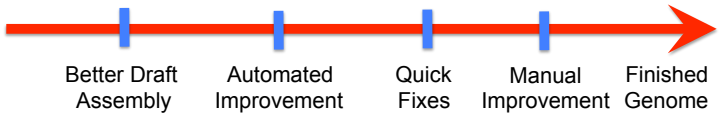
Pilon:
Automated Assembly Improvement

Bruce Walker
Director, Genome Assembly & Analysis
Broad Institute

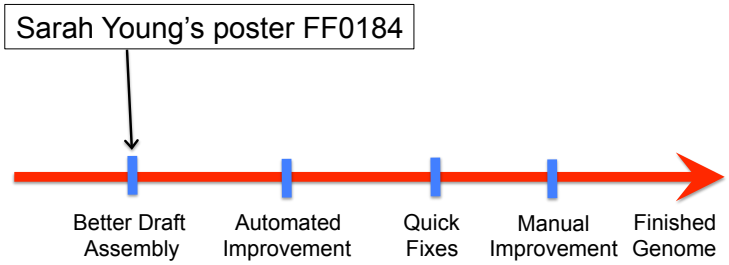
bruce@broadinstitute.org

BROAD INSTITUTE

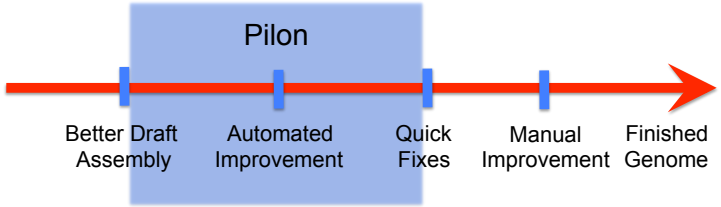
Spectrum of Genome Improvement



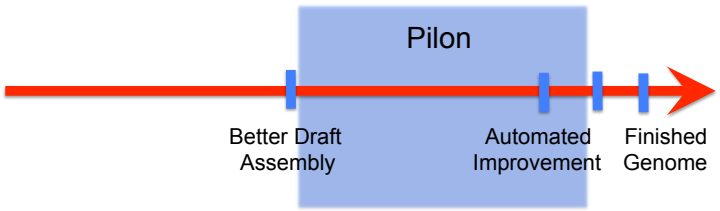
Spectrum of Genome Improvement



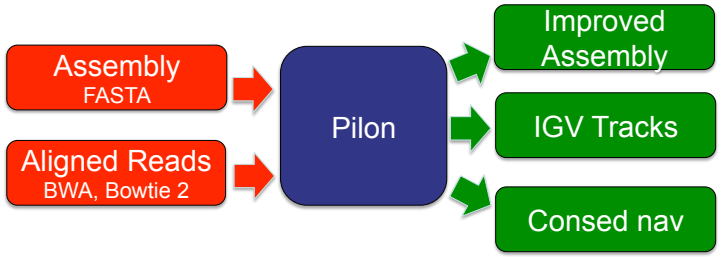
Pilon: Assembly Improvement Tool



Goal: Get Further With Automation

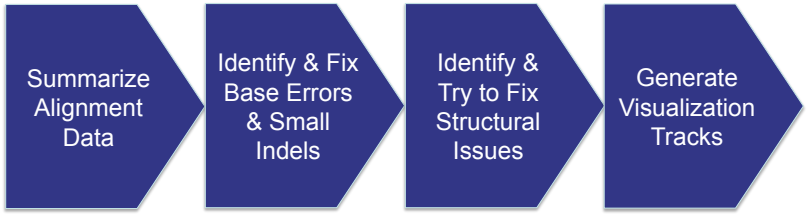


Pilon Overview



- Identifies inconsistencies between read data and assembly
- Fixes some detected differences automatically
- Reports other potential issues to target quick manual inspection

Pilon Operation



Valid and Invalid Pairs

Valid: plausible orientation and separation ("Proper Pair")

Invalid: bogus orientation, separation, or unmapped mate

Invalid doesn't mean pair is bad -- it might be the assembly!

