Introduction to bioinformatics (NGS data analysis)

Alexander Jueterbock

June 2016

Got your sequencing data - now, what to do with it?

- File size: several Gb
- Number of lines: >1,000,000

```
@M02443:17:000000000-ABPBW:1:1101:12675:1533 1:N:0:1
TCGATAATTCTTACTTTCTCTCTGGTCTGAGCGTTTCACATCAACGACAAGCTCGA
TTCTTCCTTTTCTCTTTTTTTCTTCTCTTCCTCTTTTTTCCTTTTCTCCCTCTTCT
TTTTTTTTTCTTCTT
+
8B6-@-,CFFED9CFAE@@C6;@,CFEEF9<@6FGGF9F<CC,,CB,@::8CF,6+
,,3733>>@@,,,388@,,8*,773333,3,333738,*,,,,,,76,,2,,2,,2
0*).1.))(0*)***
@M02443:17:000000000-ABPBW:1:1101:18658:1535 1:N:0:1
TCCCTAATTCTCTGTCTTCAAATTTTCCTTCTCTAAATCGTCCCTCGTTTCTACCT
TTTCTTGTTTTTTTATTTCCTCCTCTTCCTTTTTTACTTCCACCTTCTTTTCTGCC
TTTTCTTCTTTTTCT
+
-<<9-@CCEF9CE-<,,,,,;,,<C,=,6,C9,C<=C,,,;,86C,6:C,,,;<;,,
,,,,5,5:,,9++4,,,:,,,,,,,,,,,38,853,5,,3,,7,,,6,,,,,7,,,,
+0,()+++)11.*)*
```

Before library preparation

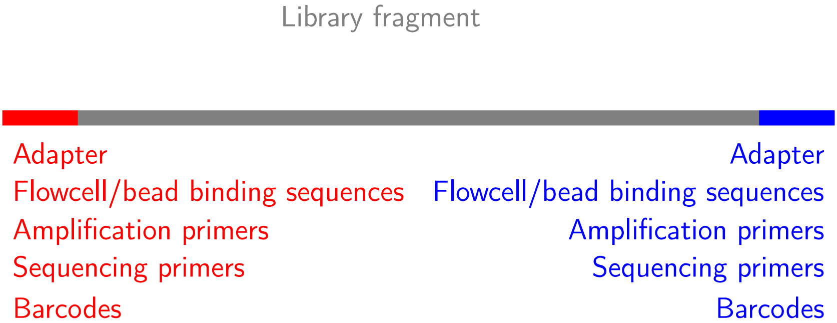
What you need to know to steer your way through the analysis

- Research question
    - Identify adaptive genes
    - *De novo* genome assembly
    - Population genetic structure
    - Phylogenetic relation
- Experimental design
    - Number of individuals
    - Treatment of samples (e.g. heat stress)
- Sample collection
    - Samples degraded (e.g. stored in Formalin)
    - Tissue (reproductive, vegetative)

Library preparation

- DNA-seq, RNA-seq, Bis-Seq, Chip-Seq. . .
  - RNA reads (which lack introns) require splice-aware mappers.
  - Bis-seq changes GC ratio (bisulphite converts cytosine to uracil, but leaves 5-methylcytosine unaffected)
  - Chip-Seq enriches binding-sites of DNA-associated proteins
- Pooled samples?
  - Demultiplexing
  - Remove barcodes
- Adapter sequences that have to be trimmed off?
- Targeted coverage

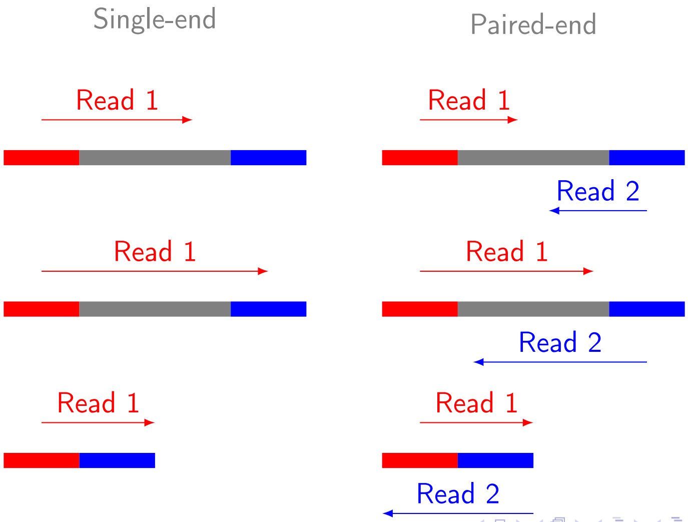
Single- or Paired end sequencing, read length



Library fragment

Adapter	Adapter
Flowcell/bead binding sequences	Flowcell/bead binding sequences
Amplification primers	Amplification primers
Sequencing primers	Sequencing primers
Barcodes	Barcodes

Single- or paired-end sequencing, read length - why does it matter



Expected read lengths and sequencing qualities for common sequencing platforms

Platform	Max. length	Reads/run	Consideration
Illumina Miseq	2x300	25 million	
Illumina NextSeq	2x150	400 million	
Illumina HiSeq	2x150	5 billion	
Roche 454	700	0.7 million	High error rate
Ion PGM 318 chip	200-400	4-5.5 million	
PacBio RSII	14,000	0.47 million	High error rate
SoliD 5500xl W	2x100	266 million	Low error rate Color-space

Primary analysis

- Demultiplexing
- Adapter trimming
- Quality control

Demultiplexing of pooled samples (if barcoded inline)

