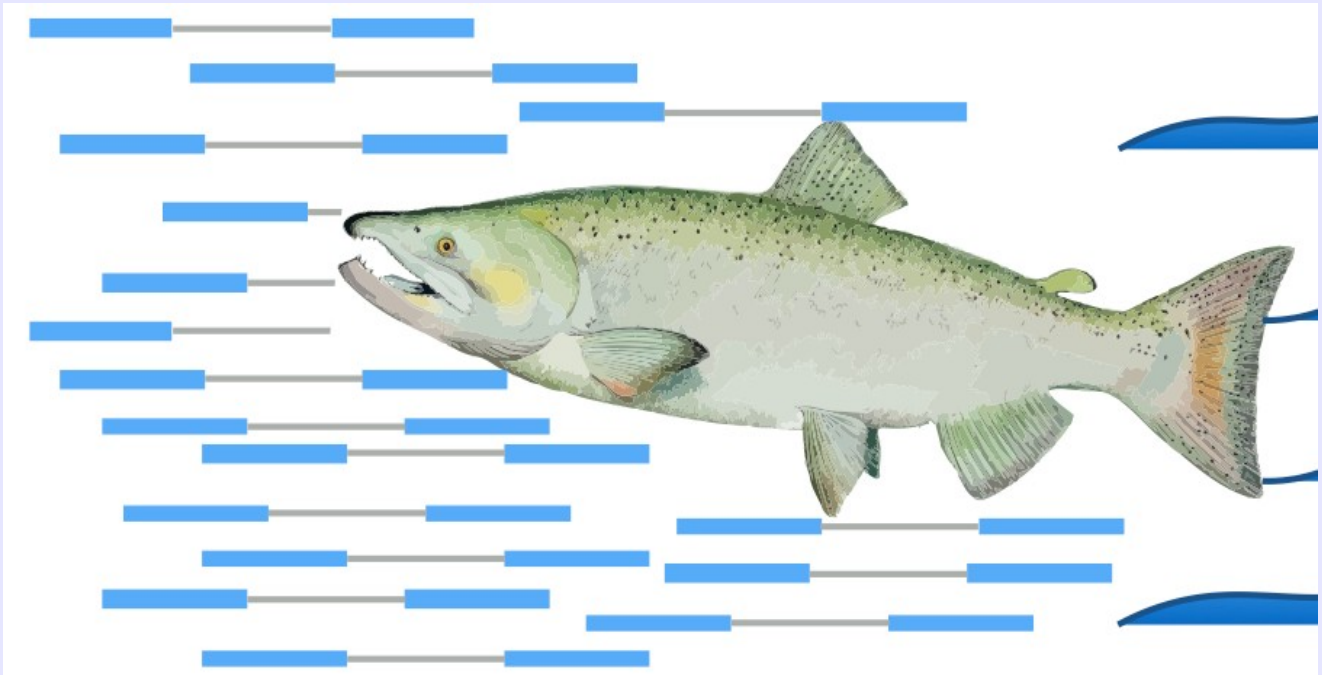
Finding the optimal cell basis for transcriptome assembling

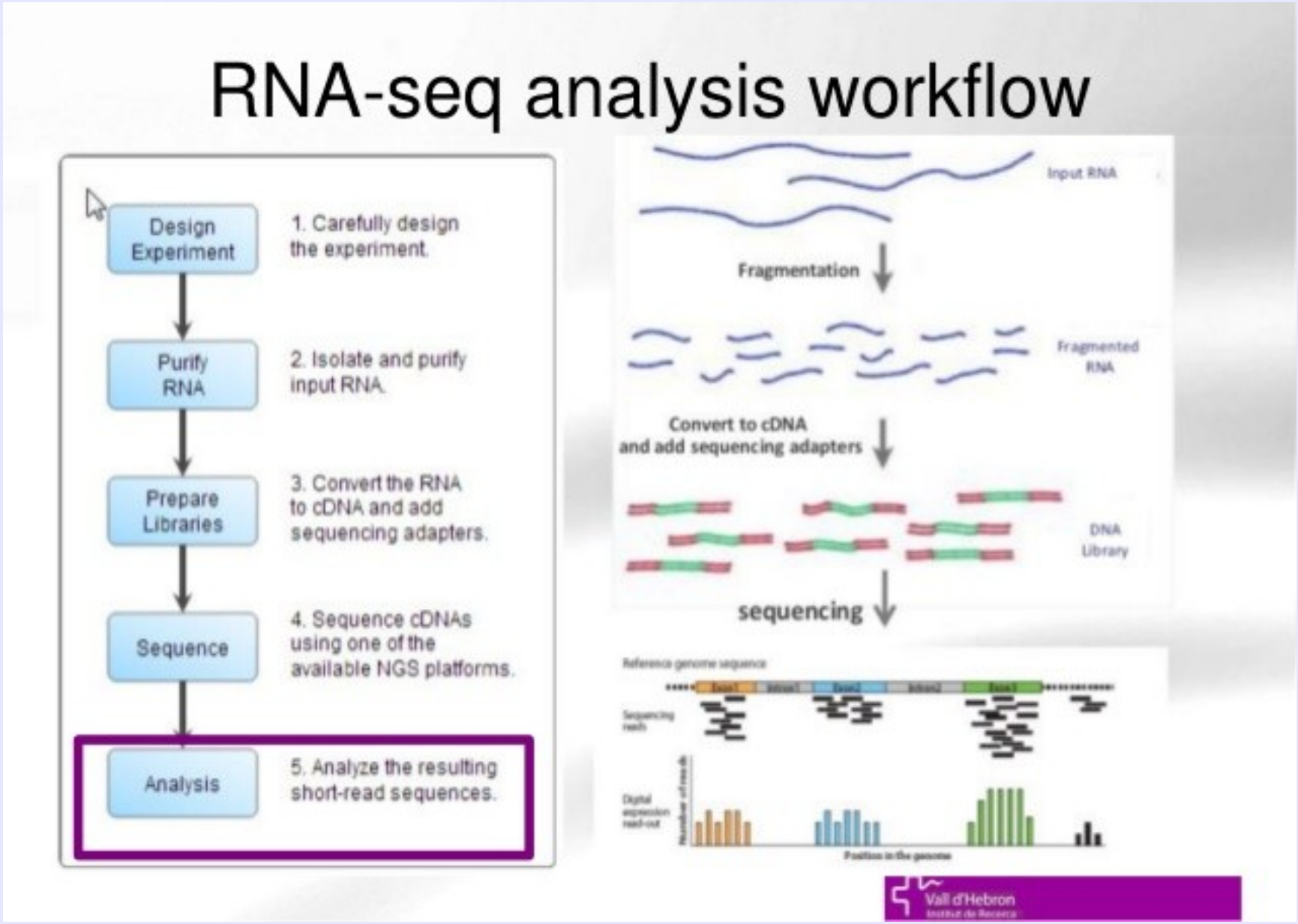
Student: Erik Zhivkoplias
Research advisor: Alexander Predeus



