Genomic file types

Quality Assessment and Control of Sequence Data

Workshop on Genomics, 2012

File types overview

http://evomics.org/resources/file-formats/

- Fasta/fasta qual
- Fastq
- SAM

Text files

- BAM
- sff

Binary files

- …
- …

Fasta

- Most basic file format to represent nucleotide or amino-acid sequences
- Each sequence is represented by:
  - A single description line (shouldn't exceed 80 characters):
    - Starts with ">"
    - Followed by the sequence ID, and a space, then
    - More information (description)
  - The sequence, over one or several lines (the number of characters per line is generally 70 or 80, but it doesn't matter)

```
>Protein1 Description of protein 1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGK
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEER
>DNA1 Description of dna segment 1
AACTCTCGCGTAGCTCAGAGAAGAGCTTGATCGATCGTGCTGCTGCTA
CCGCTAGTAGCTGTAGATCGTGCTAGTCAGCATCGATGCTAGCTAGCT
```

Qual (aka fasta qual)

- Fasta-like quality format
- Always paired with a fasta file (sequences with same ids, same order)
- Description line as in fasta format
- Qualities: a number for each base in the corresponding fasta, separated by spaces
- Can be gzip-ped and used as such by some programs

```
>DNA1 quality of dna segment 1
23 23 22 24 23 12 3 34 23 23 22 21 23 34 32 9 14
22 24 23 12 3 34 23 23 22 21 23 34 32 9 14 25 21
```

Quality - Phred scores

- Most common representation of qualities
- Related to the probability of errors (*P*) in a particular base

$$Q = -10 \log_{10} P$$

$$P = 10^{\frac{-Q}{10}}$$

Phred score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
...	...	

- Solexa runs < 1.3 use a different calculation:
  - Equivalent for high quality
  - Different for low quality (negative values of Q allowed)

FastQ

- A more compact format to store sequence and qualities
- Normally on 4 lines:
  - “@” and the sequence ID
  - Sequence
  - “+” (and the sequence ID)
  - The quality score

