High-throughput sequencing

Technologies and applications

Lecture outline

- Introduce recent sequencing technologies (focus on Illumina/Solexa sequencing)
- Discuss applications of high-throughput sequencing
- More detailed discussion of gene expression analysis using sequencing (RNA-seq) and comparison to arrays

Highthroughput sequencing technologies (partial list)

- 454 Life Sciences/Roche high-throughput pyrosequencing
- Applied Biosystems SOLiD sequencing
- Illumina/Solexa
- Single Molecule Real Time Sequencing (SMRT)
- Many others continuously being developed…

454 highthroughput pyrosequencing

- uses sequencing by synthesis
- growing sequences are attached to beads
- one type of nucleotide supplied at a time
- detect incorporation of the nucleotide by measuring emission of pyrophosphate when base is incorporated
- emission of pyrophosphate is quantitative (can be used to infer that multiple nucleotides have been incorporated)

Pyrogram:

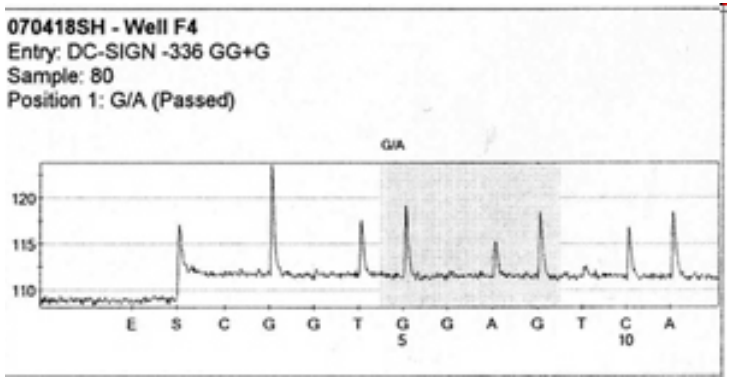


Image from Wikipedia

ABI Sequencing by Oligonucleotide Ligation and Detection (SOLiD)

