci (e.g., STRaM_v18XY_flank3_v1. fasta), 5' reference flanking sequences of loci (e.g., STRaM_v18XY_flank5_v1. fasta), Marker name and genomic locations (e.g., STRaM_markers_v18_list_v21.tabular).

Note: Raw data file should be uploaded before the reference genome file.

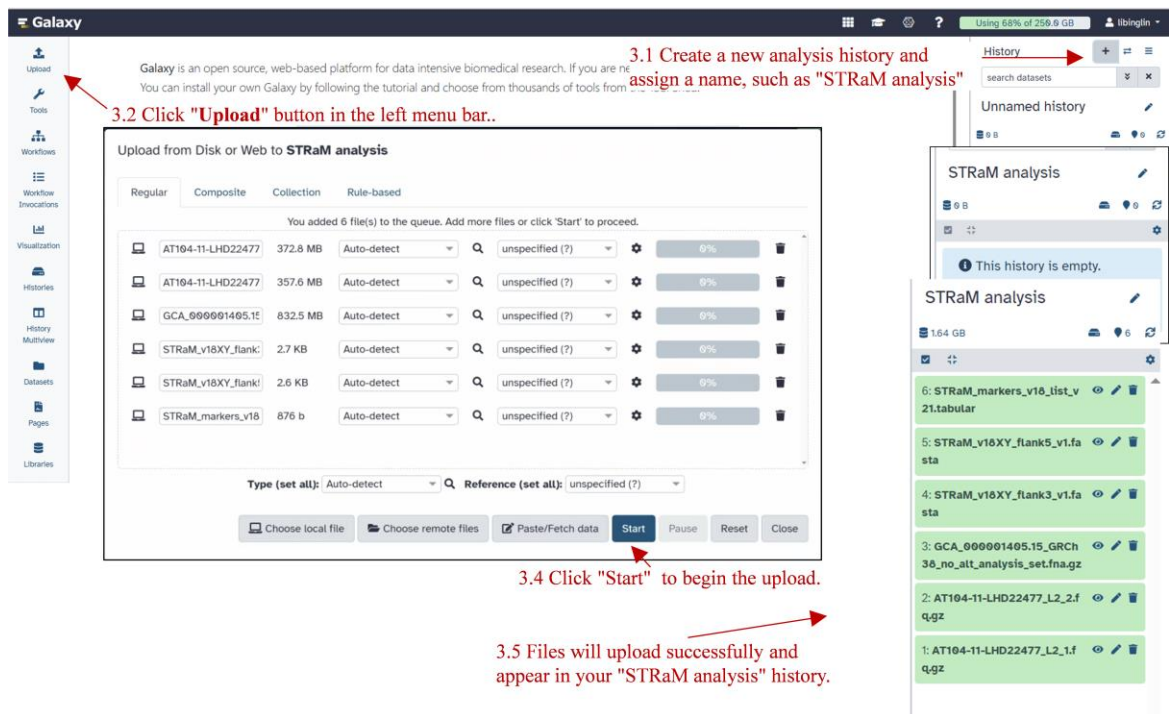


Figure 3

4. Run STRaM analysis workflow

4.1 Click "workflow" and run the uploaded STRaM workflow file (Figure 4).

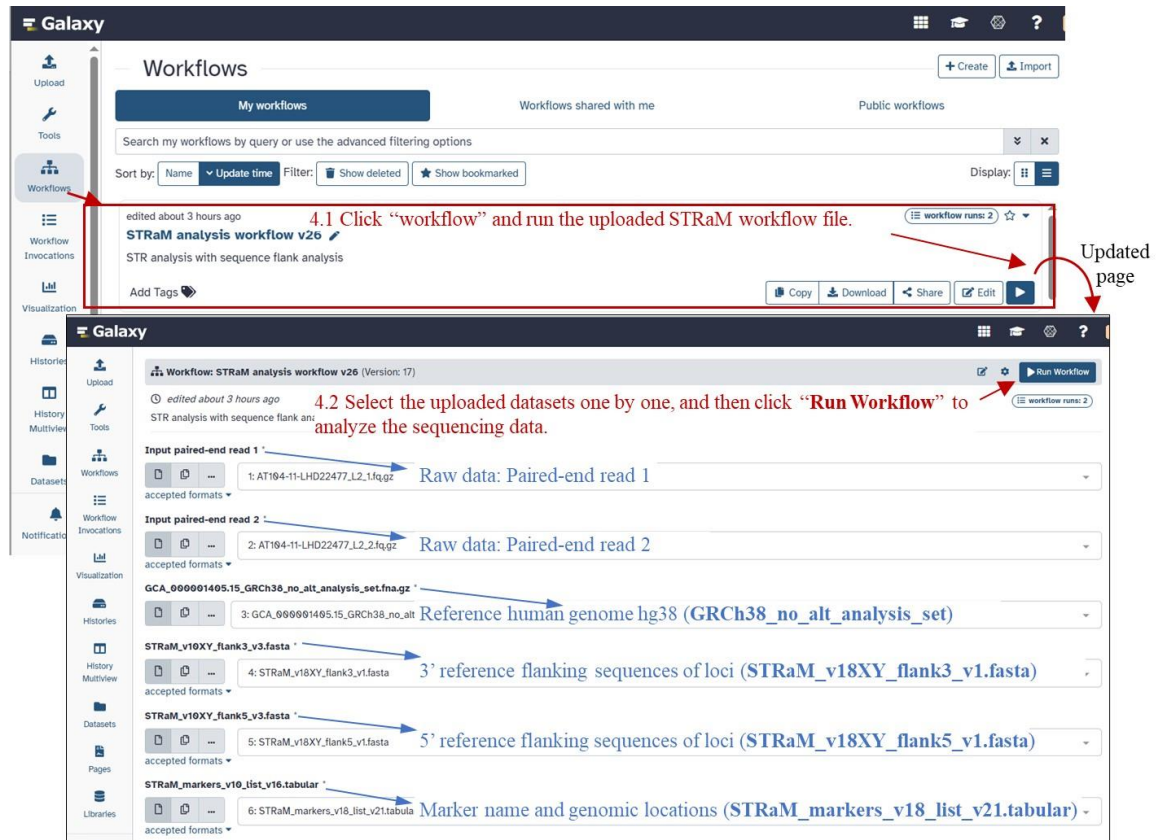


Figure 4

4.2 Select the uploaded datasets one by one, and then click "Run Workflow" to analyze the sequencing data.

5. Running

5.1 Running.

5.2 Troubleshooting.

5.3 Running Completed.

Grey: Pending; Orange: Running; Green: Success; Red: Error.

Note: Since the web platform is updated regularly, some tools may encounter errors.

Troubleshooting will be required for these tools in accordance to the error reports.

See demonstration in Figure 5.

