Next Generation Sequencing – The Role of New Sequence Technologies in Shaping the Future of Veterinary Science

Hosted by the RCVS Charitable Trust

RCVS charitable trust

Problems and pitfalls. Closing genomes, informatics, and errors

Ian Goodhead; Alistair Darby
and Neil Hall

CENTRE for GENOMIC RESEARCH
CGR

Conclusions

- Don’t be scared off!
- Probably only applicable to smaller genomes
- All of the problems discussed are tractable
- Take home message:
  - Talk to us BEFORE you design your experiment

Outline

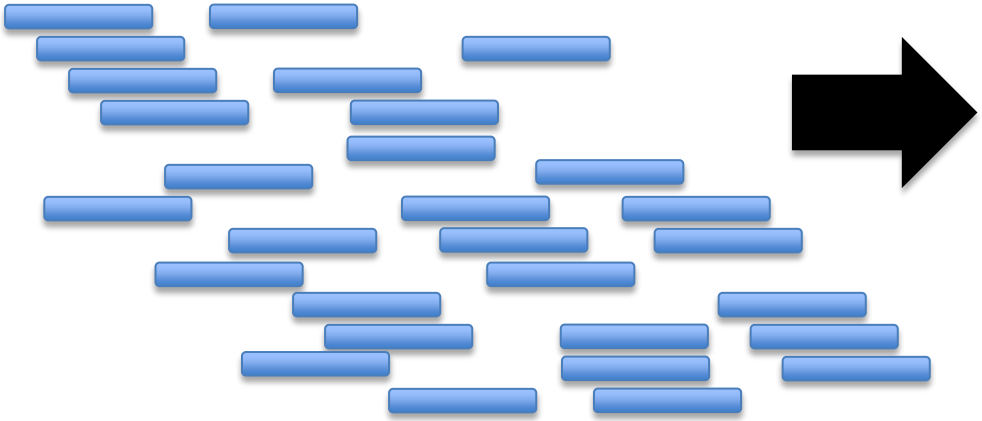
Basic concepts in genome sequencing and assembly

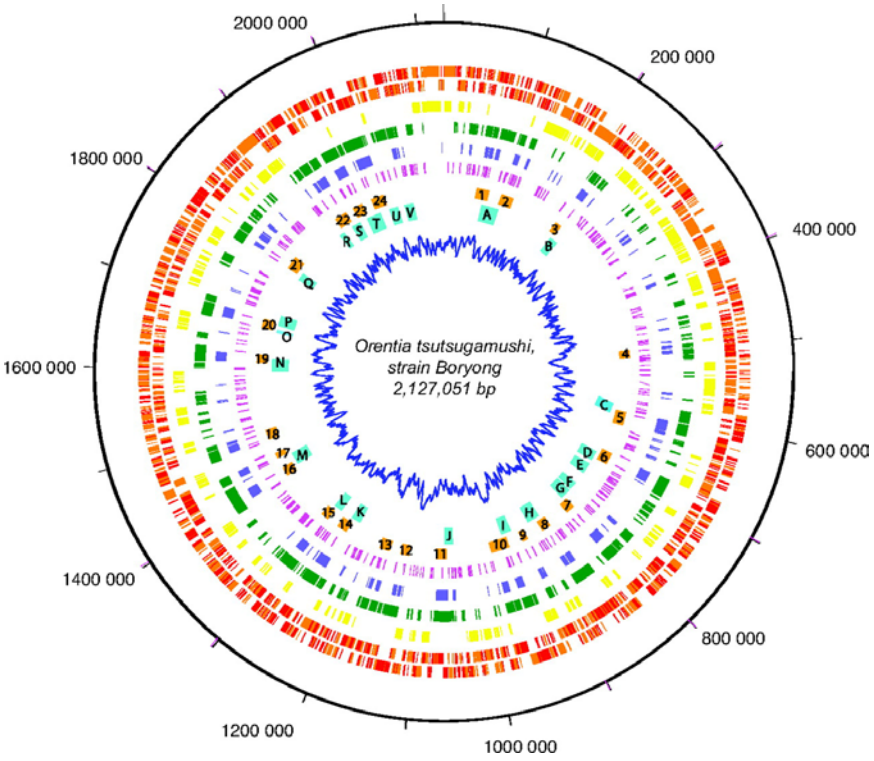
Alignment and assembly of next-generation sequencing data

Sources of error in assemblies

- Repeats
- Sequencing errors

How do you assemble a genome?