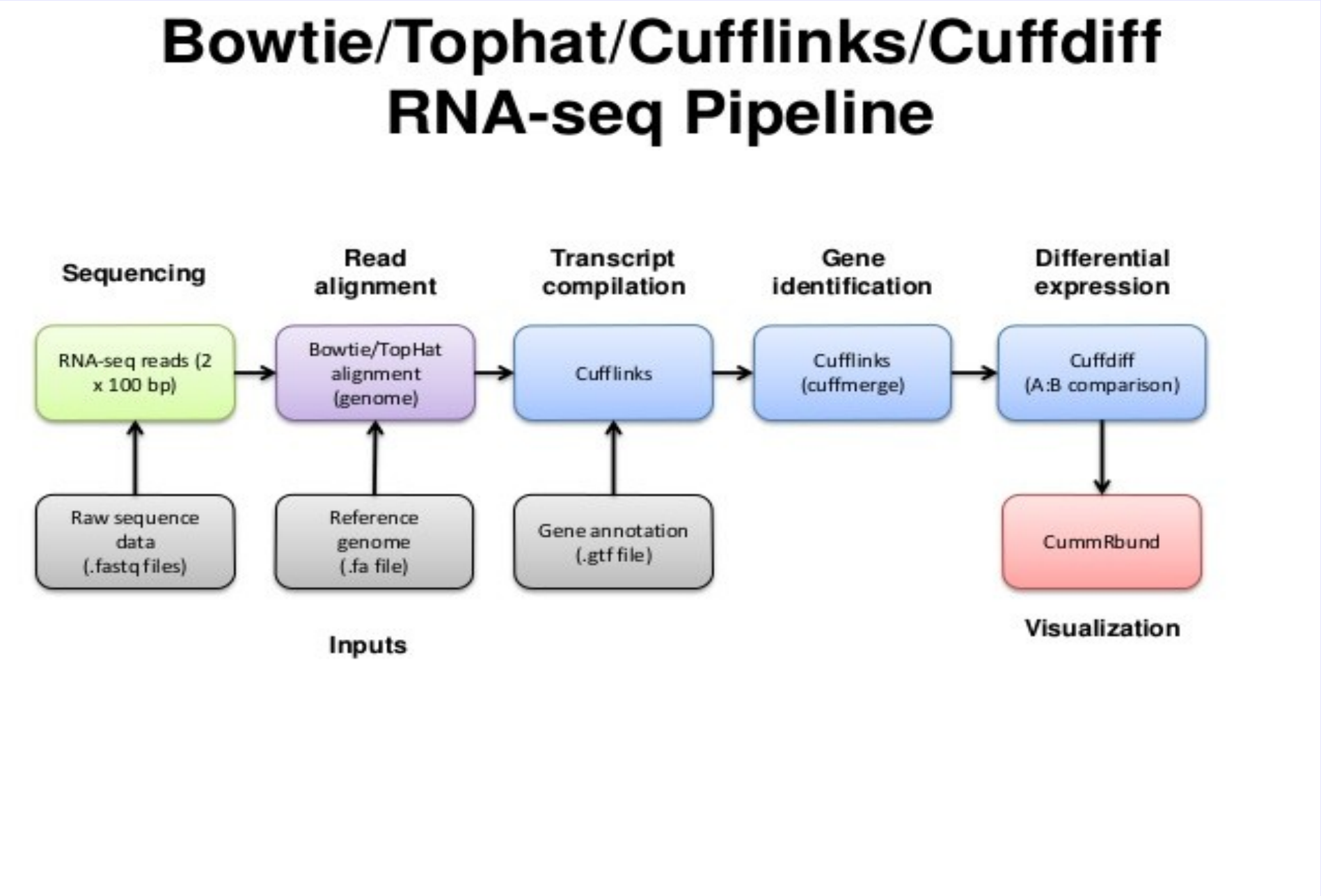
Goals to achieve

1. Get acquainted with Salmon
2. To define minumum set of cells with maximum number of expressed genes
3. Get up and running pipeline for big data meta-analyses
4. Find smth interesting

