

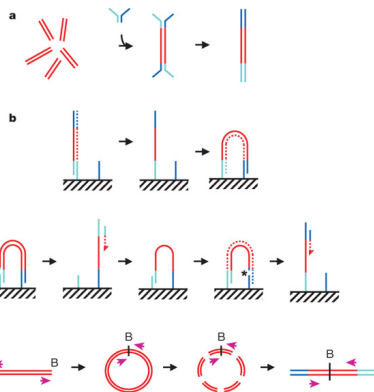
Sequence Analysis Work Flows

Martin T. Morgan¹

27-28 February 2014

Sequencing: Work flows

1. Experimental design
2. 'Wet lab' sample prep
3. Sequencing
 - ▶ 100's of millions of reads
 - ▶ 30-150 nucleotides
 - ▶ Single and paired-end
 - ▶ Bar codes, lanes & flow cells
4. Alignment
5. Analysis: DNA, RNA, epigenetics, integrative, microbiome, . . .



Bentley et al., 2008, Nature 456:
53-9

Experimental design and wet lab

- ▶ RNA-seq
 - ▶ Known gene / transcript differential expression
 - ▶ Novel transcript discovery
 - ▶ Single- versus paired-end
- ▶ ChIP-seq
- ▶ Variants
 - ▶ Germline vs. somatic
 - ▶ Exome vs. whole genome
 - ▶ SNP vs. indel vs. structural
- ▶ Copy number
 - ▶ Low vs. high coverage

RNA-seq: single versus paired end

- ▶ Most analysis now paired-end
- ▶ Reads within exons
- ▶ A single end spanning exons: 'junction reads'
- ▶ Reads spanning exons

@ERR127302.1703 HWI-EAS350_0441:1:1:1460:19184#0/1
CCTGAGTGAAGCTGATCTTGATCTACGAAGAGAGATAGATCTTGATCGTCGAGGAGATGCTGACCTTGACCT
+
HHGHHGHHHHHHHDGG<GDGGE@GDGGD<?B8??ADAD<BE@EE8EGDGA3CB85* ,77@>>CE?=896=:
@ERR127302.1704 HWI-EAS350_0441:1:1:1460:16861#0/1
GCGGTATGCTGGAAGGTGCTCGAATGGAGAGCGCCAGCGCCCCGGCGCTGAGCCGCAGCCTCAGGTCCGCCC
+
DE?DD>ED4>EEE>DE8EEEDE8B?EB<@3;BA79? ,881B?@73;1?#####
@ERR127302.1705 HWI-EAS350_0441:1:1:1460:13054#0/1
AAAACACCCTGCAATCTTTCAGACAGGATGTTGACAATGCGTCTCTGGCACGTCTTGACCTTGAACGCAAAG
+
EEDEE>AD>BBGGB8E8EEEGBGGGGGBGGGGG3G>E3*?BE??BBC8GB8?? :??GGDGDDD>D>B<GDDC8
@ERR127302.1706 HWI-EAS350_0441:1:1:1460:14924#0/1
CACCCAGTGGGGTGGAGTCGGAGCCACTGGTCTCTGCTGCTGGCTGCCTCTCTGCTCCACCTTGTGACCCAGG
+
HHHHHGEEGEEADDGDBG>GGD8EG ,<6<?AGGADFEHHC>D@<@G@>AB@B?8AA>CE@D8@B=?CC>AG
@ERR127302.1707 HWI-EAS350_0441:1:1:1461:6983#0/1
CGACGCTGACACCGGAACGGCAGCAGCAGCAGGACGATTAAGACAAGGAGGATGGCTCCACAGACGCTCATG
+
GEEGEGE@GGGGGGEGGGGGBB>G3?33?8* ; ;79?<9@?DD8@DDEE888 ; -BB? .A#####
@ERR127302.1708 HWI-EAS350_0441:1:1:1461:10827#0/1
AAAGAAGTCTTGAATAGACTGCCTCTGCTTGAGAACTTATGATGTAATTATTGCATGCTGCTAATATAC
+
GGGGGDDEBFGGGGGBE ,DAGDDGGGEEEG<EEFDECFFEEDE@<>ACEBEFDEEFE<EDC@E<EECCBEB
@ERR127302.1709 HWI-EAS350_0441:1:1:1461:7837#0/1
CAGCCACAGAACCACGGCACGGAAGACATGAGGCAGCATGCTCACGAGAGAGGTGAGGGTCTCCCCTCCAGG
+
HHGHHHH>DH: @.7@49 ;88G8>G>DDG@D>D@G@GE>@DDBDDG<A82?#####

