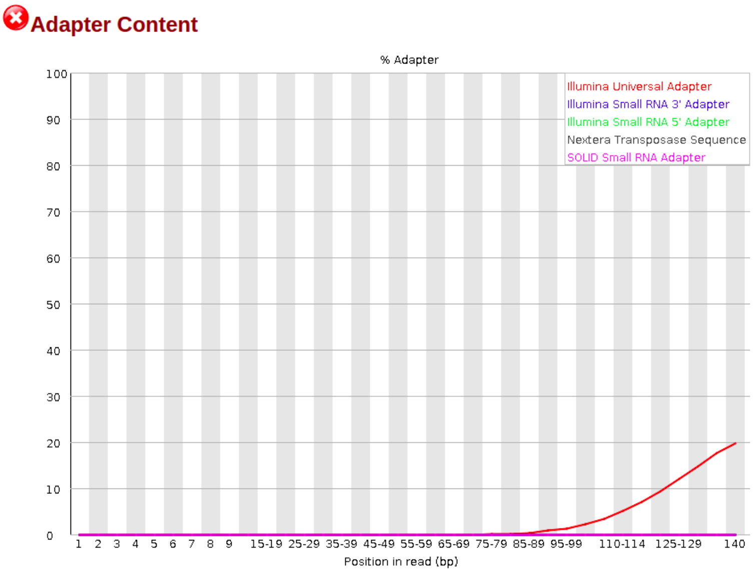
I have implemented all required steps in the snakemake workflow (source code can be found https://github.com/rove-rope/practicals-snakemake).

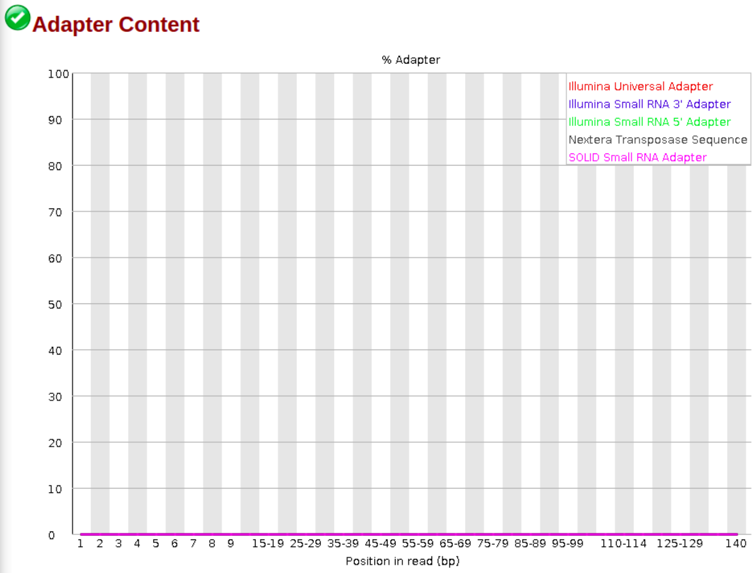
1) **bbduk adapter trim - meaning of options ref=adapters.fa ktrim=r k=23 mink=11 hdist=1 tpe tbo qtrim=r trimq=10**

ref - reference files with adapters or adapter itself;
ktrim - trim reads to remove bases matching reference kmers (r option - trim to the right),
mink - look for shorter kmers at read tips down to this length, when k-trimming or masking,
hdist - maximum hamming distance for ref kmers, tpe - when kmer right-trimming, trim both reads to the minimum length of either, tbo - trim adapters based on where paired reads overlap, qtrim - trim read ends to remove bases with quality below trimq, trimq - regions with average quality bellow this will be trimmed.

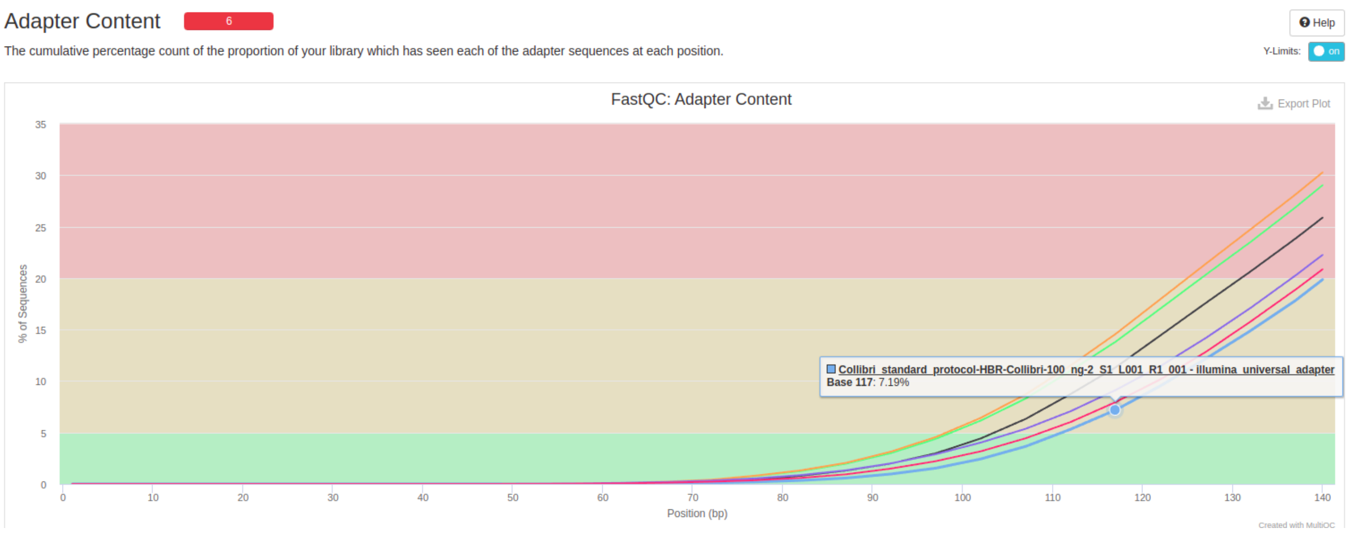
2) **Compare FastQC/MultiQC output for trimmed and not trimmed reads. What indicates that adapters were successfully removed?**

In FastQC, there is a section Adapter Content which in raw data shows that adapters exist in the data and in trimmed data shows that adapters were removed. Example of Adapter Content graphs of Collibri_standard_protocol-HBR-Collibri-100_ng-2_S1_L001_R1_001 read before and after trimming:





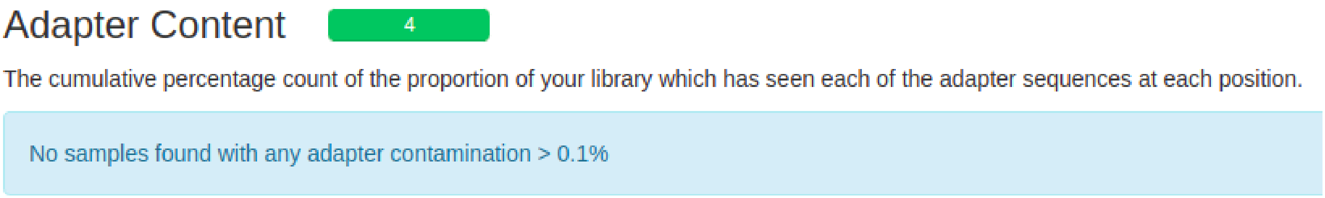
In MultiQC, there is also an Adapter Content section which identifies adapters in raw reads and shows that they were successfully trimmed afterwards. Example of Adapter Content graphs of Collibri_standard_protocol-HBR-Collibri-100_ng-2_S1_L001_R1_001 read before and after trimming (next page):



Adapter Content 6

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

Collibri_standard_protocol-HBR-Collibri-100_ng-2_S1_L001_R1_001 - illumina_universal_adapter
Base 117: 7.19%



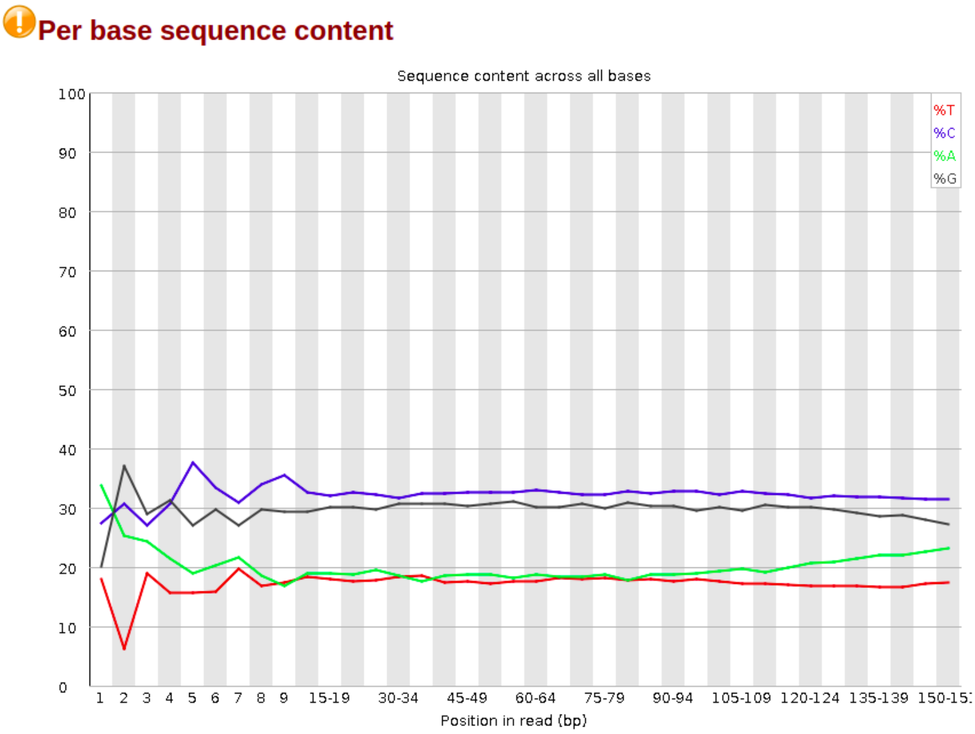
Adapter Content 4

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

No samples found with any adapter contamination > 0.1%

3) Dig into some FastQC reports, especially of files with names “Collibri...”. Are there any signs of over representation of polyA?

I expected to see polyA sequences in the Overrepresented Sequences section but did not. In the Per base sequence content section of FastQC reports for forward Collibri reads there is an increased percentage of A bases in the beginning of the read, however this may not be related as results barely change after polyA sequences are trimmed. Example of Collibri_standard_protocol-HBR-Collibri-100_ng-2_S1_L001_R1_001 read:



Per base sequence content

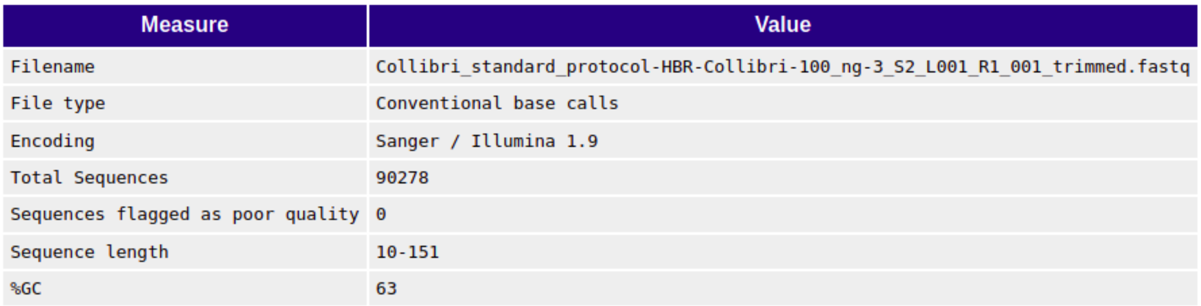
4) **How would you change the “adapters.fa” file to eliminate polyA sequences from reads?**

Add these lines to the file:
>polyA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

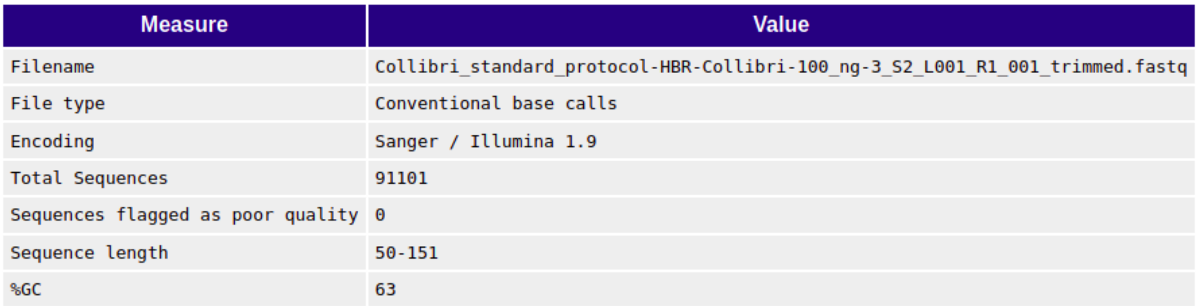
5) **How could failing to do this affect further analysis? Try out the “adapter.fa” file with addition of polyA (30A symbols)? Any changes?**