"Transcriptome" means RNA Seq



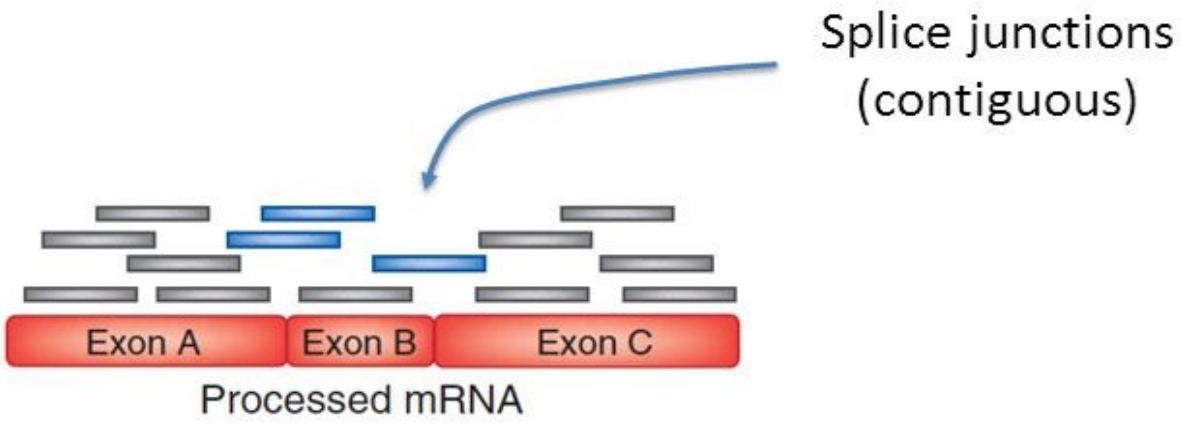
Classical pipeline



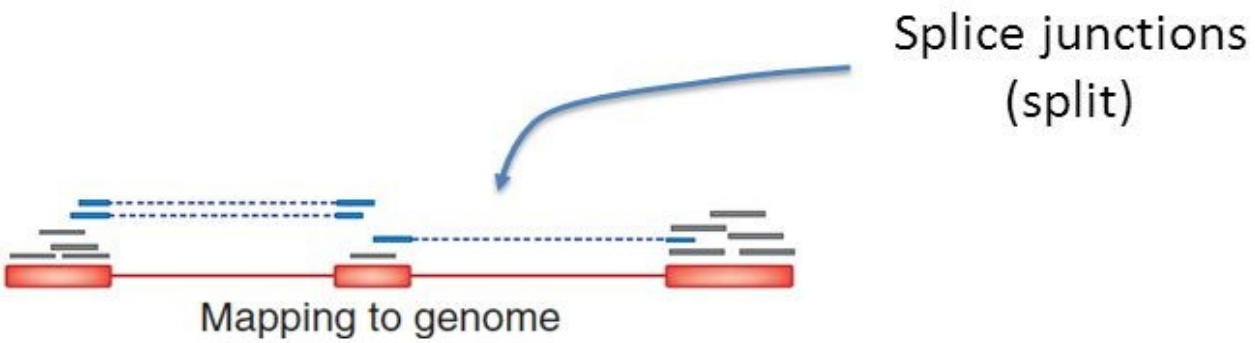
Alignment

Split read alignments for transcriptomes and structural variations

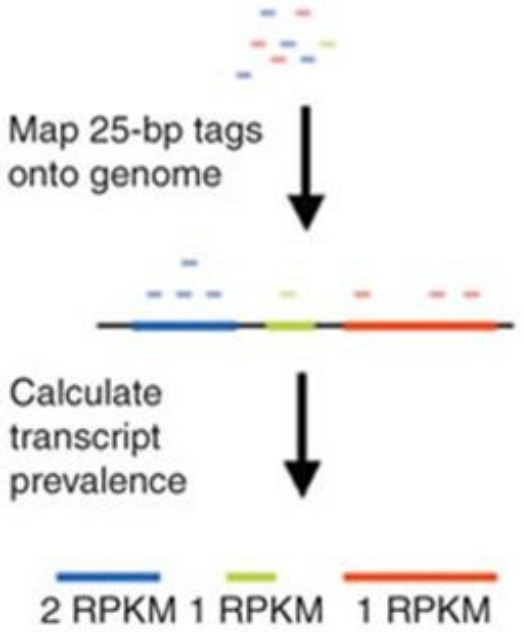
Aligning to transcripts:

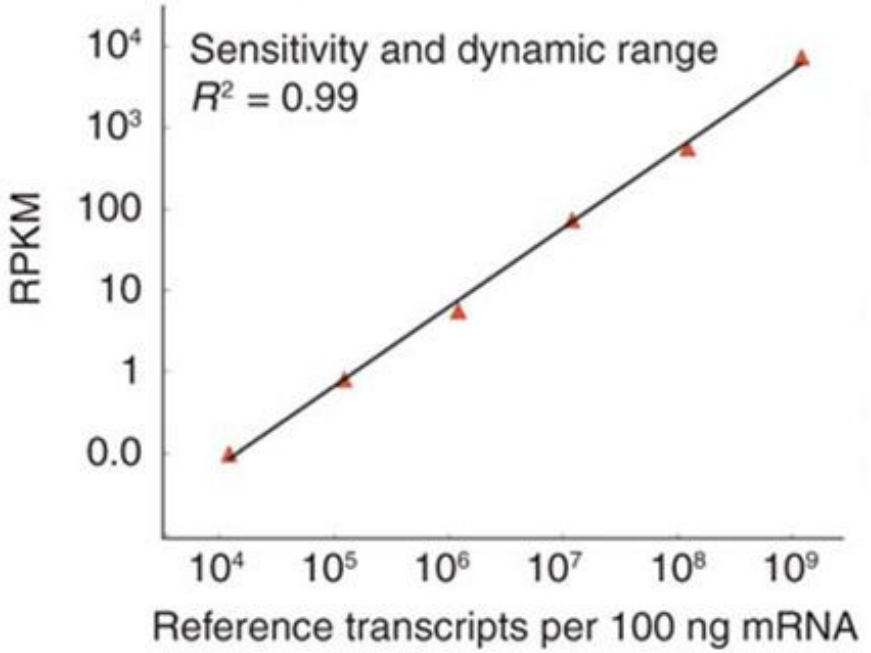


Aligning to genome:



Quantification of known transcripts



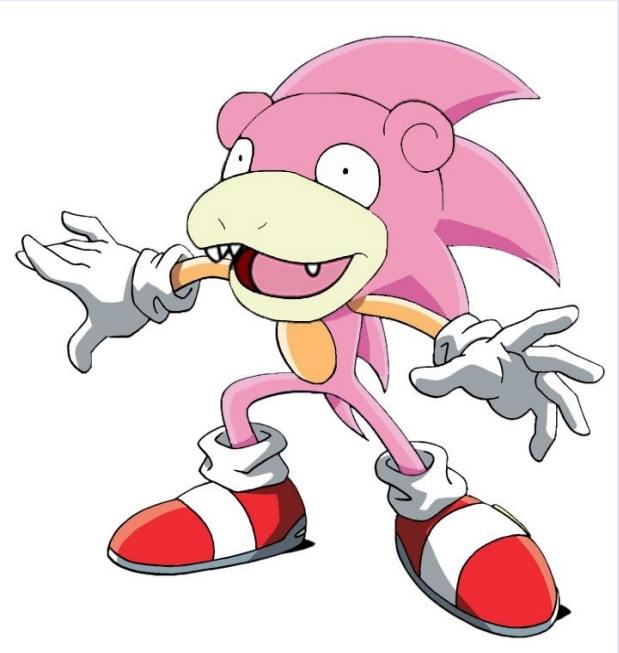


- The expression levels of known transcripts (*exon model*) are measured by the number of reads per kilobase of transcript per million mapped reads (RPKM)

