BROAD INSTITUTE

1000G indel validation: experimental design

Eric Banks
with help from Mark DePristo, Heng Li, & Ryan Poplin

Genome Sequencing and Analysis
Medical and Population Genetics Program
Broad Institute of Harvard and MIT
August 29, 2012

Experimental Design

- Importantly, we should be clear up front that we really are validating **sites** (or regions) and not **alleles** in this experiment. The plan is to use PCR to amplify 100bp regions +/- 100bp flanks (i.e. 300bp amplicons) so that theoretically multiple indel sites will get covered in any target region.

- Therefore, as long as indel calls fall within 100bp of each other, they can safely be considered overlapping and merged into a single region/interval. This means that we can skirt around the whole complicated allele overlap issue (which would require nightmarish haplotype-based resolution of all the calls).

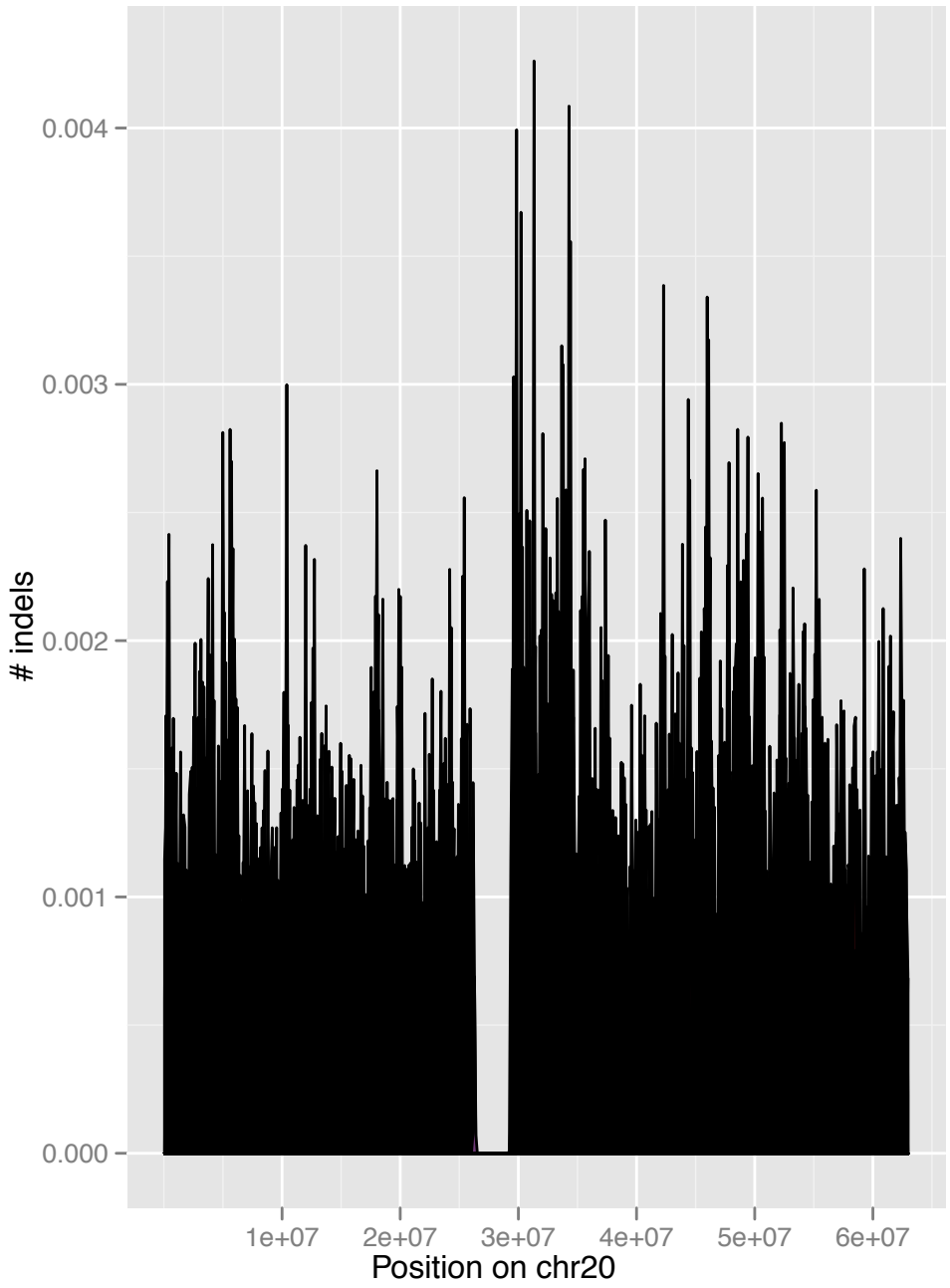
- We throw out any merged regions that are greater than 100bp (because otherwise we would need to mask out some of the indels when designing the amplicons). This set comprises < 5% of the total number of regions.

