NGS – what are we missing?

The data is there …. The standard methods are not always the best approach

James Cavalcoli, University of Michigan

Outline

- Two stories ... Examples of data which was missed

- Example 1: Genome reassembly

- Example 2: RNA-Seq

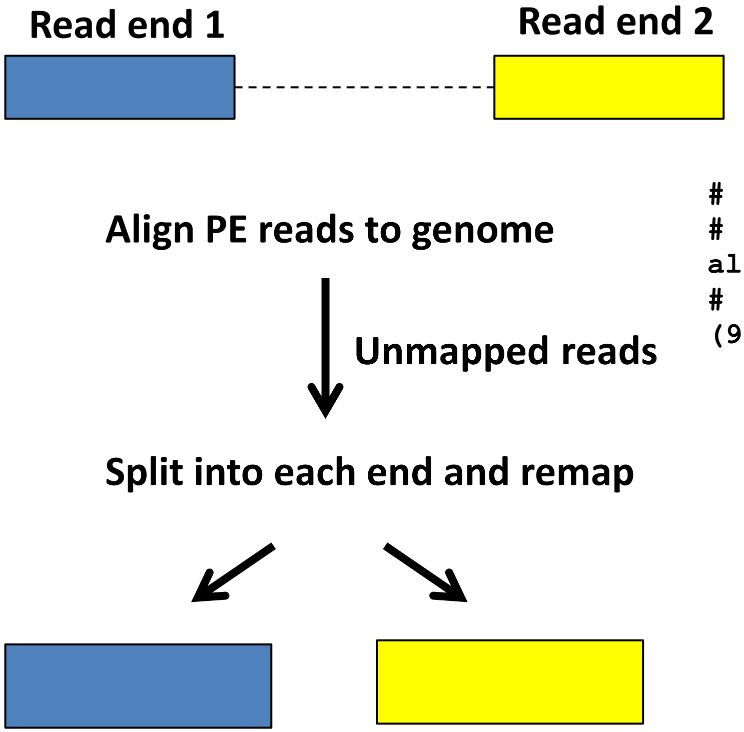
Genome Resequencing: Overview

- Mouse genome enriched for small region of interest
- Normal alignment and analysis have failed to find a mutation
- Some idea of repetitive element or structural variant as a cause.
- See Poster # **U37** for the details.

Genome Resequencing: Experimental Procedures

- Custom hybridization probes were generated to cover the genomic region (oligo array on a glass slide); 1.5 Mb region of Chr2.
- DNA from heterozygous Sd(+/-) mice was purified and hybridized to the probe sets to enrich for the region
- DNA was size selected (~300 bp) for library creation and Illumina adapters added to fragments.
- 36 cycle Paired-end sequencing was done using Illumina GAIIx yielding ~2.3x10^7 reads (median insert size was 299 nucleotides)

Initial Analysis



Read end 1

Read end 2

Align PE reads to genome

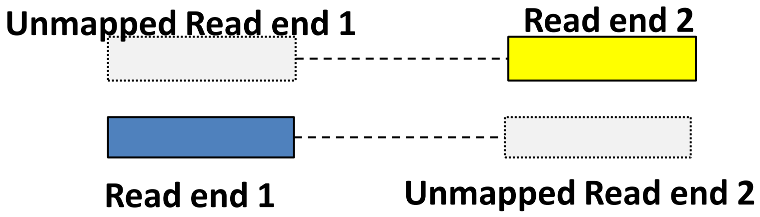
Unmapped reads

Split into each end and remap

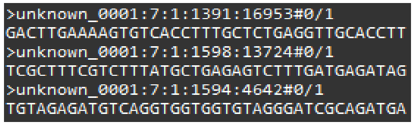
Paired ends aligned (Bowtie)

```
# reads processed: 23206800
# reads with at least one reported
alignment: 21054647 (90.73%)
# reads that failed to align: 2152153
(9.27%)
```

Identifying unmapped ends

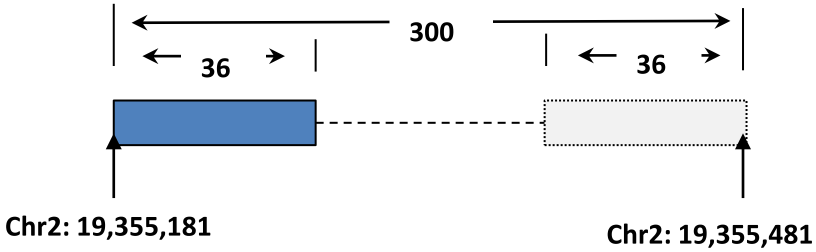


Retrieve sequence for unmapped read ends (perl script) convert to FASTA and use Exonerate to map to RepBase



