Find similarity in Galaxy tools and predict next tools in workflows

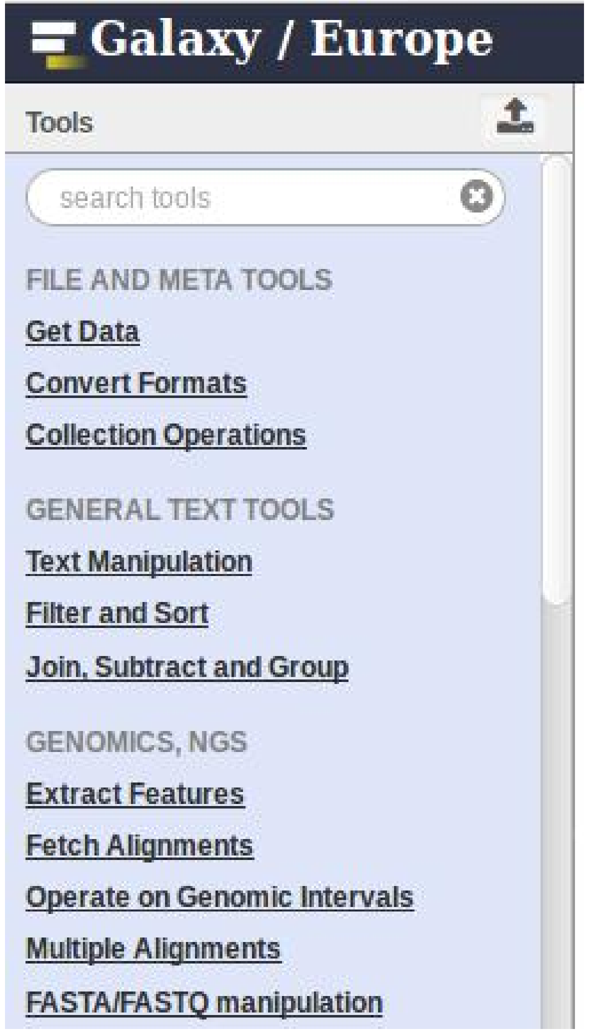
(**Master’s thesis**)