Summarize Alignment Pileups (for every locus, for every library)

- Individual base stacks
  - Allele counts
  - Allele quality sums

Summarize Alignment Pileups (for every locus, for every library)

- Individual base stacks
  - Allele counts
  - Allele quality sums

- Indels
  - Insertion events
  - Deletion events

Summarize Alignment Pileups (for every locus, for every library)

- Individual base stacks
  - Allele counts
  - Allele quality sums
- Indels
  - Insertion events
  - Deletion events

- Coverage information
  - Sequence coverage
  - Physical coverage

Summarize Alignment Pileups (for every locus, for every library)

- Individual base stacks
  - Allele counts
  - Allele quality sums
- Indels
  - Insertion events
  - Deletion events
- Coverage information
  - Sequence coverage
  - Physical coverage

- Pairing information
  - Valid pair coverage
  - Invalid pair coverage
  - Local implied insert size

Identifying Base Errors & Ambiguities

Using pileup summary data,
Pilon calls every base in the assembly as:

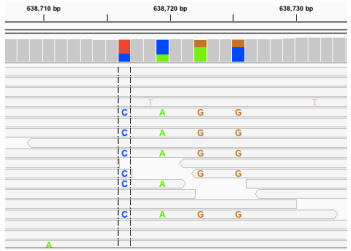
- Correct
- Incorrect
- Ambiguous
- Uncallable

Pilon will fix “incorrect” bases

Resolving Repeats Using Jump Pairs

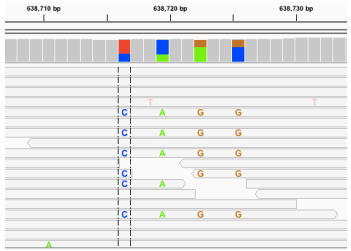
Assembled repeats often contain ambiguous bases

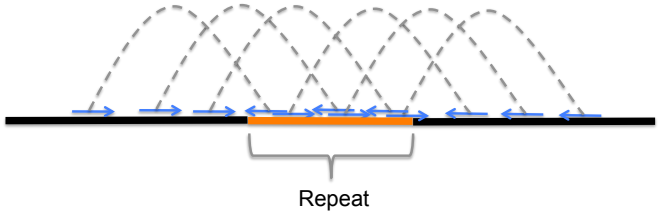
Fragment pairs are too short to disambiguate



Resolving Repeats Using Jump Pairs

Valid jump pairs can reach in to resolve

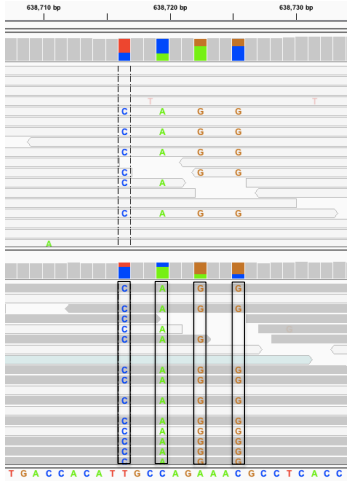




Resolving Repeats Using Jump Mates

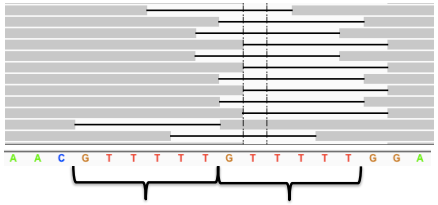
Valid jump reads (grey) agree on base corrections

Pilon will fix these 4 bases



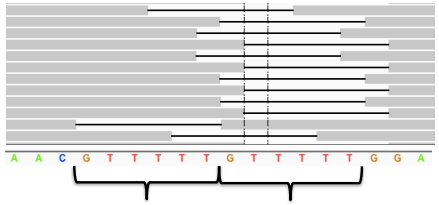
Identifying Small Indels

Small indels are identified from the alignments

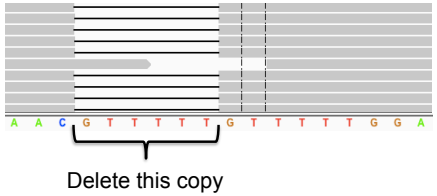


Identifying Small Indels

Small indels are identified from the alignments



Pilon shifts equivalent indels so they stack up properly

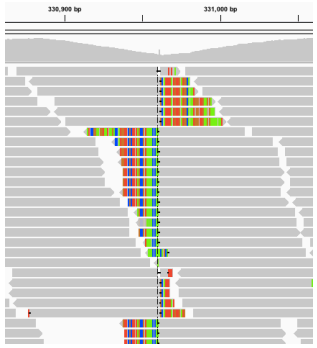


Identifying Local Misassemblies

Coverage motif: rapidly changing valid coverage on order of read length distance

- Especially "V-dips"

Obvious continuity break!

