New Sequencing Technologies and Hybrid Assemblies

Harindra M. Arachchi, Manuel Garber, Chad Nusbaum
Sarah Young, Michael C. Zody, David Jaffe, Michael Fitzgerald

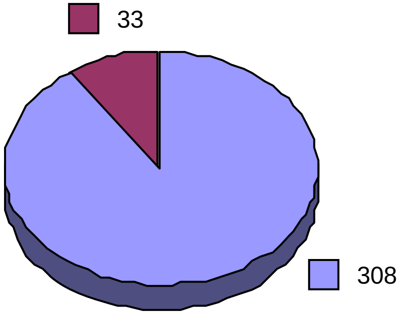
Objective

- Obtain PCR template from unclonable gaps
- Take advantage of new technologies to sequence them
- Develop methodology for closing these gaps

Finished human genome (2004)

Finished human genome build 35

341 gaps



🟥 Heterochromatic

🟦 Euchromatic

2.85 billion nucleotides

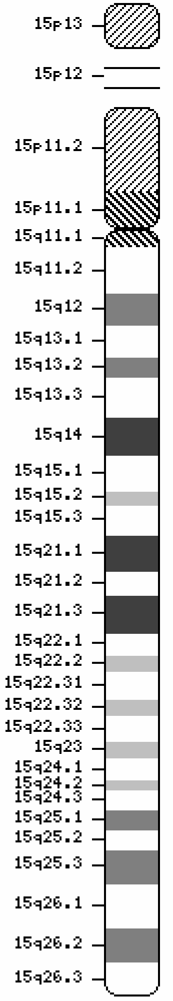
(99% of the euchromatic genome)

Unclonable

Segmental duplications

Gaps on human chromosome 15

- 10 gaps
  - 7 in segmental duplicated regions
  - 3 apparently unclonable
    - Not captured in human tiling path
    - Not captured in chimpanzee tiling path
- 53x physical coverage screened



courtesy NCBI

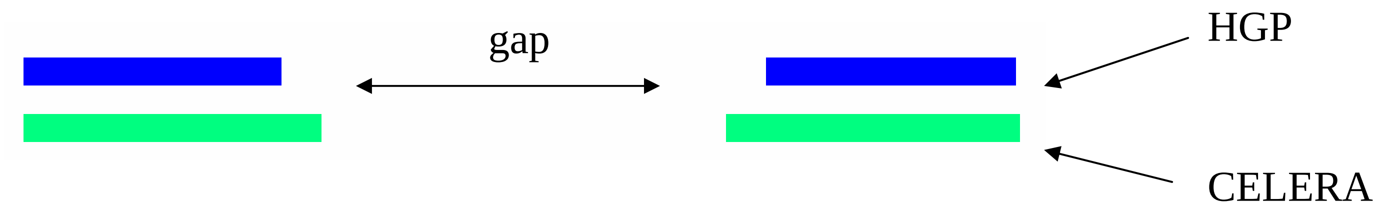
Overview.

- Process
- Analysis
- Conclusion

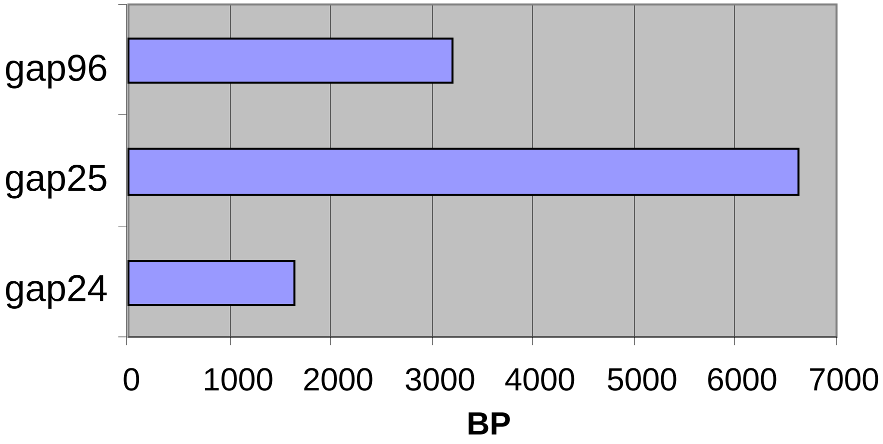
Overview.