Failure to trim polyA sequences could affect per base sequence content, read quality and overrepresentation of polyA sequences. Comparing FastQC/MultiQC results of trimmed reads where polyA sequence was trimmed vs where polyA sequence was not trimmed these differences were noticed:

- Number of Total Sequences and Sequence length was reduced after polyA trimming:

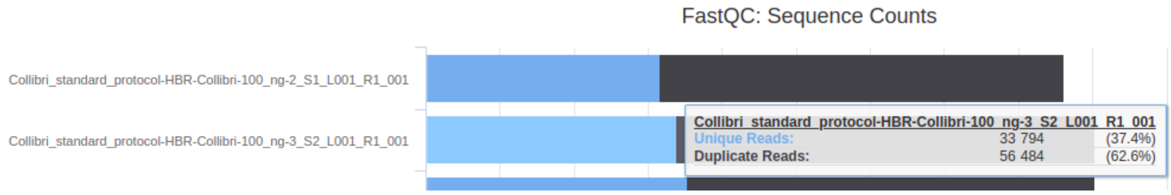


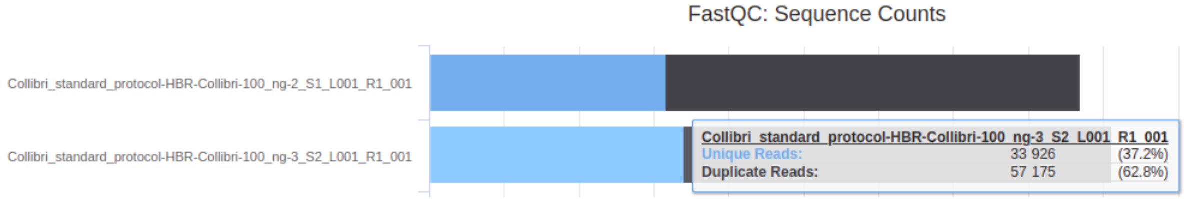
Measure	Value
Filename	Collibri_standard_protocol-HBR-Collibri-100_ng-3_S2_L001_R1_001_trimmed.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	90278
Sequences flagged as poor quality	0
Sequence length	10-151
%GC	63



Measure	Value
Filename	Collibri_standard_protocol-HBR-Collibri-100_ng-3_S2_L001_R1_001_trimmed.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	91101
Sequences flagged as poor quality	0
Sequence length	50-151
%GC	63

- Minor change in Per base sequence content and Sequence Length Distribution;
- Reduced number of some Overrepresented sequences.
- Slightly increased percentage of unique sequences:





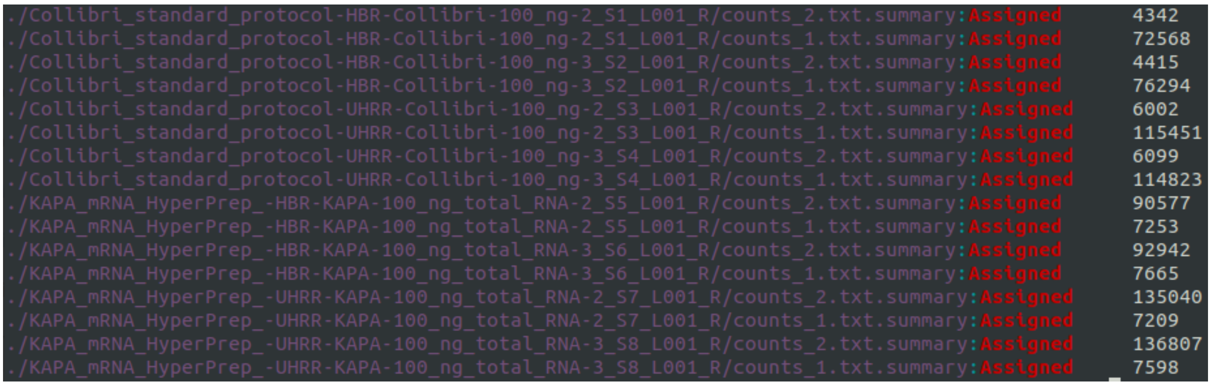
6) **[featureCounts] Please take note that either 1 or 2 value for -s is needed. That is due to strand specificity.**

The aim of strand-specific protocols is to allow an assignment of the reads to their original strand. In general this can be managed in two different ways: by attaching different adapters in a known orientation relative to the 5' and 3' ends of the RNA transcript or by marking one strand by chemical modification (e.g. bisulfite treatment).

Value s is chosen based on strand specificity: 0 (unstranded), 1 (stranded) and 2 (reversely stranded).

7) **[featureCounts] In practice for each library test either 1 or 2. The "proper" value would be the one giving higher count numbers. Which library in this set would be the special one and would be different from others?**