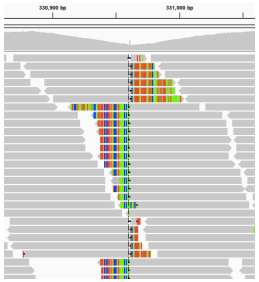


Fixing Local Misassemblies

- Find fragment pairs from the local neighborhood
- Re-assemble the local reads
- Ensure continuity with good sequence on both sides
- If no solution, optionally create contig break

Fixing Local Misassemblies

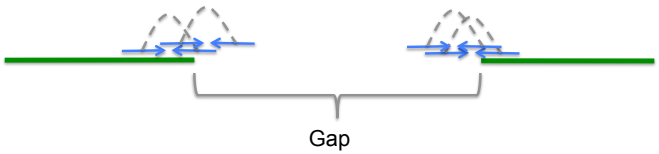
- Find fragment pairs from the local neighborhood
- Re-assemble the local reads
- Ensure continuity with good sequence on both sides
- If no solution, optionally create contig break



Pilon found 48-base deletion

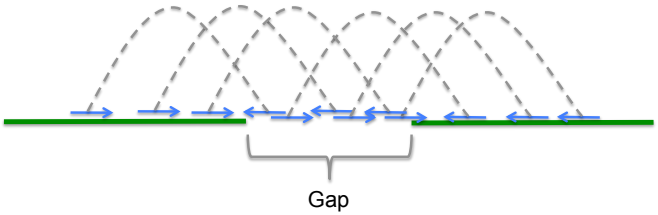
Filling Gaps
Walk in from sides

- Find fragments which align near edge of gap and their mates
- Walk in, extending contigs from each side



Filling Gaps
Find candidate jumps & assemble

- Find well-anchored and correctly oriented jumps in the neighborhood whose mates might fall in the gap



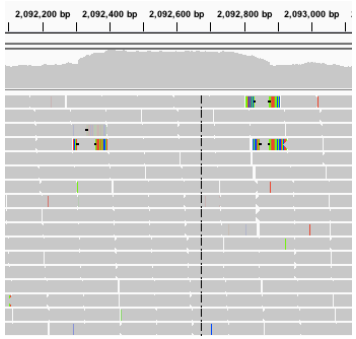
- Do a local reassembly using the frags and unplaced jump mates

Identifying Collapsed Repeats

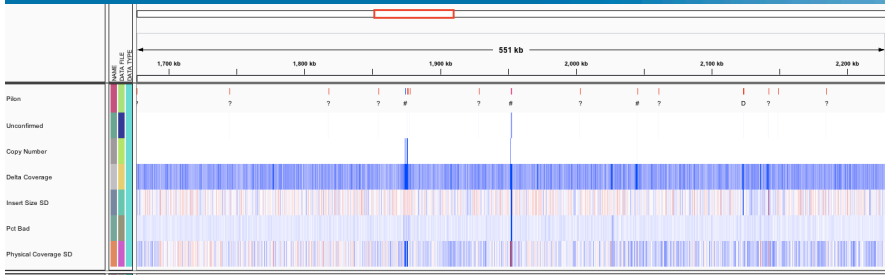
Coverage motif: rapid change to region of double (or more) coverage

- assembler included too few copies

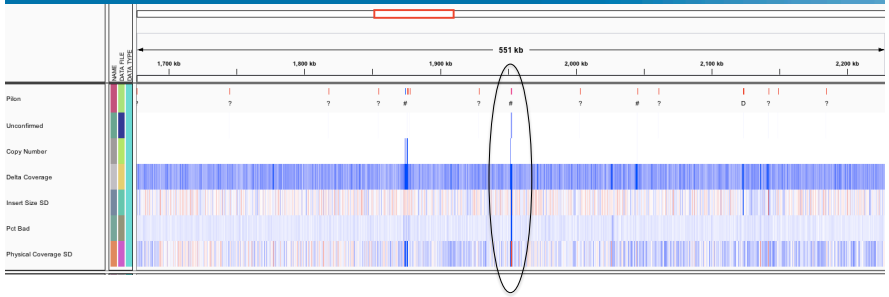
Pilon identifies collapsed repeats, but does not (yet) fix