AATTANNNNNNNNNNNNNNN	File 1
AGTCGNNNNNNNNNNNNNNN	File 2
AGTCGNNNNNNNNNNNNNNN	File 2
GCCATNNNNNNNNNNNNNNN	File 3
AATTANNNNNNNNNNNNNNN	File 1
GCCATNNNNNNNNNNNNNNN	File 3
AGTCGNNNNNNNNNNNNNNN	File 2

Trimmig: Adapter removal

Mostly 3'adapters disturb assembly and alignment

GATTTGGGGTTCAANNNNNNNATTAGTATCGAT

GATTTGGGGTTCAANNNNNNNATTAGTATCGAT

TTGGGGTTCAANNNNNNNATTAGTATCGAT

GATTTGGGGTTCAANNNNNNNATTAGTATCGAT

ATTTGGGGTTCAANNNNNNNATTAGTATCGAT

GATTTGGGGTTCAANNNNNNNATTAGTATCGAT

Fastq file - 4 lines for each read

```
@HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
GATTTGGGGTTCAAATTAGTATCGATCAAATAGTAAATCCATTTGTTCAACTC
+
!’’*((((***+))%%%++)(%%%%).1***-+*’’))**55CCF>>>>>>CC
```

1. sequence id (specifications can differ slightly between sequencing platforms)
   - =@=instrument name : flowcell lane : tile number: flowcell x coordinate : flowcell y coordinates : #barcode sequence: pair number for paired-end sequencing
2. sequence
3. + optionally followed by sequence identifier again
4. quality scores

Trimmig of low-quality bases

- Trim bases with a Phred quality score <20
- $Quality = -10 * log_{10} P$

Phred Score	Probability of incorrect base	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

Fastq file contains both sequence reads and base quality scores

Fastq file

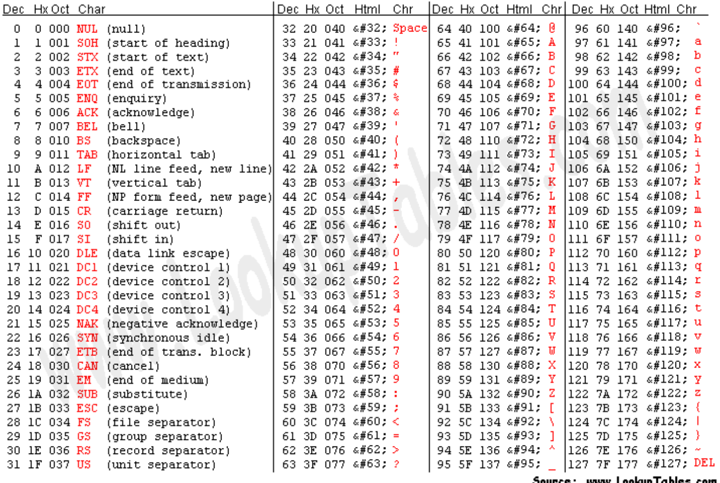
```
@SEQ_ID
GATTTGGGGTTCAAATTAGTATCGATCAAATAGTAAATCCATTTGTTCAACTC
+
!’’*((((***+))%%%++)(%%%%).1***-+*’’))**55CCF>>>>>>CC
```

Fasta file

```
>SEQ_ID
GATTTGGGGTTCAAATTAGTATCGATCAAATAGTAAATCCATTTGTTCAACTC
```

Base qualities are encoded in ascii format

ASCII stands for American Standard Code for Information Interchange. An ASCII code is the numerical representation for a character.

