Under the Hood of Alignment Algorithms for NGS Researchers

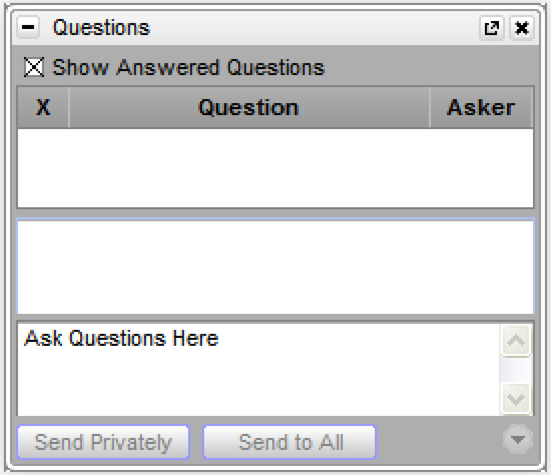
April 16, 2014

Gabe Rudy
VP of Product Development

Golden Helix

Questions during the presentation

Use the Questions pane in your GoToWebinar window



Questions

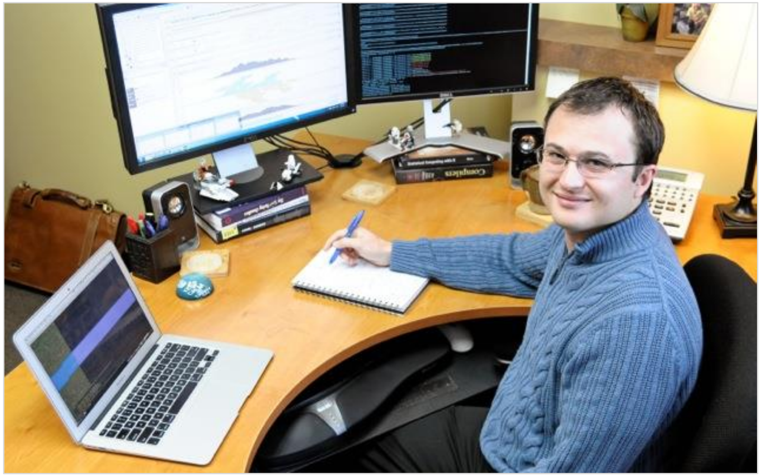
☒ Show Answered Questions

X	Question	Asker

Ask Questions Here

My Background

- **Golden Helix**
  - Founded in 1998
  - Genetic association software
  - Analytic services
  - Hundreds of users worldwide
  - Over 800 customer citations in scientific journals
- **Products I Build with My Team**
  - **SNP & Variation Suite (SVS)**
    - SNP, CNV, NGS tertiary analysis
    - Import and deal with all flavors of upstream data
  - **GenomeBrowse**
    - Visualization of everything with genomic coordinates. All standardized file formats.
  - **RNA-Seq Pipeline**
    - Expression profiling bioinformatics



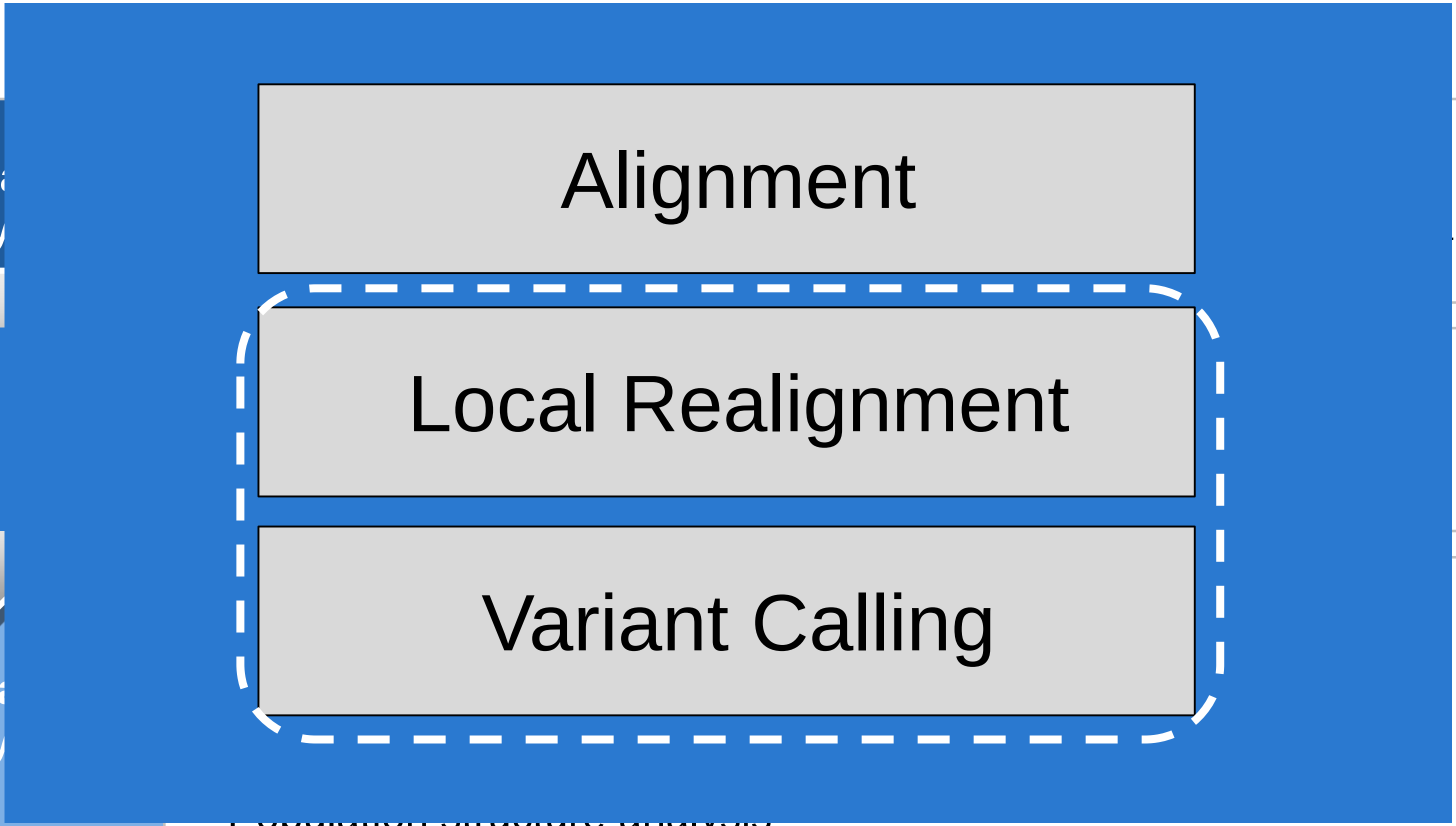
Agenda

1. Alignment 101
2. A Brief History of Time
3. Know Your CIGAR
4. It’s All about the Variants
5. Q&A

Analytics and Sequencing

Prim
Analy

l data



Alignment

Local Realignment

Variant Calling

Terti
Analy

“Sense Making”

- Population structure analysis
- Genome browser-driven exploratory analysis
- Phenotypic association testing

Golden Helix GenomeBrowse®

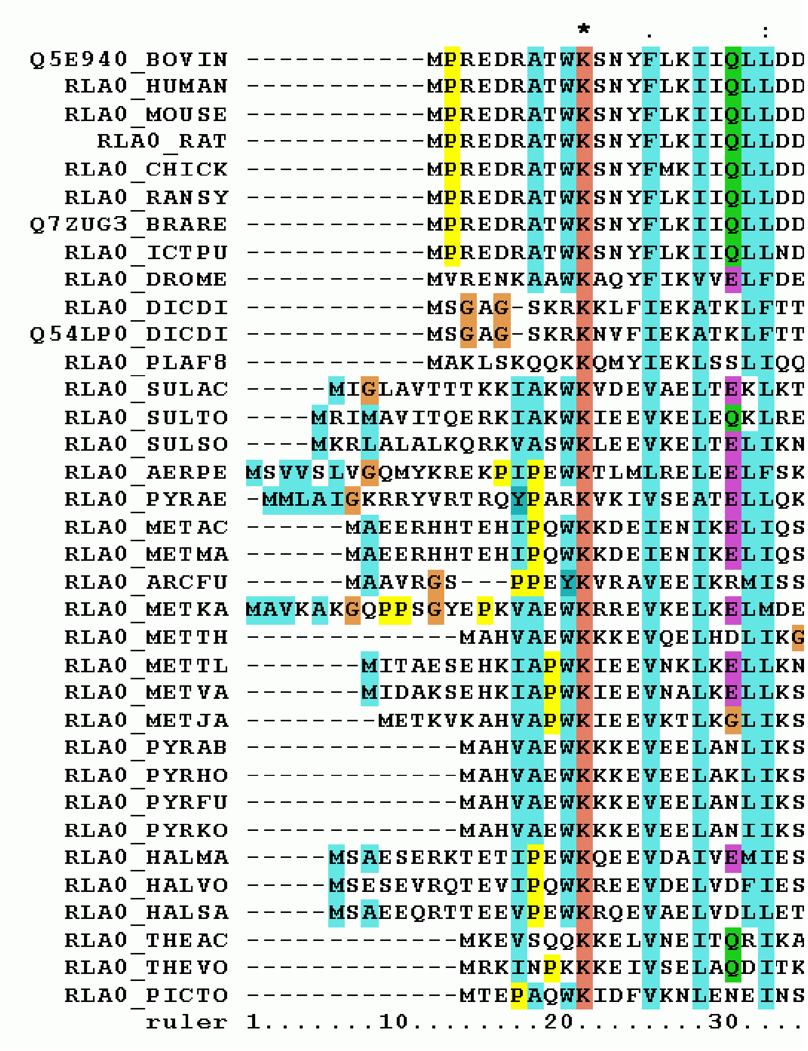
Alignment 101

Agenda

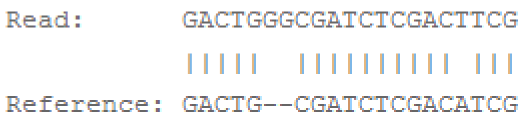
1. Alignment 101
2. **A Brief History of Time**
3. Know Your CIGAR
4. It’s All about the Variants
5. Q&A