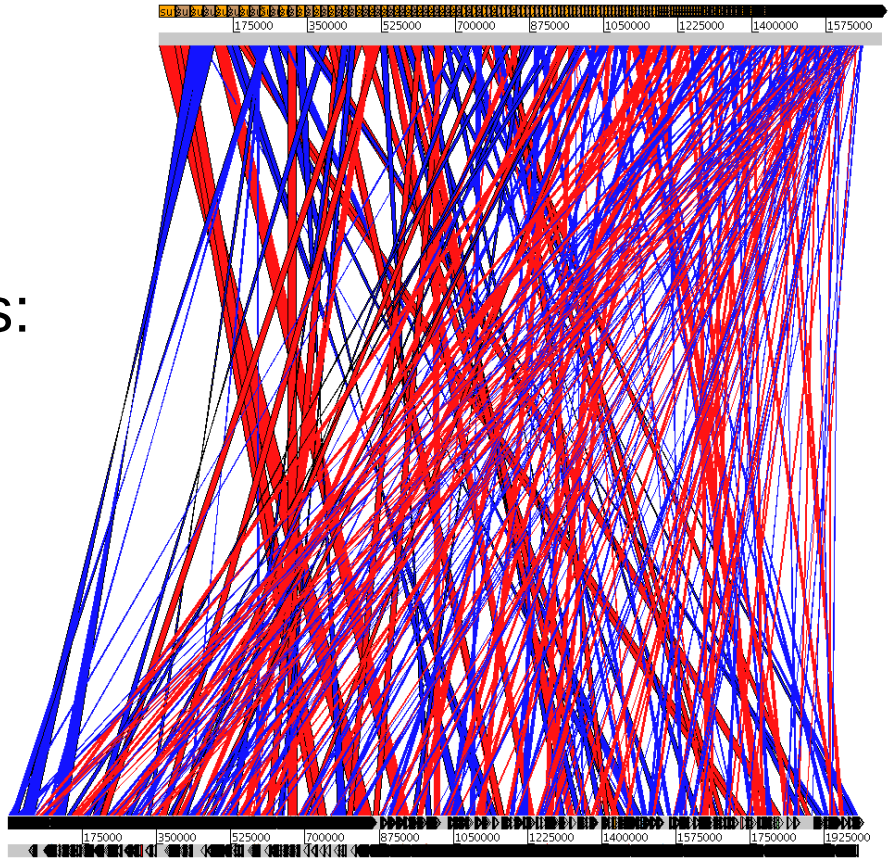
Sequencing Reads

Genome

To align or to assemble?

- Mapping / Alignment
  - Useful if you have a reference
  - Closely related
  - High quality (i.e. “finished”)
- Useful for various applications:
  - RNA-seq
  - ChIP-seq
  - Methyl-seq
  - CNV-seq
  - SNP identification

Raw Sequence



Reference Sequence

Which alignment algorithm should I use?

BFAST - Blat-like Fast Accurate Search Tool. Written by Nils Homer, Stanley F. Nelson and Barry Merriman at UCLA.
Bowtie - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads
BWA - Heng Lee's BWT Alignment program - a progression from Maq. BWA is a fast light-weighted tool that aligns short sequences to a sequence
ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author
Exonerate - Various forms of pairwise alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Sla
GenomeMapper - GenomeMapper is a short read mapping tool designed for accurate read alignments. It quickly aligns millions of reads either w
GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Ger
gnumap - The Genomic Next-generation Universal MAPper (gnumap) is a program designed to accurately map sequence data obtained from nex
MAQ - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina with preliminary functions to handle A
MOSAIK - MOSAIK produces gapped alignments using the Smith-Waterman algorithm. Features a number of support tools. Support for Roche FL
MrFAST and MrsFAST - mrFAST & mrsFAST are designed to map short reads generated with the Illumina platform to reference genome assemblie
MUMmer - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing
Novocraft - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Can support Bis-Se
PASS - It supports Illumina, SOLiD and Roche-FLX data formats and allows the user to modulate very finely the sensitivity of the alignments. Spac
RMAP - Assembles 20 - 64 bp Illumina reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bio
SeqMap - Supports up to 5 or more bp mismatches/INDELs. Highly tunable. Written by Hui Jiang from the Wong lab at Stanford. Builds available
SHRiMP - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Mi
Slider- An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignn
SOAP - SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto
SSAHA - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databas
SOCS - Aligns SOLiD data. SOCS is built on an iterative variation of the Rabin-Karp string search algorithm, which uses hashing to reduce the set o
SWIFT - The SWIFT suit is a software collection for fast index-based sequence comparison. It contains: SWIFT — fast local alignment search, guar
SXOligoSearch - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Re
Vmatch - A versatile software tool for efficiently solving large scale sequence matching tasks. Vmatch subsumes the software tool REPuter, but is
Zoom - ZOOM (Zillions Of Oligos Mapped) is designed to map millions of short reads, emerged by next-generation sequencing technology, back t

http://seqanswers.com/forums/showthread.php?t=43

Assembling ‘short’ NGS reads

- Required if no reference sequence available
- Typically uses very high coverage of short read data (eg. 50 – 150bp reads)
  - Sometimes interspersed with longer reads
- Useful for various applications:
  - *de novo* genomics
  - *de novo* transcriptomics
  - CNV-seq
  - SNP identification
- Requires some heavy-duty computing

Which assembly algorithm should I use?

* ABySS - Assembly By Short Sequences. ABySS is a de novo sequence assembler that is design
* ALLPATHS - ALLPATHS: De novo assembly of whole-genome shotgun microreads. ALLPATHS is
* Edena - Edena (Exact DE Novo Assembler) is an assembler dedicated to process the millions
* EULER-SR - Short read *de novo* assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCS
* MIRA2 - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo
* SEQAN - A Consistency-based Consensus Algorithm for De Novo and Reference-guided Seque
* SHARCGS - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and
* SSAKE - The Short Sequence Assembly by K-mer search and 3' read Extension (SSAKE) is a ge
* SOAPdenovo - Part of the SOAP suite. See above.
* VCAKE - De novo assembly of short reads with robust error correction. An improvement on e
* Velvet - Velvet is a de novo genomic assembler specially designed for short read sequencing