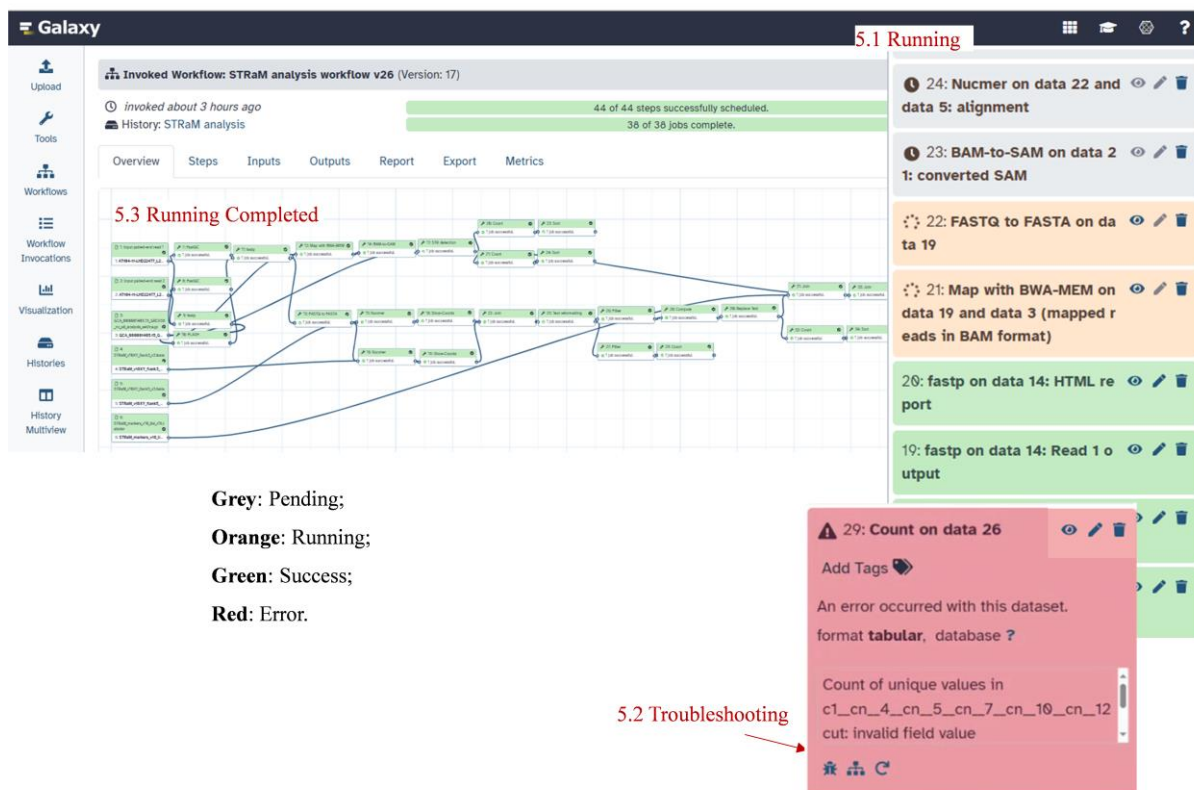


Figure 5: Schematic Flow

6. Result output

6.1 Download the output tabular.

6.2 Perform subsequent analysis to distinguish between allele and stutter signals.

6.3 STRaM profiles collection and sample assessment (See example in Figure 6).

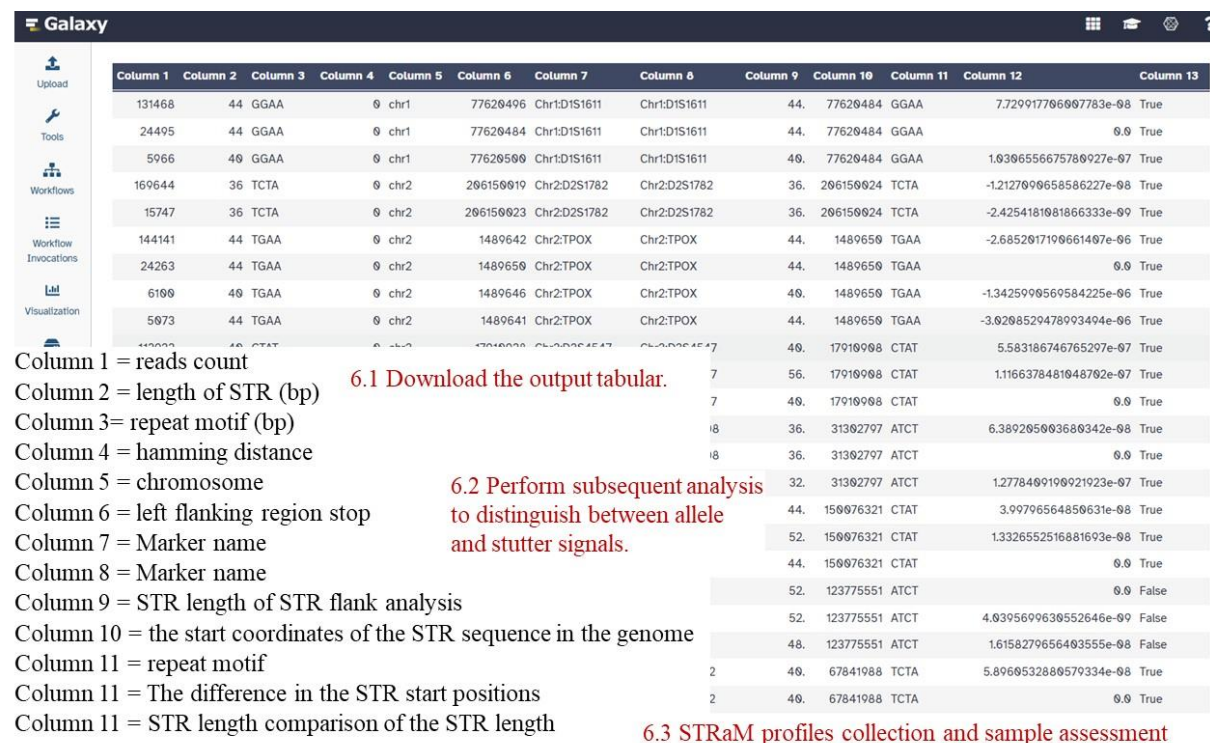


Figure 6

Part Two: Presentation of Tools in STRaM Workflow

1. Data preprocessing

1.1 *FastQC* Read Quality reports (Galaxy Version 0.72+galaxy1)

FastQC offers an efficient solution for performing quality control checks on raw sequence data coming from high throughput sequencing pipelines.

