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

Hosted by the RCVS Charitable Trust
Problems and pitfalls. Closing genomes, informatics, and errors

Ian Goodhead; Alistair Darby
and Neil Hall
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
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 per second. Written by David Langmead and Steven Schatz.

**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 database. Written by Boyce Purcell.

**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 Slack and his group.

**GenomeMapper** - GenomeMapper is a short read mapping tool designed for accurate read alignments. It quickly aligns millions of reads either with a reference or without.

**GMAP** - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Gen.

**gnumap** - The Genomic Next-generation Universal MAPper (gnumap) is a program designed to accurately map sequence data obtained from next generation DNA sequencing technologies. Support for Roche FLX.

**MrFAST and MrsFAST** - MrFAST & mrsFAST are designed to map short reads generated with the Illumina platform to reference genome assemblies.

**MUMmer** - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing various alignment programs. Authors are Michael A. Myers and Hisayoshi Miyano.


**PASS** - It supports Illumina, SOLID and Roche-FLX data formats and allows the user to modulate very finely the sensitivity of the alignments. Space needed is linear in the number of reads.

**RMAP** - Assembles 20 - 64 bp Illumina reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinformatics).

**SHRiMP** - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael A. Myers and Hisayoshi Miyano.

**Slider** - An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment.


**SSAH** - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases.

**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 of possibilities.

**SWIFT** - The SWIFT suit is a software collection for fast index-based sequence comparison. It contains: SWIFT — fast local alignment search, guardSWIFT — fast local alignment search, guard.

**SXOligoSearch** - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Reference sequences.

**Vmatch** - A versatile software tool for efficiently solving large scale sequence matching tasks. Vmatch subsumes the software tool REPuter, but is faster.

**Zoom** - ZOOM (Zillions Of Oligos Mapped) is designed to map millions of short reads, emerging by next-generation sequencing technology, back to a reference sequence.

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 designed to work with highly divergent genomes.
* **ALLPATHS** - ALLPATHS: De novo assembly of whole-genome shotgun microreads. ALLPATHS is designed for assemblies from complex genomes.
* **Edena** - Edena (Exact DE Novo Assembler) is an assembler dedicated to process the millions of short reads.
* **EULER-SR** - Short read de novo assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD.
* **MIRA** - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assembly.
* **SHARC** - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Pevzner PA.
* **SSAKE** - The Short Sequence Assembly by K-mer search and 3' read Extension (SSAKE) is a general purpose de novo assembler.
* **SOAPdenovo** - Part of the SOAP suite. See above.
* **VCAKE** - De novo assembly of short reads with robust error correction. An improvement on earlier assemblies.
* **Velvet** - Velvet is a de novo genomic assembler specially designed for short read sequencing.

The problem of repeats

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

- 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

• SFF file 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
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?

Reference Sequence

Raw Sequence

Reference Sequence
Tools now available!

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.

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)

1. Reads mapping
2. Local Assembly
3. Patching gaps

Contig A
Contig B
Contig C
Contig D
Contig E

New gap
synteny blocks

Gap closed
Gap remain open
Contig end extended
Icorn (correction at nucleotide level)

Reference:
- ATCGATGGTTTGA
- ATCGATGGTTTGA
- ATCGATGGTTTGA
- ATCGATGGTTTGA

Error:
- ATG
- TGAATTCG
- TGGACGGTGAC

Read mapping:

Updated reference:
- ATCGATGGTTTGA
- ATCGATGGTTTGA
- ATCGATGGTTTGA
- ATCGATGGTTTGA

Read mapping:

More reads are mapped with subsequent iterations:

Final reference:
- ATCGATGGTTTGA
- TGAATTCG
- TGGACGGTGAC
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 GUls

• 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

lan.goodhead@liverpool.ac.uk
Next Generation Sequencing –
The Role of New Sequence Technologies in Shaping the
Future of Veterinary Science

Hosted by the RCVS Charitable Trust