Exam Review

DISCLAIMER: This is a non-exhaustive list

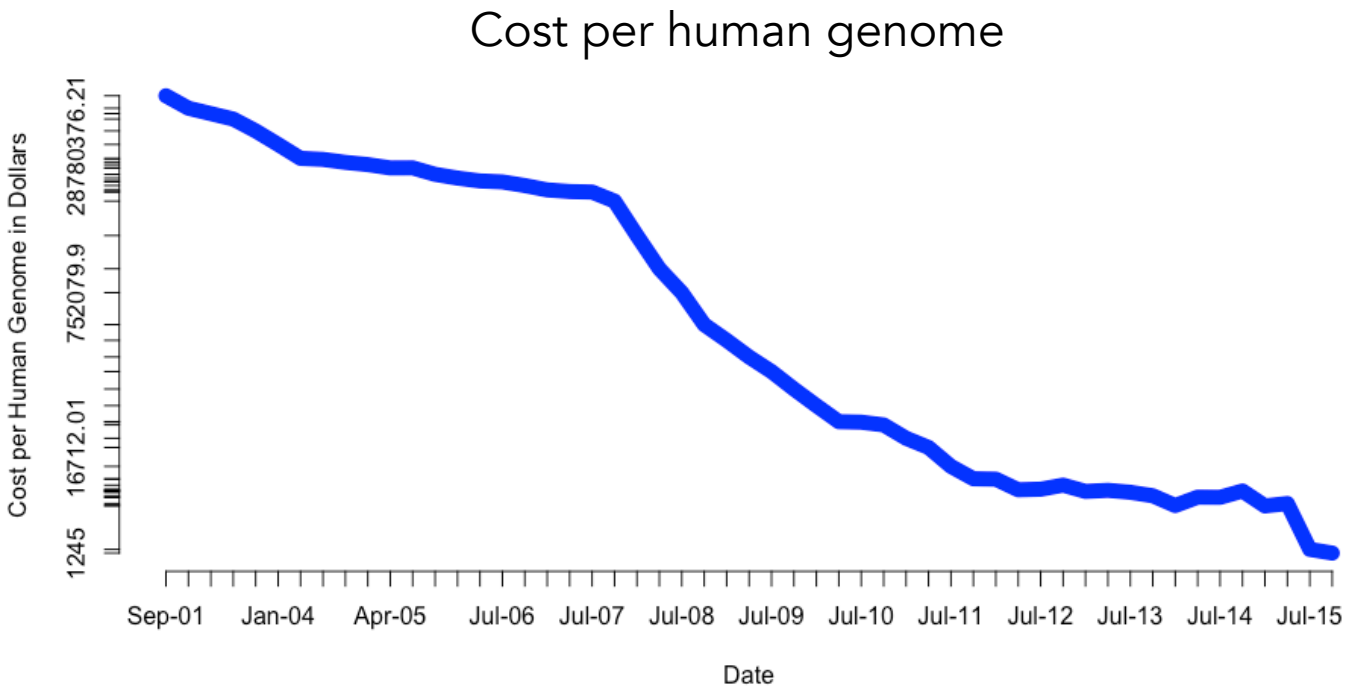
Lecture 15
Oct 5, 2016

Announcements

Lecture 1