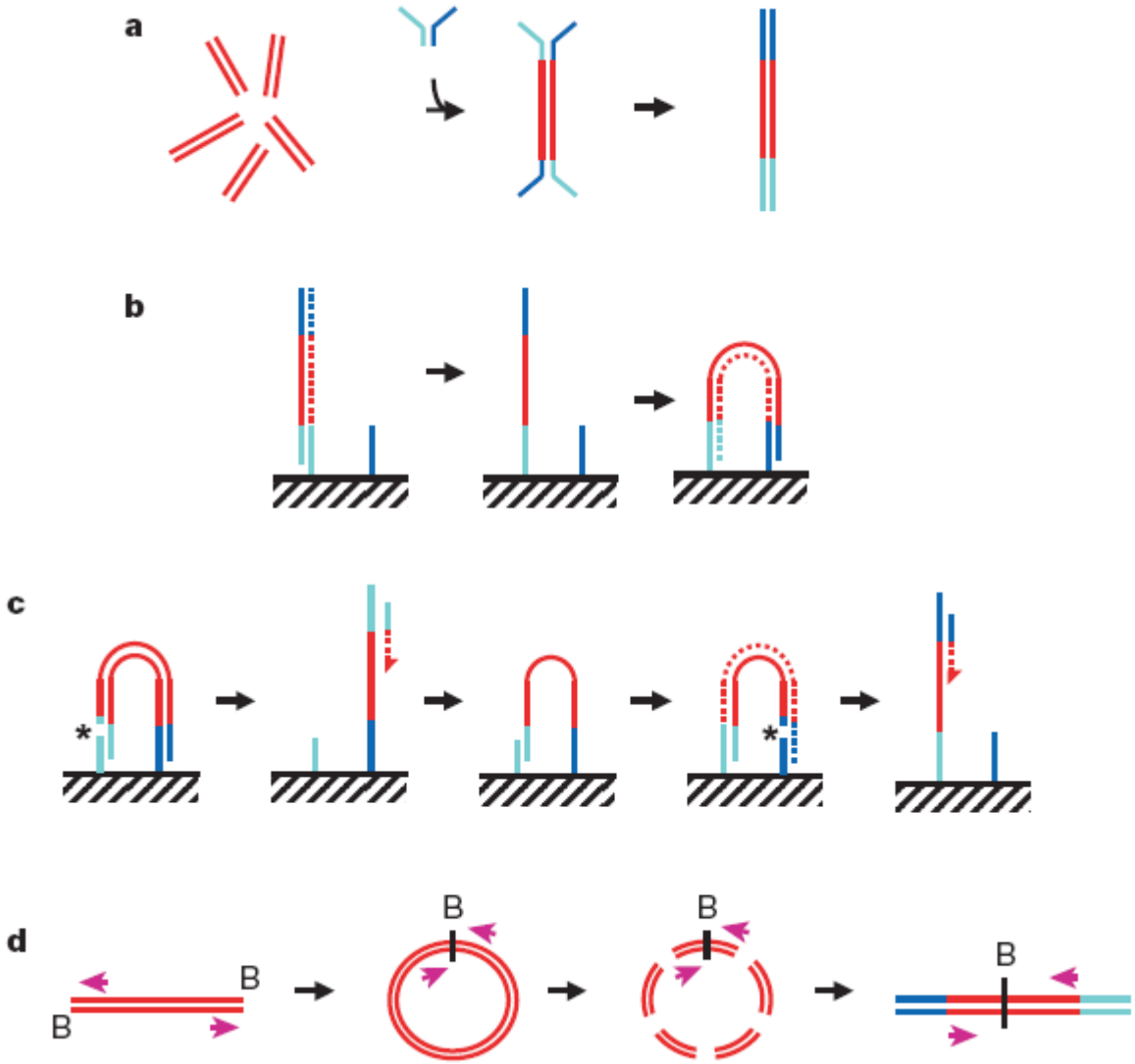
- Clonal populations of DNA fragments created using PCR (attached to magnetic beads)
- Beads attached to flow cell
- Expose beads to labelled oligonucleotides (each of the 16 dinucleotides represented in the first 2bp of the oligos with anything thereafter).
- Activity of ligase detected