Types of Alignment

- Multiple Sequence Alignment
- Phylogenic analysis
- Database Search (BLAST)
- **Pairwise Alignment**
  - Local vs Global
  - Dynamic Programing vs Word Based



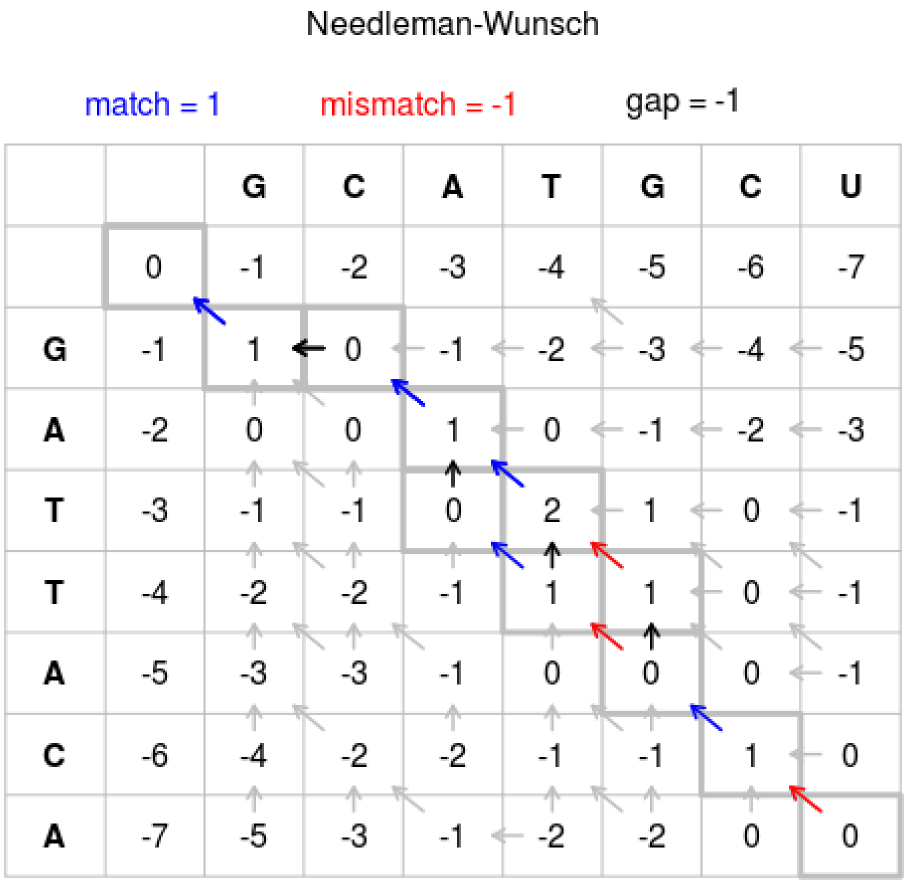
```
                                 *   .    :
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDD
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDD
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDD
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDD
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDD
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDD
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDD
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLND
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDE
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTT
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTT
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQ
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKT
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLRE
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKN
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSK
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQK
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQS
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQS
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISS
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDE
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKG
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKN
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKS
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKS
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKS
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKS
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKS
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKS
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIES
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIES
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLET
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKA
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITK
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINS
       ruler 1........10........20........30....
```



```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

Pairwise Alignment with Dynamic Programming

- **Needleman-Wunsch (1970)**
  - Dynamic programming optimal alignment of two sequences globally
  - O(n*m) space and time
  - Weighting function critical to define