Anup Kumar

Find similarity in Galaxy tools

Machine learning (ML) and natural language processing (NLP) approaches

- Paragraph Vectors
- Gradient Descent



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

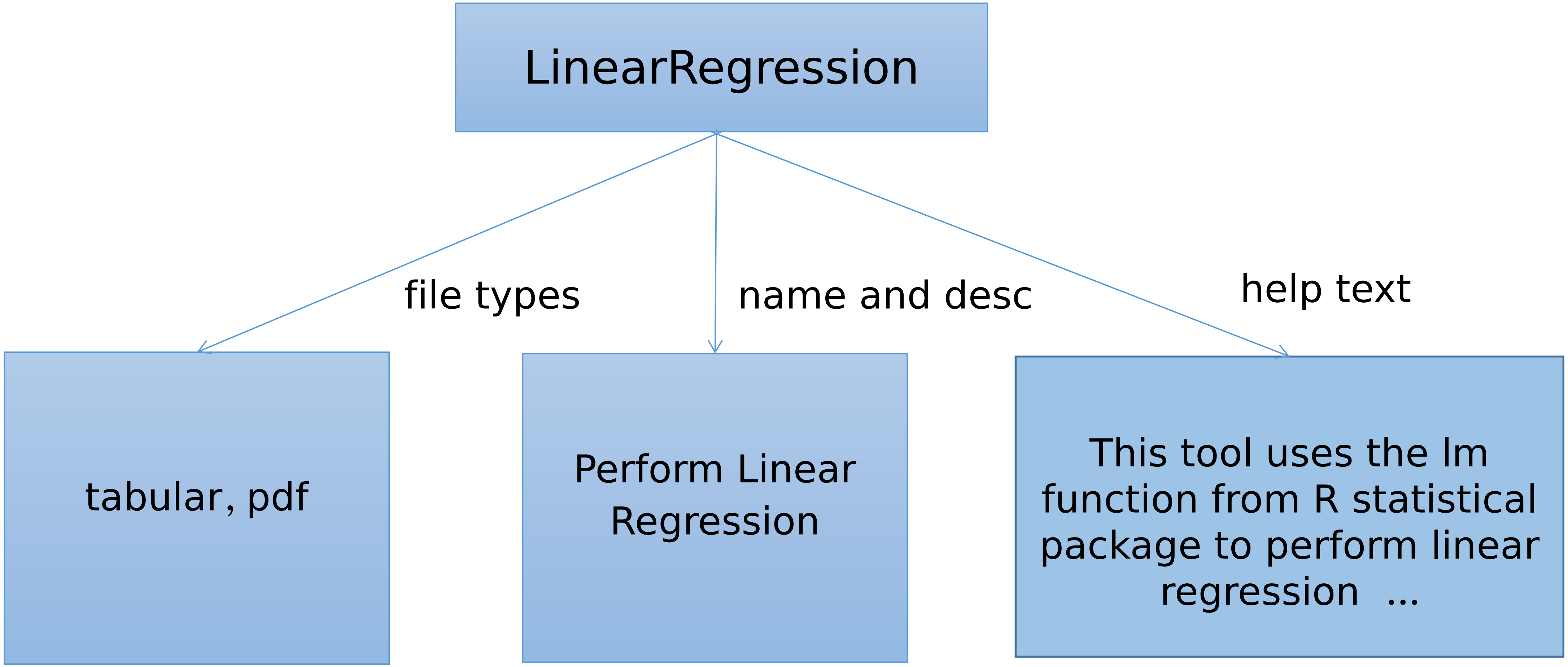
Fetch Alignments

Operate on Genomic Intervals

Multiple Alignments

FASTA/FASTQ manipulation

Tool's attributes



LinearRegression

file types

name and desc

help text

tabular, pdf

Perform Linear Regression

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

Approach

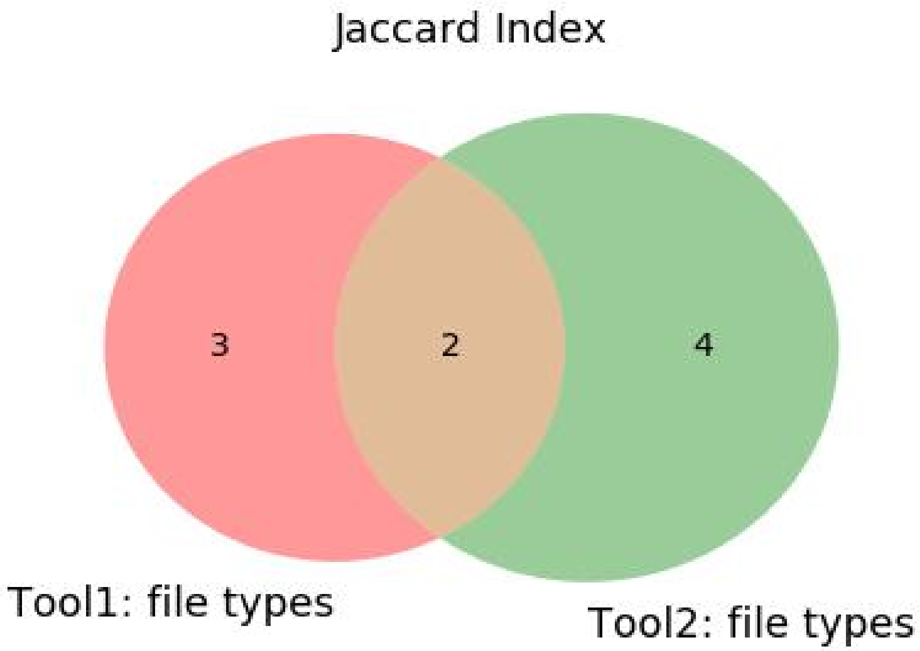
- Extract tool’s attributes
- Clean text
- Create sets of tokens
- Learn similarities
- Combine optimally

Tokens

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**‘ …	?

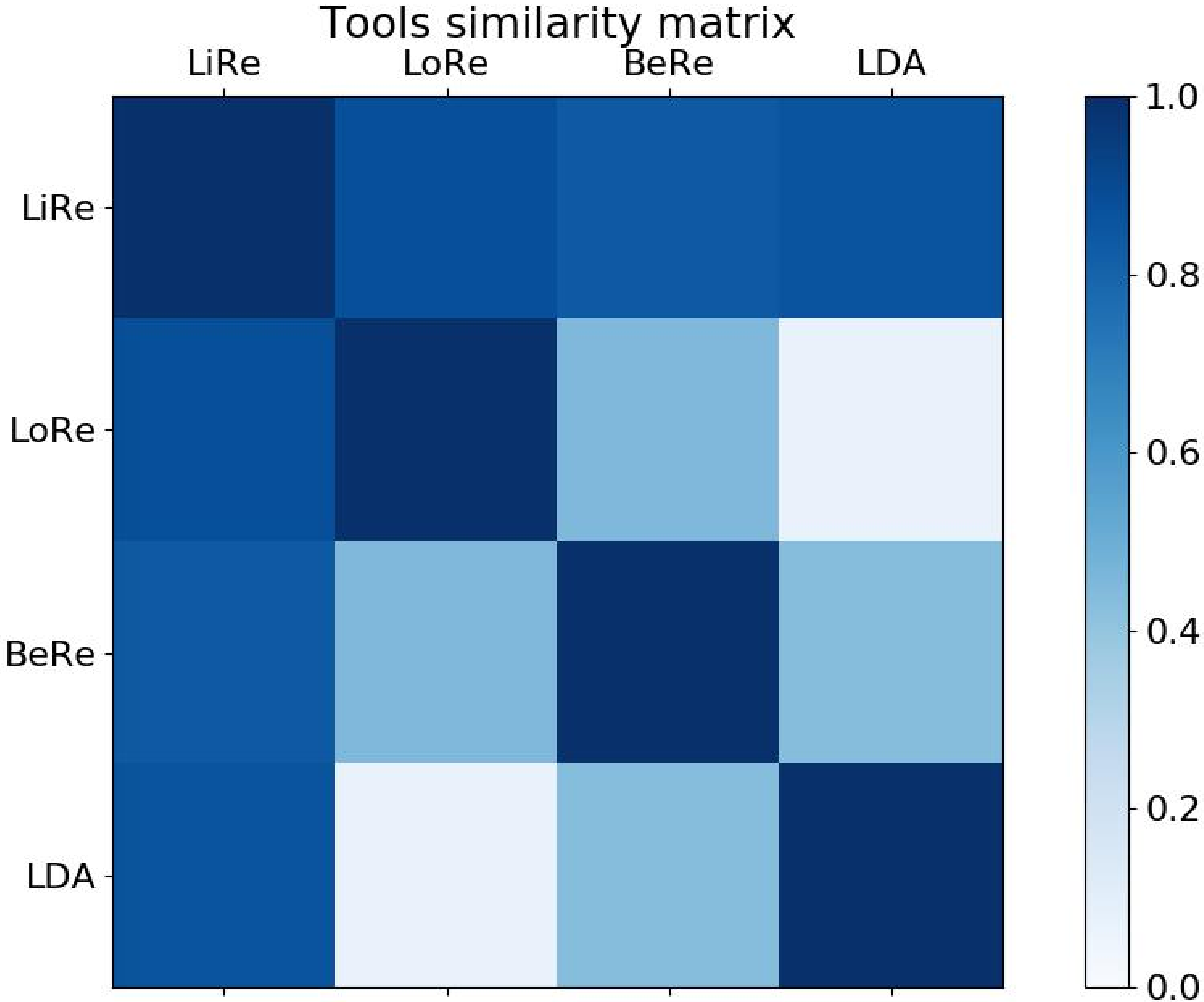
Compute similarity

- Compute Jaccard Index for input/output



- Learn dense vectors for name, description and helptext*
- [‘regress‘, ‘linear‘, ‘perform‘] = [0.98, 0.07, ... , 0.12]
- Compute cosine distance between dense vectors

*[https://cs.stanford.edu/~quocle/paragraph_vector.pdf]



- LiRe - Linear Regression
- LoRe - Logistic Regression
- BeRe - BestSubsetsRegression
- LDA - LDA Analysis

How to combine？

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

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

Optimization

s1 Input, output

1	0.34	0.65	0.9
0.34	1	.66	…
0.65	0.66	1	…
0.9	…	…	1

s2 Name, desc.

