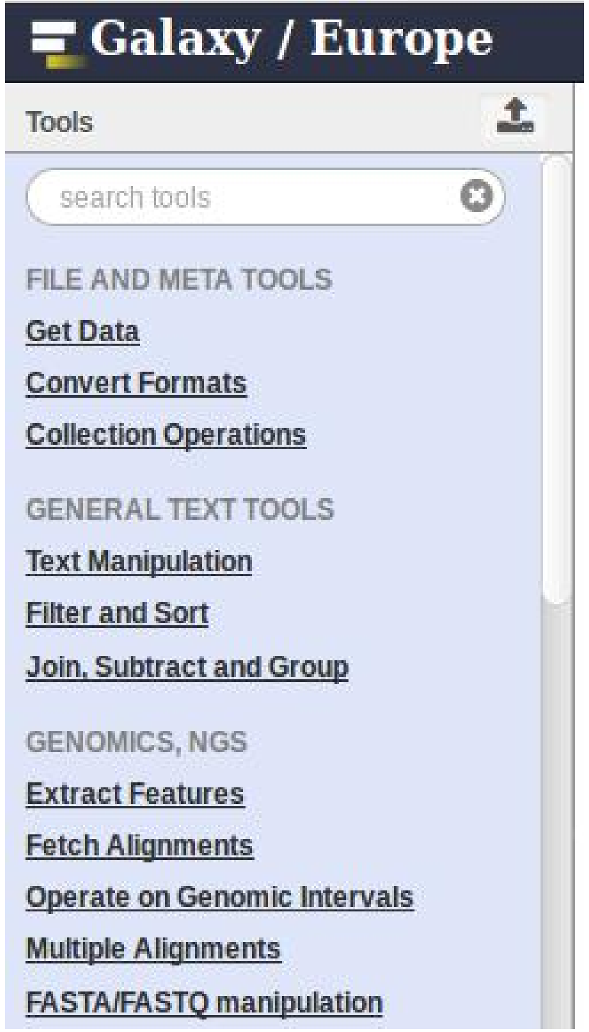
Find similarity in Galaxy tools and predict next tools in workflows

(**Master's thesis**)

Anup Kumar

Find similarity in Galaxy tools

Similarity in tools using machine learning (ML) and natural language processing (NLP) approaches