It appears that KAPA mRNA HyperPrep protocol is reversely stranded (2) and Collibri Standard Protocol is stranded (1) based on results in counts.txt.summary. I did not find any library that would have a different result - for all Collibri reads higher count numbers are achieved using -s 1 (counts_1.txt) and for all KAPA reads - using -s 2 (counts_2.txt).



```
./Collibri_standard_protocol-HBR-Collibri-100_ng-2_S1_L001_R/counts_2.txt.summary:Assigned      4342
./Collibri_standard_protocol-HBR-Collibri-100_ng-2_S1_L001_R/counts_1.txt.summary:Assigned      72568
./Collibri_standard_protocol-HBR-Collibri-100_ng-3_S2_L001_R/counts_2.txt.summary:Assigned      4415
./Collibri_standard_protocol-HBR-Collibri-100_ng-3_S2_L001_R/counts_1.txt.summary:Assigned      76294
./Collibri_standard_protocol-UHRR-Collibri-100_ng-2_S3_L001_R/counts_2.txt.summary:Assigned     6002
./Collibri_standard_protocol-UHRR-Collibri-100_ng-2_S3_L001_R/counts_1.txt.summary:Assigned     115451
./Collibri_standard_protocol-UHRR-Collibri-100_ng-3_S4_L001_R/counts_2.txt.summary:Assigned     6099
./Collibri_standard_protocol-UHRR-Collibri-100_ng-3_S4_L001_R/counts_1.txt.summary:Assigned     114823
./KAPA_mRNA_HyperPrep_-HBR-KAPA-100_ng_total_RNA-2_S5_L001_R/counts_2.txt.summary:Assigned      90577
./KAPA_mRNA_HyperPrep_-HBR-KAPA-100_ng_total_RNA-2_S5_L001_R/counts_1.txt.summary:Assigned      7253
./KAPA_mRNA_HyperPrep_-HBR-KAPA-100_ng_total_RNA-3_S6_L001_R/counts_2.txt.summary:Assigned      92942
./KAPA_mRNA_HyperPrep_-HBR-KAPA-100_ng_total_RNA-3_S6_L001_R/counts_1.txt.summary:Assigned      7665
./KAPA_mRNA_HyperPrep_-UHRR-KAPA-100_ng_total_RNA-2_S7_L001_R/counts_2.txt.summary:Assigned     135040
./KAPA_mRNA_HyperPrep_-UHRR-KAPA-100_ng_total_RNA-2_S7_L001_R/counts_1.txt.summary:Assigned     7209
./KAPA_mRNA_HyperPrep_-UHRR-KAPA-100_ng_total_RNA-3_S8_L001_R/counts_2.txt.summary:Assigned     136807
./KAPA_mRNA_HyperPrep_-UHRR-KAPA-100_ng_total_RNA-3_S8_L001_R/counts_1.txt.summary:Assigned     7598
```

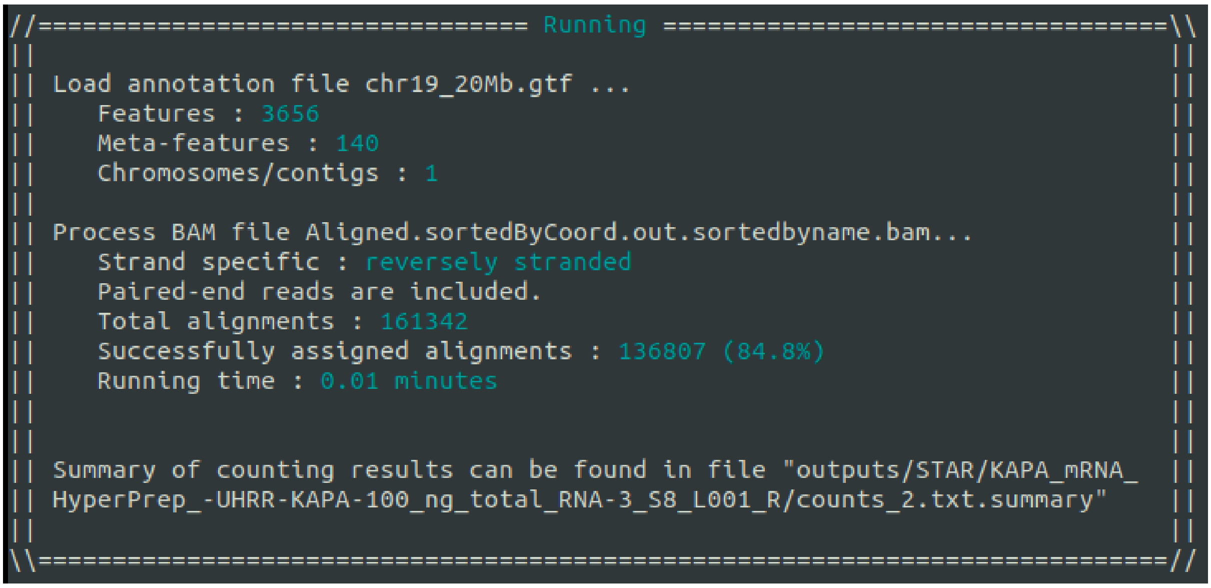
8) **[featureCounts] Collect data on alignment rate per sample (look for fraction of uniquely mapped reads). Are there any differences that could be related to tissue types and/or sample preparation methods?**

Protocol	Tissue type	Repeat	Unique alignments	Percentage
Collibri	HBR	2	72568	83.4%
Collibri	HBR	3	76294	83.6%
Collibri	UHRR	2	115451	76.9%
Collibri	UHRR	3	114823	76.7%
KAPA	HBR	2	90577	87.0%
KAPA	HBR	3	92942	87.1%
KAPA	UHRR	2	135040	85.2%
KAPA	UHRR	3	136807	84.8%

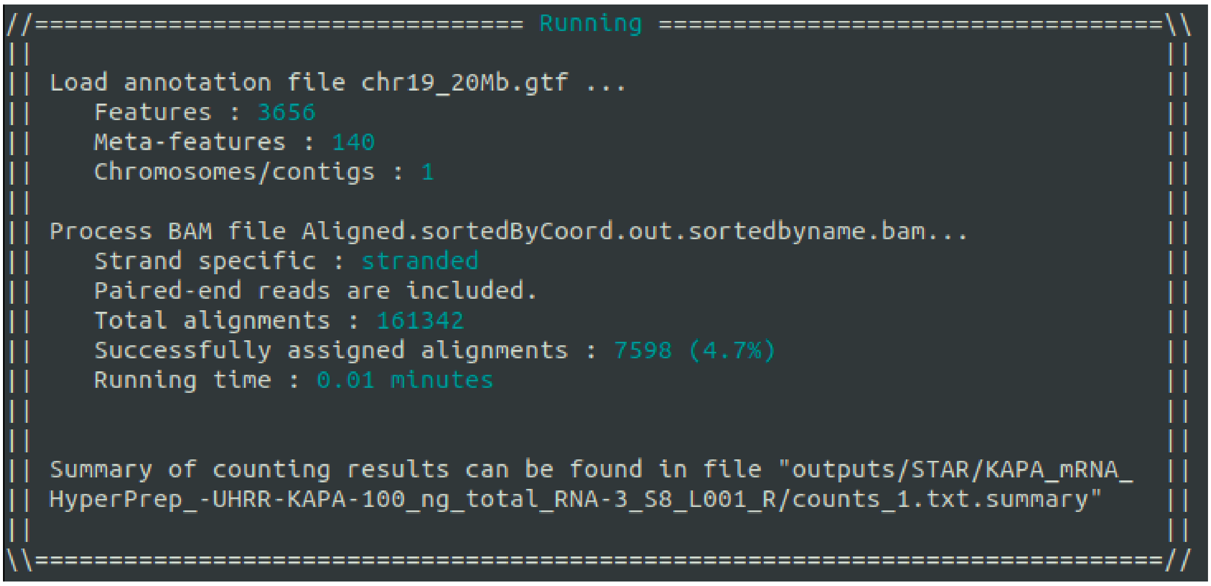
Based on the sample preparation method, samples prepared using Collibri protocol have lower alignment rate per sample than samples prepared using KAPA protocol. Based on tissue type, UHRR samples (originating from several cancerous cell lines) have lower alignment rate per sample than HBR (normal) samples.

