# A Hadoop Based Genomic Structural Variation Detection Pipeline

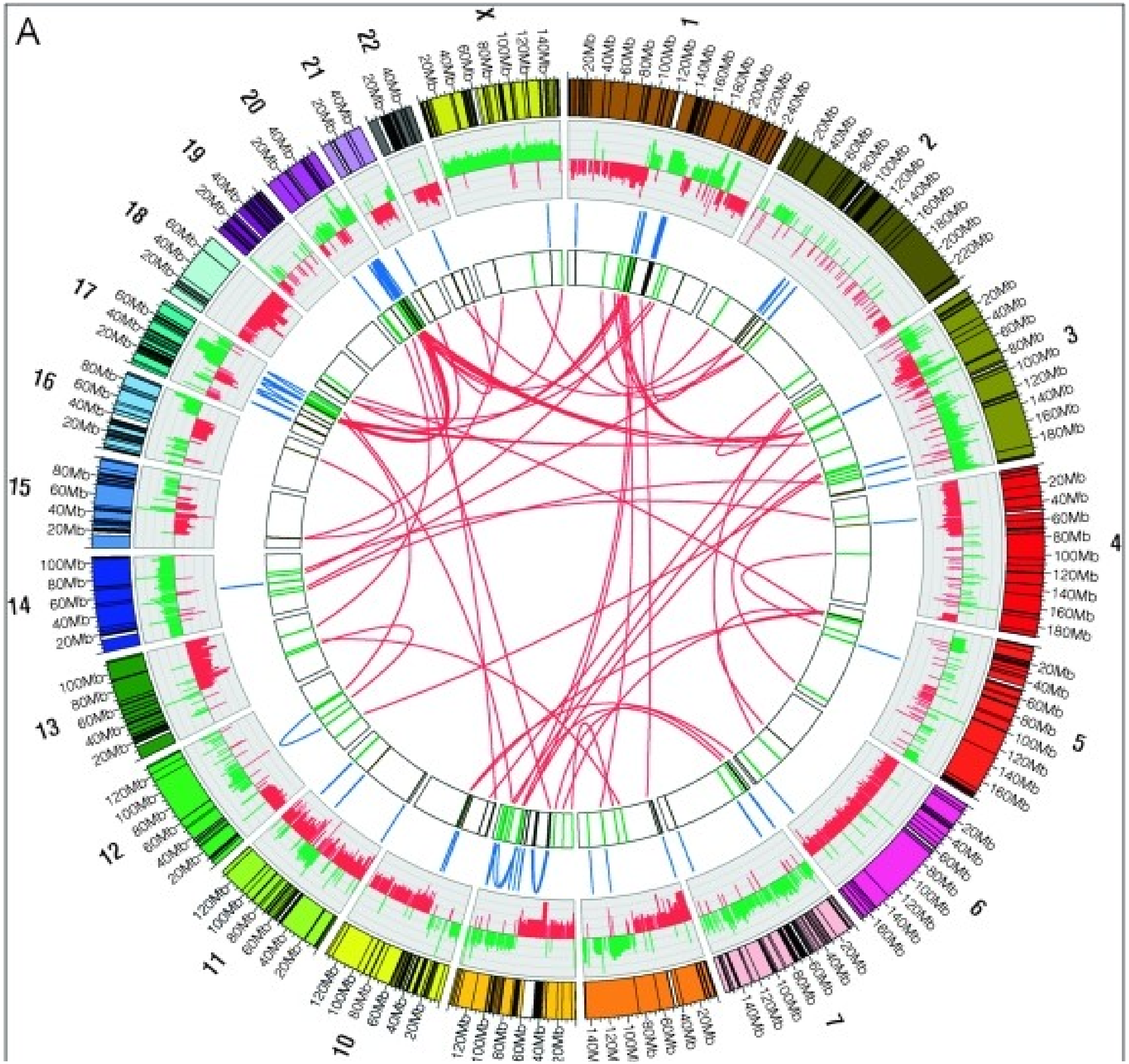
Chris Whelan
CS 506 – Problem Solving with Large Clusters

