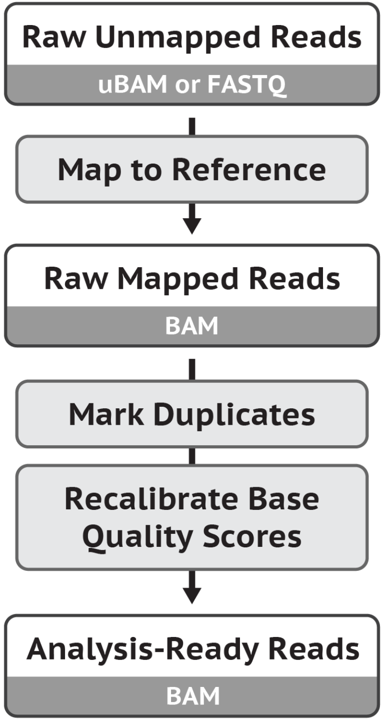
GATK Best Practices for Variant Discovery

Marking duplicates

Removing non-independent observations

http://software.broadinstitute.org/gatk/

Data Pre-processing for Variant Discovery



Raw Unmapped Reads

uBAM or FASTQ

Map to Reference

Raw Mapped Reads

BAM

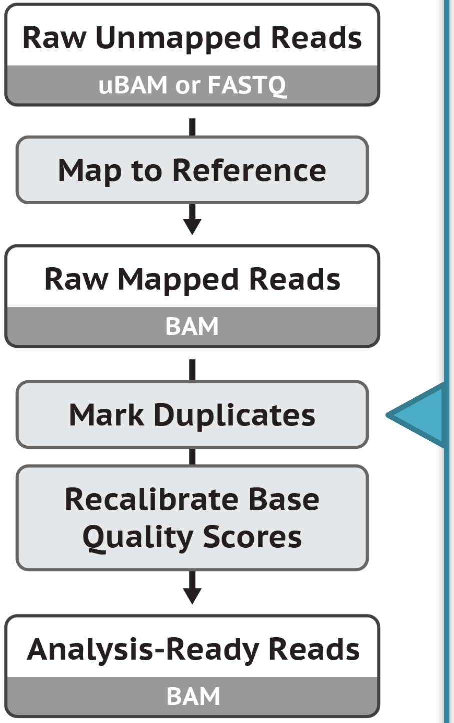
Mark Duplicates

Recalibrate Base Quality Scores

Analysis-Ready Reads

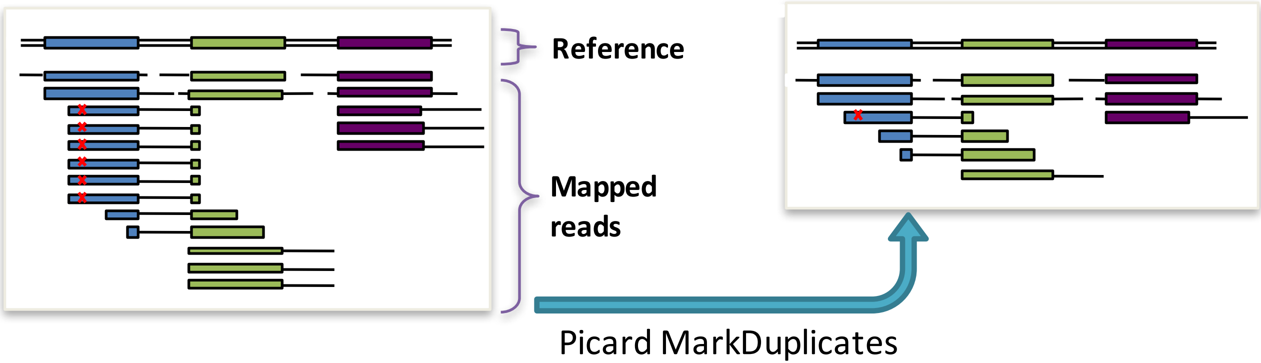
BAM

Mark duplicates to mitigate duplication artifacts



Duplicates = **non-independent measurements** of a sequence fragment