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGA
+
!''*((((***+))%%%++)(%%%%
```

Example taken from Wikipedia

Fastq (ctd.)

- Quality score:
  - ASCII encoding of phred scores
  - Sanger has one scale
  - Illumina has 3 differents
- Can be gzip-ped and used as such by some programs

ASCII decimal codes

Dec	Symbol	Dec	Symbol	Dec	Symbol
32	Space	64	@	96	`
33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46		78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r

FastQ – quality values

- Solexa picked different quality definition and ranges over time, all different from Sanger values
- Ask your sequence provider!
- Guessing by getting the range of all values in all/many reads (not foolproof)

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ................................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Example taken from Wikipedia

SAM/BAM

- SAM (Sequence Alignment/Map) format represents the alignment of sequences (e.g. reads) to a reference sequence (e.g. genome)
  - Simple to read and parse (text, tab-delimited)
  - Flexible (possibility to add custom fields)
  - Compact in file size
  - Can store paired-end information
- Reference document: http://samtools.sourceforge.net/SAM1.pdf
- BAM is a binary (more compact) representation of SAM

SAM/BAM (cont.)

- Structure: two sections:
  - Header: lines starting with @, two letters, then several key:value pairs. The keys are again two letters. Contains information about the reference sequence (SQ), the libraries used ("read groups", RG), etc...
  - Sequences: one line for each read, with the following fields (among others)
    - Query (pair) name
    - Reference name
    - Position
    - Mapping quality
    - CIGAR string
    - Seq and quality
    - Tag:type:value fields