Galaxy / Europe

Tools

search tools

FILE AND META TOOLS

Get Data

Convert Formats

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

GENOMICS, NGS

Extract Features

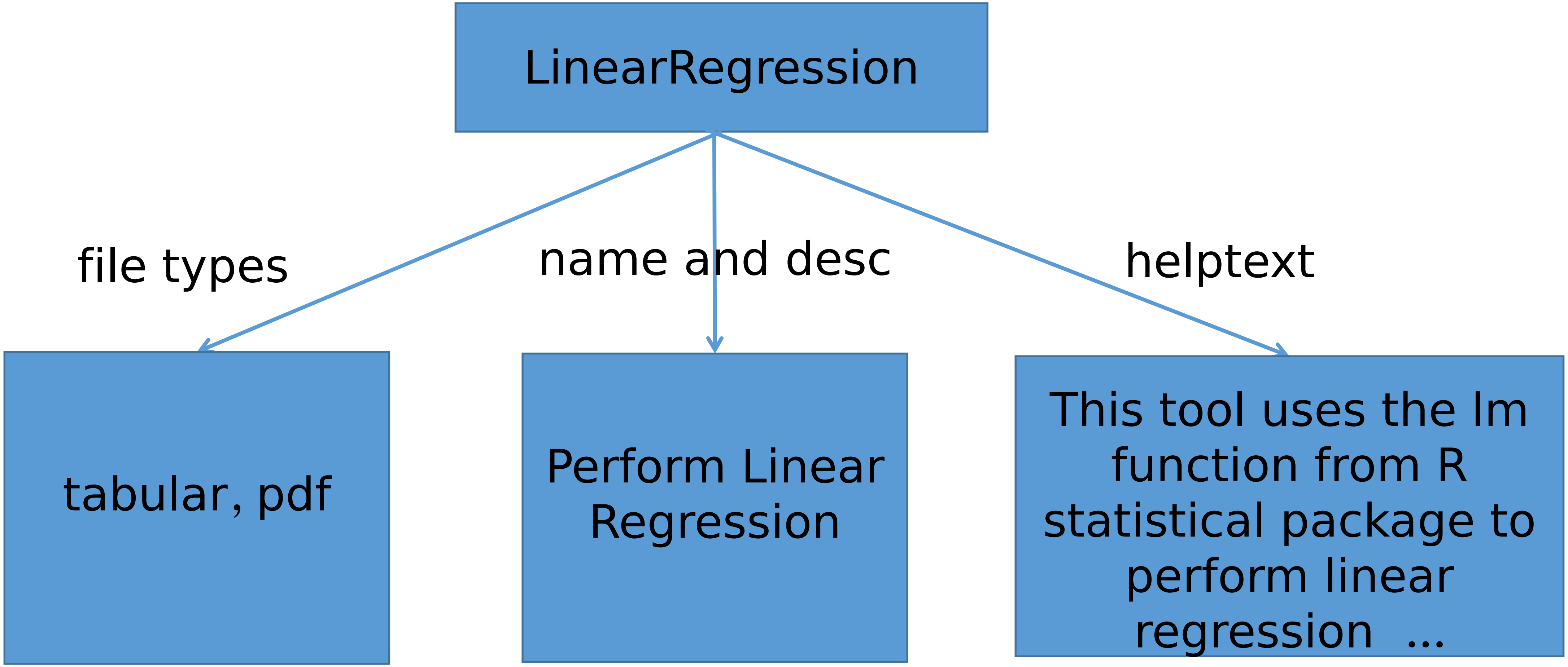
Fetch Alignments

Operate on Genomic Intervals

Multiple Alignments

FASTA/FASTQ manipulation

Tool’s attributes



LinearRegression

file types

name and desc

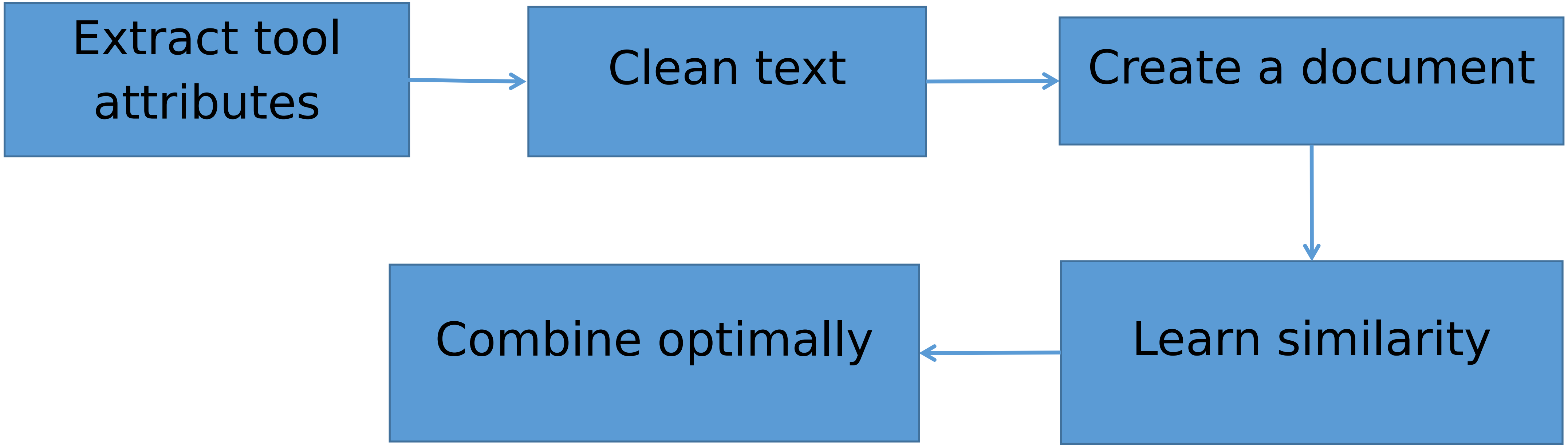
helptext

tabular, pdf

Perform Linear Regression

This tool uses the lm function from R statistical package to perform linear regression ...

Approach


Extract tool attributes
Clean text
Create a document
Learn similarity
Combine optimally

Example

Attributes/ Tools	LinearRegression	LogisticRegression	Similarity
Input, output	**‘pdf‘ , ‘tabular‘**	**‘tabular‘**	?
Name, description	**‘regress‘ , ‘linear‘ , ‘perform‘**	**‘logist‘ , ‘regress‘ , ‘perform‘**	?
Help text	**‘regress‘ , ‘assumpt‘ , ‘lm‘ , ‘statist‘ ,‘linear‘** …	**‘vif‘ ,‘regress‘ , ‘glm‘ ,‘car‘ ,‘inflat‘ , ‘function‘ ,‘statist‘ , ‘logist‘** …	?

Compute similarity

- Jaccard’s distance for input/output
- Dense vector for name, description and helptext
- Example:
- [‘**regress**‘, ‘**linear**‘, ‘**perform**‘] = [0.98, 0.07, ... , 0.12]

Similarity matrix (for name, desc.)

Tools	LinearRegression	LogisticRegression	BestSubsets Regression	lda_analy
LinearRegression	1	0.88	0.84	0.86
LogisticRegression	.88	1	0.82	0.65
BestSubsetsRegression	0.84	0.82	1	0.62
lda_analy	…..	….	….	1

How to combine？

- ３ similarity matrices，one for each attribute
- How to combine them？Take average？
- Optimal combination，learn weights for each tool
- Similarity：

$$\underset{(w_i,\ldots,w_n)}{\arg\max} \sum_{i=1}^{N} w_i \cdot s_i$$

Example

- Tool: LinearRegression
- Similarity for input/output: **sim_io** = [0.33, 0.5, 1.0, ...]
- Similarity for name, desc: **sim_nd** = [0.83, 0.09, 0.005, ...]
- Similarity for helptext: **sim_ht** = [0.45, 0.36, 0.001 .....]
- Similarity =

  argmax(w1 × sim_io + w2 × sim_nd + w3 × sim_ht)