    - Penalty matrix for mismatches
    - Penalty for gaps open and extensions (insertions, deletions)
- **Smith-Waterman (1981)**
  - NW based piecewise (local) alignment
  - Many optimizations, still O(n*m)



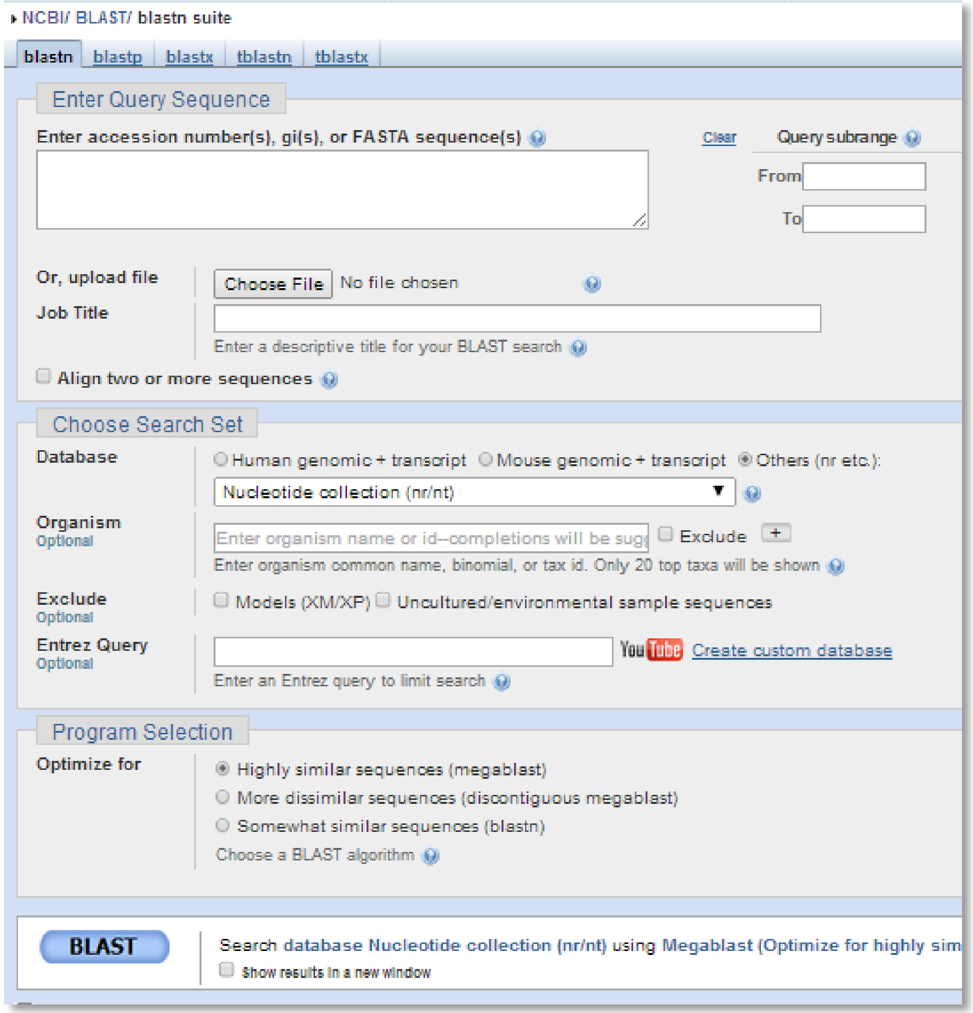
Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Pairwise Alignment with Word Methods

- Heuristics methods based on finding matching k-tuples
- Significantly more efficient when the majority of sequence will not match (database search, reference-based alignment)
- FASTA (1985), BLAST (1990) designed for large DNA/Protein searches
- New class of problem emerged with high-throughput sequencing



NCBI/ BLAST/ blastn suite

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt) ▼

Organism
Optional

Enter organism name or id--completions will be sugg □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional

YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for

● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

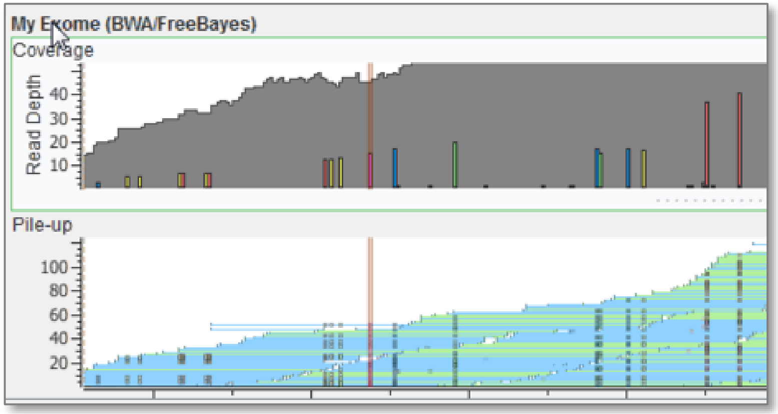
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly sim

□ Show results in a new window

Alignment Versus Assembly

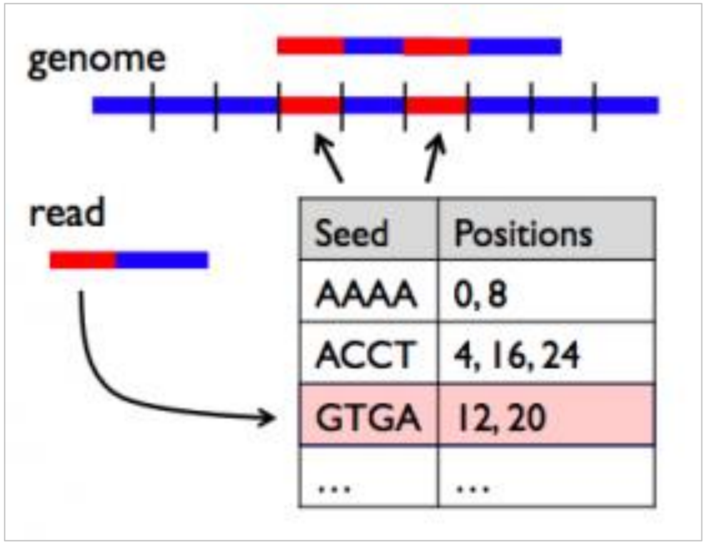
- **Assembly**
  - Orders of magnitude slower and memory intensive than alignment
  - Potentially compare every read with each other $O(n^2)$
  - Steps:
    - Merge overlapping reads into a de Bruijn graph
    - Simply the graph iteratively, construct contigs
    - Detangle with orthogonal tech (long reads, mates, optical mapping)
  - “Draft” genomes from short reads, have ~1kb sized contigs

- **Alignment**
  - Requires a finished genome for your species (draft genomes possible, but of limited utility)
  - Precompute an index of the reference genome (can be costly as you do it once)
  - Each short read uses the index to find its best placements (potentially multiple)



Hash Based Alignment Algorithms

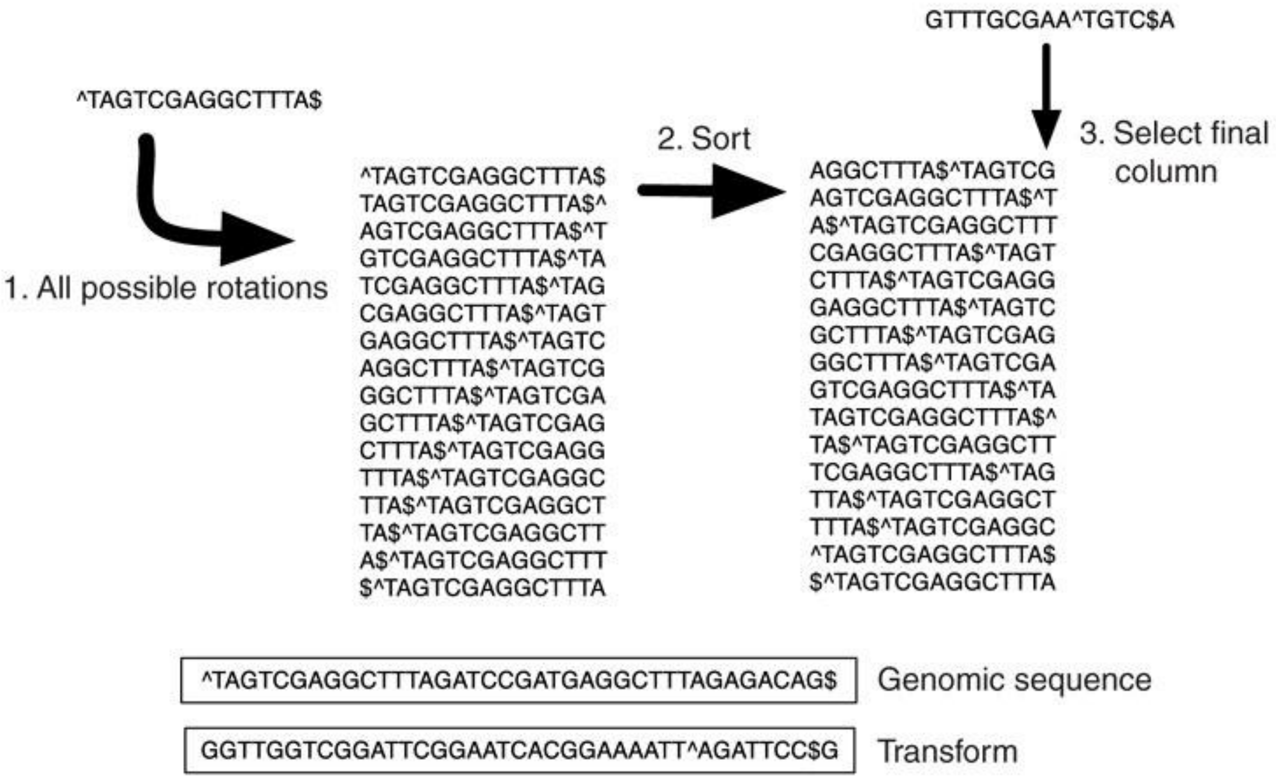
- **Hash based**
  - Pick k-mer size, build lookup of every k-mer in the reference to its positions
  - ~16GB of RAM required for hg19
  - **Seed-and-extend strategy**
  - **Popular tools:**
    - BLAST: tunable for different uses
    - MAQ (2008): Heng Li, et al
    - NovaAlign: Slower, but very accurate
    - Isaac (2013): High mem, but fast
    - MOSAIK (2014): Hash clustering+SW



Seed	Positions
AAAA	0, 8
ACCT	4, 16, 24
GTGA	12, 20
...	...

Burrows-Wheeler Transform

- BWT is a reversible permutation of characters that can be used for fast substring-searching when used with an index



^TAGTCGAGGCTTTA$

1. All possible rotations