See http://seqanswers.com/forums/showthread.php?t=10 for more

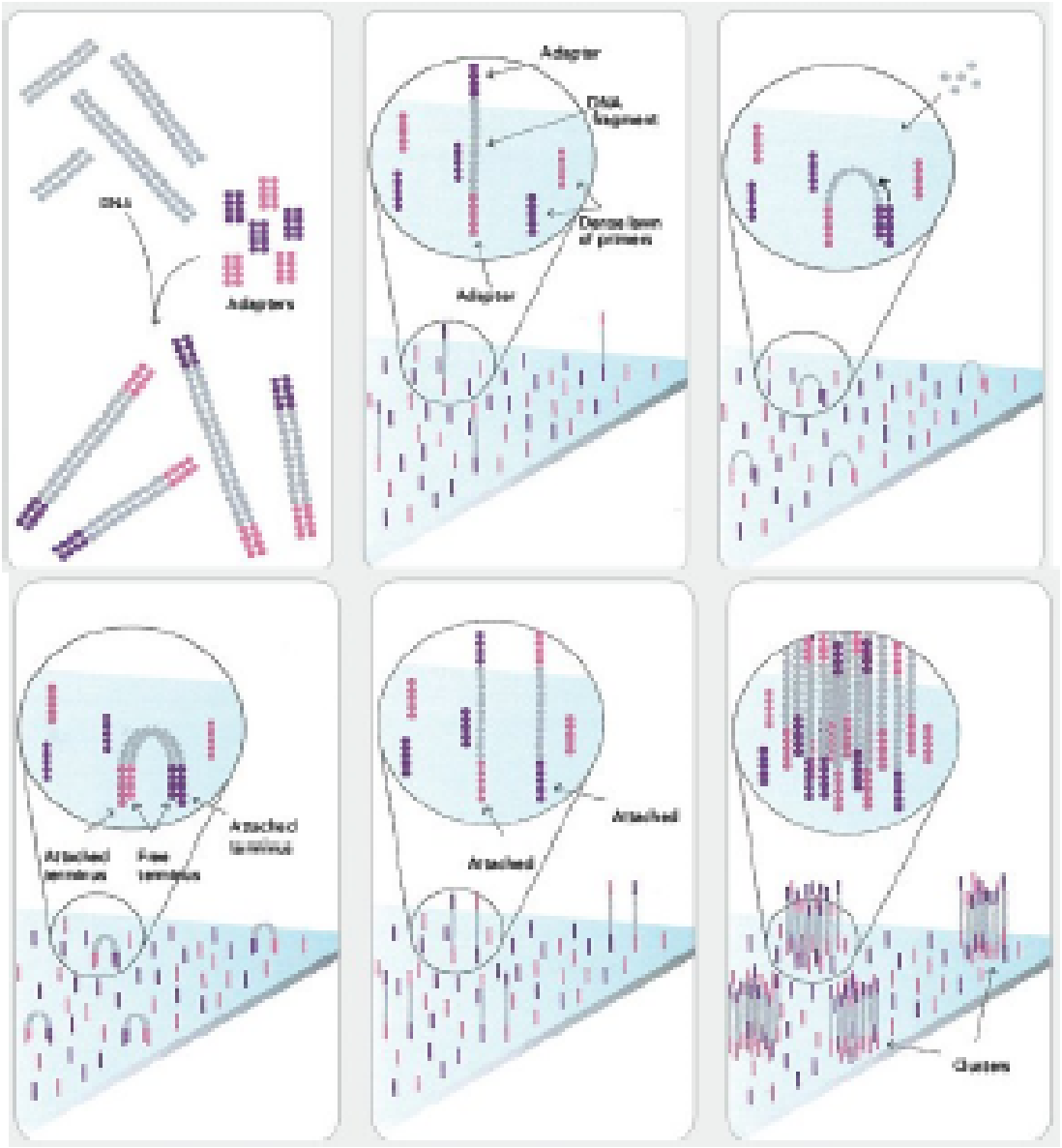
Illumina/Solexa sequencing steps

- Fragment sequences (required for most HT sequencing technologies)
  - RNA fragmentation: biased against sequence ends (5’ and 3’)
  - cDNA fragmentation: biased towards 3’ end
- ‘Bridge amplification’ forms local clusters of a given (single stranded) sequence fragment
- Sequencing by synthesis
  - second strand synthesized one base at a time
  - using flouresently labelled ‘reversible terminator’ bases
  - take an image of the ‘flow cell’
  - wash away the terminator and dye label before beginning next cycle

Illumina Sequencing (c, paired end; d, mate-pair



Bentley et al. 2008


DNA
Adapters
Adapter
DNA fragment
Adapter
Attached terminus
Free terminus
Attached
Attached
Clusters

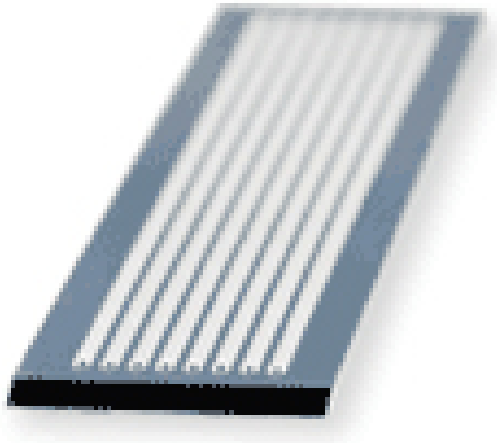
Illumina

Illumina/Solexa sequencing

The Illumina flow cell has 8 lanes

Multiple samples can be run on the same lane using ‘indexing’

(a short ‘barcode’ sequence is added to the fragments that enables the sample from which the fragment comes to be identified)



Highthroughput sequencing: comparison of technologies

- Read length
  - Illumina: ~ 100bp
  - 454: ~ 400bp
  - SOLiD: ~ 50bp

- Volume
  - Illumina: (~40 million reads per lane – Genome Analyzer; ~ 300 million – HiSeq)
  - 454: mid-range, on the order of hundreds of thousands to millions. More expensive per read than Illumina