- **Process**
- Analysis
- Conclusion

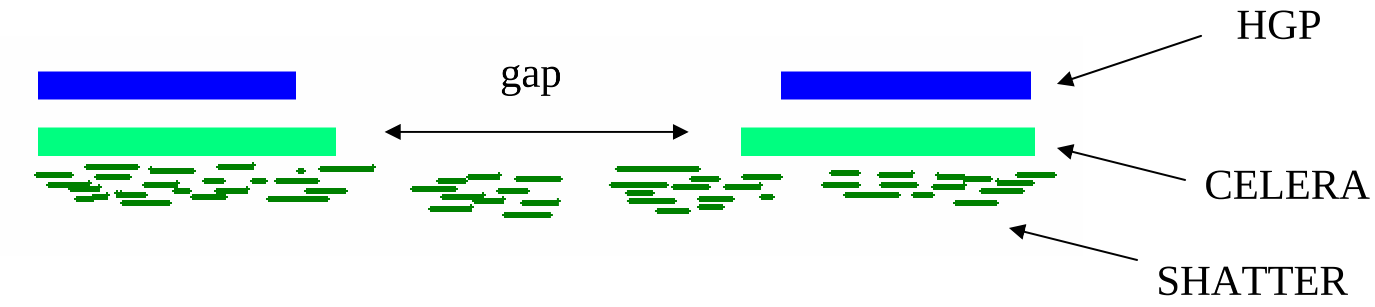
STEP 1: Align HGP & Celera



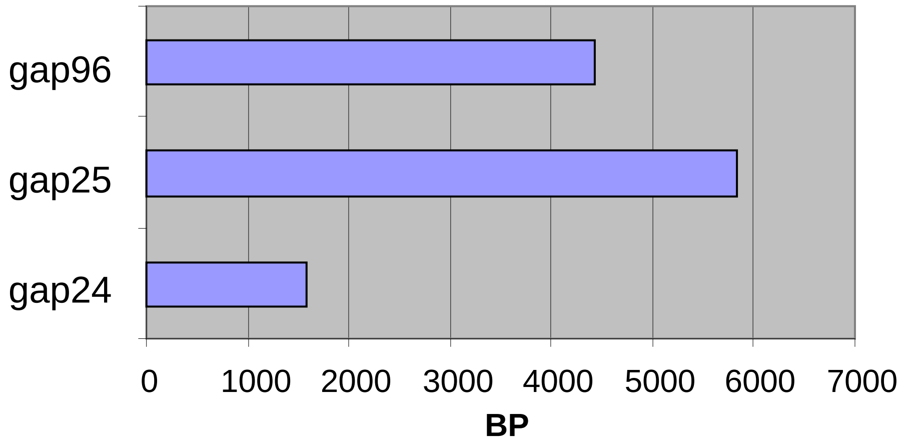
Celera coverage of gap



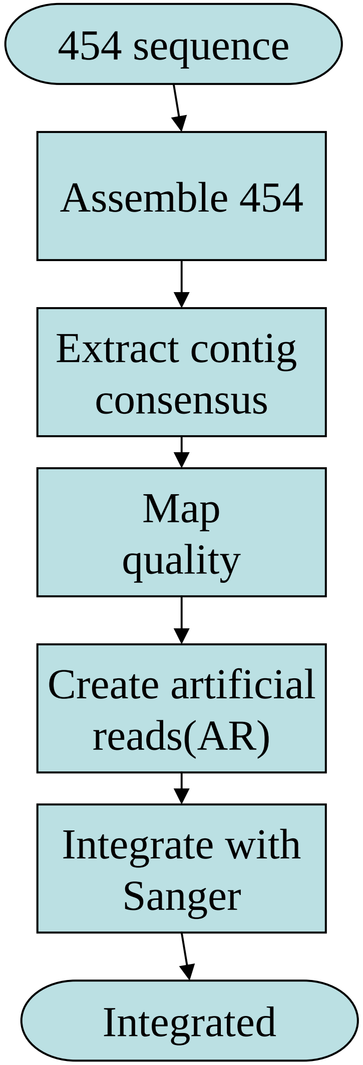
STEP 2: Integrate shatter reads



Shatter library coverage of gap



STEP 3: Integrate 454 data



454 sequence

Assemble 454

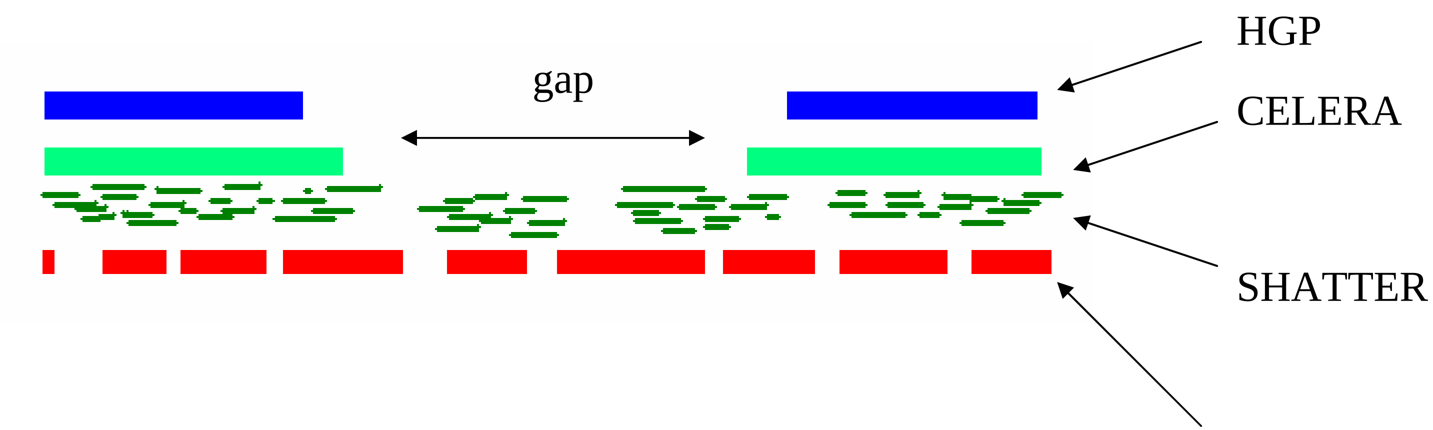
Extract contig consensus

Map quality

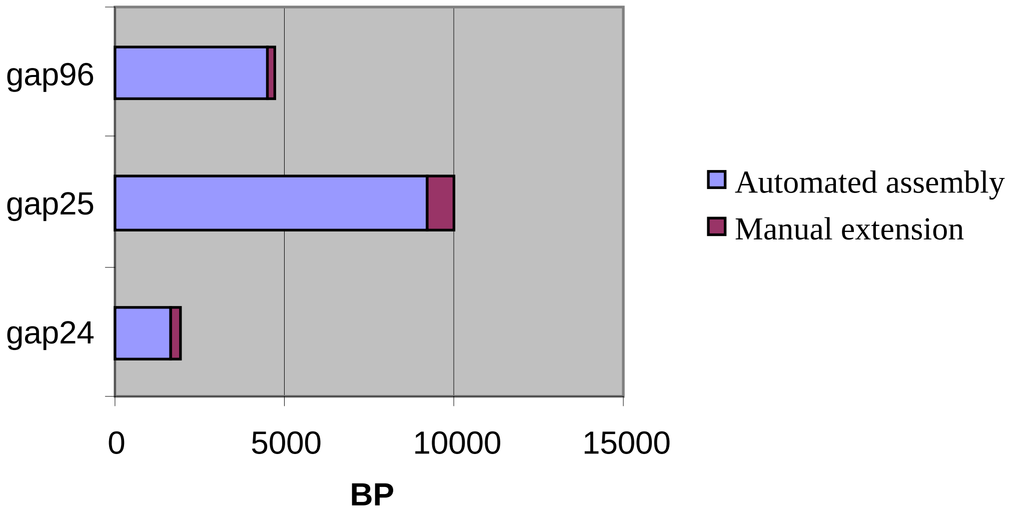