# Structural Variations

- Structural Variations are large rearrangements in a genome
  - Deletions - removing a chunk of DNA
  - Insertions - inserting a novel chunk of DNA
  - Duplications - copying a chunk of DNA
    - a.k.a copy number variation or tandem repeats
  - Inversions - reversing the order of a chunk of DNA
  - Translocations - moving a chunk of DNA from one chromosome to another

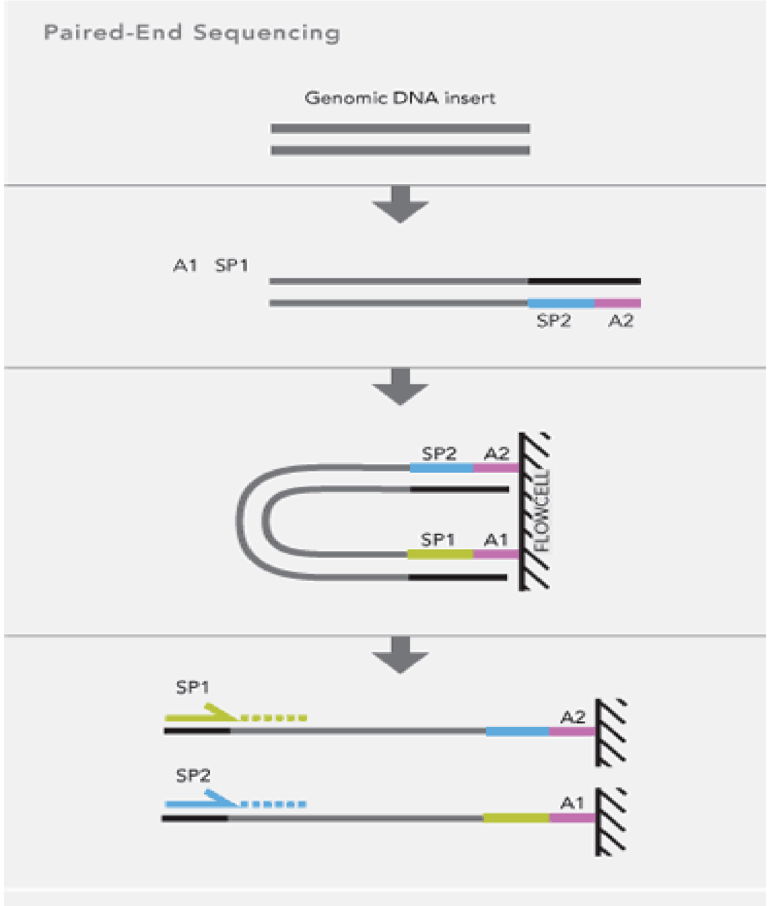
# Structural Variations

- SV's are common in cancer
- This is a map of rearrangements in a breast cancer cell line (Hampton et al. 2009, *Genome Research*)