- Cost (USD per MB): 454: 84 Illumina: 1.00 SOLiD: 5.81

Highthroughput sequencing: strand information

- Mostly not strand-specific (i.e. ambiguity around which strand the sequence read comes from)

  Strand-specific sequencing protocols are available

  Strand information can usually be inferred for spliced cDNA reads

Current list of sequencing technologies (Wikipedia; 2012)

- Lynx Therapeutics' Massively Parallel Signature Sequencing (MPSS)
- Polony sequencing
- 454 pyrosequencing
- Illumina (Solexa) sequencing
- SOLiD sequencing
- Ion semiconductor sequencing
- DNA nanoball sequencing
- Helioscope(TM) single molecule sequencing
- Single Molecule SMRT(TM) sequencing
- Single Molecule real time (RNAP) sequencing
- Nanopore DNA sequencing
- VisiGen Biotechnologies approach

The fastq format

- similar to fasta, with the addition of sequence quality scores

e.g.:

```
@PSI179204_0013:6:1:1062:14428#0/1
CACATACTCTATNAGAGAAATACCTGGTGTTATATGCTGACTTGGGAACC
+PSI179204_0013:6:1:1062:14428#0/1
dddc`dd\d`b^Bb^`Y^^a^ddda\\Q\\^b^b^\b\\_HQNRYG]^[`
```

Explanation

Line 1: @ followed by sequence title (title provides info about flow cell lane, position of the cluster, indexing and pairing)

Line 2: Sequence

Line 3: + followed by sequence title

Line 4: Quality scores (see next slide)

The fastq format

- quality scores (also called Phred scores)

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

(where p = probability of mis-called base)

In fastq format the per-base quality scores are encoded in ‘ascii’ characters.

NB: there are different conventions for calculating and encoding these quality scores. We will consider only the convention in current use by Illumina (e.g. the Sanger format is different)

For current Illumina fastq format: quality scores from 0 to 62 are encoded by (printable) ascii characters 64 to 126

0	NUL	16	DLE	32	SP	48	0	64	@	80	P	96	`	112	p
1	SOH	17	DC1	33	!	49	1	65	A	81	Q	97	a	113	q
2	STX	18	DC2	34	"	50	2	66	B	82	R	98	b	114	r
3	ETX	19	DC3	35	#	51	3	67	C	83	S	99	c	115	s
4	EOT	20	DC4	36	$	52	4	68	D	84	T	100	d	116	t
5	ENQ	21	NAK	37	%	53	5	69	E	85	U	101	e	117	u
6	ACK	22	SYN	38	&	54	6	70	F	86	V	102	f	118	v
7	BEL	23	ETB	39	'	55	7	71	G	87	W	103	g	119	w
8	BS	24	CAN	40	(	56	8	72	H	88	X	104	h	120	x
9	HT	25	EM	41	)	57	9	73	I	89	Y	105	i	121	y
10	LF	26	SUB	42	*	58	:	74	J	90	Z	106	j	122	z
11	VT	27	ESC	43	+	59	;	75	K	91	[	107	k	123	{
12	FF	28	FS	44	,	60	<	76	L	92	\	108	l	124	|
13	CR	29	GS	45	-	61	=	77	M	93	]	109	m	125	}
14	SO	30	RS	46	.	62	>	78	N	94	^	110	n	126	~
15	SI	31	US	47	/	63	?	79	O	95	_	111	o	127	DEL

Decimal ASCII Chart

Common applications of high-throughput sequencing (*-seq)

Gene expression (RNA-seq)

Protein binding to DNA (ChIP-seq)

Protein binding to RNA (Clip-seq)

DNA methylation (Methyl-seq; HELP-tagging)

Genome sequencing or re-sequencing (including targeted sequencing – e.g. T cell receptor sequencing, exome sequencing)

SNP discovery

RNA-seq

- Library preparation
    - RNA selection (polyA selection or not?)
    - (fragmentation of RNA by hydrolysis, or not?)
    - Reverse transcription
    - Fragmentation of cDNA (if RNA not already fragmented)

