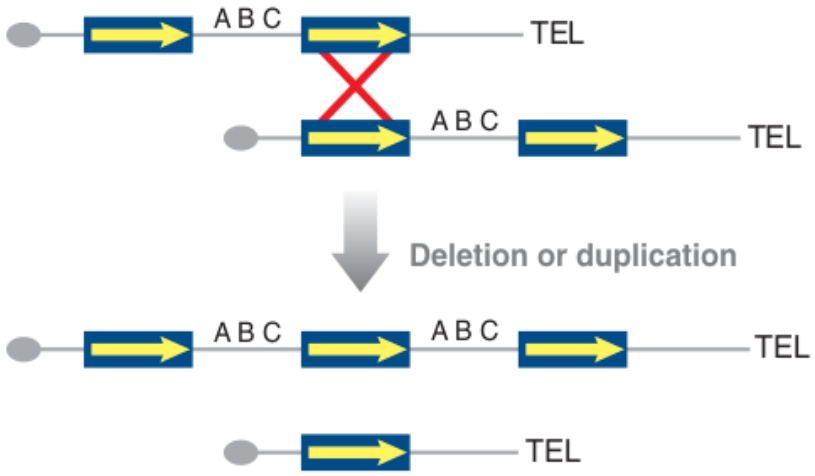
Reconstructing complex regions of genomes using long-read sequencing technology

John Huddleston
Eichler Lab
University of Washington

Segmental duplications mediate structural variation

- Large, high-identity sequence (>1Kbp, >90%)
- Predispose chromosomes to non-allelic homologous recombination (NAHR)

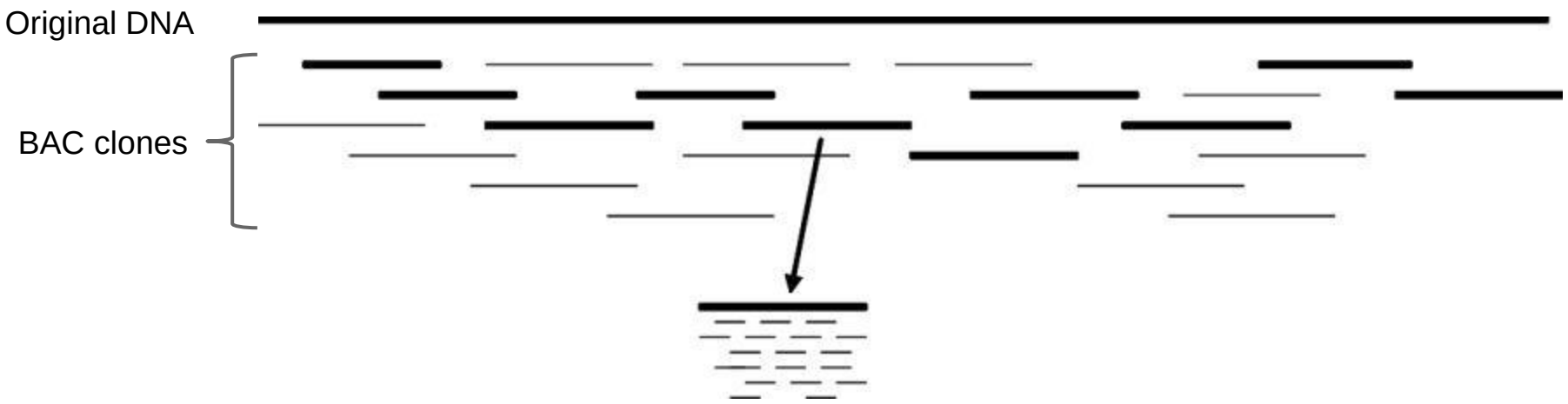


Sharp et al. 2006

Importance of high-quality references

- We need long, high-quality sequences
  - Resolve breakpoints
  - Reconstruct duplication architectures
- build37 is great but still has mistakes
- Most references aren't as finished as build37
  - Draft assemblies from short reads
  - Gaps and collapsed duplications