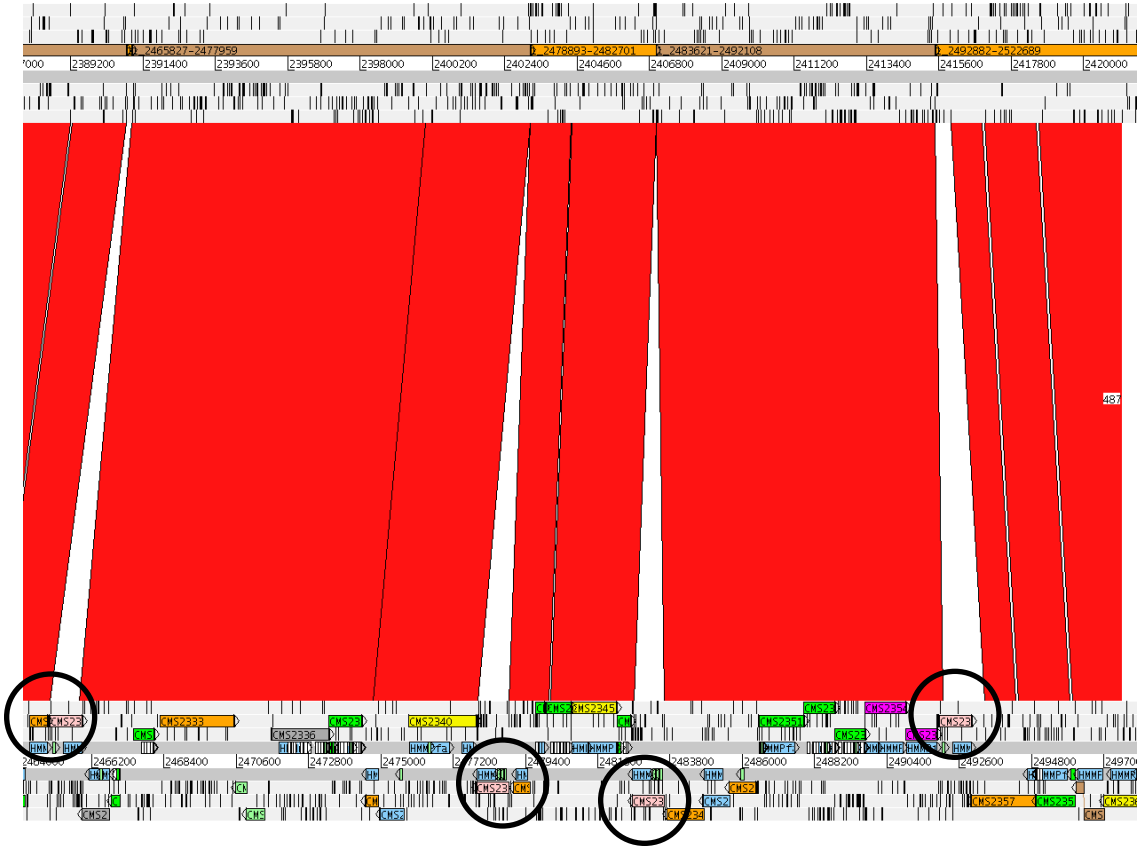
http://seqanswers.com/forums/showthread.php?t=43

The problem of repeats

NGS Alignment

***De-novo* Assembly**

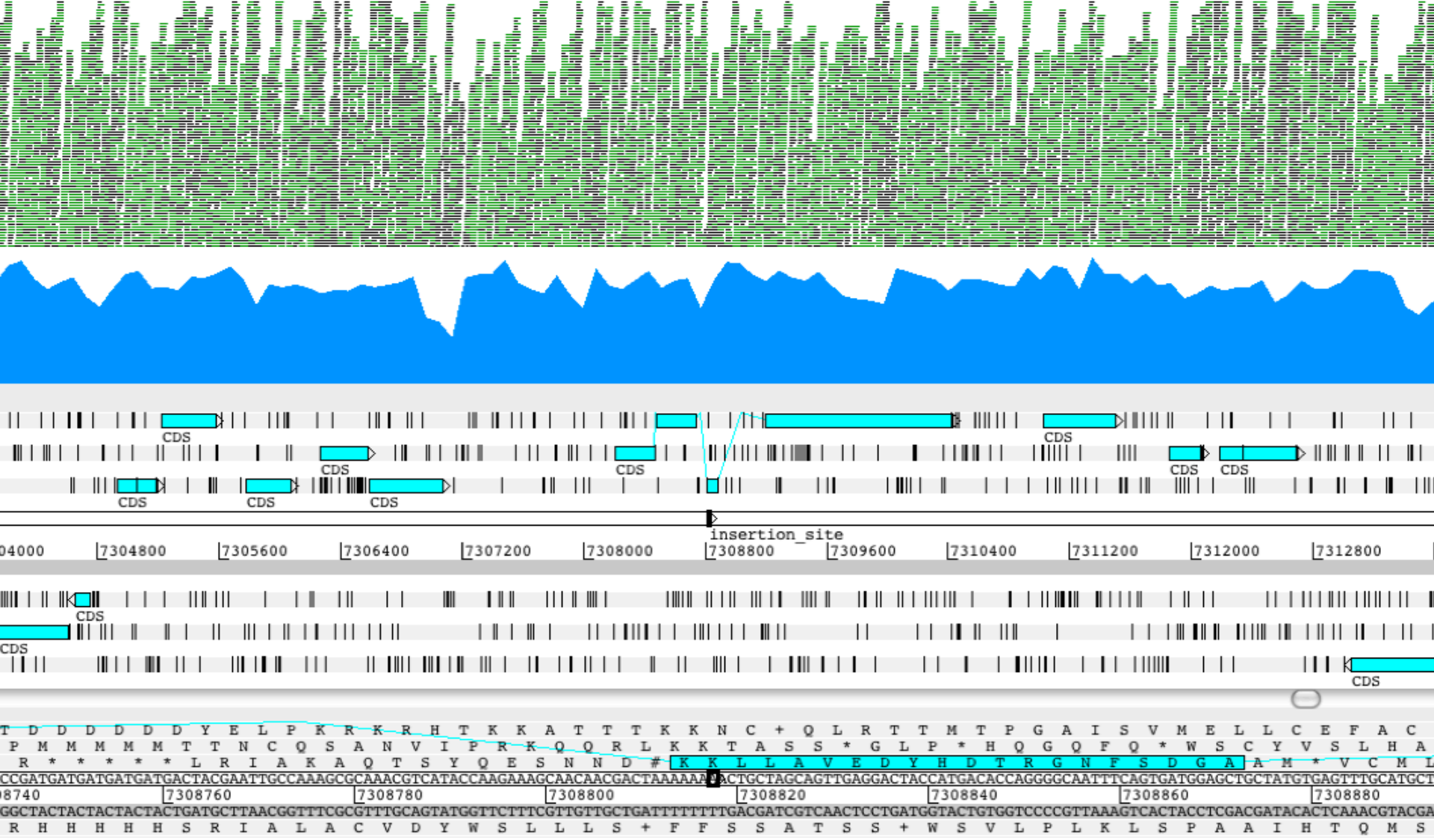
- **Assembled reads: 702,562**
- **Total number of contigs: 7,261**



