

Day 2: Read Trimming

MMB-114

Schedule

Day 1: Basics of UNIX and working with the command line

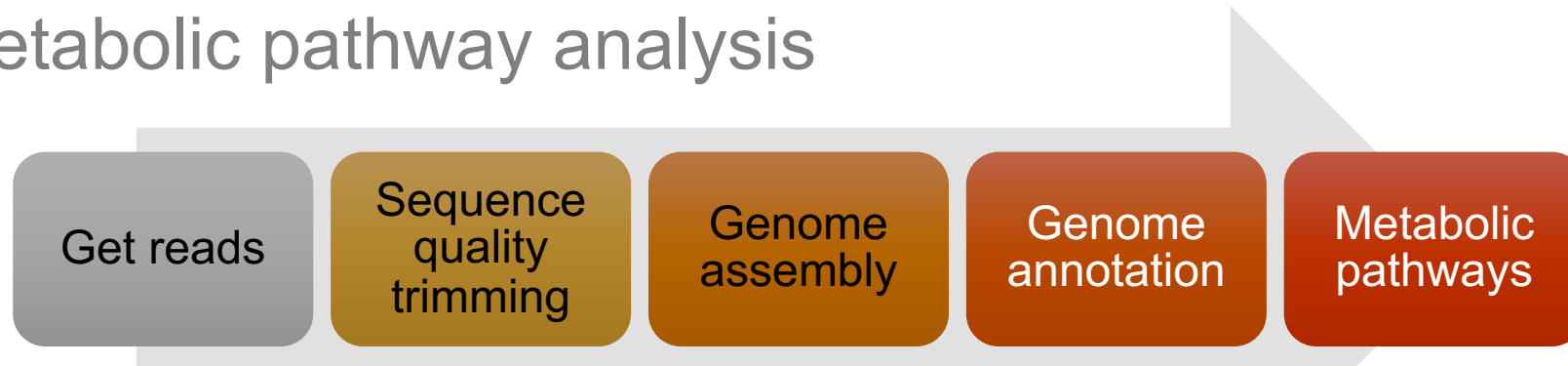
Day 2: Handling of Nanopore/Illumina data

Day 3: Check-up

Day 4: Genome assembly

Day 5: Genome annotation

Day 6: Metabolic pathway analysis



UNIX recap

- Let's go thru the UNIX exercises together
https://github.com/karkman/MMB-114_Genomics
- Any questions?

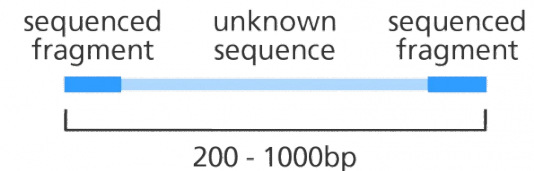
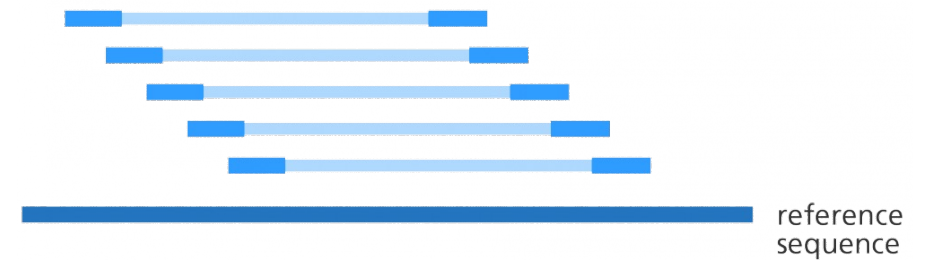
Raw Illumina sequencing data

- Our Illumina data is paired-end
(nanopore is single-end)
- What does that mean?

Single-end reads



Paired-end reads



FASTQ files

@M02764:119:000000000-C5R9K:1:1101:15139:1363 1:N:0:24

GCTAACCCCATTTGCAACGTGGTAACCTTGTAGACCGTTTTTAAAGTCGCTGAAGCAGCCACGATAAACGACATCCCGATTGAACCCCAAGGACCATGTTGACCACGCCCCAAGCTGCACCGGCATCTGCAGGTAACG
TATTGGCAACACCGGCGACCGTTAATGGACTGAGTGTAAACCTTGTCCAATCTCCATCAGGATCATGGGAACCTCAATACCTAGCGCATAGCCCACTTGTCCTCCGAGAAAAAGCCTAAGGCACCCACCCCAATTGCGG
TGTTGCAATCCCAACGATTAGTAATTTCTGTTTCCAAAAGCGCTTTGTT

