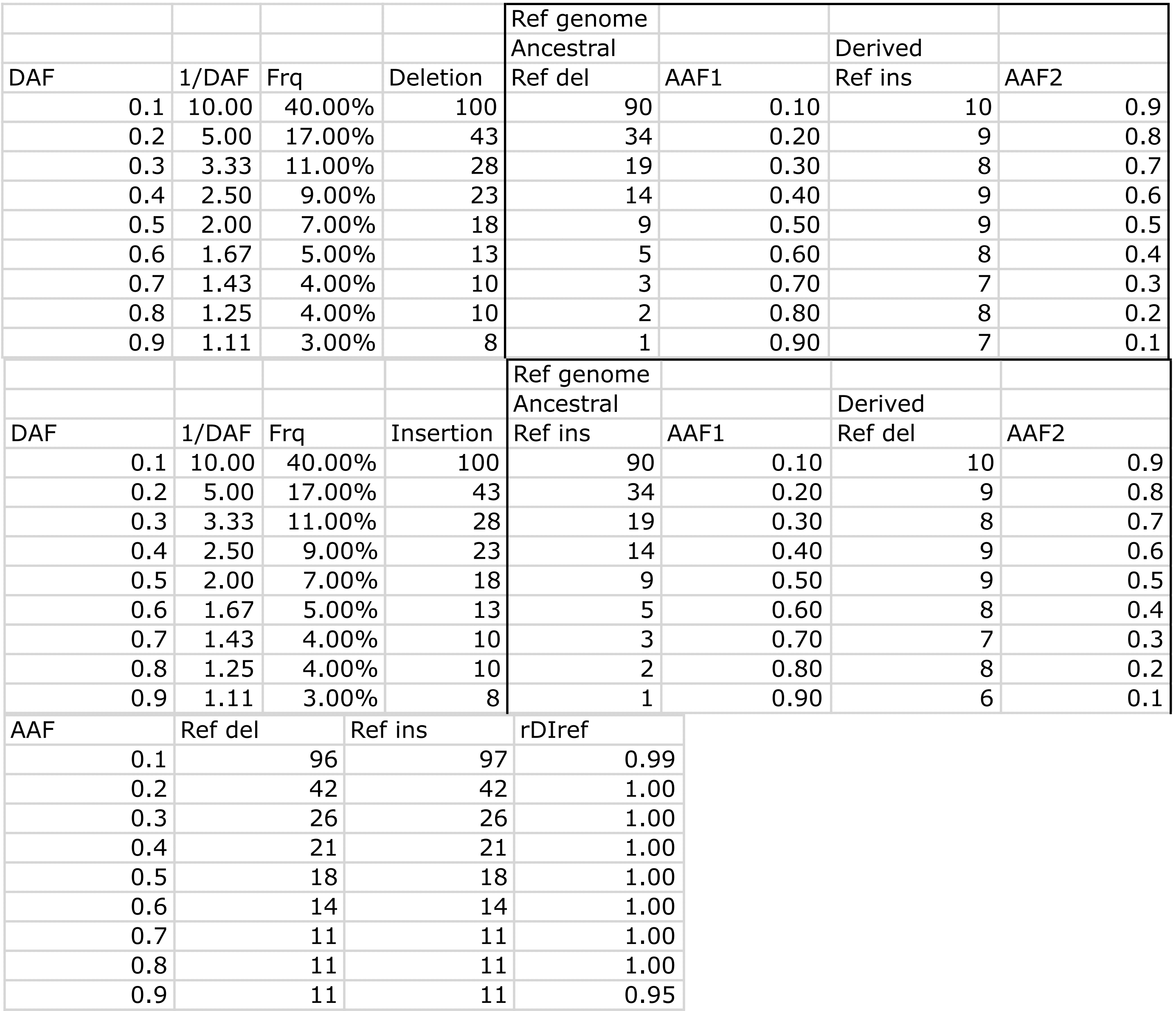
rDI (ratio of deletion to insertion) as a function of AAF