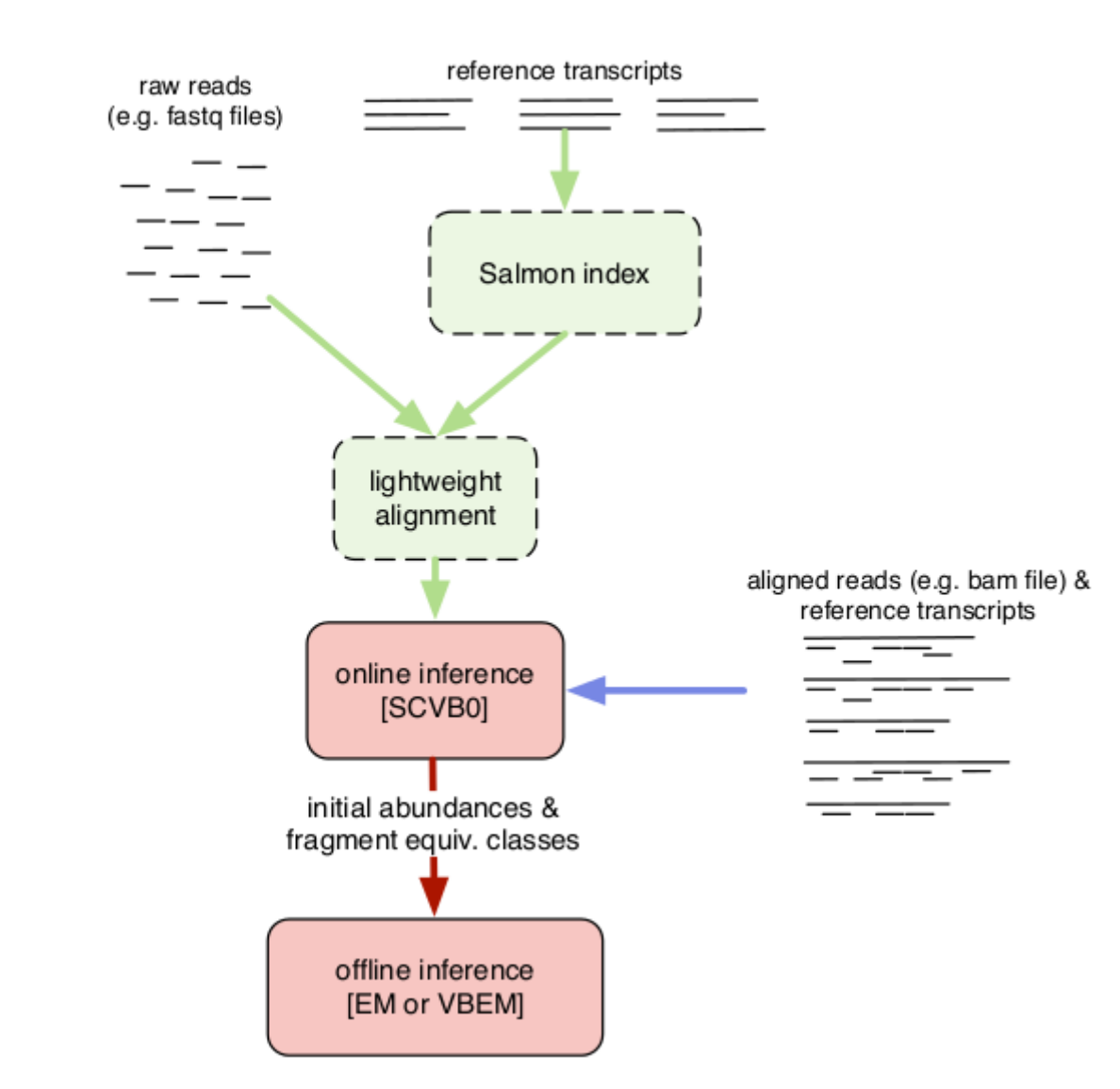
Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5, 621–628 (2008).

Problems

– Too slow!



Lightweit alignment: Salmon



1) It performs an online inference when processing fragments or alignments
2) builds equivalence classes over these fragments
3) subsequently refines abundance estimates using an offline inference algorithm on a reduced representation of the data