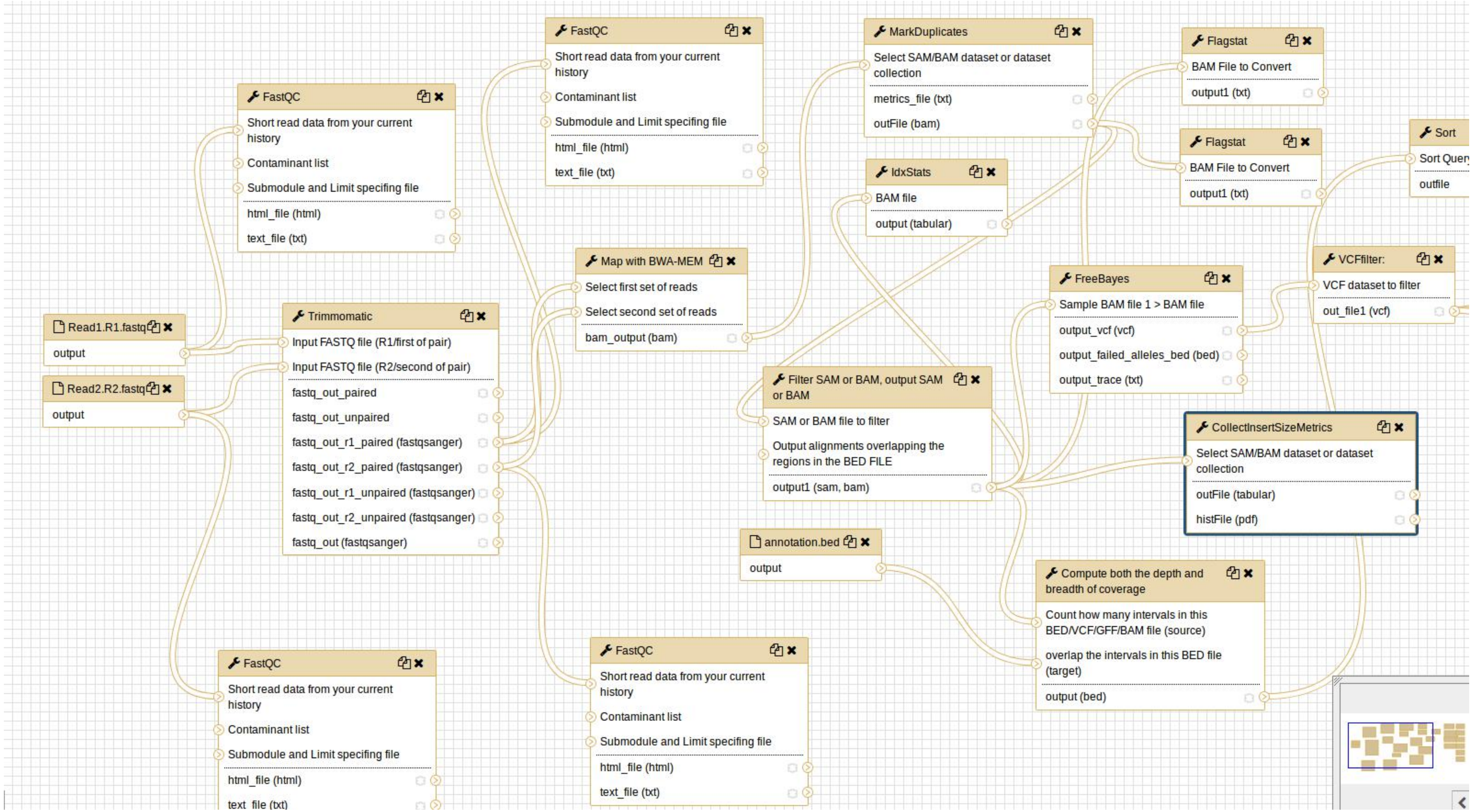
  w1 + w2 + w3 = 1

Visualizer and References

- Static website: results for ~ 1000 tools
- https://rawgit.com/anuprulez/similar_galaxy_tools/master/viz/similarity_viz.html
- https://github.com/anuprulez/similar_galaxy_tools
- https://cs.stanford.edu/%7Equocle/paragraph_vector.pdf
- https://arxiv.org/pdf/1607.05368.pdf