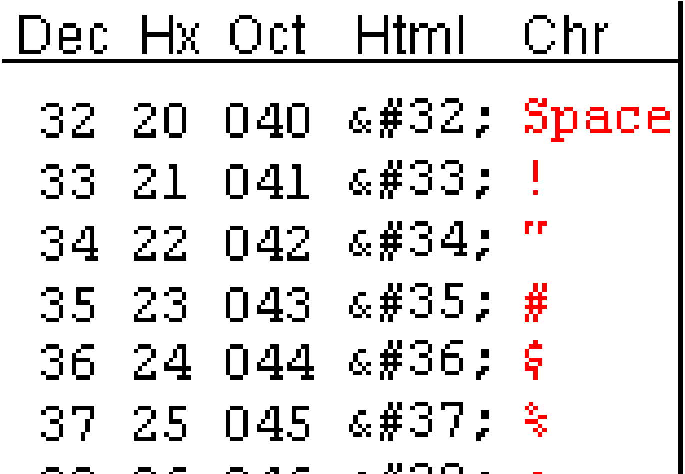


Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040	&#32;	Space	64	40	100	&#64;	@	96	60	140	&#96;	`
1	1	001	SOH (start of heading)	33	21	041	&#33;	!	65	41	101	&#65;	A	97	61	141	&#97;	a
2	2	002	STX (start of text)	34	22	042	&#34;	"	66	42	102	&#66;	B	98	62	142	&#98;	b
3	3	003	ETX (end of text)	35	23	043	&#35;	#	67	43	103	&#67;	C	99	63	143	&#99;	c
4	4	004	EOT (end of transmission)	36	24	044	&#36;	$	68	44	104	&#68;	D	100	64	144	&#100;	d
5	5	005	ENQ (enquiry)	37	25	045	&#37;	%	69	45	105	&#69;	E	101	65	145	&#101;	e
6	6	006	ACK (acknowledge)	38	26	046	&#38;	&	70	46	106	&#70;	F	102	66	146	&#102;	f
7	7	007	BEL (bell)	39	27	047	&#39;	'	71	47	107	&#71;	G	103	67	147	&#103;	g
8	8	010	BS (backspace)	40	28	050	&#40;	(	72	48	110	&#72;	H	104	68	150	&#104;	h
9	9	011	TAB (horizontal tab)	41	29	051	&#41;	)	73	49	111	&#73;	I	105	69	151	&#105;	i
10	A	012	LF (NL line feed, new line)	42	2A	052	&#42;	*	74	4A	112	&#74;	J	106	6A	152	&#106;	j
11	B	013	VT (vertical tab)	43	2B	053	&#43;	+	75	4B	113	&#75;	K	107	6B	153	&#107;	k
12	C	014	FF (NP form feed, new page)	44	2C	054	&#44;	,	76	4C	114	&#76;	L	108	6C	154	&#108;	l
13	D	015	CR (carriage return)	45	2D	055	&#45;	-	77	4D	115	&#77;	M	109	6D	155	&#109;	m
14	E	016	SO (shift out)	46	2E	056	&#46;	.	78	4E	116	&#78;	N	110	6E	156	&#110;	n
15	F	017	SI (shift in)	47	2F	057	&#47;	/	79	4F	117	&#79;	O	111	6F	157	&#111;	o
16	10	020	DLE (data link escape)	48	30	060	&#48;	0	80	50	120	&#80;	P	112	70	160	&#112;	p
17	11	021	DC1 (device control 1)	49	31	061	&#49;	1	81	51	121	&#81;	Q	113	71	161	&#113;	q
18	12	022	DC2 (device control 2)	50	32	062	&#50;	2	82	52	122	&#82;	R	114	72	162	&#114;	r
19	13	023	DC3 (device control 3)	51	33	063	&#51;	3	83	53	123	&#83;	S	115	73	163	&#115;	s
20	14	024	DC4 (device control 4)	52	34	064	&#52;	4	84	54	124	&#84;	T	116	74	164	&#116;	t
21	15	025	NAK (negative acknowledge)	53	35	065	&#53;	5	85	55	125	&#85;	U	117	75	165	&#117;	u
22	16	026	SYN (synchronous idle)	54	36	066	&#54;	6	86	56	126	&#86;	V	118	76	166	&#118;	v
23	17	027	ETB (end of trans. block)	55	37	067	&#55;	7	87	57	127	&#87;	W	119	77	167	&#119;	w
24	18	030	CAN (cancel)	56	38	070	&#56;	8	88	58	130	&#88;	X	120	78	170	&#120;	x
25	19	031	EM (end of medium)	57	39	071	&#57;	9	89	59	131	&#89;	Y	121	79	171	&#121;	y
26	1A	032	SUB (substitute)	58	3A	072	&#58;	:	90	5A	132	&#90;	Z	122	7A	172	&#122;	z
27	1B	033	ESC (escape)	59	3B	073	&#59;	;	91	5B	133	&#91;	[	123	7B	173	&#123;	{
28	1C	034	FS (file separator)	60	3C	074	&#60;	<	92	5C	134	&#92;	\	124	7C	174	&#124;	|
29	1D	035	GS (group separator)	61	3D	075	&#61;	=	93	5D	135	&#93;	]	125	7D	175	&#125;	}
30	1E	036	RS (record separator)	62	3E	076	&#62;	>	94	5E	136	&#94;	^	126	7E	176	&#126;	~
31	1F	037	US (unit separator)	63	3F	077	&#63;	?	95	5F	137	&#95;	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

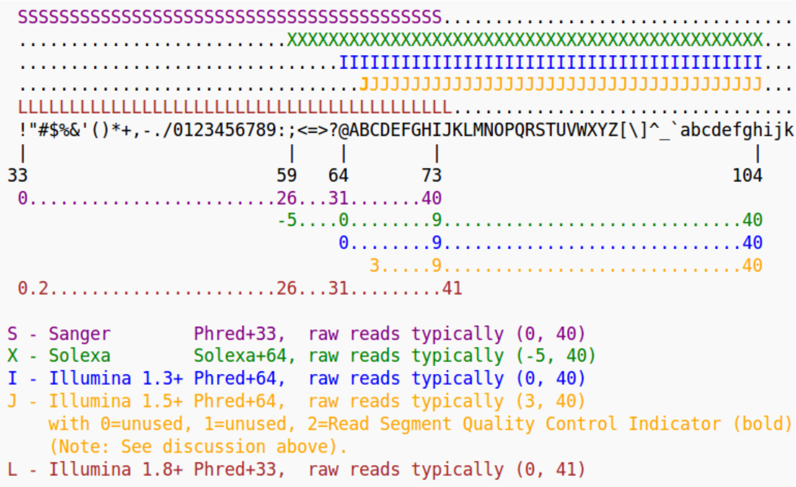
Base qualities are encoded in ascii format

ASCII stands for American Standard Code for Information Interchange. An ASCII code is the numerical representation for a character.