Repeat sequences

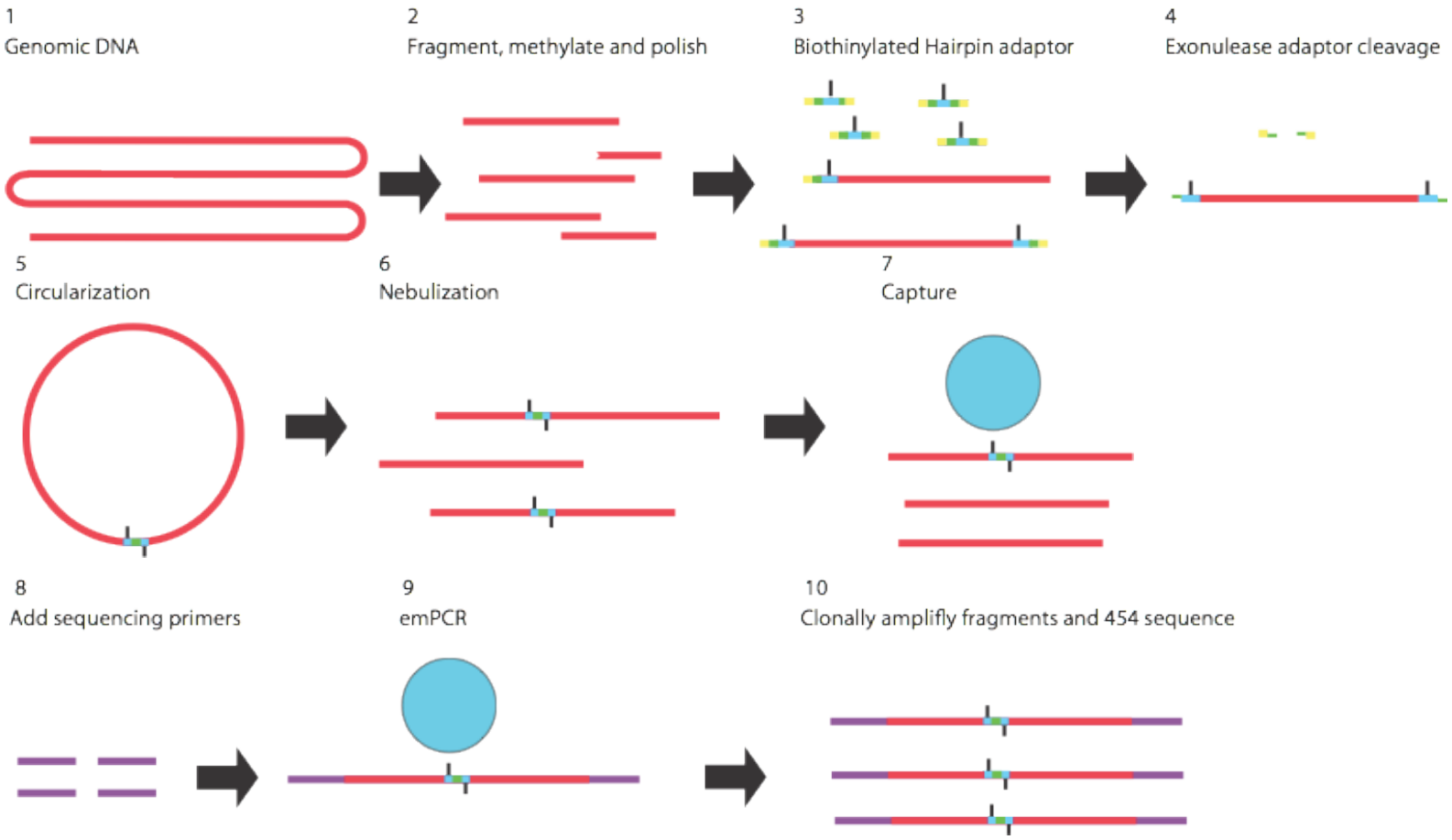
Annotated Reference Sequence

- **102 of the gaps in mapped sequence due to 975bp-long IS elements**
- **Equates to ~3% genome**



Homopolymer Errors

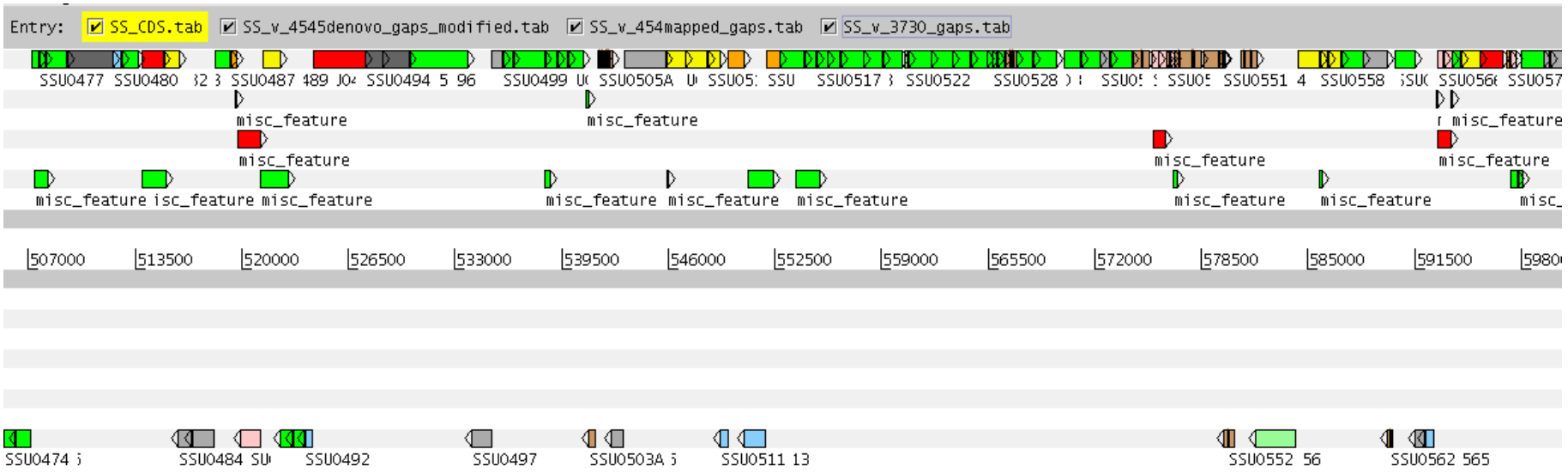
454 Mate Pairs



- Insert size 3 kb, 8 kb & 20 kb

Complementary Technologies?