+

CCCCCEFSGGGGFCGGGGGGGGGGFCCDFGGFFCCCFCC8EFF@FFCCECC@F<FGFGGGGGG7B7<C7:FFEE@EGGDGCEFFFCBFG7D,CFC,EEECGG<FG>@<F=EG=7<<EB8:+8:@@FDDFGGG8FDEC
F8FFFEFGGCEGDD:7++@FGDEE>EG9AFBCC<>>FCCFB,:8?FG;;B,?8,6B;BC@C6E6823B9C?+8C**0><++++1+6***=C:E+*:7:?7+*395CD)92*9<FFFCFFFC<FD;))1>55)* /29@)
<))00)7*1*)09))7))54-=8*9.9/5*-)-*)..3:(((43)).(

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Phred scores (quality score)

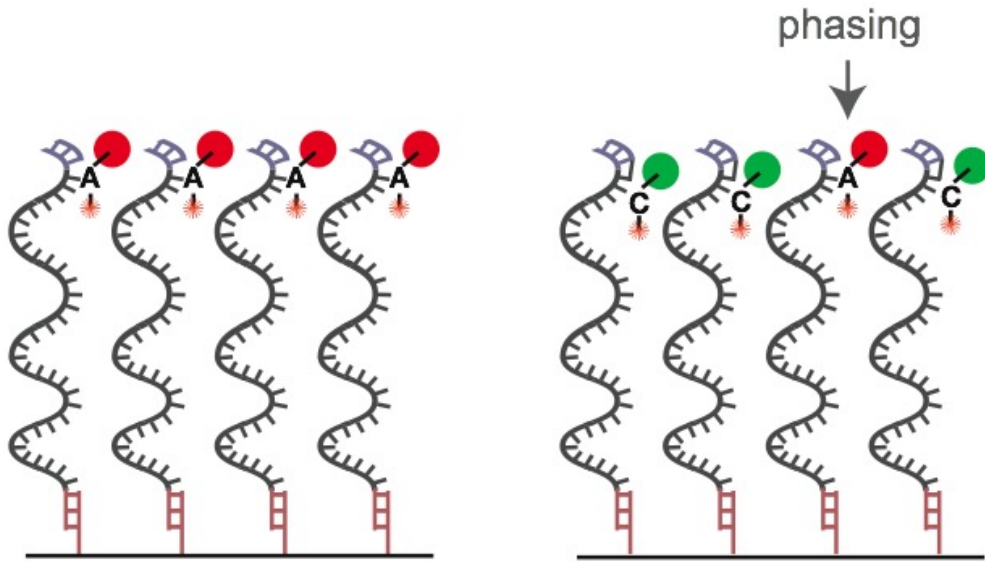
- Indicates the probability that a given base is called correctly by the sequencer
- Defined logarithmically to the base calling error probability

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

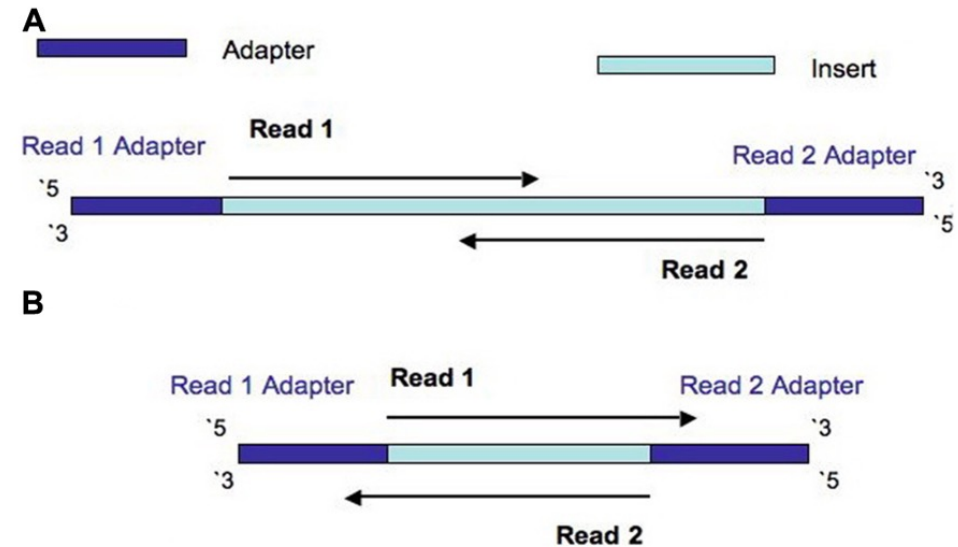
Phred Quality Score	Probability of incorrect base call	Base call accuracy	ASCII
10	1 in 10	90%	+
20	1 in 100	99%	5
30	1 in 1,000	99.9%	?
40	1 in 10,000	99.99%	

