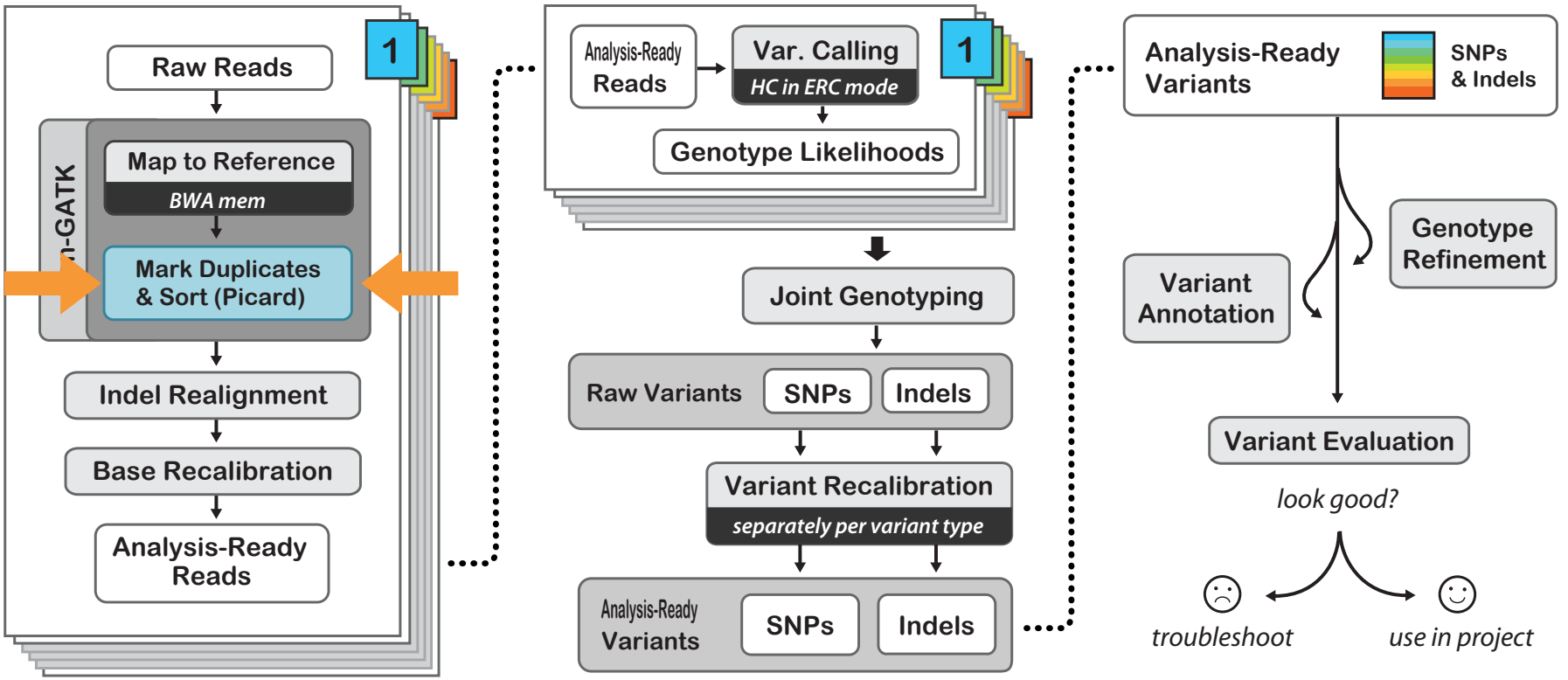
Marking duplicates

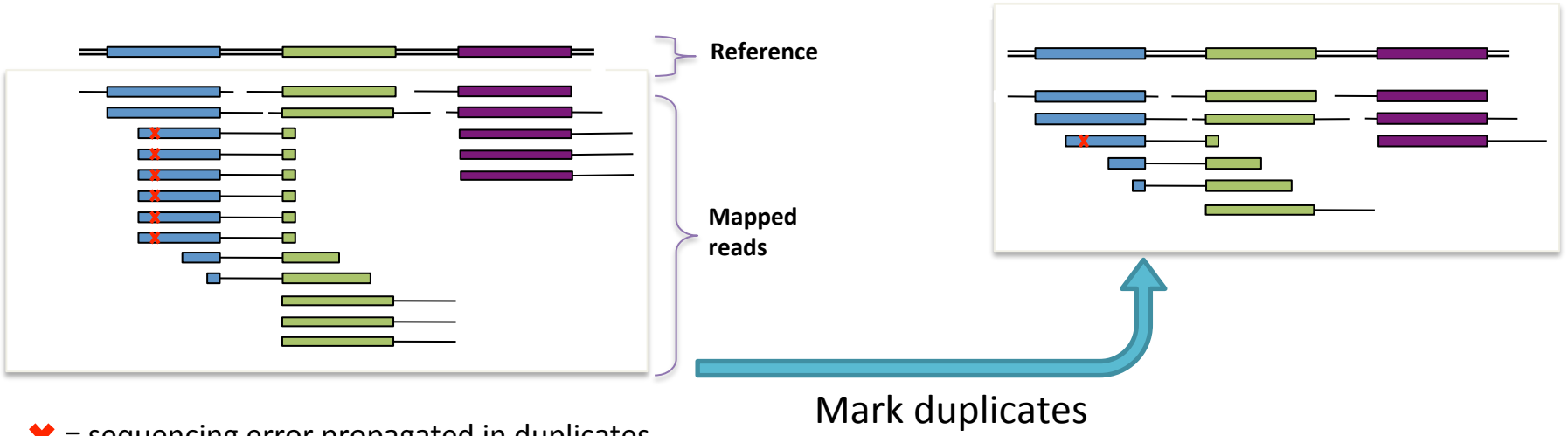
Removing non-independent observations

You are here in the GATK Best Practices workflow for **germline variant discovery**



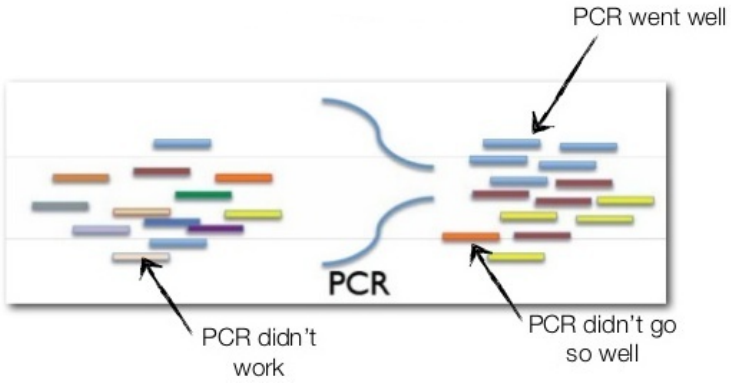
Why mark duplicates?

- Duplicates are sets of reads pairs that have the same unclipped alignment start and unclipped alignment end
- They're suspected to be **non-independent measurements** of a sequence
  - Sampled from the exact same template of DNA
  - Violates assumptions of variant calling
- What's more, errors in sample/library prep will get propagated to *all* the duplicates
  - Just pick the "best" copy – mitigates the effects of errors



How do duplication events arise?