-> Must be removed to assess support for alleles correctly

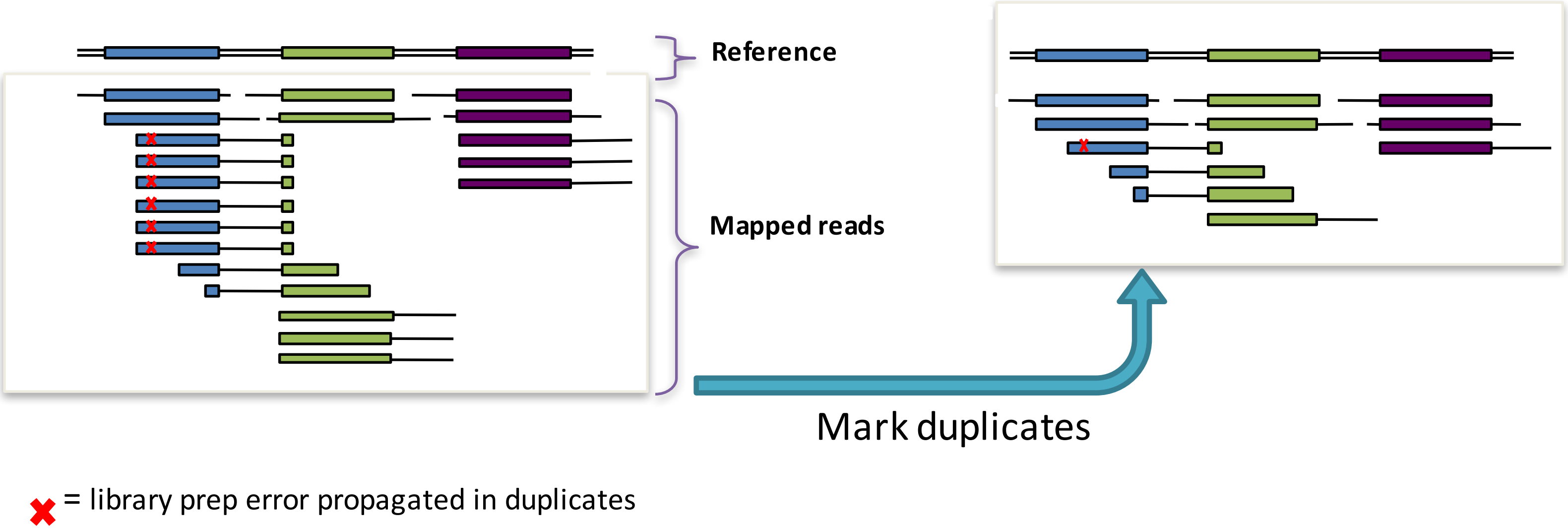


✖ = sequencing error propagated in duplicates

Why mark duplicates?

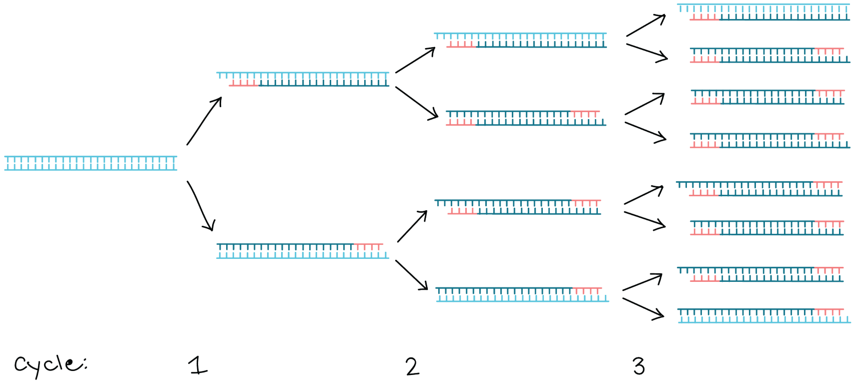
- **Non-independent measurements** of sequence
  - Sampled from single template of DNA
  - Violates independence assumptions made in variant calling
- Errors in sample/library prep are propagated to *all* the duplicates

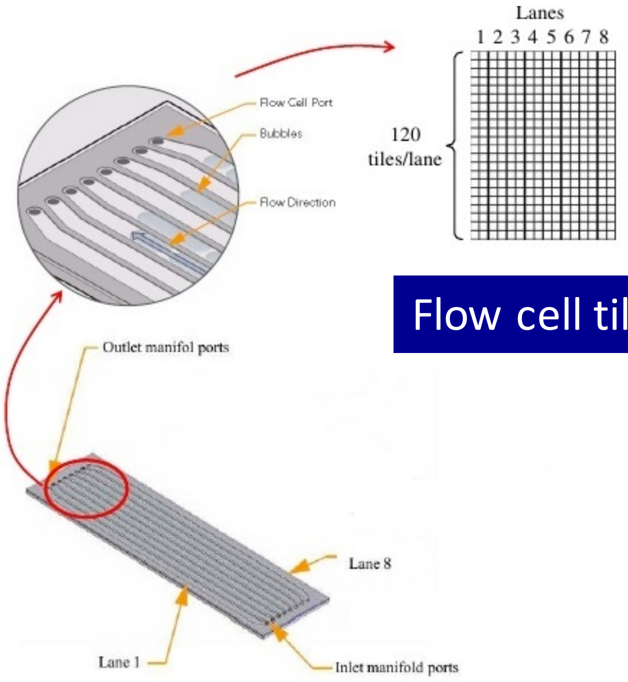
- “Best” copy – mitigates the effects of errors