1.2 *Fastp*- fast all-in-one preprocessing for FASTQ files (Galaxy Version 0.20.1+galaxy0) (Figure 7).

1.2.1 Features: *Fastp* filters out bad reads (low quality, too short, or containing too many N...), cut adapters, cut low quality bases for per read, etc. 1.2.2 Parameters set

Qualified quality phred: Inferred base detection accuracy of 99% for Q20 and 99.9% for Q30. Value: 20-30 (int [=25]).

Unqualified percent limit: If the percentage exceeds a certain threshold, the read or pair is discarded. Value 10 means 10% (int [=10]).

N base limit: if one read's number of N base is >n_base_limit, then this read/pair is discarded. Value is 1 (int [=1]).

Disable length filtering: length filtering is enabled by default.

Length required, reads shorter than this value will be discarded. Default is 50. Reads longer than this value will be discarded. Value is 300 (300bp length limit).

fastp - fast all-in-one preprocessing for FASTQ files (Galaxy Version 0.20.1+galaxy0)

Quality filtering options

Disable quality filtering
☐ No
Quality filtering is enabled by default. If this option is specified, quality filtering is disabled. (-Q)

Qualified quality phred - optional
25
The quality value that a base is qualified. Default 15 means phred quality $\geq Q15$ is qualified. (-q)

Unqualified percent limit - optional
10
How many percents of bases are allowed to be unqualified (0~100). Default 40 means 40%. (-u)

N base limit - optional
1
If one read's number of N base is $>n_base_limit$, then this read/pair is discarded. Default is 5. (-n)

Length filtering options

Disable length filtering
☐ No
Length filtering is enabled by default. If this option is specified, length filtering is disabled. (-L)

Length required - optional
50
Reads shorter than this value will be discarded. Default is 15. (-l)

Maximum length - optional
300
Reads longer than this value will be discarded. Default is 0 and means no limitation. (--length_limit)

Figure 7

1.3 **FLASH** adjust length of short reads (Galaxy Version 1.2.11.4) (Figure 8).

1.3.1 **FLASH** (Fast Length Ajustment of Short reads) is an accurate and fast tool to merge paired-end reads from DNA fragments shorter than twice the read length.

FLASH adjust length of short reads (Galaxy Version 1.2.11.4)

Input data structure
Individual datasets

Forward reads *
111: fastp on data 2 and data 1: Read 1 output
accepted formats ▼

Reverse reads *
112: fastp on data 2 and data 1: Read 2 output
accepted formats ▼

Minimum overlap - optional
6
The minimum required overlap length between two reads to provide a confident overlap. (--min-overlap)

Maximum overlap - optional
100
Maximum overlap length expected in approximately 99% of read pairs. Overlaps longer than the maximum overlap parameter are still considered as good overlaps, but the mismatch density is calculated over the first max_overlap bases in the overlapped region rather than the entire overlap. (--max-overlap)

Maximum mismatch density - optional
0.25
Maximum allowed ratio between the number of mismatched base pairs and the overlap length. Two reads will not be combined with a given overlap if that overlap results in a mismatched base density higher than this value. (--max-mismatch-density)

Combine read pairs in both orientations
☐ No
FLASH uses the same parameters when trying each orientation. If a read pair can be combined in either orientation, the better-fitting one will be chosen using the same scoring algorithm that FLASH normally uses. (--allow-outies)

Output a text rendering of the histogram
☐ No

Save FLASH log file
☐ No

Figure 8

1.3.2 Parameters set

Minimum overlap: The minimum required overlap length between two reads for reliable merging, value is 6 (int [=6]);

Maximum overlap: Maximum overlap length refers to the maximum number of overlapping rectangles with at least one common point, value is 100 (int [=100]);

Maximum mismatch density: Maximum allowed ratio between the number of mismatched base pairs and the overlap length. Reads won't merge if their overlap exceeds this mismatch density. value is 0.25 (int [=0.25]).

2. STR analysis

2.1 Map with BWA-MEM: map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.1)

BWA-MEM is an alignment algorithm that maps sequence reads or long query sequences to large reference genomes (e.g., human). It automatically selects between local and end-to-end alignment modes, supports paired-end reads and performs chimeric alignment. All parameters use default settings.

2.2 BAM-to-SAM: convert BAM to SAM (Galaxy Version 2.0.1)

Converts BAM dataset to SAM using the samtools view command.

