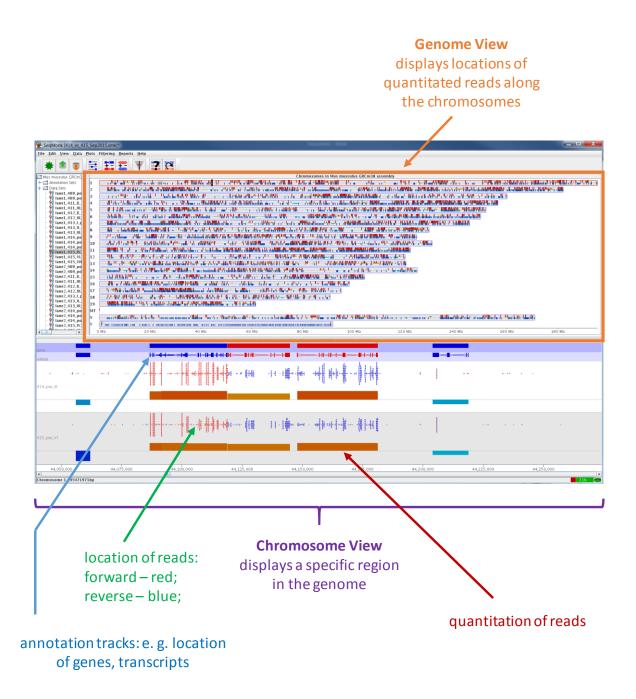
# **Exercises: Explore Before You Analyse**

For the following exercises you will be given a series of screenshots from Seqmonk which is a tool to visualise and analyse high-throughput mapped sequence data. Here are some quick explanations on what you can see:



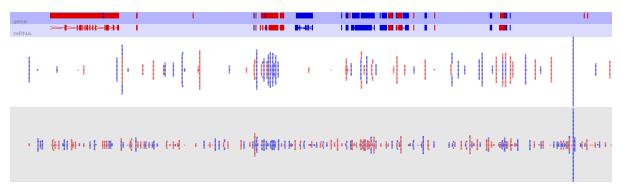
Here are some real life examples of RNA-Seq data. Look at each one of them (location, number and direction of reads) and answer the following questions:

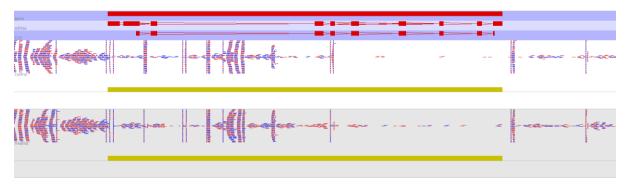
- 1) It there anything obviously wrong with the data?
- 2) Would the data be quantified by an analysis package? Would it give you a sensible answer?
- 3) If there is something wrong, what could you do to fix it?

#### Example 1

mRNA	1484 <b>1-</b> =14	•	н	1	1915-1911		H>	⊨₩ <b>■</b>	Hdd
	****	ŧ	H .	• 4	-4	111	· · · · · · · · · · · · · · · · · · ·	4 ( <b>1</b>	
• •	┝╋╬┟╴┟╽		-	-		1 i i i i i i i i i i i i i i i i i i i		p	• • • • • • • • • • • • • • • • • • •
c	H∰ ŀ				C	(₩	4 I	ŀ H	1
E		a de la composición de		1.1		•••••	· • • • • • • • • • • • • • • • • • • •	F H	1. A A A A A A A A A A A A A A A A A A A
			-	• • • •	· ·	+ + + + + + + + + + + + + + + + + + +	· · · · · · · ·	11.1	
Constant and the second		÷.			11	4	19 (19 (19 (19 (19 (19 (19 (19 (19 (19 (	1	• · · · · · · · · · · · · · · · · · · ·
		<b>†</b> -	×			<b>₩</b> ~ → -	1.00	les et al.	1 (1) (1) (1) (1) (1) (1) (1) (1) (1) (1
		4	× •		1.1	(H- · · ·		в 🔶 -	
	<b>***</b> *	÷ •			4	() 이 이 이 이 이 이 이 이 이 이 이 이 이 이 이 이 이 이 이	8 - 1 - 1 - <b>1</b> - 1	P 4	2