- Gerton proposed the deletion to insertion ratio (rDI) as a proxy for call quality
- For a given mutation, the probability that the reference allele is the derived allele is equal to the alternate allele frequency (AAF). This is true for insertions and deletions.
- Under the neutral model, the distribution of AAF across all polymorphic loci is the same for insertions and deletions. However, if the 'true' rDI (rDI_t) deviates from 1, this will impact the relative number of insertion and deletion loci for which the reference allele is ancestral. Such that, the rDI_{ref} (ratio reference del to ref ins) will vary as a function of both the AAF and rDI_t.
- We expect the distribution of rDI_{ref}as a function of AAF to vary for contexts with different rDI_t.
- To test for calling errors, we can contrast the distribution of rDI_{ref} for known indels, i.e. identified by independent methods, to those identified by 1K genomes (novel).

Example: relationship DAF to AAF for indels is complex

- To illustrate the relationship between AAF and DAF, imagine a scenario in which we have N=200 indels at a frequency of 0.10
- We observe that the rDI_t is 1:1, therefore we have 100 deletions and 100 insertions
- Given that the probability that the reference allele is ancestral is simply 1-allele frequency, we estimate that of the 100 deletions, 90 are reference deletions and 10 are reference insertions.
- However, the 90 reference deletions will be associated with an AAF of 0.10, and the 10 reference insertions will be associated with an AAF of 0.9
- Similarly, of the 100 insertions there wil be 90 reference insertions assicated with an AAF of 0.10, and 10 reference deletions at an AAF of 0.9
- Therefore, at AAF of 0.10, the rDI_{ref} is 1.
- We can extend this logic to all frequencies, and calculate the number of reference deletions and reference insertions per AAF.
- In the case where the rDI_t is 1:1, we will have an rDI_{ref} of 1 at all AAF.

Example : relationship DAF to AAF for $rDI_t = 1$



				Ref genome			
				Ancestral		Derived	
DAF	1/DAF	Frq	Deletion	Ref del	AAF1	Ref ins	AAF2
0.1	10.00	40.00%	100	90	0.10	10	0.9
0.2	5.00	17.00%	43	34	0.20	9	0.8
0.3	3.33	11.00%	28	19	0.30	8	0.7
0.4	2.50	9.00%	23	14	0.40	9	0.6
0.5	2.00	7.00%	18	9	0.50	9	0.5
0.6	1.67	5.00%	13	5	0.60	8	0.4
0.7	1.43	4.00%	10	3	0.70	7	0.3
0.8	1.25	4.00%	10	2	0.80	8	0.2
0.9	1.11	3.00%	8	1	0.90	7	0.1