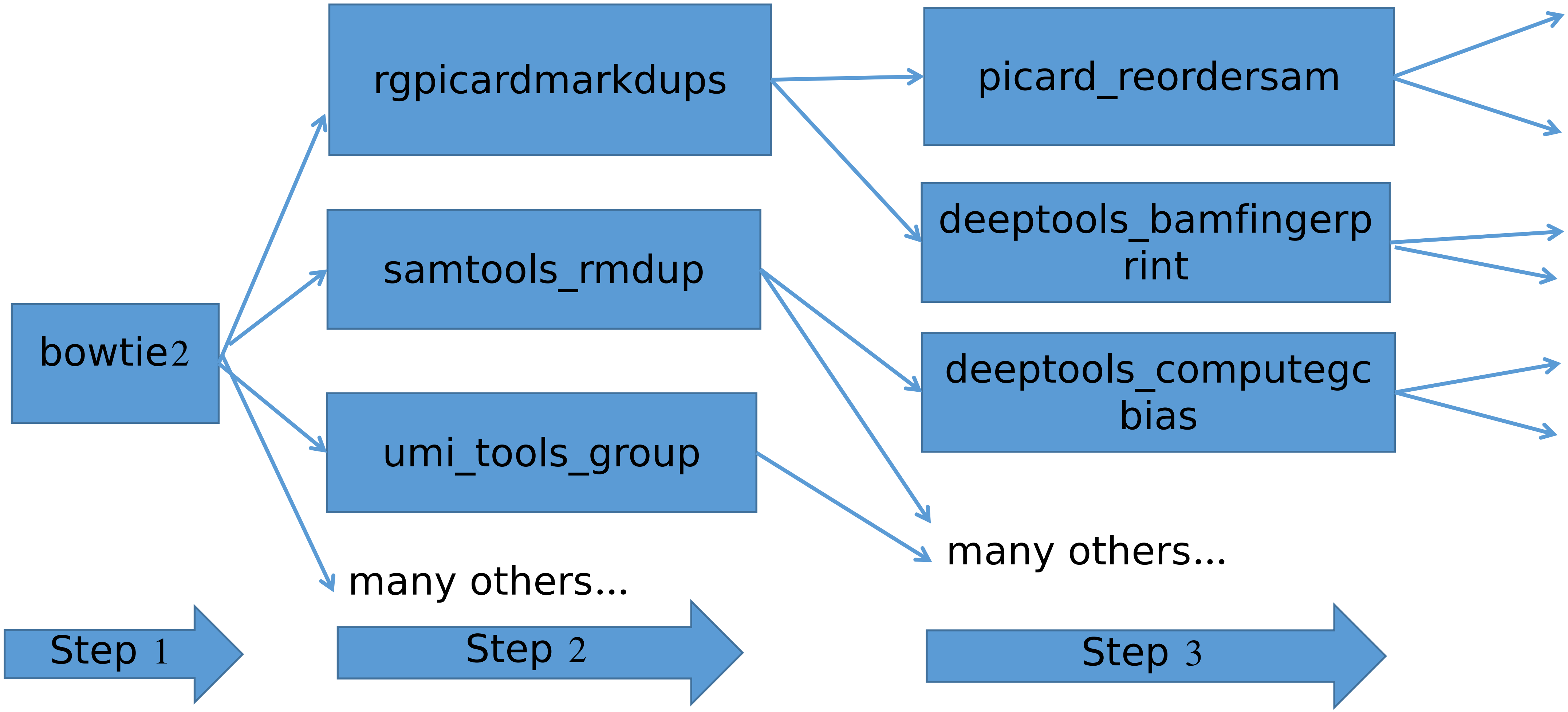
Predict next tools in Galaxy workflows

Galaxy workflow



Read1.R1.fastq
output
Read2.R2.fastq
output
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Trimmomatic
Input FASTQ file (R1/first of pair)
Input FASTQ file (R2/second of pair)
fastq_out_paired
fastq_out_unpaired
fastq_out_r1_paired (fastqsanger)
fastq_out_r2_paired (fastqsanger)
fastq_out_r1_unpaired (fastqsanger)
fastq_out_r2_unpaired (fastqsanger)
fastq_out (fastqsanger)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Map with BWA-MEM
Select first set of reads
Select second set of reads
bam_output (bam)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
MarkDuplicates
Select SAM/BAM dataset or dataset collection
metrics_file (txt)
outFile (bam)
IdxStats
BAM file
output (tabular)
Filter SAM or BAM, output SAM or BAM
SAM or BAM file to filter
Output alignments overlapping the regions in the BED FILE
output1 (sam, bam)
annotation.bed
output
Flagstat
BAM File to Convert
output1 (txt)
Flagstat
BAM File to Convert
output1 (txt)
FreeBayes
Sample BAM file 1 > BAM file
output_vcf (vcf)
output_failed_alleles_bed (bed)
output_trace (txt)
CollectInsertSizeMetrics
Select SAM/BAM dataset or dataset collection
outFile (tabular)
histFile (pdf)
Compute both the depth and breadth of coverage
Count how many intervals in this BED/VCF/GFF/BAM file (source)
overlap the intervals in this BED file (target)
output (bed)
Sort
Sort Query
outfile
VCFfilter:
VCF dataset to filter
out_file1 (vcf)

Next tools？



bowtie2

rgpicardmarkdups

samtools_rmdup

umi_tools_group

many others...

picard_reordersam

deeptools_bamfingerprint

deeptools_computegcbias

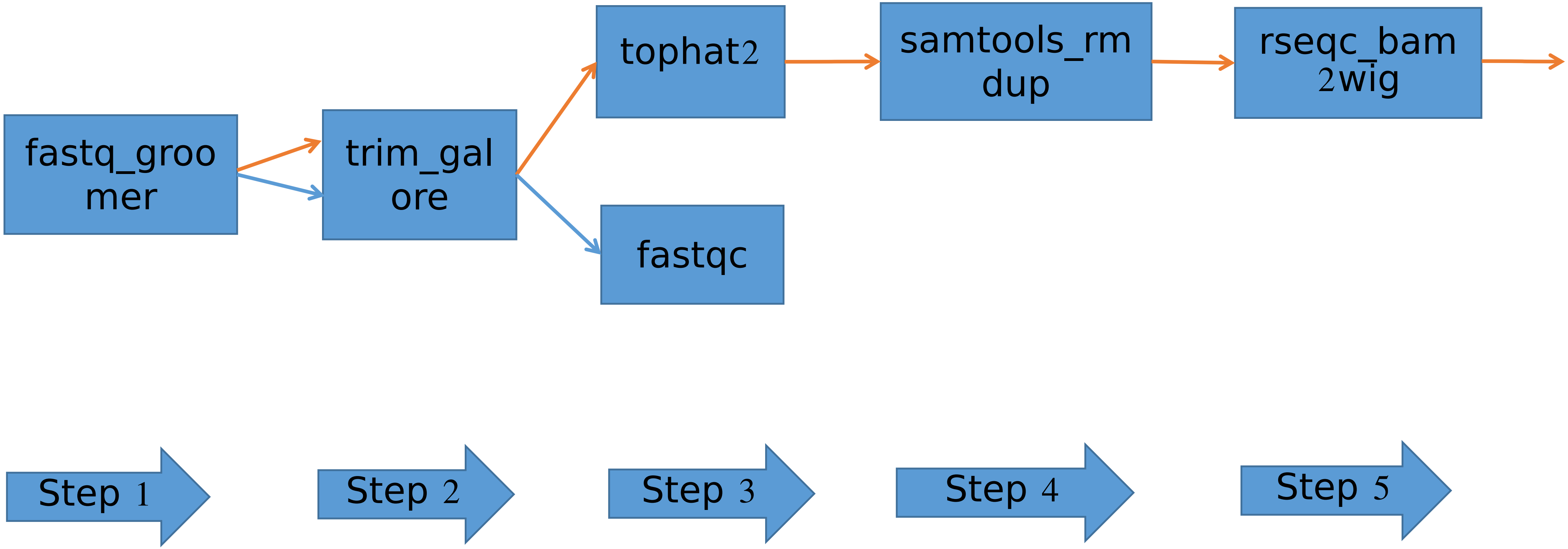
many others...