Dec	Hx	Oct	Html	Chr
32	20	040	&#32;	Space
33	21	041	&#33;	!
34	22	042	&#34;	"
35	23	043	&#35;	#
36	24	044	&#36;	$
37	25	045	&#37;	%

ASCII encodings of sequencing platforms



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS..................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII...
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ...
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijk
 |                         |    |        |                              |
33                        59   64       73                            104
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

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)
```

Figure: Quality score encodings

Quality control tool: FastQC

Informs on:

- Base quality
- Duplication
- Overrepresentation of sequences
  - contamination?
  - adapters?
- GC content (should be around 50%, in Bis-Seq lower)

Quality before trimming

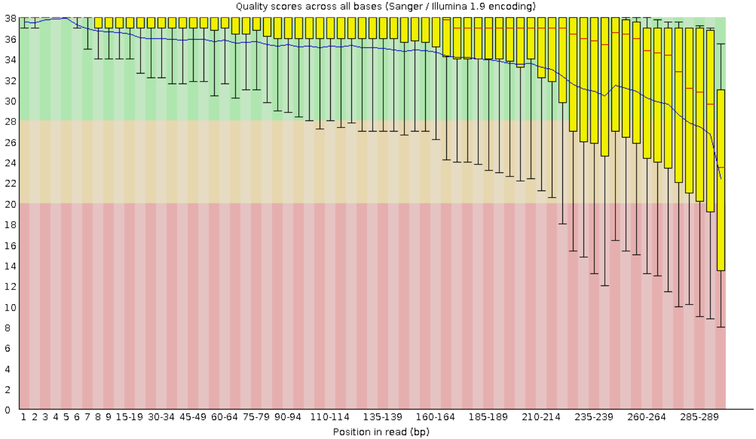


Figure: Base-quality generally decreases with increasing sequencing length

Quality after trimming

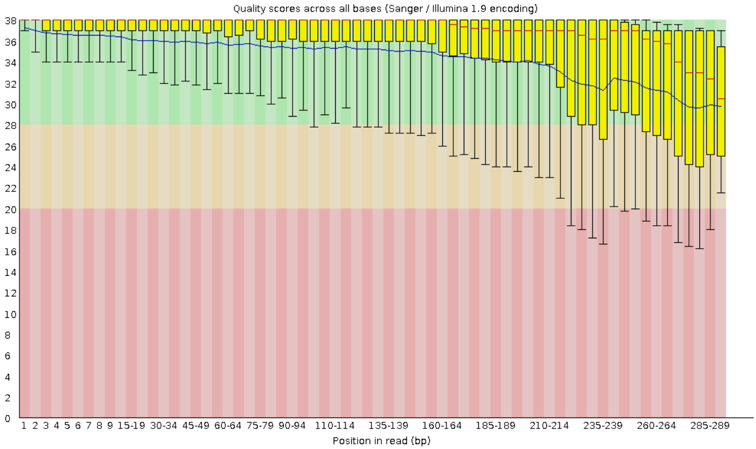
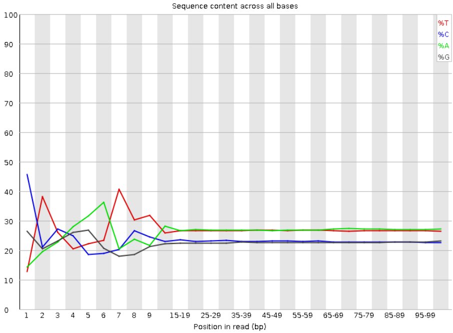


Figure: Quality after trimming

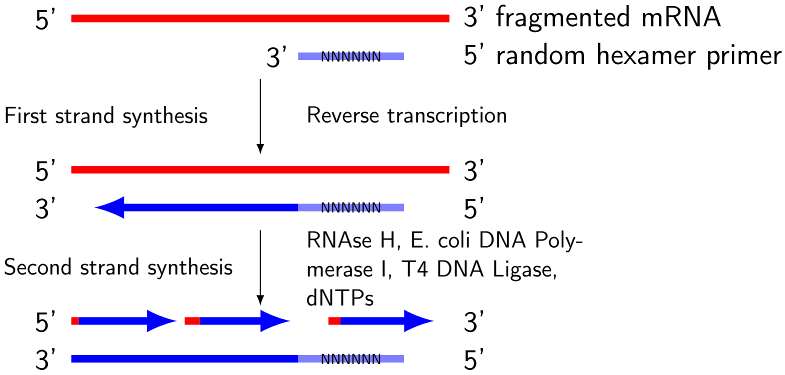
Sequence bias

For example in:

- First bases of Illumina RNAseq due to 'random' hexamer primers for reverse transcription
- RADseq fragments (cutting sites)



Hexamer primers for cDNA synthesis cause sequence bias



PCR Duplicates

Duplicates are generally removed in quantitative analyses (e.g. RNA-seq)

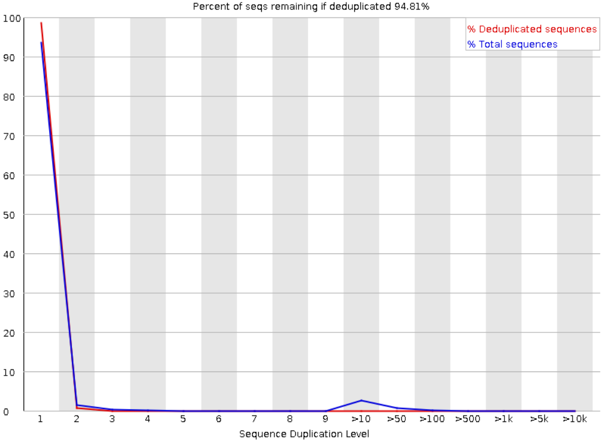


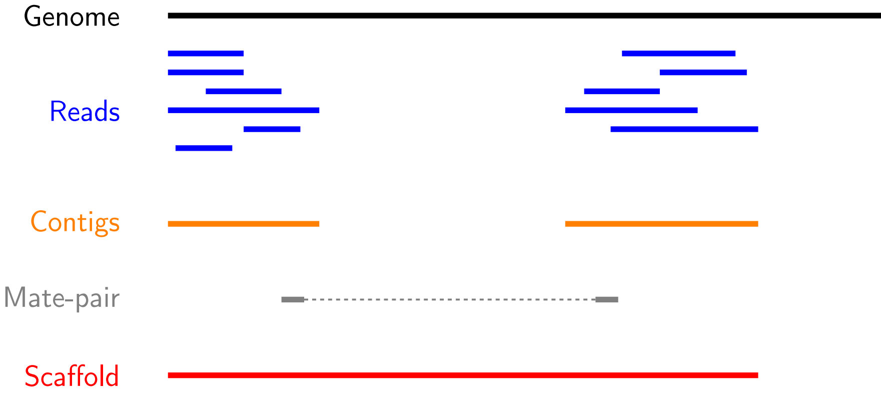
Figure: Duplication levels (FastQC output)

De novo assembly

Task: Look for overlapping regions and create contigs (contiguous sequences)

- Genome assembly software
  - SOAP de NOVO
  - Velvet
  - MIRA (we use this one in the course)
- Transcriptome assembly software
  - Review: Martin and Wang, (2011)
  - Trinity
  - MIRA

De novo assembly: Step by step



De novo assembly: The N50 metric

N50 is a single measure of the contig length size distribution in an assembly

- Sort contigs in descending length order
- Size of contig above which the assembly contains at least 50% of the total length of all contigs

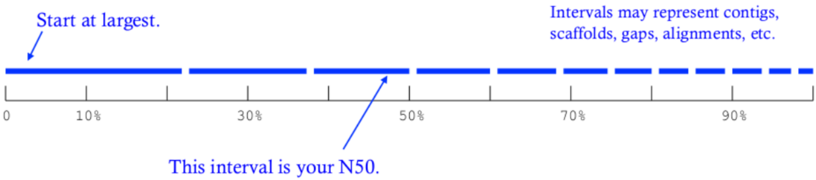
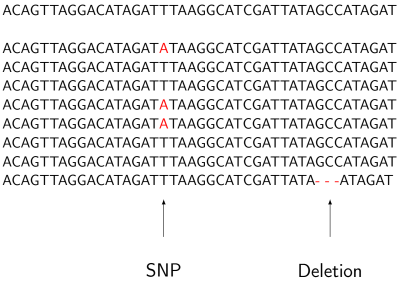

Start at largest.
Intervals may represent contigs, scaffolds, gaps, alignments, etc.
0
10%
30%
50%
70%
90%
This interval is your N50.

Figure: From Kane, N.C.

Mapping against reference genome/transcriptome

- Main purposes:
  - Identify variants (SNPs, InDels)



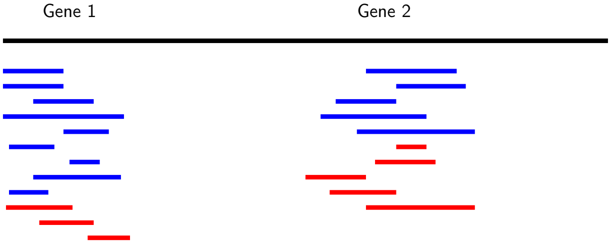
```
ACAGTTAGGACATAGATTTAAGGCATCGATTATAGCCATAGAT