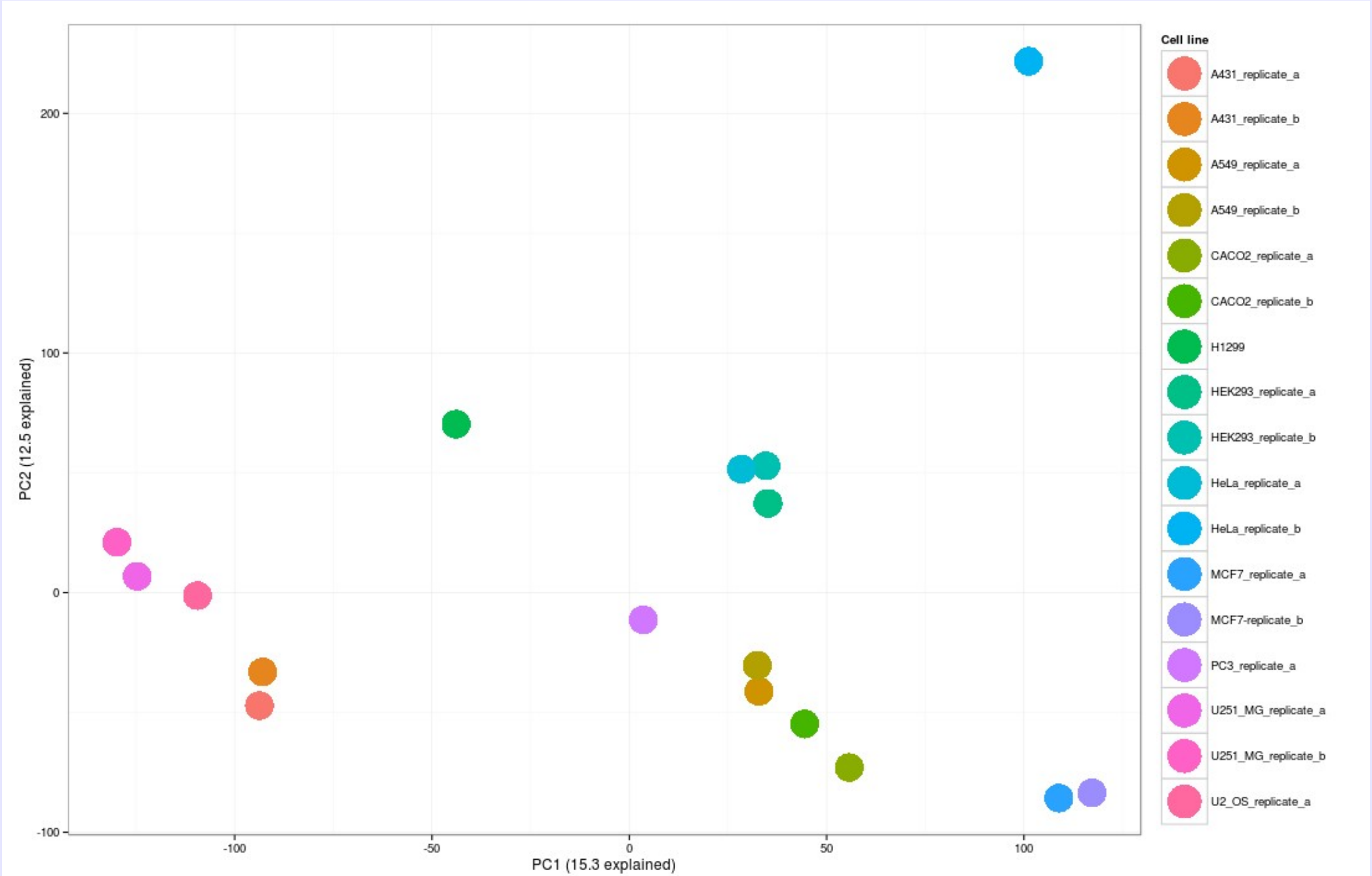
The object: HumanProteinAtlas, RNA Seq Data (32 tissues and 45 cell lines mentioned*)

*only 11 is available at the moment (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA183192)

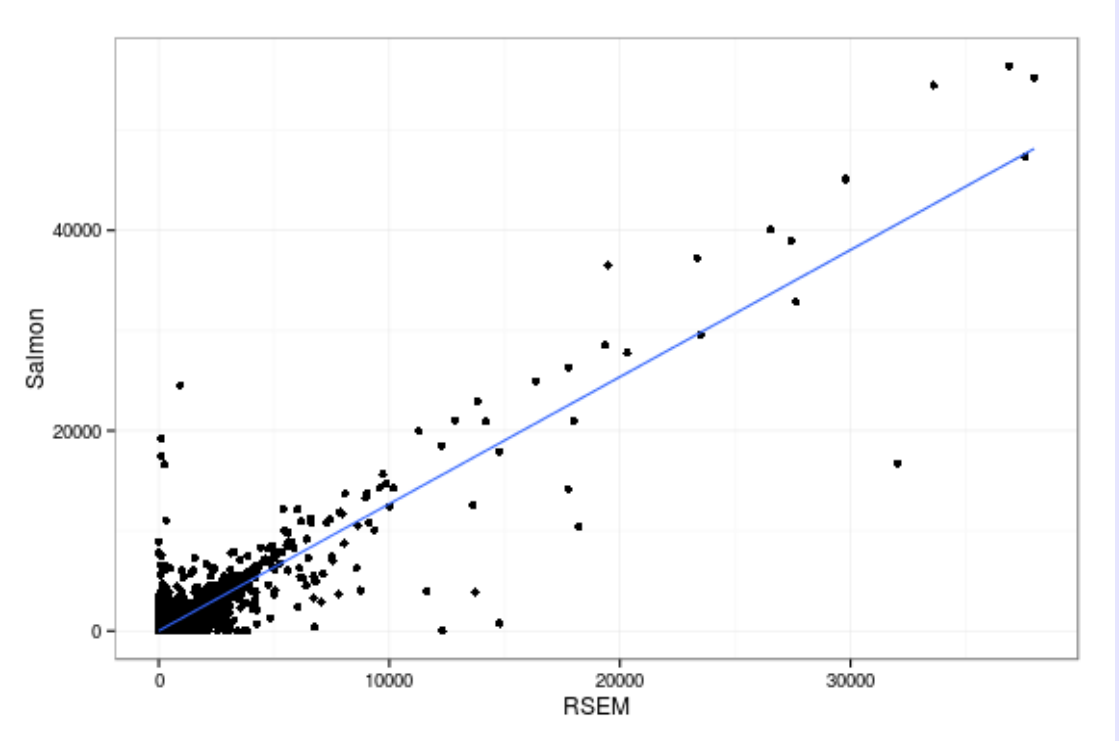
Table 1. Cell Lines Used for Subcellular Localization of All Human Proteins[a]

Cell line	Tissue origin	Description	Expressed genes (FPKM >1)	Gene expression (%)
A-431	Skin	Epidermoid carcinoma cell line	13.637	68
U-251 MG	Brain	Glioblastoma cell line	13.128	65
U-2 OS	Bone	Osteosarcoma cell line	15.478	76
A-549	Lung	Lung carcinoma cell line	13.849	69
CACO-2	Colon	Colon adenocarcinoma cell line	13.796	68
HEK 293	Embryonal kidney	Embryonal kidney cell line, transformed by adenovirus type 5	14.413	71
HeLa	Cervix	Cervical epithelial adenocarcinoma cell line	14.061	70
Hep-G2	Liver	Hepatocellular carcinoma cell line	13.724	68
MCF-7	Pleural effusion	Metastatic breast adenocarcinoma cell line	13.686	68
PC-3	Bone marrow	Metastatic poorly differentiated prostate adenocarcinoma cell line	13.889	69
RT-4	Urinary bladder	Urinary bladder transitional cell carcinoma cell line	14.210	70