The positions of the gaps differ between the two technologies:

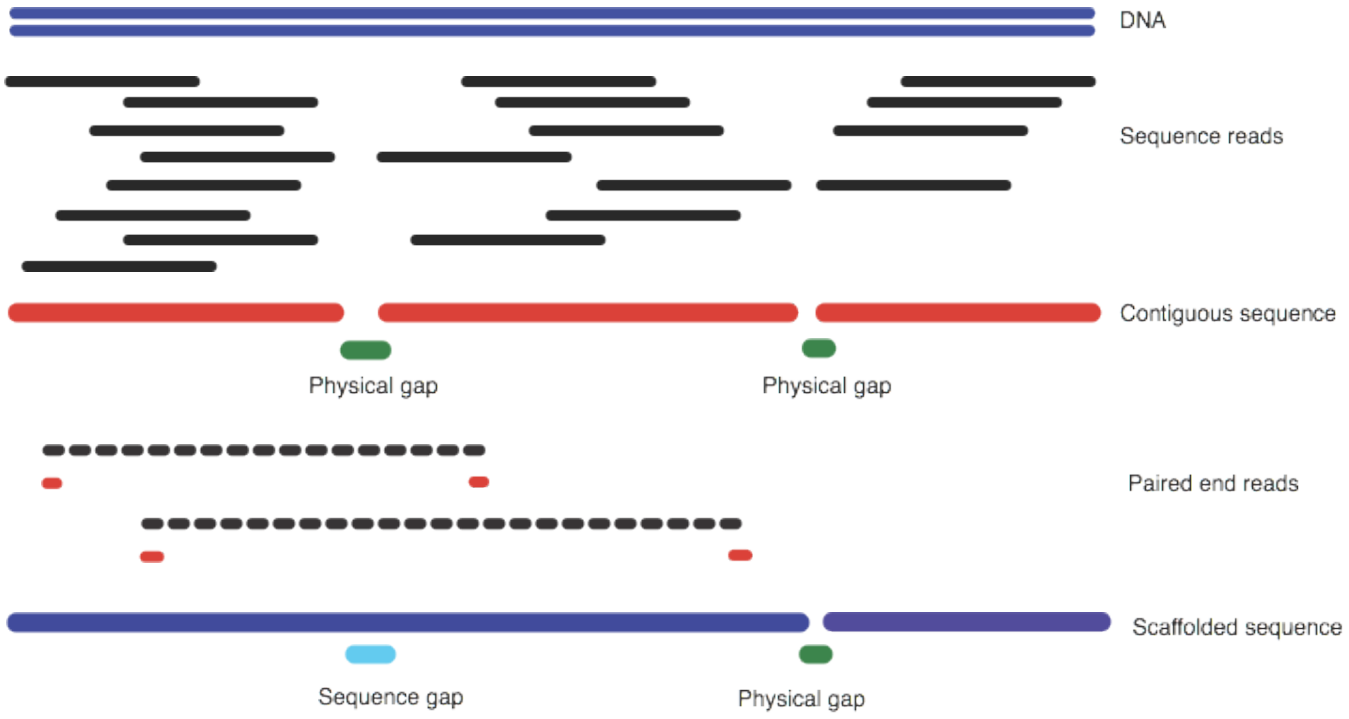


Schematic showing the positions of the gaps present in the different assemblies.

Shotgun Sequencing

1. Generate reads
2. Find overlaps
3. Identify contigs

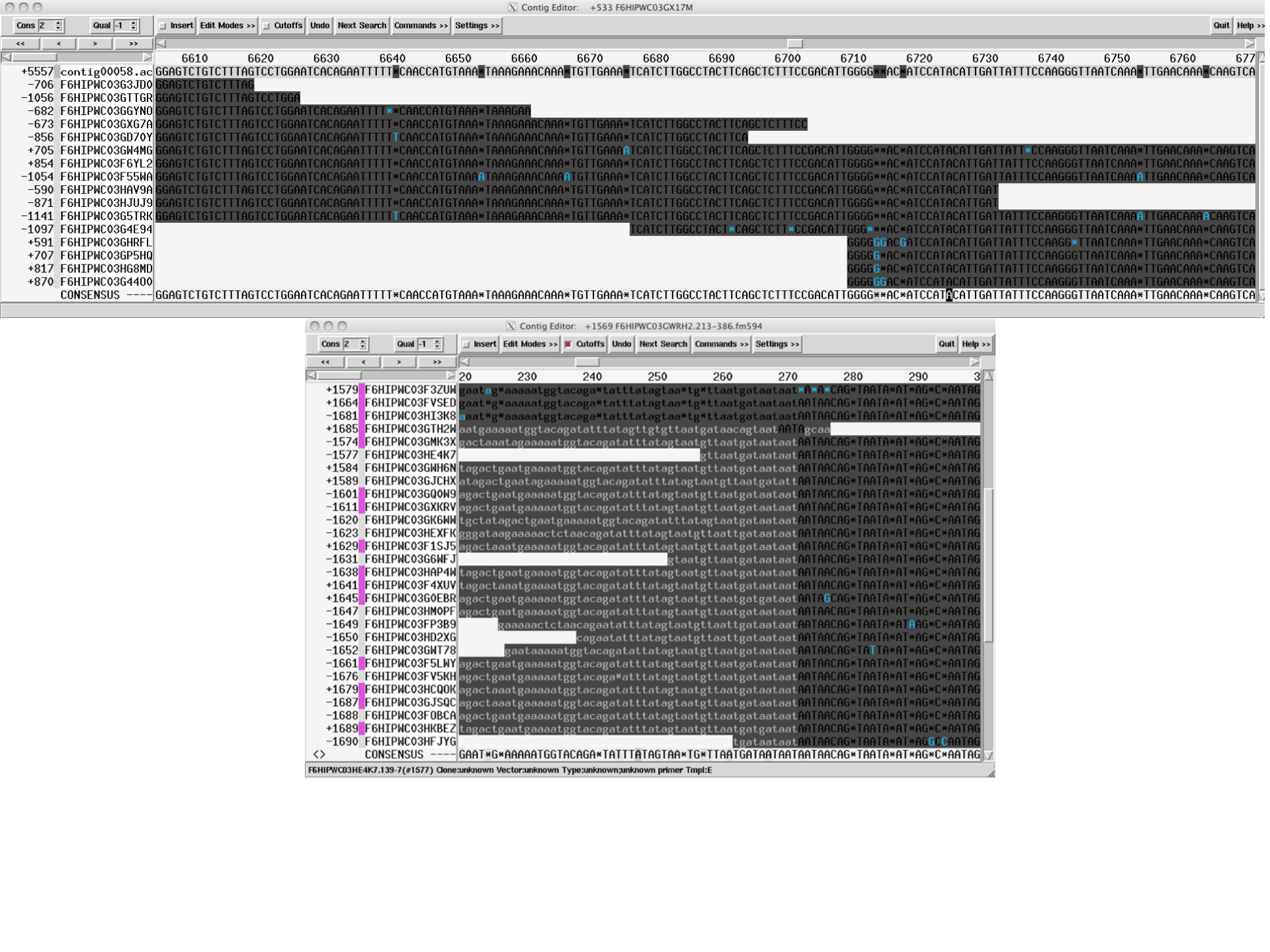
3. Scaffold Contigs (long range data)