9) [featureCounts] What strand specificity settings should be used for each sample preparation method. Illustrate by one example what happens if a wrong setting is used?

As mentioned before, KAPA mRNA HyperPrep protocol is reversely stranded (2) and Collibri Standard Protocol is stranded (1) based on results in counts.txt.summary. If the wrong setting is used, the number of successfully assigned alignments is significantly reduced.



```
//================================= Running ==================================\\
||                                                                            ||
|| Load annotation file chr19_20Mb.gtf ...                                    ||
||    Features : 3656                                                         ||
||    Meta-features : 140                                                     ||
||    Chromosomes/contigs : 1                                                 ||
||                                                                            ||
|| Process BAM file Aligned.sortedByCoord.out.sortedbyname.bam...             ||
||    Strand specific : reversely stranded                                    ||
||    Paired-end reads are included.                                          ||
||    Total alignments : 161342                                               ||
||    Successfully assigned alignments : 136807 (84.8%)                       ||
||    Running time : 0.01 minutes                                             ||
||                                                                            ||
||                                                                            ||
|| Summary of counting results can be found in file "outputs/STAR/KAPA_mRNA_  ||
|| HyperPrep_-UHRR-KAPA-100_ng_total_RNA-3_S8_L001_R/counts_2.txt.summary"    ||
||                                                                            ||
\\============================================================================//
```

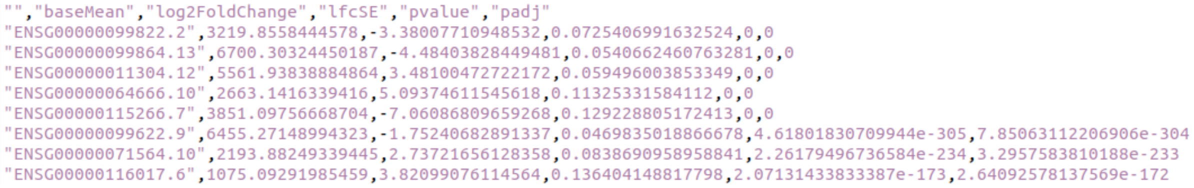


```
//================================= Running ==================================\\
||                                                                            ||
|| Load annotation file chr19_20Mb.gtf ...                                    ||
||    Features : 3656                                                         ||
||    Meta-features : 140                                                     ||
||    Chromosomes/contigs : 1                                                 ||
||                                                                            ||
|| Process BAM file Aligned.sortedByCoord.out.sortedbyname.bam...             ||
||    Strand specific : stranded                                              ||
||    Paired-end reads are included.                                          ||
||    Total alignments : 161342                                               ||
||    Successfully assigned alignments : 7598 (4.7%)                          ||
||    Running time : 0.01 minutes                                             ||
||                                                                            ||
||                                                                            ||
|| Summary of counting results can be found in file "outputs/STAR/KAPA_mRNA_  ||
|| HyperPrep_-UHRR-KAPA-100_ng_total_RNA-3_S8_L001_R/counts_1.txt.summary"    ||
||                                                                            ||
\\============================================================================//
```

10) DE analysis - The output should be:

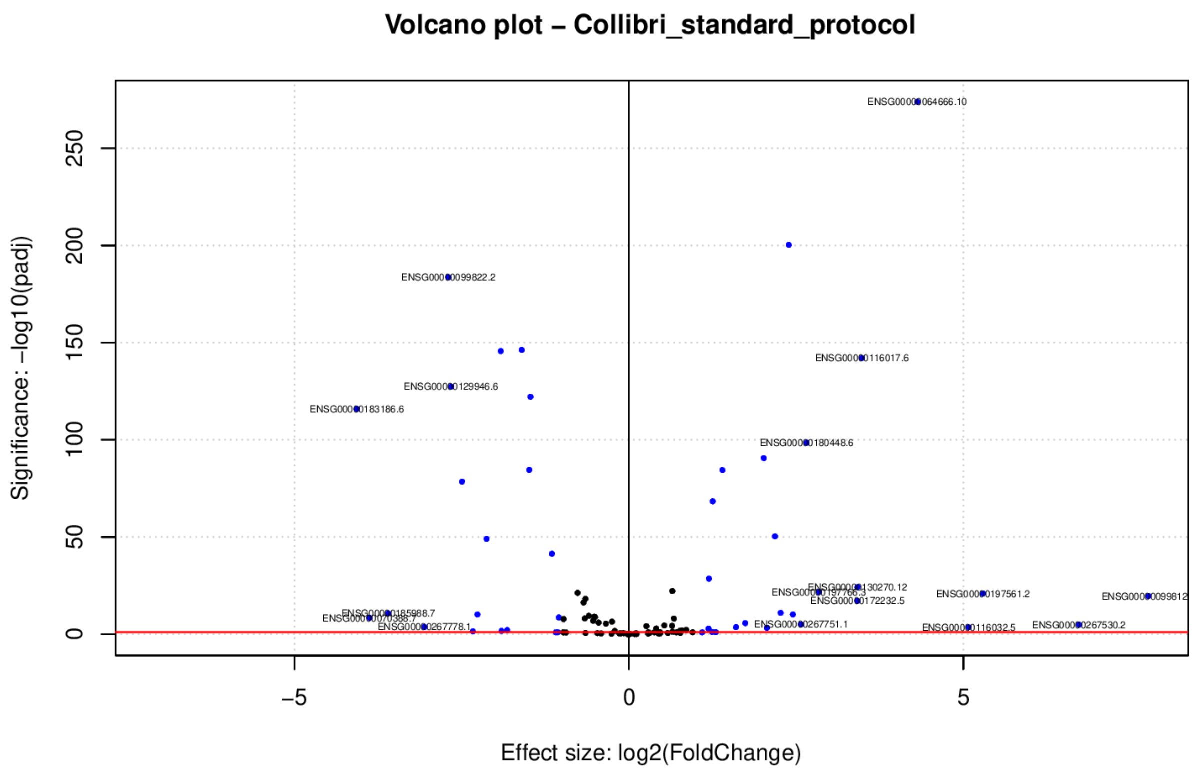
a) a list of DE genes with padj values.