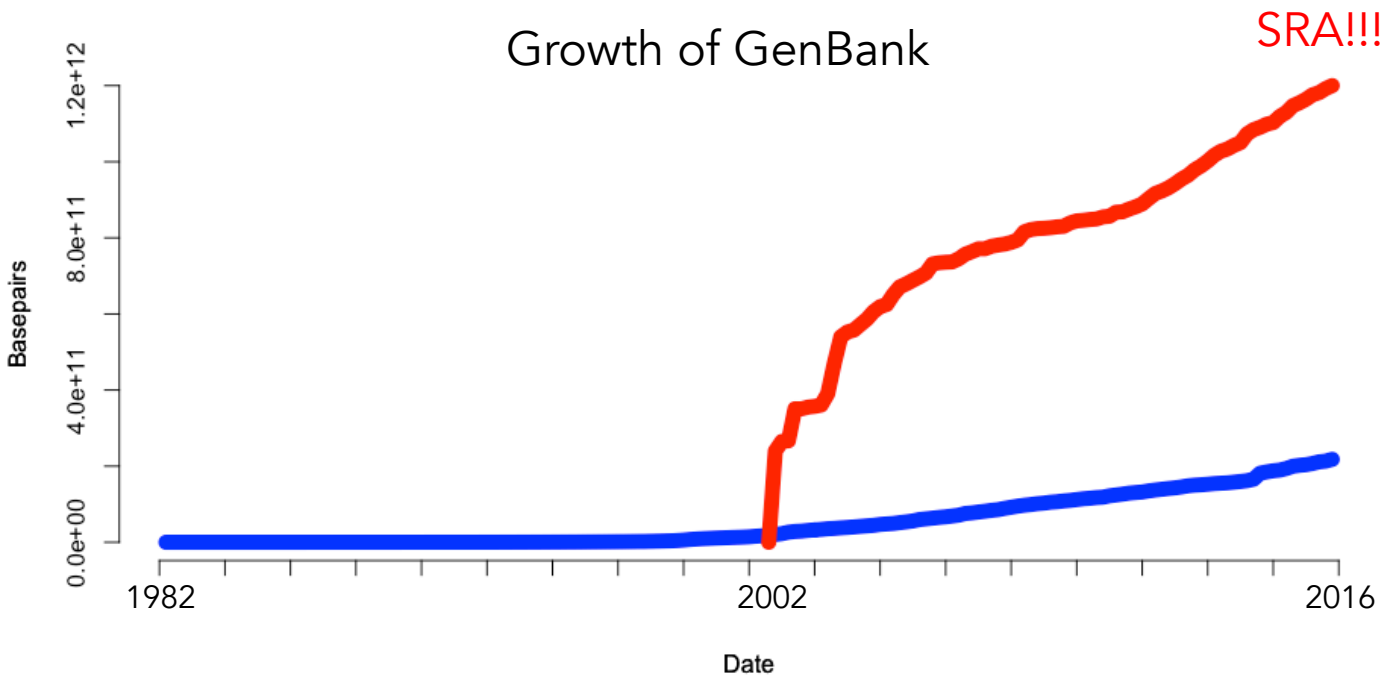
Bioinformatics

- Why now??



Bioinformatics

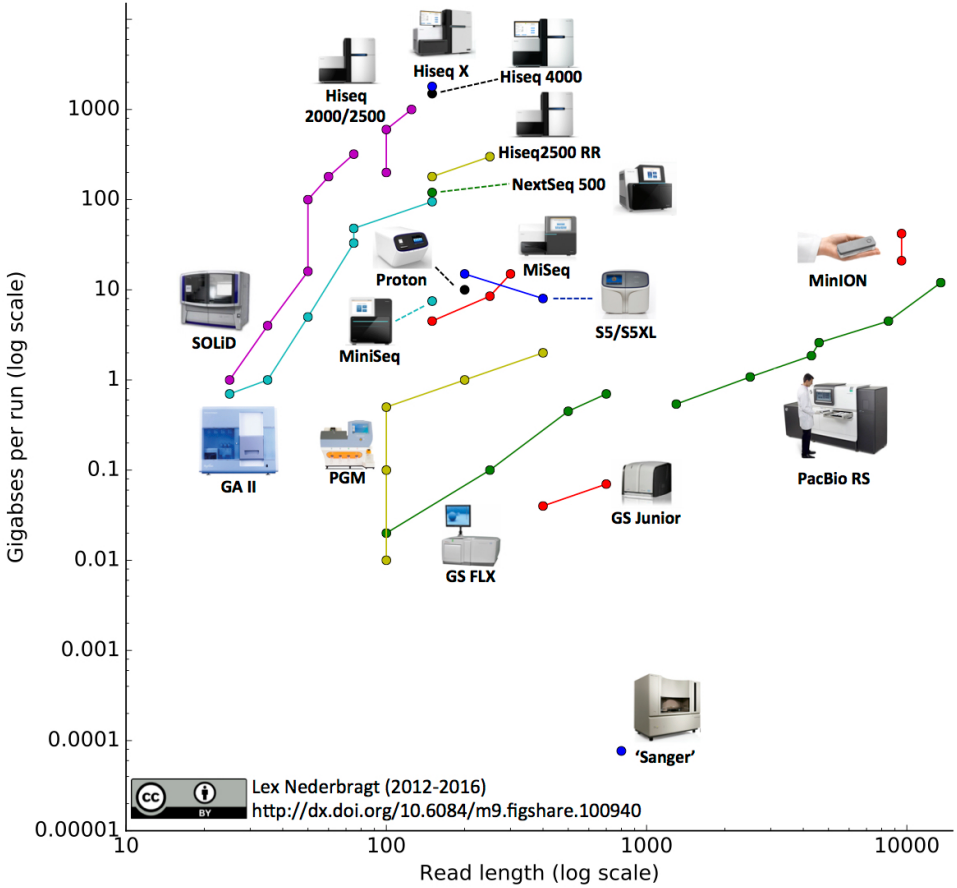
- Why now??