How to close gaps?
Closing NGS generated genomes

- SFFfile split 454 MID tags
  - Multiple genomes in a single run may be preferable?
- Assembly with Newbler
- Convert ACE format to GAP
- Edit in GAP

With thanks to Alistair Darby


Contig Editor: +533 F6HIPWC03GX17M
Cons 2
Qual -1
Insert
Edit Modes >>
Cutoffs
Undo
Next Search
Commands >>
Settings >>
Quit
Help >>
6610 6620 6630 6640 6650 6660 6670 6680 6690 6700 6710 6720 6730 6740 6750 6760 677
+5557 contig00058.ac
-706 F6HIPWC03G3JD0
-1056 F6HIPWC03GTTGR
-682 F6HIPWC03GGYNO
-673 F6HIPWC03GXG7A
-856 F6HIPWC03GD70Y
+705 F6HIPWC03GW4MG
+854 F6HIPWC03F6YL2
-1054 F6HIPWC03F55WA
-590 F6HIPWC03HAV9A
-871 F6HIPWC03HJUJ9
-1141 F6HIPWC03G5TRK
-1097 F6HIPWC03G4E94
+591 F6HIPWC03GHRFL
+707 F6HIPWC03GP5HQ
+817 F6HIPWC03HG8MD
+870 F6HIPWC03G44O0
CONSENSUS ----
Contig Editor: +1569 F6HIPWC03GWRH2.213-386.fm594
Cons 2
Qual -1
Insert
Edit Modes >>
Cutoffs
Undo
Next Search
Commands >>
Settings >>
Quit
Help >>
20 230 240 250 260 270 280 290 3
+1579 F6HIPWC03F3ZUW
+1664 F6HIPWC03FVSED
-1681 F6HIPWC03HI3K8
+1685 F6HIPWC03GTH2W
-1574 F6HIPWC03GMK3X
-1577 F6HIPWC03HE4K7
+1584 F6HIPWC03GWH6N
+1589 F6HIPWC03GJCHX
-1601 F6HIPWC03GQOW9
-1611 F6HIPWC03GXKRV
-1620 F6HIPWC03GK6WW
-1623 F6HIPWC03HEXFK
+1629 F6HIPWC03F1SJ5
-1631 F6HIPWC03G6WFJ
-1638 F6HIPWC03HAP4W
+1641 F6HIPWC03F4XUV
+1645 F6HIPWC03G0EBR
-1647 F6HIPWC03HMOPF
-1649 F6HIPWC03FP3B9
-1650 F6HIPWC03HD2XG
-1652 F6HIPWC03GWT78
-1661 F6HIPWC03F5LWY
-1676 F6HIPWC03FV5KH
+1679 F6HIPWC03HCQOK
-1687 F6HIPWC03GJSQC
-1688 F6HIPWC03FOBCA
+1689 F6HIPWC03HKBEZ
-1690 F6HIPWC03HFJYG
<> CONSENSUS ----
F6HIPWC03HE4K7.139-7(#1577) Clone:unknown Vector:unknown Type:unknown;unknown primer Tmpl:E