Create artificial reads(AR)

Integrate with Sanger

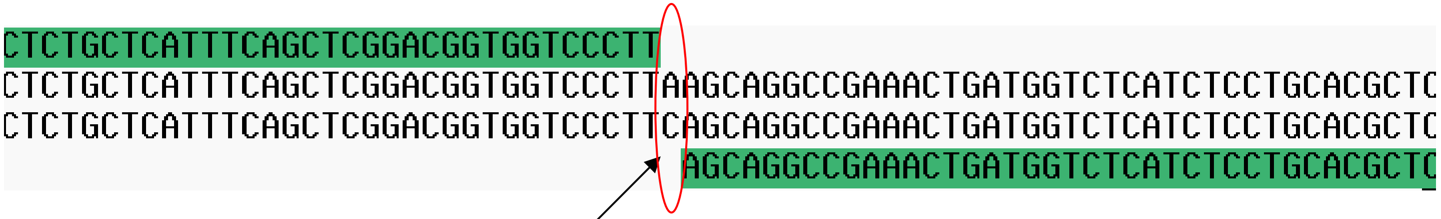
Integrated



454 coverage of gap



Newbler assembly breaks at SNPs



```
CTCTGCTCATTTCAGCTCGGACGGTGGTCCCTT
CTCTGCTCATTTCAGCTCGGACGGTGGTCCCTTAAGCAGGCCGAAACTGATGGTCTCATCTCCTGCACGCTC
CTCTGCTCATTTCAGCTCGGACGGTGGTCCCTTCAGCAGGCCGAAACTGATGGTCTCATCTCCTGCACGCTC
                                  AGCAGGCCGAAACTGATGGTCTCATCTCCTGCACGCTC
```