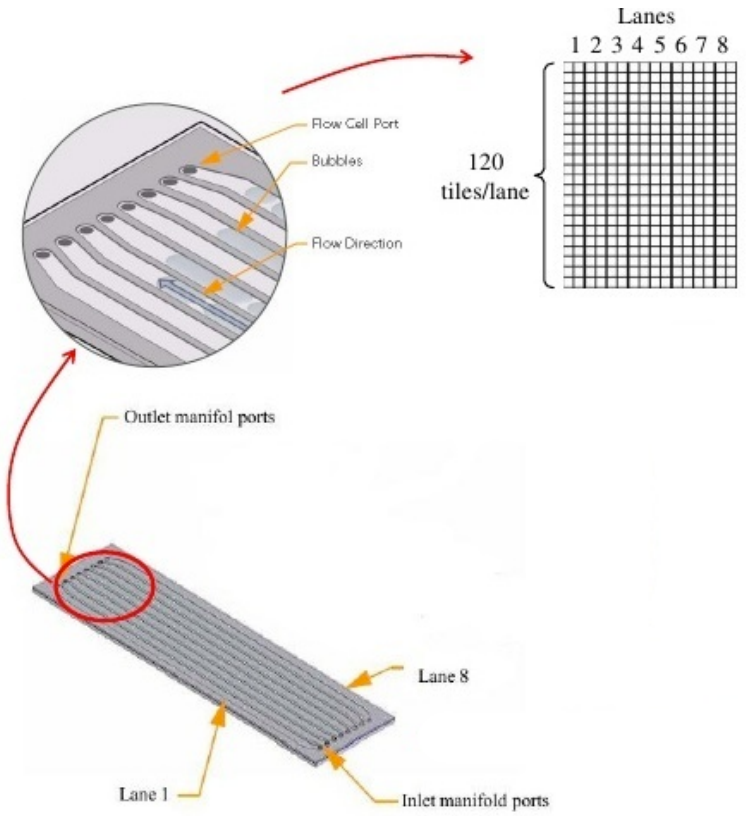
PCR duplicates



Optical duplicates

Read names have the following form:

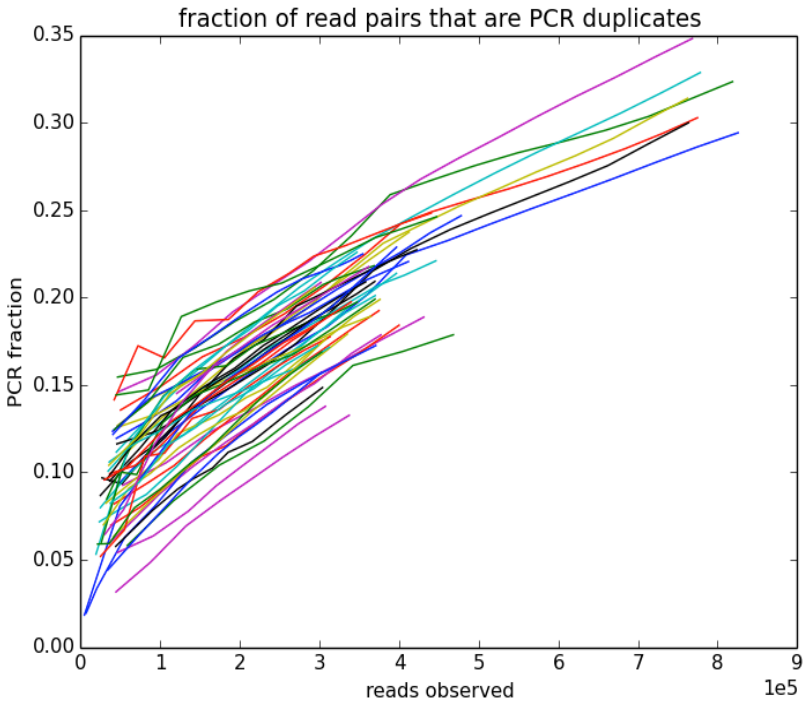
`@identifier:lane:tile:x:y`



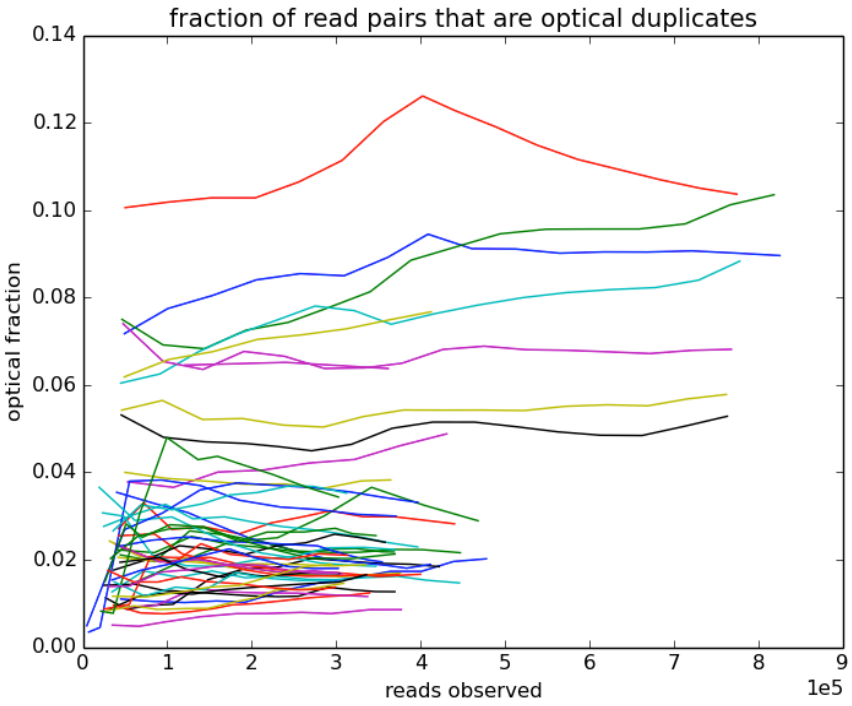
http://www.slideshare.net/jandot/next-generation-sequencing-course-part-2-sequence-mapping
http://www.slideshare.net/cosentia/illumina-gaiix-for-high-throughput-sequencing

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

PCR duplicates



Optical duplicates



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
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)