ACAGTTAGGACATAGATATAAGGCATCGATTATAGCCATAGAT
ACAGTTAGGACATAGATTTAAGGCATCGATTATAGCCATAGAT
ACAGTTAGGACATAGATTTAAGGCATCGATTATAGCCATAGAT
ACAGTTAGGACATAGATATAAGGCATCGATTATAGCCATAGAT
ACAGTTAGGACATAGATATAAGGCATCGATTATAGCCATAGAT
ACAGTTAGGACATAGATTTAAGGCATCGATTATAGCCATAGAT
ACAGTTAGGACATAGATTTAAGGCATCGATTATAGCCATAGAT
ACAGTTAGGACATAGATTTAAGGCATCGATTATA- - -ATAGAT
```

SNP

Deletion

Mapping against reference genome/transcriptome

- Main purposes:
  - Quantify gene expression

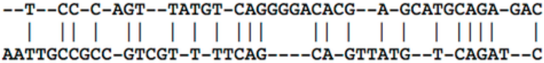


Population 1

Population 2

Mapping: global alignment

- Implemented in e.g. BWA, Bowtie2
- Needleman-Wunsch algorithm
- Aligns sequences in their full length
- Used for multiple sequence alignment when sequences are similar

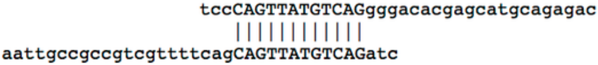


```
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || |  | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C
```

Figure: Global alignment from rosalind.info

Mapping: local alignment

- Smith-Waterman algorithm
- Clipping of terminal unmatched bases
- Only aligned bases contribute to the alignment's score
- Used to target smaller portions of genes with high similarity



```
                  tccCAGTTATGTCAGgggacacgagcatgcagagac
                     ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Figure: Local alignment from rosalind.info

