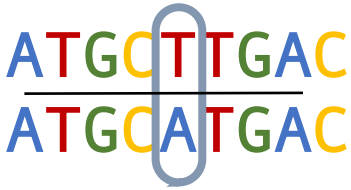
RNA-DNA differences identification



- Giovanni Quinones Valdez

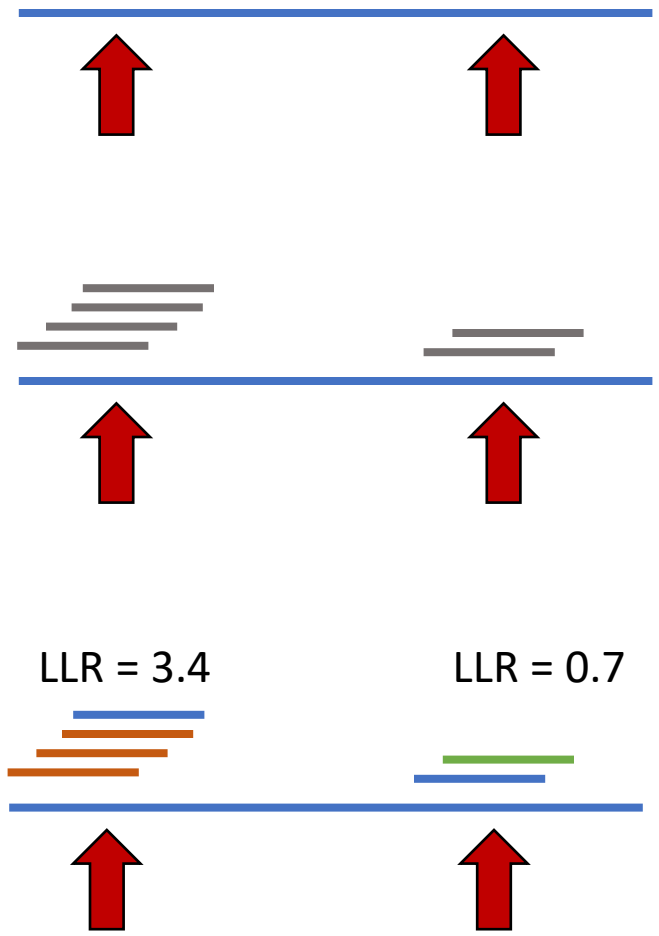
- Lab meeting: 10/04/2018

Motivation

- To create a pipeline to identify RNA-DNA differences (mismatches) that can be used by all lab members.
- To modify the current scripts to accept bam files as input, as they are the main output format from aligners.
- Include reads with Soft Clipping and InDels (2.5 – 3% of the reads).
- To improve run time and memory usage.
- To merge information from multiple chromosomes to improve the log likelihood ratio estimation (more power).