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

sff

- Binary format provided by 454
- Contains
  - A header with information on the run (name, key sequence, number of reads, etc.)
  - For each read:
    - Name, length of the read
    - Clipping information (quality and adaptor)
    - Numeric representation of the flowgrams (454 equivalent to chromatograms)
    - Base sequence called from flowgrams
    - Qualities

Why quality control is important

Some of the artefacts/problems we can detect with QC:

- Sequence quality
- Nucleotide composition
- Technial issues
- Overrepresented sequences
- Adaptor sequence presence
- ...

LOOK AT YOUR SEQUENCES
SPOT “STRANGE” BEHAVIOURS
TRY TO EXPLAIN THEM

Sequence quality

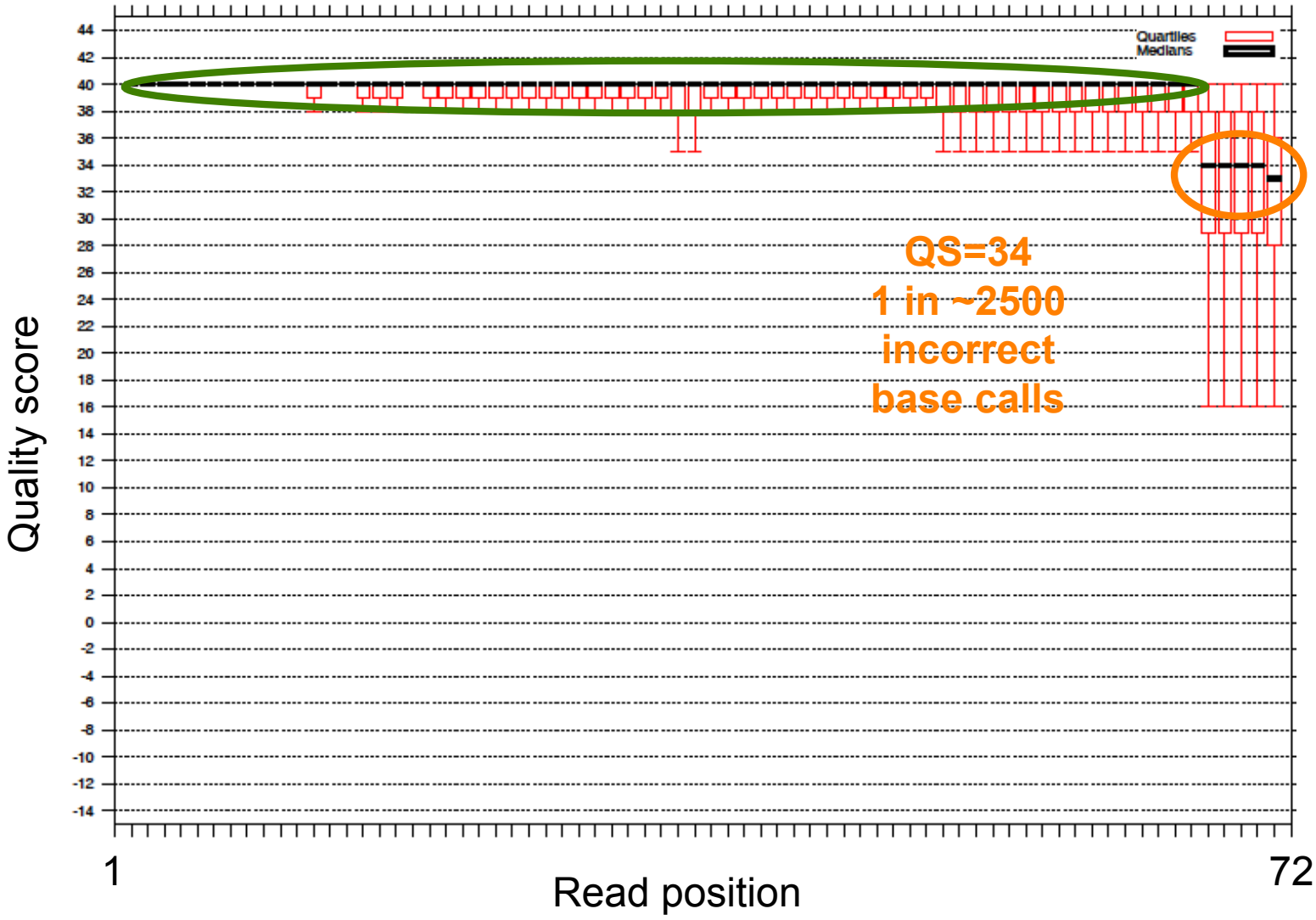
```
@ILLUMINA-GA_0000:1:1:2771:1022#0/1
TGACATNAAGCACTGTAGCTCATCTCGTATGCCGTCTT
+ILLUMINA-GA_0000:1:1:2771:1022#0/1