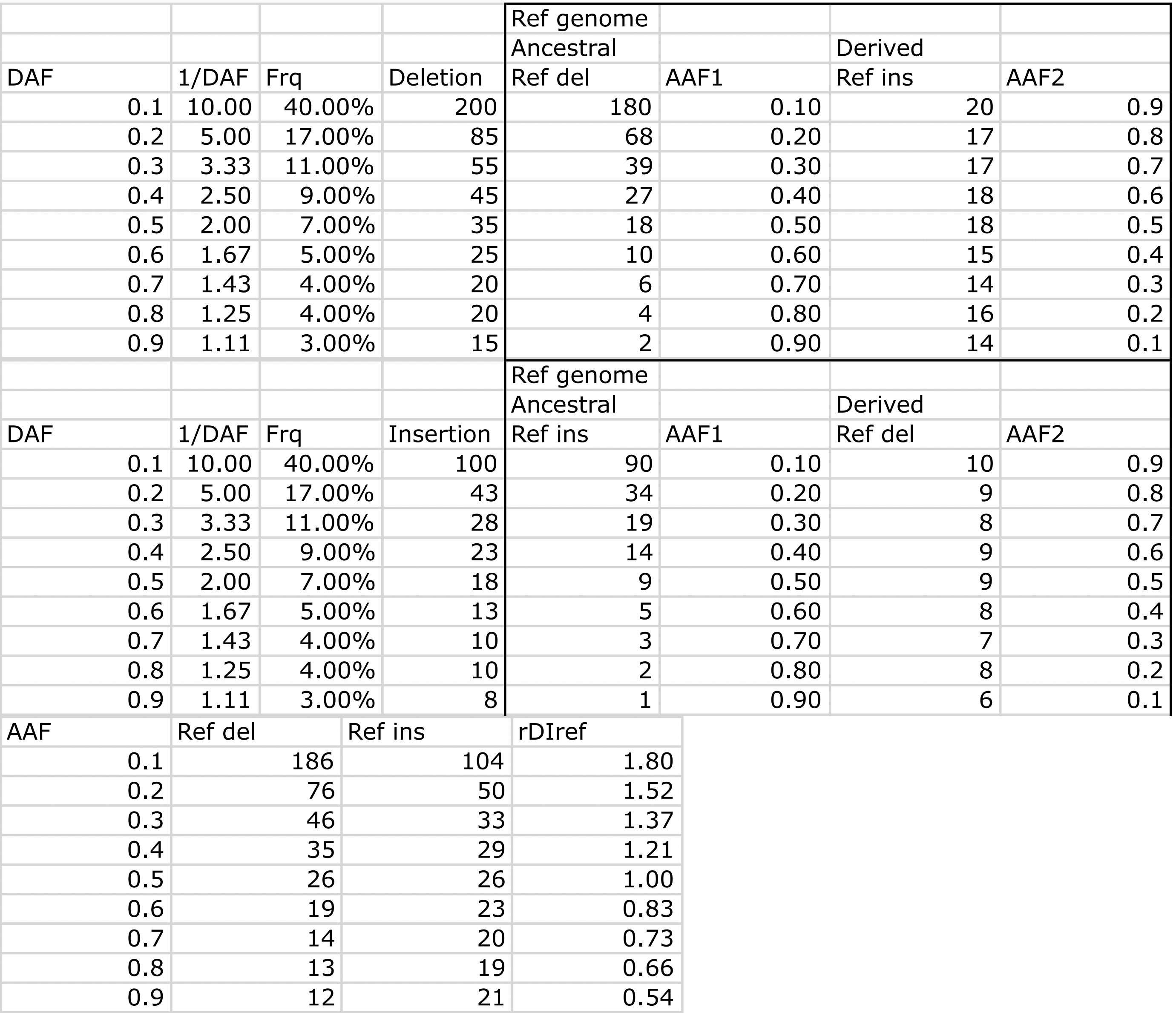
				Ref genome			
				Ancestral		Derived	
DAF	1/DAF	Frq	Insertion	Ref ins	AAF1	Ref del	AAF2
0.1	10.00	40.00%	100	90	0.10	10	0.9
0.2	5.00	17.00%	43	34	0.20	9	0.8
0.3	3.33	11.00%	28	19	0.30	8	0.7
0.4	2.50	9.00%	23	14	0.40	9	0.6
0.5	2.00	7.00%	18	9	0.50	9	0.5
0.6	1.67	5.00%	13	5	0.60	8	0.4
0.7	1.43	4.00%	10	3	0.70	7	0.3
0.8	1.25	4.00%	10	2	0.80	8	0.2
0.9	1.11	3.00%	8	1	0.90	6	0.1

AAF	Ref del	Ref ins	rDIref
0.1	96	97	0.99
0.2	42	42	1.00
0.3	26	26	1.00
0.4	21	21	1.00
0.5	18	18	1.00
0.6	14	14	1.00
0.7	11	11	1.00
0.8	11	11	1.00
0.9	11	11	0.95

Example2: relationship DAF to AAF for indels when rDI is not 1:1

- Now, lets say that we have N=300 indels at a frequency of 0.10
- We observe that the rDI_t is 2:1, therefore we have 200 deletions and 100 insertions
- We estimate that of the 200 deletions, 180 correspond to reference deletions at an AAF of 0.10, and 20 are reference insertions at an AAF of 0.90.
- Similarly, of the 100 insertions there wil be 90 reference insertions assicated with an AAF of 0.10, and 10 reference deletions at an AAF of 0.9
- Here, at AAF of 0.10 we calculate the rDI_{ref} to be 1.7. At AAF of 0.9, we calculate the rDI_{ref} is 0.5
- We can extend this logic to all frequencies, and calculate the number of reference deletions and reference insertions per AAF.
- In the case where the rDI_t is not 1, the rDI_{ref} depends not only on the AAF but the rDI_t as well