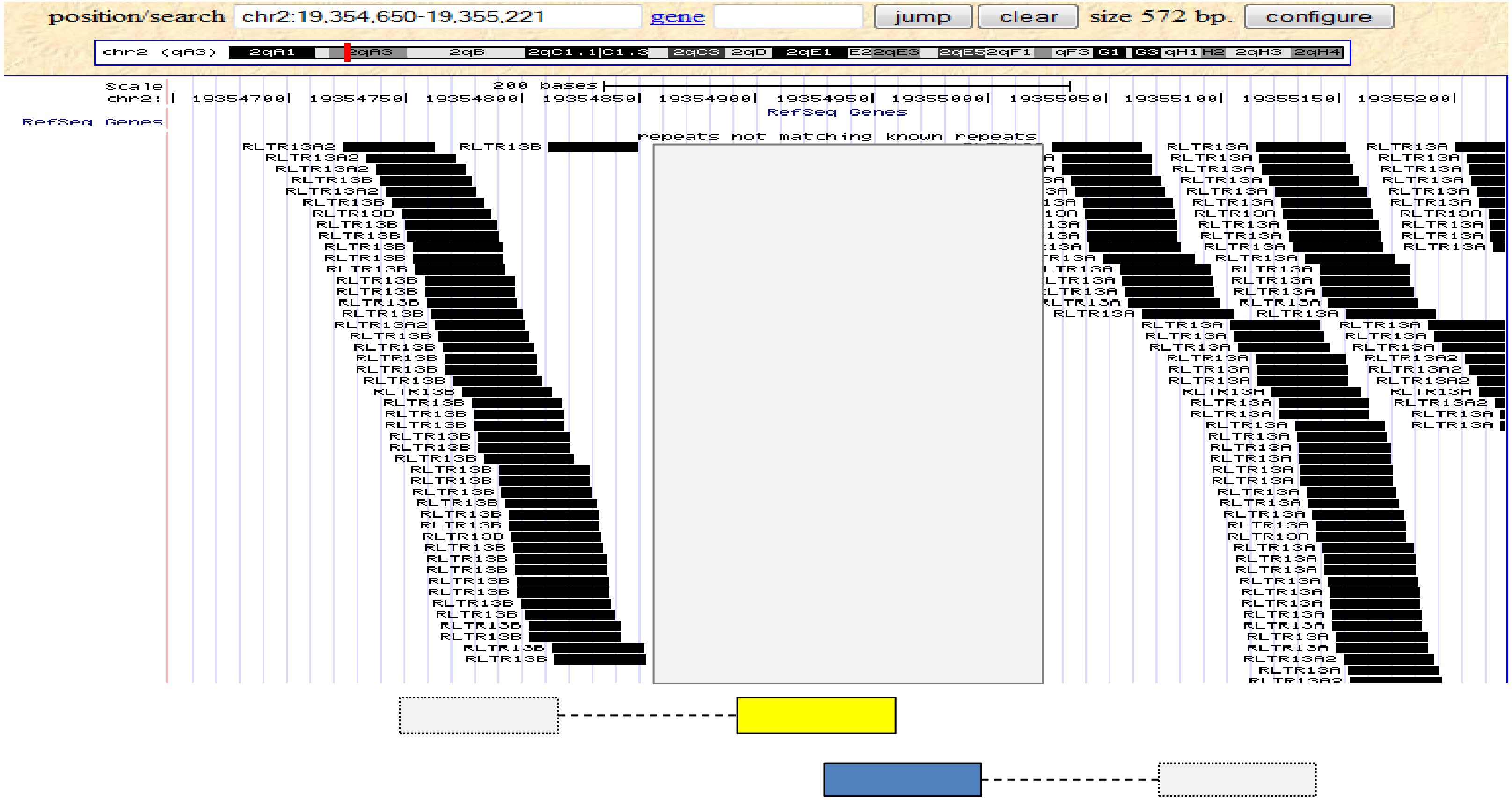
```
>unknown_0001:7:1:1391:16953#0/1
GACTTGAAAAGTGTCACCTTTGCTCTGAGGTTGCACCTT
>unknown_0001:7:1:1598:13724#0/1
TCGCTTTCGTCTTTATGCTGAGAGTCTTTGATGAGATAG
>unknown_0001:7:1:1594:4642#0/1
TGTAGAGATGTCAGGTGGTGGTGTAGGGATCGCAGATGA
```

Establish Approximate genomic location for repeat



Integrating and merging

- Using Perl scripts to integrate and merge genomic locations of:
  - Paired end naming information
  - mapped reads ends
  - predicted repeat locations
  - known repeat locations in mm9 (filtering for simple repeats).
- The result is a list of repeat elements at a new genomic location (with 5 or more reads coverage)



position/search chr2:19,354,650-19,355,221 gene jump clear size 572 bp. configure

Multiple reads representing the 2 LTRs of RLTR13 are found to flank a small genomic region (170 bp). To find the insertion site, this genomic segment and the LTR region were used to fish out original reads to find reads which overlap the genome and the LTR.