faaWa]B\]`^b`Vcdfd_f_cd_f[d_bfaSadddfb
@ILLUMINA-GA_0000:1:1:3203:1022#0/1
TGAGATNAAGCACTGTAGCTCTATCTCGTATGCCGTCT
+ILLUMINA-GA_0000:1:1:3203:1022#0/1
dcgga^BY_^\b]b`ggggffgegggdeggggggegggg
@ILLUMINA-GA_0000:1:1:4878:1023#0/1
TGAGGTNGTAGGTTGTATAGTATCTCGTATGCCGTCTT
+ILLUMINA-GA_0000:1:1:4878:1023#0/1
cdaed[BWa\Z]\\\ffffdfffdfffffdfffffffff
@ILLUMINA-GA_0000:1:1:5393:1022#0/1
TTCACTNATGAGAGCATTGTTCTGAGCATCTCGTATGC
+ILLUMINA-GA_0000:1:1:5393:1022#0/1
hhhhheBdeeffffchhhhhhhhhfgfhhhffffefff
@ILLUMINA-GA_0000:1:1:5523:1022#0/1
TGAGGTNGTAGGTTGTATAGTTATCTCGTATGCCGTCT
+ILLUMINA-GA_0000:1:1:5523:1022#0/1
ff]cf[B^X_bb^bbggggfgggg_ggfggcfcffaff
….
….
```

pos	mean	Q1	med	Q3
1	38.01	38	39	40
2	37.73	38	39	40
3	37.53	38	39	40
4	37.69	38	39	40
5	37.41	37	39	40
6	37.63	38	39	40
7	37.52	38	39	40
8	36.95	36	38	40
9	37.31	37	39	40
10	37.59	38	39	40
11	37.61	38	39	40
12	37.28	36	39	40
13	37.71	38	39	40