1	0.56	0.6	0.9
0.56	1	…	…
…	…	1	…
…	…	…	1

s3 Helptext

1	0.6	0.90	0.7
0.6	1	…	…
0.9	…	1	…
0.7	…	…	1

$$Minimize\,([\,1.0\,,\,1.0\,,\,1.0\,,\,\ldots\,,\,1.0\,]\;-\;[w1\cdot s1\;+\;w2\cdot s2\;+\;w3\cdot s3])$$

where $w1 + w2 + w3 = 1$

Example

- Tool: LinearRegression
- Similarity for input/output: **s1** = [0.33, 0.5, 1.0, ...]
- Similarity for name, desc: **s2** = [0.83, 0.09, 0.005, ...]
- Similarity for helptext: **s3** = [0.45, 0.36, 0.001 .....]
- Optimal weights: w1, w2 and w3
- Similarity:

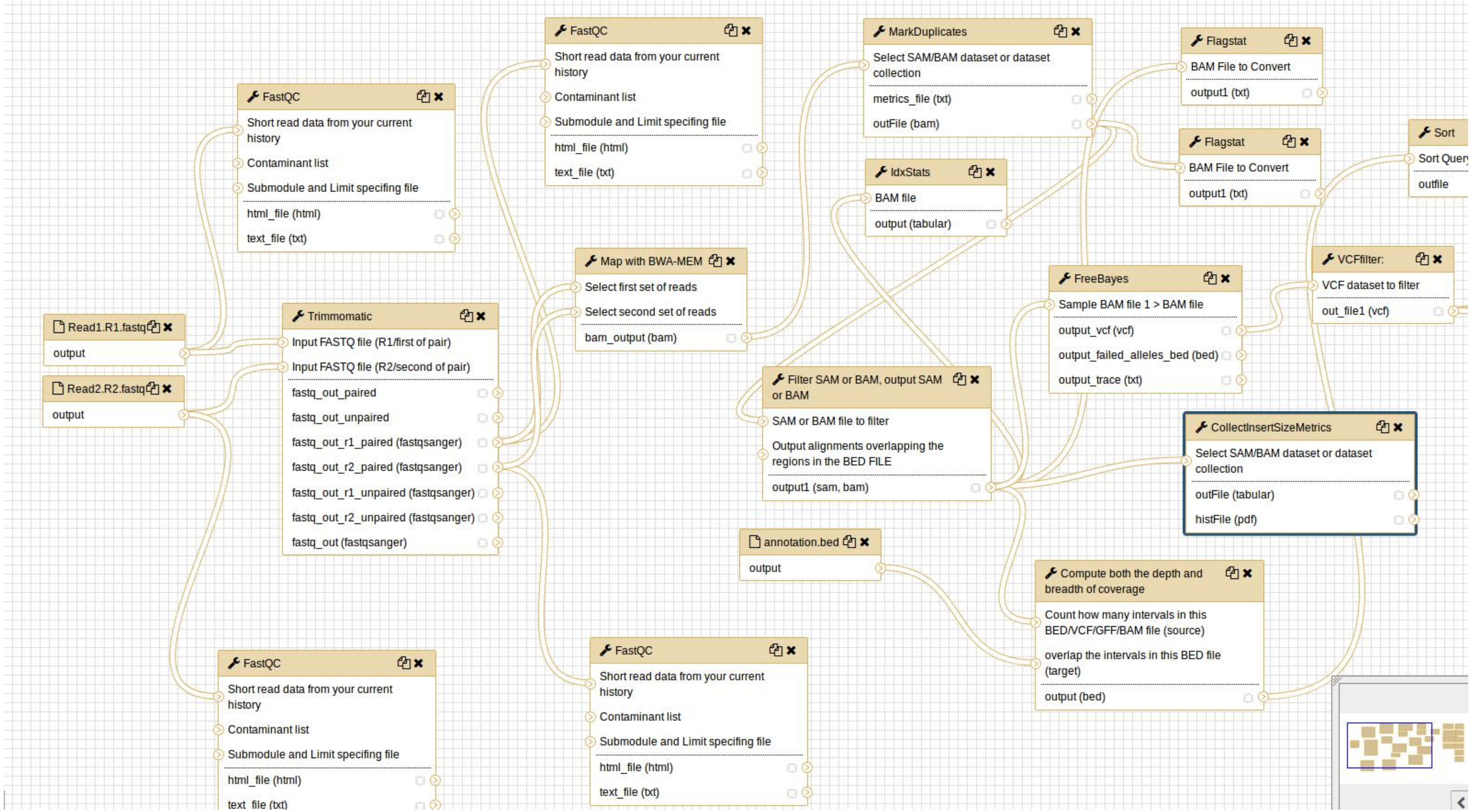
$$[w1 \cdot s1 \; + \; w2 \cdot s2 \; + \; w3 \cdot s3]$$