Bacterial artificial chromosome (BAC) libraries



- DNA inserted into bacterial hosts
- 100-300 Kbp sequence

Image after http://www.cbcb.umd.edu/research/bac-by-bac.jpg

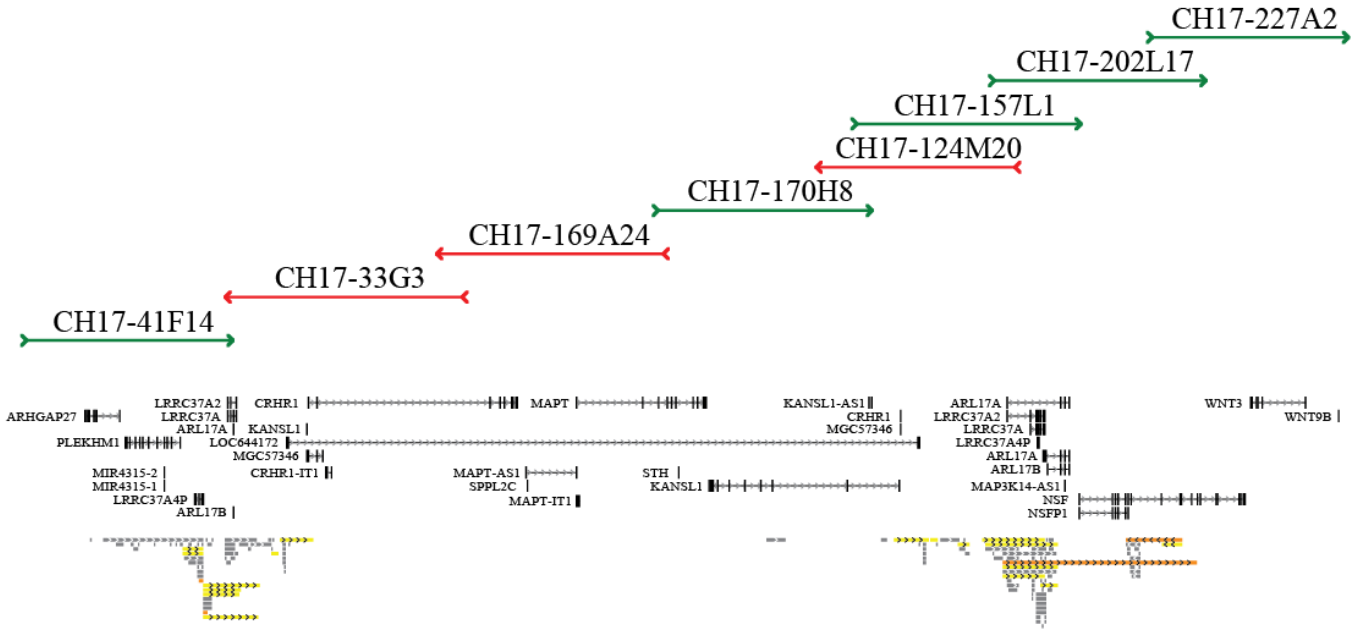
Sanger sequencing

- High-quality, finished Sanger assemblies
  - 1 contig per BAC with no gaps
  - Q50 (99.999% accuracy)
- $3,000-6,000 and 1-2 months per BAC
- Sequencing centers are discontinuing Sanger sequencing

Alternatives for BAC sequencing

- Local Sanger sequencing and assembly
- Illumina sequencing and *de novo* assembly of short reads
- PacBio sequencing and *de novo* assembly of long reads