Splice-aware alignment of RNAseq reads to the genome

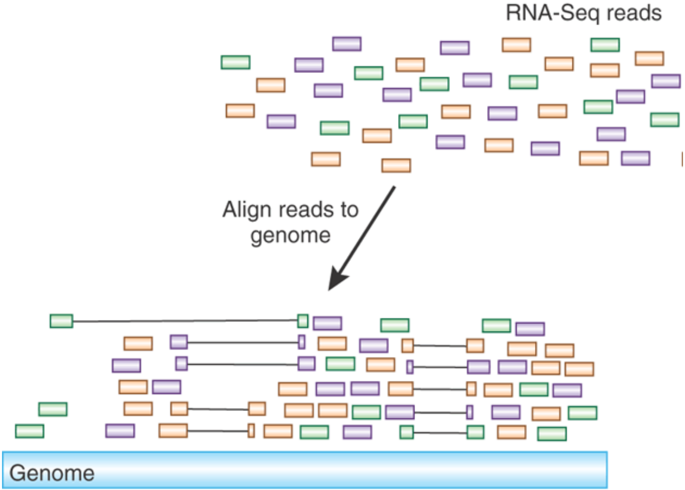
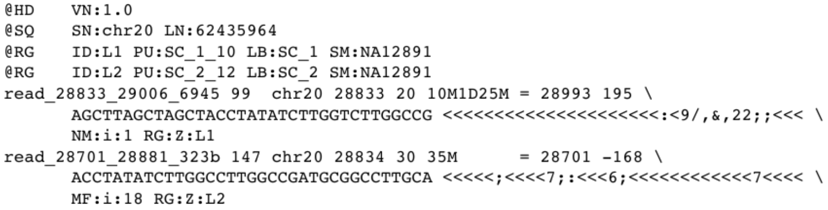


Figure: Adapted from Haas and Zody, (2010)

Mapping: SAM/BAM files example

Output format of most alignment programs

- Header lines preceded by `@`
- One tab-delimited line per read



```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
       AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
       NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
       ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<<<7<<<< \
       MF:i:18 RG:Z:L2
```

Figure: Example from `http://samtools.sourceforge.net/SAM1.pdf`

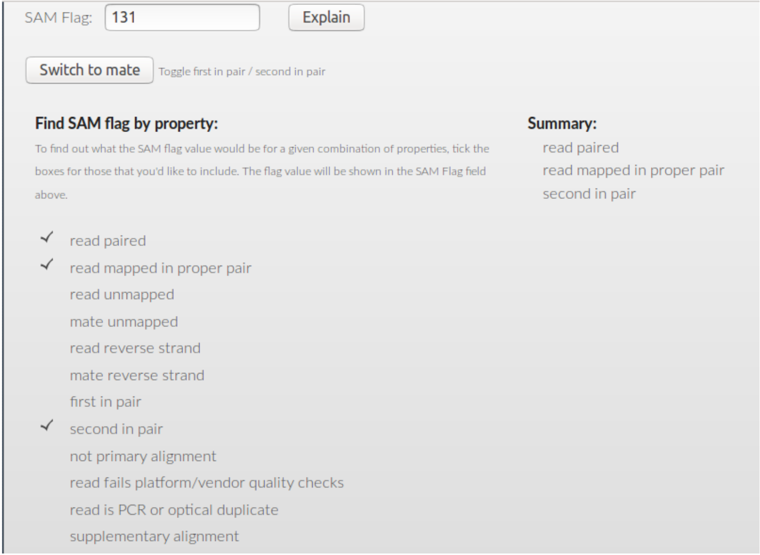
- SAM files are large
- BAM: Compressed binary versions, not human-readable

Mapping: Mandatory fields in SAM files

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	[0,2^{31}-1]	Position of the mate/next read
9	TLEN	Int	[-2^{31}+1,2^{31}-1]	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Explanation of the flag field (click here: Link1, Link2)

Mapping: Easy decoding of SAM flags



SAM Flag: 131

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- ✓ read paired
- ✓ read mapped in proper pair
- read unmapped
- mate unmapped
- read reverse strand
- mate reverse strand
- first in pair
- ✓ second in pair
- not primary alignment
- read fails platform/vendor quality checks
- read is PCR or optical duplicate
- supplementary alignment

Summary:

read paired
read mapped in proper pair
second in pair

Mapping: CIGAR string in SAM files

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Mapping: CIGAR string example

RefPos:	1	2	3	4	5	6	7		8	9	10	11	12	13	14	15	16
Ref:	C	C	A	T	A	C	T		G	A	A	C	T	G	A	C	T
Read:					A	C	T	A	G	A	A		T	G	G	C	T

CIGAR: 3M1I3M1D5M

Variant calling

Consistent mismatches in the alignment indicate:

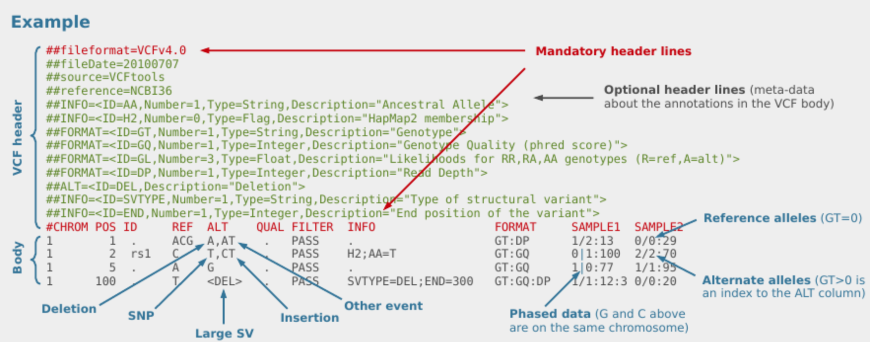
- Single Nucleotide Polymorphisms (SNPs)
- Insertions/Deletions (InDels)

VCF file format

Variant call format

- described in `http://www.1000genomes.org/node/101`
- informs on location and quality of each SNP