```
^TAGTCGAGGCTTTA$
TAGTCGAGGCTTTA$^
AGTCGAGGCTTTA$^T
GTCGAGGCTTTA$^TA
TCGAGGCTTTA$^TAG
CGAGGCTTTA$^TAGT
GAGGCTTTA$^TAGTC
AGGCTTTA$^TAGTCG
GGCTTTA$^TAGTCGA
GCTTTA$^TAGTCGAG
CTTTA$^TAGTCGAGG
TTTA$^TAGTCGAGGC
TTA$^TAGTCGAGGCT
TA$^TAGTCGAGGCTT
A$^TAGTCGAGGCTTT
$^TAGTCGAGGCTTTA
```

2. Sort

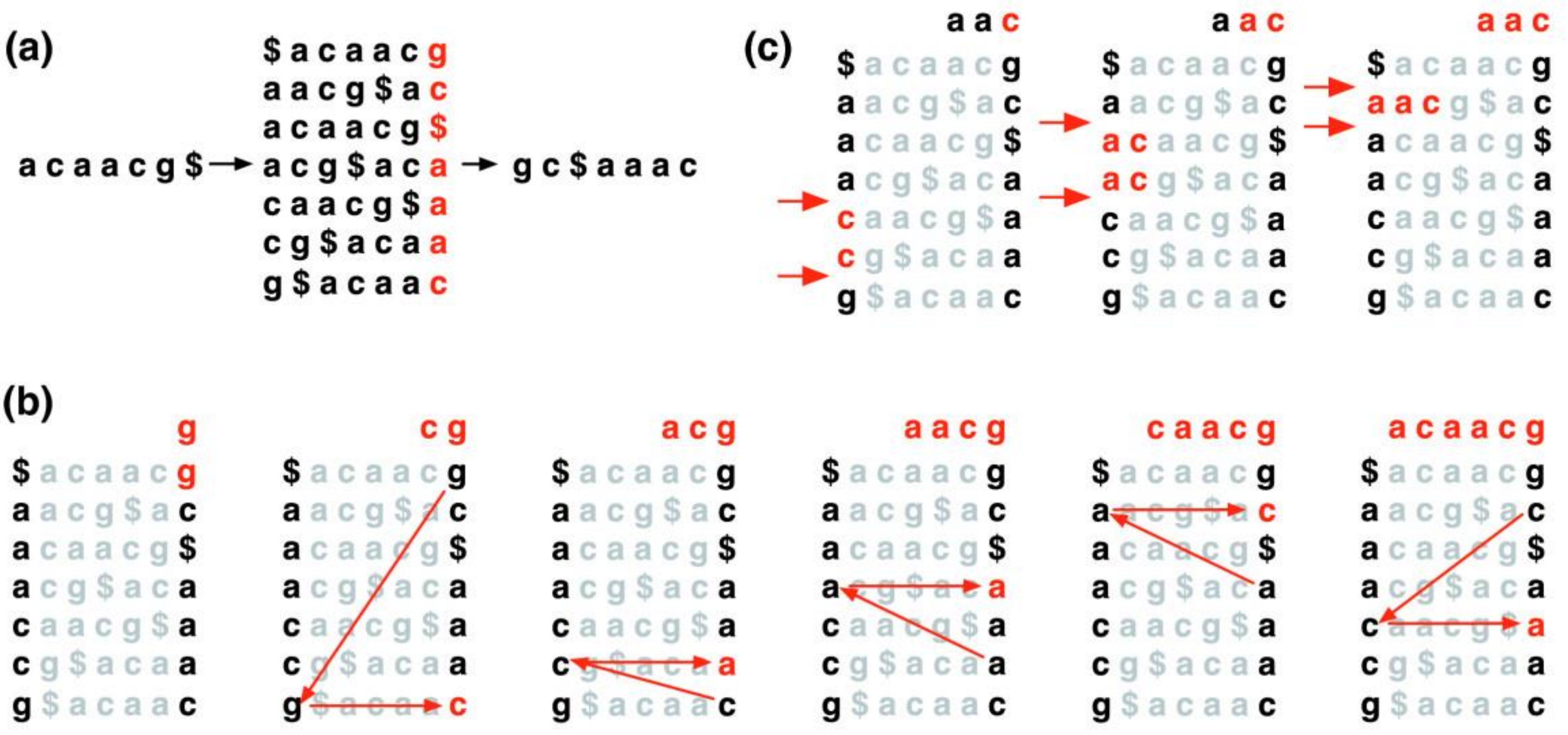
```
AGGCTTTA$^TAGTCG
AGTCGAGGCTTTA$^T
A$^TAGTCGAGGCTTT
CGAGGCTTTA$^TAGT
CTTTA$^TAGTCGAGG
GAGGCTTTA$^TAGTC
GCTTTA$^TAGTCGAG
GGCTTTA$^TAGTCGA
GTCGAGGCTTTA$^TA
TAGTCGAGGCTTTA$^
TA$^TAGTCGAGGCTT
TCGAGGCTTTA$^TAG
TTA$^TAGTCGAGGCT
TTTA$^TAGTCGAGGC
^TAGTCGAGGCTTTA$
$^TAGTCGAGGCTTTA
```

3. Select final column

GTTTGCGAA^TGTC$A

^TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG$	Genomic sequence
GGTTGGTCGGATTCGGAATCACGGAAAATT^AGATTCC$G	Transform

BWT



acaacg$ → $acaacg, aacg$ac, acaacg$, acg$aca, caacga, cgacaa, g$acaac → gc$aaac

g

cg

acg

aacg

caacg

acaacg

$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac

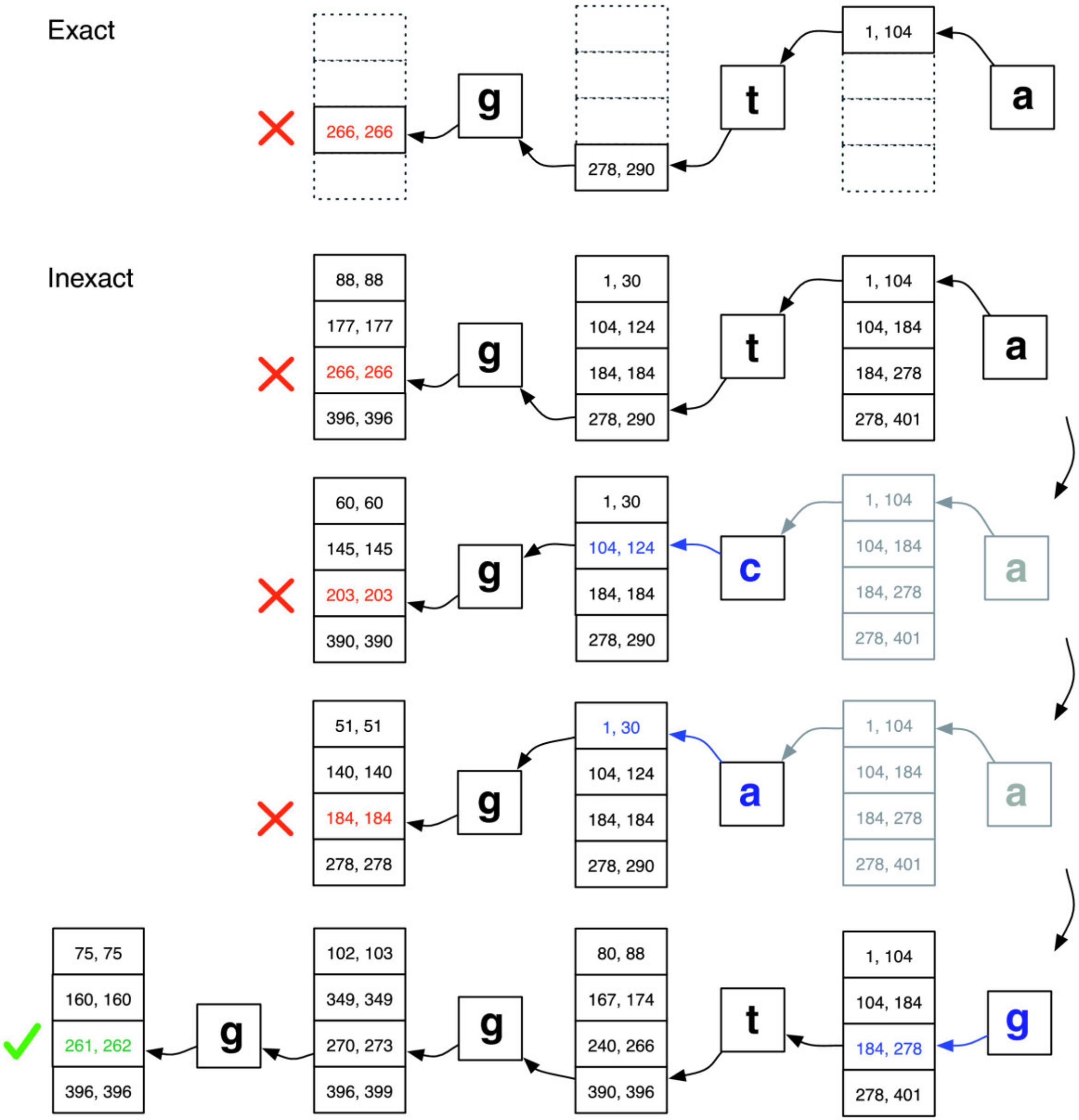
aac

aac

aac

$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac

Backtracking – query ‘ggta’ with 1 mismatch



Exact

✗	266, 266	g	278, 290	t	1, 104	a

Inexact

	88, 88		1, 30		1, 104	
	177, 177	g	104, 124	t	104, 184	a
✗	266, 266		184, 184		184, 278	
	396, 396		278, 290		278, 401	

	60, 60		1, 30		1, 104	
	145, 145	g	104, 124	c	104, 184	a
✗	203, 203		184, 184		184, 278	
	390, 390		278, 290		278, 401	

	51, 51		1, 30		1, 104	
	140, 140	g	104, 124	a	104, 184	a
✗	184, 184		184, 184		184, 278	
	278, 278		278, 290		278, 401	

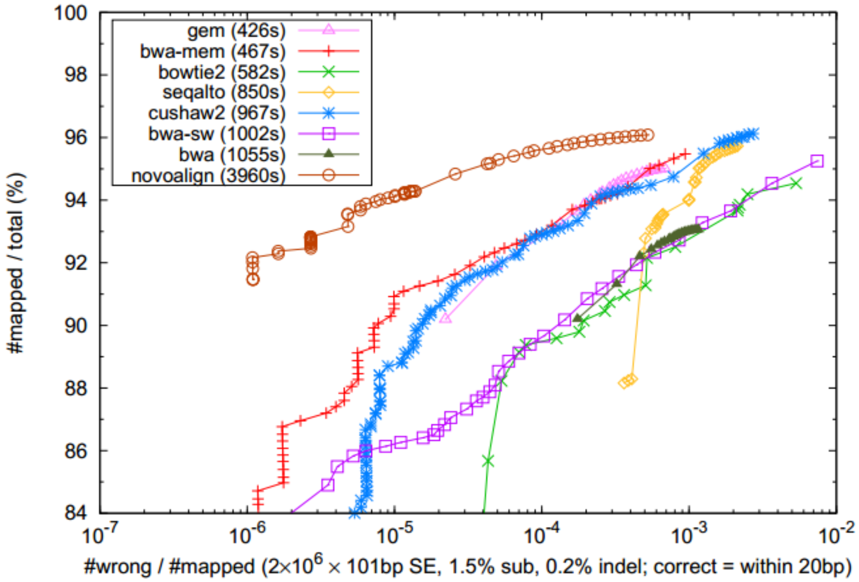
	75, 75		102, 103		80, 88		1, 104	
	160, 160	g	349, 349	g	167, 174	t	104, 184	g
✓	261, 262		270, 273		240, 266		184, 278	
	396, 396		396, 399		390, 396		278, 401	

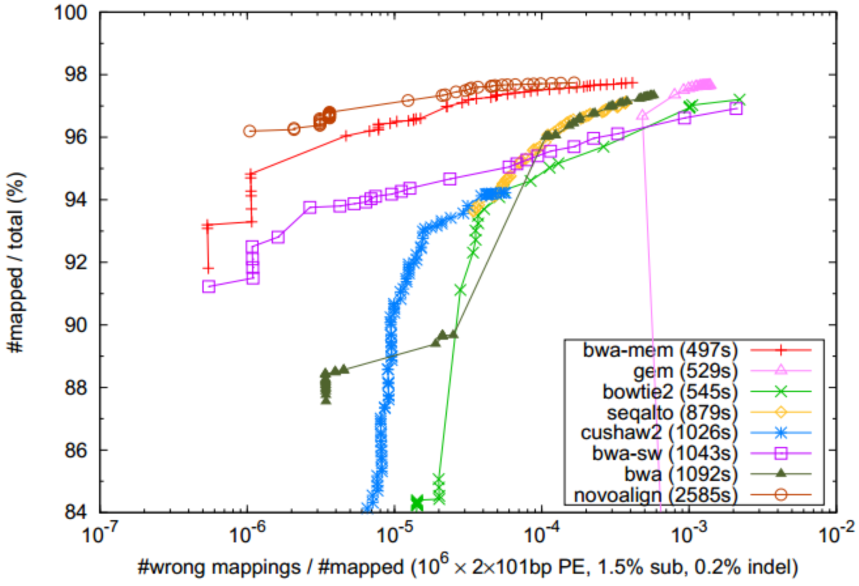
BWT Based Algorithms

- **Compute a FM index of the reference**
  - Requires only ~1.5GB to hold in RAM of hg19
- **Requires a back-tracking algorithm to account for mismatches and gaps**
- **Designed for speed**
  - BowTie, BWA, SOAP2 (2009)
  - Order of magnitude less RAM and Time
- **More recent algorithms are often a hybrid:**
  - BWA-SW (2010)
  - Bowtie2 (2012)
  - BWA-MEM (2013)

BWA and Friends

- **BWA (backtrack) - 2009**
  - Very mature, handles short reads up to 100bp
- **BWA-SW (Smith Waterman) - 2010**