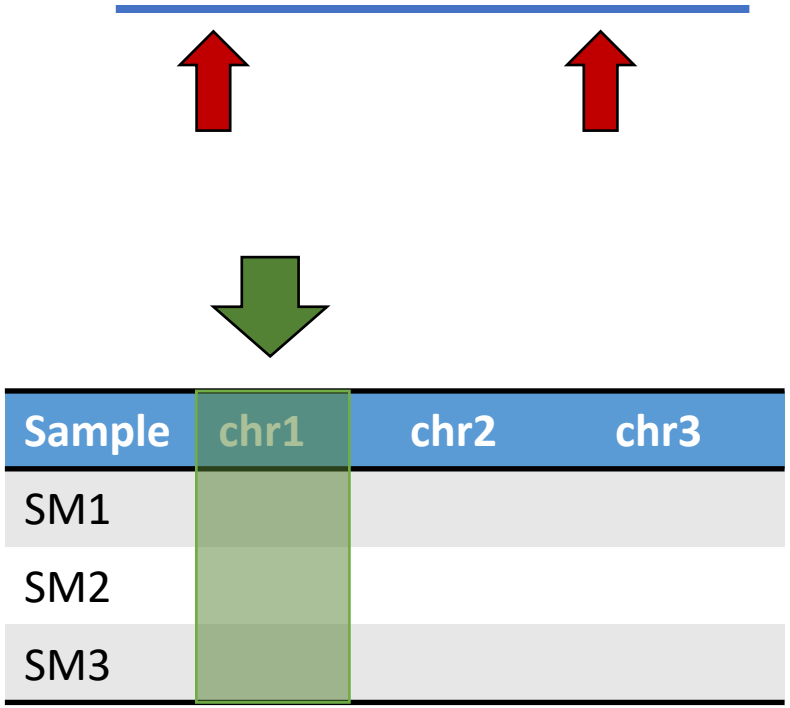
Methods

- **Step1**
  - **Identify the coordinates of mismatches**
  - **Pool mismatches from multiple samples**
- **Step2**
  - **Identify the reads overlapping the mismatch coordinates**
  - **Filter reads**
- **Step3**
  - **Calculate editing ratio**
  - **Calculate Log Likelihood Ratio (LLR)**



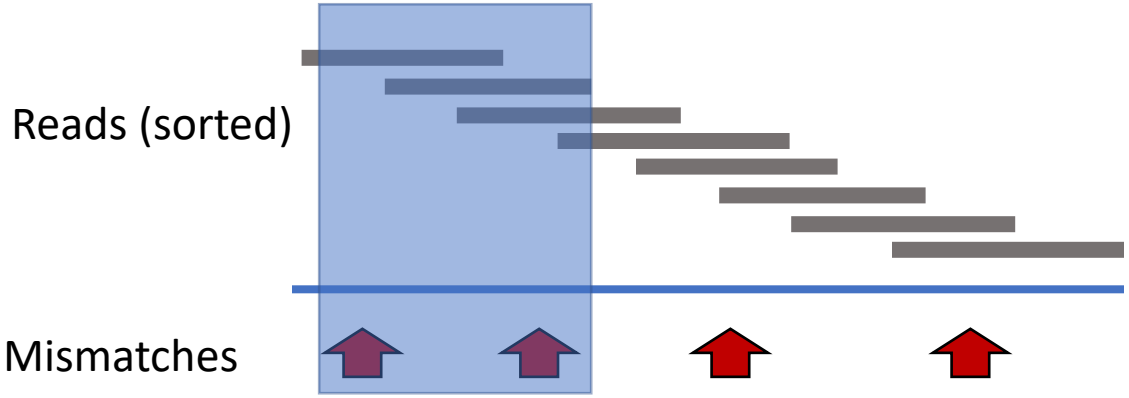
Methods

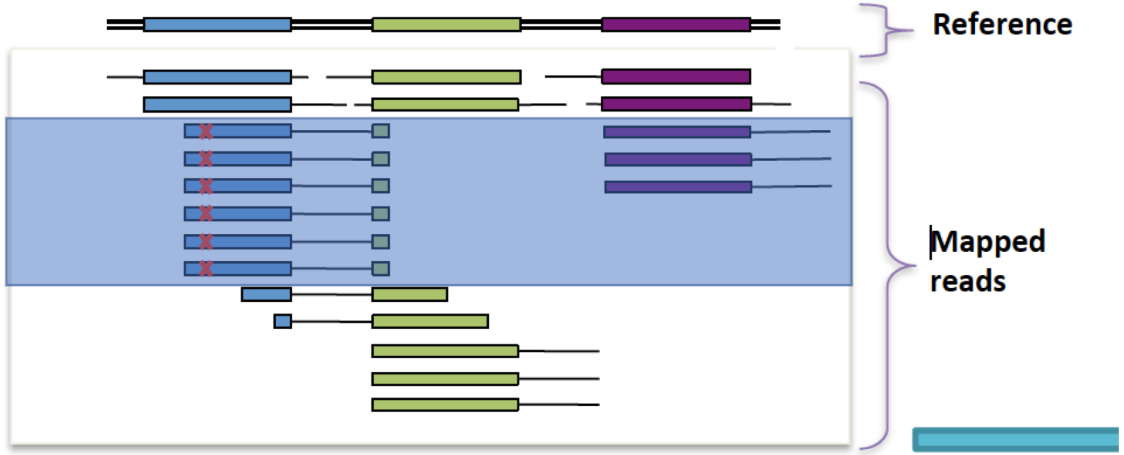
- **Step1**
  - **Identify the coordinates of mismatches.**
  - **Approach:**
    - Compare the Ref sequence to Read Sequence base by base.
  - **Pool mismatches from multiple samples.**
    - **One chromosome at the time for all samples**
  - **Filters:**
    - Position from ends (> 5 nt)
    - Read Quality (Phred = 33, > 20)
    - Non secondary, proper-paired
    - Minimum Read coverage per editing type per locus.