Lecture 2

Lecture 2

Sequencing Platforms



Lecture 3

Sequencing Platforms

	Illumina	PacBio	ONT
Read Length			
Error Rate			
Throughput			
Expense			

Lecture 4

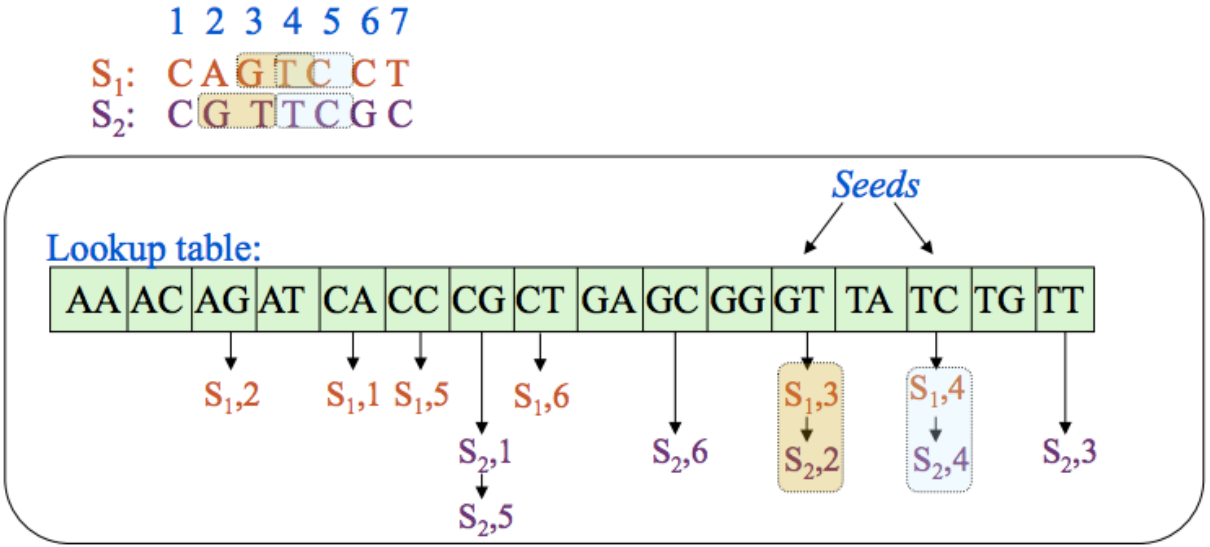
Mechanisms of Evolution

- Gene Flow
- Mutation
- Drift
- Natural Selection

Lecture 5

BLAST

2. Filter low complexity and identify seeds



BLAST

Stats

$$E = Kmne^{-\lambda S}$$

Smith Waterman

$$V[i,j] = \max \begin{cases} V[i-1,j] + s(x[i-1], -) \\ V[i,j-1] + s(-, y[j-1]) \\ V[i-1,j-1] + s(x[i-1], y[j-1]) \\ 0 \end{cases}$$