2.3 STR detection for short read, reference and mapped data (Galaxy Version 1.0.0) (Figure 9).

2.3.1 This tool identifies both simple and interrupted STRs.

STR detection for short read, reference, and mapped data (Galaxy Version 1.0.0)

Select reference file *

3: GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz

Motif size of microsatellites of interest (e.g. Mononucleotide microsatellite =1) (must be less than 10) *

4

Consider microsatellites with a partial motif?

☐ No

Minimal length (bp) of microsatellite sequence reported *

12

Do not report candidate repeat intervals that have left flanking region less than (bp): *

8

Do not report candidate repeat intervals that have right flanking region less than (bp): *

8

Hamming threshold of microsatellite, If greater than 0, interrupted microsatellites will also be reported *

15

Consider all candidate intervals in a sequence. If not check, only the longest one will be considered

☐ No

Show the entire flanking regions

☒ Yes

Additional Options

Email notification

☐ No

Send an email notification when the job completes.

Run Tool

Figure 9

2.3.2 Parameters set

Motif size of microsatellites of interest (e.g. tetranucleotide microsatellite = 4),

Minimal length (bp) of microsatellite sequence reported, value is 12 (int [=12]).

Do not report candidate repeat intervals with left/right flanking region shorter than (bp), value is 8 (int [=8]).

Hamming threshold of microsatellites: A hamming distance > 0 returns both simple and interrupted STRs. Note: Hamming distance (hamming threshold of microsatellites) needs to be adjusted based on STR markers. Example: VWA is a complex structured STR locus, value is 15 (int [=15]).

2.3.3 **Output** (SAM format input)

Column 1 = length of STR (bp)

Column 2 = length of left flanking region (bp)

Column 3 = length of right flanking region (bp)

Column 4 = repeat motif (bp)

Column 5 = hamming distance

Column 6 = read name (**reads ID**)

Column 7 = read sequence with soft masking of STR

Column 8 = read quality (the same Phred score scale as input)

Column 9 = read name (The same as column 6)

Column 10 = chromosome

Column 11 = left flanking region start

Column 12 = left flanking region stop

Column 13 = STR start as infer from pair-end

Column 14 = STR stop as infer from pair-end

Column 15 = right flanking region start

Column 16 = right flanking region stop

Column 17 = STR length in reference

Column 18 = STR sequence in reference

2.4 **Count and Sort**

2.4.1 *Count* occurrences of each record (Galaxy Version 1.0.3)

Without sequence: Column 1, 4, 5, 10, 12;

With sequence: Column 1, 4, 5, 7, 10, 12.

2.4.2 Sort data in ascending or descending order (Galaxy Version 1.1.1)

3. STR flanking analysis

3.1 *Nucmer* Align two or more sequences (Galaxy Version 4.0.0beta2+galaxy1) (Figure 10).

Figure 10

3.1.1 *Nucmer* is for the all-vs-all comparison of nucleotide sequences contained in multi-FastA data files. Run *Nucmer* against 2 STR flanking reference fasta files respectively.

3.1.2 Parameters set

Anchoring use default.

Break Length: Sets the distance an alignment extension will attempt to extend through low-scoring regions before termination, default is 200 (int [=200]).

Minimum Cluster Length: Sets the minimum required length for a match cluster, value is 15 (int [=15]).

Maximum Diagonal Difference: Sets the maximum allowed diagonal difference between adjacent anchors in a cluster, default is 5 (int [=5]).

Maximum Diagonal Difference: Sets the maximum diagonal difference between adjacent anchors as a differential fraction of gap length, default is 0.12 (int [=0.12]).

Direction: Choose a direction of Query Sequence: Use only the forward strand of the Query sequences.

Maximum Gap Distance: Sets the maximum gap between adjacent matches in a cluster, value is 60 (int [=60]).

Minimum Match Length: Sets the minimum length of a single exact match, value is 15 (int [=15]).

Minimum Alignment Length: Minimum length of an alignment, after clustering and extension, value is 0 (int [=0]).

3.2 **Show-Coords**: Parse delta file and report coordinates and other information (Galaxy Version 4.0.0beta2+galaxy1)

3.2.1 Run *Show-Coords* to show all the alignments by *Nucmer* for reads.

3.2.2 Parameters set (Figure 11).

Identity: Sets minimum percent identity to display, set value is 75 (int [=75]).

Minimum Alignment Length: Sets minimum alignment length to display, set value is 15 (int [=15]).

Sorting strategy for output: Sorts output lines by query IDs and coordinates.