14	37.62	38	39	40
15	37.54	37	39	40
16	37.97	38	39	40
17	37.44	37	39	40
18	37.49	38	39	40
19	37.36	37	39	40

Sequence quality

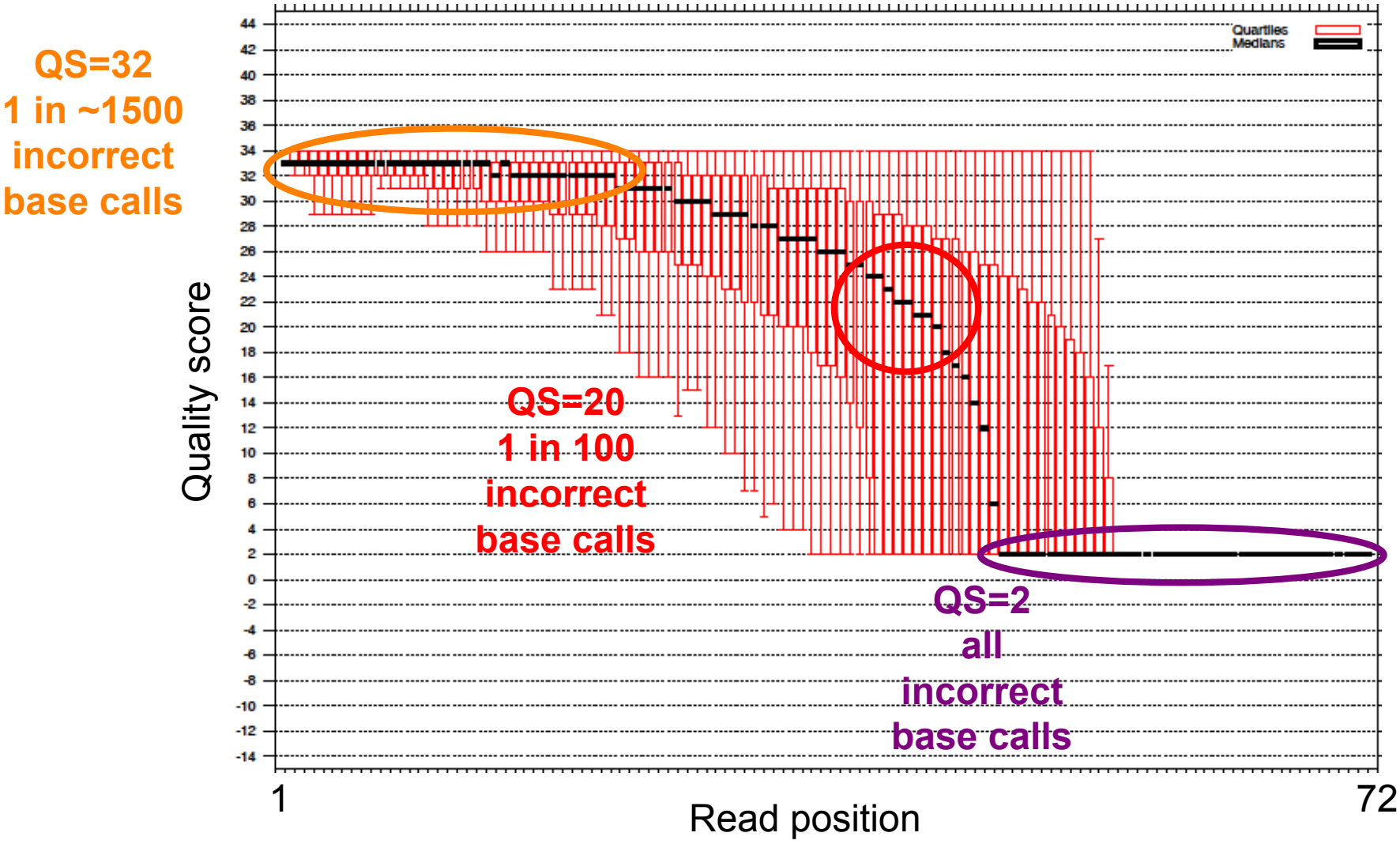
GOOD QUALITY RUN

QS=40
1 in 10000
incorrect
base calls



Sequence quality

BAD QUALITY RUN

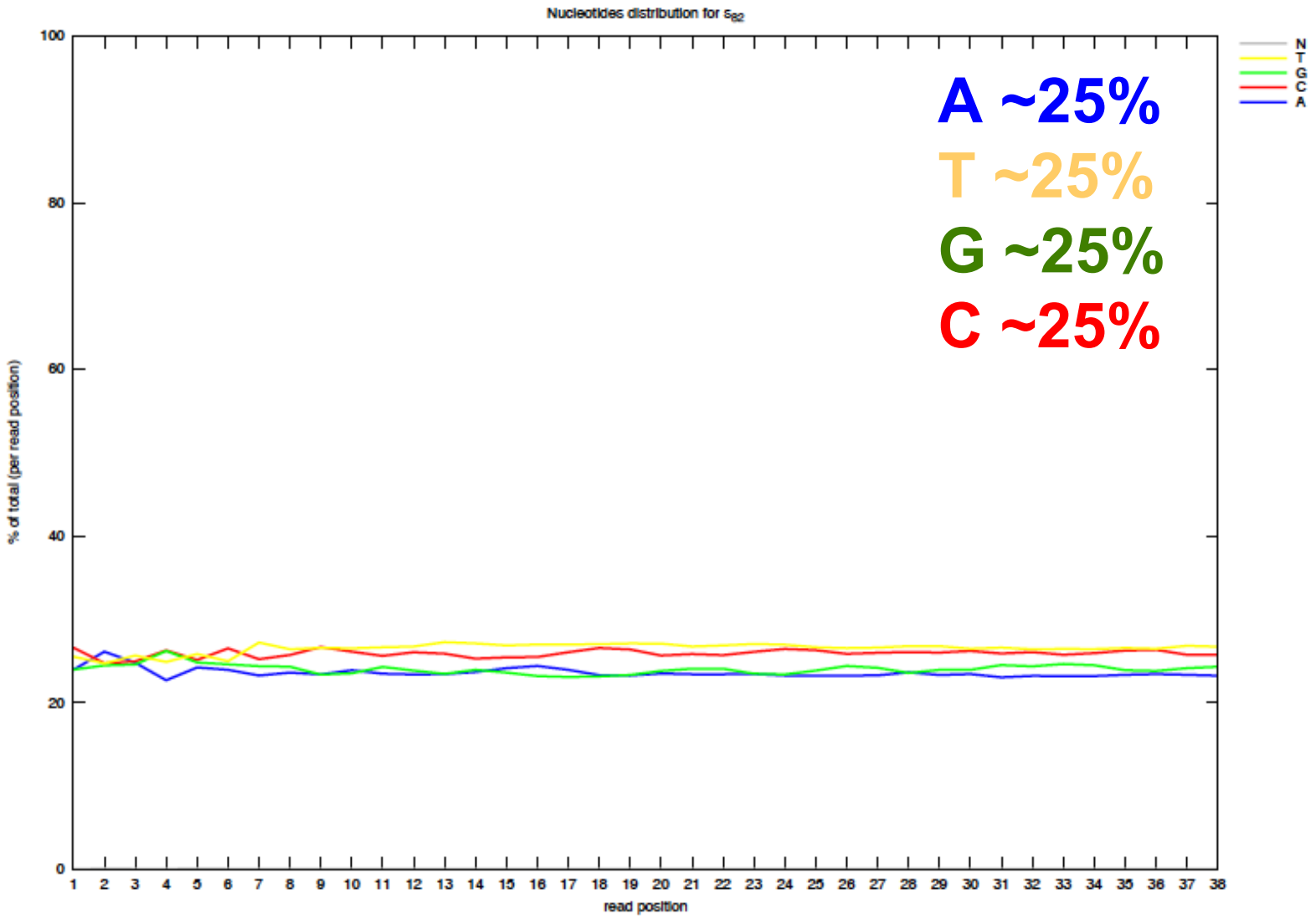


Nucleotide composition

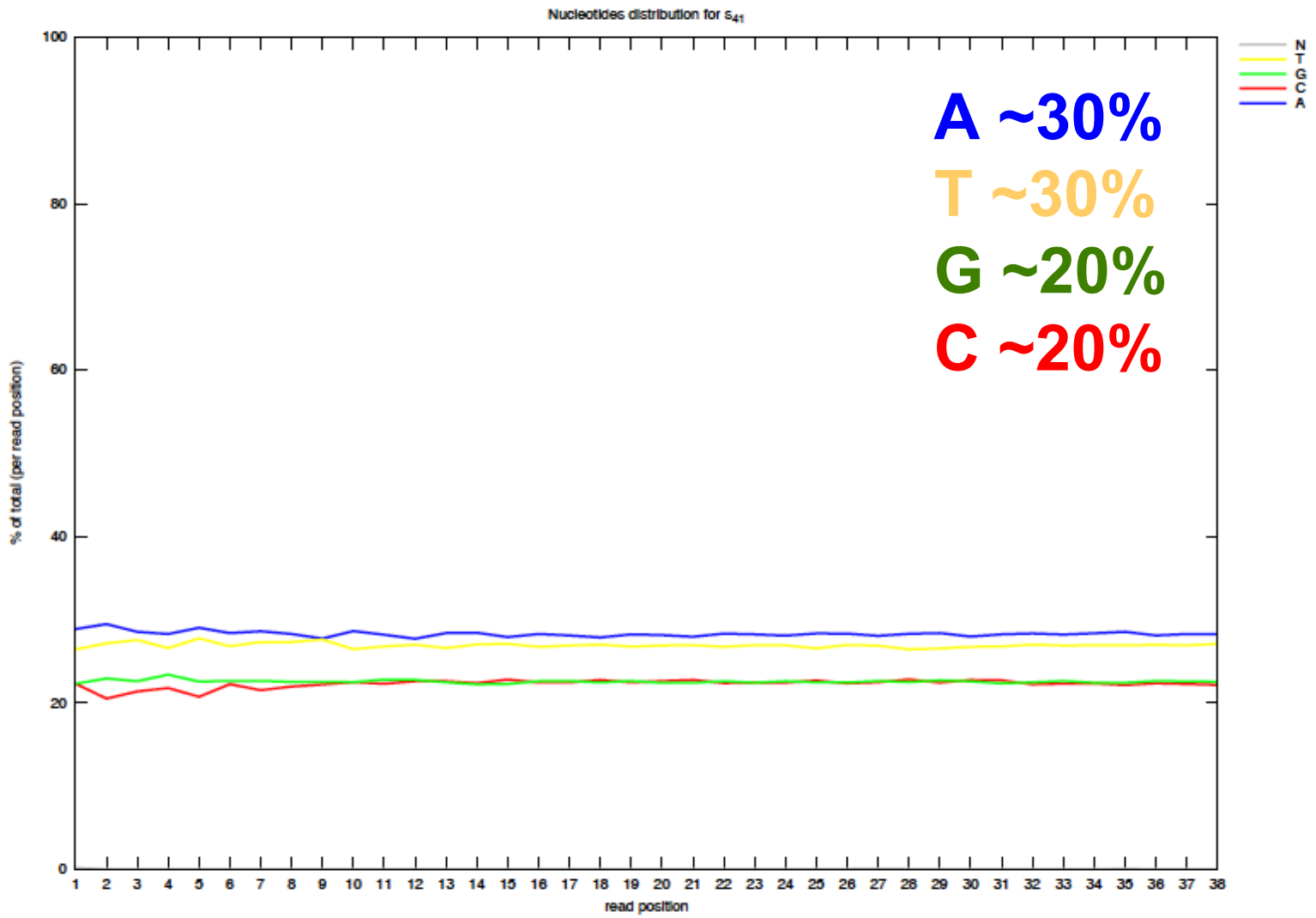
```
@ILLUMINA-GA_0000:1:1:2771:1022#0/1
TGACATNAAGCACTGTAGCTCATCTCGTATGCCGTCTT
+ILLUMINA-GA_0000:1:1:2771:1022#0/1
faaWa]B\]`^b`Vcdfd_f_cd_f[d_bfaSadddfb
@ILLUMINA-GA_0000:1:1:3203:1022#0/1
TGAGATNAAGCACTGTAGCTCTATCTCGTATGCCGTCT
+ILLUMINA-GA_0000:1:1:3203:1022#0/1
dcgga^BY_^\b]b`ggggffgegggdeggggggegggg
@ILLUMINA-GA_0000:1:1:4878:1023#0/1
TGAGGTNGTAGGTTGTATAGTATCTCGTATGCCGTCTT
+ILLUMINA-GA_0000:1:1:4878:1023#0/1
cdaed[BWa\Z]\\\ffffdfffdfffffdfffffffff
@ILLUMINA-GA_0000:1:1:5393:1022#0/1
TTCACTNATGAGAGCATTGTTCTGAGCATCTCGTATGC