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	CA	G	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

What this means for downstream analysis

- 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

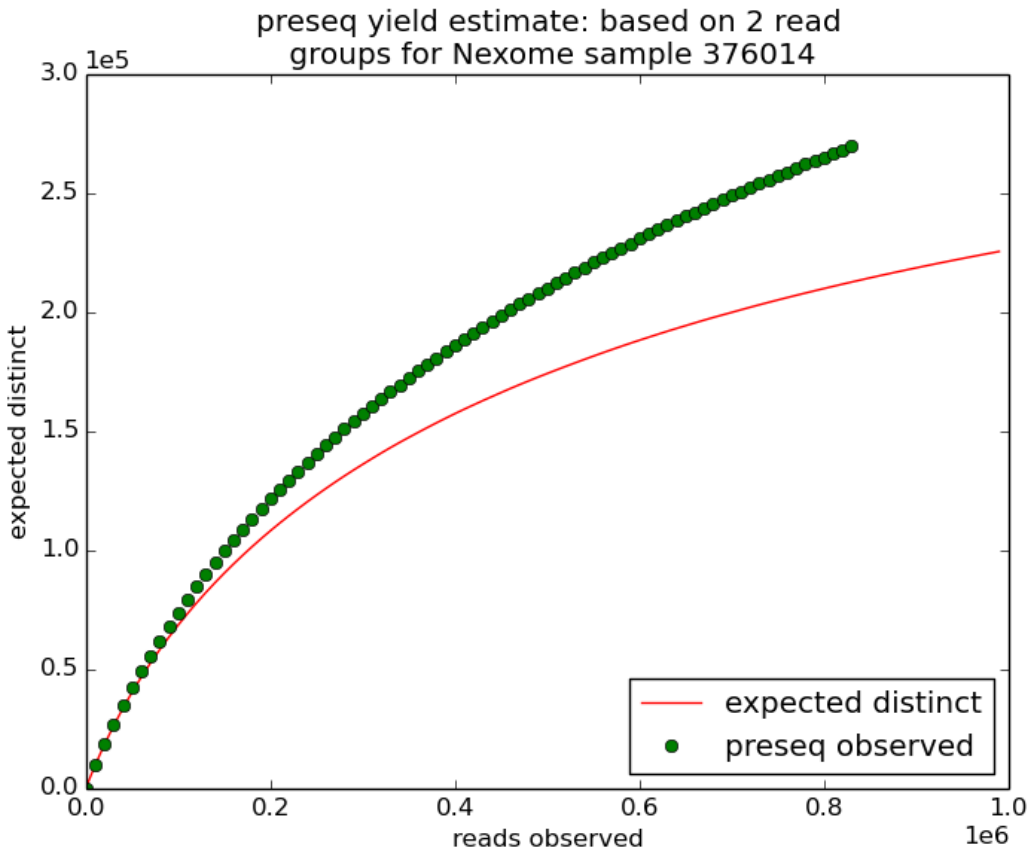
- Amplicon sequencing
  -> all reads start at same position by design

- RNAseq allele-specific expression analysis
  (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 Po(λ). 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 $p_k = e^{-\lambda}\lambda^k/k!$.