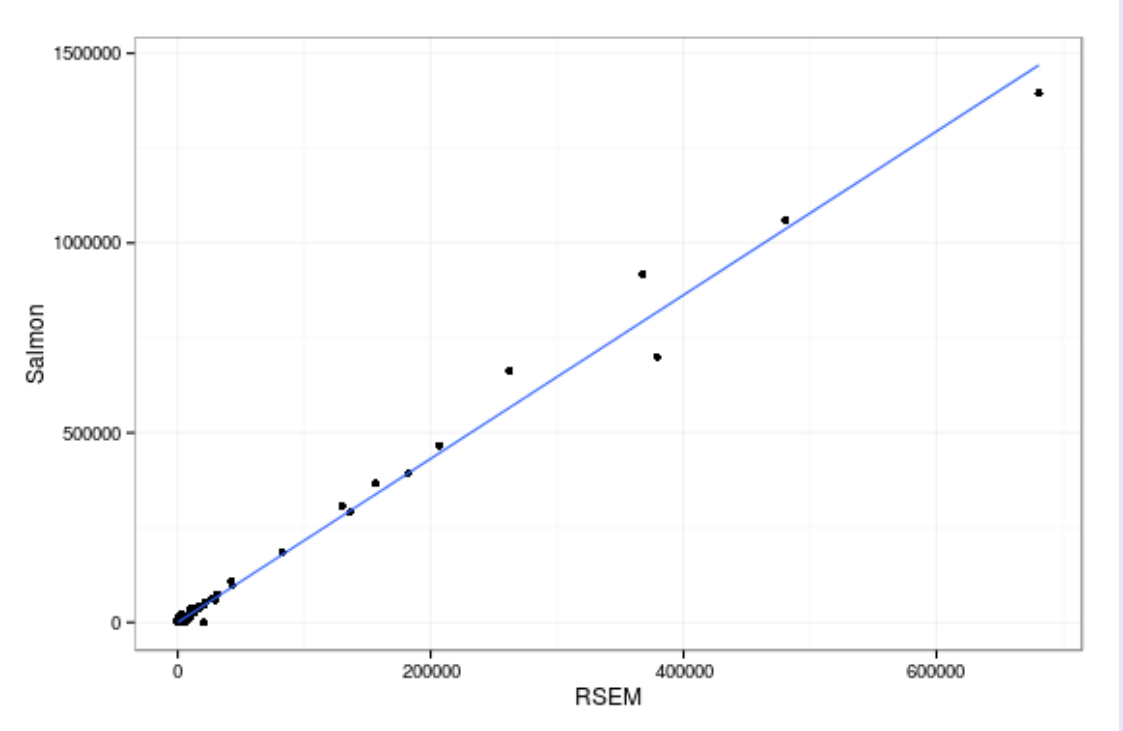
Principal Component Analysis



A-549 replicate_a
Cor w/RSEM is
0.9135

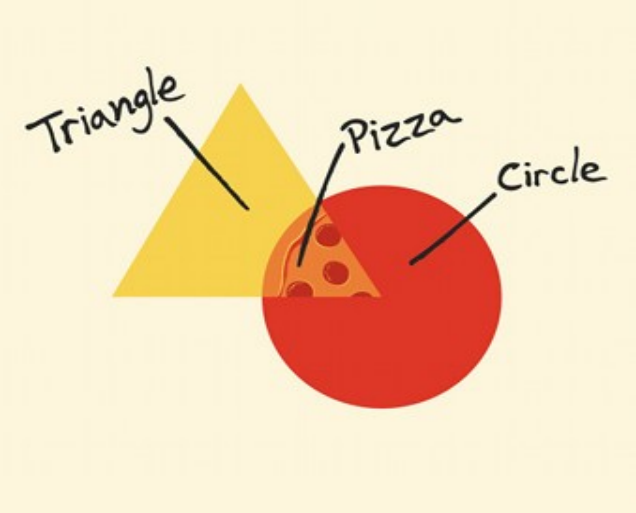


HeLa replicate_b
Cor w/RSEM is
0.9950



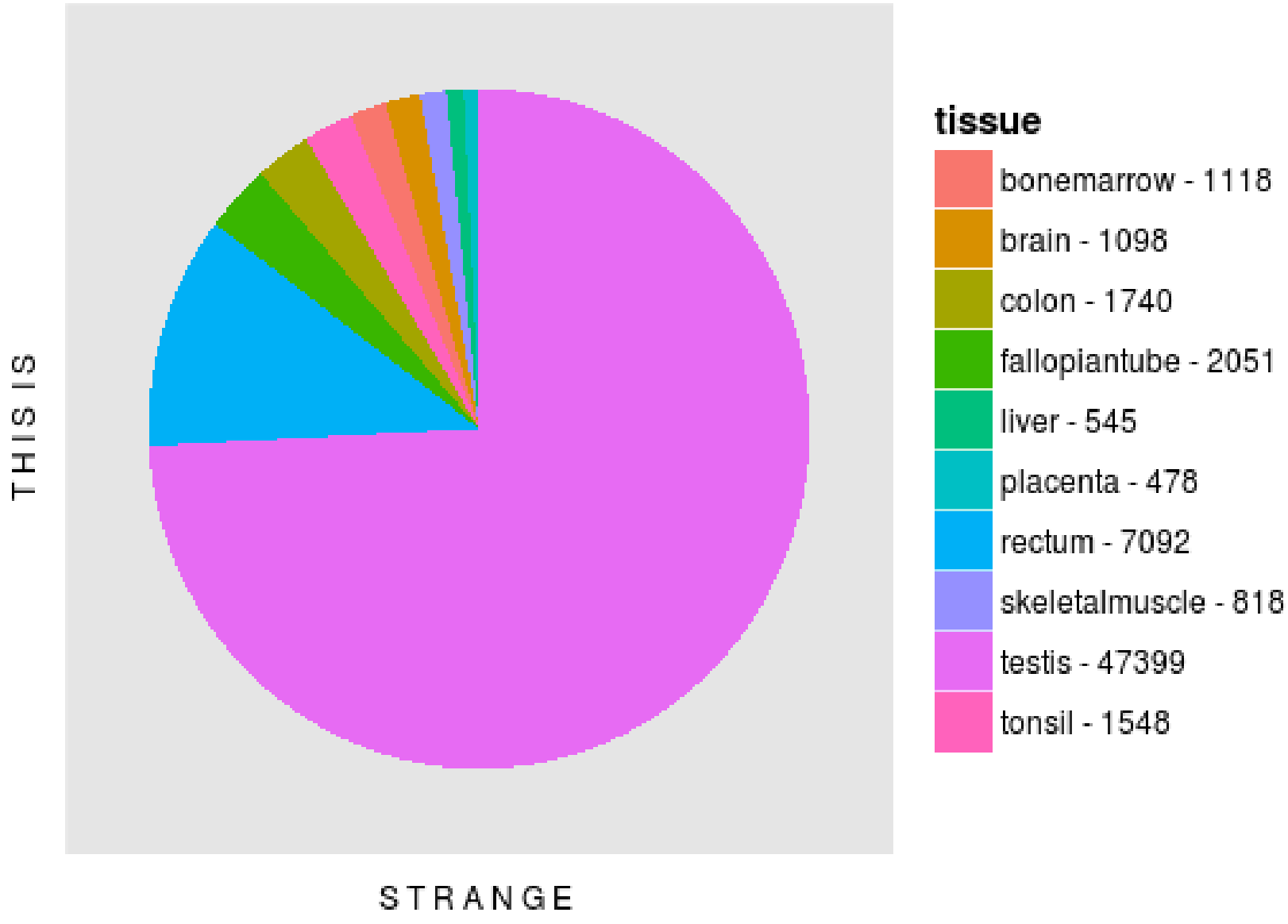
The whole idea... err.. was...