+ILLUMINA-GA_0000:1:1:5393:1022#0/1
hhhhheBdeeffffchhhhhhhhhfgfhhhffffefff
@ILLUMINA-GA_0000:1:1:5523:1022#0/1
TGAGGTNGTAGGTTGTATAGTTATCTCGTATGCCGTCT
+ILLUMINA-GA_0000:1:1:5523:1022#0/1
ff]cf[B^X_bb^bbggggfgggg_ggfggcfcffaff
….
….
```

pos	A_Count	C_Count	G_Count	T_Count
1	4184726	3636289	2993640	14529850
2	2493259	4490289	13722137	4661065
3	12276591	6845747	3622752	2625158
4	3611989	4517290	12764502	4476465
5	11248562	3968447	6464472	3688770
6	3094389	3153655	6099499	13022698
7	4923585	3544477	11822757	5079405
8	11866464	1042283	6207172	6254332
9	8870719	3488704	2745084	10252623
10	5375998	2761606	12917981	4314650
11	3043455	11638364	6835895	3852527
12	12629424	5073041	4632904	3034882
13	2545268	10564820	6711226	5548937
14	3752988	2794955	3207436	15614698
15	4694143	4729795	13525064	2420856
16	3859216	3854697	3303337	14352850
17	12274317	2566690	4261912	6267332