Sanger data

454 data

Haplotype difference

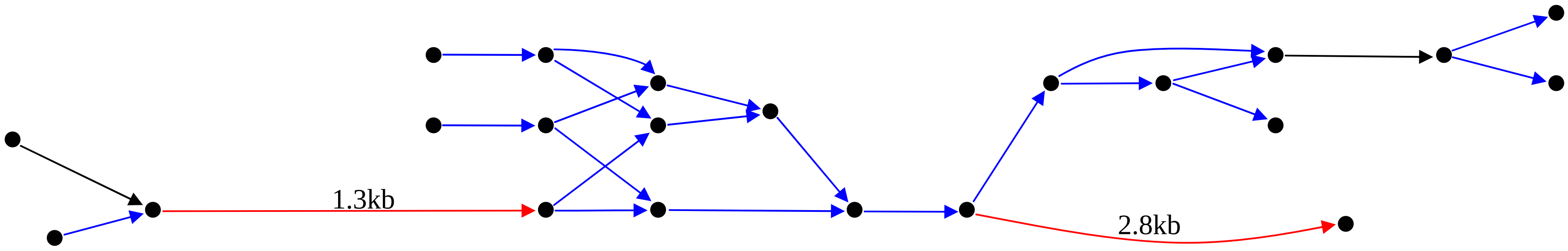
454 Assembly: ALLPATHS

ALLPATHS assembly algorithm

-Finds all shared kmers between reads

-Uses shared kmers to build graph across the data

-Graphs represents contigs as edges, branch points as nodes



David Jaffe

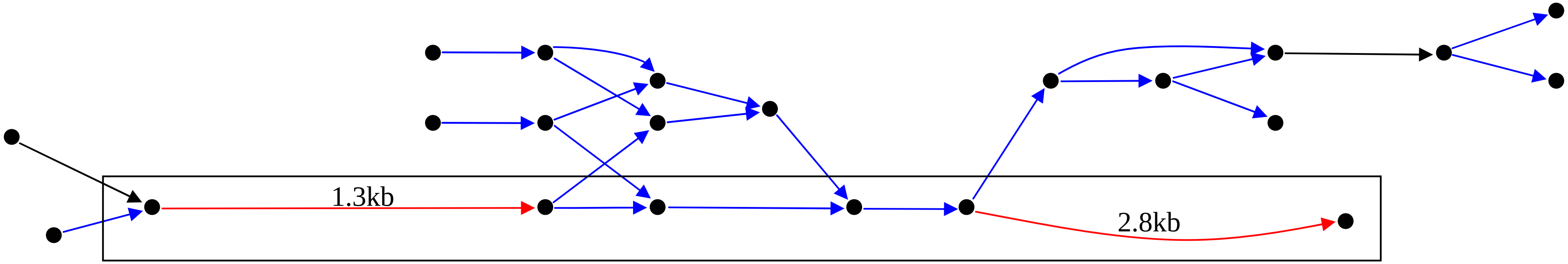
454 Assembly: ALLPATHS

ALLPATHS assembly algorithm

-Finds all shared kmers between reads

-Uses shared kmers to build graph across the data

-Graphs represents contigs as edges, branch points as nodes



Complete path across region: spans a gap

David Jaffe

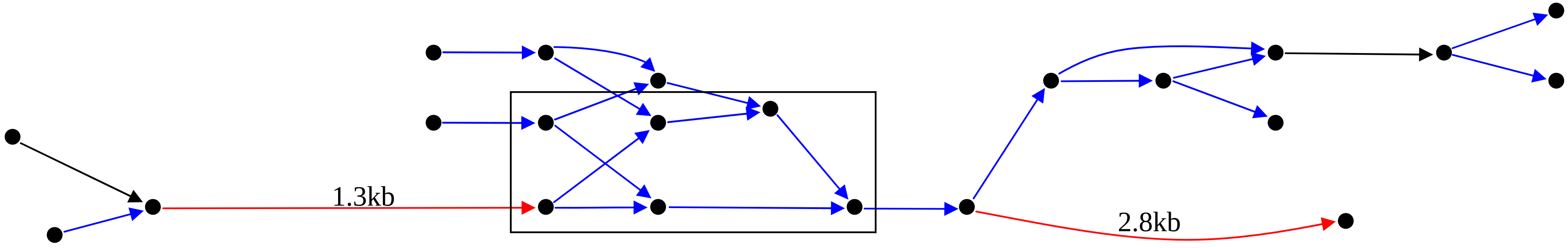
454 Assembly: ALLPATHS

ALLPATHS assembly algorithm

-Finds all shared kmers between reads

-Uses shared kmers to build graph across the data

-Graphs represents contigs as edges, branch points as nodes