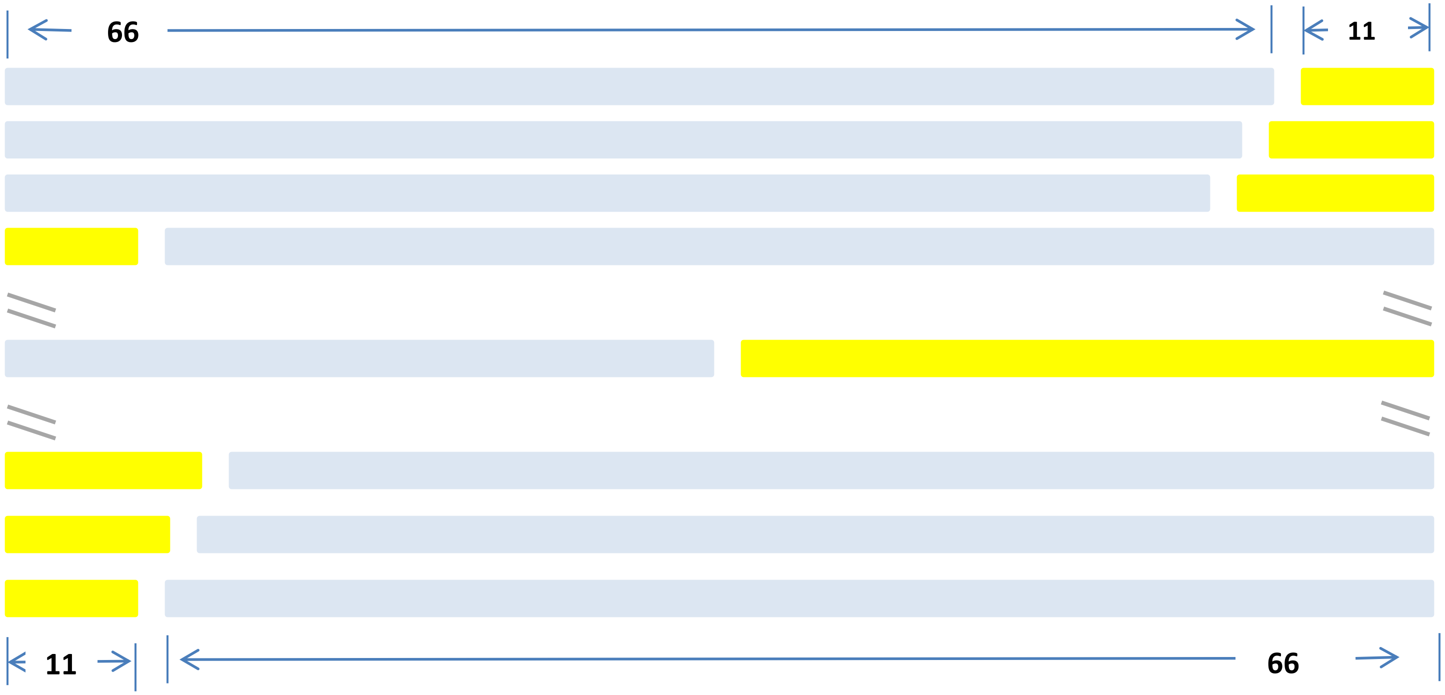
Example 1 key take home message

- Unmapped PE reads contain single ends which map perfectly
- Some of the reads map to repetitive elements which can be mapped to known positions relative to their paired end read