- **BWA-MEM (Max Exact Matches) - 2013**
  - >70bp read length recommended, but up to 1Mbp
  - Seed and extend with SW
  - Allowable error rate adjust with sequence length
  - Finds larger gaps
  - Faster! Generally supersedes BWA-SW





Algorithm Comparison

	BWA	BWA-SW	BWA-MEM	Bowtie	Bowtie2	NovaAlign	MOSAIK	Isaac	Tmap
Affiliation	Heng Li			U of Maryland		Novacraft	Boston College	Illumina	Ion Torrent
First Published	2009	2010	2013	2009	2012	-	2014	2013	-
Read Length	<100	70bp-1Mbp		<100	>50				
Gapped Alignments				No					
Trimming				No					
Error Rates Allowed	Low	High	Med	Low	Med	Med	High	Low	Med
Chim Reads	No	Yes	**Yes**	No	Opt	Opt	**Yes**	No	No
Mem Usage	Med	Med	Med	Low	Low	Low	Med	**High**	Med
Speed	Med	Med	Fast	Fast	Fast	**Slow**	Fast	Fast	Fast

Agenda

1 Alignment 101

2 A Brief History of Time

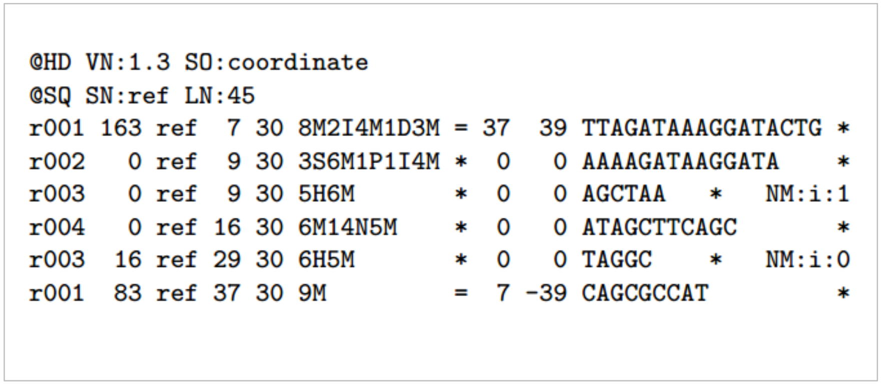
3 **Know Your CIGAR**

4 It’s All about the Variants

5 Q&A

SAM/BAM

- **Spec defined by bwa/samtools author** Heng Li, aka Li H, aka lh3.
- **SAM is text version** (easy for any program to output)
- **BAM is binary/compressed version** with indexing support
- Alignment encoded in CIGAR code of matches, insertions, deletions, gaps and clipping
- **Can have any custom flags set** by alignment tool (mix of standard and custom two-letter tags)



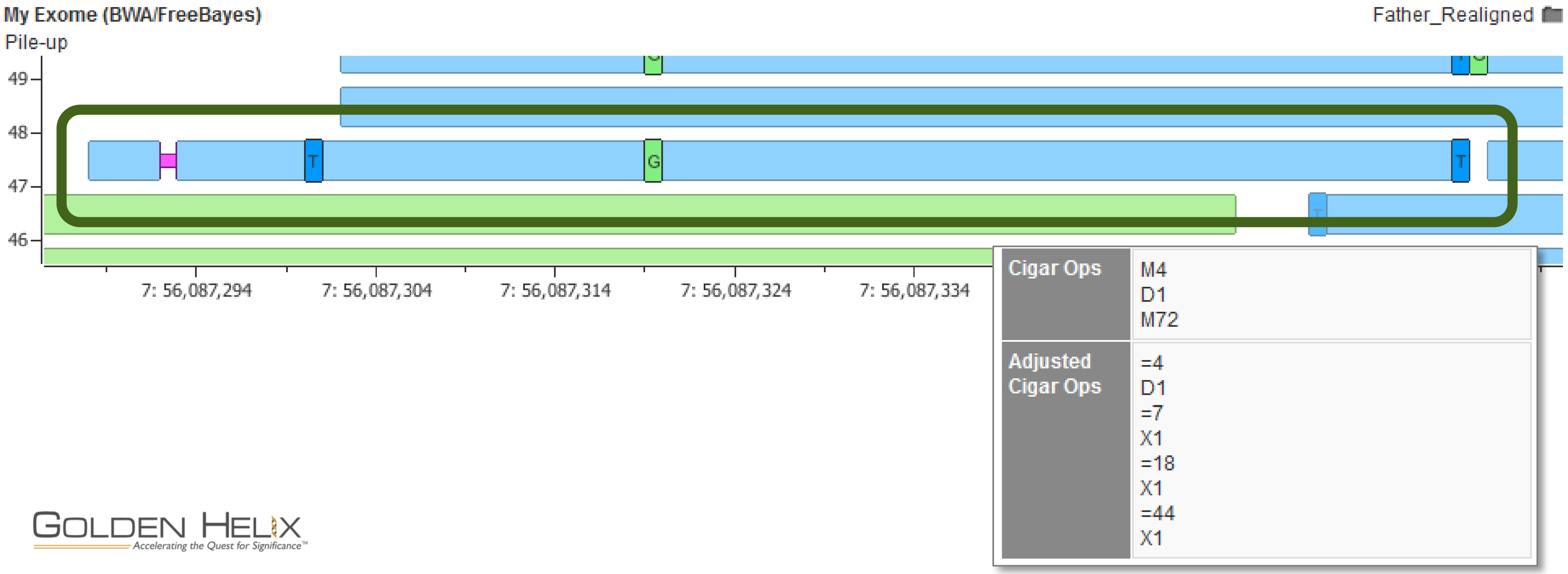
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Key Fields

- Chr, position
- **Mapping quality**
- **CIGAR**
- Name/position of mate
- Total template length
- Sequence
- Per-Base Quality Scores

CIGAR String

Op	Description	Used
M	alignment match (can be sequence match or mismatch)	by default
I	insertion to the reference	InDel
D	deletion from the reference	InDel
N	skipped region from the reference	spanning intron
S	soft clipping (not-aligned)	per-base quality drops to threshold to trim read
H	hard clipping (not-present in reference)	chimeric reads, breakpoints, end of seq
P	padding (silent deletion from padded reference)	multiple sequence alignments (not common)
=	sequence match	when compared to ref
X	sequence mismatch	when compared to ref



Cigar Ops	M4 D1 M72
Adjusted Cigar Ops	=4 D1 =7 X1 =18 X1 =44 X1

Golden Helix GenomeBrowse®

Different Alignment Outcomes

Agenda

1 Alignment 101

2 A Brief History of Time