- High-throughput sequencing

- Quality control and analysis

RNA-seq: Key advantages over arrays

- Transcripts not specified *a priori*

- Greater ‘dynamic range’

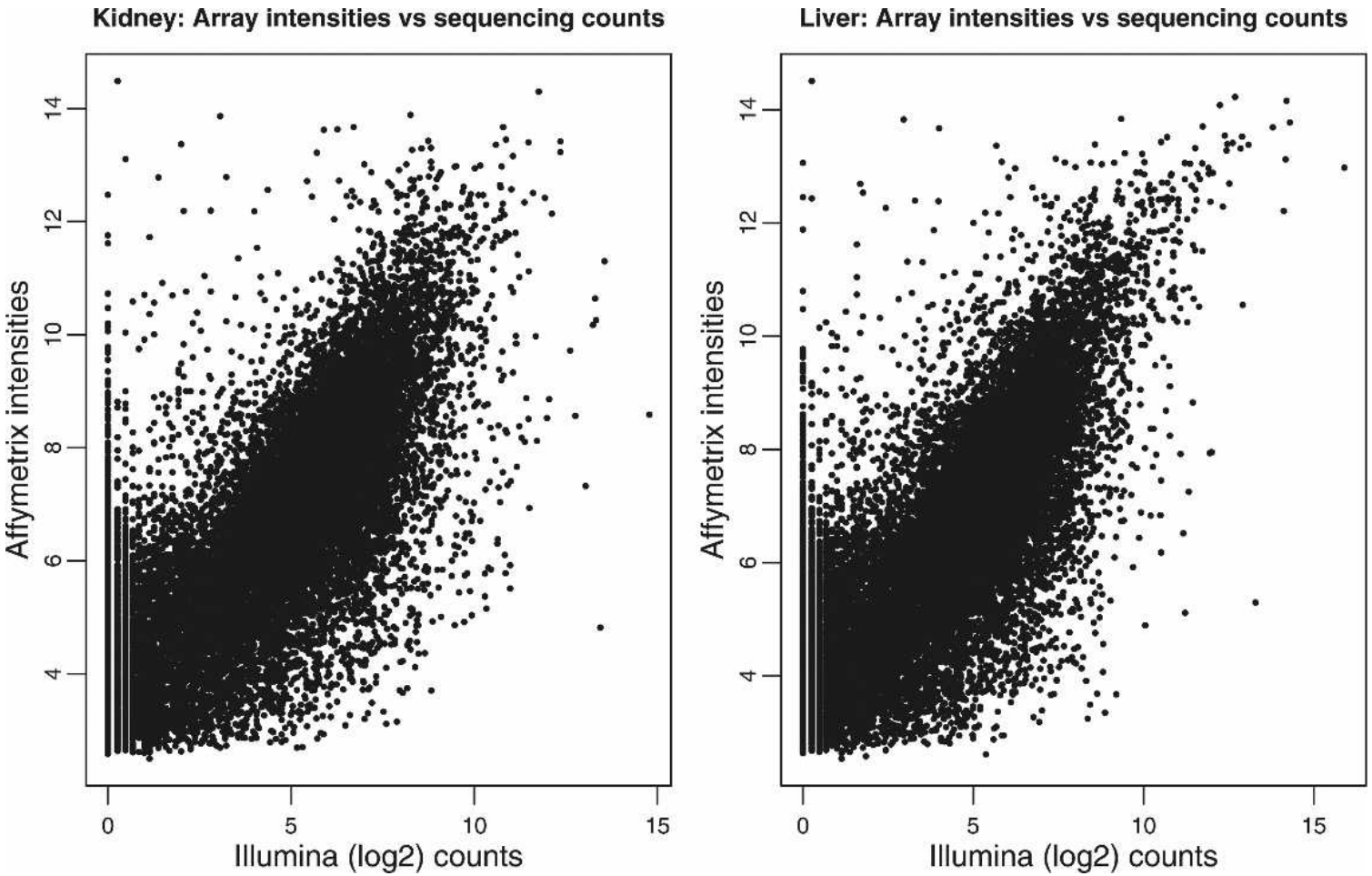
- More reproducible

- Possible to compare across different labs and experiments (poses normalization challenges with arrays)