Test case: CH17 clones from 17q21

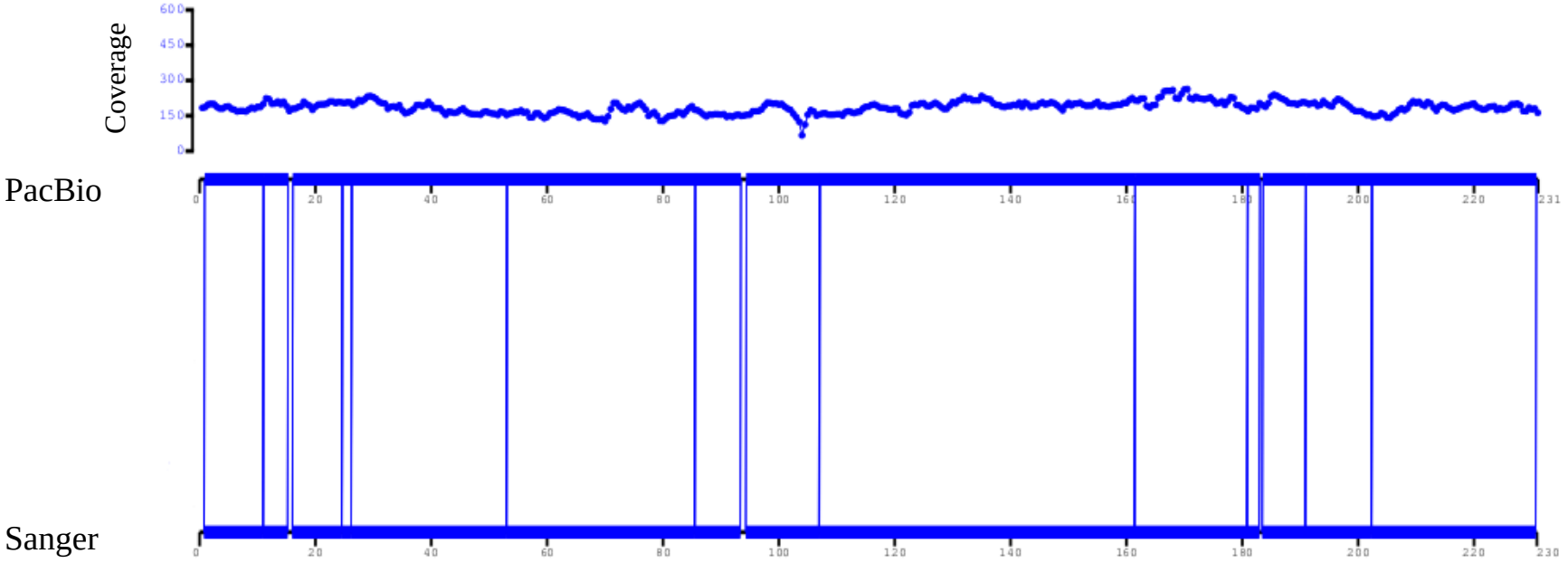


- Illumina reads
- Sanger assemblies
- PacBio reads

PacBio assembly workflow

1. Mask vector sequence in PacBio reads
2. Assemble with HGAP
3. Apply Quiver
4. Screen assembly for bacteria

Alignment of PacBio and Sanger sequences for CH17-157L1



99.994% identity, 12 insertions, 3 deletions, 0 mismatches

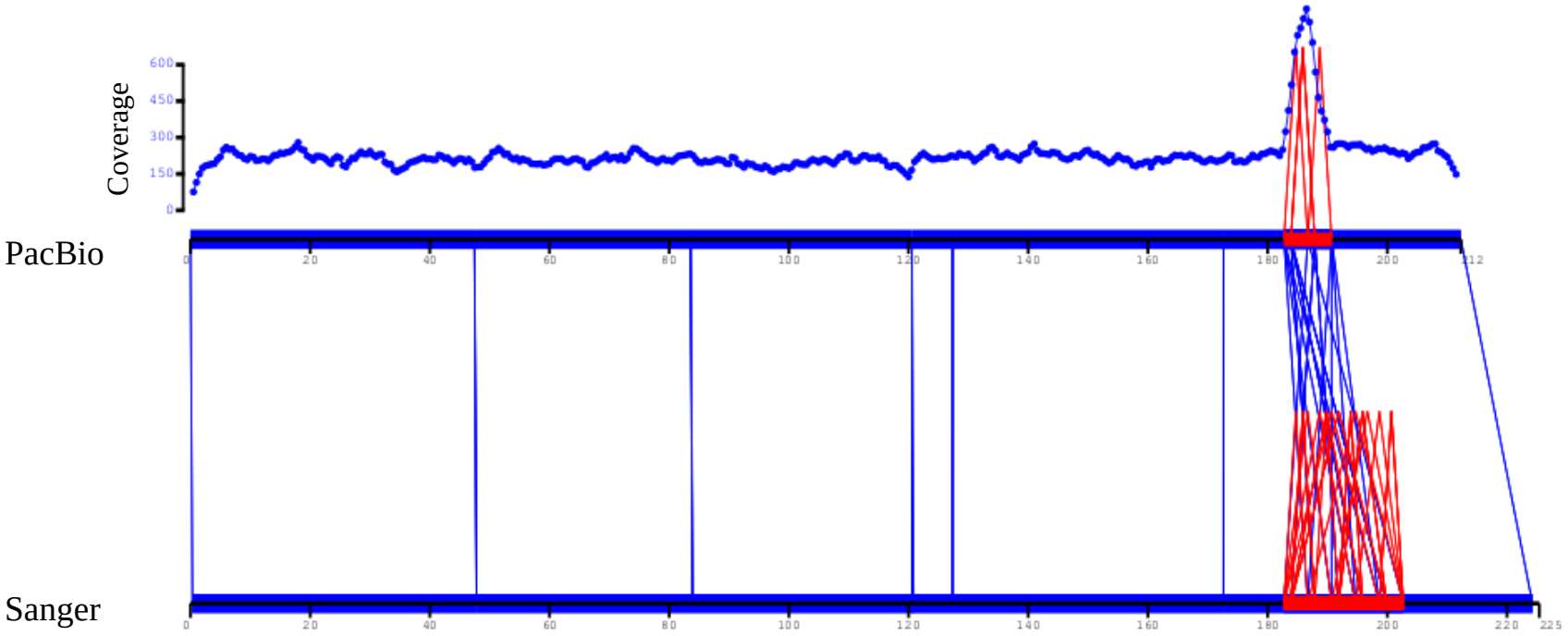
CH17 alignment summary

Clone	Substitutions	PacBio Insertions	Sanger Insertions	Identity
CH17-124M20	0	15	4	0.99991
CH17-157L1	0	12	3	0.99994
CH17-169A24	18	27	6	0.99979
CH17-170H8	0	13	0	0.99994
CH17-202L17	0	2	1	0.99999
CH17-227A2	0	0	3	0.99999
CH17-33G3	2	5	4	0.99996
CH17-41F14	4	349	8,344	0.96303

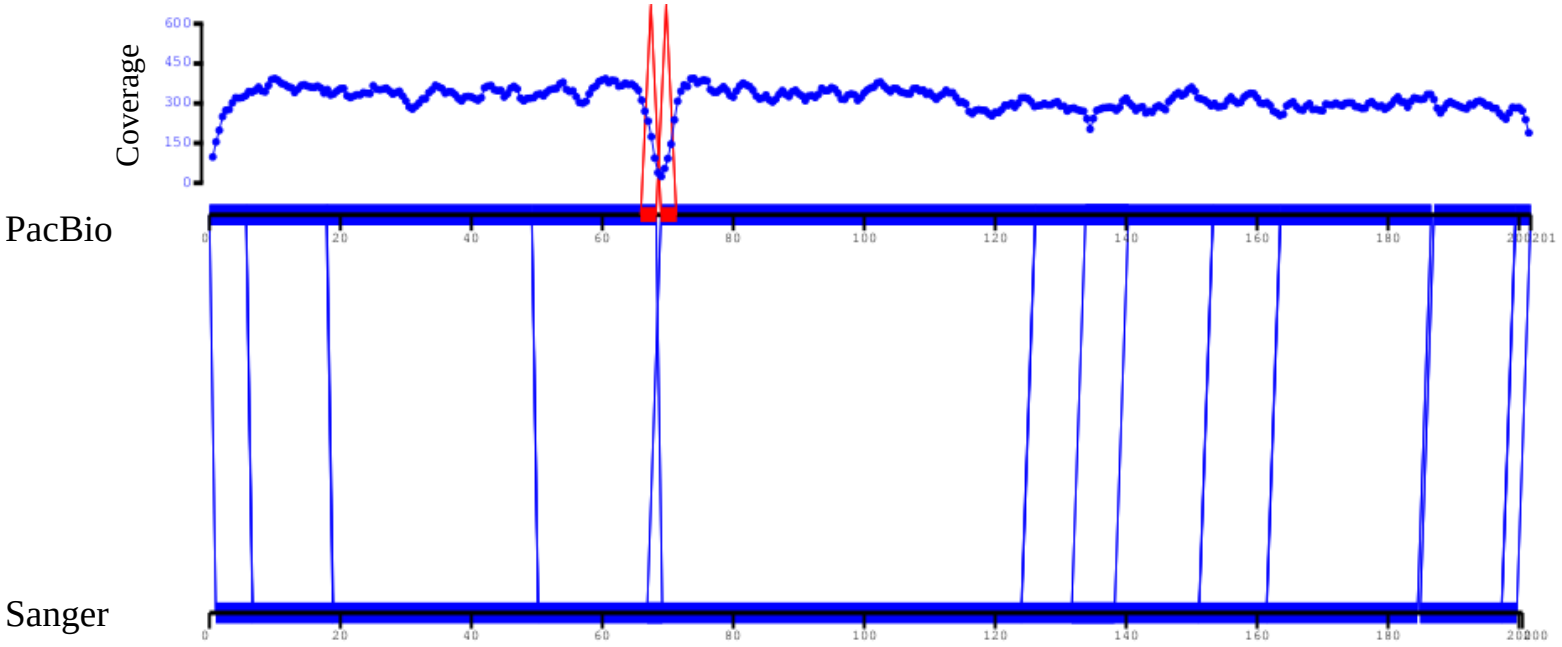
Types of assembly problems

- Large insertions or deletions
- Small mismatches
  - Substitutions
  - Homopolymer and dipolymer indels
  - Other indels