3 Know Your CIGAR

4 It’s All about the Variants

5 Q&A

Mapping and Calling Variants on the Human Genome

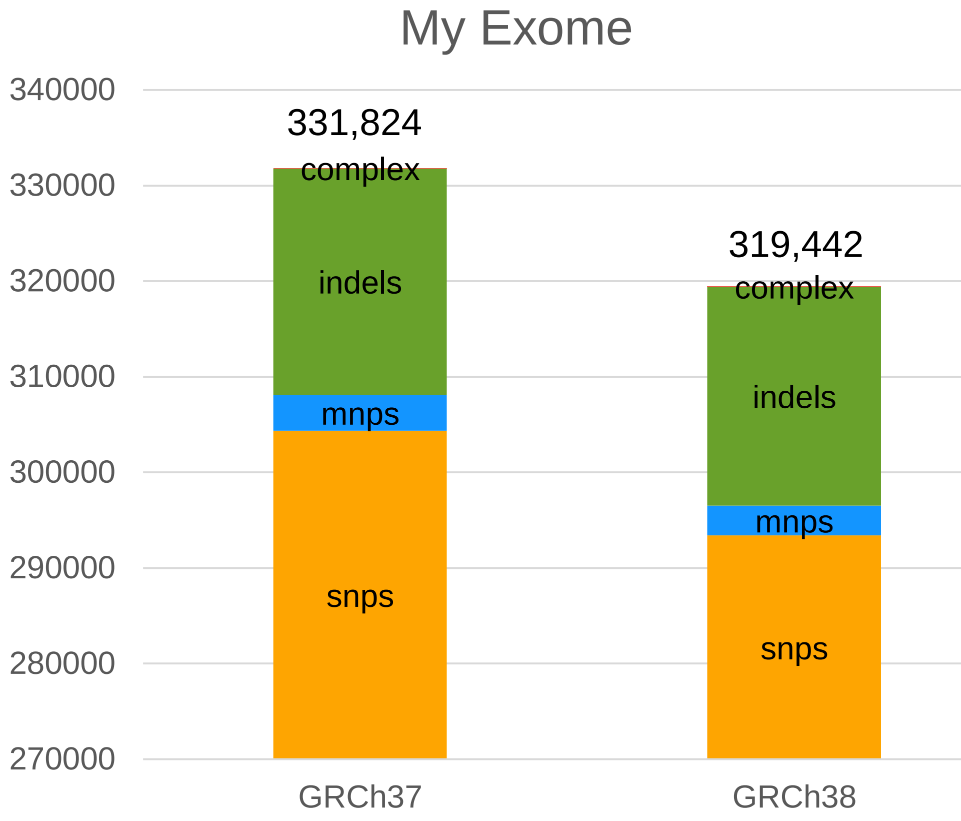
- **Classes of Confounders:**
  - Issues with the **Reference Assembly**:
    - Sequence under-represented (exact match not in human reference, so get poor match)
    - Tiling issues creating artificial splices
  - **Repeated** regions and Low **Mapping Quality** Regions:
    - Over 50% of the genome is repetitive
    - Low sequence “complexity” or “information density” means short reads cannot uniquely map. “Mappability”
  - Interference with larger classes of variation: **Structural Variation**
    - Calling genotypes of SNPs/short-InDels in a deletion
    - Inversion/Translocation/CNV break points
  - Disagreement in **Representing Complex Variants**

Responsibility of the Alignment Algorithm?

- **Placing reads in the right part of the genome**
- **Providing accurate mapping quality scores**
  - Often need to empirically train an aligner to produce Gaussian spectrum of scores
- **Providing the best data to variant callers**
- **What Variant Callers expect?**
  - Multi-mapped read status (often filtered out by MQ=0)
  - Mate-pair mapping information
  - Just “localizing” the read?
  - Consistently described gapped alignments?

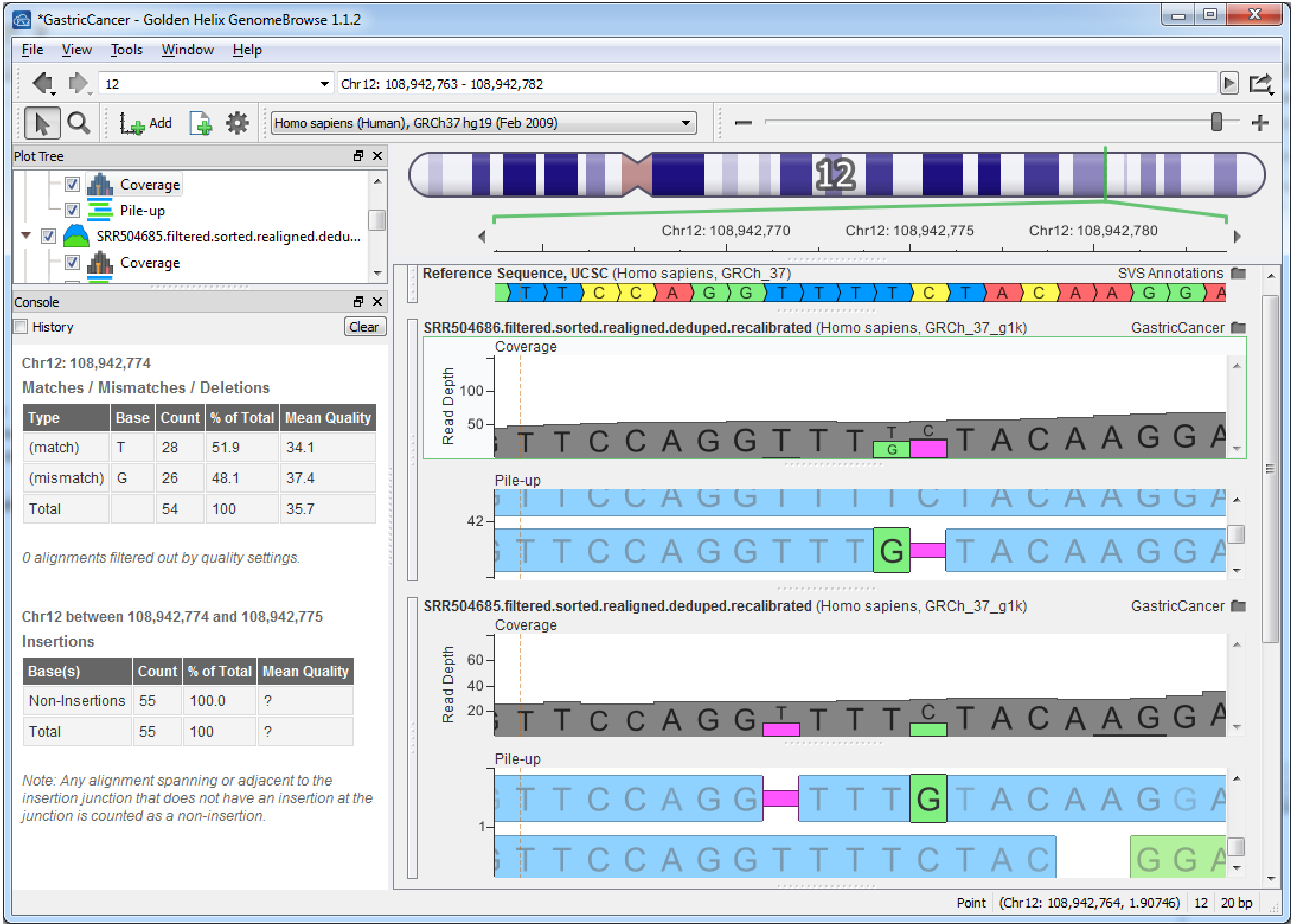
GRCh38 – Here Now, but still Waiting

- **A better human reference**
  - Revised Cambridge Reference Sequence (rCRS) MT
  - Has centromere models
  - ~2000 incorrect alleles fixed
  - ~100 assembly gaps updated
- **No Gene Annotations**
  - RefSeqGene - Feb 2014
  - Ensembl Q4 2014
- **No Variant Annotations**
  - Re-align 1000 Genomes and NHLBI 6500?
  - dbSNP?



	GRCh37	GRCh38
Ts/Tv	2.06558	2.10171

InDel Alignment: Watch for ambiguities


*GastricCancer - Golden Helix GenomeBrowse 1.1.2
File View Tools Window Help
12
Chr12: 108,942,763 - 108,942,782
Add
Homo sapiens (Human), GRCh37 hg19 (Feb 2009)
Plot Tree
Coverage
Pile-up
SRR504685.filtered.sorted.realigned.dedu...
Coverage
Console
History
Clear
12
Chr12: 108,942,770
Chr12: 108,942,775
Chr12: 108,942,780
Reference Sequence, UCSC (Homo sapiens, GRCh_37)
SVS Annotations
T T C C A G G T T T T C T A C A A G G A
SRR504686.filtered.sorted.realigned.deduped.recalibrated (Homo sapiens, GRCh_37_g1k)