Tracks Identify Potential Issues
500Kb View

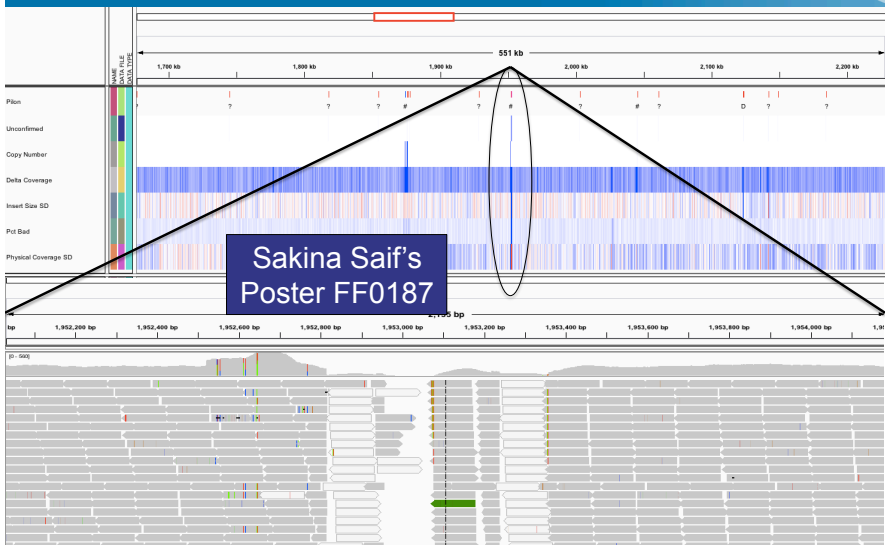


Tracks Identify Potential Issues
500Kb view of Pilon tracks

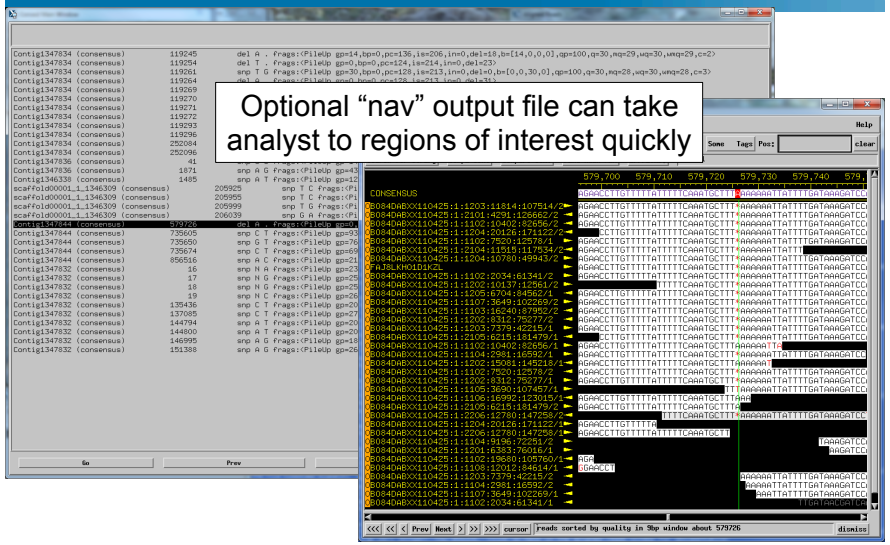


Tracks Identify Potential Issues
500Kb view zoomed in to 2Kb view

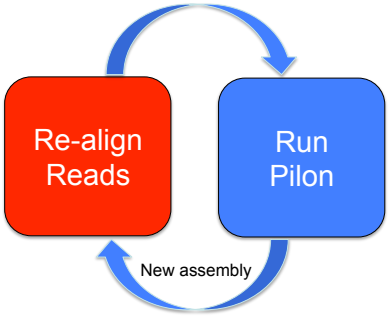
Tracks Identify Potential Issues
500Kb view zoomed in to 2Kb view



Consed Integration

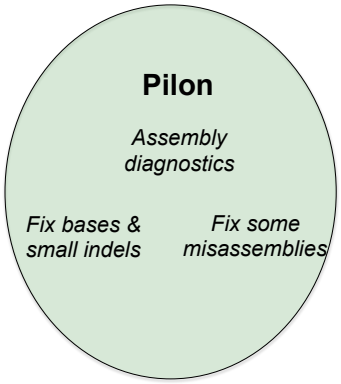


Iterative Improvement



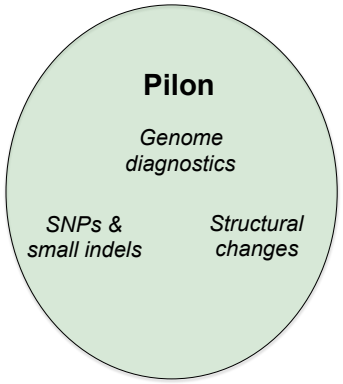
- Some issues might only be revealed by other fixes
- Data might not have the “reach” to identify or fully fix some things in one pass

