exceRpt updates

RK + JR

2015 - 07 - 22

agenda

1. **updates to endogenous alignment**

2. **support for *N random barcodes**

updates to endogenous alignment

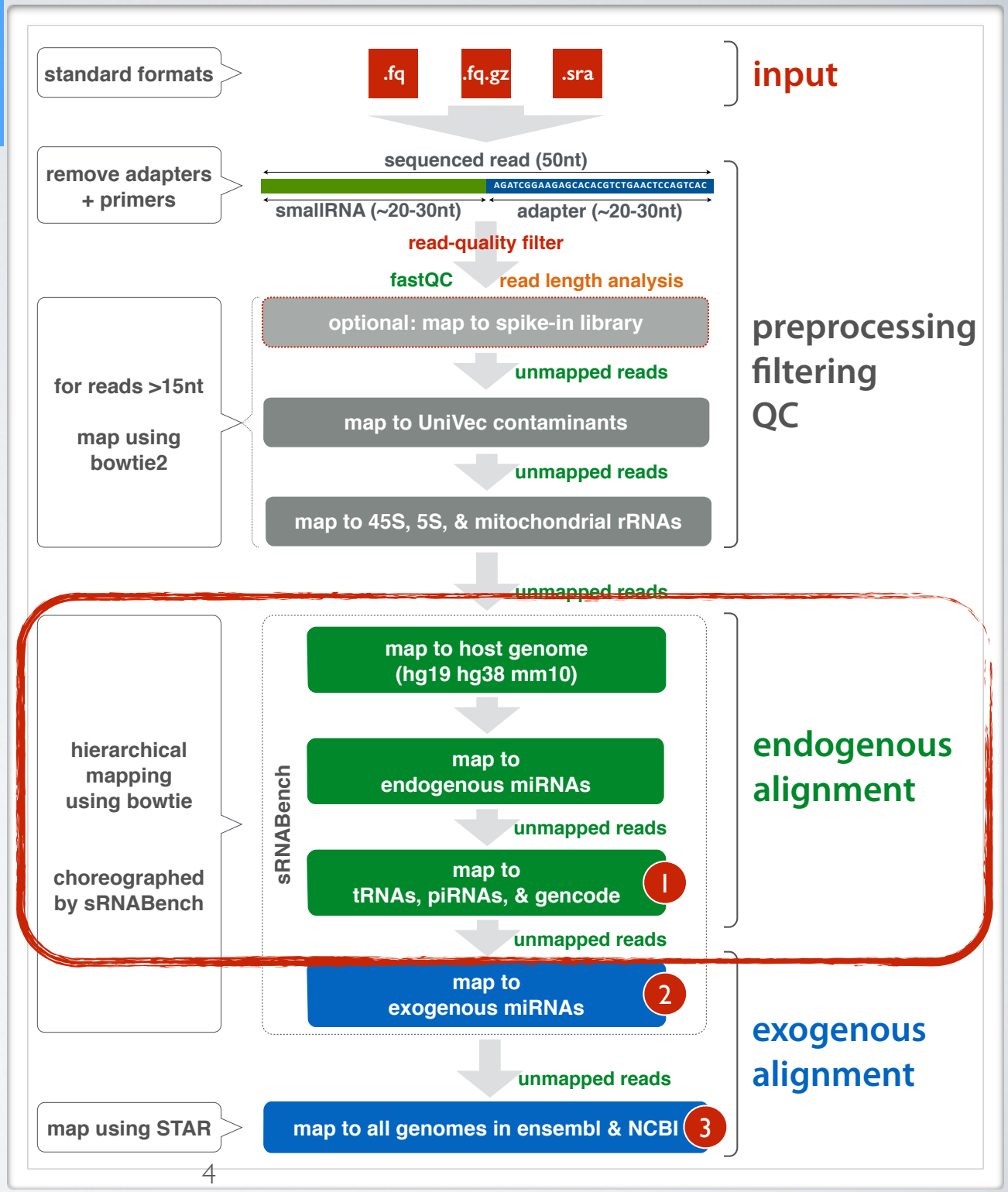
- **sRNABench** was not performing as desired for reads **multi-mapping to non-miRNA libraries**
- causing problems for **tRNA/piRNA** quants in particular
- decided on complete overhaul- replaced sRNABench with **custom endogenous alignment & quantification**
- some of **Anna Krichevsky's** samples show differences with analysis done at the BGI:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		20_3_1_5M_	20_3_140mL_	20_3_140mL_	20_3_140mL_	GBM4_1_5M	GBM4_130m	GBM4_130m	GBM4_130ml	GBM8_1_5M	GBM8_120m	GBM8_120m	GBM8_120ml
2	Genboree miRNA_sense	95,216	298,464	273,853	42,253	31,694	525,689	854,375	127,101	19,544	239,378	171,377	54,900
3	BGI miRNA	77,935	230,221	168,017	15,230	15,182	83,268	603,115	66,669	17,369	73,891	114,595	29,827
4													
5	Genboree piRNA_sense	89,583	73,652	13,710	33,040	69,970	19,833	8,691	43,206	63,115	58,943	13,824	34,698
6	BGI piRNA	158,751	165,836	222,554	967,647	178,782	38,182	34,590	307,976	166,622	153,096	39,610	807,981
7													
8	Genboree tRNA_sense	842	3,032	1,319	3,481	2,153	477	165	3,097	3,183	2,760	260	8,655
9	BGI tRNA	339,445	1,142,863	1,111,092	5,275,495	547,931	174,482	83,684	1,210,903	757,849	907,304	102,407	4,363,225
10													