VCF file information



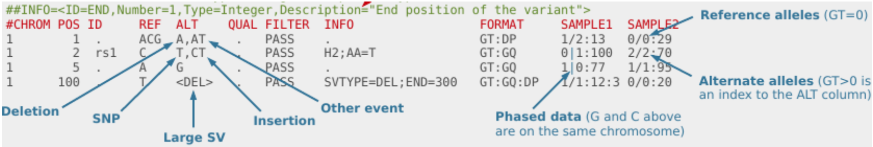
Example
VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL  FILTER  INFO                FORMAT    SAMPLE1   SAMPLE2
1        1 .      ACG  A,AT   .     PASS    .                   GT:DP     1/2:13    0/0:29
1        2 rs1    C    T,CT   .     PASS    H2;AA=T             GT:GQ     0|1:100   2/2:70
1        5 .      A    G      .     PASS    .                   GT:GQ     1|0:77    1/1:95
1      100 .      T    <DEL>  .     PASS    SVTYPE=DEL;END=300  GT:GQ:DP  1/1:12:3  0/0:20
```

Figure: VCF file info from `http://vcftools.sourceforge.net/VCF-poster.pdf`

Phased alleles are on the same chromosome strand

VCF file information



```
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF ALT     QUAL FILTER  INFO                 FORMAT      SAMPLE1    SAMPLE2
1        1 .      ACG A,AT     .   PASS    .                    GT:DP       1/2:13     0/0:29
1        2 rs1    C   T,CT     .   PASS    H2;AA=T              GT:GQ       0|1:100    2/2:70
1        5 .      A   G        .   PASS    .                    GT:GQ       1|0:77     1/1:95
1      100 .      T   <DEL>    .   PASS    SVTYPE=DEL;END=300   GT:GQ:DP    1/1:12:3   0/0:20
```

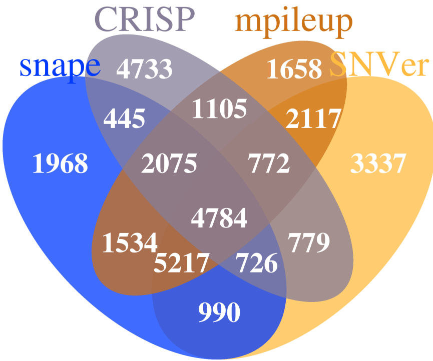
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

Figure: VCF file info from `http://vcftools.sourceforge.net/VCF-poster.pdf`

Phased alleles are on the same chromosome strand

Identified SNPs vary between programs/algorithms

Venn diagram of the number of SNPs (coverage >400) called with four programs from the same alignment file (ddRAD tags mapped against the genome of Guppy).



Differential gene expression analysis

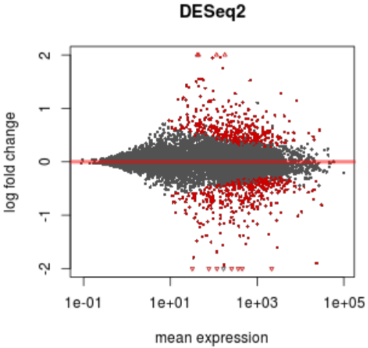


Figure: Log2 fold-change of expression over the mean of counts normalized by size factors. Differentially expressed genes (p<0.1) are red.

From the DESeq2 R package documentation

Clustering

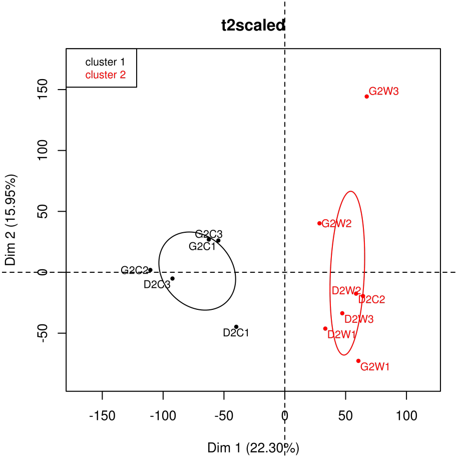


Figure: Multivariate grouping of stressed (W) and control (C) seagrass samples. Most variation is explained by the first principle component