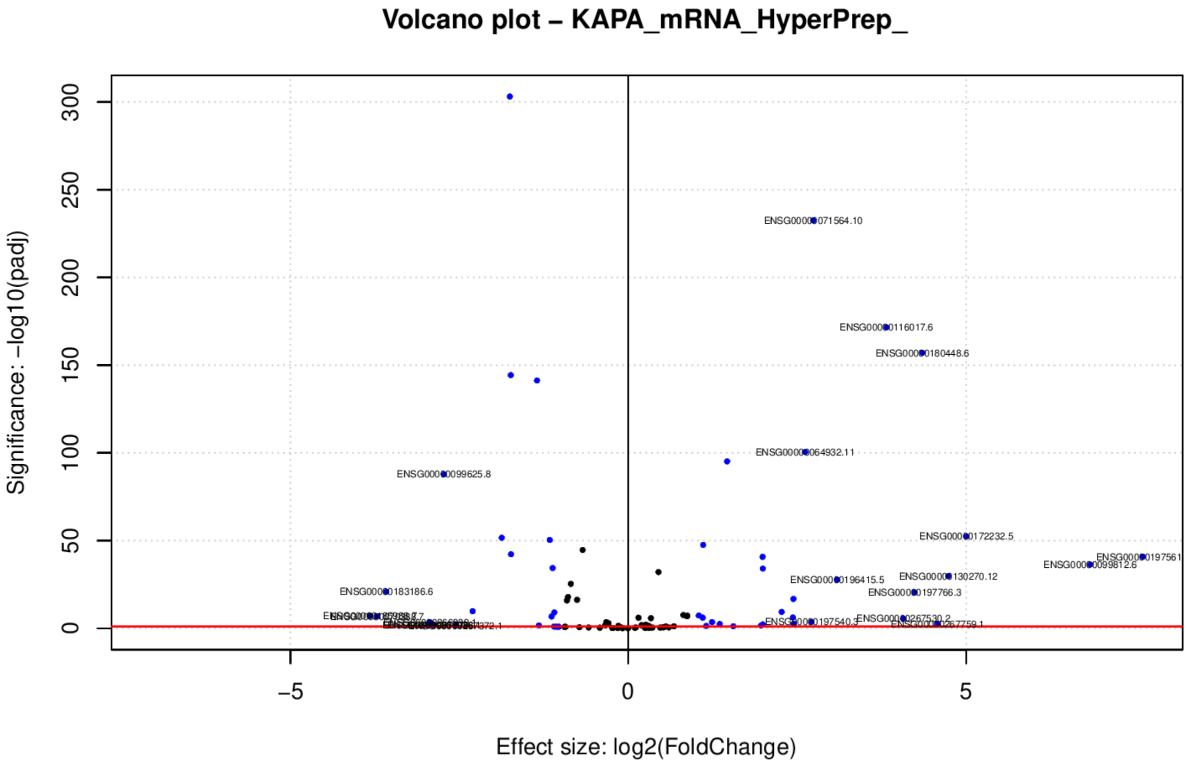
Stored in resLFC.csv files. Example:



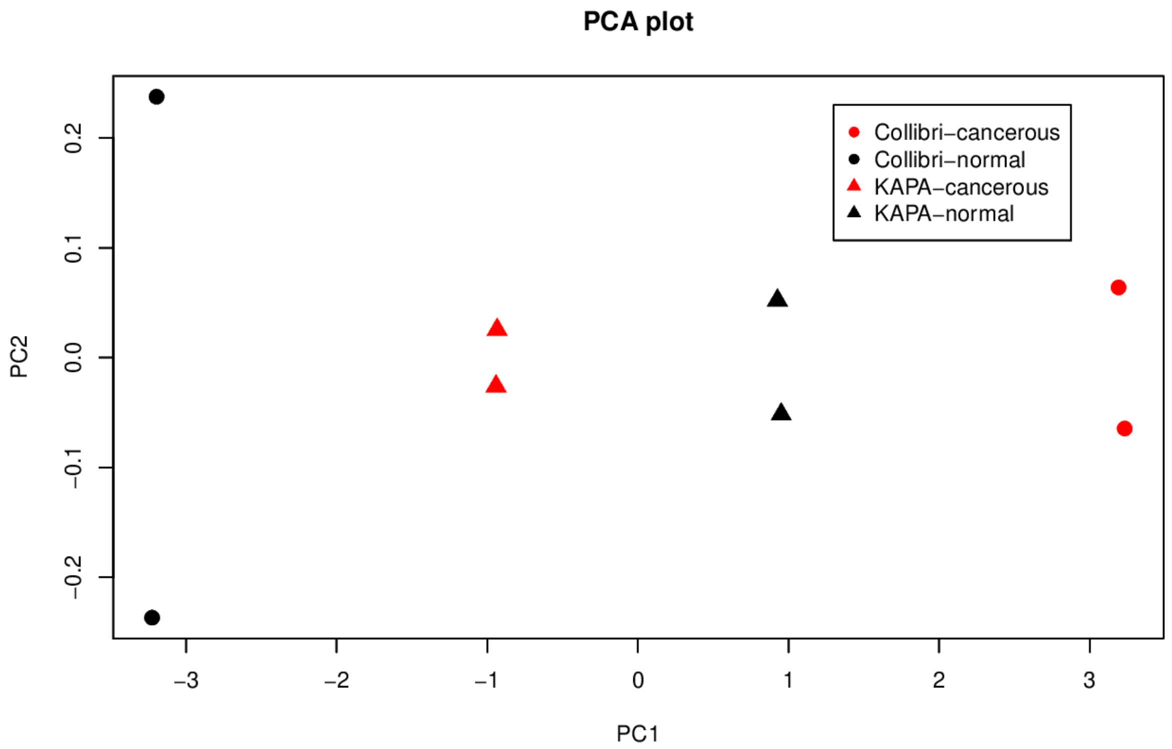
```
"","baseMean","log2FoldChange","lfcSE","pvalue","padj"
"ENSG00000099822.2",3219.8558444578,-3.38007710948532,0.0725406991632524,0,0
"ENSG00000099864.13",6700.30324450187,-4.48403828449481,0.0540662460763281,0,0
"ENSG00000011304.12",5561.93838884864,3.48100472722172,0.059496003853349,0,0
"ENSG00000064666.10",2663.1416339416,5.09374611545618,0.11325331584112,0,0
"ENSG00000115266.7",3851.09756668704,-7.06086809659268,0.129228805172413,0,0
"ENSG00000099622.9",6455.27148994323,-1.75240682891337,0.0469835018866678,4.61801830709944e-305,7.85063112206906e-304
"ENSG00000071564.10",2193.88249339445,2.73721656128358,0.0838690958958841,2.26179496736584e-234,3.2957583810188e-233
"ENSG00000116017.6",1075.09291985459,3.82099076114564,0.136404148817798,2.07131433833387e-173,2.64092578137569e-172
```

b) Vulcan Plot





11) **Create a PCA plot exploring how different sample preparation plots cluster based on differentially expressed genes (use genes that are DE for all sample preparation methods).**