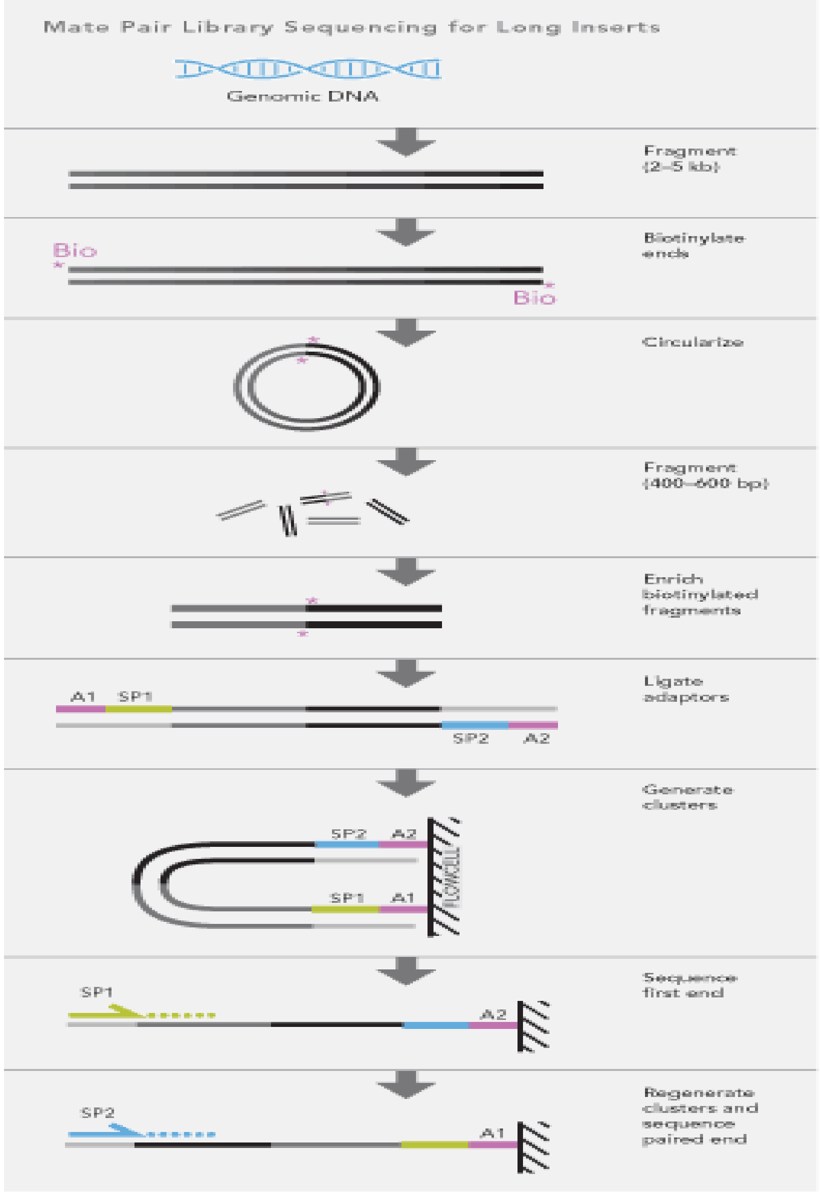


# Paired End Short Read Data

Paired-end and mate-pair sequencing give pairs of reads an approximate distance apart