Example : relationship DAF to AAF for $rDI_t = 2$

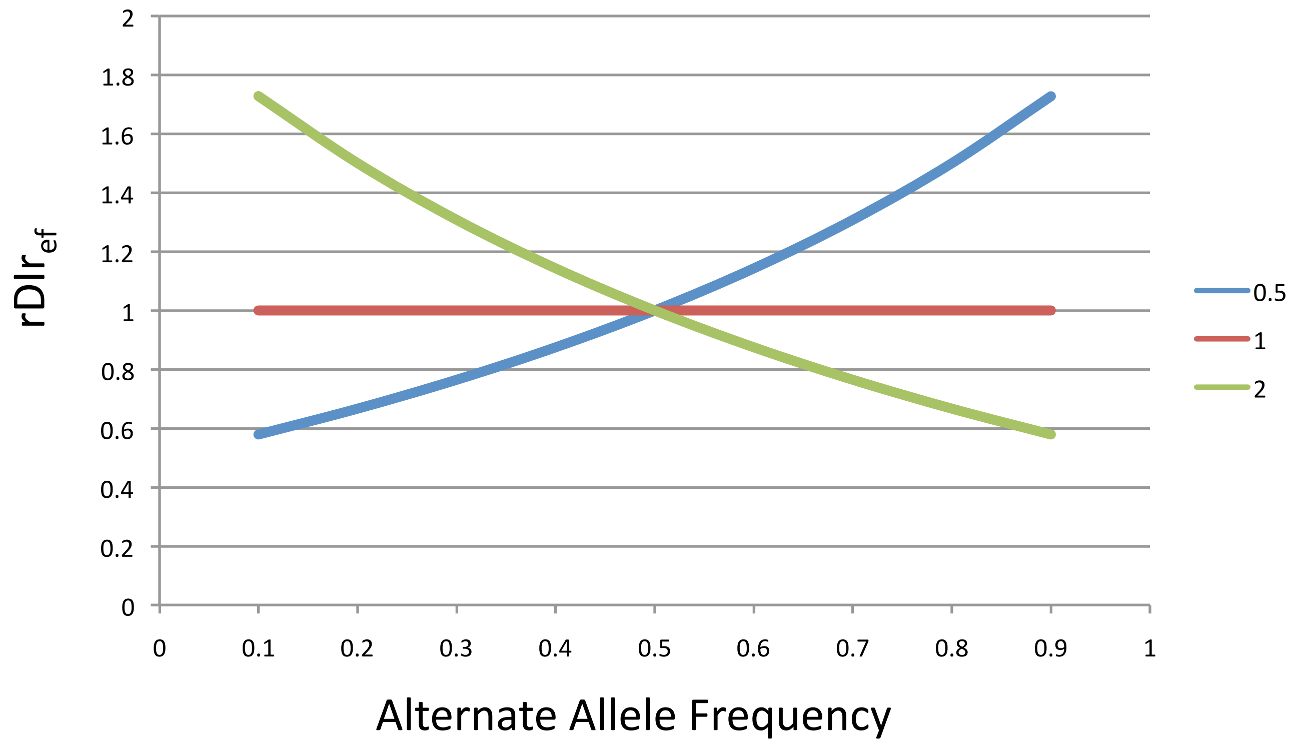


				Ref genome			
				Ancestral		Derived	
DAF	1/DAF	Frq	Deletion	Ref del	AAF1	Ref ins	AAF2
0.1	10.00	40.00%	200	180	0.10	20	0.9
0.2	5.00	17.00%	85	68	0.20	17	0.8
0.3	3.33	11.00%	55	39	0.30	17	0.7
0.4	2.50	9.00%	45	27	0.40	18	0.6
0.5	2.00	7.00%	35	18	0.50	18	0.5
0.6	1.67	5.00%	25	10	0.60	15	0.4
0.7	1.43	4.00%	20	6	0.70	14	0.3
0.8	1.25	4.00%	20	4	0.80	16	0.2
0.9	1.11	3.00%	15	2	0.90	14	0.1