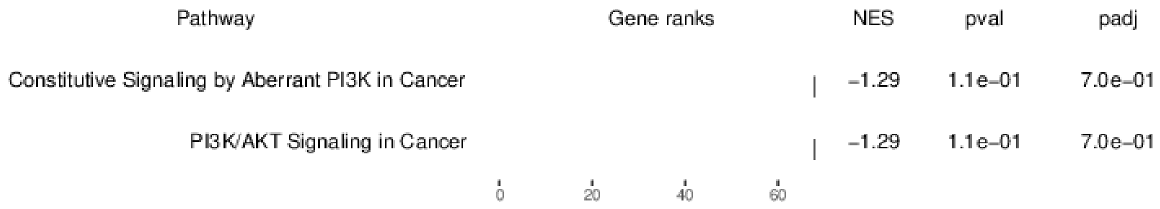


12) **Which fold changes you should use - shrunken or not for the enrichment analysis?**

LFC – shrunken log2 fold change – is a count ratio model to estimate fold changes. The shrunken fold changes are useful for ranking genes by effect size and for visualization. Additionally, for functional analysis tools such as GSEA which require fold change values as input you would want to provide shrunken values. Therefore, LFC should be used.

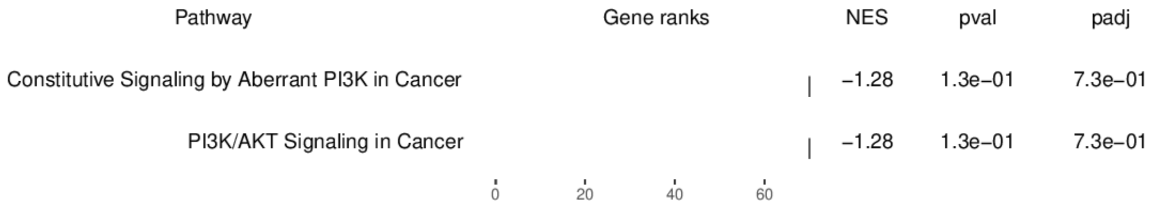
13) **Make a R script testing enrichment for the Reactome pathways. Are there any cancer related enriched pathways?**

Collibri



Pathway	Gene ranks	NES	pval	padj
Constitutive Signaling by Aberrant PI3K in Cancer		−1.29	1.1e−01	7.0e−01
PI3K/AKT Signaling in Cancer		−1.29	1.1e−01	7.0e−01

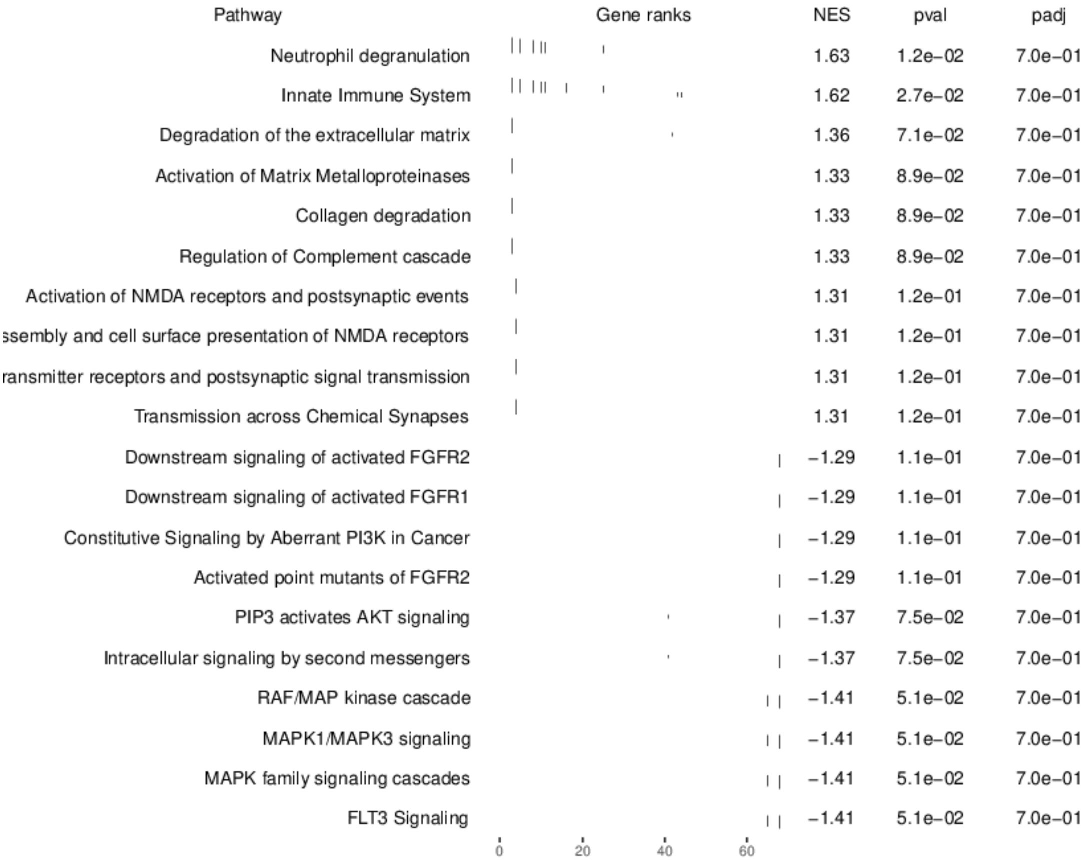
KAPA



Pathway	Gene ranks	NES	pval	padj
Constitutive Signaling by Aberrant PI3K in Cancer		−1.28	1.3e−01	7.3e−01
PI3K/AKT Signaling in Cancer		−1.28	1.3e−01	7.3e−01