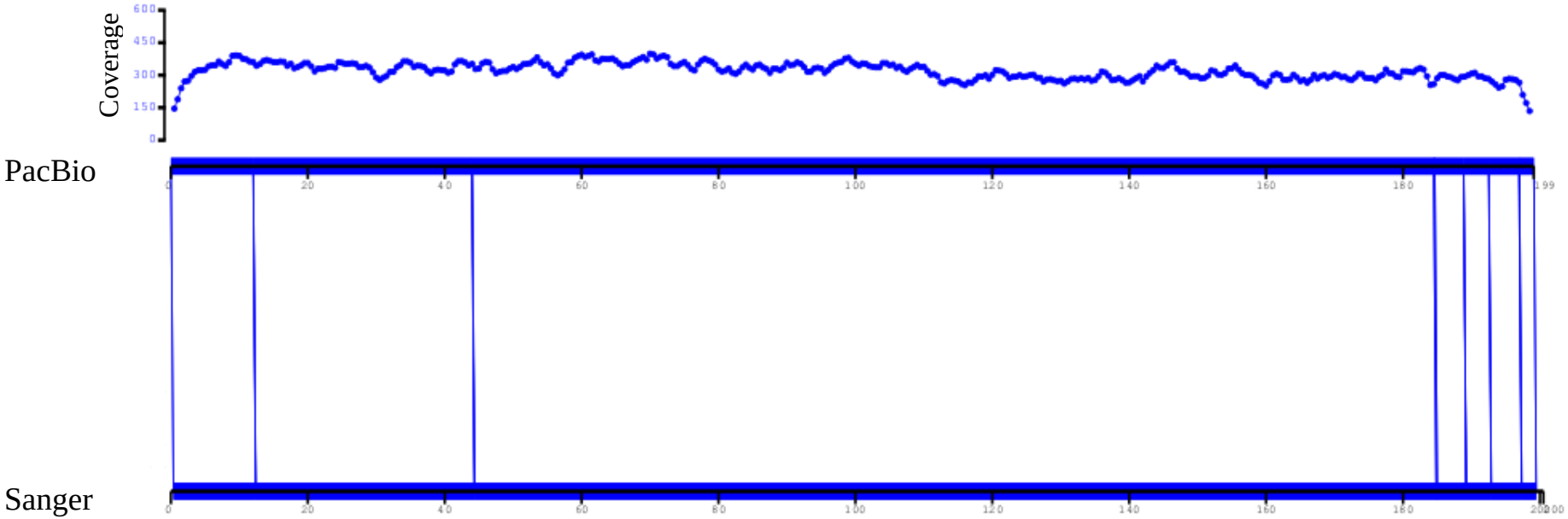
CH17-41F14


Coverage
600
450
300
150
0
PacBio
Sanger
0
20
40
60
80
100
120
140
160
180
200
220
225

CH17-227A2 (with Allora)



CH17-227A2 (with HGAP)



Analyzing small mismatches

1. Align assemblies with BLASR
2. Find mismatches in events < 50 bp long
3. Annotate mismatches by context
    - Homopolymer
    - Dipolymer

Small mismatches (< 50 bp)

Clone	% Identity	Total mismatches	Homopolymer	Dipolymer
CH17-124M20	0.99991	19	11	7
CH17-157L1	0.99994	15	11	1
CH17-169A24	0.99979	51	2	3
CH17-170H8	0.99994	13	11	0
CH17-202L17	0.99999	3	1	1
CH17-227A2	0.99999	3	2	0
CH17-33G3	0.99996	11	6	0
CH17-41F14	0.96303	10	3	1
Total	-	125	47	13

Validating mismatches with Illumina reads

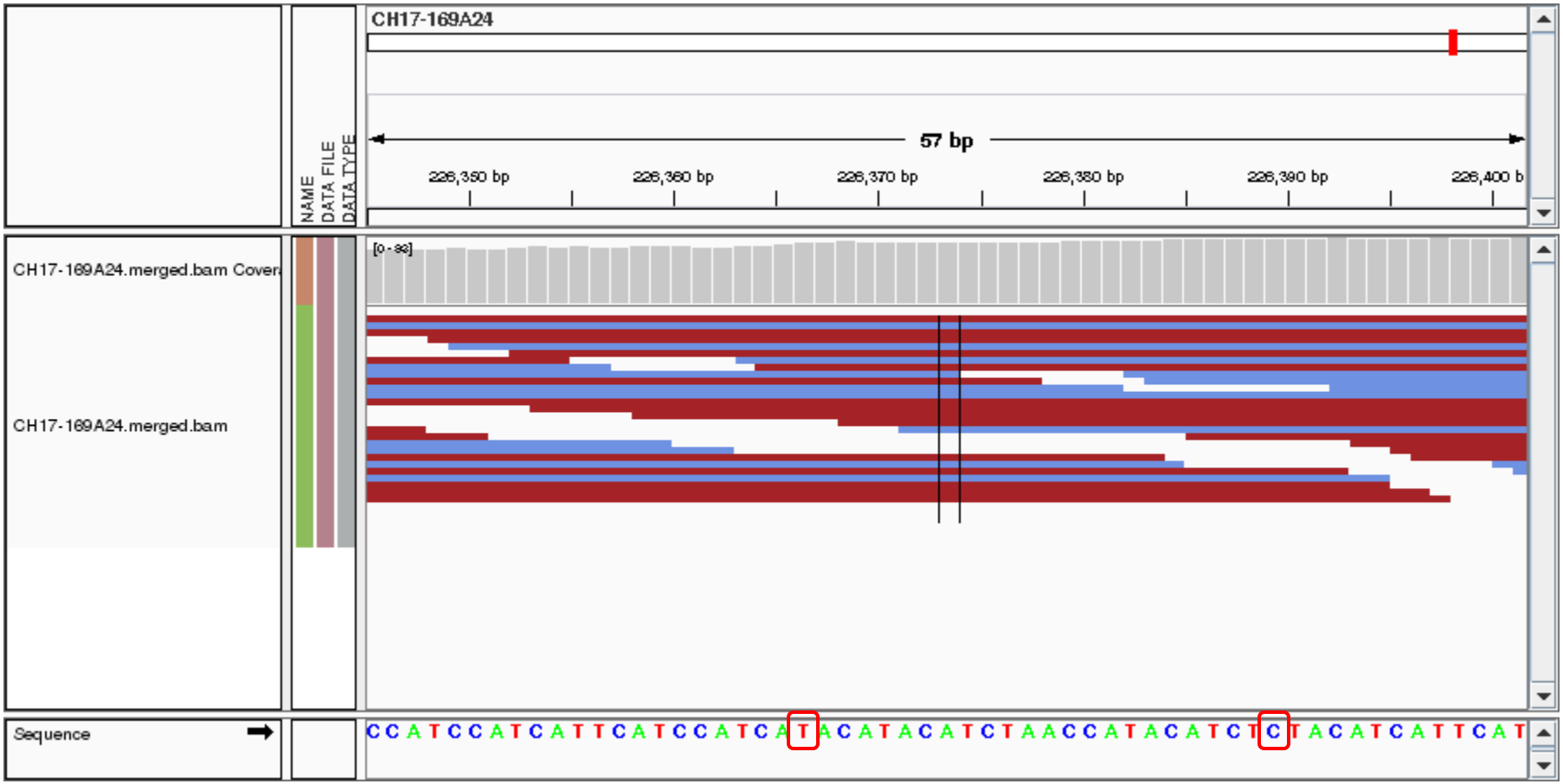
- Map 76 bp Illumina reads to assemblies
- Require perfect mappings of single reads
- Count read support for mismatches
  - At least one read anchored in complex sequence
  - No support for alternate assembly's sequence