Step 1

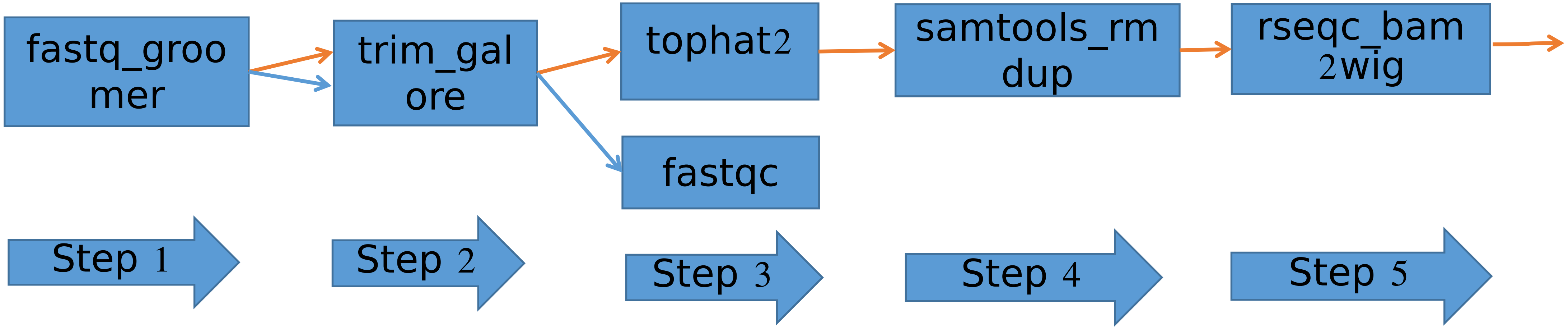
Step 2

Step 3

Workflow as a sequence



Data preprocessing

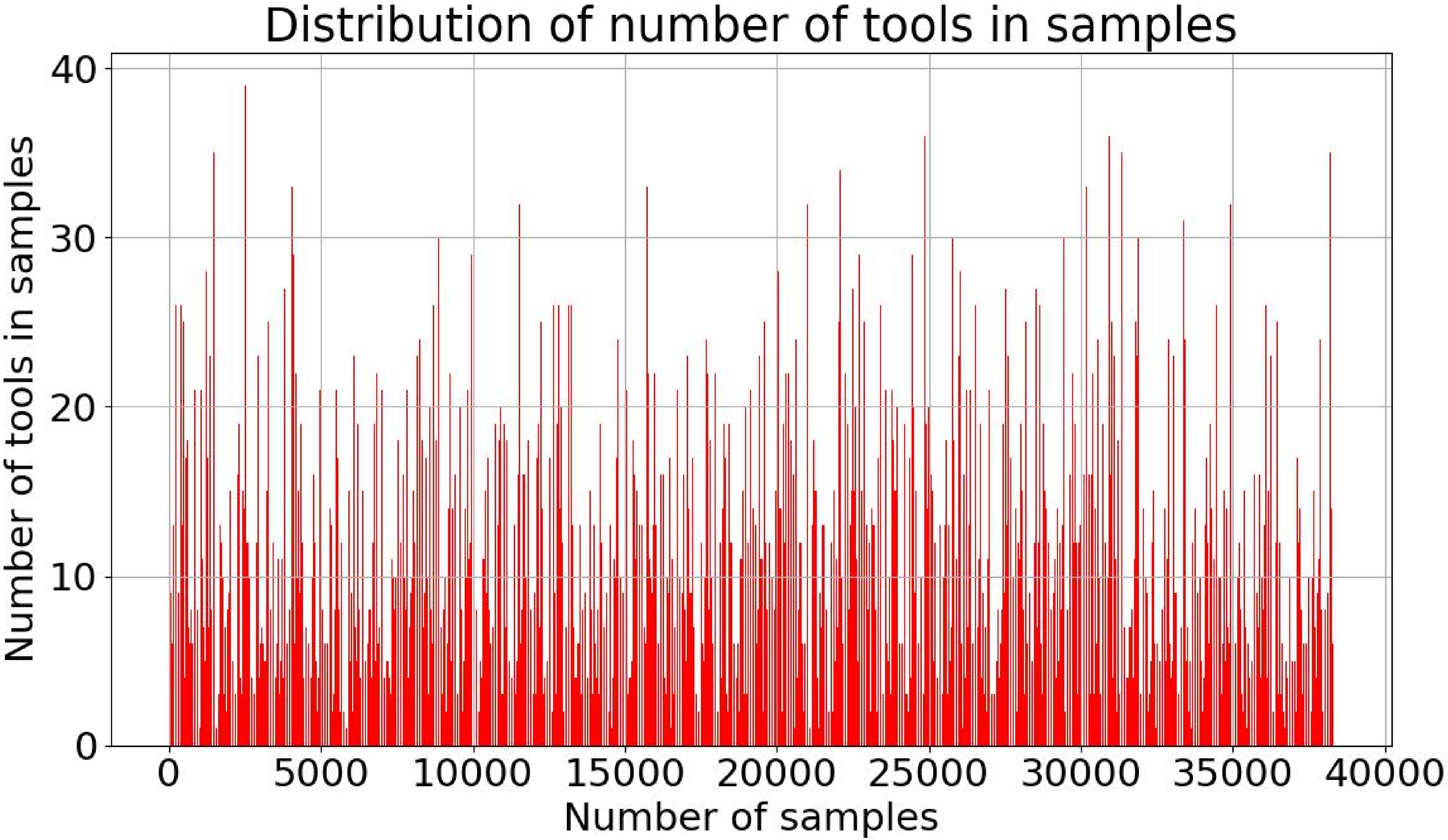


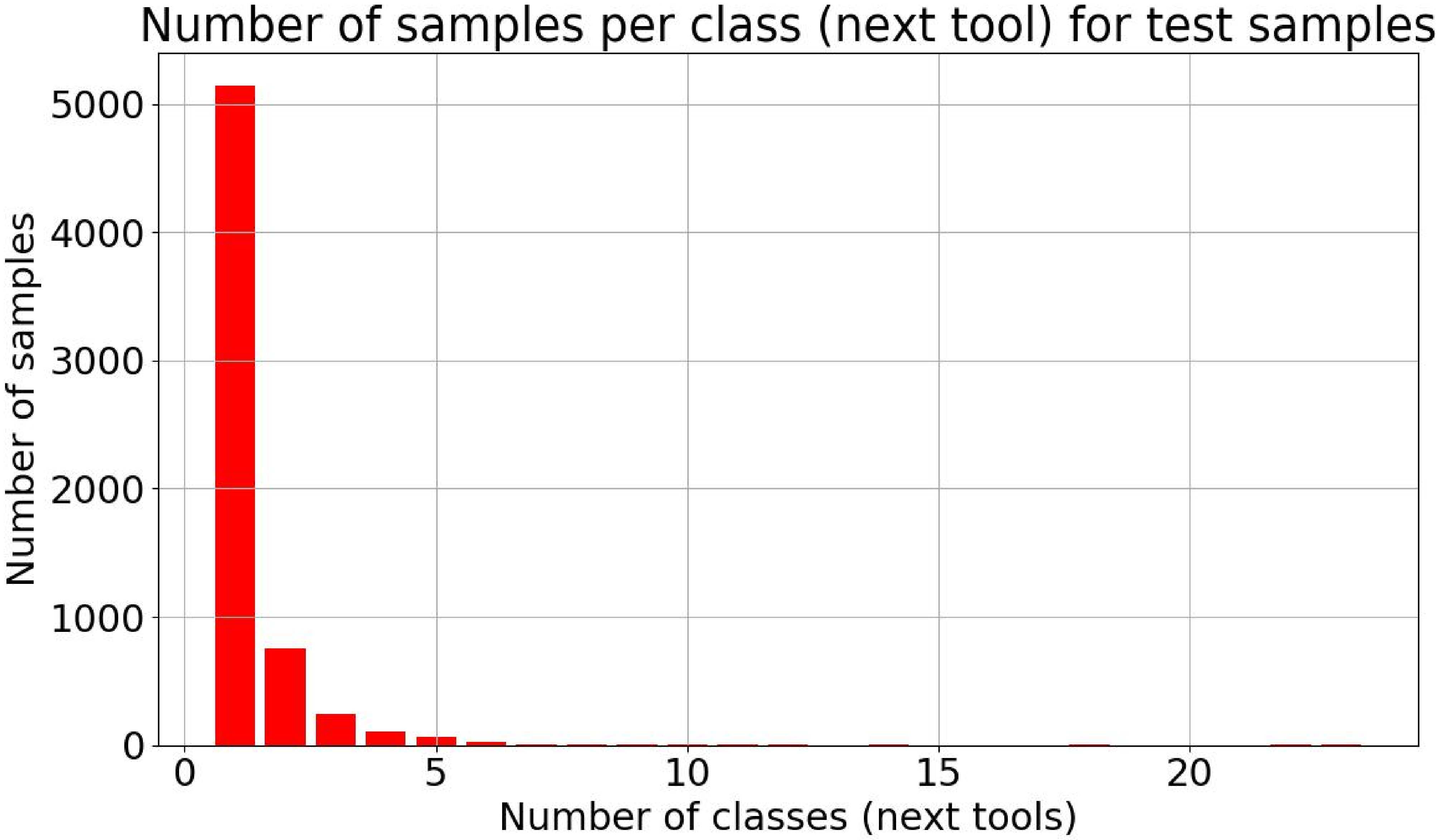
- fastq_groomer, trim_galore (Step 1)
- fastq_groomer, trim_galore, tophat2, fastqc (Step 2)
- fastq_groomer, trim_galore, tophat2, samtools_rmdup (Step 3)
- fastq_groomer, trim_galore, tophat2, samtools_rmdup, rseqc_bam2wig (Step 4)