	Name.of.transcript	adrenal	appendix	bonemarrow	brain	colon	duodenum
1	ENST00000371059.7	3.7766683333	10.457026333	4.389171250	3.519710000000	12.6631363	7.6461275
2	ENST00000563752.5	0.0000000000	0.000000000	0.000000000	0.000000000000	0.0000000	0.0000000
3	ENST00000623597.3	0.0000000000	0.000000000	0.277845000	2.671563333333	2.2776887	0.0000000
4	ENST00000366544.5	40.1404033667	57.350966667	0.001598447	485.359666666667	48.0234000	7.8992625
5	ENST00000585745.5	0.0000000000	0.000000000	0.000000000	0.000000000000	0.0000000	0.0000000
6	ENST00000456900.1	4.2249516667	3.639695000	0.000000000	25.850400000000	9.7282787	4.4333550
7	ENST00000497055.1	2.8296716667	0.689799167	0.311446250	39.396300000000	1.3275913	0.0000000
8	ENST00000361745.10	7.9828666930	0.000000000	0.000000000	26.492600000000	5.5403252	0.0000000
9	ENST00000400899.2	0.0000000000	0.000000000	0.198630000	0.000000000000	0.1612163	0.0000000
10	ENST00000489188.1	6.0176580000	11.208545000	34.418751250	27.573400000000	27.9829062	14.6421500
11	ENST00000593011.5	0.2163666667	0.000000000	0.000000000	0.000000000000	0.0000000	0.0000000
12	ENST00000470278.5	12.4424407420	18.218565000	3.602437500	36.795466666667	21.8007087	0.9952600
13	ENST00000367184.2	0.0000000000	0.317681667	0.366832500	0.479143333333	1.2685775	0.0000000
14	ENST00000367207.7	0.0000000000	0.000000000	0.162492500	0.000000000000	0.0000000	0.0000000
15	ENST00000372990.5	0.2884133333	15.555643333	0.000000000	108.964366666667	512.5403875	4.3575975
16	ENST00000368630.7	144.3063333333	205.517000000	453.362000000	219.378033333333	304.5144250	88.3060500
17	ENST00000526773.2	13.7495166667	5.064496667	1.771304875	18.885077666667	34.3793000	83.8631500
18	ENST00000456298.5	1.1203000000	2.399515000	0.000000000	1.174013333333	10.7381538	2.8003000
19	ENST00000422198.1	0.0000000000	0.000000000	0.000000000	0.000000000000	0.0000000	0.0000000
20	ENST00000532087.1	0.2150350000	1.070963333	0.000000000	0.000000000000	1.3171775	0.0000000
21	ENST00000611702.4	1.1091066667	2.406916667	0.932672121	3.725120000000	10.8144800	0.0000000
22	ENST00000539444.5	0.0000000000	1.515263333	0.000000000	0.000000000000	0.5828912	0.0000000
23	ENST00000628080.1	0.0000000000	0.000000000	0.000000000	0.000000000000	0.0000000	0.0000000