FASTQ files

```
@ERR127302.1709 HWI-EAS350\_0441:1:1:1461:7837\#0/1
```

▶ (Anntoation), machine, flow cell, tile, coordinates, end
CAGCCACAGAACCACGGCACGGAAGACATGAGGCAGCATGCTCACGAGA...

▶ DNA sequence, usually A, C, T, G, N (uncalled)

```
HHGHHHH>DH:@.7@49;88G8>G>DDG@D>D@G@GE>@DDBDDG<A82...
```

▶ Quality (approximately, $-\log_{10}(p)$) encoded as ASCII characters: 40 is approximately $p = 0.0001$, 30 is $p = 0.001$, etc.

▶ Different encodings – [wikipedia](#)

▶ Higher in the alphabet is better

##	!	"	#	\$	%	&	'	()	*	+	,	-	.	/	0	1
##	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
##	2	3	4	5	6	7	8	9	:	;	<	=	>	?	@	A	B
##	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
##	C	D	E	F	G	H	I										
##	34	35	36	37	38	39	40										

Quality assessment

Assessment, e.g., `fastqc`

- ▶ Read length, duplication
- ▶ Nucleotide use: per cycle, GC content, N content, ...
- ▶ Quality: per cycle, per read
- ▶ Consistency across samples, without obvious treatment-specific associations

Remediation, e.g., `trimmomatic`

- ▶ Crop e.g., leading / trailing artifacts of sequencing protocol
- ▶ Trim based on quality

Alignment

Bowtie / Tophat / Cufflinks

- ▶ Bowtie2 – alignment
- ▶ Tophat – splice junction mapper
- ▶ Cufflinks / *cummeRbund* – isoform assembly & quantification

Other aligners

- ▶ SNAP – fast and accurate
- ▶ *subread* / *Rsubread* – memory efficient
- ▶ GSNAP / GMAP / *gmapR* – flexible and high quality alignments

BAM files

- ▶ Visualization with, e.g., [IGV](#)
- ▶ SAM / BAM (and other) [specifications](#)

```
ERR127302.25553011 403 chr14 19413639 1 72M
=          19413589          -122
CAAAGAATTGATTGAATTCATCAGGGCTAAAATCTCCAAAAATATACTGCGG...
!#&&"%&$&%&%&&%&"%***&'(')'')')%('#++++++*)'&%+***++...
AS:i:-2 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1
MD:Z:7C64 YT:Z:UU NH:i:3 CC:Z:= CP:i:20145991
HI:i:0
```

Field	Name	Value
1	QNAME	Query (read) NAME
2	FLAG	Bitwise FLAG, e.g., strand of alignment
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string
7	MRNM	Mate Reference sequence NaMe
8	MPOS	1-based Mate POSition
9	ISIZE	Inferred insert SIZE
10	SEQ	Query SEQUENCE on the reference strand
11	QUAL	Query QUALity
12+	OPT	OPTional fields, format TAG:VTYPE:VALUE

Analysis

Bioconductor

- ▶ RNA-seq differential expression (e.g., *DESeq2*, *edgeR*, *DEXSeq*)
- ▶ ChIP-seq designed experiments (e.g., *DiffBind*)
- ▶ Calling (e.g., *VariantTools*, *h5vc*, *Rorient*) and working with (*VariantAnnotation*) variants
- ▶ Copy number (e.g., *cn.mops*, *ExomeSeq*)
- ▶ Visualization (e.g., *Gviz*)

Other common software

- ▶ ChIP-seq peak calling
- ▶ Calling variants, e.g., *GATK*
- ▶ Novel transcript discovery
- ▶ Genome browsers, e.g., *IGV*