Quality filtering and adapter removal

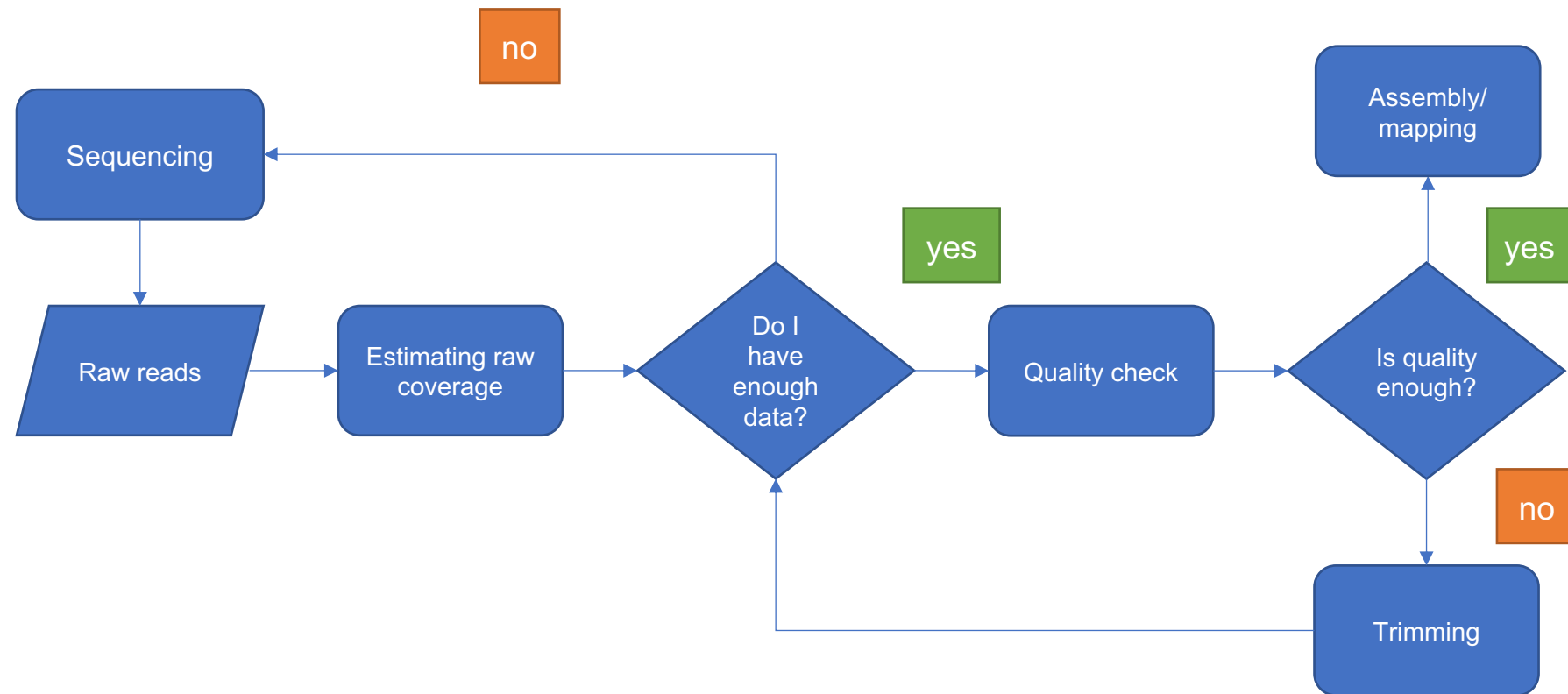
Phasing



Adapter read-through



Read quality assessment



Additional basic UNIX commands

Compressing/decompressing

- tar
- gunzip

Inspect files

- head
- tail
- less/more

Operations

- Piping (|)
- Redirection (>)

Remember:

- Commands have to be typed in a single line, one at a time
 - Backslash (\)
- Whenever possible, type everything, don't copy and paste
 - Tabulator!

Additional notes about Puhti

Interactive partition

- `sinteractive`
- Max. running time: 7 days (default 24 h)
- Computing capacity up to 8 cores (default 1)
- Memory capacity of up to 76 GB (default 2 GB)

The module system

- `module load biokit`
- Convenient way to manage applications

Time to take a look at our genome data

https://github.com/karkman/MMB-114_Genomics

(Day 2: Read trimming)