Example 2: RNA-Seq

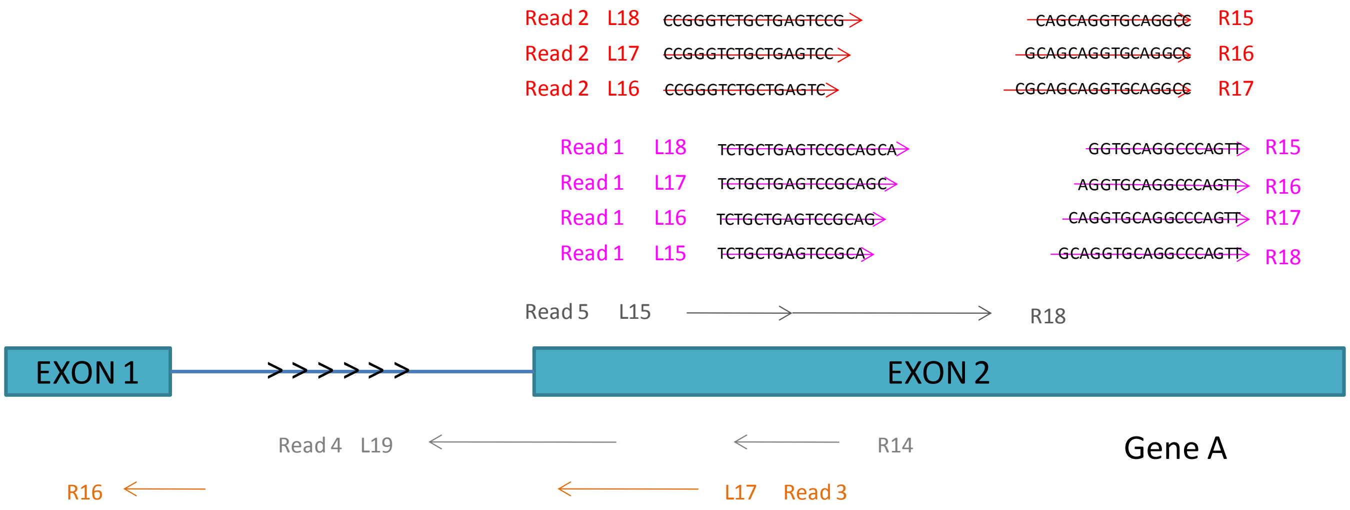
- Investigating a mechanism involving novel splicing of mRNA (not novel splice junction) in mice
- Because of the splice size (e.g 26 bp in one gene), gapping of reads didn’t work.
- Developed novel method to differentially split reads into series of overlapping sub-reads to map novel junctions.
- See Poster #**A57** for details

- RNA-Seq on 2 samples:
  - mouse embryonic fibroblasts (MEF) cells (gene KO, and heterozygote) treated to induce splicing
  - 77 nucleotide single-end reads (trimmed)
- Mapped to mouse genome (+ known splice junctions)
- Unmapped reads then processed into split-reads
  - maintaining FastQ format and modifying readnames to include split information (e.g. L66, R11)



Remapping Split Reads

- All the split reads are put into a FastQ file and Re-aligned to mouse genome.
- Determine which pairs both aligned perfectly (0-mismatch), uniquely to regions within a certain distance (40kb), and on the same strand/chromosome
- Remove inversely mapped pairs which are false positives (Reads 3 and 4 below) and filter out adjacent read pairs (splice region length < 2 bp; read 5).
- The splice region is defined as the interval between the nearest ends of two split halves from the same read.
- Cluster all reads which surround a novel splice location and determine exact boundaries.


Read 2 L18 CCGGGTCTGCTGAGTCCG CAGCAGGTGCAGGCC R15
Read 2 L17 CCGGGTCTGCTGAGTCC GCAGCAGGTGCAGGCC R16
Read 2 L16 CCGGGTCTGCTGAGTC CGCAGCAGGTGCAGGCC R17
Read 1 L18 TCTGCTGAGTCCGCAGCA GGTGCAGGCCCAGTT R15
Read 1 L17 TCTGCTGAGTCCGCAGC AGGTGCAGGCCCAGTT R16
Read 1 L16 TCTGCTGAGTCCGCAG CAGGTGCAGGCCCAGTT R17
Read 1 L15 TCTGCTGAGTCCGCA GCAGGTGCAGGCCCAGTT R18
Read 5 L15 R18
EXON 1
EXON 2
Read 4 L19 R14
Gene A
R16
L17 Read 3

Example 2 –take home message

- Splitting reads has advantage, but sometimes the gaps are larger than normal gapping will support
- Unmapped reads remain a valuable source of data.

Acknowledgements

- Example 1:
  - Dept. of Pediatrics, UM
    - Chris Vlangos, Post-Doc
    - Katie Keegan, Asst. Prof.

- Example 2:
  - Center for Comp. Biology Med. (CCMB)
    - Yongsheng Bai, PhD
    - Maureen Sartor, Asst. Prof.
  - Dept. Biolog. Chemistry
    - Justin Hassler, Doctoral student
    - Randy Kaufman, Professor