Substitutions

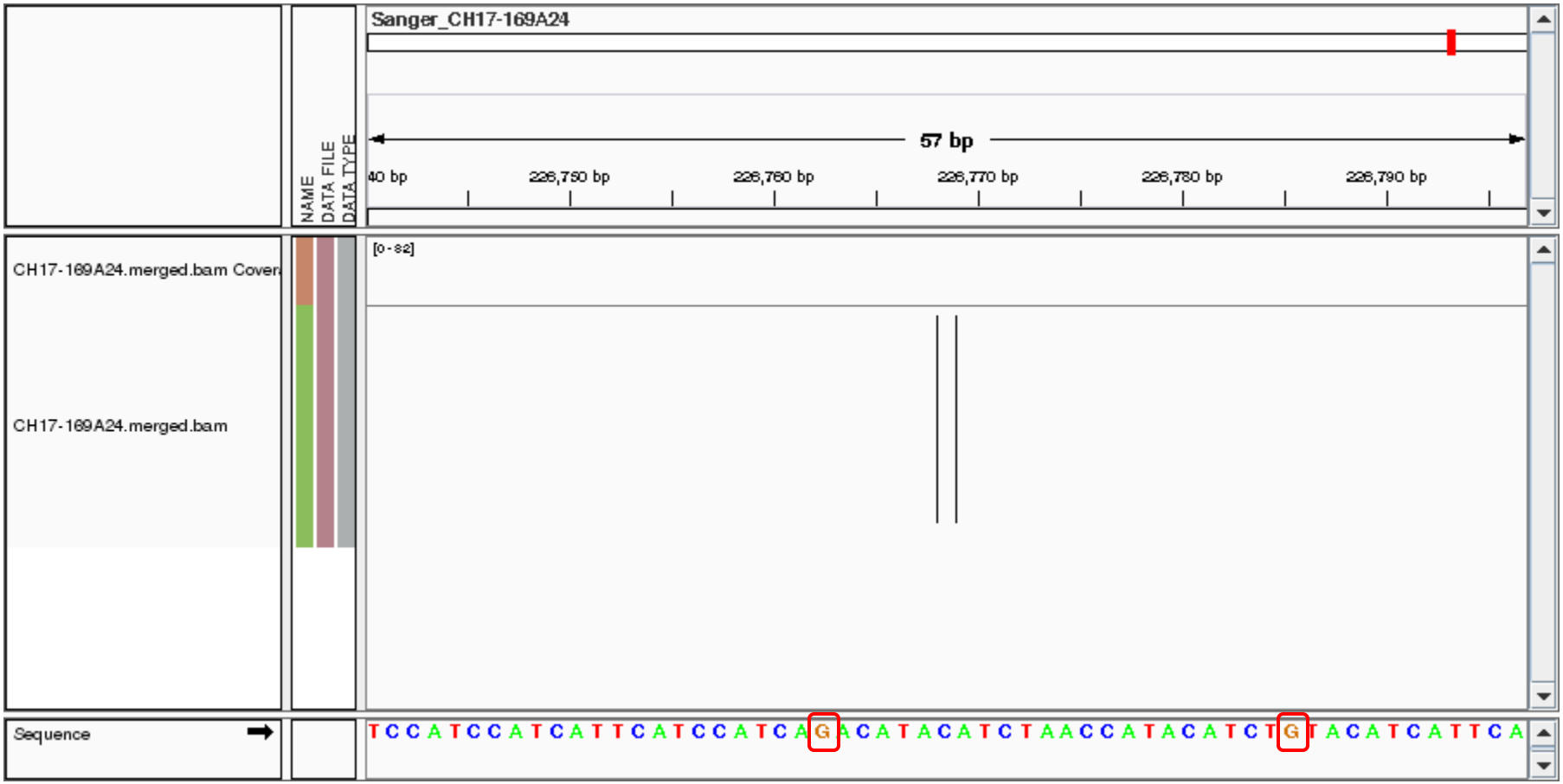
CH17-169A24

```
PB   226348  CCATCATTCATCCATCATACATACATCTAACCATACATCTCTACATCATT
             ||||||||||||||||*||||||||||||||||||||*|||||||||
SG   226744  CCATCATTCATCCATCAGACATACATCTAACCATACATCTGTACATCATT
```