Visualizer and References

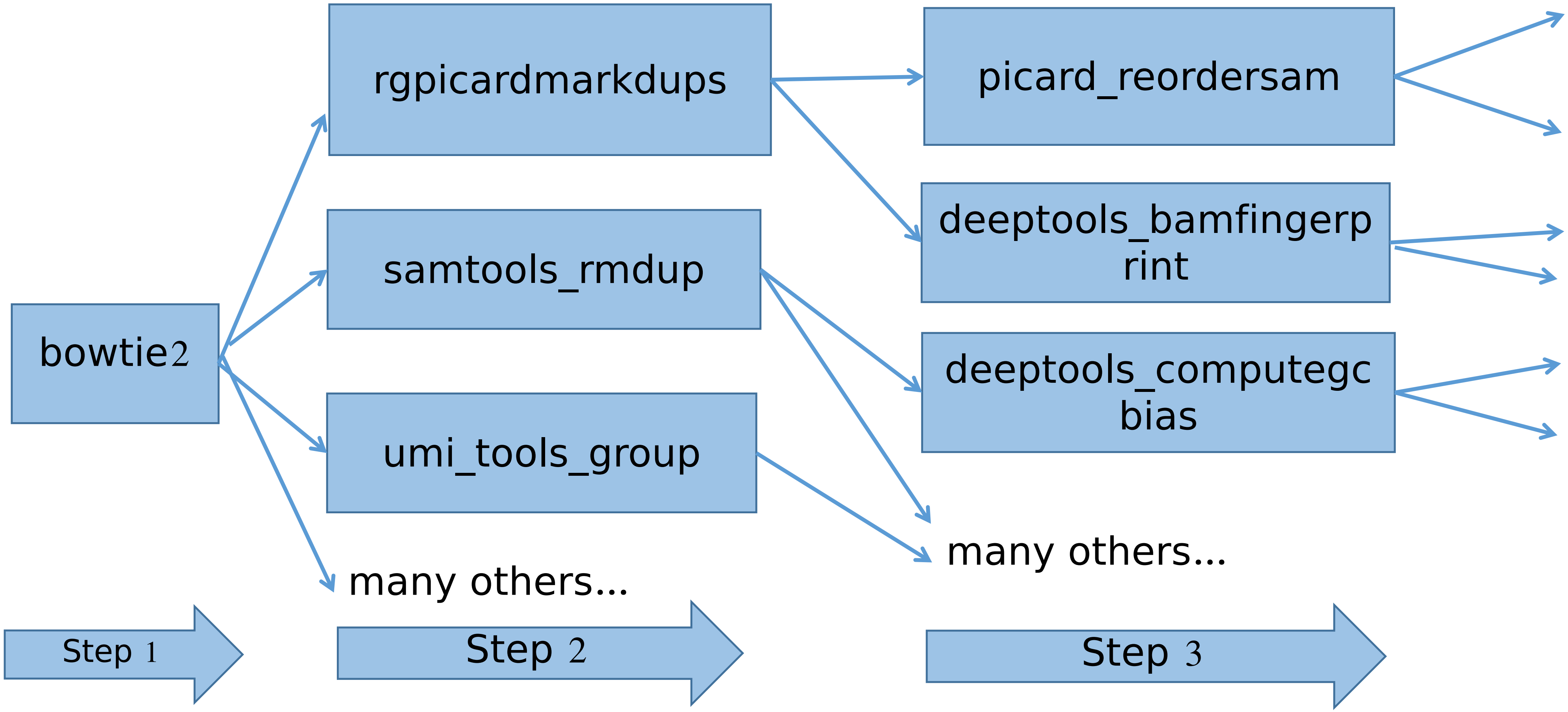
- Static website: results for ~ 1000 tools
- https://rawgit.com/anuprulez/similar_galaxy_tools/master/viz/similarity_viz.html
- https://github.com/anuprulez/similar_galaxy_tools
- https://cs.stanford.edu/%7Equocle/paragraph_vector.pdf
- https://arxiv.org/pdf/1607.05368.pdf

Predict next tools in Galaxy workflows

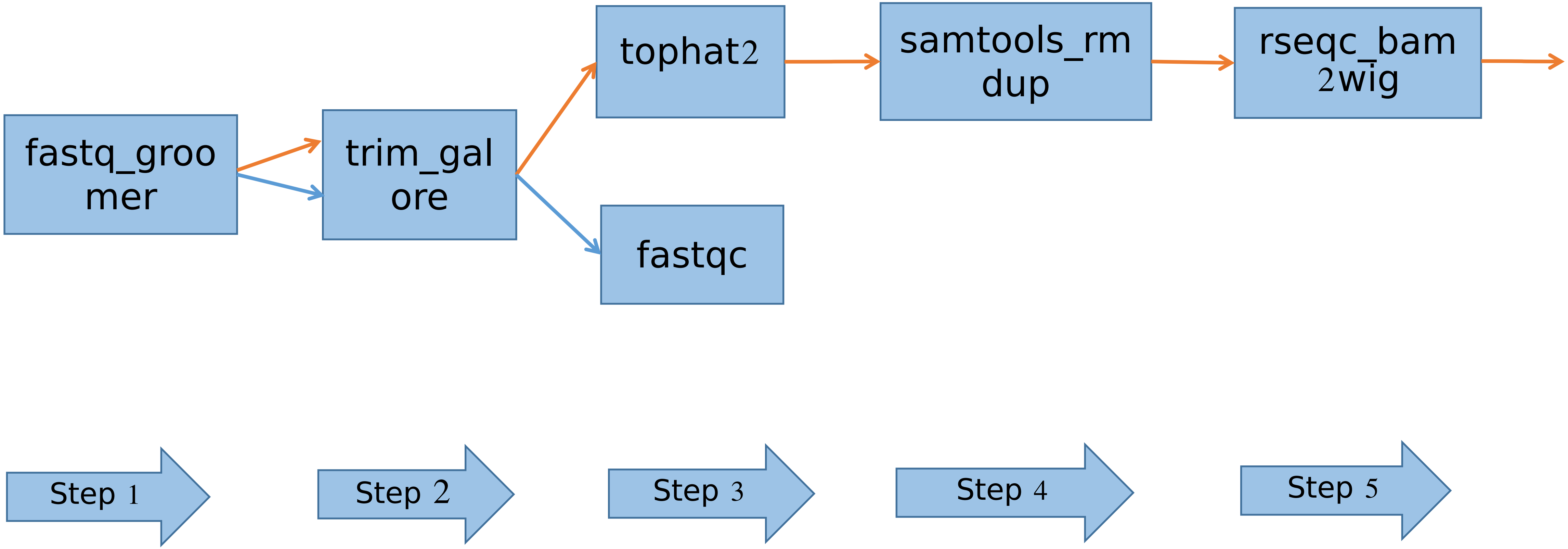
Galaxy workflow


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)
Flagstat
BAM File to Convert
output1 (txt)
Flagstat
BAM File to Convert
output1 (txt)
Sort
Sort Query
outfile
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
IdxStats
BAM file
output (tabular)
Map with BWA-MEM
Select first set of reads
Select second set of reads
bam_output (bam)
VCFfilter:
VCF dataset to filter
out_file1 (vcf)
FreeBayes
Sample BAM file 1 > BAM file
output_vcf (vcf)
output_failed_alleles_bed (bed)
output_trace (txt)
Read1.R1.fastq
output
Read2.R2.fastq
output
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)
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)
CollectInsertSizeMetrics
Select SAM/BAM dataset or dataset collection
outFile (tabular)
histFile (pdf)
annotation.bed
output
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)
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)