Branched path = two haplotypes

David Jaffe

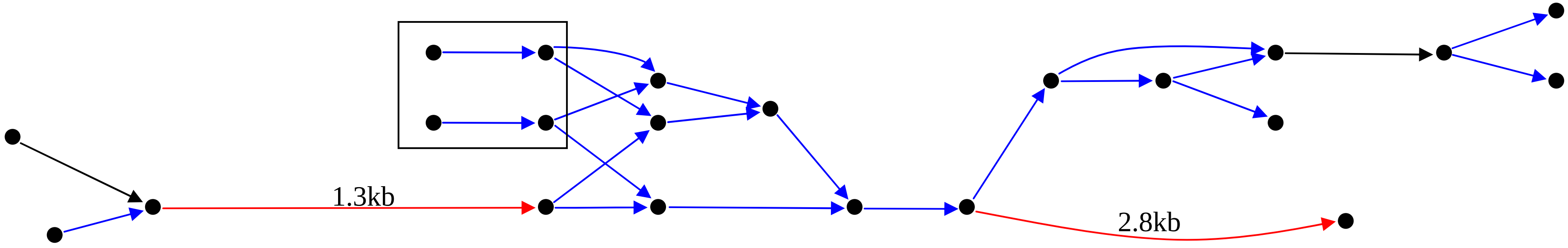
454 Assembly: ALLPATHS

ALLPATHS assembly algorithm

-Finds all shared kmers between reads

-Uses shared kmers to build graph across the data

-Graphs represents contigs as edges, branch points as nodes



PCR slippage in repeat

David Jaffe

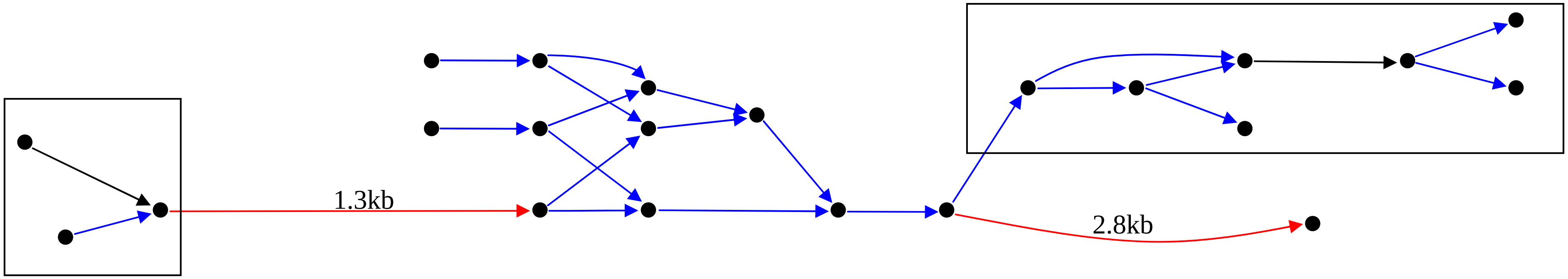
454 Assembly: ALLPATHS

ALLPATHS assembly algorithm

-Finds all shared kmers between reads

-Uses shared kmers to build graph across the data

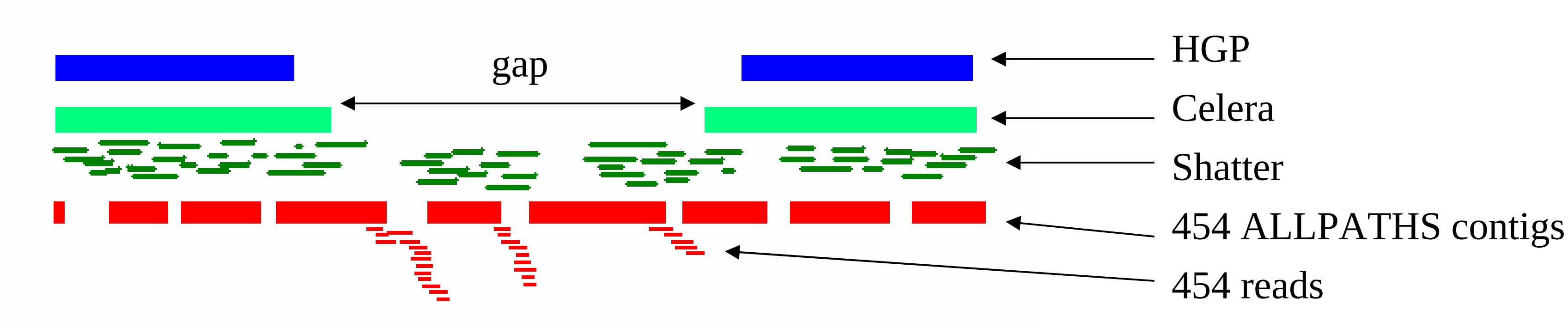
-Graphs represents contigs as edges, branch points as nodes



Artifacts from messy PCR

David Jaffe

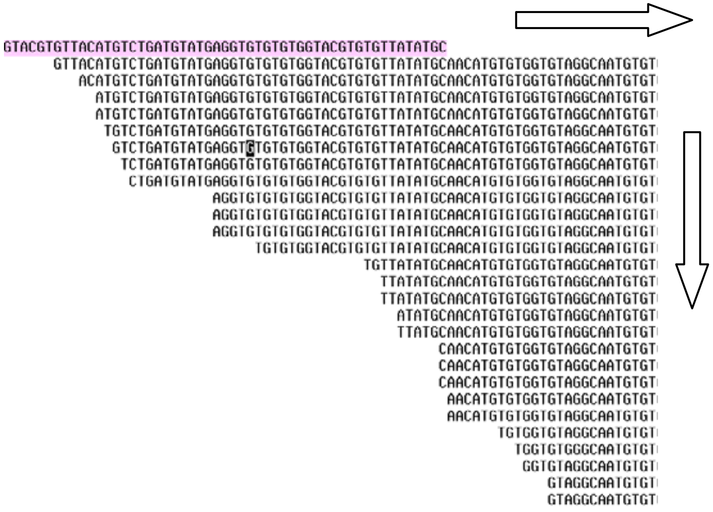
Manual extension of contigs



- Identify unassembled reads computationally by string search
- Manually insert reads into assembly