Data preprocessing

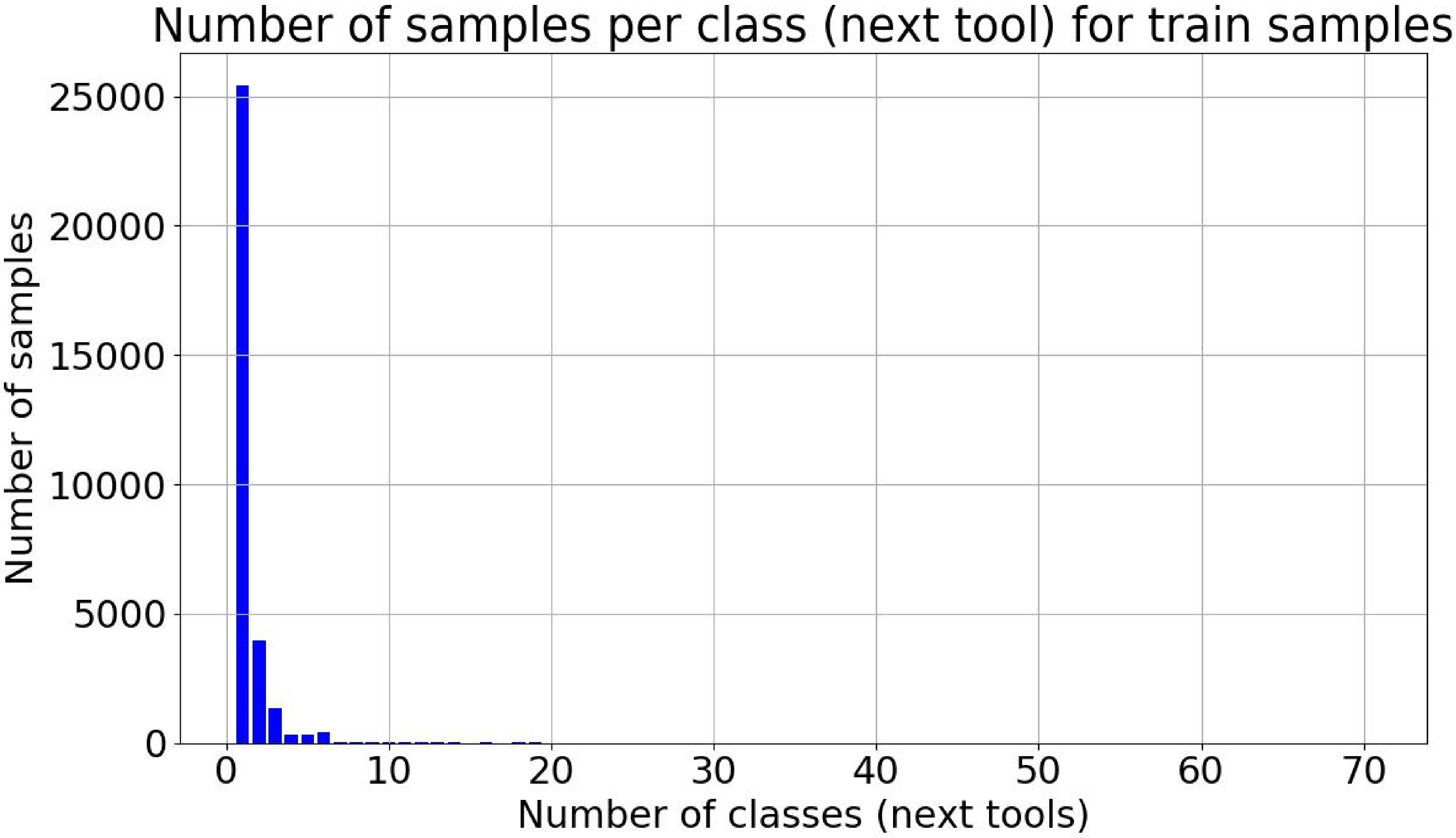
- Assign a unique index to each tool
- {‘fastq_groomer‘: 1, ‘trim_galore‘: 2, ‘tophat2‘: 3, ‘samtools_rmdup‘: 4, ‘rseqc_bam2wig‘: 5, ‘fastqc‘: 6}
- Training samples:

Training sample	Label (next tools)
fastq_groomer, trim_galore	tophat2, fastqc
1 ,2	3, 6
1 ,2 ,3	4


Distribution of number of tools in samples
Number of tools in samples
0
10
20
30
40
Number of samples
0
5000
10000
15000
20000
25000
30000
35000
40000

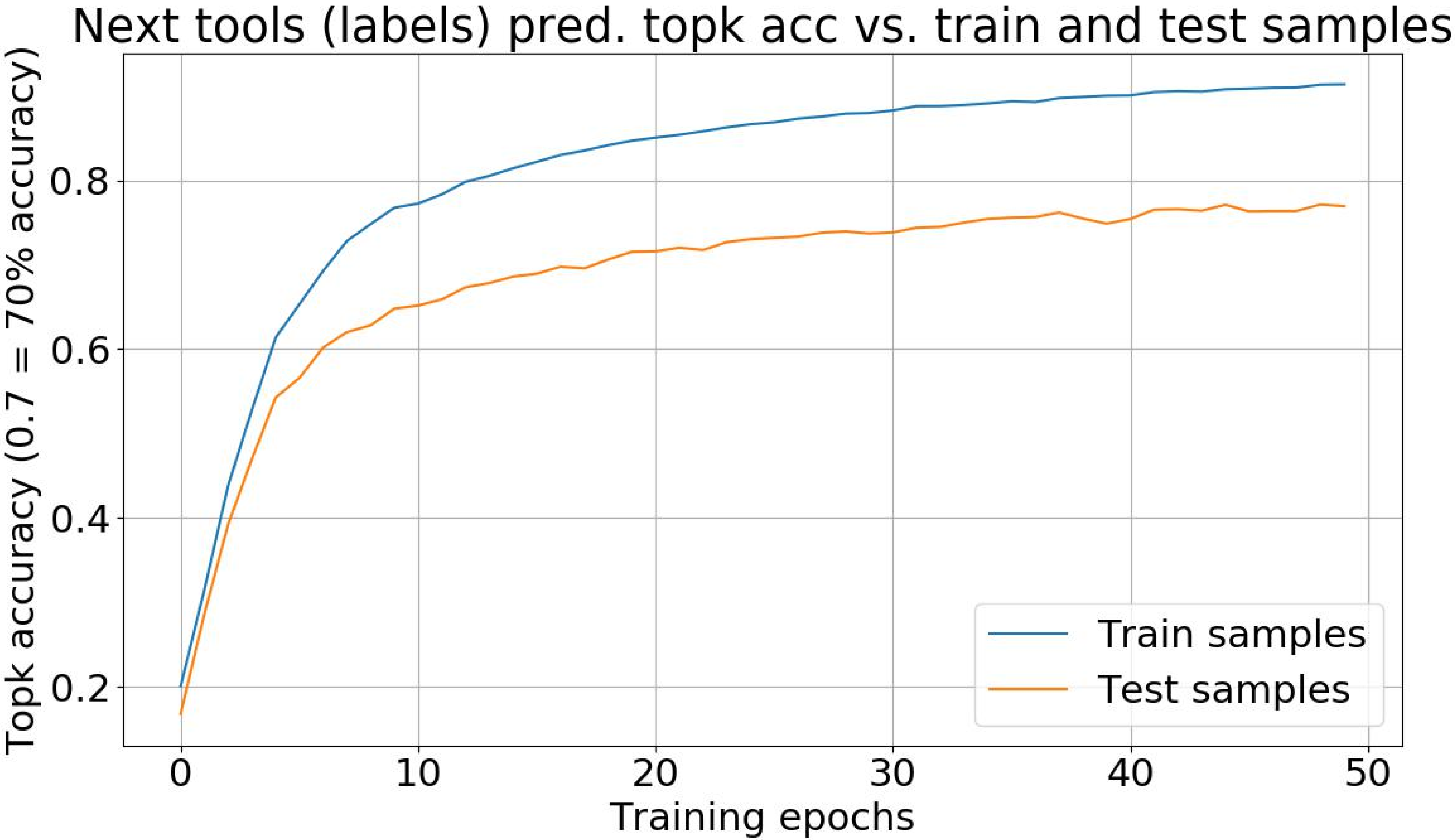


Number of samples per class (next tool) for test samples
Number of samples
5000
4000
3000
2000
1000
0
0
5
10
15
20
Number of classes (next tools)