Methods

- **Step2**
  - **Identify the reads overlapping the mismatch coordinates.**
    - Dynamic sub setting of mismatch list according to read coordinates. Aided by **binary search** of closest MM
    - Faster then pileup approach (more common)
  - **Filter reads (same as before)**
    - Additionally, we filter for PCR duplicates. We select unique start and end coordinates for every pair of reads.





✖ = sequencing error propagated in duplicates

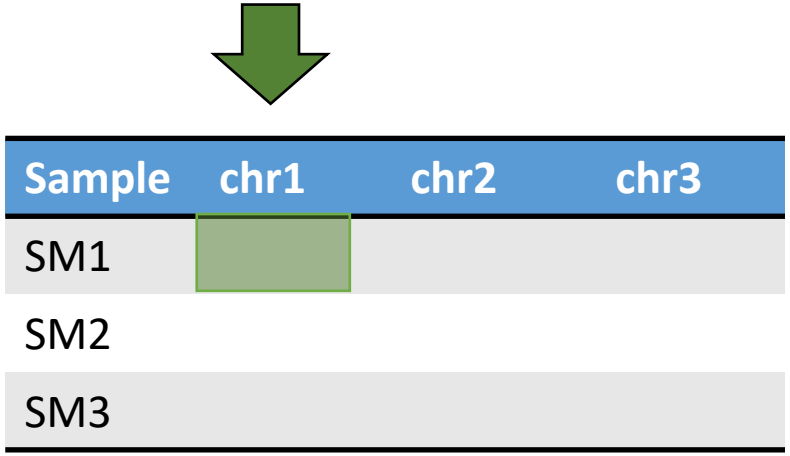
GATK: Marking PCR Duplicates

Methods

- **Step2**
  - **LLR calculation.**
    - This step counts the number reads that for each condition (reference base, quality, read position) where the mismatch happens.

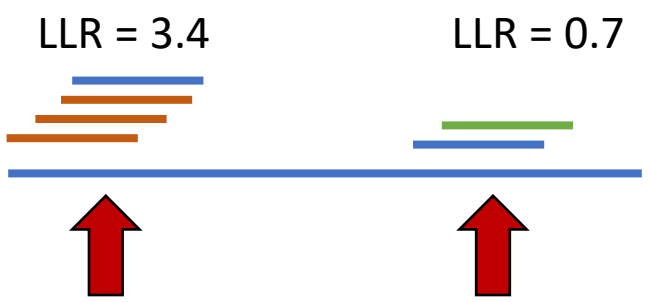
$$\boldsymbol{Count}\{ref = G\}\{Pos_{read} = 50\}\{q = 20\}\{b = C\} = \boldsymbol{n}$$

  - **Per chromosome per sample.**
  - **Memory.**
    - PCR removal is on the spot, so no need for intermediate files for this step.
    - Saving intermediate information in binary pickle files and binary arrays.
      - MM coordinates files: 4.7x compression
      - MM read information files: > 4x compression



Methods

- **Step3**
  - **LLR calculation.**
    - B = Nucleotide in read ; q = quality of the nucleotide in read.
    - It calculates the most likely editing ratio and its confidence based on the quality of the reads containing the mismatch.



$$P(b, q \mid ref = G, Pos_{read} = 50)$$

$$P(b = C, q = 20 \mid ref = G, Pos_{read} = 50) = \frac{Count\{ref = G\}\{Pos_{read} = 50\}\{q = 20\}\{b = C\}}{\sum_{base} Count\{ref = G\}\{Pos_{read} = 50\}\{q = 20\}\{b = base\}}$$