	ϵ	T	A	T	A	T	G	C	G	G	C	G	T	T	T
ϵ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	2	0	2	2	0	2	0	0	0
G	0	0	0	0	0	0	2	0	2	4	0	2	0	0	0
T	0	2	0	2	0	2	0	0	0	0	0	0	4	2	2
A	0	0	4	0	?										
T	0														
G	0														
C	0														
T	0														
G	0														
G	0														
C	0														
G	0														
C	0														
T	0														
A	0														

$s(a,b)$

	A	C	G	T	-
A	2	-4	-4	-4	-6
C	-4	2	-4	-4	-6
G	-4	-4	2	-4	-6
T	-4	-4	-4	2	-6
-	-6	-6	-6	-6	

Lecture 6

PAM versus Divergence

PAM0 PAM1 PAM250 PAM∞

Less divergent More divergent

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp

Lecture 7

fastQ

@HSQ-7001360:67:H88RHADXX:1:1101:1448:2158 1:N:0:CAGATC

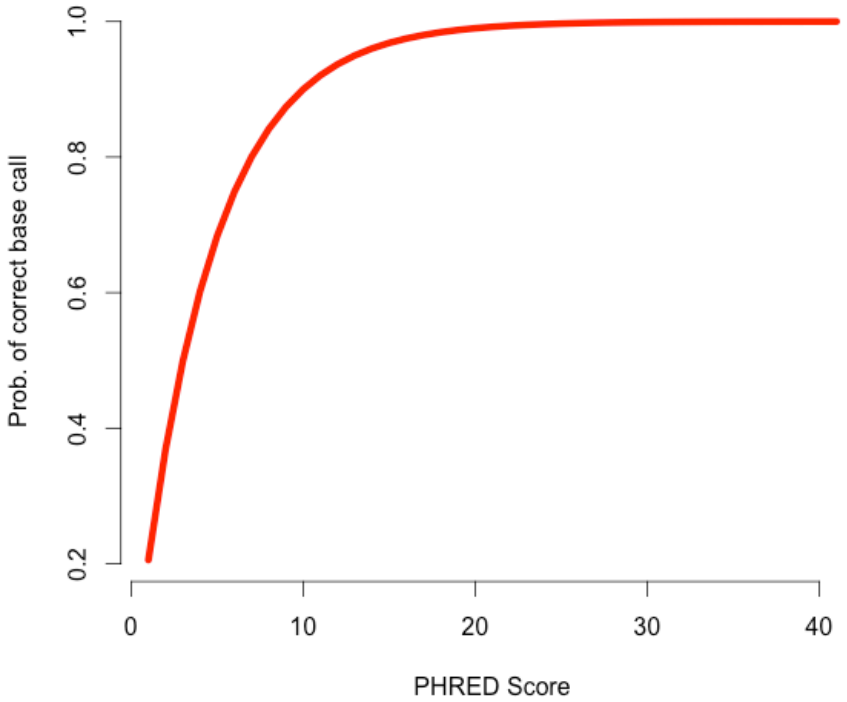
- HSQ-7001360= Instrument name
- 67= run ID
- H88RHADXX=Flowcell ID
- 1=lane 1
- 1101=tile number
- 1448= x coordinate
- 2158= y coordinate
- 1=left read
- N=not filtered
- 0=control bit -> (not used anymore)
- CAGATC= adapter sequence