- Results are easily interpretable

- Abundances of whole mRNA isoforms can be inferred (very difficult with arrays)

RNA-seq versus arrays

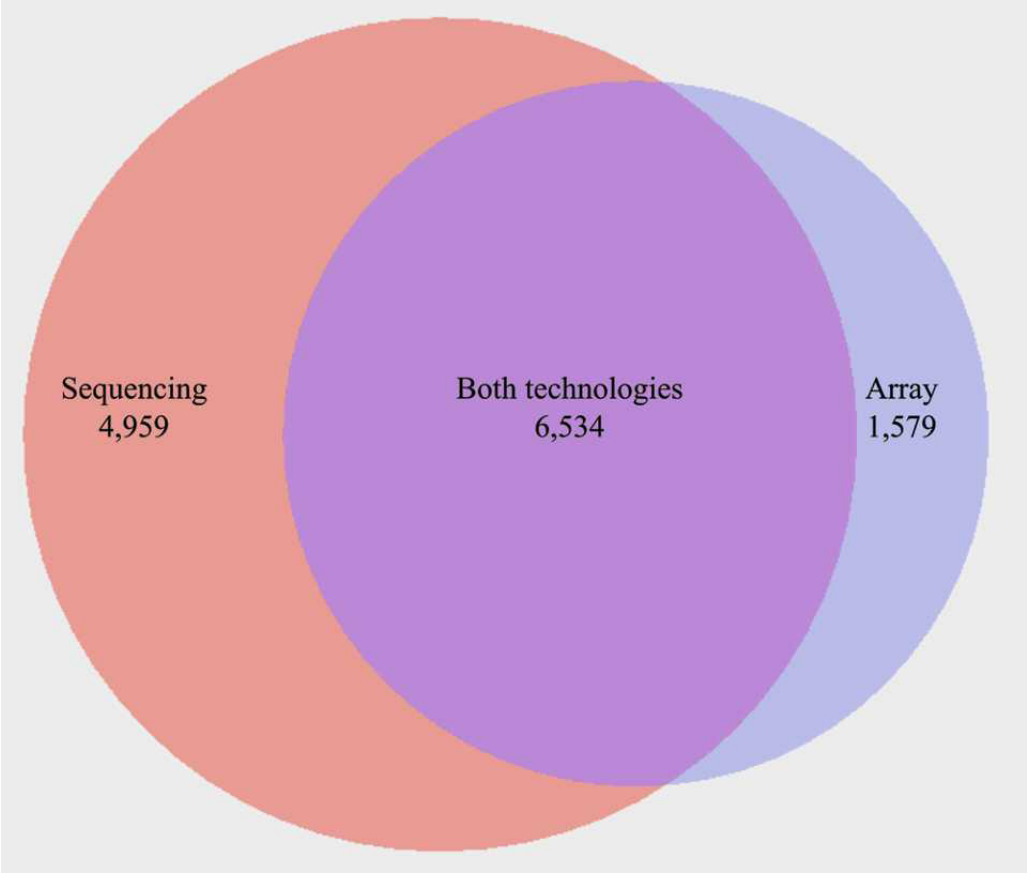


e.g Marioni et al. found better agreement between RNA-seq and qrtPCR compared to between array and qrtPCR

Fig from Marioni et al. Genome Research

RNA-seq versus arrays: Differential Expression

Marioni *et al.* also found more DE genes using RNA-seq than arrays



Marioni et al. Genome Research

RNA-seq: data processing and analysis steps

1. Quality control
2. Mapping or assembly
    - Reads mapped to reference genome (e.g. TopHat)
    - Deal with multiply mapping reads (e.g. exclude or weight?)
    - Alternatively reads can be assembled into transcripts (*c.f.* genome assembly)
3. Mapping quality control
    - Determine mapping duplication rate (potential PCR bias)
    - Exclude duplicate reads?
4. Transcript discovery/assembly/identification
5. Estimate expression levels