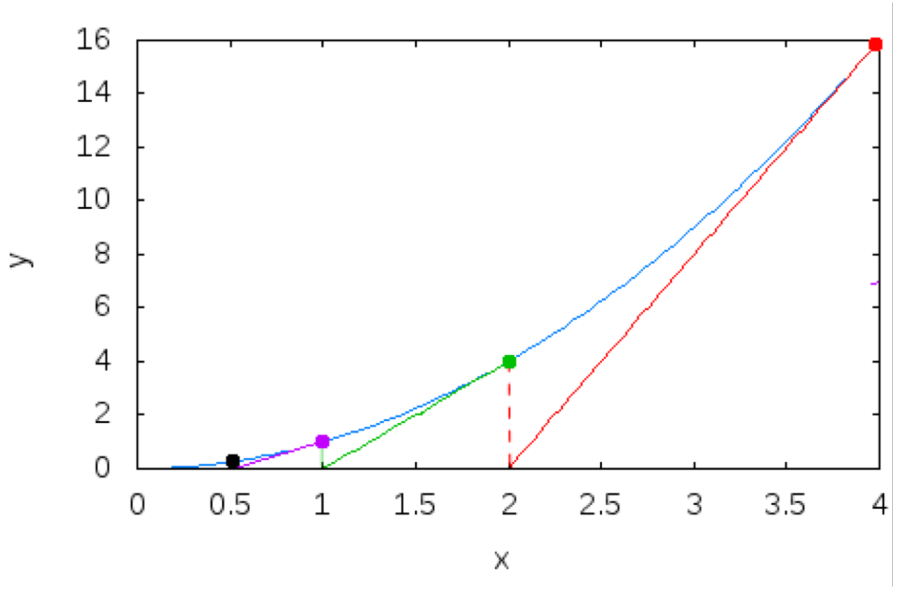
$$Likelihood(\boldsymbol{f}) = \prod_{all\ reads} P(b_i, q_i \mid ref = G, Pos_{read} = P_i) * \boldsymbol{f} + P(b_i, q_i \mid ref = C, Pos_{read} = P_i) * (1 - \boldsymbol{f})$$

$$LLR = \log_{10}\left(\frac{Likelihood(f = f_{max})}{Likelihood(f = 0)}\right)$$