#### Example 2

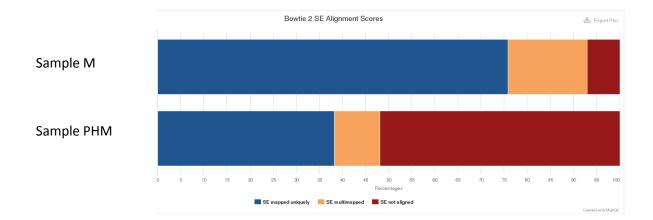




### Example 4

			Mus musculus GRCm38	8 chr1:93168340-93978412 (810 kbp)		_
ene IRNA						
	■>## E>=##E>##>###E># ########################	> <b>#######</b> F### <b>##</b> ~##		╾╋╪╋╋╤═╼╾┼╋┿╪╋ <b>╗╌</b> ╡		∎ ⊨≠⊶∎
a				·····	· ••• ••• ••••••••••••••••••••••••••••	e F
м -	an an an an an an ann ann ann an 1970 an an 1961 an an	# <b>#{(</b> ]# €• …		· · · · · · · · · · · · · · · · · · ·	and a second of a second difference	
93,200	93,300,000	93,400,000	93,500,000	93,600,000	93,700,000 93,800,000	93,900,000

### Additional information: Mapping efficiencies for the data shown in the screenshot above



			Homo sapiens GRCh38_ERCC (	hr1:140897908-162185118 (21	.2 Mbp)		
gene							
mRNA							
cos	II III <b>B</b> HA I <b>NA B</b> HA						
A1	• - K-1 - P	19 T B	11 {\$\$1	- +4  I	1 M At K + - M C - C -	10 - 10 - H - H	• • • • • • • • • • • • • • • • • • •
A10	· · · · ((	to Real a	1 4 × 4 + (44)	4 K (+ )	·	∰i ( ) ( )	
A11			1 (1) (1) (1) (1) (1) (1) (1) (1) (1) (1	1. S.	and the second second		1
A12	100 C 100 C 100 C	tere de la server	ः त्य के का भी की	Hile et al.	I - 4 H I + I AM - H I AM -		e e 14 📭 🚯 🕸 e
A2	e #	(i i i i i i i i i i i i i i i i i i i	ात्त्र न म सम्बद्ध में न सूच	<b>A</b> 101 - 1	· I weller in Clean-Cale	the state of the	(· · · · · · · · · · · · · · · · · · ·
A3	· · 0· ···)	{∝⊢ -# ( }	and a state of the 🗰		i i∰ -∳ ∥ €∮ -∲I∰ + ₹⊕ Ker-l	<b>€</b> 1 € 1	• • • • • • • • •
A4	1999 C 1999 (#	( · · · · · · · · · · · · · · · · · · ·	en (an en in huild) - fre	a di se	o ke Kion e kata i	1	in the part of the second
A5	1.5 (F) (F) (F) (F)	e e contra	the second state of the second	el-1≬∳		↓ 1 ↓ 1 ×	e en 🎼 🕴 🕸 de re
142,5	00,000 145,000,000	147,500,000	150,000,000	152,500,000	155,000,000	157,500,000	160,000,000

The next three examples show an RNA-Seq QC plot (Seqmonk) which provides summarised information on the location of reads. Where would you expect reads to be located? Are there any red flags?

The following metrics are given:

**Percent in Gene:** What percentage of reads fall into genes. A low value here would indicate potential DNA contamination

**Percent in Exons:** What percentage of reads in genes fall into exons. A low value here would indicate either DNA contamination, or possibly just a high proportion of unprocessed transcripts

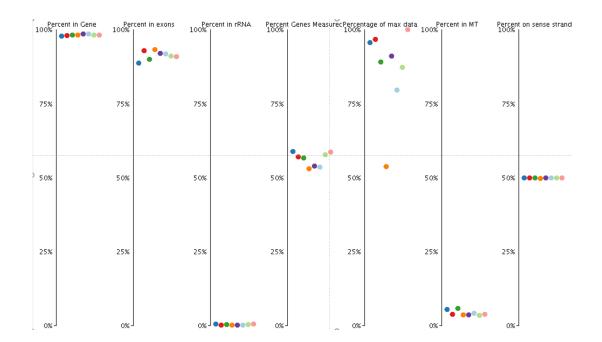
**Percent in rRNA:** What percentage of reads fall into rRNA. This will only account for mapped reads falling into annotated rRNA features, so will normally be much less than the actual amount of rRNA in the starting library

Percent Genes Measured: What percentage of genes have any exonic reads at all.

**Percentage of max data:** The highest number of reads in a data set is set to 100%. What percentage of reads do the other samples have.

Percent in MT: What percentage of reads fall into the Mitochondrion

**Percent on sense strand**: What percentage of reads fall onto the sense strand - depending on the type of library you've made this value could be expected to be near 0%, 100% or 50%



# Example 7