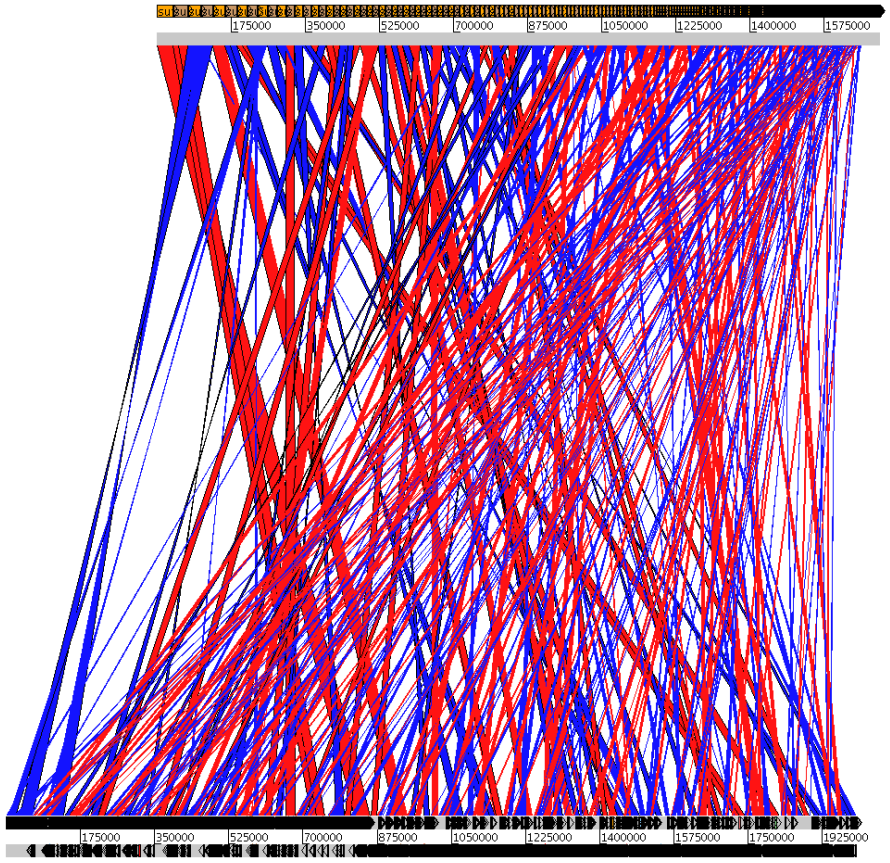
How to close gaps?
Closing NGS generated genomes

- SFFfile split 454 MID tags
  - Multiple genomes in a single run may be preferable?
- Assembly with Newbler
- Convert ACE format to GAP
- Edit in GAP
  - Examine ‘cutoff data’ due to high sequence depth
  - Design Primers and Sanger sequence gaps
  - Combine with ‘other’ NGS datasets for SNP calling

With thanks to Alistair Darby

Remember this?

Raw Sequence



Reference Sequence

Tools now available!

wellcome trust sanger institute

Search for Enter search here...

Home | Research | Scientific resources | Work & study | About us

Mouse | Zebrafish | Data | Software | Databases | Technologies | Talks & training

PAGIT - Post Assembly Genome Improvement Toolkit

Tools to generate automatically high quality sequence by ordering contigs, closing gaps, correcting sequence errors and transferring annotation.

With the advent of next generation sequencing a lot of effort was put into developing software for mapping or aligning short reads and performing genome assembly. For genome assembly the problem of generating a draft assembly (i.e. a set of unordered contigs) has now been very well addressed - but for users who need high quality assemblies for their analyses there are still unresolved issues: this is where PAGIT is used.

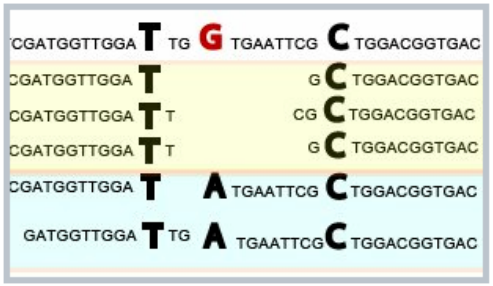
PAGIT addresses the need for software to generate high quality draft genomes. It is based on a series of programs that we developed:

1. ABACAS, that is able to contiguate contigs from a *de novo* assembly against a closely related reference.
2. IMAGE, an iterative approach for closing gaps in assembled genomes using mate pair information. It is able to close gaps left open by the assembler in a draft genome, even when using the same data sets as used by the original assembler.
3. iCORN, that enables errors in the consensus sequence to be corrected by iteratively mapping reads to the current assembly.
4. RATT, a tool to transfer the annotation from a reference genome, or an earlier assembly, onto the latest assembly.

PAGIT bundles these software and makes them more accessible for users.

We have a mailing list for announcements and questions. PAGIT mailing list.

Please note that we submitted a protocol paper that will explain each step of the toolkit. Extra care must be taken, when working with genome bigger than 200mb.



[Genome Research Limited]

Overview | Download | ABACAS | IMAGE | iCORN | RATT | FAQ

How to Get PAGIT:

We have bundled the four tools together with some other helpful scripts. In the download area they can be downloaded as precompiled versions, or pre-installed on a virtual machine.

Abacas (order and orient scaffolds)