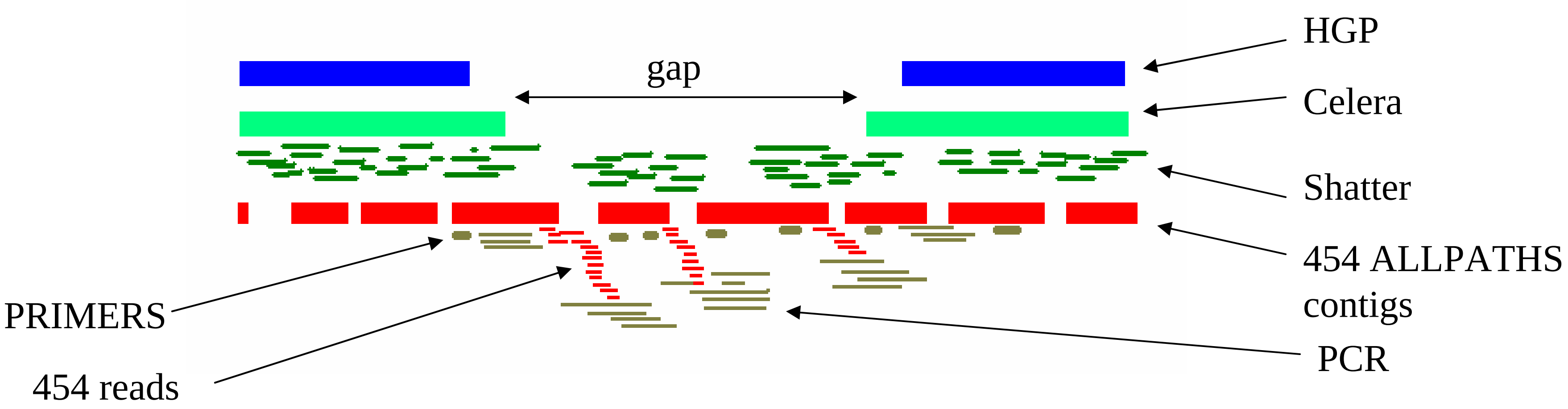
454 contig GTCT

"Fished" in 454 reads ATGT



```
GTACGTGTTACATGTCTGATGTATGAGGTGTGTGTGGTACGTGTGTTATATGC
      GTTACATGTCTGATGTATGAGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
         ACATGTCTGATGTATGAGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
           ATGTCTGATGTATGAGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
           ATGTCTGATGTATGAGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
            TGTCTGATGTATGAGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
             GTCTGATGTATGAGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
              TCTGATGTATGAGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
               CTGATGTATGAGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
                         AGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
                         AGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
                         AGGTGTGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
                              TGTGTGGTACGTGTGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
                                           TGTTATATGCAACATGTGTGGTGTAGGCAATGTGT
                                             TTATATGCAACATGTGTGGTGTAGGCAATGTGT
                                             TTATATGCAACATGTGTGGTGTAGGCAATGTGT
                                               ATATGCAACATGTGTGGTGTAGGCAATGTGT
                                               TTATGCAACATGTGTGGTGTAGGCAATGTGT
                                                    CAACATGTGTGGTGTAGGCAATGTGT
                                                    CAACATGTGTGGTGTAGGCAATGTGT
                                                    CAACATGTGTGGTGTAGGCAATGTGT
                                                     AACATGTGTGGTGTAGGCAATGTGT
                                                     AACATGTGTGGTGTAGGCAATGTGT
                                                           TGTGGTGTAGGCAATGTGT
                                                             TGGTGTGGGCAATGTGT
                                                              GGTGTAGGCAATGTGT
                                                                 GTAGGCAATGTGT
                                                                 GTAGGCAATGTGT
```