exceRpt

- automatic pre-processing and QC of sequence reads
- absolute quantitation by quantification of exogenous spike-in sequences
- explicit rRNA filtering & QC
- quantify many different smallRNA types
- choice of 3 end-points

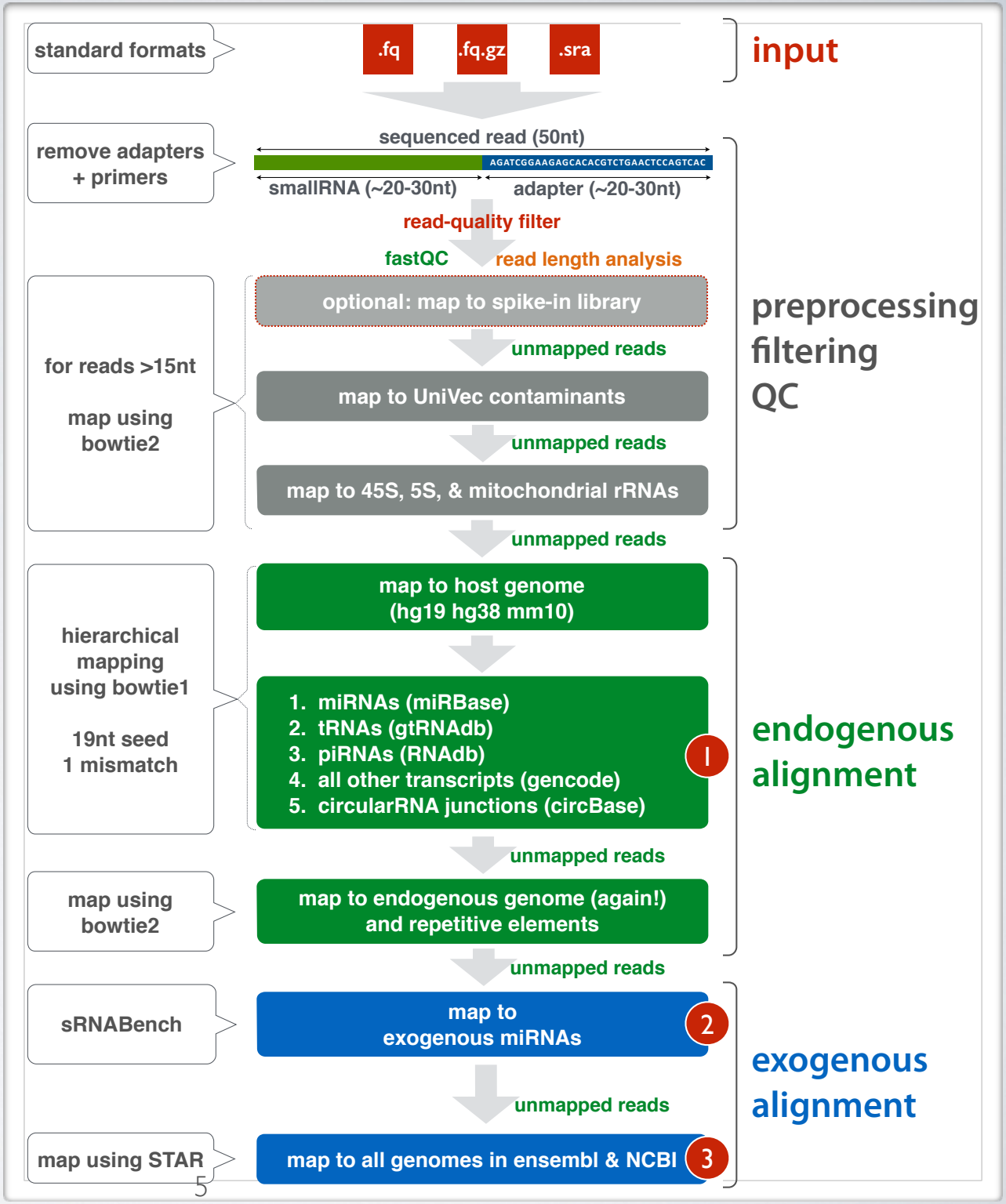
1 2 3



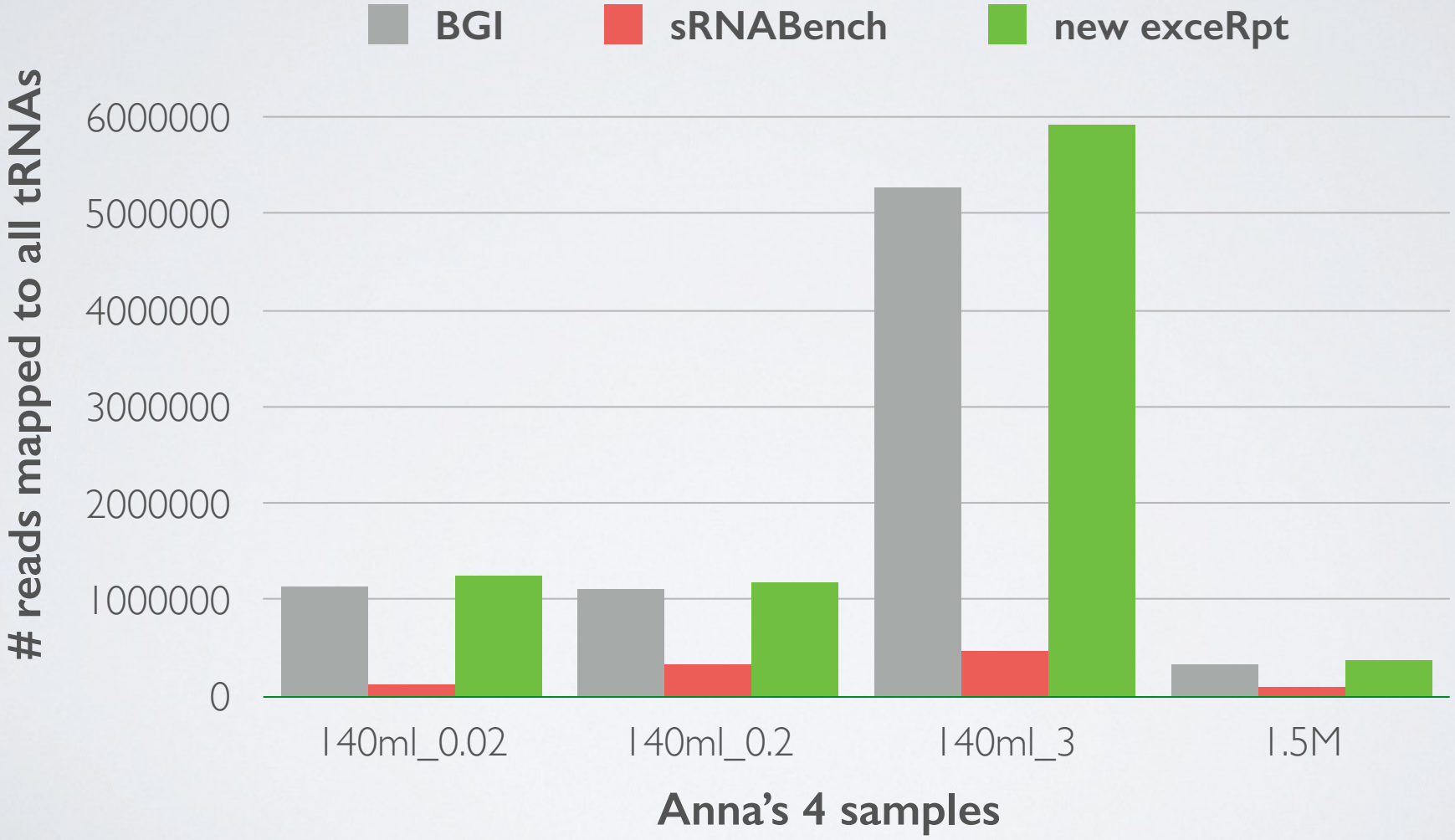
exceRpt

- automatic pre-processing and QC of sequence reads
- absolute quantitation by quantification of exogenous spike-in sequences
- explicit rRNA filtering & QC
- quantify many different smallRNA types
- choice of 3 end-points

1 2 3



Anna Krichevsky's missing tRNAs



advantages

- much more **reliable quantification of non-miRNA** libraries
- full use of **read qualities** during alignment
- can **prioritise alignments** to different classes of RNA
