

Software engineering

Steven Salzberg

What is software engineering? Why does it matter?

Math is one thing

$$z = \frac{x}{y}$$

Programming is another

if $y \neq 0$ then $\left\{ z = \frac{x}{y} \right\}$

Why do we need to understand genomics software?



DNA

ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT



transcription

RNA

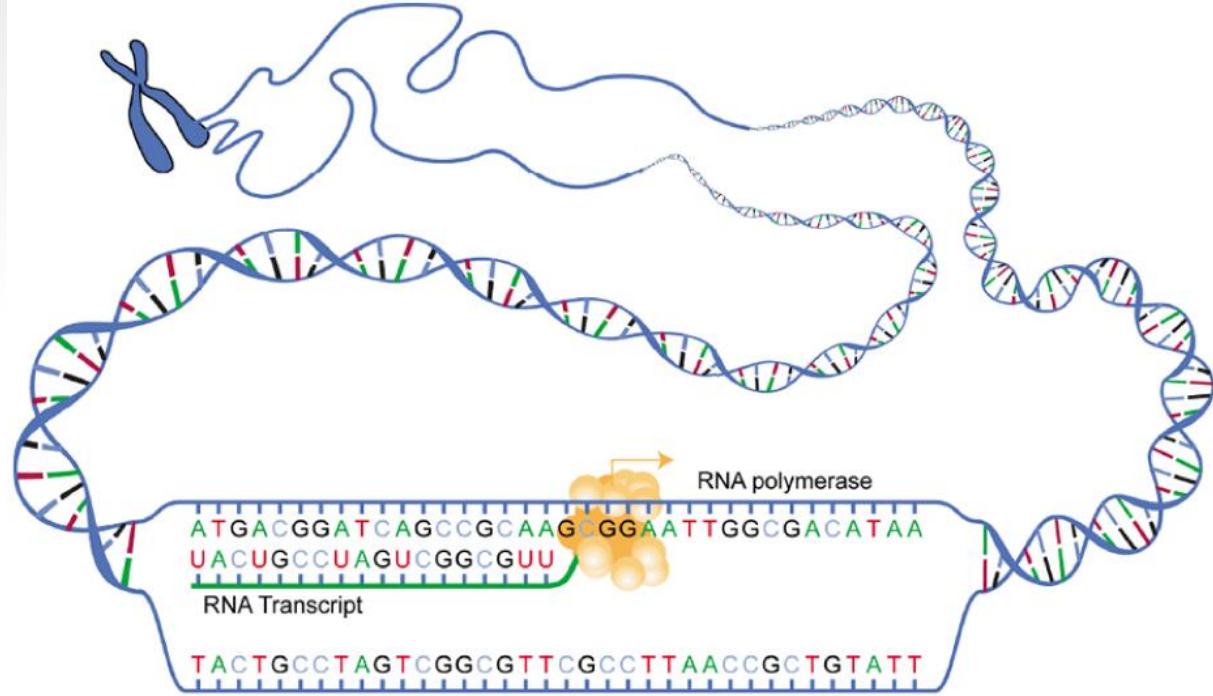
AUCAGUCGAUCACCGAU



editing?

AUCAGUCG**C**UCACCGAU

How can we detect RNA editing?



Alignment reveals RNA-DNA differences

H.sapiens	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA	primates	mammals
H.sapiens13G>A	UGGCC.	GAUUUUUGGACACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
H.sapiens14C>A	UGGCC.	GAUUUUUGGAACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
P.troglodytes	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
P.paniscus	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
G.gorilla	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
M.nemestrina	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
M.mulatta	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
S.labiatus	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	--GUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
M.musculus	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UGUGAGCAAUCAUGUGUAGUGCCAAUAU	GGGAAA		
R.norvegicus	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
C.lupus	UGGCC.	GAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
B.taurus	UGGCC.	GACUUUUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAAA		
L.africana	UGGCCG	GAUUCUGGCACUAGC.	ACAUUUUUGGCUU.	GUGUCUCUCC.	GC.	UCUGAGCAAUCAUGUGCAGUGCCAAUAU	GGGAGA		
M.domestica	UGGCC.	CGUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	CUGUCUCUCU.	GC.	UCUGAGCAAUCAUGUGUAGUGCCAAUAU	GGGAAA	teleost fish	
O.anatinus	UGGCC.	UCUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	UUUGUCUCUCU.	GC.	UCUGAGCAAUCAUGUGUAGUGCCAAUAU	GGGAAA		
O.latipes	UGGCC.	CGUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	UUUUUUUUUGGUG.	UCUGAGCAAUCAUGUGUGGUGCCAAUAU	GGGACA			
G.aculeatus	UGGCC.	CGUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	UUUUUUUUUGGUG.	UCUGAGCAAUCAUGUGUGGUGCCAAUAU	GGGACA			
T.rubripes	UUGCC.	CAUUUUUGGCACUAGC.	ACAUUUUUGGUUU.	CGUUUUUUUGGUG.	UUUGAGCAAUCAUGUGUAGUGCCAAUAU	GGGACA			
T.rubripes2	UCGCC.	CAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	CUGUAUGUAU.	AC.	UUUGAGCAAUAUAGUGUAGUGCCAAUAU	AGGAGA		
T.nigroviridis	UCGCC.	CAUUUUUGGCACUAGC.	ACAUUUUUGGUUU.	C-GUUUUUG-	GC.	UGUGAGCAAUCGUGUGCAGUGCCAAUCU	AGGACA		
T.nigroviridis2	UUGCC.	CAUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	CUGUAUAUAU.	AC.	UUUGAGCAAUAUAGUGUAGUGCCAAUAU	AGGAGA		
D.rerio	UUGCC.	UGUUUUUGGCACUAGC.	ACAUUUUUGGCUU.	UUUUUAUAUAU.	AC.	CUUGAGCAAUAUAGUGUAGUGCCAAUAU	GGGACA		

Are they real?

Mostly....

...but 1 mis-alignment in 1 million

...yields 100s of errors in large
genomic data sets



Trust, but verify

Understand your algorithms and
how they can go wrong

Just because it runs doesn't mean
it's free of bugs



Photograph of praying mantis by Annika Salzberg