3.2.3 Output is tabular (Figure 12).

[S1] Start of the alignment region in the reference sequence

[E1] End of the alignment region in the reference sequence

[S2] Start of the alignment region in the query sequence

[E2] End of the alignment region in the query sequence

[LEN 1] Length of the alignment region in the reference sequence, measured in nucleotides

[LEN 2] Length of the alignment region in the query sequence, measured in nucleotides

[% IDY] Percent identity of the alignment, calculated as (number of exact matches) / ([LEN 1] + insertions in the query)

[LEN R] Length of the reference sequence


[LEN Q] Length of the query sequence

[COV R] Percent coverage of the alignment on the reference sequence, calculated as [LEN 1] / [LEN R]

[COV Q] Percent coverage of the alignment on the query sequence, calculated as [LEN 2] / [LEN Q]





[REF TAG] The reference FastA ID (Marker name)

[QUERY TAG] **The query FastA ID** (Read ID)

 **Show-Coords** Parse delta file and report coordinates and other information (Galaxy Version 4.0.0beta2+galaxy1)

Tool Parameters

Match file from Nucmer *

accepted formats ▼

Merge

☒ No

Merges overlapping alignments regardless of match dir or frame and does not display any identity information. (-b)

Identity *

Set minimum percent identity to display (-I)

Minimum Alignment Length *

Set minimum alignment length to display (-L)

Annotate

☒ No

Annotate maximal alignments between two sequences, i.e. overlaps between reference and query sequences (-o)

Sorting strategy for output *

Additional Options

Email notification

☒ No

Send an email notification when the job completes.


 **Run Tool**

Figure 11

3.3 *Join* two files (Galaxy Version 1.1.2)

This tool merges columns from two flanking alignment files using matching read IDs (The query FastA ID).

3.3.1 Output a new tabular with 25 Column

column 1: **The query FastA ID (Reads ID)** 5' flank alignment

column 2: Start of the alignment region in the reference sequence

column 3: End of the alignment region in the reference sequence

column 4: Start of the alignment region in the query sequence

[S1]	[E1]	[S2]	[E2]	[LEN 1]	[LEN 2]	[% IDY]	[LEN R]	[LEN Q]	[COV R]	[COV Q]	[REF TAG]	[QUERY TAG]
1	30	103	132	30	30	100.00	30	192	100.00	15.62	Chr10:D10S1426	A00599:717:H7CGTDSXF:2:1101:10004:30185
1	30	139	168	30	30	100.00	30	198	100.00	15.15	Chr9:D9S926	A00599:717:H7CGTDSXF:2:1101:10013:17863
1	30	85	114	30	30	100.00	30	136	100.00	22.06	Chr20:D20S482	A00599:717:H7CGTDSXF:2:1101:10013:27258
4	19	74	89	16	16	100.00	30	180	53.33	8.89	Chr16:D16S539-RC	A00599:717:H7CGTDSXF:2:1101:10013:35712
3	18	112	127	16	16	100.00	30	180	53.33	8.89	Chr12:VWA-RC	A00599:717:H7CGTDSXF:2:1101:10013:35712
1	30	118	147	30	30	96.67	30	180	100.00	16.67	Chr12:VWA-RC	A00599:717:H7CGTDSXF:2:1101:10013:35712
1	30	142	171	30	30	100.00	30	191	100.00	15.71	Chr14:D14S306-RC	A00599:717:H7CGTDSXF:2:1101:10013:6026
1	30	97	126	30	30	100.00	30	139	100.00	21.58	Chr6:D6S1282	A00599:717:H7CGTDSXF:2:1101:10013:7999
1	30	89	118	30	30	100.00	30	140	100.00	21.43	Chr20:D20S482	A00599:717:H7CGTDSXF:2:1101:10022:9392
1	30	85	114	30	30	100.00	30	187	100.00	16.04	Chr2:D2S1782	A00599:717:H7CGTDSXF:2:1101:10059:25864
1	30	92	121	30	30	100.00	30	167	100.00	17.96	Chr3:D3S4547	A00599:717:H7CGTDSXF:2:1101:10068:3176
1	30	119	148	30	30	100.00	30	180	100.00	16.67	Chr12:VWA	A00599:717:H7CGTDSXF:2:1101:10068:3834
1	30	117	146	30	30	100.00	30	151	100.00	19.87	Chr3:D3S4547-RC	A00599:717:H7CGTDSXF:2:1101:10077:26835
1	30	111	140	30	30	100.00	30	179	100.00	16.76	Chr18:D18S1358	A00599:717:H7CGTDSXF:2:1101:10077:7513
1	30	88	117	30	30	100.00	30	138	100.00	21.74	Chr17:D17S1298	A00599:717:H7CGTDSXF:2:1101:10086:36652
1	30	78	107	30	30	100.00	30	170	100.00	17.65	Chr21:D21S1409	A00599:717:H7CGTDSXF:2:1101:10104:12727
1	30	113	142	30	30	100.00	30	195	100.00	15.38	Chr11:D11S2364	A00599:717:H7CGTDSXF:2:1101:10113:29684
1	30	90	119	30	30	100.00	30	156	100.00	19.23	Chr16:D16S539	A00599:717:H7CGTDSXF:2:1101:10113:31626
1	20	138	157	20	20	100.00	30	202	66.67	9.90	Chr14:D14S306-RC	A00599:717:H7CGTDSXF:2:1101:10122:17707