Assembly Improvement



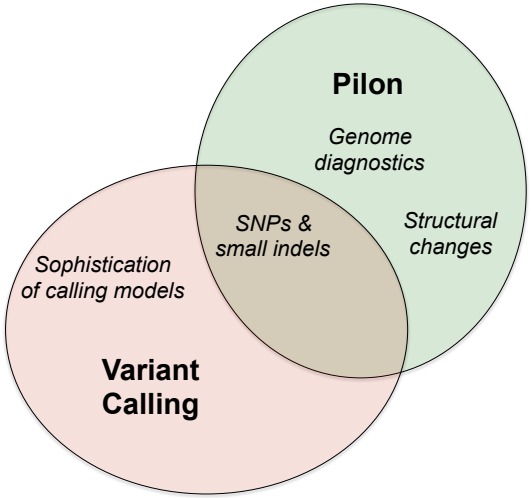
Pilon takes assembly and morphs it to conform to evidence in sequencing data

Assembly Improvement or …?

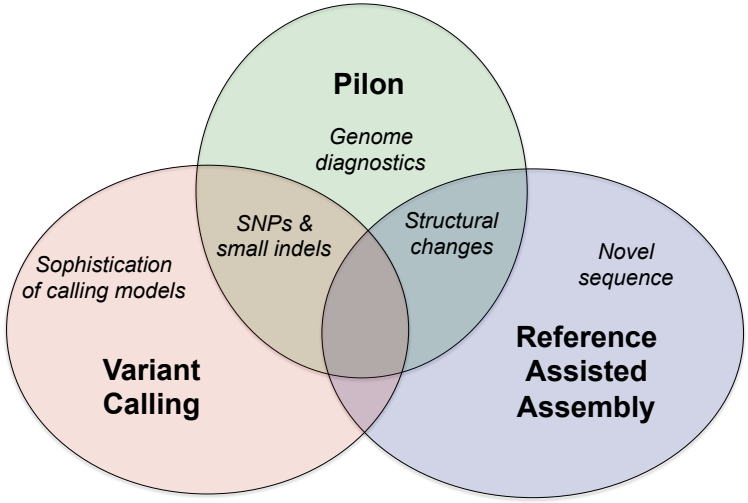


Pilon takes ~~assembly~~ reference genome and morphs it to conform to evidence in sequencing data

Assembly Improvement or …?



Assembly Improvement or …?



Streptococcus pneumoniae TIGR4 Improving a Finished Reference

- ALLPATHS-LG hybrid assembly development
  - Shuangye Yin's talk FF0170
  - Measure accuracy *vs* finished references
  - Realized that often assembler was "right", not reference

Streptococcus pneumoniae TIGR4 Improving a Finished Reference

- ALLPATHS-LG hybrid assembly development
  - Shuangye Yin's talk FF0170
  - Measure accuracy *vs* finished references
  - Realized that often assembler was "right", not reference

- Pilon called the same set of 57 SNP and small indel events
  - Plus one truly polymorphic base
- Pilon also identified two collapsed repeats in the reference
  - ALLPATHS-LG had inserted a tandem copy of each

Schizosaccharomyces octosporus Improving Genes using RNA-Seq

- "Great" hybrid assembly using Sanger and 454 data
  - 19 contigs, 11 scaffolds, scaffold N50 of 1.5Mb
- …but genes were riddled with frame shifts

Schizosaccharomyces octosporus Improving Genes using RNA-Seq

- "Great" hybrid assembly using Sanger and 454 data
  - 19 contigs, 11 scaffolds, scaffold N50 of 1.5Mb
- …but genes were riddled with frame shifts

- Aligned Illumina RNA-Seq data to assembly
- Pilon corrected 127 of 164 genes with frame shifts

Pilon Implementation

- Portable: runs in Java VM
- Written in the Scala programming language
- Uses open source picard/samtools Java libraries
- Fast (~5 minutes on bacterial genome with >200x coverage)

- Availability:
  - still under development
  - will be released under open source license this summer

bruce@broadinstitute.org

Final Thoughts

- There is often more in the data than assemblers can utilize

- Pilon is part of our larger vision for advanced assembly analysis
  - Sean Sykes' presentation next!