Where does the duplication come from?

- **LIBRARY DUPLICATES**
  - **Increases with PCR cycles**
- **OPTICAL DUPLICATES**
  - **Are nearby clusters on a flow cell lane**





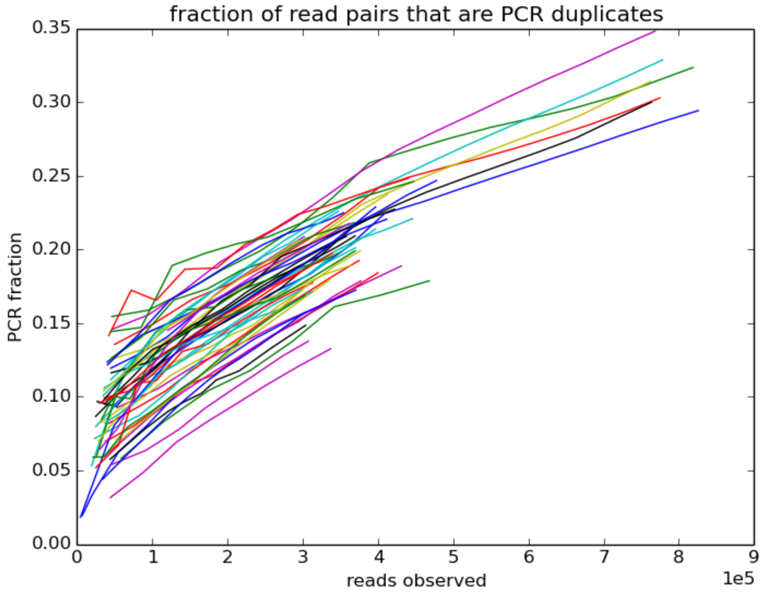
Flow cell tiles

https://www.khanacademy.org/science/biology/biotech-dna-technology/dna-sequencing-pcr-electrophoresis/a/polymerase-chain-reaction-pcr

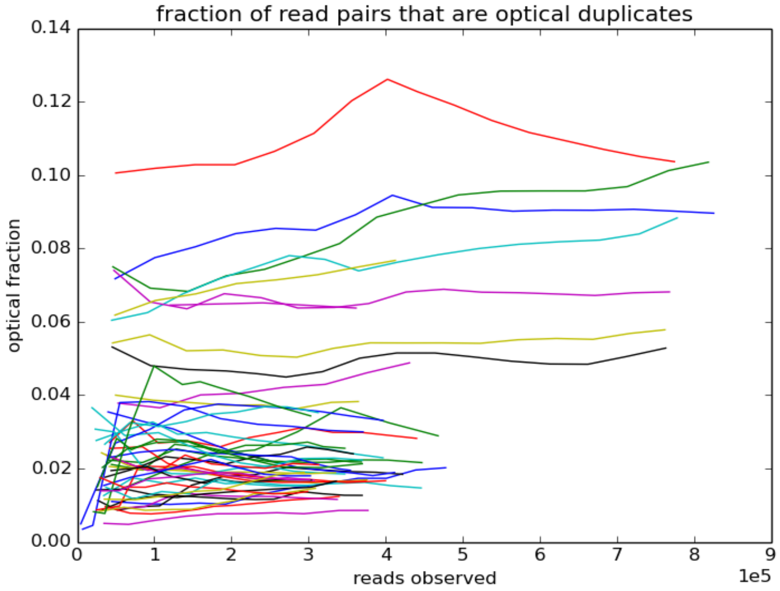
http://www.slideshare.net/jandot/next-generation-sequencing-course-part-2-sequence-mapping
http://www.slideshare.net/cosentia/illumina-gaiix-for-high-throughput-sequencing

Optical and PCR duplication events arise at different rates as a sequencing experiment proceeds

PCR duplicates



Optical duplicates



Duplicates are flagged the same but can be tagged differently (DT)