- The following is the list of the 4 LWK samples that will be used for validation. Note that all of their LC and exome sequencing is Illumina (the corresponding sequencing centers are given in parentheses):
  NA19311 (LC: Sanger, Ex:WashU)
  NA19332 (LC: Broad, Ex: Broad)
  NA19385 (LC: Illumina, Ex: WashU)
  NA19457 (LC: Sanger, Ex: WashU)

Site Selection

- Ideally we would like to take the union of all 7 input callsets, but we don't want any one center to dominate the selection process (i.e. if it made an excessive number of center-unique calls), so we decided to cap the number of unique calls permitted for any given center at 25.

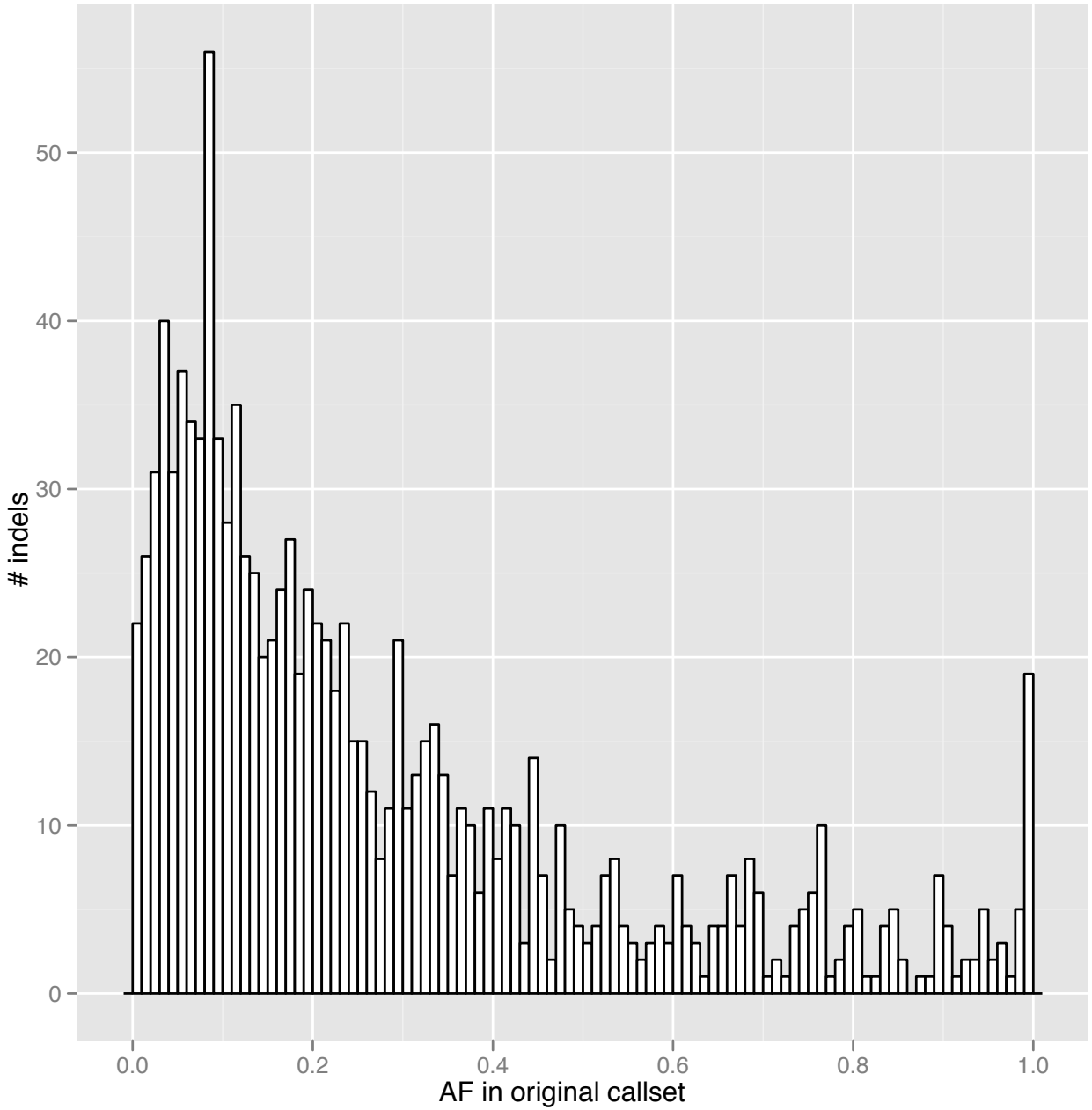
- We also want to ensure that the list of variants to validate is not dominated by high frequency events (it should model the original AF spectrum). Note that this is non-trivial given the experimental design because each region could comprise multiple variants.



The merged indel regions from which we select our target 250 appear evenly distributed over chr20

Summary of center contributions for the 250 indel validation sites

Center	% of total regions that include calls from this center	Number (%) of validation regions that include calls from this center	Number of validation regions that represent center-unique calls
Oxford_Cortex	16%	44 (18%)	2
Pindel	32%	90 (36%)	0
BC	44%	126 (50%)	5
Oxford_Platypus	57%	126 (50%)	21
Broad_assembly	52%	138 (55%)	5
Sanger	54%	149 (60%)	11
Broad_mapping	73%	192 (77%)	22



The distribution of allele frequencies of all indels called within the 250 validation regions approximately models the original spectrum