STEP 4: PCR to verify 454



gap

HGP

Celera

Shatter

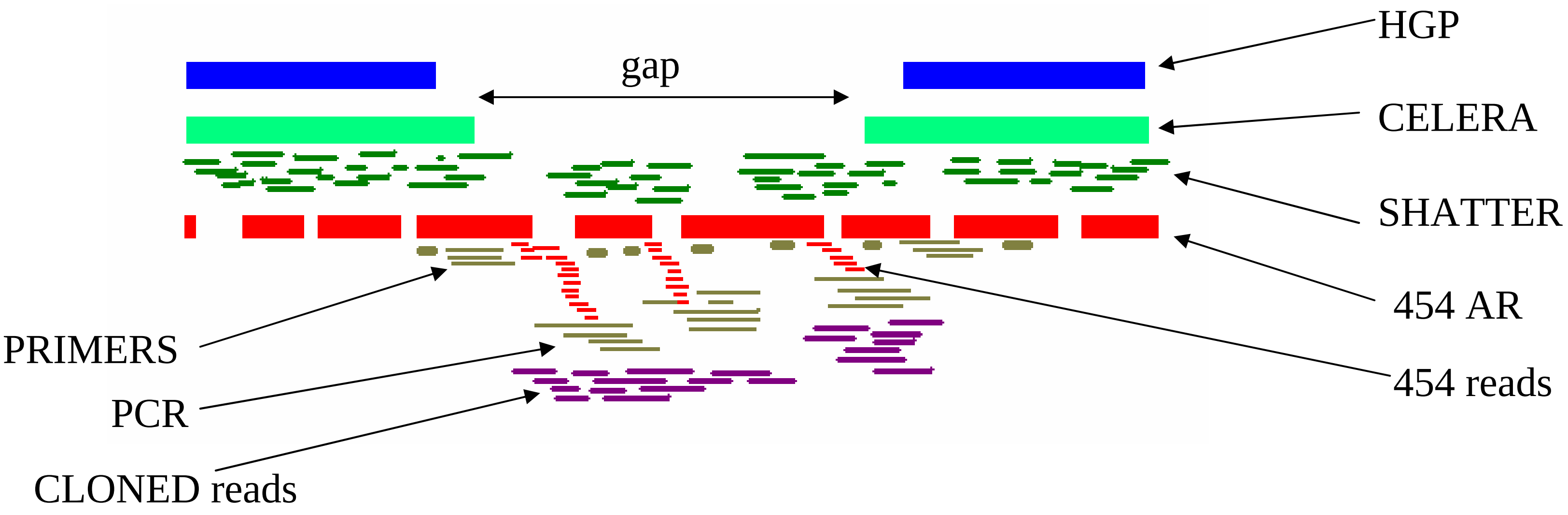
454 ALLPATHS contigs

PCR

PRIMERS

454 reads

STEP 5: Cloning PCR fragments



gap

HGP

CELERA

SHATTER

454 AR

454 reads

PRIMERS

PCR

CLONED reads

Overview

- Process
- **Analysis**
- Conclusion

Hard finishing….

- Sequence flanking all 3 gaps looks normal
- Problem is nasty repeat in gap
  - Unable to clone
  - Can sequence by 454 but Newbler can't assemble
  - ALLPATHS gave better assembly
  - Had to finish the job by hand