0x400 flag
DT:SQ

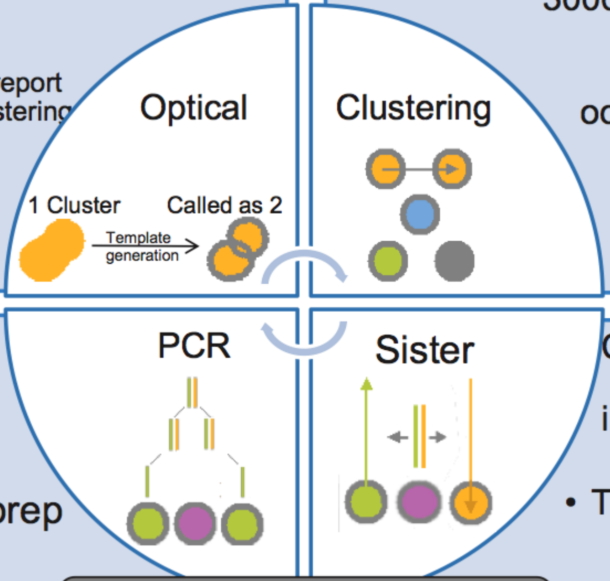
- A single cluster that has falsely been called as two by RTA
  - Third party tools may report patterned flow cell clustering duplicates as optical duplicates

Not on Patterned Flow Cells

0x400 flag
DT:SQ

- Duplicates in nearby wells on HiSeq 3000/4000
  - During cluster generation a library occupies two adjacent wells

Unique to Patterned Flow Cells



0x400 flag
DT:LB

- Duplicate molecules that arise from amplification
- during sample prep

0x400 flag
DT:LB

Complement strands of same library form independent clusters

- Treated as duplicates by some informatic pipelines

Present on all Illumina platforms

http://core-genomics.blogspot.fi/2016/01/almost-everything-you-wanted-to-know.html

How do we identify duplicate reads?

- Dupes might come from the same input DNA template, so we will assume that reads will have same start position on reference
  - “Where was the first base that was sequenced?”
  - For paired-end (PE) reads, same start for both ends
- Identify duplicate sets, then choose representative read based on base quality scores and other criteria

But there’s a catch (or two)…

- BWA sometimes “clips” bases from the ends of the alignment (when the alignment there is poor)
  - Need to use SAM flags + CIGAR string to determine the unclipped 5’ end
- Fragments mapped to the reverse strand are specified by their 3’ position, instead of 5’

Identify duplicates using orientation + “unclipped” 5’ position

Pos	1	2	3	4	5	6	7	8	9
Ref	T	A	G	C	C	G	A	T	C
r1	T	A	G	C	C	G	A		
r2	T	A	G	C	C	G	A		
r3	T	A	—	C	CAG		A		
r4	T	A	G	C	C	H	H		
r5	T	A	G	C	C	G	A	T	C
r6	S	S	G	C	C	G	A		
r7			G	C	C	G	A		

Blue maps to forward strand
Red maps to reverse strand
Grey bases are clipped

Underlined is the expected 5’ start of the read, given the mapping

What are the duplicate sets?

Identify duplicates using orientation + “unclipped” 5’ position

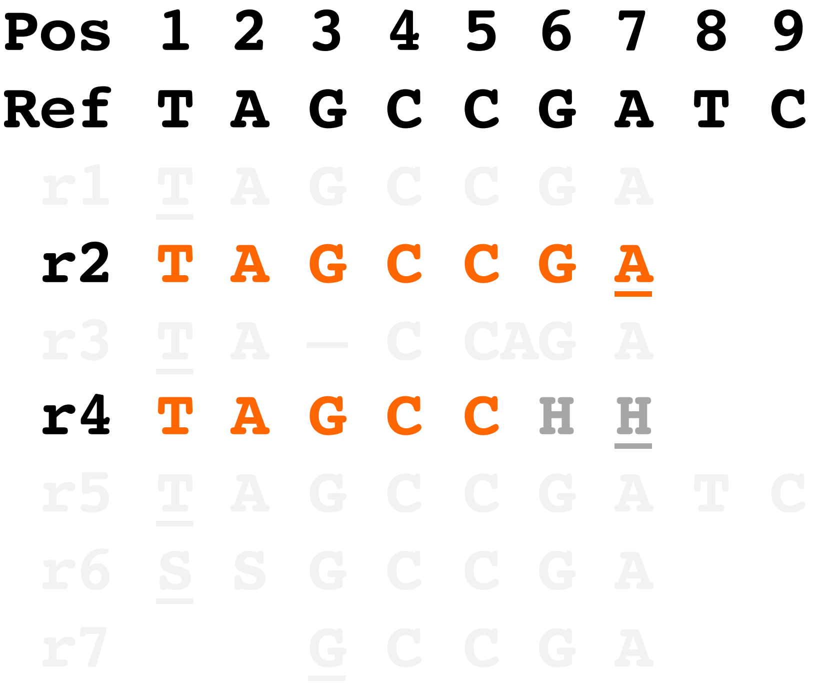
Pos	1	2	3	4	5	6	7	8	9
Ref	T	A	G	C	C	G	A	T	C
r1	T	A	G	C	C	G	A		
r2	T	A	G	C	C	G	A		
r3	T	A	—	C	CAG		A		
r4	T	A	G	C	C	H	H		
r5	T	A	G	C	C	G	A	T	C
r6	S	S	G	C	C	G	A		
r7			G	C	C	G	A		