- output genome **alignments in BAM/WIG** for viewing in a browser
- much **better control over memory** usage
- fully **modular species databases**
- do not need bowtie1/ViennaRNA on the **PATH**
- **faster*!**

*for most samples

agenda

1. **updates to endogenous alignment**

2. **support for *N random barcodes**

support for *N random barcodes

- several investigators moving toward use of random sequences to detect amplification artefacts:

  3' adapter - 5' **NNNN**ATCACCGACTGCCCATAGAGAGG 3'
  5' adapter- 5' GGCCAAGGCG**NNNN** 3'

  sequenced read- 5' **NNNN**<insert smRNA>**NNNN**ATCACCGACTGCCCATAGAGAGG 3'

- with **two** 4N random barcodes we potentially get 4^8=65,536 unique reads for each **insert** sequence

- need to support **identification**, **removal**, and **quantification** of these random sequences - also, do they help **normalisation**?

*N random barcodes - app

- created a Java app to read a fastq and remove the *N barcodes
- for this example:

sequenced read- 5' **NNNN**<insert smRNA>**NNNN**ATCACCGACTGCCCATAGAGAGG 3'

- the command to run would be:

```
java ProcessFastqWithRandomBarcode –n 4 ––5p ––3p ––stats clippedReads.fq > clippedReads.noRand.fq 2> clippedReads.stats
```

 # bases in random barcode

look for barcode at 5’ of read

look for barcode at 3’ of read

compute barcode frequency statistics

input reads

output reads

output statistics

*N random barcodes - statistics

- as well as removing barcodes, we can use them to help **normalise the read-counts** for amplification artefacts:

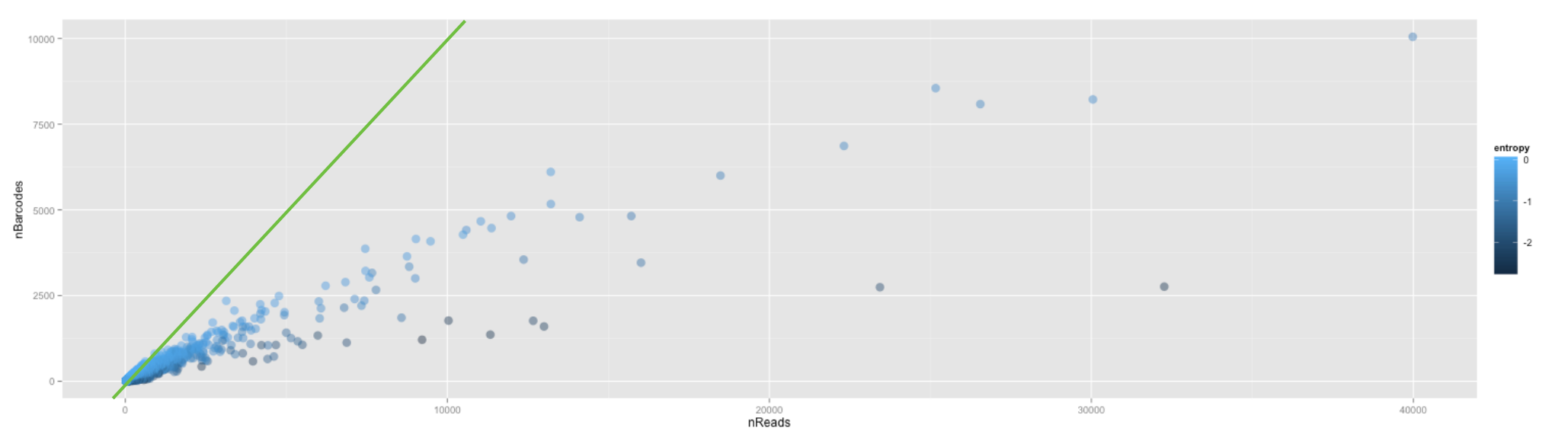
	seq	nReads	nBarcodes	entropy	barcode_1	barcode_2	barcode_3
1	ACGGCCCTGGCGGAGCGCTGAGAAGACGGTCGAACTTGACTATCT	39980	10048	-0.815	TGGGTCTC 150	TAGGTCTC 148	TTGGTCTC 98
2	CCCCACAACCGCGCTTGACTAGCTTGCTGTTT	32263	2760	-2.392	GTGGGGGA 1157	GTGAGGGA 1017	GTGGGGGG 616
3	AACTTGACCGCTCTGACCA	30047	8221	-0.580	GGCGAGTT 108	GGCAAGTT 62	GGTGAGTT 59
4	CGGCCGGGCGCGAC	26551	8085	-0.704	GTGGGGCC 103	GGGGGGCC 81	TAGGGGTA 65
5	ACGGCCCTGGCGGAGCGC	25164	8550	-0.584	TAAGCGTA 77	TAGGCTCG 51	TAGGCGTA 46
6	GAGGCGTCCAGTGCGGTAACGCGAC	23434	2744	-1.906	GTGGGCTC 498	GTGGGCTT 397	GTGGGCAC 384

- we calculate an **relative entropy score** (*KL divergence*) for each unique insert, more negative values indicate more 'order', less randomness, to the barcodes:

	seq	nReads	nBarcodes	entropy	barcode_1	barcode_2	barcode_3
1911	TGAATCACCGACTGCCCACTAGAGAGGCTGAGACTGCCAAGGCACACAGGGGA	190	10	-2.743	TGATTAGG 176	GTGATAGG 6	GTGGTAGG 1
2626	TCAAATCACCGACTGCCCACTAGAGAGGCTGAGACTGCCAAGGCACACAGGGGA	142	13	-2.702	GTGATAGG 123	GAGATAGG 6	ATGATAGG 2
25	CCCCCCACTGCTAAATTTGACTGGCTT	9215	1211	-2.698	GTGGGGGG 913	GTGGGGGA 715	GTGTGGGG 385
554	AGTAATCACCGACTGCCCACTAGAGAGGCTGAGACTGCCAAGGCACACAGGGGA	574	39	-2.642	GTGTTAGG 325	GGATTAGG 58	GTGCTAGG 45
2680	TTAAATCACCGACTGCCCACTAGAGAGGCTGAGACTGCCAAGGCACACAGGGGA	139	10	-2.611	GTGATAGG 126	GAGATAGG 5	GTGCTAGG 1
2382	ACCTACACCGACTGCCCATAGAGAGGCTGAGACTGCCAAGGCACACAGGGGA	154	25	-2.491	GTGTTAGG 106	GTGCTAGG 11	GAGTTAGG 7