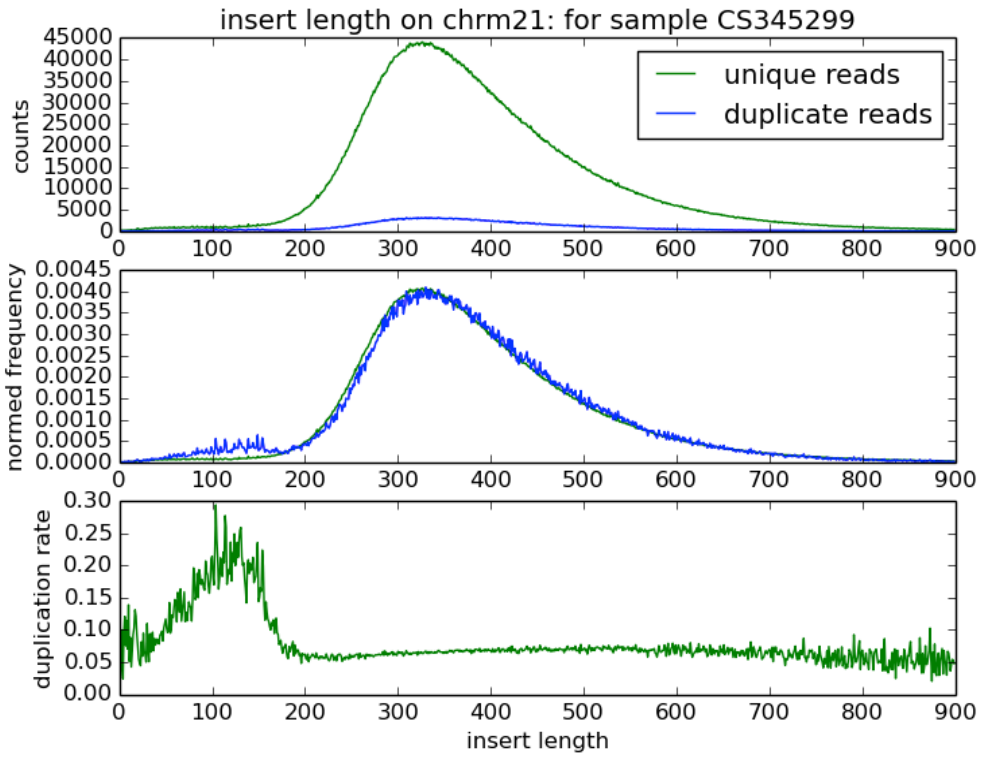
Estimated fraction of duplicates

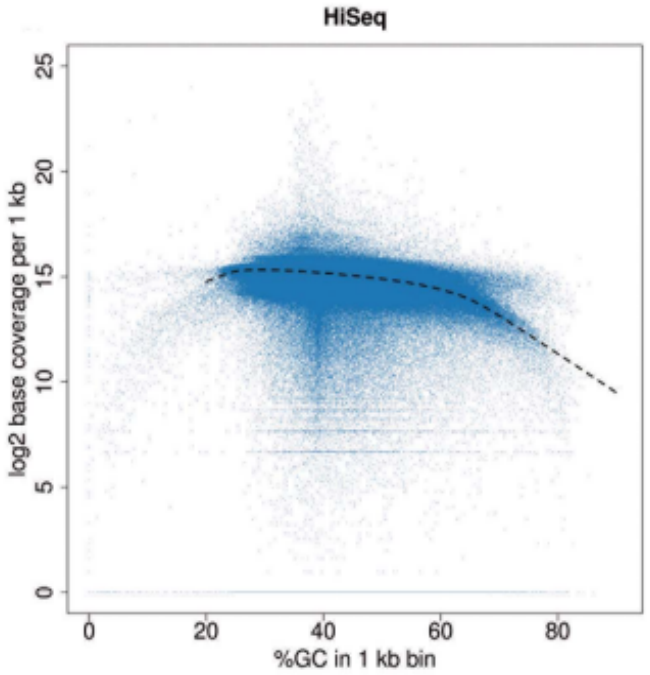
$$d \simeq 1 - \frac{N}{m}\left(1 - e^{-m/N}\right)$$