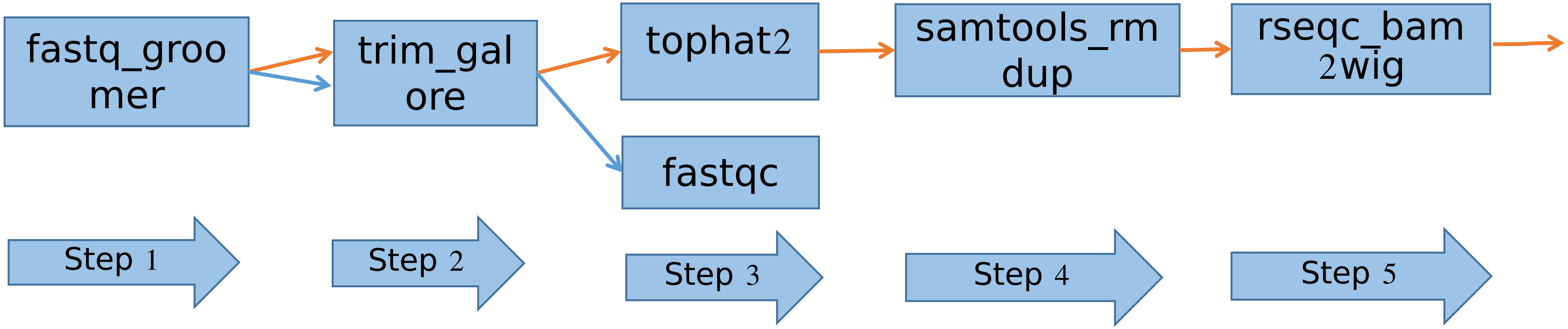
Next tools？


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

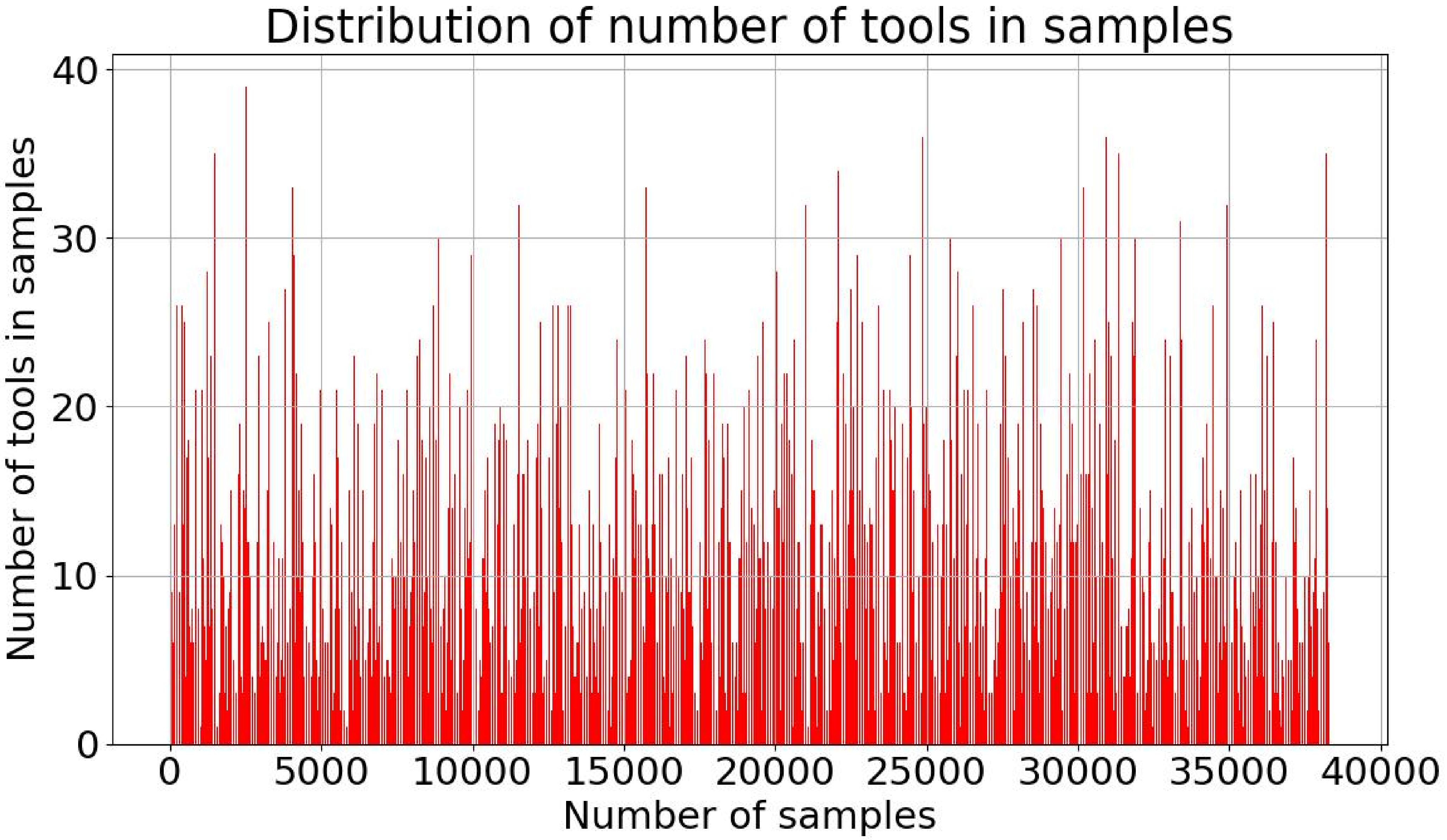


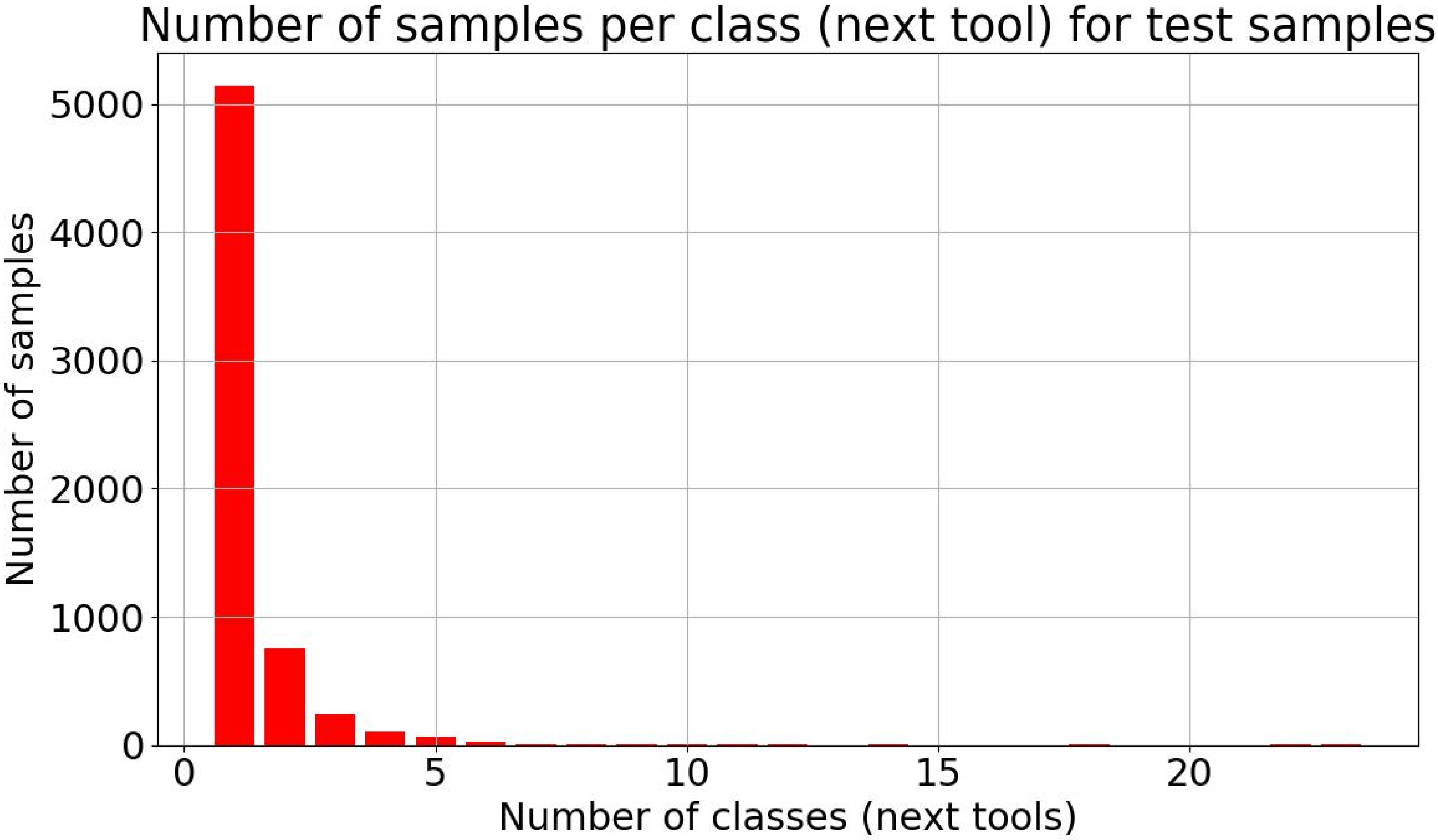
- fastq_groomer, trim_galore (Step 1)
- fastq_groomer, trim_galore, tophat2, fastqc (Step 2)
- fastq_groomer, trim_galore, tophat2, samtools_rmdup (Step 3)
- fastq_groomer, trim_galore, tophat2, samtools_rmdup, rseqc_bam2wig (Step 4)