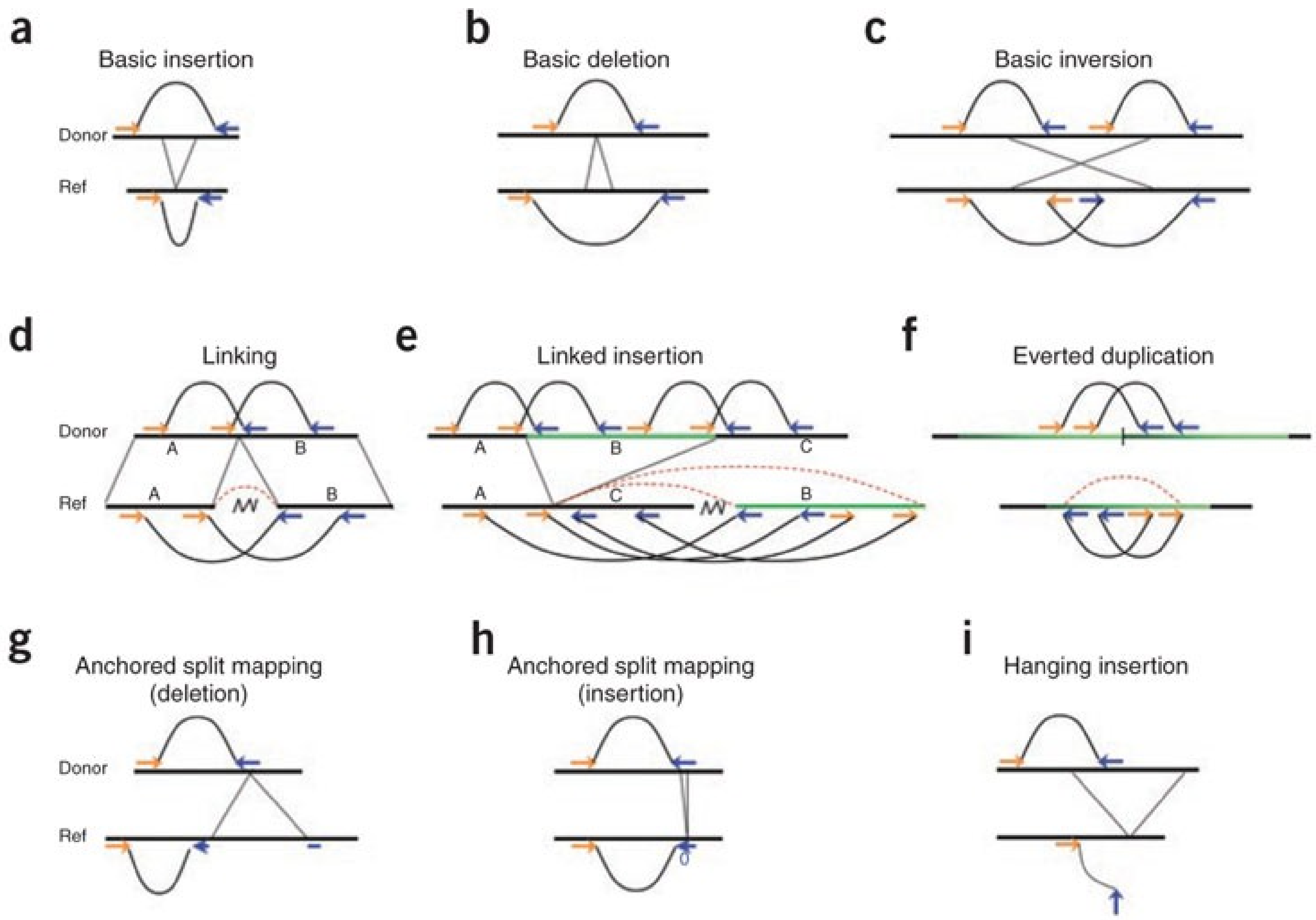


200-300 bp



3000-4000 bp

# Finding Variations with Mate Pair Data



Medvedev et al 2009, Nature Methods

# Current Strategies

- General framework of existing approaches:
  - Map reads to the genome
    - Some reads have many possible mappings; some approaches use multiple mappings (VariationHunter, HYDRA), most do not
  - Remove any pairs for which the aligner did not find a concordant mapping
    - Concordant: within some range of the mean insert size
  - Attempt to cluster the remaining discordant read pair mappings
  - Clusters containing a certain number of read pairs are used to call SV's

# Goals

- Want to use all possible mappings for a read pair
- Don't throw away evidence:
  - Concordant pairs
  - Secondary alignments
- Use Hadoop to manage the large number of possible mappings
- Examine the evidence for an SV at each location in the genome

# My Pipeline

- Map 1: Align each read in a pair independently to the genome with Novoalign – allow up to 10 alignments for each end
  - Emit (Read Pair ID, alignment)
- Reduce 1: Produce all possible pairs of read alignments
  - Emit (Read Pair ID, Read1 Alignments X Read2 Alignments)
- Map 2: For each paired alignment:
  - Compute an SV score: evidence from this alignment for a structural variation
  - Divide the span that this alignment covers into windows
  - Emit (Window location, SV score)
- Reduce 2: Sum all SV scores for each window
- Finally, visualize distribution of scores across the genome and call SV's at the peaks

# Scoring Variant Likelihood at Each Locus

- Novoalign assigns a posterior probability to each possible mapping:

$$P(A_i|R,G) = \frac{P(R|A_i,G)}{P(R|N,G) + \sum_i P(R|A_i,G)}$$

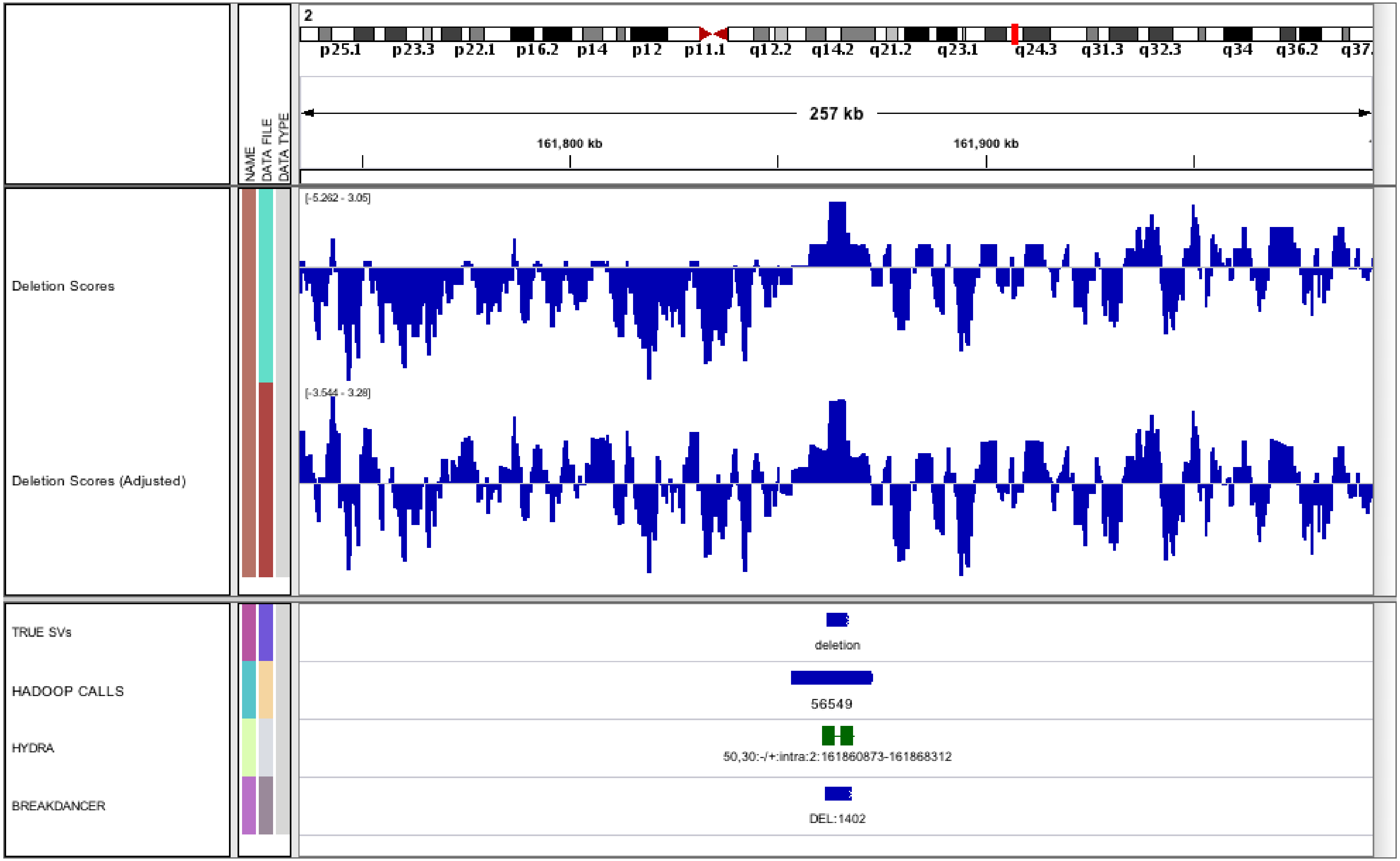
- Based on an aligned insert size *s* and the expected distribution of insert sizes ~ *N(μ,σ)*, call the likelihood of the alignment pair spanning a deletion $P(S < s - \alpha\sigma)$
- Let the read's vote *v* be 1 if *p* > .5; -1 otherwise
- Each read contributes a score of

$$P(A_{end1}|R,G) * P(A_{end2}|R,G) * v$$

# Experiments

- Using a simulation pipeline:
  - Add random structural variations to a genome sequence (Human Chromosome 2)
  - Simulate 10,000,000 Illumina mate pairs from the modified genome sequence
- Compare to output from two published SV callers: HYDRA and BreakDancer
- Using a very simple sliding window to call peaks in my output distribution: from each value subtract the average of the values in the neighboring 50kb

# Visualizing Results



# Evaluating Results

| | True Positives | False Positives |
|---|---|---|
| New Workflow | 24 | 122 |
| HYDRA | 28 | 52 |
| BreakDancer | 35 | 70 |

- True SV's: 75
- Too many false positives for my method
- Need better peak calling technique
- Did identify 7 SV's that both HYDRA and BreakDancer missed

# Future Work

- Better method for calling peaks from deletion scores
  - Anecdotally, many actual SV's had peaks in the score distribution, but didn't make the threshold
- Add scoring functions for other types of SV
  - Scoring function in the current implementation is designed for deletions but suprisingly does well with insertions and inversions
- Optimize parameters of scoring functions
- Allow more mappings
- Expand to multiple chromosomes