Acknowledgments

Genome Assembly & Analysis

Sarah Young
Sean Sykes
Sakina Saif
Amr Abouelleil
Harindra Arachchi
Aaron Berlin
Sante Gnerre
Alma Imamovic
Jerome Naylor
Terry Shea

Computational Research & Development

David Jaffe
Iain MacCallum
ALLPATHS-LG team

Finishing

Mike FitzGerald

Integrative Genomics Viewer

Jim Robinson, et al.

Genome Sequencing & Analysis Program

Chad Nusbaum

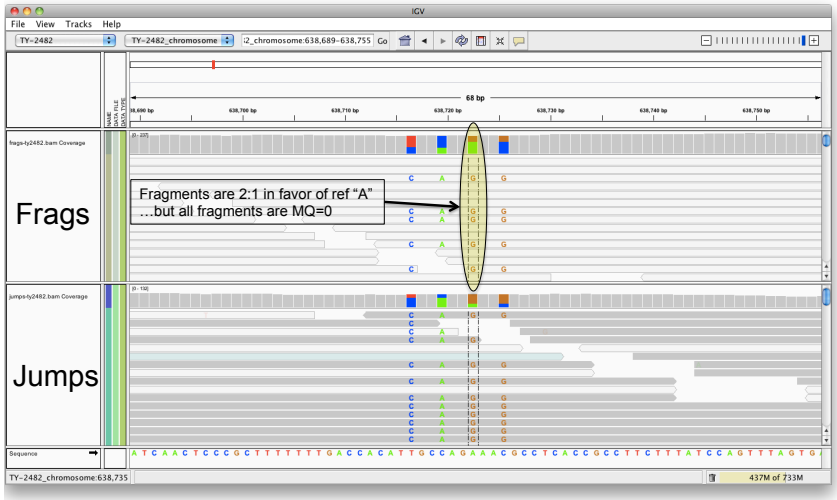
Funding

National Human Genome Research Institute

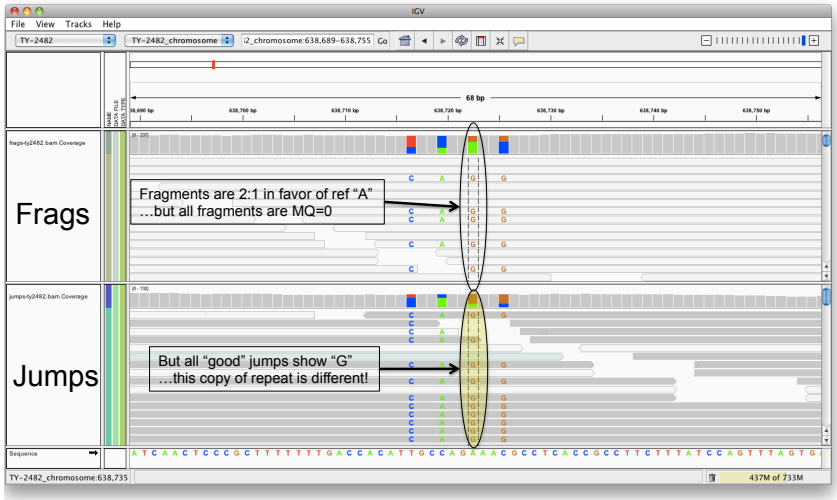
National Institute of Allergy and Infectious Diseases

Extra Slides

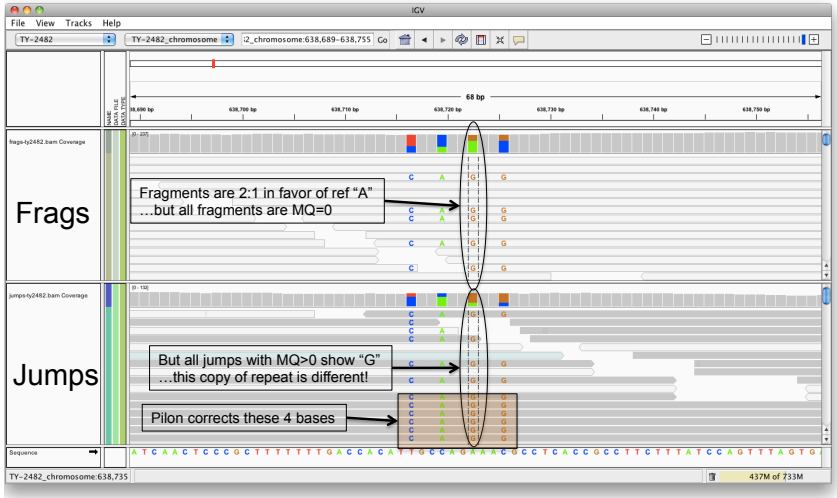
Mixed Repeat: Resolving Ambiguities Using Jumps