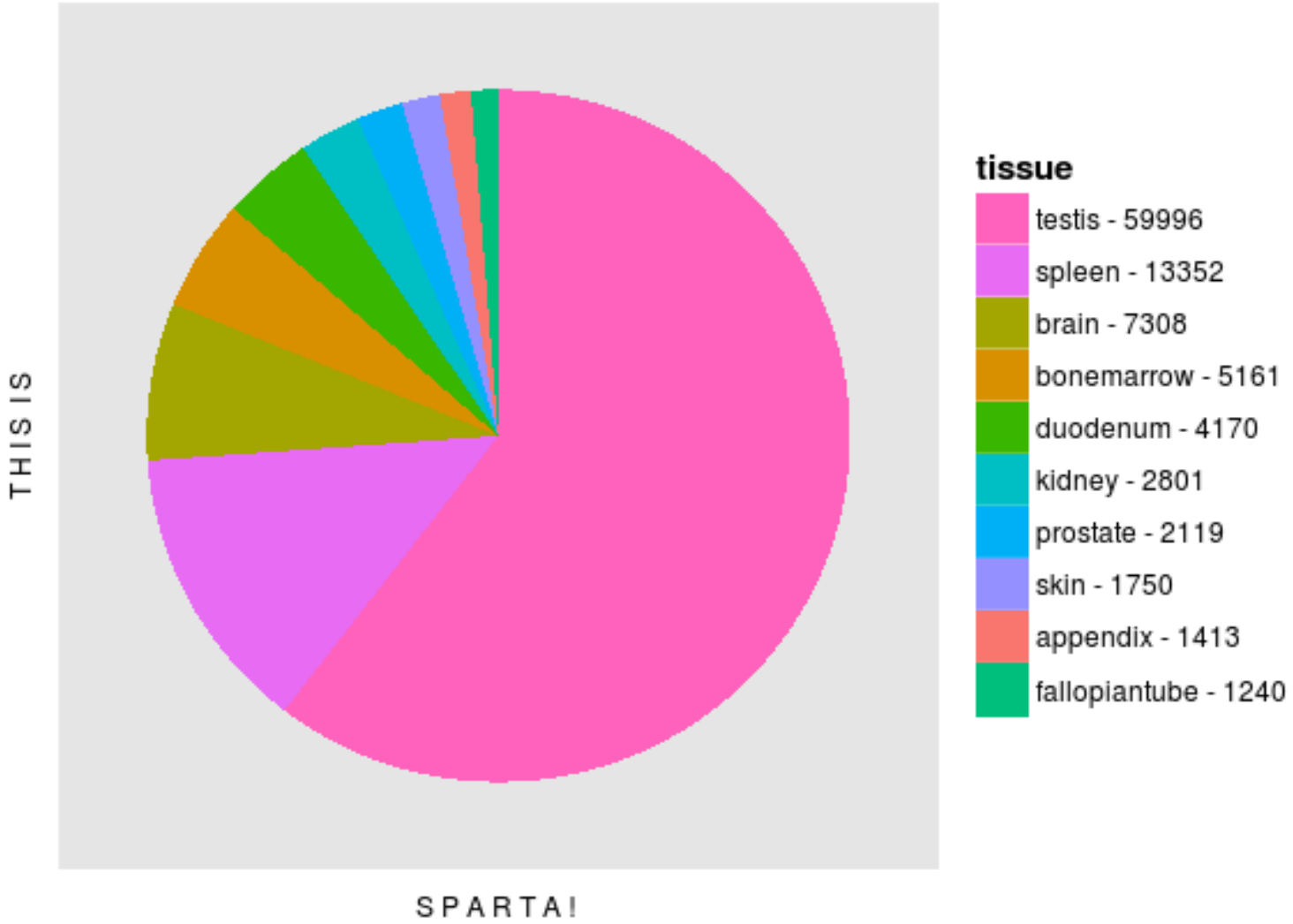


– To sum up all data (replics and individuals)
– Normalise it.
– Define the threshold for "good-covered" transcripts
– Count the number of "good-covered" transcripts
– To sum up tissues like sets one after another

Smth went wrong ;(

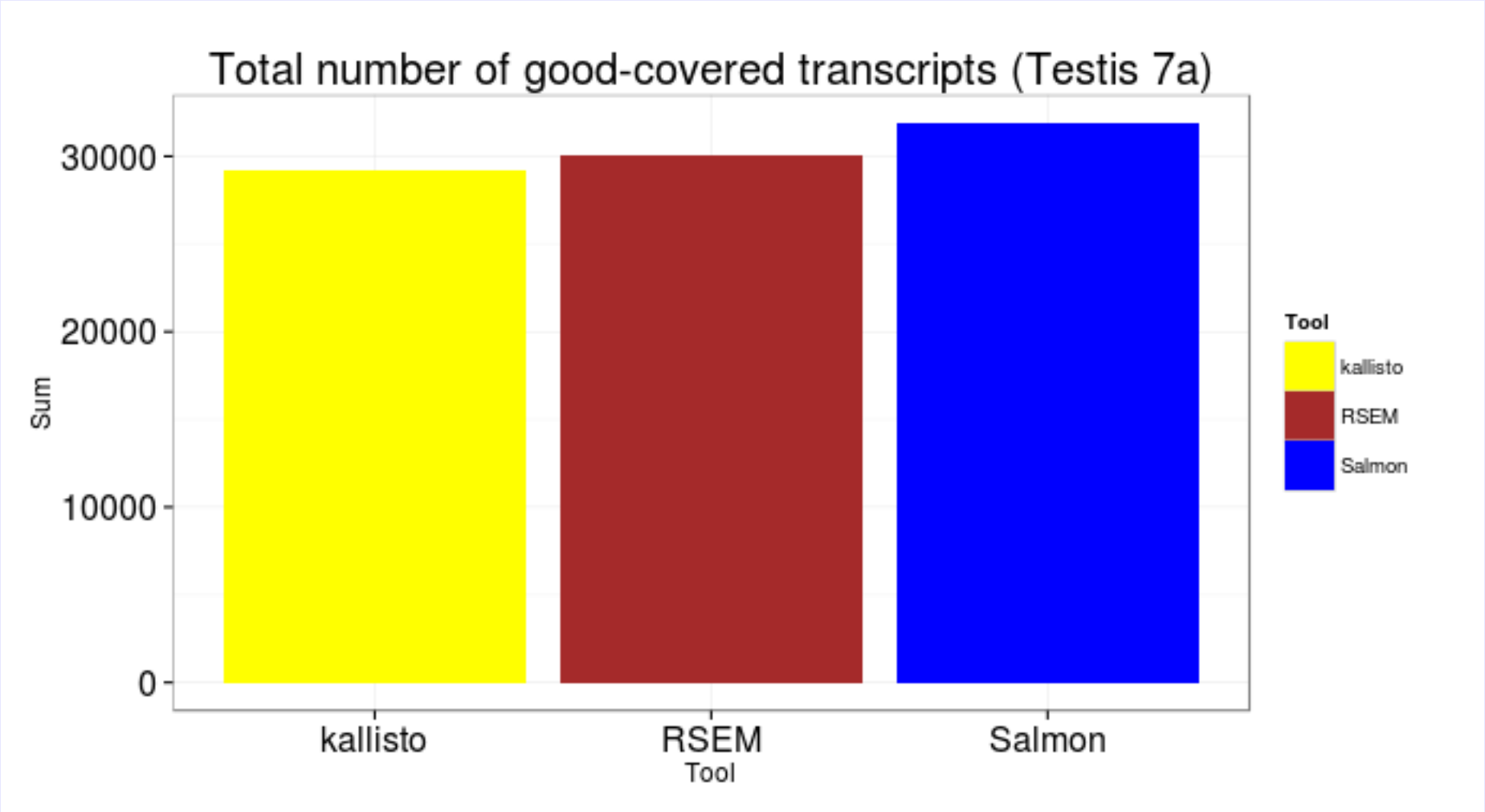


Yeah! That's better




SAGE
SIMMS

What about RSEM?



Results

1. Got acquainted with Salmon
2. The minumum set of cells with maximum number of expressed transcripts based on E-MTAB-2836 is defined
3. Got "so-so" but running pipeline for big data meta-analyses
4. Learned lots of interesting staff

Tnanks for listenin!