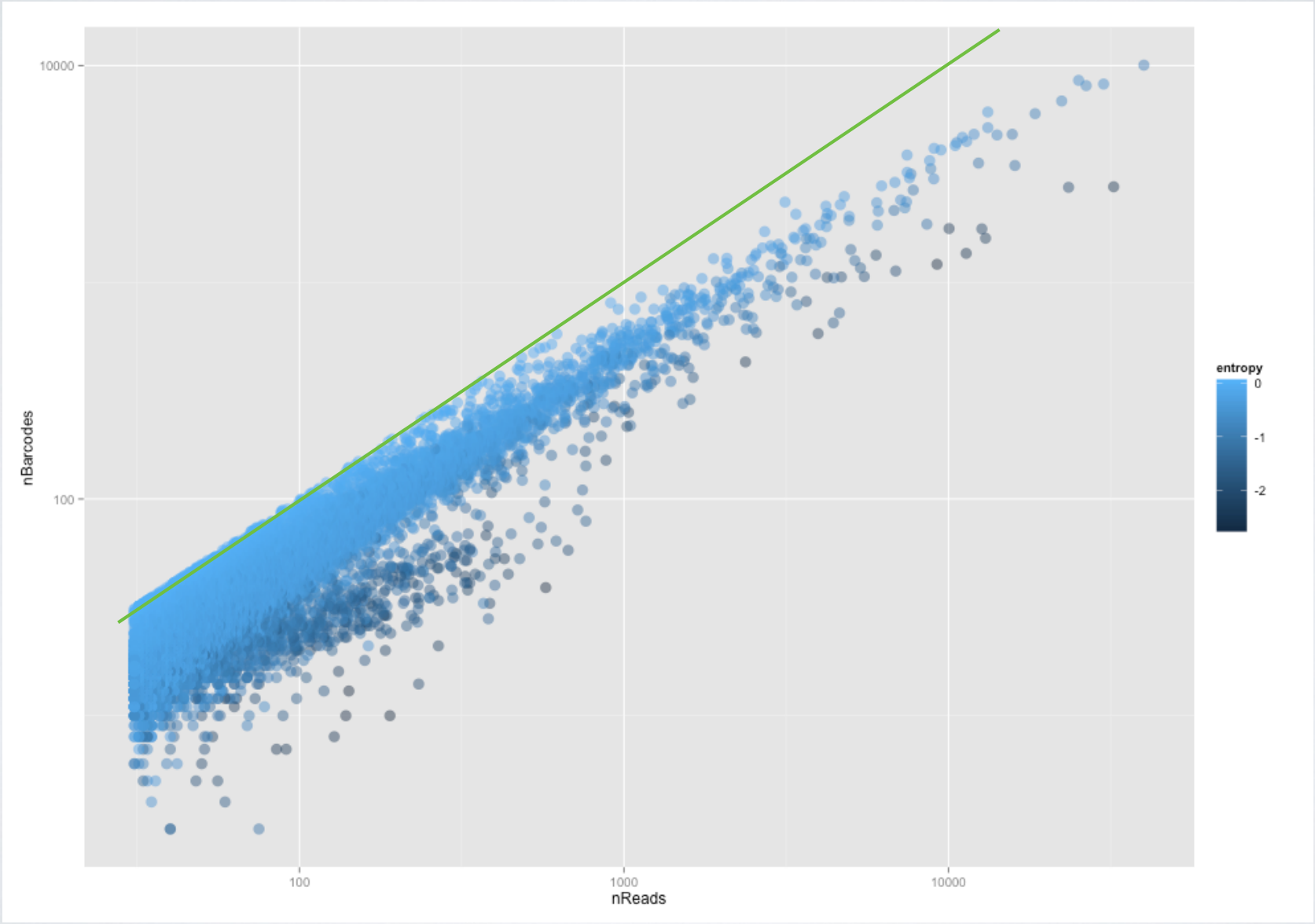
*N random barcodes - counts

- easy to plot # reads against number of barcodes:

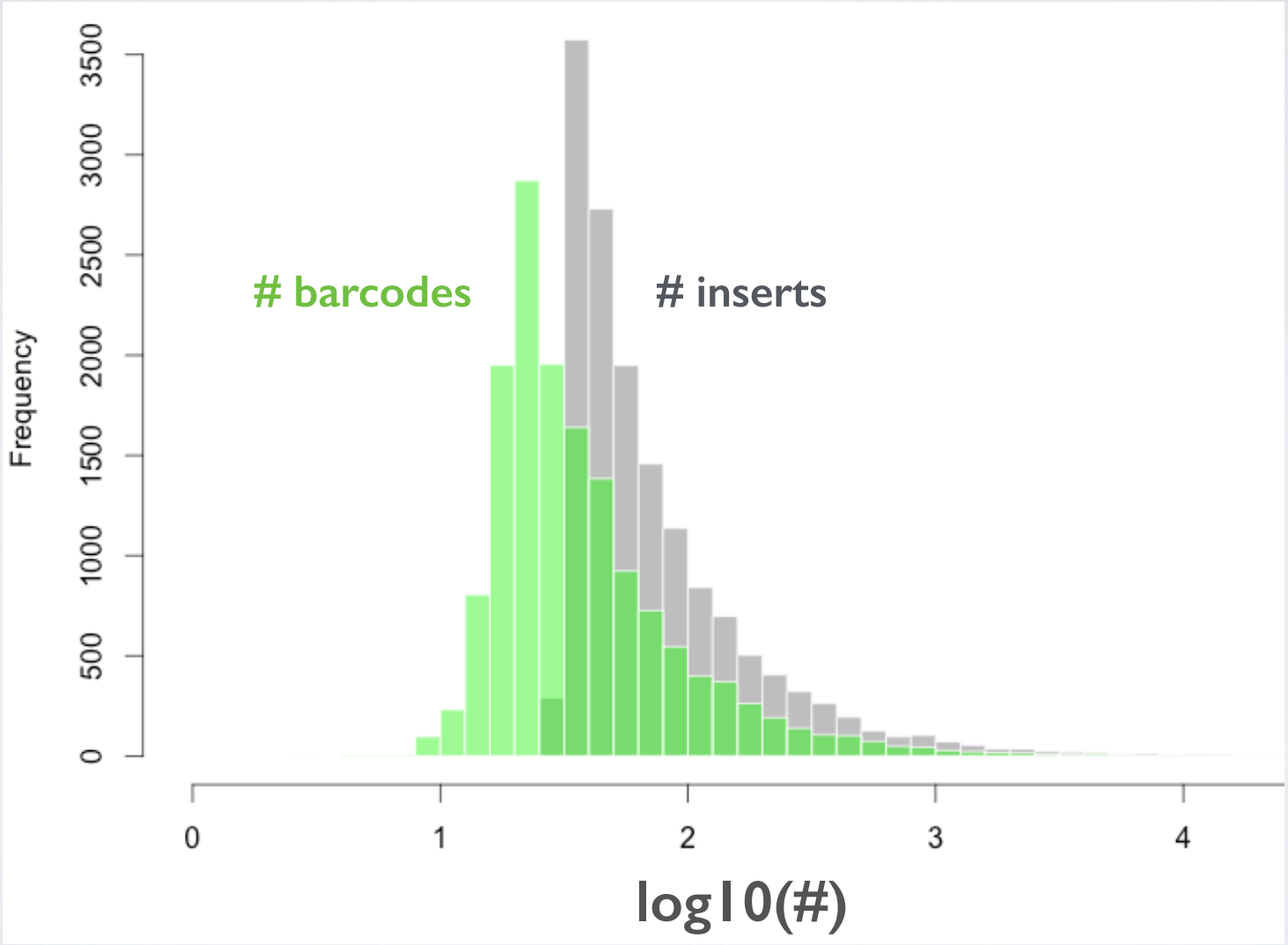


- some inserts have many more reads than they 'should' get according to the barcodes

*N random barcodes - counts



*N random barcodes - counts



agenda

1. **updates to endogenous alignment**

2. **support for *N random barcodes**

3. **to-do**
   - include random barcode removal/analysis in pipeline
   - further explore barcode quant for normalisation & QC
   - modify R script for multi-sample merge for new pipeline
   - compute reference-sequence similarity matrix/network