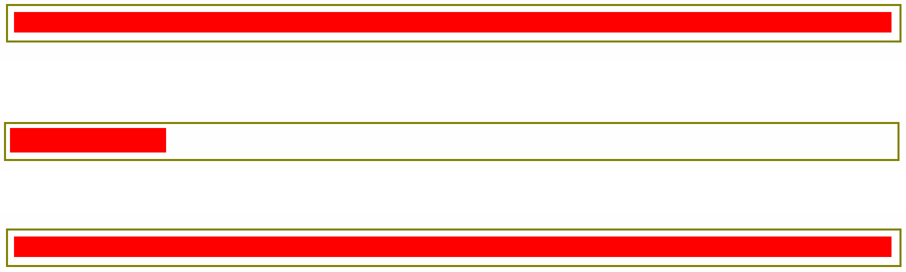
Gap25

Status: Closed

Size: 10225bp

Unclonable: 21.6%

Total 454 coverage: 98%



Repeat motif: GGTGTTTGTGTGTGTATGGT

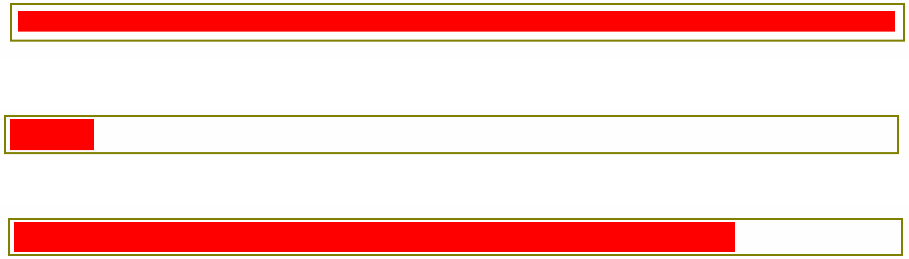
Gap96

Status: Closed

Size: 5474bp

Unclonable: 14%

Total 454 coverage: 83.7%



Repeat motif: TATGTGTGTGGCATGTGTGGT

Gap24

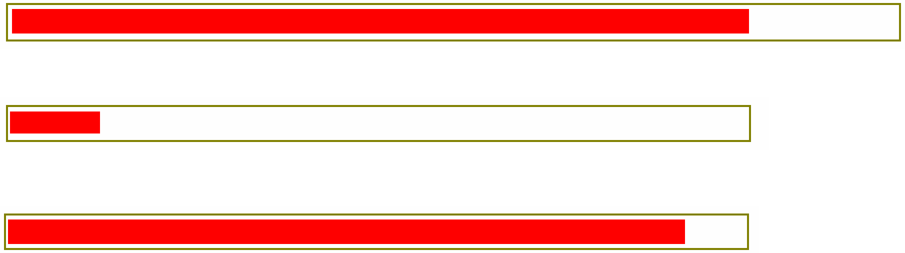
Status: Active

Deletions complications

Size: 2539bp**

Unclonable: 14%

Total 454 coverage: 92%



Repeat motif:

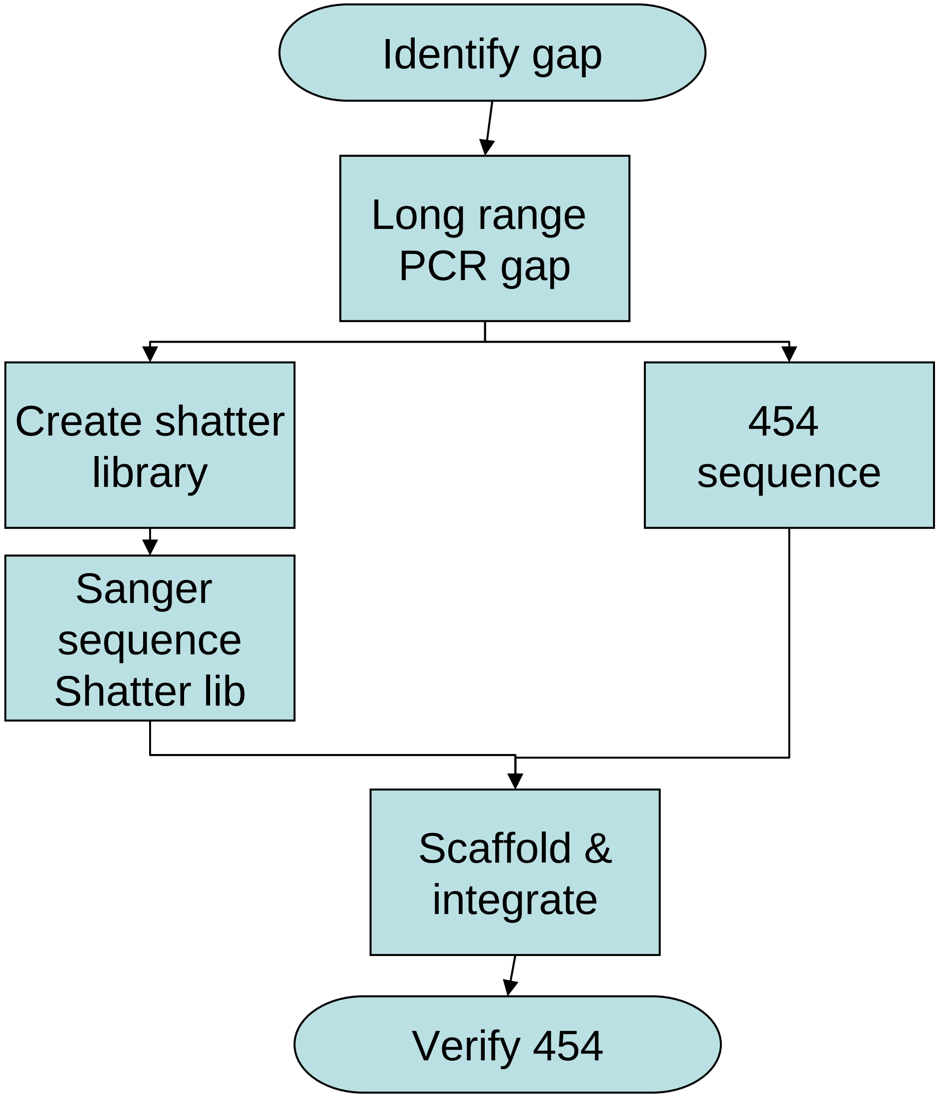
TGTATGGTGTGTGGCGTGTG

**sequence captured so far

Overview

- Process
- Analysis
- **Conclusion**

Summary



HGP unclonable remaining:126

- Not due to copy number variations
- In unique regions

How do we define “finished” ?

Thank you!

Manuel Garber
Sarah Young
Chad Nusbaum
Michael C. Zody

Michael Fitzgerald
Will Lee
David Jaffe
Lisa Zembek
Niall Lennon

Cristyn Kells
James Bonfield

Broad Institute
Special Projects Group
Finishing Group
Genome Sequencing Platform