Blue maps to forward strand
Red maps to reverse strand
Grey bases are clipped

Underlined is the expected 5’ start of the read, given the mapping

So…what are the duplicate sets?
☞ **r1, r3, r5, r6** (start at position 1)

Identify duplicates using orientation + “unclipped” 5’ position



Pos	1	2	3	4	5	6	7	8	9
Ref	T	A	G	C	C	G	A	T	C
r1	T	A	G	C	C	G	A		
r2	T	A	G	C	C	G	A		
r3	T	A	—	C	CAG		A		
r4	T	A	G	C	C	H	H		
r5	T	A	G	C	C	G	A	T	C
r6	S	S	G	C	C	G	A		
r7			G	C	C	G	A		

Blue maps to forward strand
Orange maps to reverse strand
Grey bases are clipped

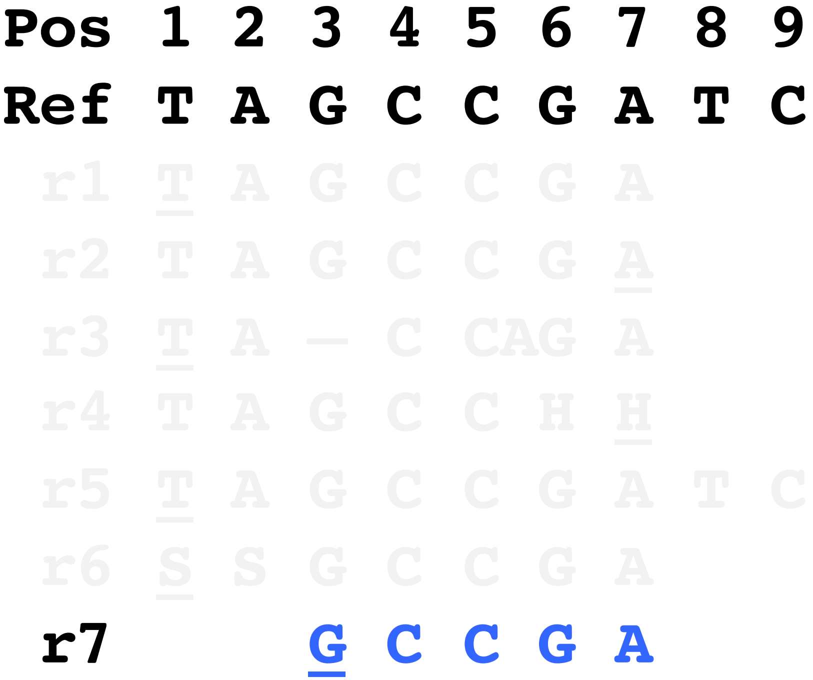
Underlined is the expected 5’ start of the read, given the mapping

So…what are the duplicate sets?