				Ref genome			
				Ancestral		Derived	
DAF	1/DAF	Frq	Insertion	Ref ins	AAF1	Ref del	AAF2
0.1	10.00	40.00%	100	90	0.10	10	0.9
0.2	5.00	17.00%	43	34	0.20	9	0.8
0.3	3.33	11.00%	28	19	0.30	8	0.7
0.4	2.50	9.00%	23	14	0.40	9	0.6
0.5	2.00	7.00%	18	9	0.50	9	0.5
0.6	1.67	5.00%	13	5	0.60	8	0.4
0.7	1.43	4.00%	10	3	0.70	7	0.3
0.8	1.25	4.00%	10	2	0.80	8	0.2
0.9	1.11	3.00%	8	1	0.90	6	0.1

AAF	Ref del	Ref ins	rDIref
0.1	186	104	1.80
0.2	76	50	1.52
0.3	46	33	1.37
0.4	35	29	1.21
0.5	26	26	1.00
0.6	19	23	0.83
0.7	14	20	0.73
0.8	13	19	0.66
0.9	12	21	0.54

Estimated distribution of rDI_{ref} as a function of AAF and rDI_t



$$(1) rDI_{ref}(p) = \frac{(1-p)rDI_t + p}{(1-p) + prDI_t}$$

Where p is equal to Alternate allele frequency, and rDI_t is the ratio of ‘true’ deletion:insertion

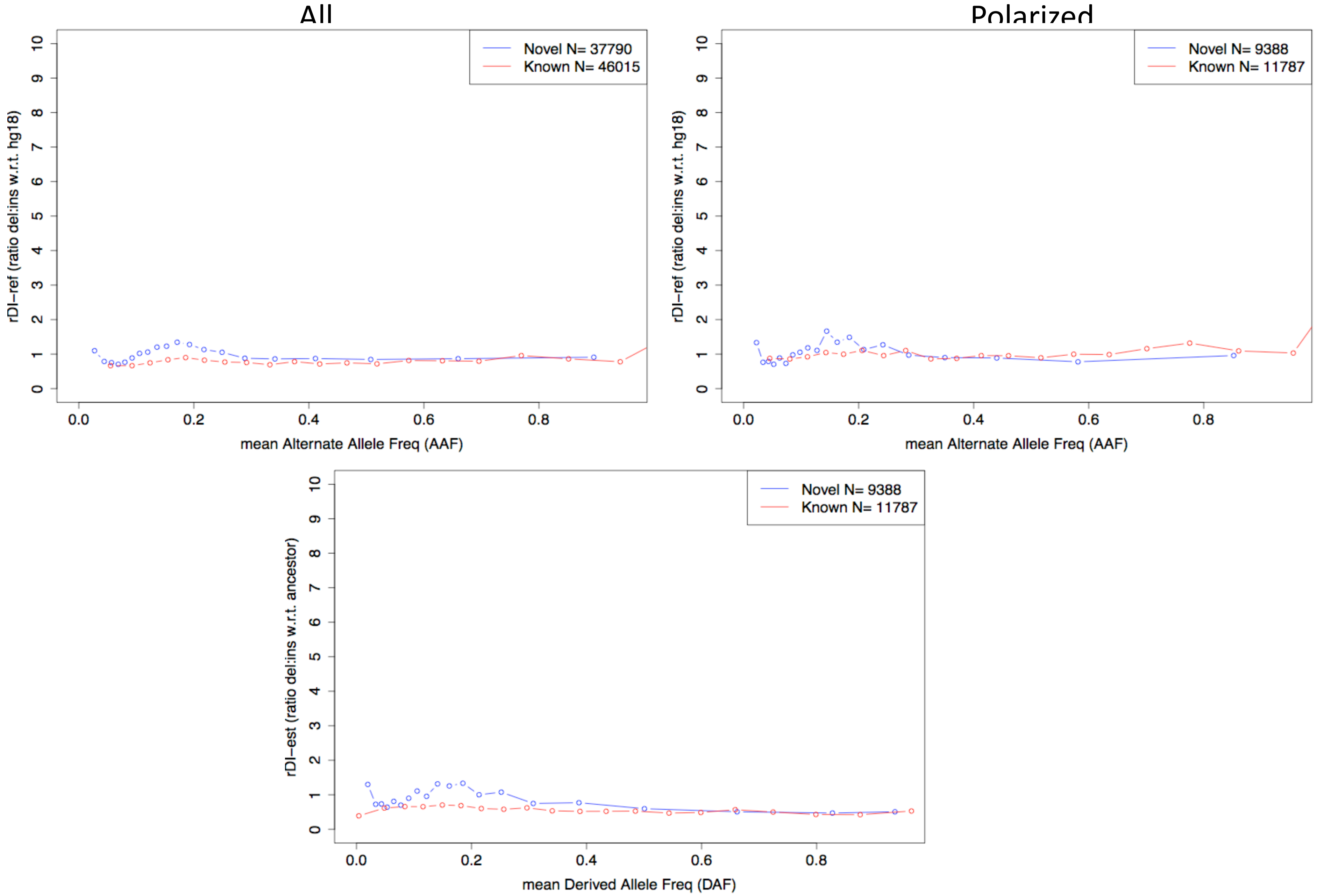
rDI (ratio of deletion to insertion) as a function of DAF