Substitutions



Substitutions

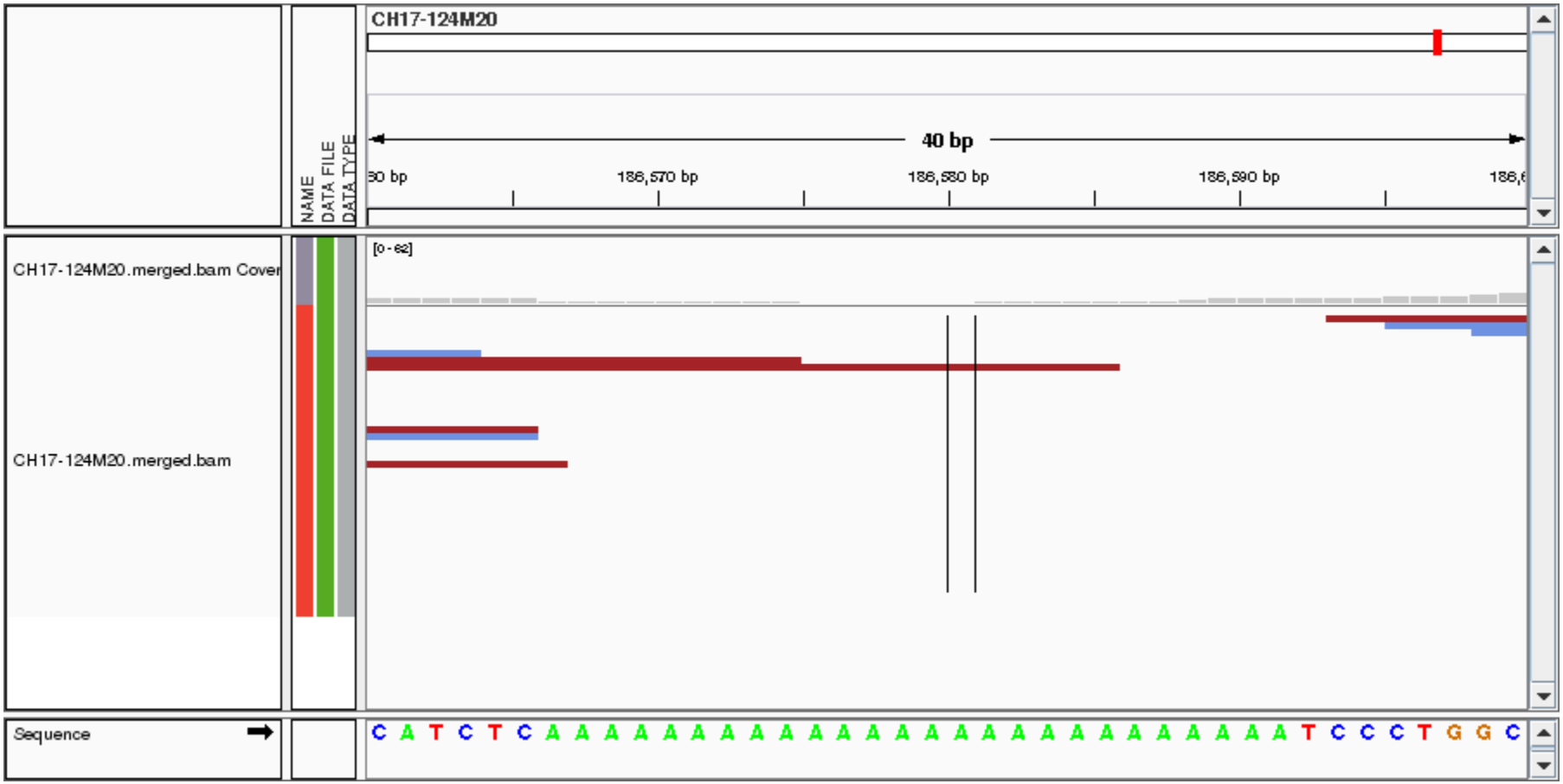


Homopolymers

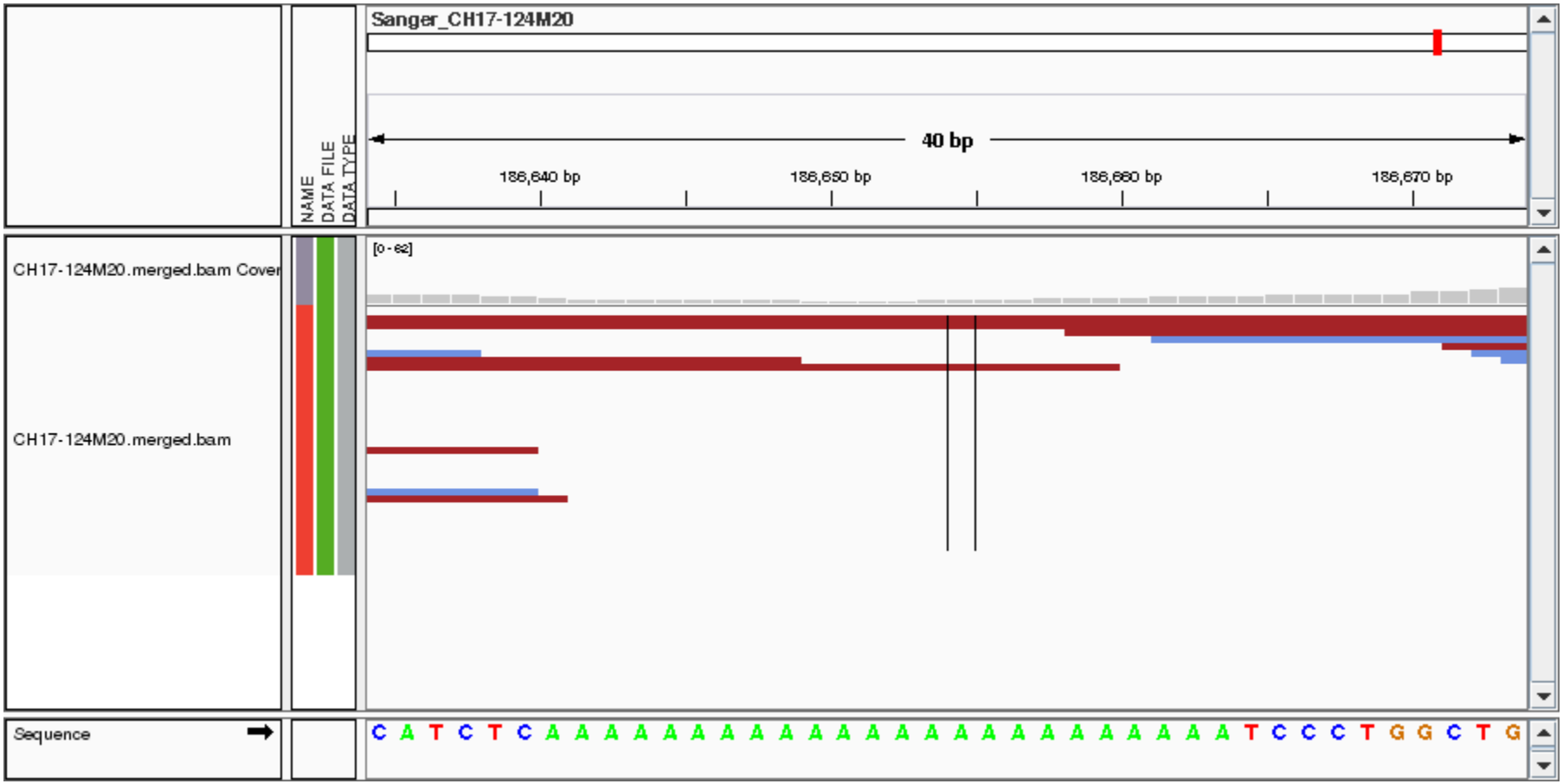
```
CH17-124M20

PB   186548  GAGCAAGACTCCATCTCAAAAAAAAAAAAAAAAAAAAAAAAAATCCCTGG
              ||||||||||||||||  ||||||||||||||||||||||||||||
SG   186622  GAGCAAGACTCCATCTC--AAAAAAAAAAAAAAAAAAAAAAAAATCCCTGG
```