☞ **r1, r3, r5, r6** (start at position 1)

☞ **r2, r4** (start at position 7)

Identify duplicates using orientation + “unclipped” 5’ position



Pos	1	2	3	4	5	6	7	8	9
Ref	T	A	G	C	C	G	A	T	C
r1	T	A	G	C	C	G	A		
r2	T	A	G	C	C	G	A		
r3	T	A	—	C	CAG		A		
r4	T	A	G	C	C	H	H		
r5	T	A	G	C	C	G	A	T	C
r6	S	S	G	C	C	G	A		
r7			G	C	C	G	A		

Blue maps to forward strand
Orange maps to reverse strand
Grey bases are clipped

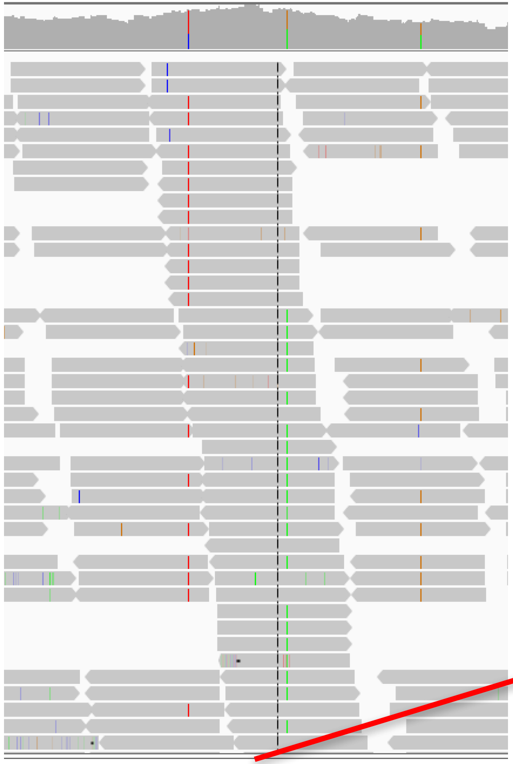
Underlined is the expected 5’ start of the read, given the mapping

So…what are the duplicate sets?

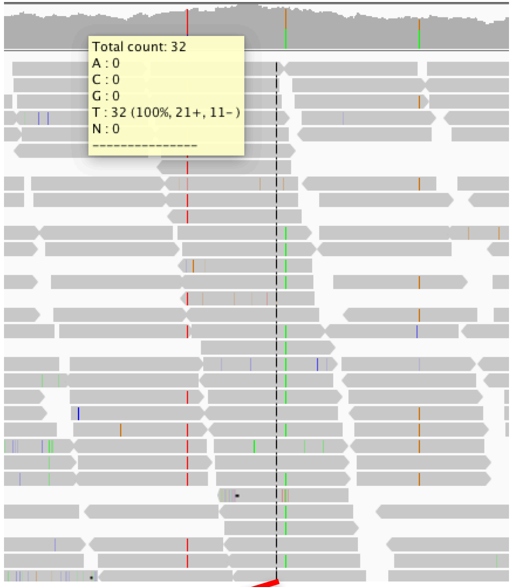
- ☞ **r1, r3, r5, r6** (start at position 1)
- ☞ **r2, r4** (start at position 7)
- ☞ **r7** (starts at position 3)

So now we have mapped, sorted, and *deduped* reads

Showing duplicate reads



Hiding duplicate reads



```
Total count: 32
A : 0
C : 0
G : 0
T : 32 (100%, 21+, 11- )
N : 0
---------------
```

- Duplicate status is indicated in SAM flag
- Duplicates are **not removed**, just tagged (unless you request removal)
- Downstream tools can read the tag and choose to ignore those reads
- Most GATK tools ignore duplicates by default

Use cases where you may *NOT* want to mark duplicates

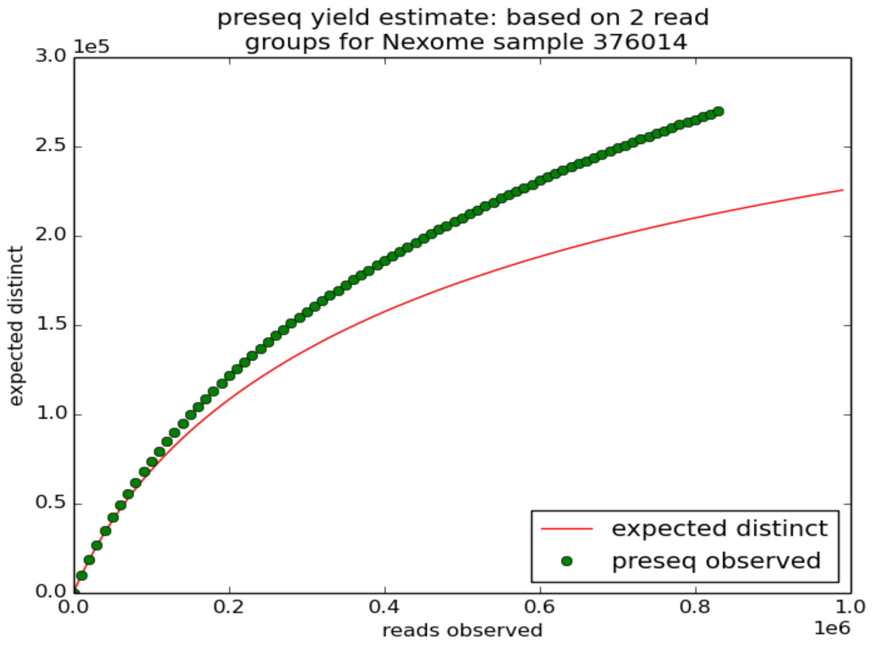
- Amplicon sequencing (or any other LC with non-random starts)
  - All reads start at same position by design
  - If have UMI can use UMIAwareMarkDuplicatesWithMateCigar