- According to neutral model, wherein the probability of fixation is equal for insertions and deletions, the DAF reflects the age of mutation. Hence, we expect the rDI-est (ratio estimated del to estimated insertion, based on polarization) to be constant irrespective of DAF.
- If polarization errors, we expect to see deviation from expected distribution; polarization errors may also be context-dependent.
- Differences in rDI-est for known vs. novel indels, which are polarized using the same methodology, also likely reflect calling errors.

Methods to investigate rDI (ratio of deletion to insertion) as a function of AAF, DAF

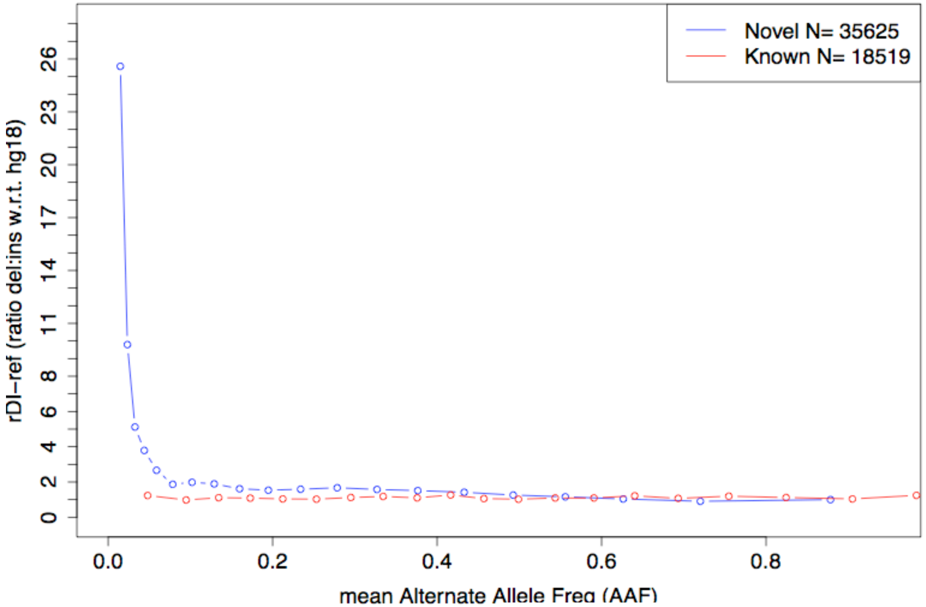
- Focus on indels in ancestral repeats (ARs), to avoid selection
- Only polymorphic indels considered (alternate allele observed in at least 1 individual per pop)
- Distinguish Novel vs. Known (within 50-bp of dbSNP allele of same size and type; according to VCF annotation)
- Indels are partitioned into 10 equal-sized bins according to increasing AAF/DAF frequency
- Read depth obtained from NF and NR annotations

HR context (CEU)