18	3047662	6016803	10623984	5675723
19	4562389	9049534	3894678	7842744

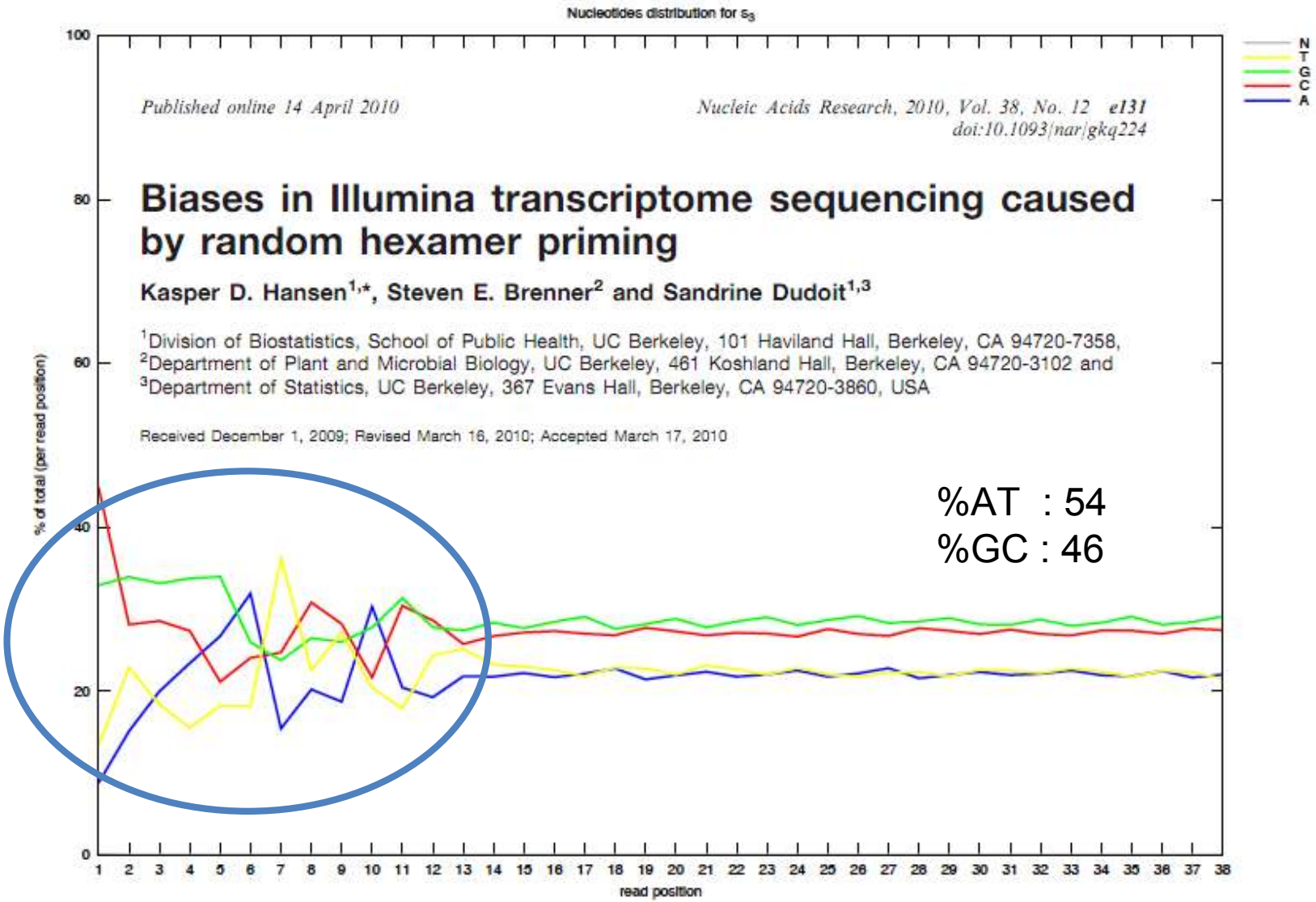
Nucleotide composition



Nucleotide composition



Technical issues



Published online 14 April 2010

Nucleic Acids Research, 2010, Vol. 38, No. 12 e131
doi:10.1093/nar/gkq224

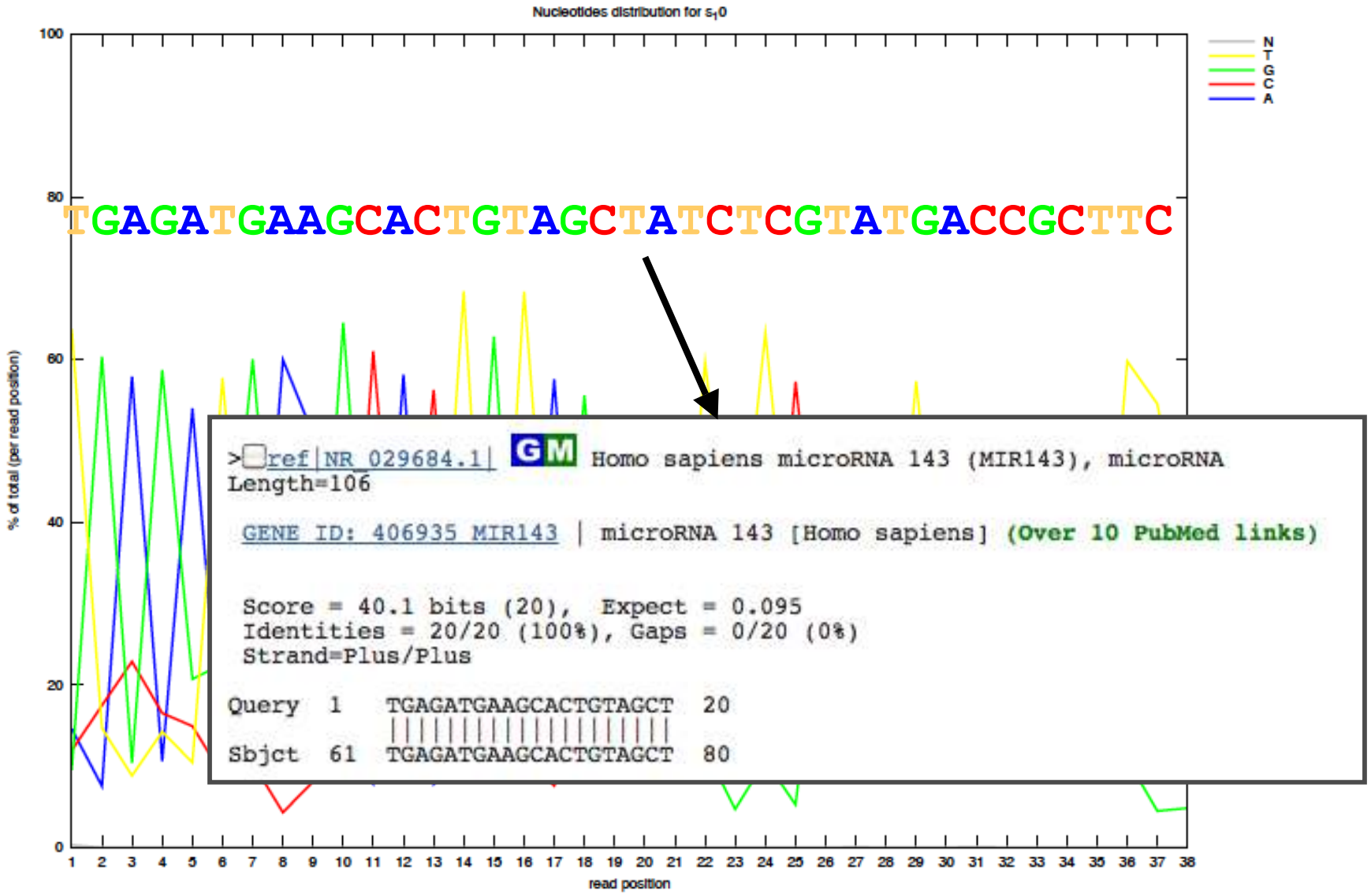
Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

[1]Division of Biostatistics, School of Public Health, UC Berkeley, 101 Haviland Hall, Berkeley, CA 94720-7358,
[2]Department of Plant and Microbial Biology, UC Berkeley, 461 Koshland Hall, Berkeley, CA 94720-3102 and
[3]Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA

Received December 1, 2009; Revised March 16, 2010; Accepted March 17, 2010

Overrepresented sequences



Exercise: preliminary quality control of raw sequences

http://evomics.org/learning/quality-assessment-and-control-of-sequence-data/