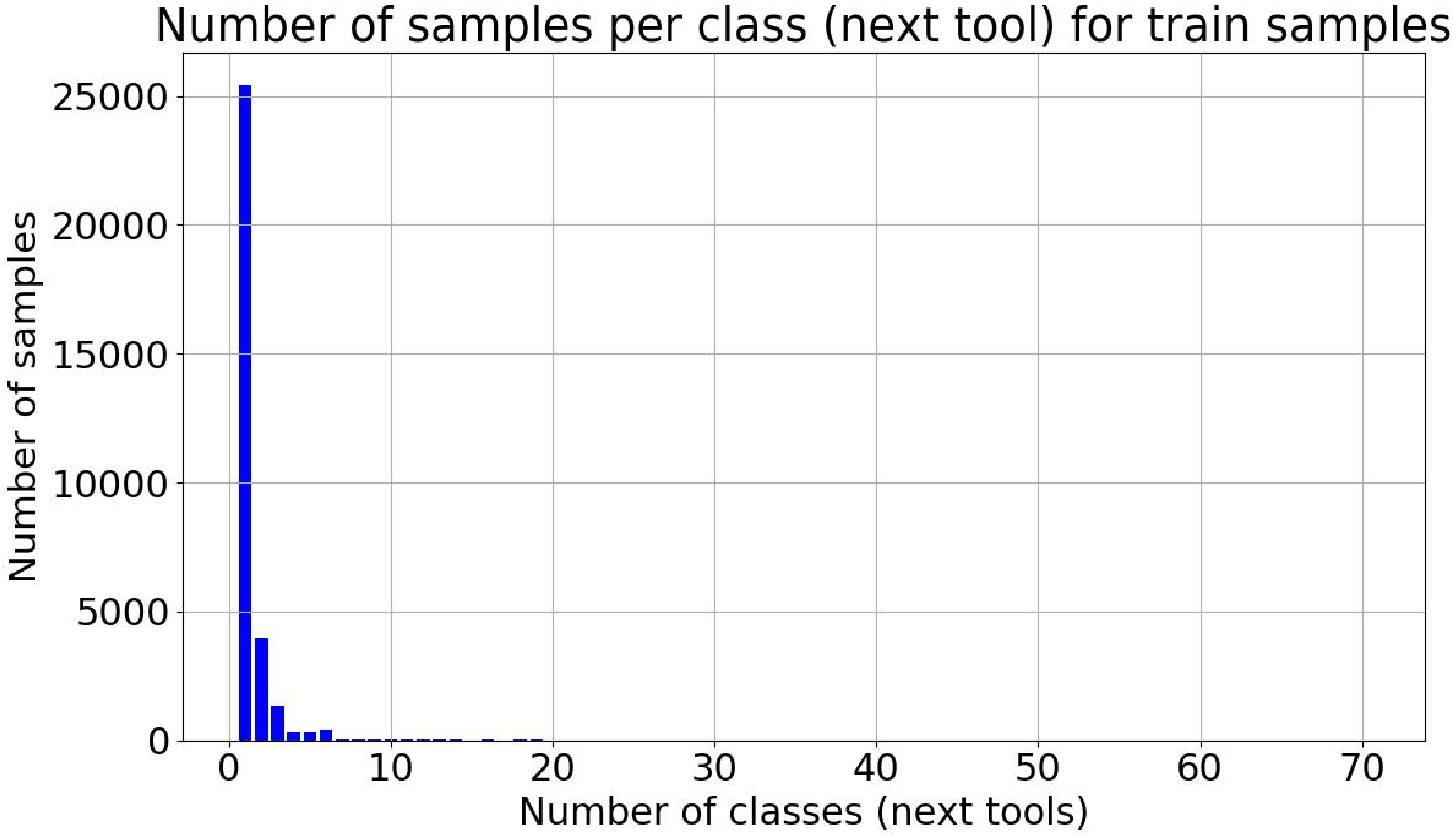
Data preprocessing

‘fastq_groomer‘: 1, ‘trim_galore‘: 2, ‘tophat2‘: 3,
‘samtools_rmdup‘: 4, ‘rseqc_bam2wig‘: 5, ‘fastqc‘: 6

Sample	Label (next tool(s)/classes)
1 ,2 (fastq_groomer, trim_galore)	3, 6 (tophat2, fastqc)
1 ,2 ,3	4
….	….
….	….

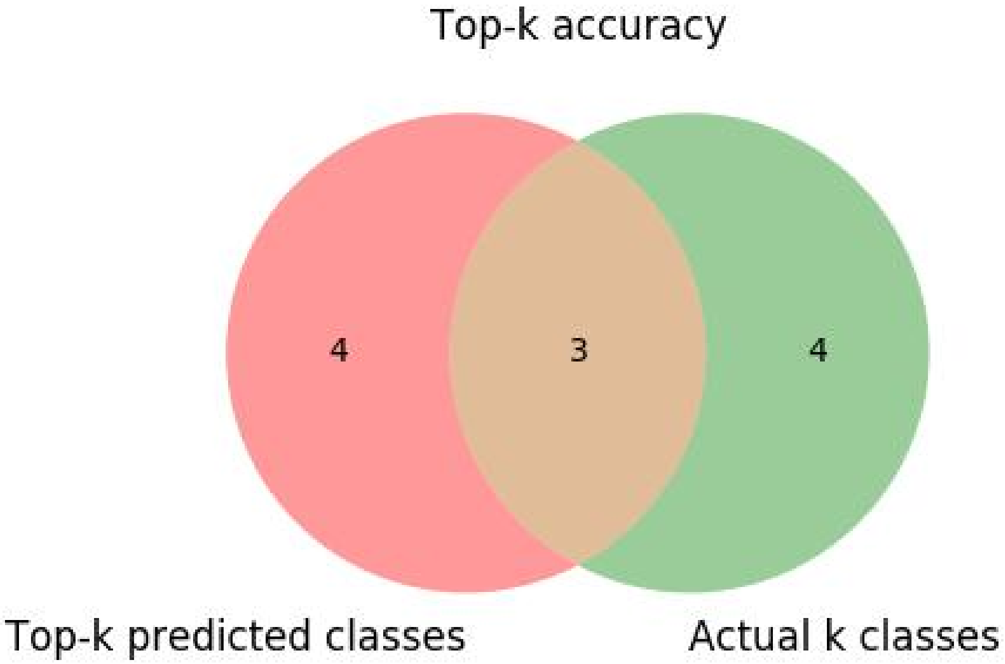