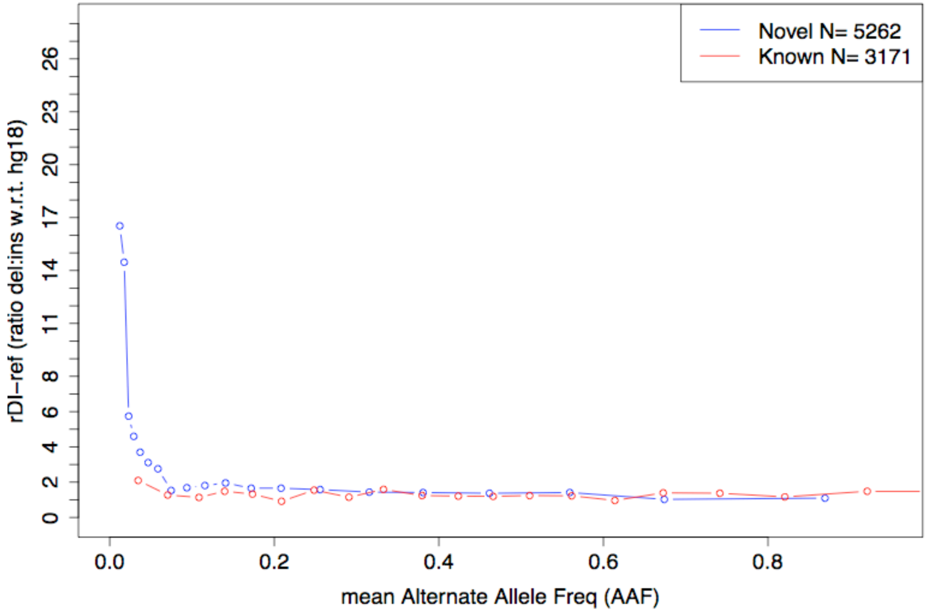


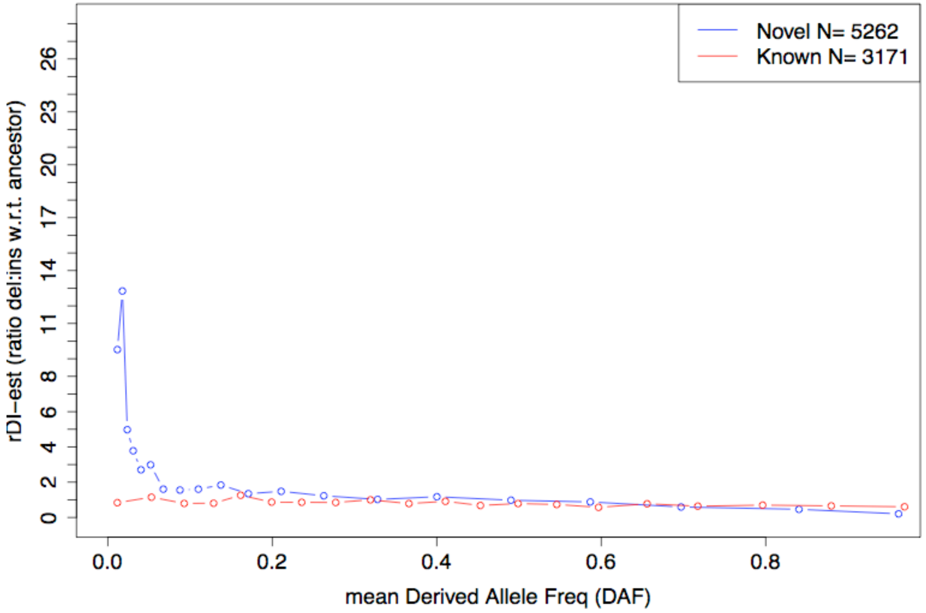
TR context (CEU)