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



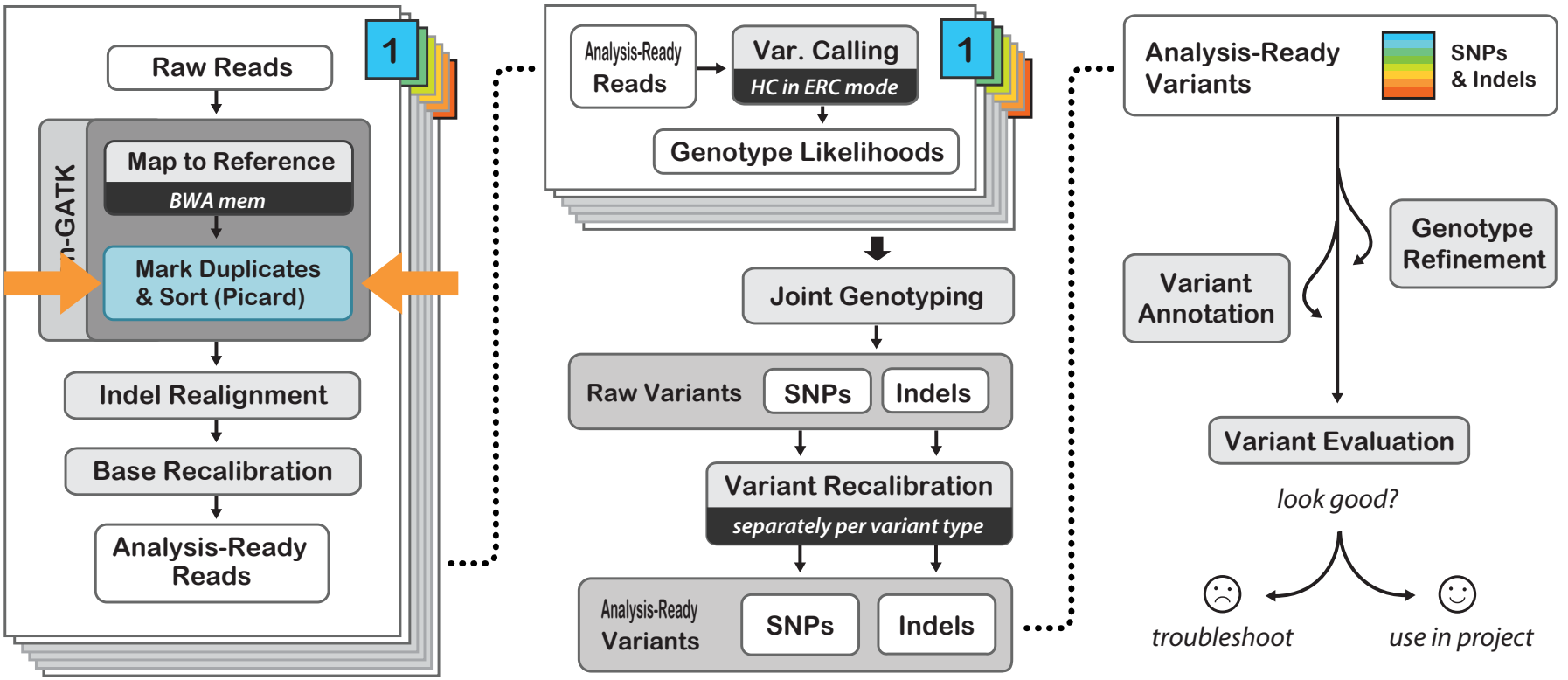


Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies

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

You are here in the GATK Best Practices workflow for **germline variant discovery**



Further reading

http://www.broadinstitute.org/gatk/guide/best-practices

http://broadinstitute.github.io/picard/