Distribution of number of tools in samples
Number of tools in samples
0
10
20
30
40
Number of samples
0
5000
10000
15000
20000
25000
30000
35000
40000

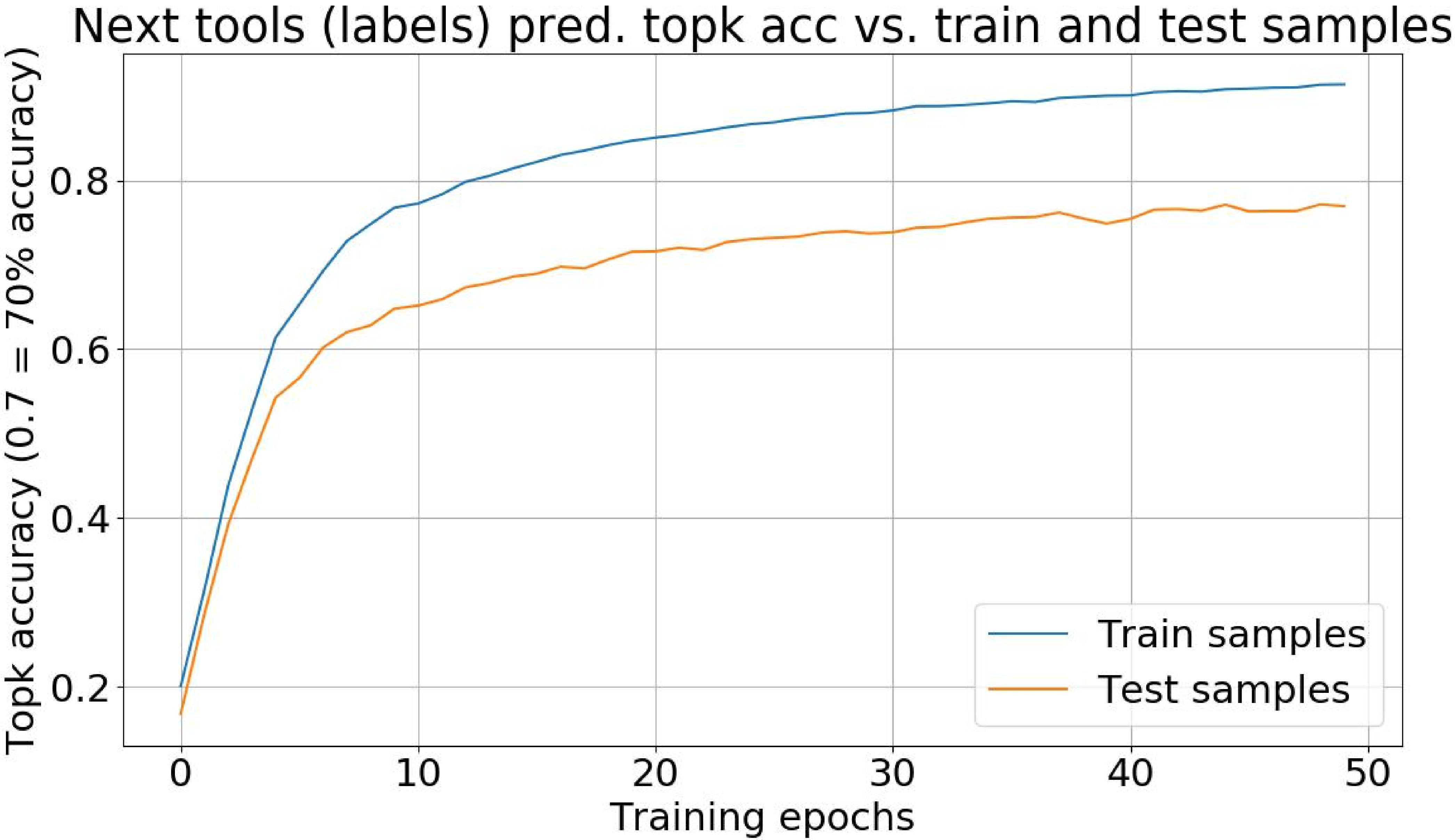
Number of samples per class (next tool) for test samples
Number of samples
0
1000
2000
3000
4000
5000
0
5
10
15
20
Number of classes (next tools)