All

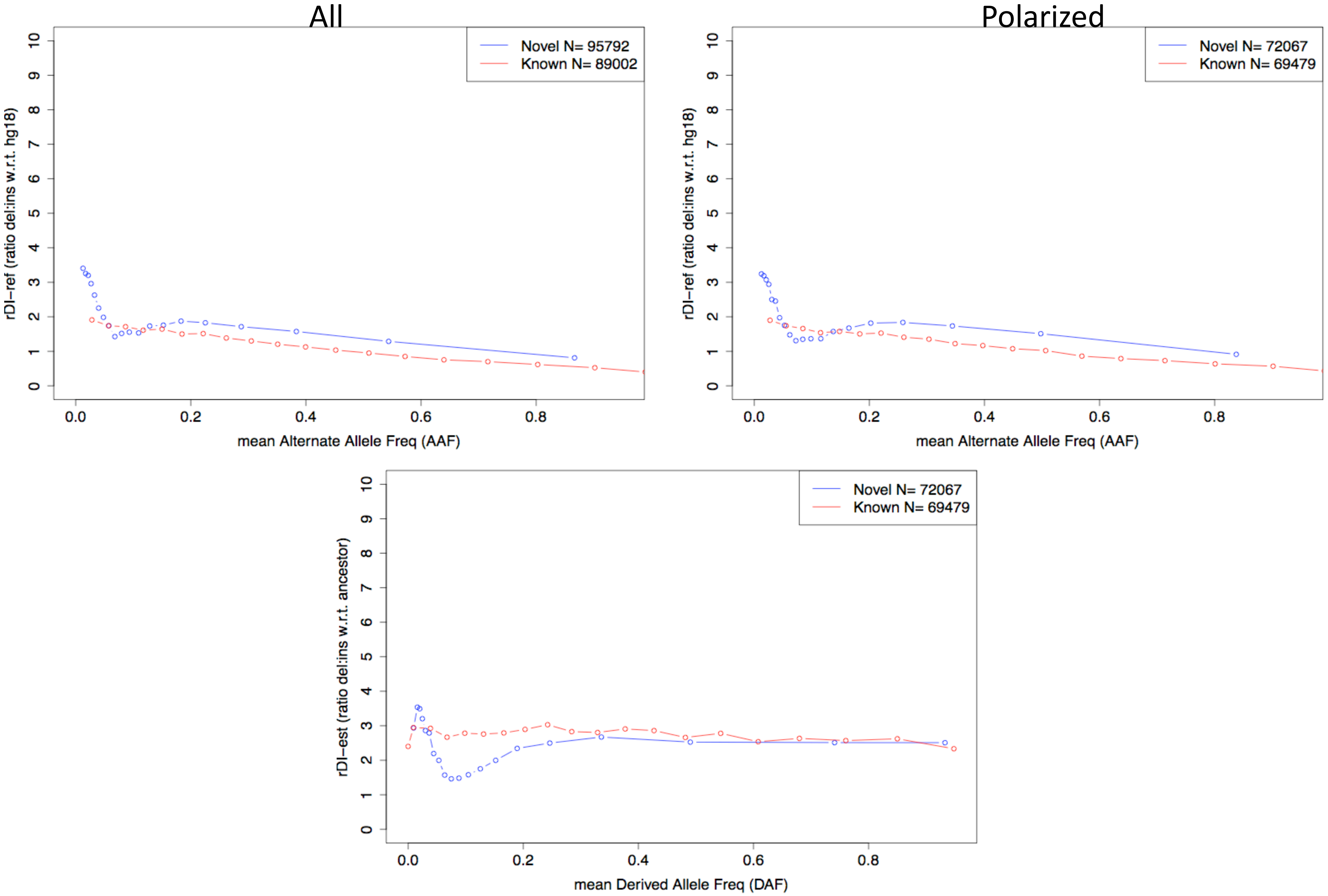


Polarized





NR context (CEU)



Conclusions

- HRs: OK: Novel and Known show very similar distributions of rDI-ref at ~1.
- TRs: Novel indels show a high rDI-ref at low AAF, whereas the distibution for Known events is roughly constant at ~1. This implies that, at low AAF, Novel events have disproportionately high number of either false positive deletions or false negative insertions. Since we don't see the trend for the known ones, we can conclude that the pattern is due to FPs.
- NRs: the distribution of rDI-ref for Novel events appears complex (and differs from the known ones) up to AAF of ~0.15 … is it possible that detection varies with the allele frequency – differently for insertions and deletions?
- Same pattern in the 3 pops

- Read coverage is lower for Novel indels compared to Known ones (expected??)
- This bias is much more pronounced for NR indels >> HR>TR
- The mean read coverage is low (20-60 depending on indel category): is that normal?