Visualizing differential expression

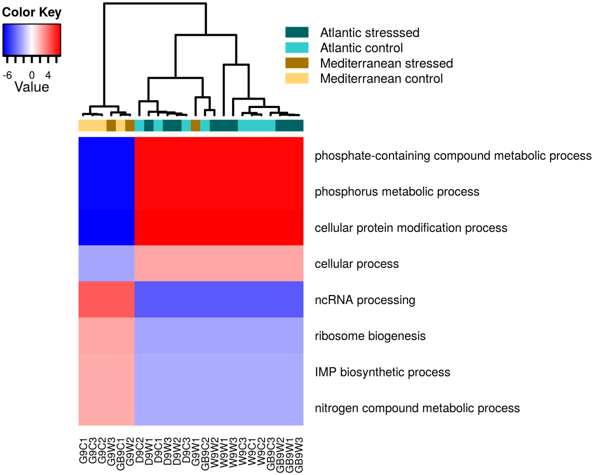
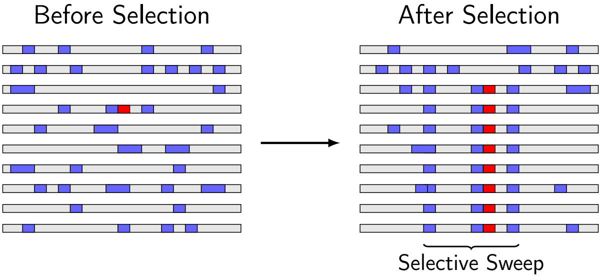


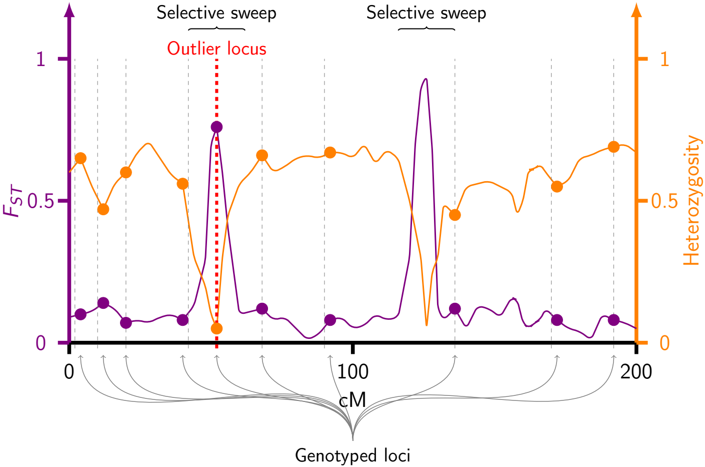
Figure: Heatmap of functions that were differentially expressed between Atlantic and Mediterranean seagrass samples.

Outlier analysis



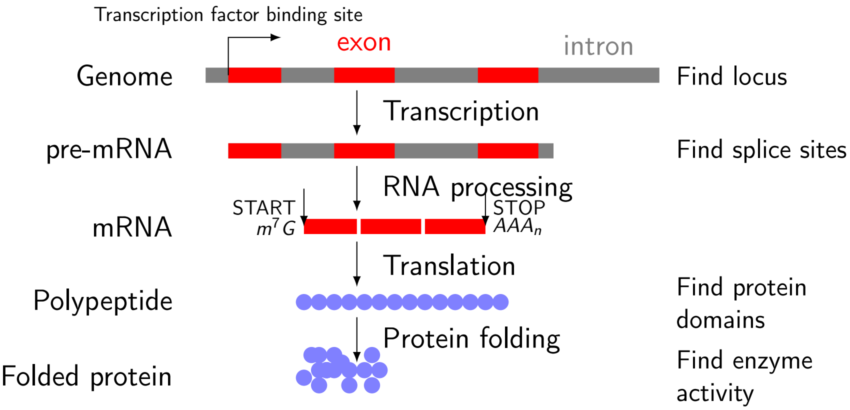
Based on Vitti et al., (2012)

Outlier detection



Eukaryote genome annotation

Identify the strcuture and functional role



Transcription factor binding site

exon

intron

Genome — Find locus

Transcription

pre-mRNA — Find splice sites

RNA processing

START m^7G

STOP AAA_n

mRNA

Translation

Polypeptide

Find protein domains

Protein folding

Folded protein

Find enzyme activity

Gene ontologies

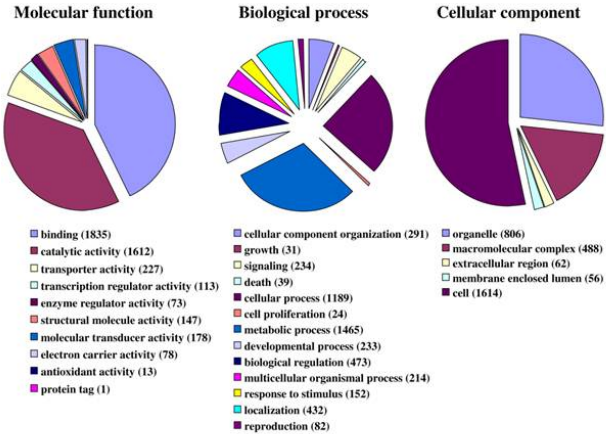


Figure: GO terms of unigenes in a moth genome

(Jacquin-Joly et al., 2012)

Cloud of GO term enrichments

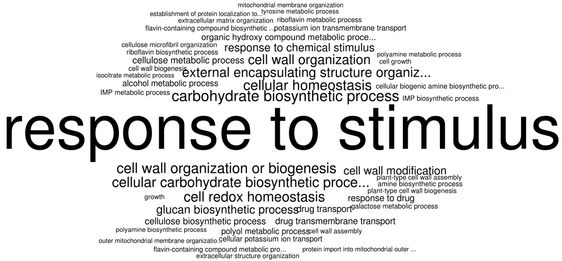


Figure: Term cloud of heat-responsive functions in seagrass

Bioinformatics-Practical

- Unix Tools (Martin)
- Trimming and Quality Control (Alexander)
- Genome Assembly (Alexander)
- Mapping and Variant Calling (Martin)

References

- Haas, BJ and MC Zody (2010). "Advancing RNA-seq analysis". In: *Nature biotechnology* 28.5, pp. 421–423.
- Hansen, KD, SE Brenner, and S Dudoit (2010). "Biases in Illumina transcriptome sequencing caused by random hexamer priming". In: *Nucleic acids research* 38.12, e131–e131.
- Jacquin-Joly, E, F Legeai, N Montagné, C Monsempes, MC François, J Poulain, et al. (2012). "Candidate chemosensory genes in female antennae of the noctuid moth Spodoptera littoralis". In: *International journal of biological sciences* 8.7, p. 1036.
- Martin, J and Z Wang (2011). "Next-generation transcriptome assembly". In: *Nature Reviews Genetics*.
- Vitti, JJ, MK Cho, SA Tishkoff, and PC Sabeti (2012). "Human evolutionary genomics: ethical and interpretive issues". In: *Trends in Genetics* 28.3, pp. 137–145.