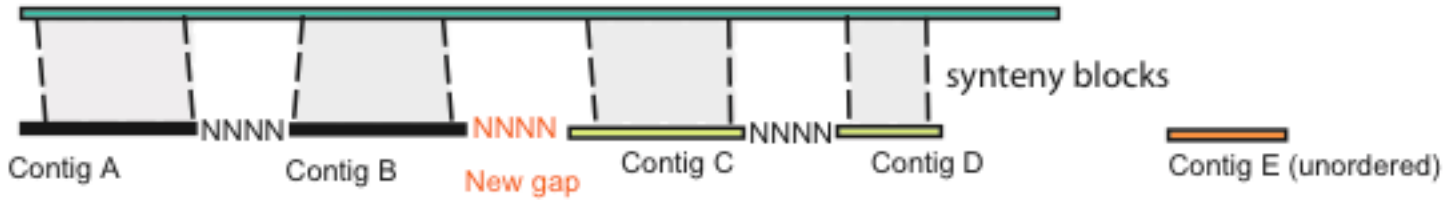
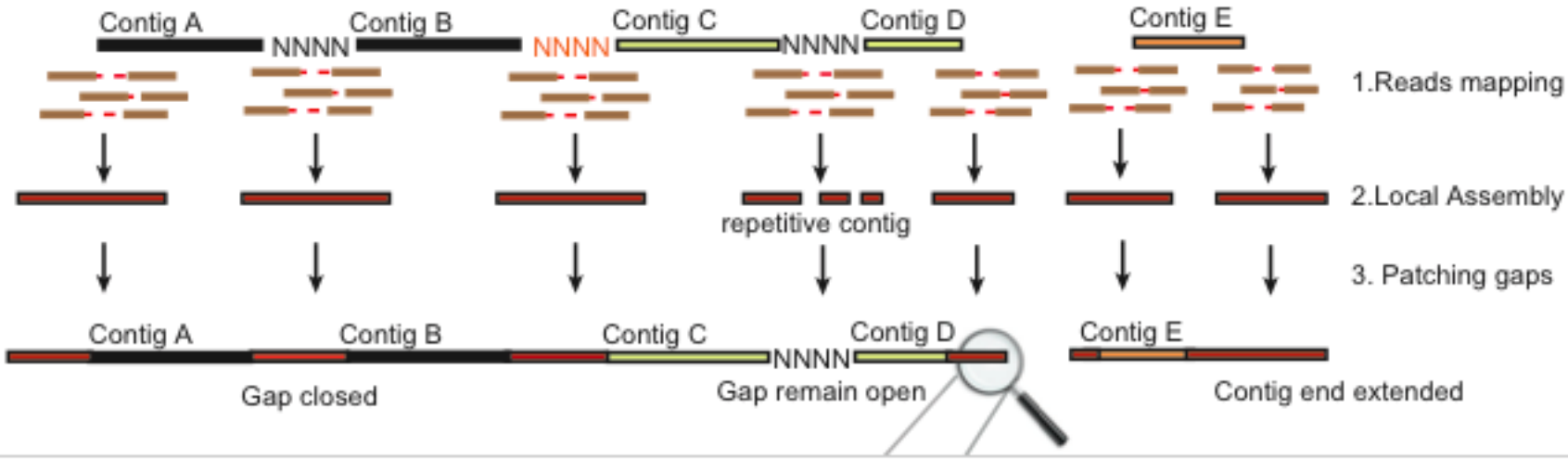
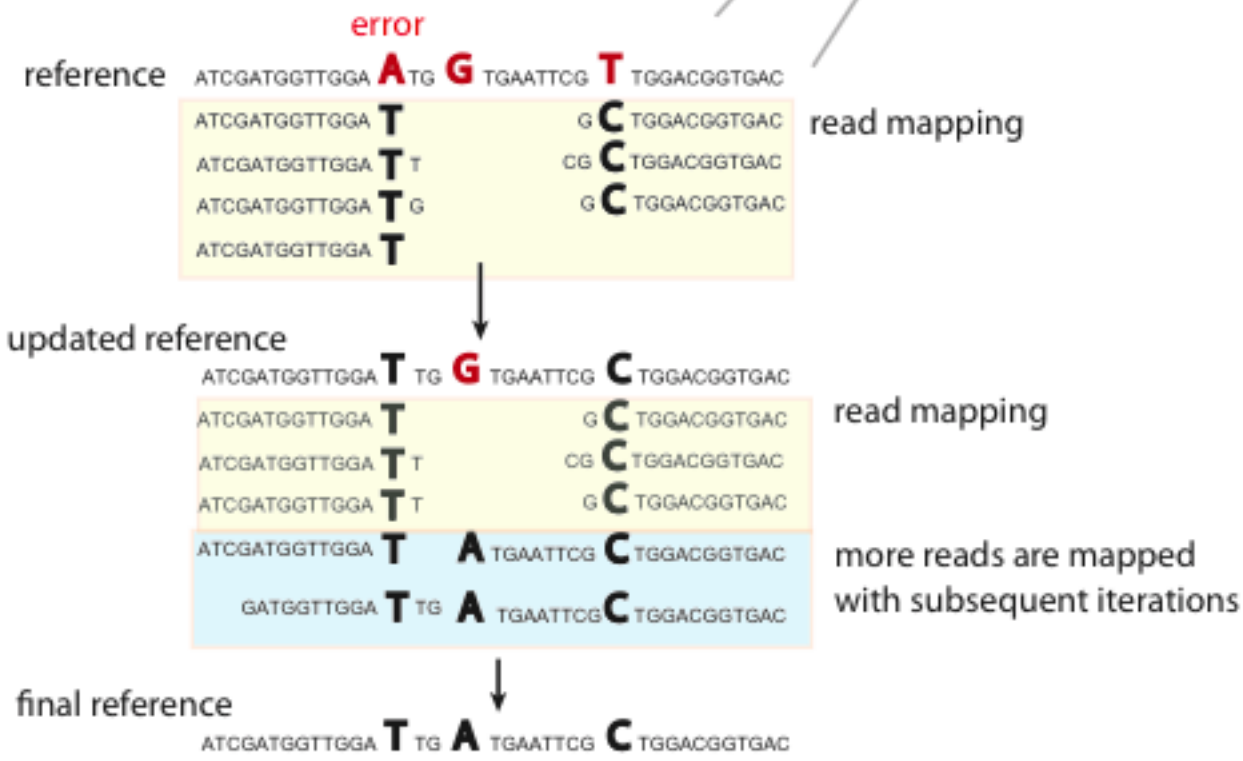


Image (gaps closing)



Icorn (correction at nucleotide level)



Conclusions

- To assemble or to align?
  - Largely down to whether you have an acceptable reference sequence
- Which analysis software to use
  - Publicly available? It’s free!
  - Commercial? Nice GUIs
- Don’t be scared off!
- All of the problems discussed are tractable

What haven’t I covered?

- Experimental design
- Sample preparation
  - “Rubbish in / Rubbish out”
- How do I extract my useful data?
  - Genome Annotation
  - SNP extraction
- How do I write my Nature paper?
- Take home message:
  - Talk to us BEFORE you design your experiment

Acknowledgements

- Centre for Genomic Research
- Wellcome Trust Sanger Institute
- www.seqanswers.com

  SEQanswers: An open access community for collaboratively decoding genomes

  Bioinformatics (2012)

  doi: 10.1093/bioinformatics/bts128

Ian.goodhead@liverpool.ac.uk

CENTRE for GENOMIC RESEARCH
CGR

Next Generation Sequencing – The Role of New Sequence Technologies in Shaping the Future of Veterinary Science

Hosted by the RCVS Charitable Trust

RCVS charitable trust