RNA-seq

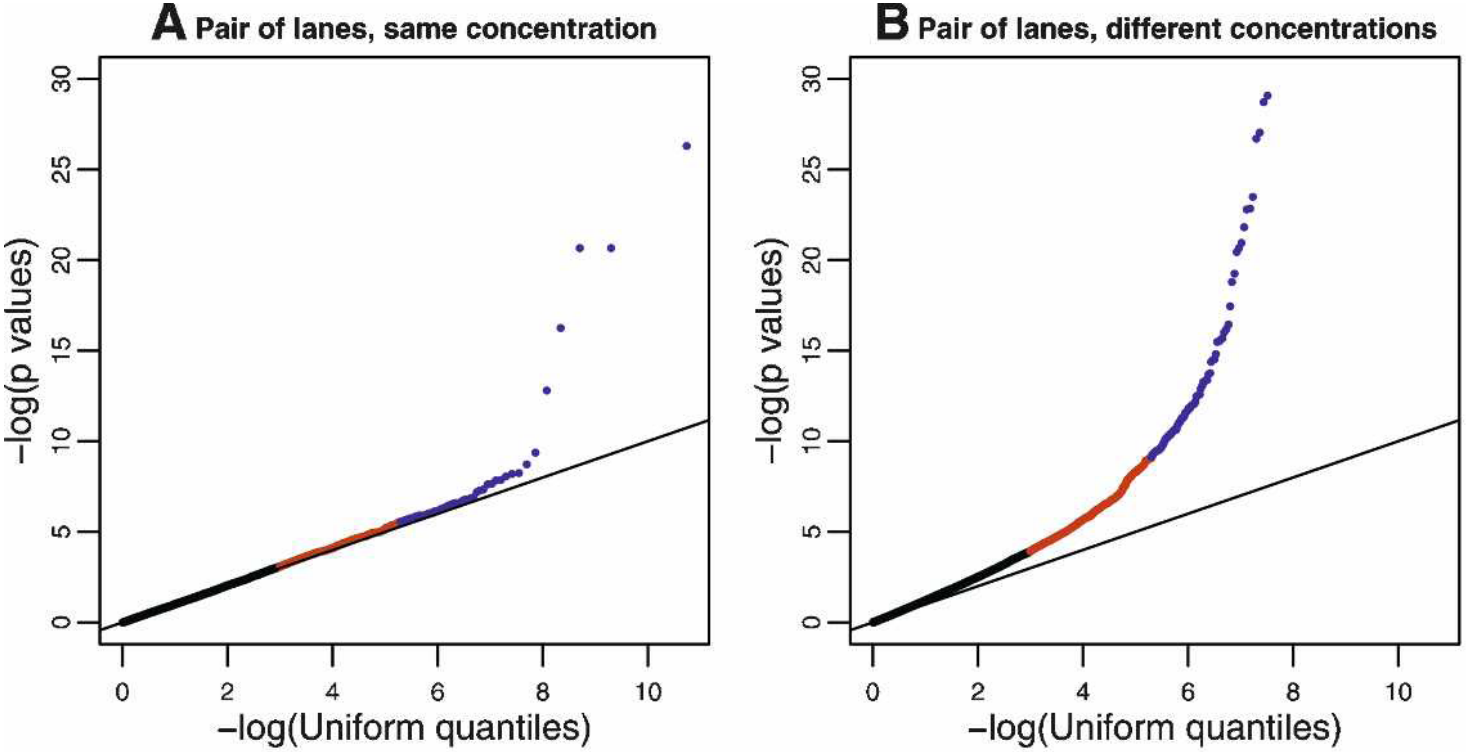
Measures of gene expression with RNA-seq

RPKM: Reads per kilobase (of transcript) per million mapped reads

FPKM: Fragments per kilobase (of transcript) per million mapped reads

(difference between RPKM and FPKM has to do with the fact that more than one read can be generated from the same fragment – e.g. in the case of paired-end sequencing).

Lane effects



Possible resolution: divide samples across lanes (using multiplexing)

Fig from Marioni et al. Genome Research