GastricCancer
Coverage
Read Depth
100
50
T T C C A G G T T T T G C T A C A A G G A
Pile-up
42
T T C C A G G T T T G T A C A A G G A
SRR504685.filtered.sorted.realigned.deduped.recalibrated (Homo sapiens, GRCh_37_g1k)
GastricCancer
Coverage
Read Depth
60
40
20
T T C C A G G T T T T C T A C A A G G A
Pile-up
1
T T C C A G G T T T G T A C A A G G A
T T C C A G G T T T T C T A C G G A
Point
(Chr12: 108,942,764, 1.90746)
12
20 bp

Chr12: 108,942,774

Matches / Mismatches / Deletions

Type	Base	Count	% of Total	Mean Quality
(match)	T	28	51.9	34.1
(mismatch)	G	26	48.1	37.4
Total		54	100	35.7

0 alignments filtered out by quality settings.

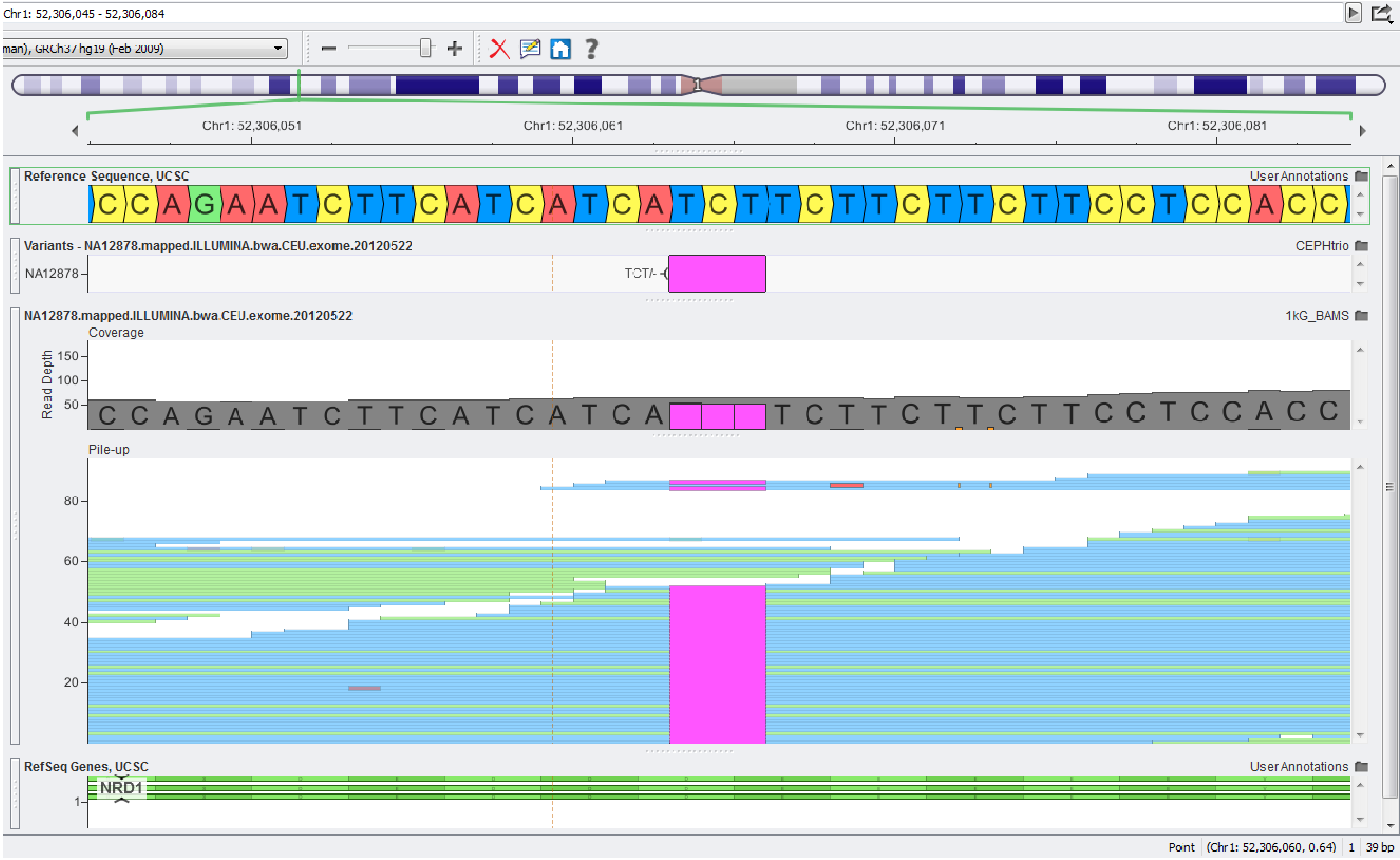
Chr12 between 108,942,774 and 108,942,775

Insertions

Base(s)	Count	% of Total	Mean Quality
Non-Insertions	55	100.0	?
Total	55	100	?

Note: Any alignment spanning or adjacent to the insertion junction that does not have an insertion at the junction is counted as a non-insertion.

InDel Alignment: Watch for repeats and read ends



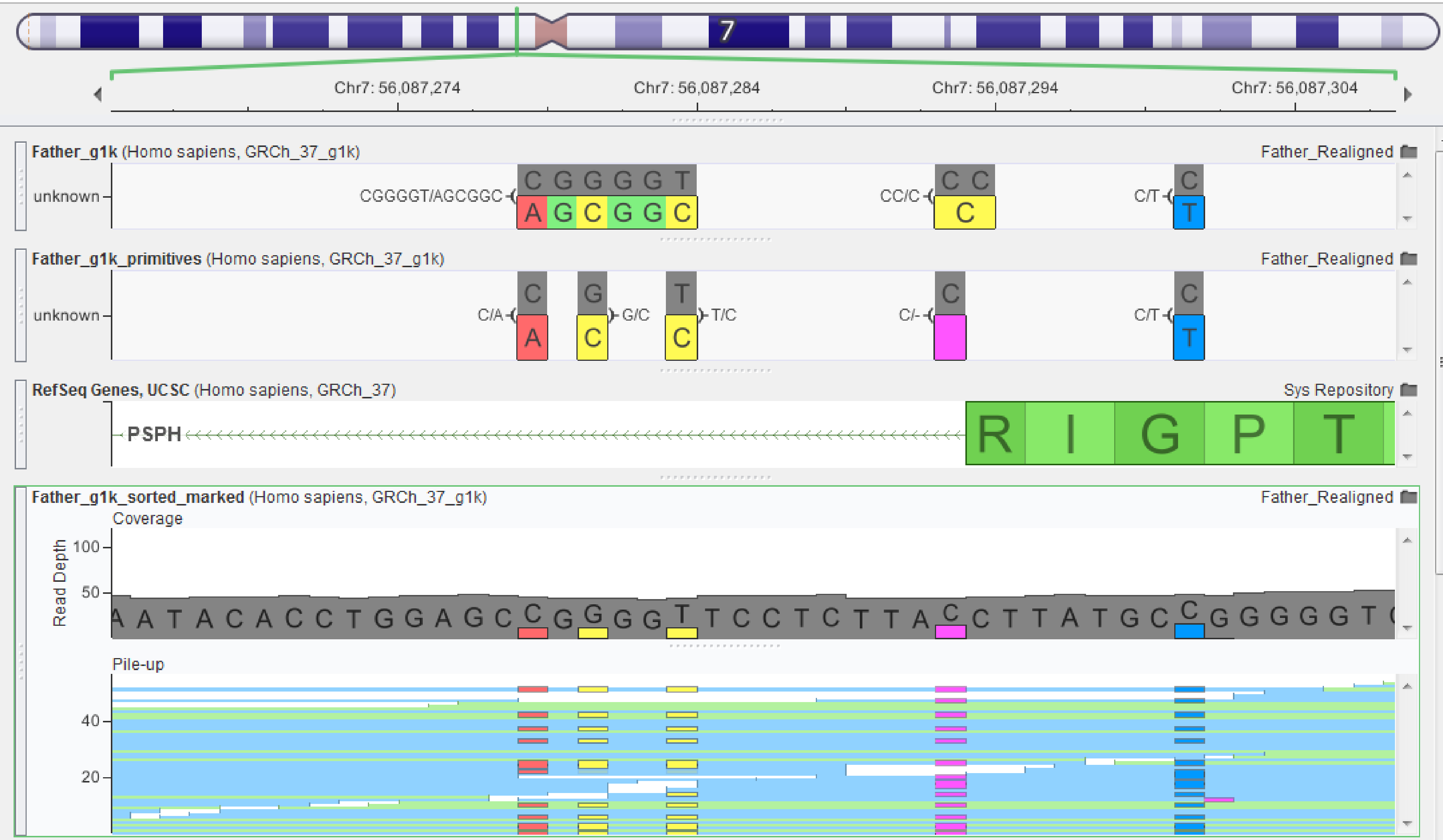
Chr1: 52,306,045 - 52,306,084

man), GRCh37 hg19 (Feb 2009)

Point | (Chr1: 52,306,060, 0.64) | 1 | 39 bp

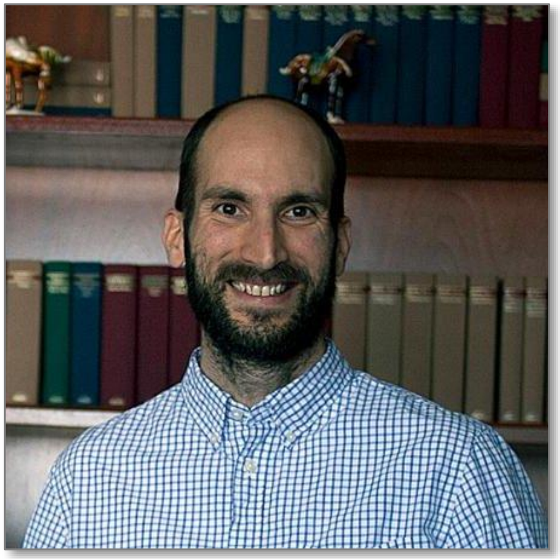
Chr1: 52,306,051
Chr1: 52,306,061
Chr1: 52,306,071
Chr1: 52,306,081
Reference Sequence, UCSC
User Annotations
C C A G A A T C T T C A T C A T C A T C T T C T T C T T C T T C C T C C A C C
Variants - NA12878.mapped.ILLUMINA.bwa.CEU.exome.20120522
CEPHtrio
NA12878
TCT/-
NA12878.mapped.ILLUMINA.bwa.CEU.exome.20120522
1kG_BAMS
Coverage
Read Depth
150
100
50
C C A G A A T C T T C A T C A T C A T C T T C T T C T T C C T C C A C C
Pile-up
80
60
40
20
RefSeq Genes, UCSC
User Annotations
1
NRD1

MNP vs Allelic Primitives


7
Chr7: 56,087,274
Chr7: 56,087,284
Chr7: 56,087,294
Chr7: 56,087,304
Father_g1k (Homo sapiens, GRCh_37_g1k)
Father_Realigned
unknown
CGGGGT/AGCGGC
C G G G G T
A G C G G C
CC/C
C C
C
C/T
C
T
Father_g1k_primitives (Homo sapiens, GRCh_37_g1k)
Father_Realigned
unknown
C/A
C
A
G
C
G/C
T
C
T/C
C/-
C
C/T
C
T
RefSeq Genes, UCSC (Homo sapiens, GRCh_37)
Sys Repository
PSPH
R I G P T
Father_g1k_sorted_marked (Homo sapiens, GRCh_37_g1k)
Father_Realigned
Coverage
Read Depth
100
50
A A T A C A C C T G G A G C C G G G G T T C C T C T T A C C T T A T G C C G G G G G T
Pile-up
40
20