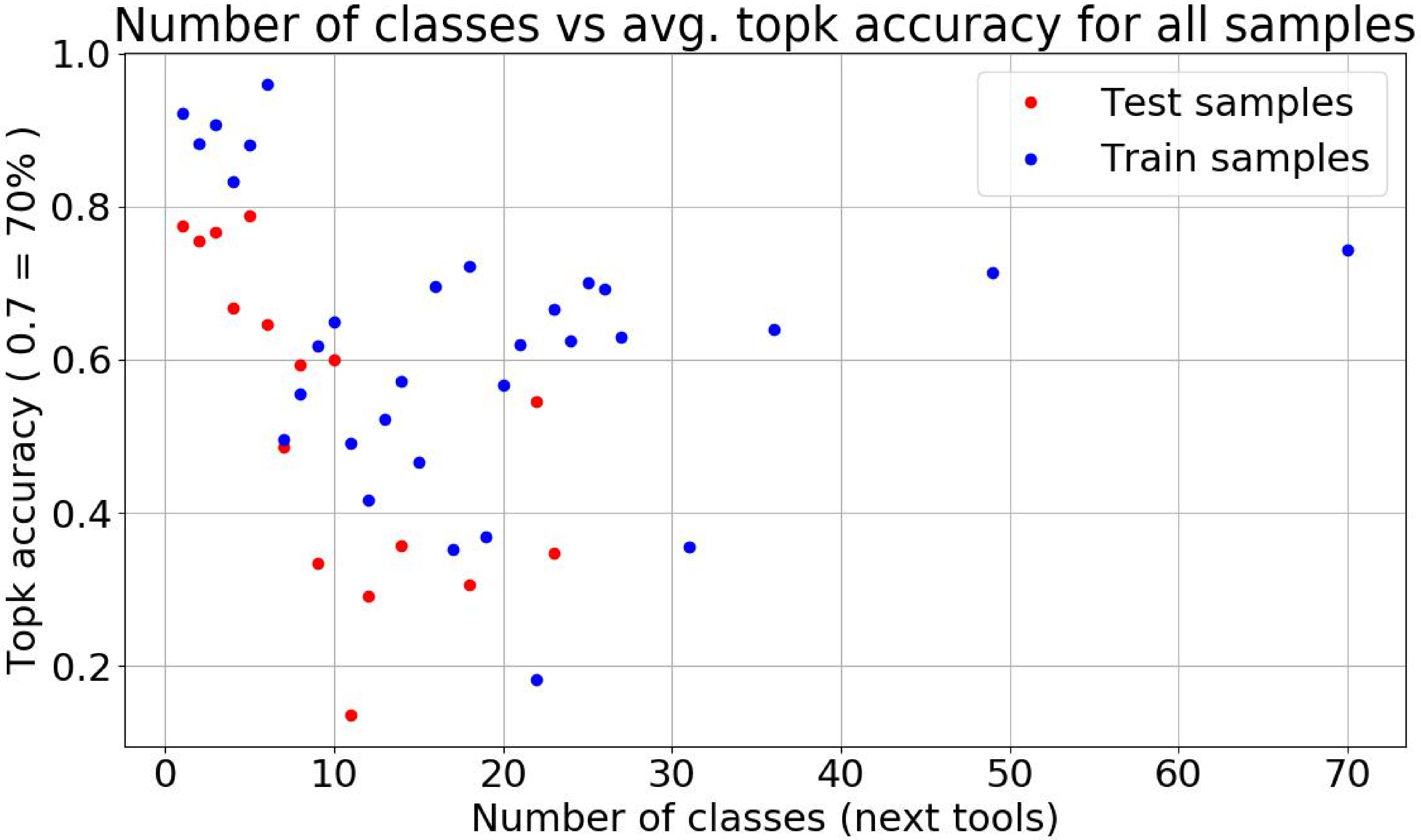

Number of samples per class (next tool) for train samples
Number of samples
25000
20000
15000
10000
5000
0
0
10
20
30
40
50
60
70
Number of classes (next tools)

Classification

- Multi label, multi class classification
- Long short term memory (LSTM) networks
- Topk accuracy




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
Topk accuracy (0.7 = 70%)
1.0
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
Number of classes (next tools)
● Test samples
● Train samples

Next steps

- Data balancing/ augmentation
- Different activations
- Compatibility constraint
- Bayesian Inference
- Convolution

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

Questions？