Figure 12

column 5: End of the alignment region in the query sequence

....

column 13: The reference FastA ID (12, Marker name) 3' flank alignment

column 14: Start of the alignment region in the reference sequence

column 15: End of the alignment region in the reference sequence

column 16: Start of the alignment region in the query sequence

...

column 25: The reference FastA ID (12, Marker name)

3.4 *Text reformatting* with awk (Galaxy Version 1.1.2)

Run *Text reformatting* with awk to remove reads with mismatched 5' flanking and 3' flanking alignment. AWK Program: {if(\$13 == \$25){print \$0}}.

3.5 *Filter* data on any column using simple expressions (Galaxy Version 1.1.1)

3.5.1 Select Amelogenin X/Y reads using filter: c3>30 and c15>30. Count reads by column 13 or 25 (Marker name).

3.5.2 Remove reads unaligned to both STR ends using filter: c3==30 and c14==1.

3.6 *Compute* an expression on every row (Galaxy Version 1.6)

Calculate the STR length of STR flanking analysis by invoking the *Compute* expression on every row. Add the expression: c16-c5-1 and the output **column 26** STR length with schematic as shown (Figure 13).

3.7 *Replace* Text in entire line (Galaxy Version 1.1.2)

Remove -RC tags for the names of reverse-complement flanking sequences. Find pattern: "-RC" replace with: " " (blank).

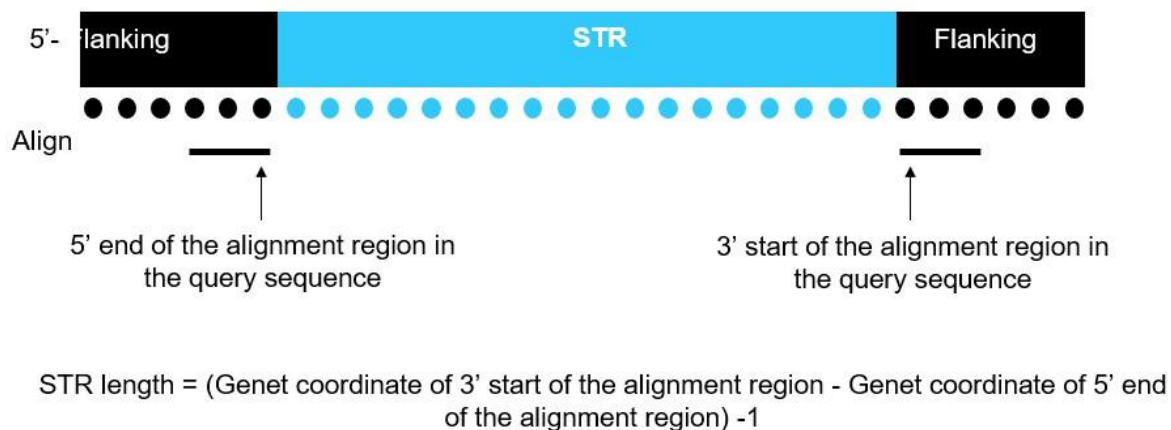


Figure 13: Calculation of STR length in the STR flanking analysis

3.8 *Count* (Galaxy Version 1.0.3) and *Sort* (Galaxy Version 1.1.1)

3.8.1 *Count* STR reads by column 13 (or column 25) and column 26

Note: Marker name: Column 13 and column 25; STR length: column 26.

3.8.2 *Sort* STR read counts according to chromosomes and markers.

3.9 *Join* two files (Galaxy Version 1.1.2)

Join the Step 3.7 output tabular and "**Marker name and genomic locations**" (**STRaM_markers_v18_list_v21.tabular**) by **Marker name**.