Genome In a Bottle

- NIST Sponsored, Community Effort
- NA12878 cell line, sequenced many platforms, read lengths and sample preps
- Create “ensembl” variant call set
- Create many making regions
  - Regions not able to make consensus call
  - Repeat and low-complexity regions
  - SV in NA12878
- Variants, BED and alignment data available

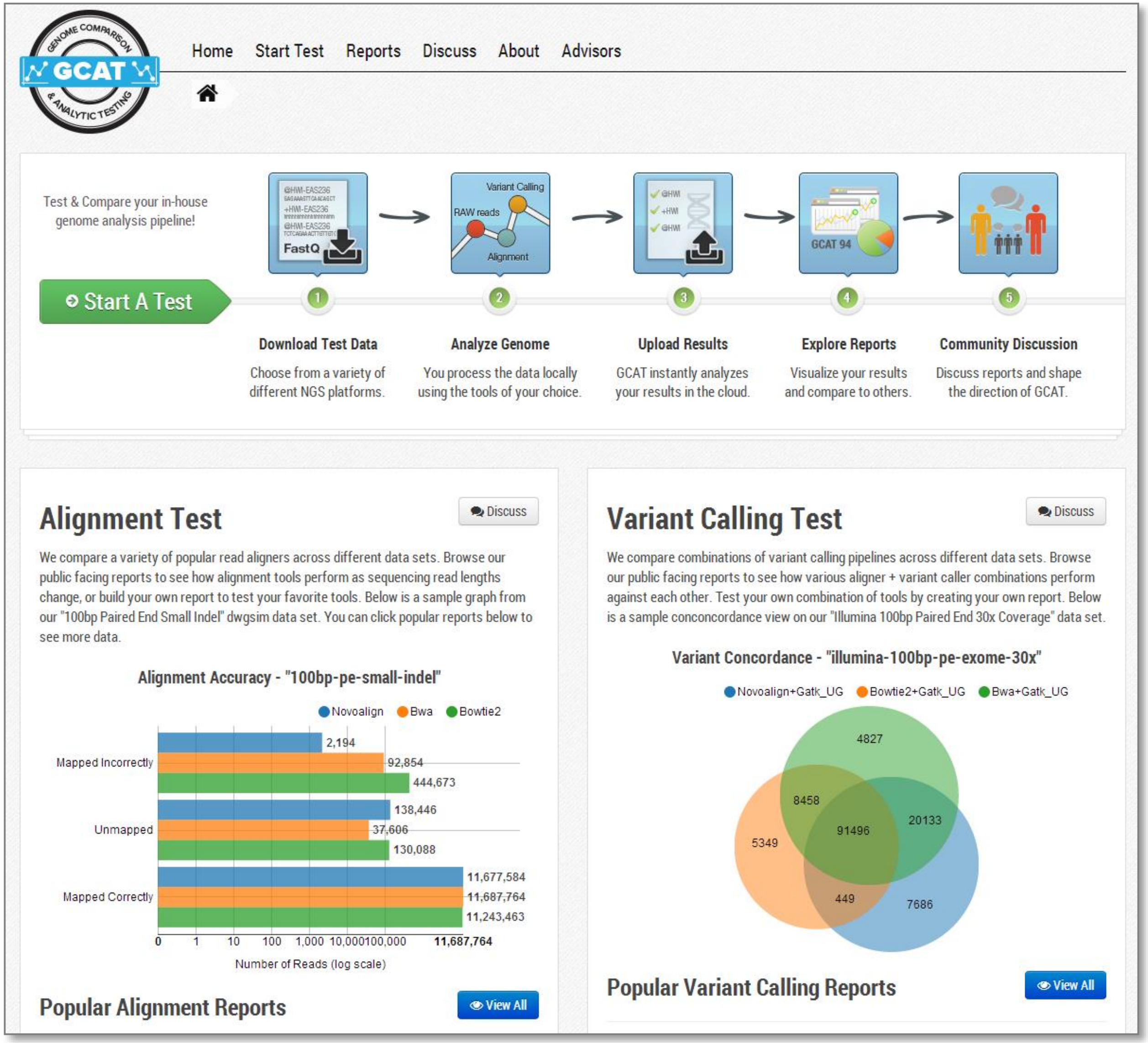


Justin Zook

Genome in a Bottle Consortium

Resources

- **GCAT**
- **Benchmarks**
  - Alignment
  - Variant Calling
  - GIAB Truth Set
  - Various bench samples
- **Interactive filtering**



Home Start Test Reports Discuss About Advisors

Test & Compare your in-house genome analysis pipeline!

Start A Test

1. **Download Test Data**
   Choose from a variety of different NGS platforms.
2. **Analyze Genome**
   You process the data locally using the tools of your choice.
3. **Upload Results**
   GCAT instantly analyzes your results in the cloud.
4. **Explore Reports**
   Visualize your results and compare to others.
5. **Community Discussion**
   Discuss reports and shape the direction of GCAT.

Alignment Test

Discuss

We compare a variety of popular read aligners across different data sets. Browse our public facing reports to see how alignment tools perform as sequencing read lengths change, or build your own report to test your favorite tools. Below is a sample graph from our "100bp Paired End Small Indel" dwgsim data set. You can click popular reports below to see more data.

Alignment Accuracy - "100bp-pe-small-indel"

Popular Alignment Reports

View All

Variant Calling Test

Discuss

We compare combinations of variant calling pipelines across different data sets. Browse our public facing reports to see how various aligner + variant caller combinations perform against each other. Test your own combination of tools by creating your own report. Below is a sample conconcordance view on our "Illumina 100bp Paired End 30x Coverage" data set.

Variant Concordance - "illumina-100bp-pe-exome-30x"

Popular Variant Calling Reports

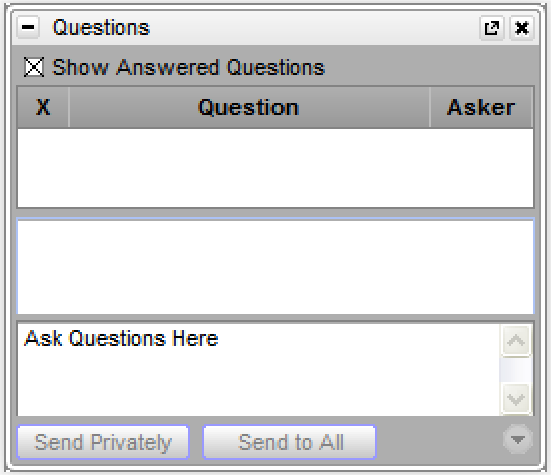
View All

Golden Helix GenomeBrowse®

Some GIAB Examples

Questions?

Use the Questions pane in your GoToWebinar window


Questions
☒ Show Answered Questions
X
Question
Asker
Ask Questions Here
Send Privately
Send to All