By Dr. Grace Xiao

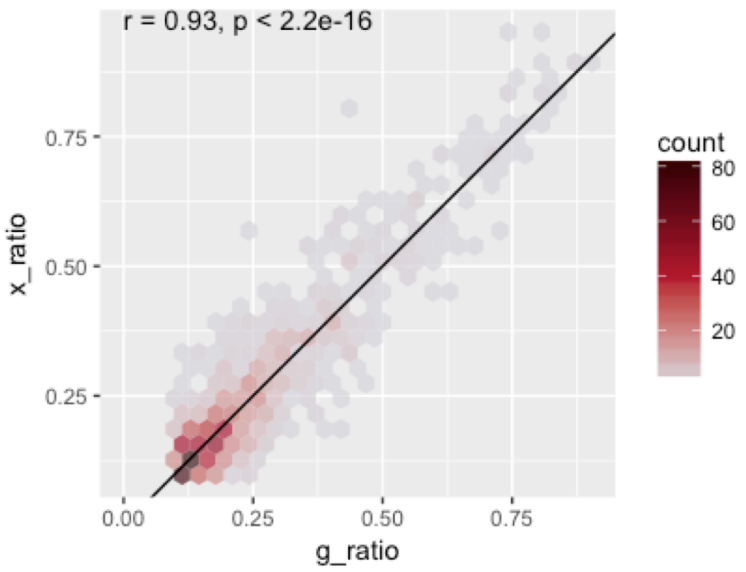
Methods

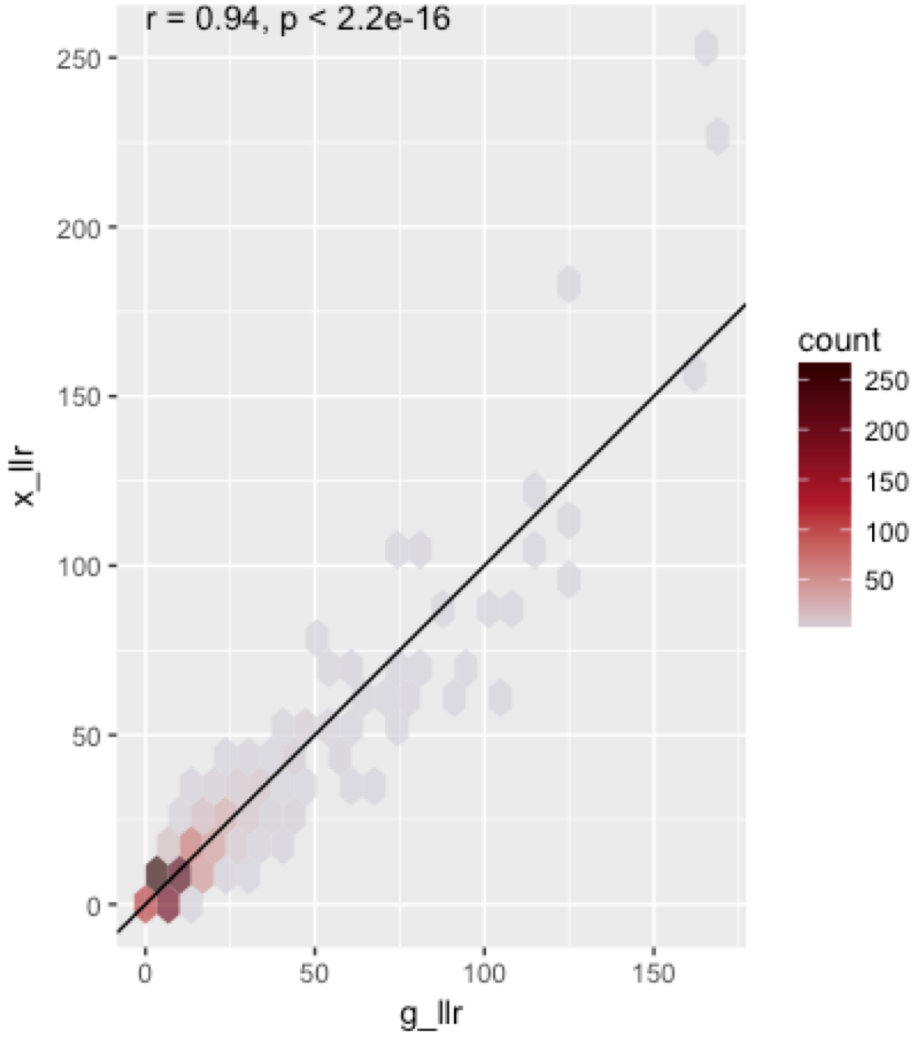
- **Step3**
  - **Calculate Log Likelihood ratio (run time reduction)**
    - **Newton's method:** Instead of trying multiple values of f to try $\mathbf{f_{max}}$ (1000 iterations) we use Newton's approach to find the zeros of the derivative (~3 iterations).
    - **Avoid redundant information:** Multiple reads have the same probability value (same mismatch type, same quality bin, same read position bin). No need to repeat the operation multiple times.
  - The calculation is done with reads obtained form all chromosomes



Sample	chr1	chr2	chr3
SM1			
SM2			
SM3			

Results: comparison of editing ratios and LLR



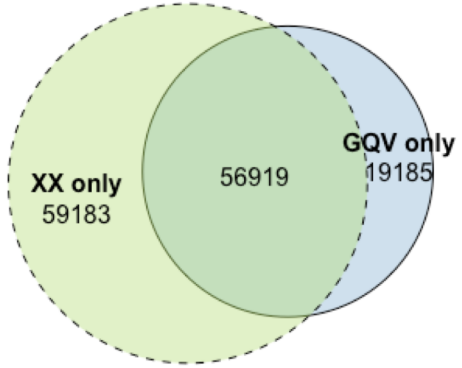


- Increasing accuracy with increasing read coverage cutoff
- (Tracey's) Nicotine dataset (editing ratio vs LLR)

Conclusion

- The new pipeline is separated into 3 main steps to minimize the running time and allow the user to select the number of jobs to merge together.
- Lowered I/O load to the cluster with lower memory requirement.
- Greater power for LLR calculation.
- Overall, very reproducible editing ratio and LLR values.
- Low overlap of sites identified, greatly increased however by considering higher coverage thresholds.
- 7-filters script is also modified to run faster and not need intermediate files.

All sites after 7-filters



All sites after 7-filters (XX > 2 edited reads)