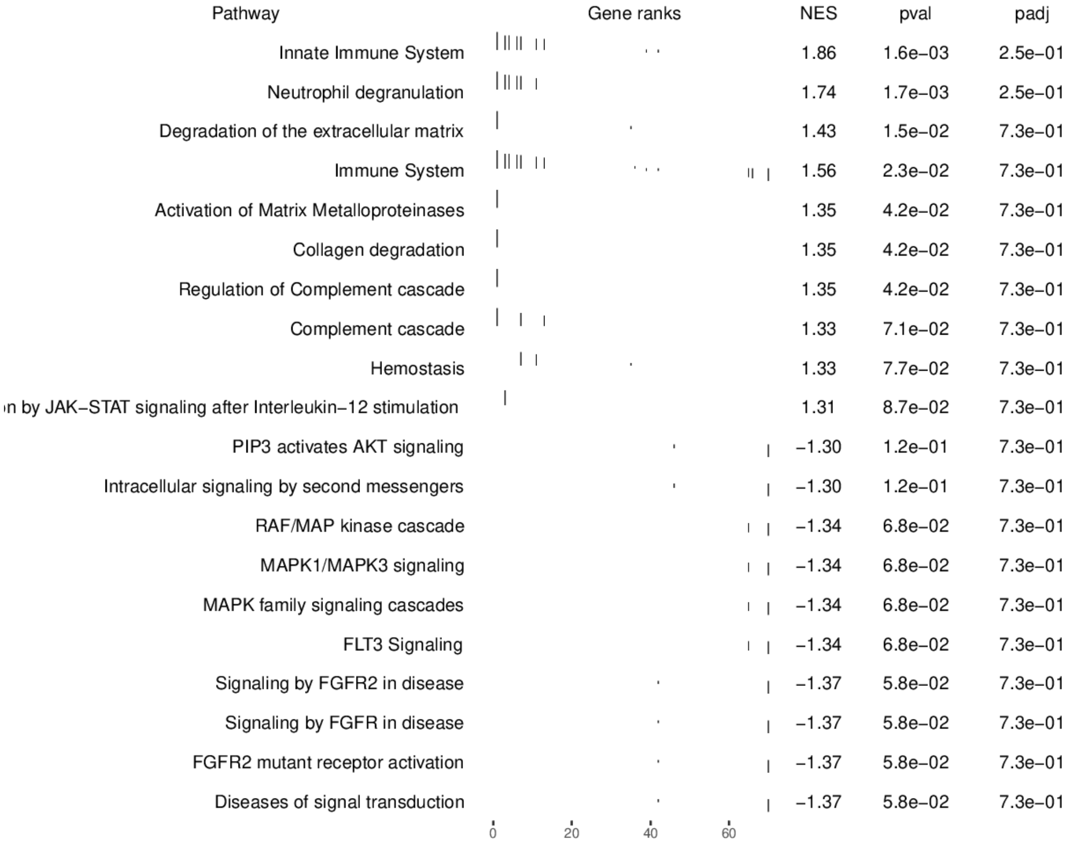
14) **Are there any differences in results matching different sample preparation kits?**

Collibri



Pathway	NES	pval	padj
Neutrophil degranulation	1.63	1.2e−02	7.0e−01
Innate Immune System	1.62	2.7e−02	7.0e−01
Degradation of the extracellular matrix	1.36	7.1e−02	7.0e−01
Activation of Matrix Metalloproteinases	1.33	8.9e−02	7.0e−01
Collagen degradation	1.33	8.9e−02	7.0e−01
Regulation of Complement cascade	1.33	8.9e−02	7.0e−01
Activation of NMDA receptors and postsynaptic events	1.31	1.2e−01	7.0e−01
ssembly and cell surface presentation of NMDA receptors	1.31	1.2e−01	7.0e−01
ransmitter receptors and postsynaptic signal transmission	1.31	1.2e−01	7.0e−01
Transmission across Chemical Synapses	1.31	1.2e−01	7.0e−01
Downstream signaling of activated FGFR2	−1.29	1.1e−01	7.0e−01
Downstream signaling of activated FGFR1	−1.29	1.1e−01	7.0e−01
Constitutive Signaling by Aberrant PI3K in Cancer	−1.29	1.1e−01	7.0e−01
Activated point mutants of FGFR2	−1.29	1.1e−01	7.0e−01
PIP3 activates AKT signaling	−1.37	7.5e−02	7.0e−01
Intracellular signaling by second messengers	−1.37	7.5e−02	7.0e−01
RAF/MAP kinase cascade	−1.41	5.1e−02	7.0e−01
MAPK1/MAPK3 signaling	−1.41	5.1e−02	7.0e−01
MAPK family signaling cascades	−1.41	5.1e−02	7.0e−01
FLT3 Signaling	−1.41	5.1e−02	7.0e−01

KAPA



Pathway	NES	pval	padj
Innate Immune System	1.86	1.6e−03	2.5e−01
Neutrophil degranulation	1.74	1.7e−03	2.5e−01
Degradation of the extracellular matrix	1.43	1.5e−02	7.3e−01
Immune System	1.56	2.3e−02	7.3e−01
Activation of Matrix Metalloproteinases	1.35	4.2e−02	7.3e−01
Collagen degradation	1.35	4.2e−02	7.3e−01
Regulation of Complement cascade	1.35	4.2e−02	7.3e−01
Complement cascade	1.33	7.1e−02	7.3e−01
Hemostasis	1.33	7.7e−02	7.3e−01
ın by JAK−STAT signaling after Interleukin−12 stimulation	1.31	8.7e−02	7.3e−01
PIP3 activates AKT signaling	−1.30	1.2e−01	7.3e−01
Intracellular signaling by second messengers	−1.30	1.2e−01	7.3e−01
RAF/MAP kinase cascade	−1.34	6.8e−02	7.3e−01
MAPK1/MAPK3 signaling	−1.34	6.8e−02	7.3e−01
MAPK family signaling cascades	−1.34	6.8e−02	7.3e−01
FLT3 Signaling	−1.34	6.8e−02	7.3e−01
Signaling by FGFR2 in disease	−1.37	5.8e−02	7.3e−01
Signaling by FGFR in disease	−1.37	5.8e−02	7.3e−01
FGFR2 mutant receptor activation	−1.37	5.8e−02	7.3e−01
Diseases of signal transduction	−1.37	5.8e−02	7.3e−01