Mixed Repeat: Resolving Ambiguities Using Jumps



Mixed Repeat: Resolving Ambiguities Using Jumps



GAEMR

- Genome Assembly Evaluation, Metrics, and Reporting
- Suite of assembly analysis tools
- Any assembler, any input data
- Releasing to community this summer!

Visualization Tracks

- "Good" and "bad" pair sequence coverage
- Delta coverage (how fast sequence coverage is changing)
- Good pair physical coverage
- Mean insert size of good pairs in physical coverage
- Percentage of bad pairs
- Mapping quality weighted by sequence base quality
- Sequence base quality weighted by mapping quality
- GC percentage
- Corrected bases, unconfirmed bases
- High-level summary feature track

Information Gleaned from Pileups

- Pilon first looks at pileups of aligned read data
  - "Knows" about paired-end fragments and jumps (mate pair)
- Utilizes notion of "good" and "bad" pairs
  - Good alignments and "proper pair" identification important
  - "Good": plausible orientation and separation
- Computes lots of summary metrics at every locus in assembly:
  - Physical & sequence coverages
  - Counts and quality score sums of alleles
  - Counts of implied deletions and insertions at locus
  - Insert sizes, …

Planned Enhancements

- Clone collapsed repeats and find homes for copies (e.g., tandem, in gaps)
- Support for non-paired data (e.g., PacBio)
- Circularity detection (bacterial chromosomes, plasmids)
- Improve scaffolding diagnostics, implement fixing
- Assembly of novel sequence (unmapped reads)

Spectrum of Genome Improvement

Assembly algorithms have improved

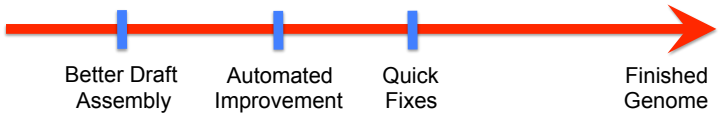
- Even with inexpensive sequencing, draft assemblies are often quite good (>100Kb contigs, >1Mb scaffolds, >Q40)
- With the right combinations of data, near-perfect *de novo* bacterial genomes are possible (latest ALLPATHS-LG paper)

Spectrum of Genome Improvement

Automated improvement

- Assemblers still make mistakes
- *There may be more in the data than the assembler could see*
- How can we mine the data to make the assembly better?

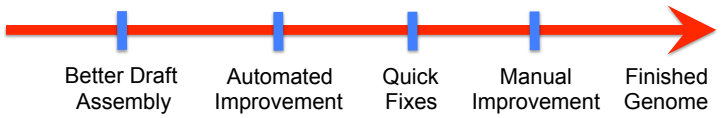
Spectrum of Genome Improvement



Automated analysis can pinpoint potential issues

- Lower overhead for manual fixing

Spectrum of Genome Improvement



Devote additional resources as appropriate to the project:

- Add additional data types to generate better assemblies
- Additional manual investigation and improvement

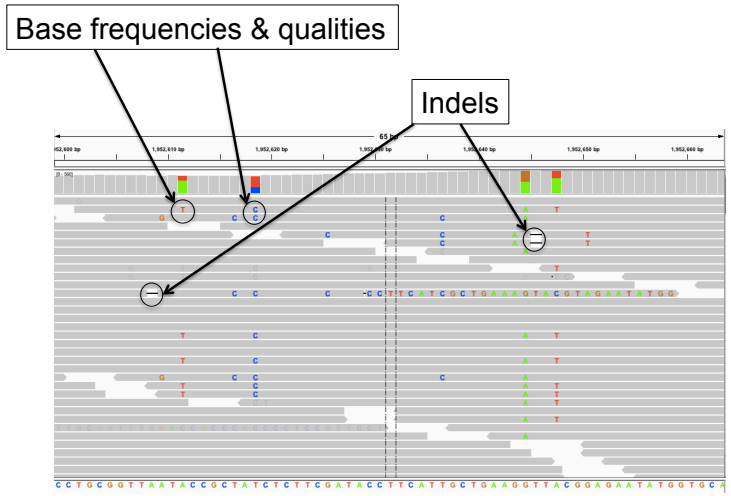
Traditional Path to Better Genomes: Finishing