3.9.1 The Marker name now is column 1, The query FastA ID (**Reads ID**) now is column 2.

3.9.2 Three new columns have been added, column 27 for the chromosome, column 28 for the start coordinates of the STR sequence in the genome, and column 29 for the repeat motif.

4. Error-sensing analysis

4.1 *Join* two files (Galaxy Version 1.1.2)

Join the output tabular of Step 2.3.3 (STR detection, Column 6) and Step 3.9 (STR flanking analysis) by **Reads ID**.

Output a new tabular with 46 Columns

Column 1 = read name (**reads ID**)

STR analysis

Column 2 = length of STR (bp)

Column 3 = length of left flanking region (bp)

Column 4 = length of right flanking region (bp)

Column 5 = repeat motif (bp)

Column 6 = hamming distance

Column 7 = read sequence with soft masking of STR

Column 8 = read quality (the same Phred score scale as input)

Column 9 = read name (**reads ID**)

Column 10 = chromosome

Column 11 = left flanking region start

Column 12 = left flanking region stop

Column 13 = STR start as infer from pair-end

Column 14 = STR stop as infer from pair-end

Column 15 = right flanking region start

Column 16 = right flanking region stop

Column 17 = STR length in reference

Column 18 = STR sequence in reference

Column 19 = Marker name

STR flanking analysis

5' flanking sequence align information

Column 20 = Start of the alignment region in the reference sequence

Column 21 = End of the alignment region in the reference sequence

Column 22 = Start of the alignment region in the query sequence

Column 23 = End of the alignment region in the query sequence

Column 24 = Length of the alignment region in the reference sequence, measured in nucleotides

Column 25 = Length of the alignment region in the query sequence, measured in nucleotides

Column 26 = Percent identity of the alignment, calculated as (number of exact matches) / ([LEN 1] + insertions in the query)

Column 27 = Length of the reference sequence

Column 28 = Length of the query sequence

Column 29 = Percent coverage of the alignment on the reference sequence, calculated as [LEN 1] / [LEN R]

Column 30 = Percent coverage of the alignment on the query sequence, calculated as [LEN 2] / [LEN Q]

3' flanking sequence align information

Column 31 = Start of the alignment region in the reference sequence

Column 32 = End of the alignment region in the reference sequence

Column 33 = Start of the alignment region in the query sequence

Column 34 = End of the alignment region in the query sequence

Column 35 = Length of the alignment region in the reference sequence, measured in nucleotides

Column 36 = Length of the alignment region in the query sequence, measured in nucleotides

Column 37 = Percent identity of the alignment, calculated as (number of exact matches) / ([LEN 1] + insertions in the query)

Column 38 = Length of the reference sequence

Column 39 = Length of the query sequence

Column 40 = Percent coverage of the alignment on the reference sequence, calculated as [LEN 1] / [LEN R]

Column 41 = Percent coverage of the alignment on the query sequence, calculated as [LEN 2] / [LEN Q]

Column 42 = Marker name

Column 43 = STR length of STR flank analysis

Column 44 = chromosome

Column 45 = the start coordinates of the STR sequence in the genome

Column 46 = repeat motif

4.2 *Count* (Galaxy Version 1.0.3)

Count reads by column 2, 5, 6, 10, 12, 19, 42, 43, 45, 46

Output a new tabular with 11 Columns

Column 1 = reads count

Column 2 = length of STR (bp)

Column 3= repeat motif (bp)

Column 4 = hamming distance

Column 5 = chromosome

Column 6 = left flanking region stop

Column 7 = Marker name

Column 8 = Marker name

Column 9 = STR length of STR flank analysis

Column 10 = the start coordinates of the STR sequence in the genome

Column 11 = repeat motif

4.3 Compare the information of STR and STR flanking analyses

Compute an expression on every row (Galaxy Version 1.6) or download the Count file (4.2) and import it into Excel for data evaluation.

4.3.1 Genomic coordinates: The difference in the STR start position (DSPs) "(c6-c10)/300"

4.3.2 STR length comparison of the STR length (CSL) between the STR analysis and the STR flanking analysis. "c2==c9".

4.3.3 Read counts extracted independently by the STR analysis (Step 2.4.2) and the STR flanking analysis (Step 3.8.2) from separately merged reads.

5. Sample assessment

5.1 STR profile contains loci with their repeat unit numbers, followed by similarity index (SI) analysis.

5.2 Read count distribution per locus determines qualification status, with qualified locus percentage calculated as purity index (PI).

5.3 Relative level/percentage analysis for targeted gene sequences serves as editing/mutation index (EMI).