Lecture 9

Quality Scores

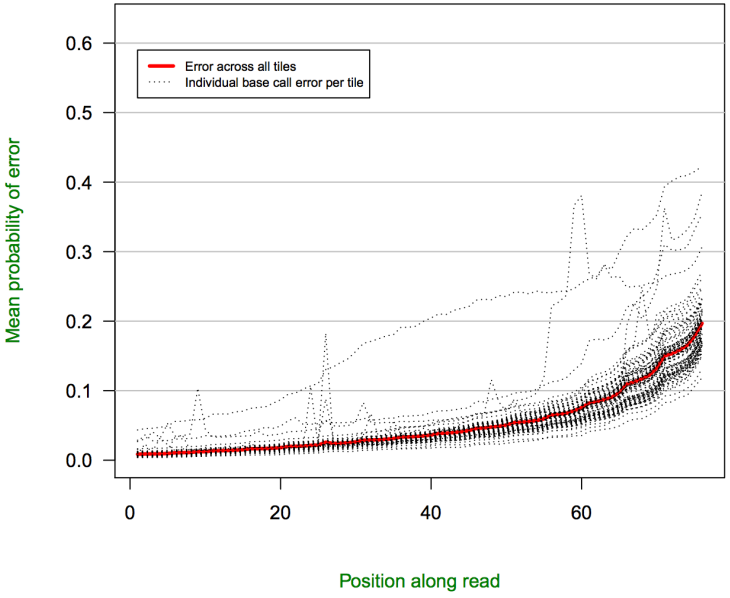
$$p_{correct} = 1 - [10^{-\left(\frac{Q}{10}\right)}]$$

Q=Phred score



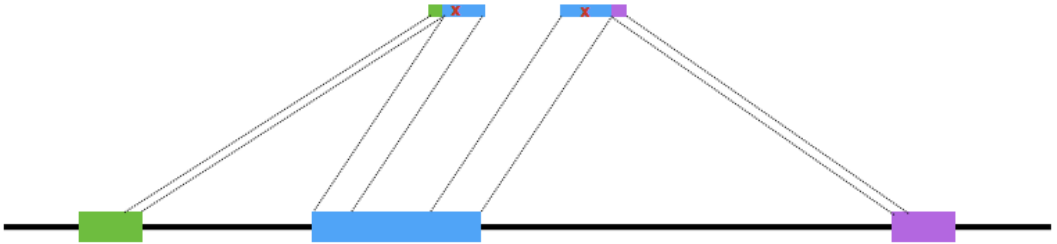
Quality Scores

What should we do about this?



Lecture 10

Spliced Alignment



Splice junctions might be known, or *unknown.*

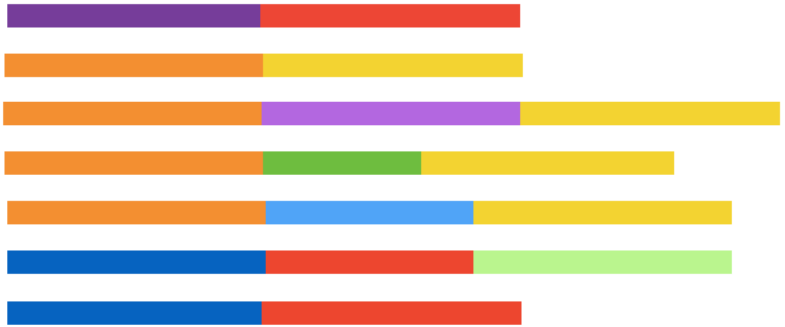
Overlap of read with exon may be *very short*, sequence is ambiguous (e.g. 10 bases).

Sequence of read might be repetitive in the genome.

Aligning reads to a Transcriptome

Consider the following scenario:

Transcripts



Read

Lecture 11

MAPPING - BWT

A	B	A	A	B	A	
$	a	b	a	a	b	a
a	$	a	b	a	a	b
a	a	b	a	$	a	b
a	b	a	$	a	b	a
a	b	a	a	b	a	$
b	a	$	a	b	a	a
b	a	a	b	a	$	a

Lecture 12

SAM/BAM FILE CONTENTS

Alignment Fields

- Col1
- Col2
- Col3
- Col4
- Col5
- Col6
- Col7
- Col8
- Col9
- Col10
- Col11

Lecture 13

Error Correction