Result: very good (Q50+, sometimes Q60) genomes

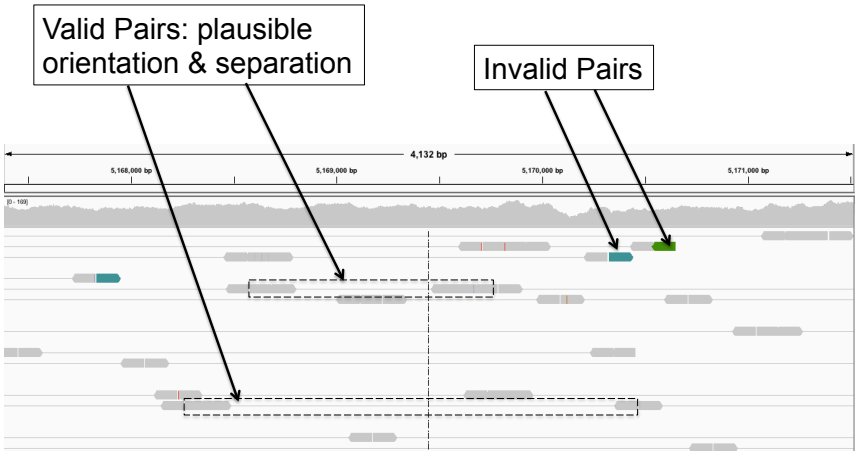
Escherichia coli O104:H4 (European Outbreak)

- C227-11 (from German outbreak) vs TY-2482 (BGI reference)
  - 117 base changes (95 chromosomal, 2 pTY1, 20 pTY2)
  - 6 small insertions
  - 1 larger insertion (53b)
  - 7 small deletions
  - 18 larger deletions (36-144b)
  - Other potential structural differences

Information Gleaned from Pileups Bases & Small Indels



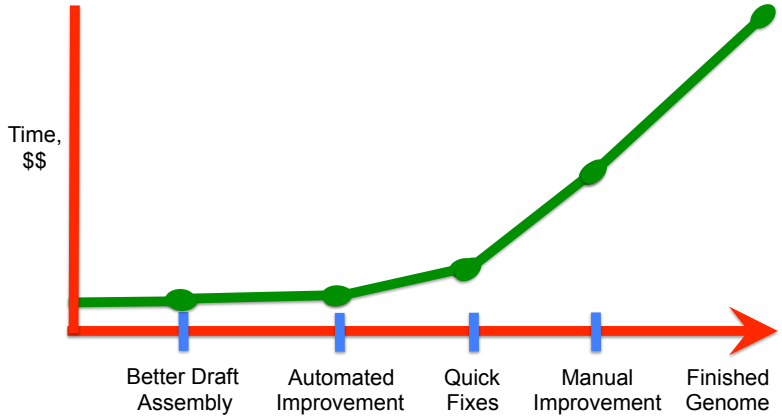
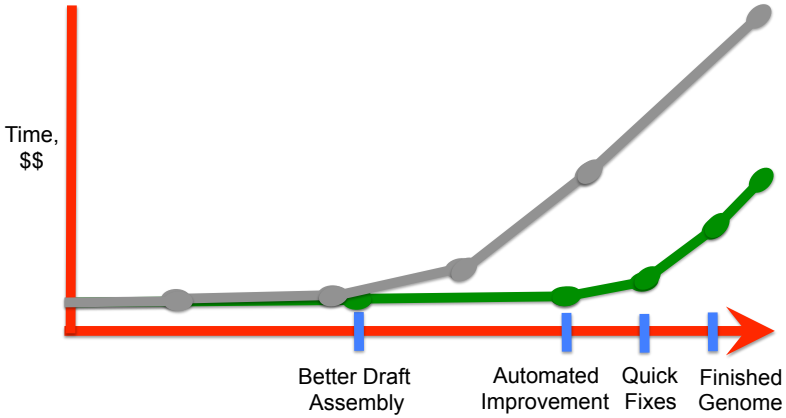
Information Gleaned from Pileups
Paired Alignments



Information Gleaned from Pileups
Per-Base Metrics

- Sequence coverage
- Physical coverage
- Local mean insert size

Spectrum of Genome Improvement



Spectrum of Genome Improvement