- RNAseq (allele-specific) expression analysis
  (alternatively, ASEReadCounter can disable DuplicateFilter)

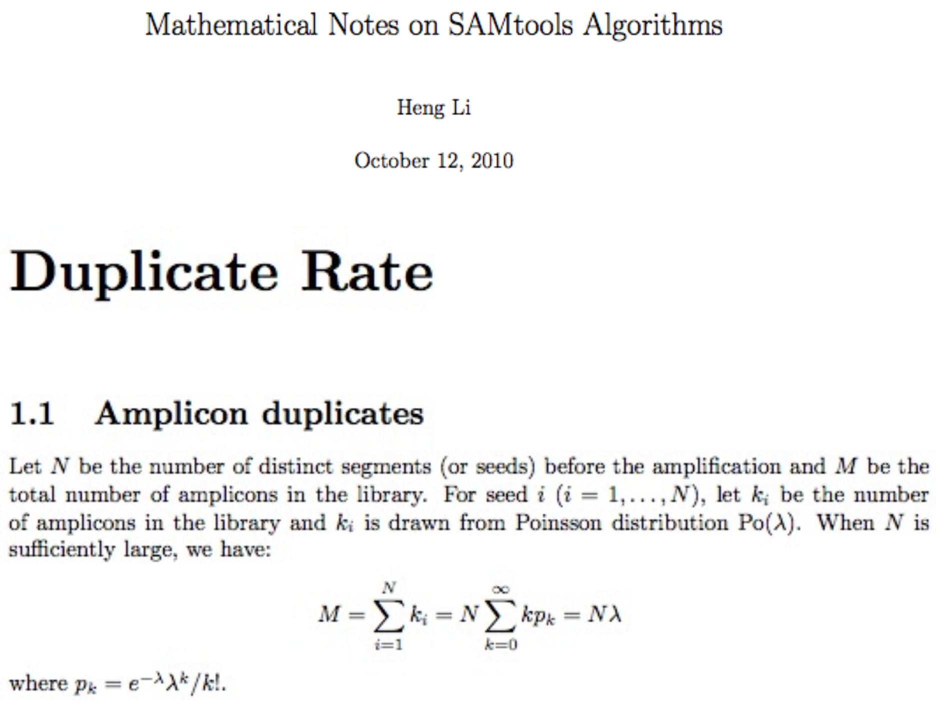
Add-on: Predicting the complexity of a sequencing experiment

Complexity analysis depends on:

- **Estimated library size**
- **Return on Investment (ROI) calculations**



Estimation of library size and duplication in Picard



Mathematical Notes on SAMtools Algorithms

Heng Li

October 12, 2010

Duplicate Rate

1.1 Amplicon duplicates

Let N be the number of distinct segments (or seeds) before the amplification and M be the total number of amplicons in the library. For seed i ($i = 1, \ldots, N$), let k_i be the number of amplicons in the library and k_i is drawn from Poinsson distribution $\mathrm{Po}(\lambda)$. When N is sufficiently large, we have:

$$M = \sum_{i=1}^{N} k_i = N \sum_{k=0}^{\infty} k p_k = N\lambda$$

where $p_k = e^{-\lambda}\lambda^k/k!$.