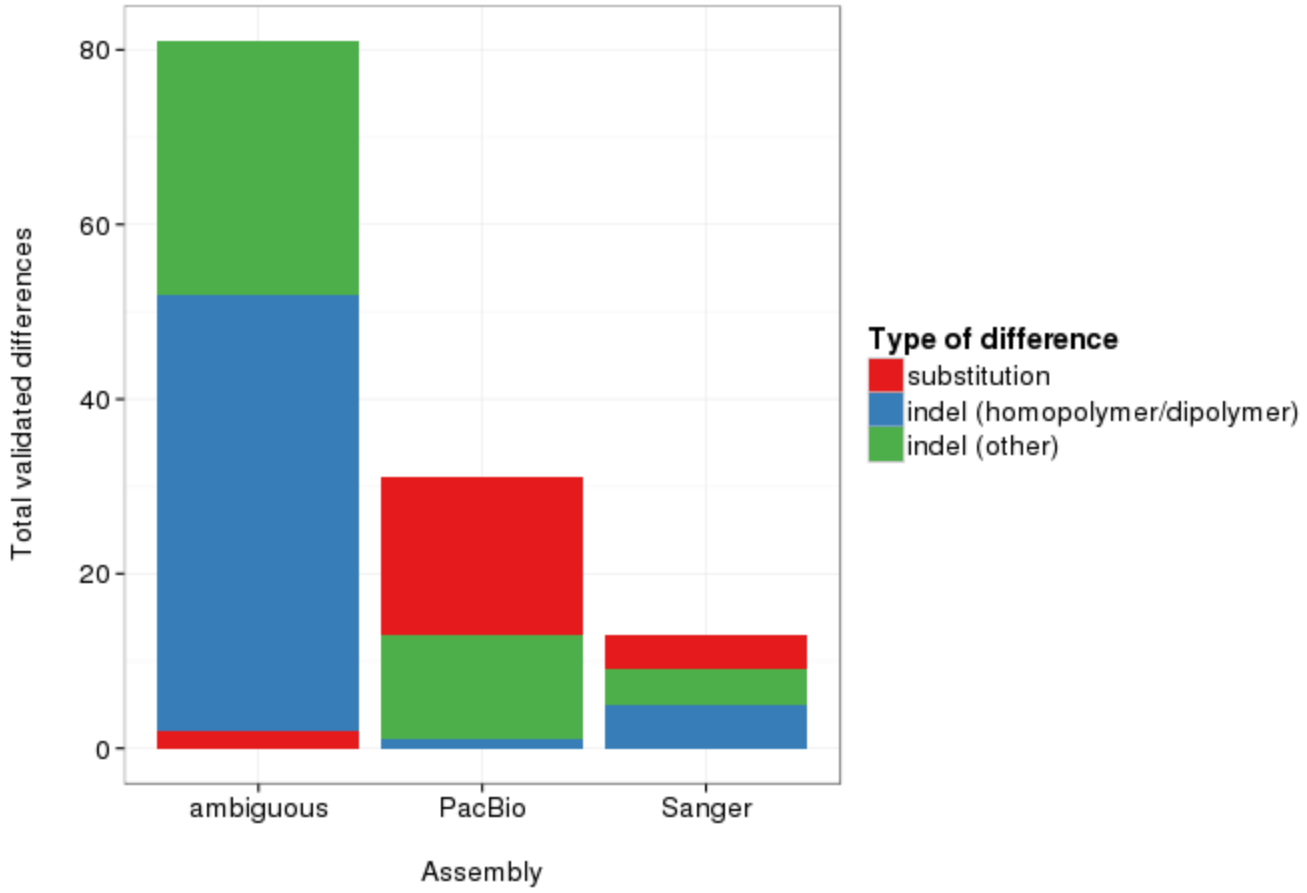
Homopolymers



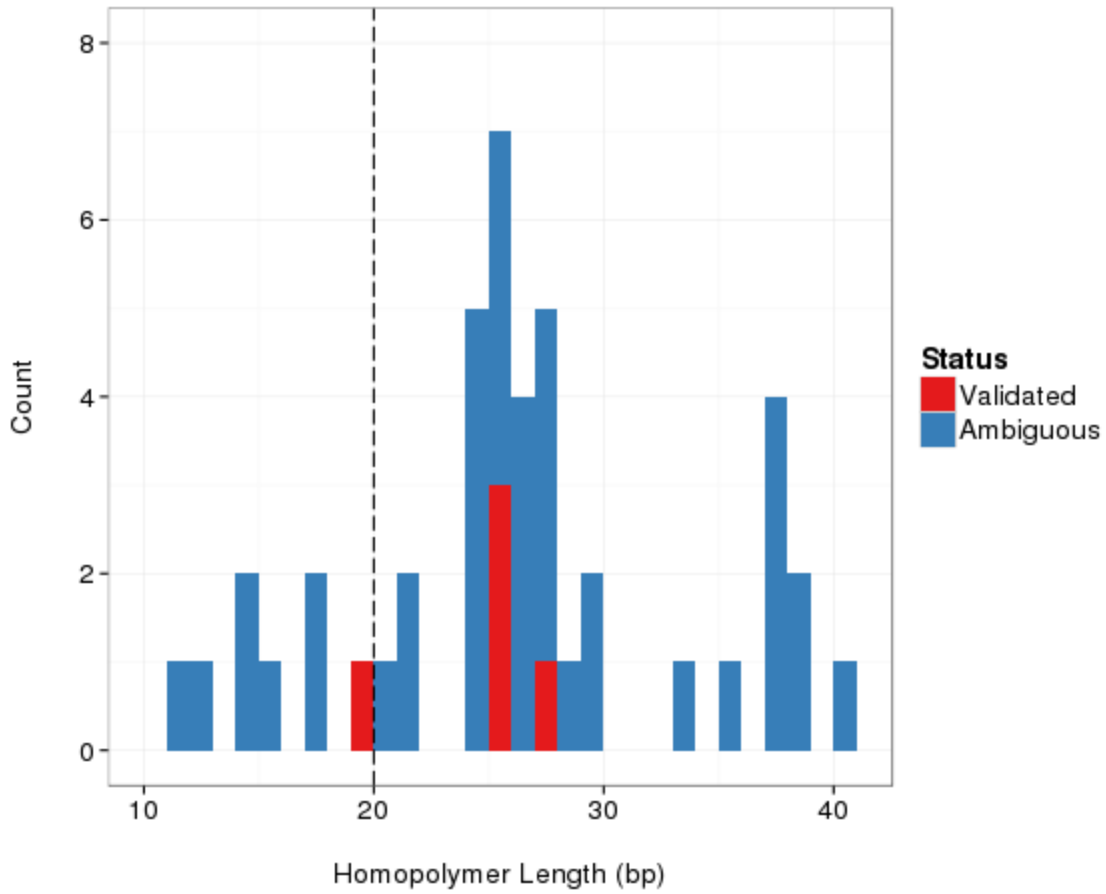
Homopolymers



Validated differences



Homopolymer lengths



Known problems by technology

PacBio

- indels more likely than mismatches
- uniform error distribution (sequencing)

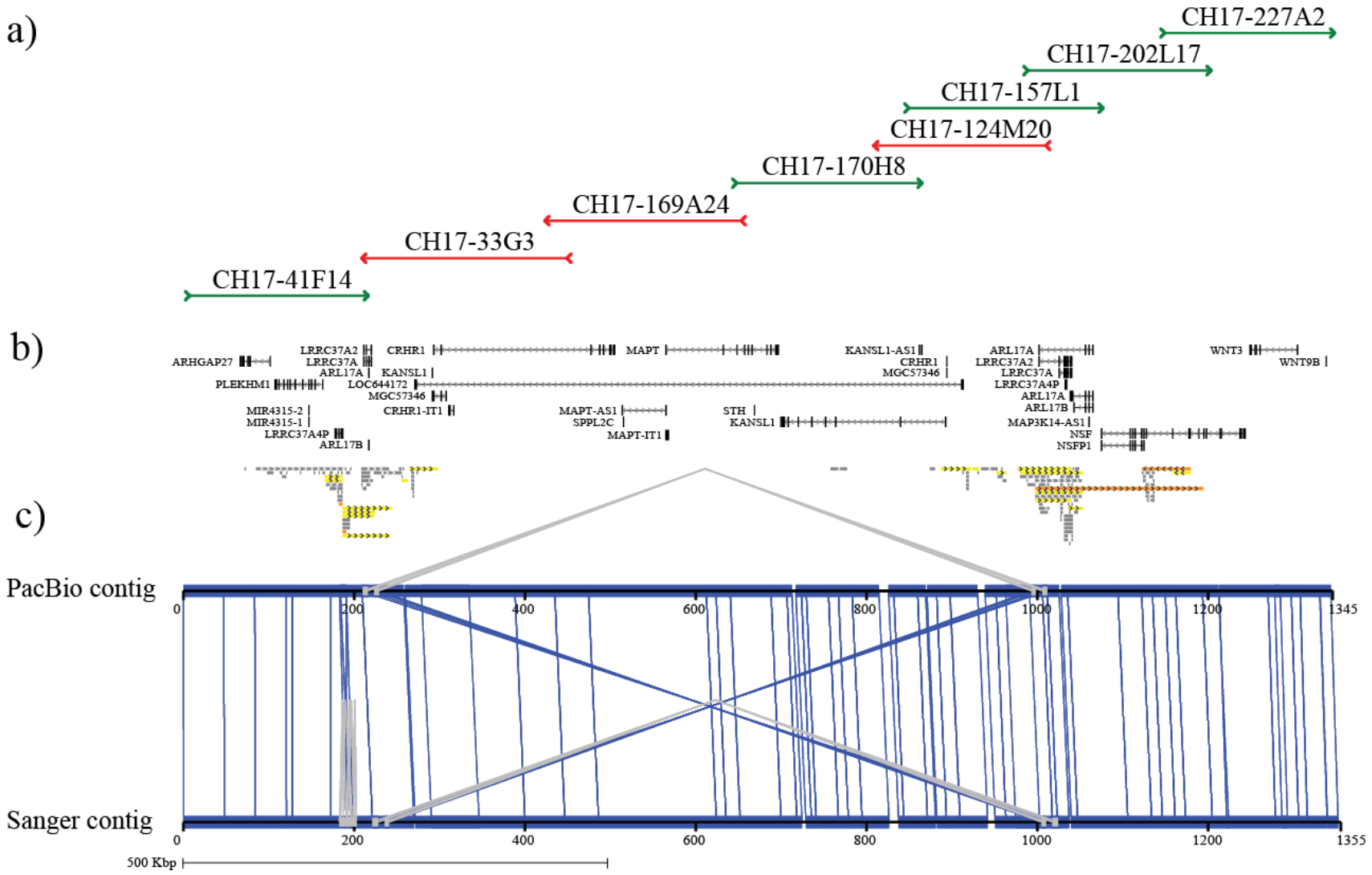
Sanger

- compression in GC-rich sequence
- poly-A and AT slippage

Illumina

- Substitutions more likely than indels
- Low coverage in GC- and AT-rich sequence
- Homopolymer limit of ~20 bp

The final product



Conclusions

- *De novo* assembly of BACs with PacBio can resolve complex regions
  - One contig per BAC
  - Q40-50 quality (99.99-99.999% accuracy)
  - 2 days of sample prep, 1 day of assembly/finishing
- Remaining problems
  - Collapsed duplications
  - Very long homopolymer sequences

Acknowledgements

Eichler lab

Francesca Antonacci

Mark Chaisson

Maika Malig

Megan Dennis

Peter Sudmant

Evan Eichler

PacBio team

Lawrence Hon

David Alexander

Aaron Klammer

Swati Ranade

Nick Sisneros

Steve Turner

UW Genome Sciences

Brenton Munson

Skylar Thompson

Liz Young