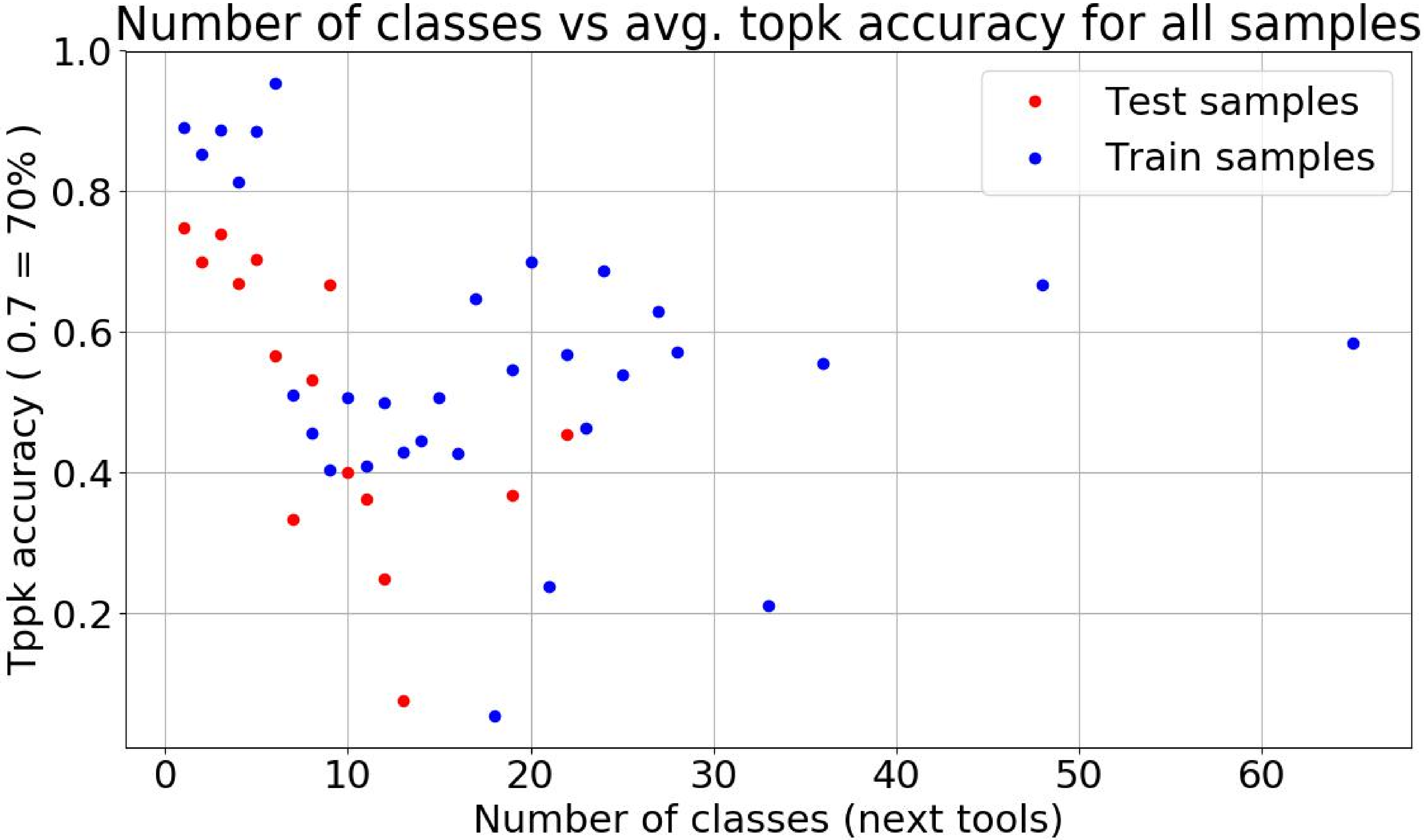

Number of samples per class (next tool) for train samples
Number of samples
25000
20000
15000
10000
5000
0
Number of classes (next tools)
0
10
20
30
40
50
60
70

Classification

- Multi label, multi class classification
- Long range dependencies samples
- Long short term memory (LSTM) networks
- Topk accuracy
- # of top k tools in actual k next tools ÷ k actual next tools



Next tools (labels) pred. topk acc vs. train and test samples
Topk accuracy (0.7 = 70% accuracy)
0.2
0.4
0.6
0.8
0
10
20
30
40
50
Training epochs
Train samples
Test samples


Number of classes vs avg. topk accuracy for all samples
Tppk accuracy (0.7 = 70%)
Number of classes (next tools)
● Test samples
● Train samples
1.0
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60

Next steps

- Convolution
- Balance the samples
- Different activations
- Compatibility constraint

References

- https://github.com/anuprulez/similar_galaxy_workflow
- https://arxiv.org/pdf/1511.03677.pdf
- https://arxiv.org/pdf/1604.04573.pdf
- https://arxiv.org/pdf/1506.00019.pdf

Thank you for your attention