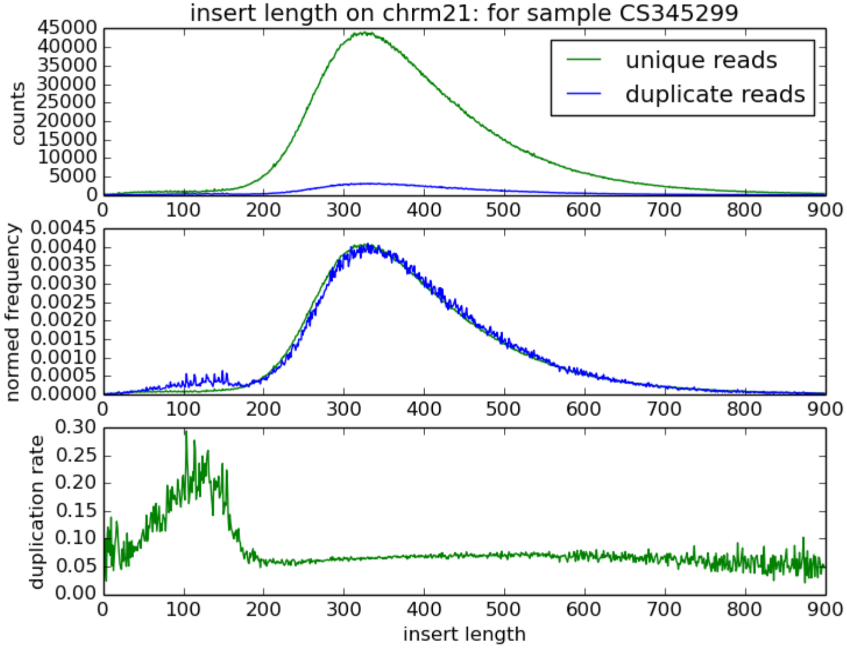
Estimated fraction of duplicates

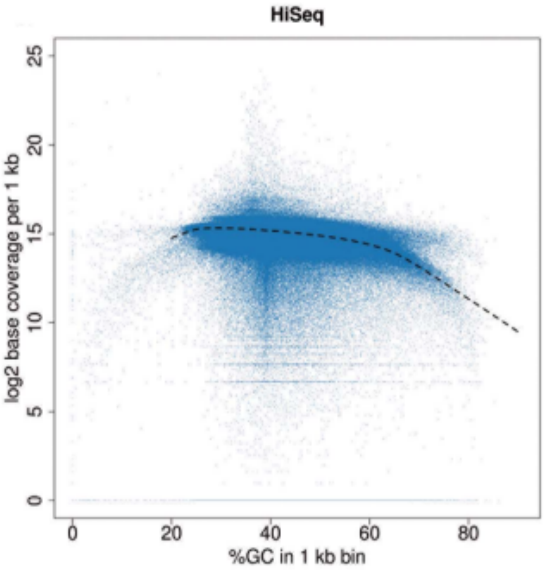
$$d \simeq 1 - \frac{N}{m}\left(1 - e^{-m/N}\right)$$

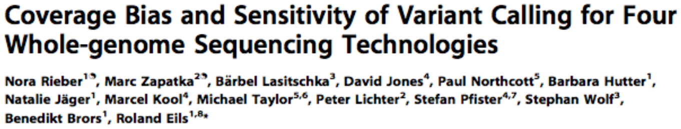
Assumptions

- all reads are drawn from the same Poisson distribution Po(λ)
- the occurrence of duplication events depends on underlying concentration of inserts in the library

Active research to improve library size estimation





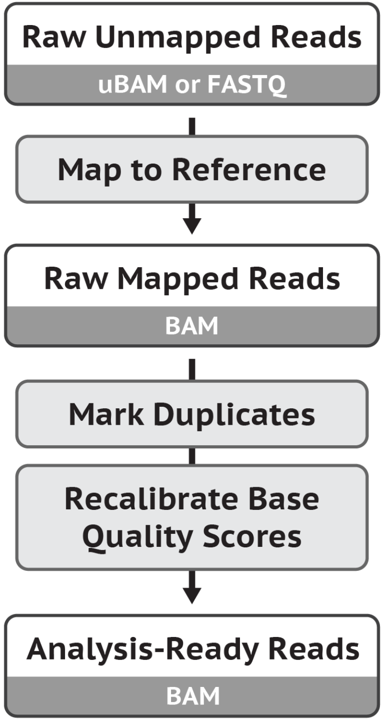


Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies

Nora Rieber[1¶], Marc Zapatka[2¶], Bärbel Lasitschka[3], David Jones[4], Paul Northcott[5], Barbara Hutter[1], Natalie Jäger[1], Marcel Kool[4], Michael Taylor[5,6], Peter Lichter[2], Stefan Pfister[4,7], Stephan Wolf[3], Benedikt Brors[1], Roland Eils[1,8*]

- Rate of duplication varies with insert size length
- Duplications rates also likely vary with GC content

Data Pre-processing for Variant Discovery


Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM