Supplementary material

michafla

May 19, 2015

Contents

1 GSNAP/VariantTools pipeline

1.1 GSNAP parameters

We align the reads to the hg19 reference with GSNAP version 2013-03-31, according to the parameters:

```
-M 2 -n 10 -B 2 -i 1 --pairmax-dna=1000 \
--terminal-threshold=1000 --gmap-mode=none \
--clip-overlap
```

We restrict to uniquely aligned reads, where a read is considered multi-mapping if there is at least one additional alignment with a score that is within 2 of the score of the best alignment. When paired ends overlap (ie. when the fragment is shorter than 150 nt) the overlapping region is split between the two ends, but not double counted. Potential PCR duplicates are removed with Picard [Picard, 2014]. We tallied the nucleotides at each position, ignoring indels and base calls of quality below 23.

We then apply the filter based on the binomial likelihood ratio, as described in the main text. It effectively excludes all variants with alt frequency less than 0.04. We guard against false positives that can arise in low coverage regions where a single variant read would exceed our minimum threshold by requiring that at least two different reads contain the variant. We discard any variant overlapping or within 1nt of a homopolymer that is longer than 6nt. Lastly, we filter variants that are clumped together on the chromosome. Figure `fig:2`B and C show the effect of these characteristics on the FDR. More details on the second filter are provided in Section S1.3.

We masked out simple and satellite repeats (derived from the UCSC RepeatMasker track) from the tallies, because mapping to those regions is one of the major sources of error [Li, 2011], and we believe that developing a better understanding of those regions would require specialized sequencing, which is outside of our scope. These regions together cover about 1.6% of the genome.

1.2 Likelihood ratio test

The likelihood ratio method is often used to choose between two competing models. To apply the method one considers the probability of observing the data we did under different probability models. In the case of variant

calling, at any locus where there are variants present one can consider them to have arisen due to changes in the genome, i.e. true variants, or due to sequencing or alignment errors. We, and others, use a Binomial model but this could easily be extended to allow for over dispersion. We let Model 1 (M1) be that the variant arose due to sequencing errors, alignment errors or other errors. While Model 2 (M2) is that the variant is present in the genome being sequenced at some frequency.

So that the likelihood ratio for any particular locus is then proportional to the ratio of the probabilities of the two models:

$$LRT = \frac{P(D|M1)}{P(D|M2)} = \frac{p_1^x(1-p_1)^{n-x}}{p_2^x(1-p_2)^{n-x}},$$

where n is the coverage at the locus and x is the number of variant reads.

We select appropriate choices for p_1 and p_2 and compute the LRT directly at each locus. When the ratio exceeds 1 then the probability for the observed data under M1 is greater and we would choose M1. Similarly if the ratio is less than 1, then the probability of the observed data under M2 is greater and we would choose M2. The LRT is identical to a Bayes factor.

1.3 Filtering out alignment artifacts

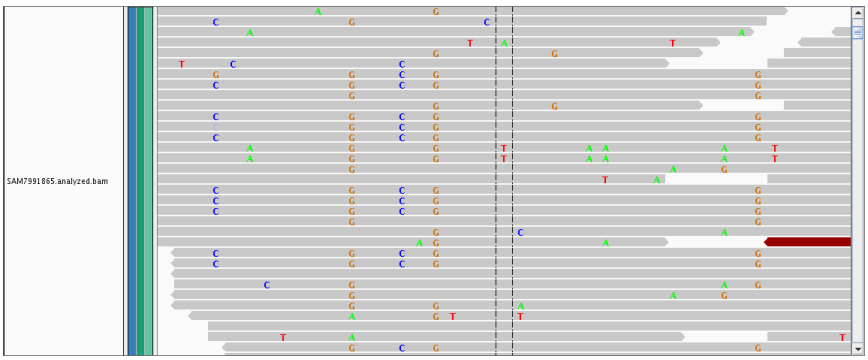

Figure S1: An IGV screenshot of a region enriched for likely erroneous variant calls.

We have identified two factors related to alignment that affect the variant calls: 1) self-similar regions and 2) the presence of indels and/or other structural variants.

When two or more genomic regions are similar, there may not be sufficient information in the read sequence for an accurate alignment. If the regions are

exactly identical in the reference, an aligner will typically discard the read or indicate that the read maps to multiple locations. It is easy to ignore such ambiguous cases; however, the sample genome may differ slightly from the reference leading to localized enrichment of alignment errors. The IGV screenshot in Figure S1 illustrates a region that is enriched for mismatches. We describe how we addressed this problem in Section 7.

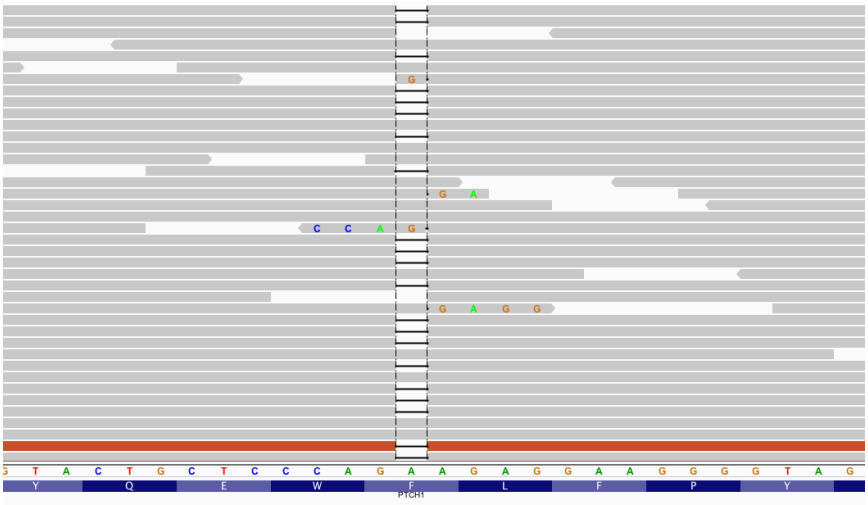

Figure S2: An IGV screenshot illustrating the generation of false SNVs due to mis-alignment of indels.

The second challenge to accurate alignments is the presence of indels and other structural variants. The aligner has difficulty detecting an indel when it is at one of the ends of the read, because there is insufficient information on one side of the event. Figure S2 is an IGV screenshot of alignment errors near an indel. While indel realignment resolves many of these artifacts, we have found that homopolymer regions are particularly challenging. We implemented a filter that excluded variants falling within a homopolymer of length greater than or equal to some minimum. To select the optimal minimum, we fit logistic regression models predicting the true positive status from the homopolymer length (`hp.length`), which we treated as either a linear predictor, or as an indicator for when the length exceeded some minimum, like 7. The latter model uses less information but more directly translates to our cutoff-based filter. The first term in these two formulas is an indicator for whether the variant is overlapping a homopolymer, while the second term treats the length.

Length as linear predictor: $TP \sim I(dtn.hp <= 1) + hp.length$
Length as cutoff indicator: $TP \sim I(dtn.hp <= 1) + I(hp.length > 7)$

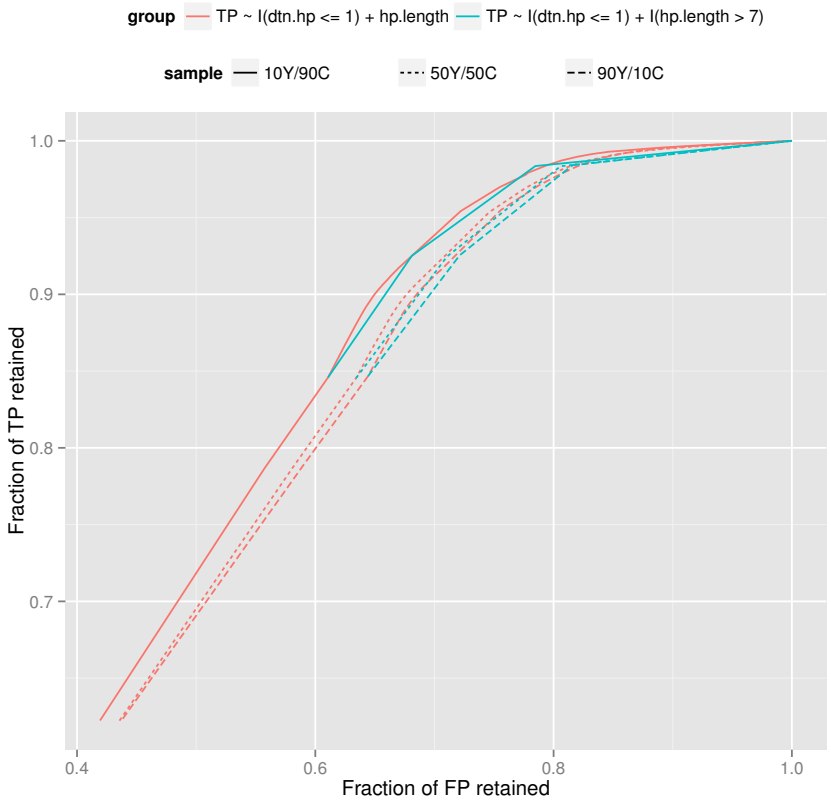

Figure S3: ROC curves comparing the two logistic regression models predicting the true positive status from whether a variant falls within a homopolymer, and the length of the homopolymer. The first model treats the length as a linear predictor, and the second as an indicator for whether the lenth exceeds some cutoff. The second model is simpler, easier to operationalize, and performs comparably.

Figure S3 compares the two models using ROC curves. From the curves, we conclude that the model with the indicator term performs favorably to the model with the linear predictor, and select 7 as our minimum cutoff for the filter.

2 GATK-based pipeline

To align short read data we used the BWA short read aligner 0.5.9 [Li and Durbin, 2009]. Default settings were used with 12 threads, a mismatch penalty of 3, a gap opening penalty of 11 and a gap extension penalty of 4. This algorithm uses the Burrows-Wheeler Transform (BWT) for gapped global alignment and has support for paired end reads. The hg19 human reference genome was used with the BWA aln command. To help reduce PCR amplification biases the Java based Picard package (v 0.68) (`http://picard.sourceforge.net/index.shtml`) was used with default settings running Java 6.0.22 with 12 Gb of memory allocated. In the case of duplicate coordinates the read with the highest mapping quality was retained. To help reduce possible false positive variants caused by poor initial alignments, local realignment was carried out using GATK (v1.0.4906) [DePristo et al., 2011]. For single nucleotide variant calling the GATK UnifiedGenotyper was used with the following flags:

```
-l INFO -mbq 20 -mmq 30 -nt 8 -G Standard -A AlleleBalance
```

For indels the GATK IndelGenotyperV2 was used with the following flag: `-l INFO`. Both the UnifiedGenotyper and IndelGenotyperV2 used DbSNP Build130 [Sherry et al., 2001] using the `-D` flag.

3 Self-chain scores

3.1 Source

To obtain an indicator of whether a region shared similarity with another region in the genome, we downloaded the hg19 self-chain track from the UCSC genome browser database. According to UCSC:

> This track shows alignments of the human genome with itself, using a gap scoring system that allows longer gaps than traditional affine gap scoring systems. The system can also tolerate gaps in both sets of sequence simultaneously. After filtering out the "trivial" alignments produced when identical locations of the genome map to one another (e.g. chrN mapping to chrN), the remaining alignments point out areas of duplication within the human genome.

We denoted any locus where the self-chain score was larger than zero as "self-chained". This definition was used in all figures and calculations that

present self-chain data. About 6% of the genome is self-chained, including 17% of coding regions. The enrichment in coding regions likely reflects homology within gene families.

3.2 Association with coverage

Since we have observed that alignment artifacts often result in coverage that is below or above expectation, we investigated whether the self-chain score for a region was associated with coverage. We extracted the unique and multimapping coverage over coding regions from the 50Y/50C mixture, and segregated the coding regions depending on self-chain status. We then averaged the coverage for each of the contiguous regions in the two sets. As shown in Figure S4, we found that self-chains tend to have lower unique coverage, and that virtually all multimapping coverage (again, over coding regions) is within self-chained regions.

3.3 Association with Platinum Genomes high-confidence regions

The Platinum Genomes consortium has annotated about 70% of the genome as being a source of high-confidence variant calls, while only about 16% of the genomic positions we designated as self-chained were also declared high-confidence. This is confirmation that the self-chained regions tend to generate lower quality variant calls.

4 Extra binomial variation

Other groups have observed extra binomial variation (EBV) in the distribution of sample-wise fractions of reads mapping to genomic features in the context of pairwise sample comparisons [Plagnol et al., 2012] [Gerstung et al., 2012]. We wondered whether EBV was present in our data set. We computed the observed and expected (under the binomial) variance in allele frequencies for each coverage from 50 to 80. The expected variance was computed from $np(1-p)$, where n is the coverage and p is the sample mean of the frequencies for the coverage value. We used the sample mean instead of the theoretical frequency, because there was significant error in the actual mixture proportions.

As shown in Figure S5, we found in the 50Y50C YRI-specific variants that the EBV was very small and was largely independent of genomic context. However, we did notice slight overdispersion in regions with very low

coverage, which suggested that there might be an association between EBV and mappability. To investigate, we again used the YRI-specific hets in 50Y50C sample conditional on their self-chained status, as shown in Figure S6. Unsurprisingly, the variation in the self-chained regions is much higher than expected under the binomial model. While for those variants with self-chain score of zero the there is little evidence of over-dispersion.

5 GSNAP/VT and BWA/GATK frequency concordance

While there are discrepancies at specific positions, the overall concordance between the GSNAP and BWA variant frequencies was very good, as shown in Figure S7.

6 CEU deletion analysis

We extracted the validated CEU deletions [Mills et al., 2011] from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/paper_data_sets/companion_papers/mapping_structural_variation/MasterValidation.Pilot2.all.leftmost.061510.txt` and compared the YRI-only SNVs inside those deletions to the rest in terms of their coverage and alt frequencies.

To investigate these trends for individual variants, we restricted our analysis to a single validated deletion at chr14:80106289-80115049 (8761nt) that harbored 14 unique YRI-specific variants (2 het, 12 hom) In Figure S8, we observe increasing coverage with increasing YRI mixture proportion.

7 Weighted neighborhood score filter

We observed that in some cases false positive calls tended to cluster while true positive calls rarely do. This suggests some form of filtering calls when the local density of variants is too high. A simple filter would rule out any call with more than say 3 or 4 non reference calls in a 75nt window. However, in practice the method did not provide sufficient flexibility.

For any variant we identified all called neighboring variants within 75nt (37nt in the 5-prime direction and 37nt in the 3-prime direction). We then computed $S = \sum_j 1/\sqrt{d_j}$, where the sum is over all neighboring variants and d_j is the distance to the j^{th} one. We then used $S = 0.1$ as the threshold

so that variants where $S > 0.1$ we identified as FPs. See the next section for the rationale.

7.1 Rationale

To develop a score representative of local neighborhood density, we first examined the local decay function by fitting a logistic regression to the the data, treating those variants in the reference genome as TPs, and all others as FPs. We first fit

$$Y = \sum_{j=1}^{37} \beta_j X_j,$$

where X_j is an indicator of whether there were any called variants at a distance j from the variant of interest. Since fitting logistic regression models to these many data is prohibitive, we needed a simpler formula. Fitting the regression separately to the three mixtures and then plotting the estimated β_j showed that the coefficients dropped off at a rate that was reasonably approximated by a quadratic function, so we adopted the simplified approach $S = \sum_j 1/\sqrt{d_j}$.

To select the cutoff on S, we generated an ROC curve using the concordance with the published genotypes. From that, we determined that 0.1 corresponded to an acceptable compromise between sensitivity and specificity.

7.2 Left and right sided counts

We suspected that loci near the edges of a region with a high number of variants might not be handled well. They would have only half the right number of non-reference calls in their window. To investigate this, we made separate counts to the left and right of the target and then worked with min(L, R) and max(L, R) as the features. This made no material improvement to the ROC curve.

7.3 Cauchy-logistic regression

One key aspect was the use of logistic regression with a Cauchy link function. The logistic model is not a perfect fit. If for example our calls of TP and FP were incorrect then we anticipate that the variance will be somewhat inflated. Using a quasibinomial did not ameliorate the situation, but rather yielded absurdly high overdispersion values (such as 10^9) and no significant coefficients. We also explored the `family = quasibinomial(link=cauchit)`.

The slow decay of the Cauchy tails mimics the use of a mixture with epsilon probability of a bad label, but without requiring all the EM machinery needed to fit a mixture.

7.4 Association with mapping issues

While we have empirically shown that excluding variants in crowded neighorhoods is a useful trade-off between sensitivity and specificity, it is also important to understand the underlying error process. We suspect that the error is mostly mapping artifact, and Figure S9 supports this by showing an enrichment for self-chains in the regions with high (> 0.1) score compared to the rest.

8 Undercalling in the 1000G YRI genotypes

As shown in Figure S11, the frequencies for the CEU-specific variants in the 90Y/10C tend to be much higher than expected. Figure S10 shows that the frequency distributions have modes that match what one would expect for variants that are het or hom in the YRI while also present in the CEU. While we observed this presumed undercalling for YRI, we did not observe the phenomenon for CEU.

The 1000G applied a sample-specific mask prior to calling genotypes, as documented here: `http://www.1000genomes.org/faq/why-only-85-genome-assayable`. In summary, there were four main reasons for masking: no coverage, abnormally high coverage, more than 20% of the reads had a MAPQ of zero, and N in the reference. For the YRI, 78% of the masked 1000G calls fell into the zero MAPQ category. We found that 70% of our discordant calls that tracked the YRI mixture proportion fell into the mask, and 97% of those were in the zero MAPQ regions.

9 Predicted genotypes

Due to our suspicions that the published genotypes, in particular those of the YRI, were incomplete, we developed a strategy to distinguish actual FPs from the true variants that were missing from the genotypes. For each variant, we predicted a genotype from the cross-mixture alt frequency trend. Each predicted genotype is a combination of the source genotypes, which we encode in the format YRI/CEU, where 0 indicates wildtype, 0.5 indicates heterozygous and 1 indicates homozygous alt. For example 0.5/0 would be

heterozygous in YRI and wildtype in CEU, and 0/1 would be wildtype in YRI and homozygous alt in CEU. If an observed frequency trend tracks any of the trends expected for a true variant, we claim that the variant may indeed be real. When the trend does behave as expected, we mark the genotype as no-call (NC) and consider it very likely to be a FP.

Table S1: Tabulation of the base changes for the FPs, conditioned on whether an FP was classified as tracking or non-tracking. The transitions are highlighted in bold. The Ti/Tv was closer to expectations (2.15) for the tracking vs. the non-tracking (1.37), in agreement with our claim that the tracking variants are likely to be TPs.

Change	Non-tracking	Tracking
A/C	9691	16819
A/G	19504	64855
A/T	7746	14808
C/A	9674	22396
C/G	7448	22133
C/T	27482	99095
G/A	26709	98671
G/C	7317	21823
G/T	9798	22293
T/A	7690	14713
T/C	19843	64337
T/G	8503	16936

To predict a genotype from a frequency trend, we first binned the frequencies according to the red dashed lines in Figure S11. We then classified the variants into predicted genotypes based on the bin values. Given the evident variability in the 50Y/50C mixture, we allowed for variants to spill over into adjacent bins. For example, variants in the bottom bin in the 10Y/90C mixture, the second bin the 50Y/50C and middle bin in 90Y/10C were classified as heterozygous (0.5) in the YRI and wildtype/no-call (0) in the CEU. Any call that did not match a pattern corresponding to one of the 8 possible genotype combinations was marked as no-call (NC). These are very likely to to be FP. See Figure 4B in the main text for examples of variant frequency trends that correspond to each predicted genotype.

In Figure S12A, we find that the NC variants unsuprisingly have the highest FDR (about 0.6) in terms of concordance with the published genotypes. We observe the second highest FDR (about 0.2) for the predicted

genotypes corresponding to het/hom in the YRI and wildtype in the CEU. This agrees with the notion that a significant number of alt calls are missing from the YRI genotypes. We also found that the transition to transversion ratio (Ti/Tv) better matched expectations for the tracking (2.15) than the non-tracking (1.37); see Table S1.

Figure S12B demonstrates an association between the predicted genotype and the self-chain status, an indicator of mapping artifact. This is further evidence that the frequency trend is useful as a predictor of whether a call is correct and helps fill the gaps in the published individual genotypes, in particular the 1000G calls for the YRI.

We were surprised by the number of variants that appeared 0/0 in our data, but were called as heterozygous in the CEU (0/0.5). Figure S13 shows that CEU hets that VariantTools missed, as well as CEU hets that show a WT frequency trend, are much more likely to be inside self-chain regions than calls with well-behaved frequencies that were detected by VariantTools. This puts those calls in question.

Figure S14 shows the self-chain fraction for each predicted genotype, restricted to variants from the 50Y/50C sample in coding regions and conditional on whether the call is concordant with the published genotypes (TP) or not (FP). FP sites are much more likely to be self-chained, although the trend is not as strong for the YRI-alts, likely because they are enriched for true calls that were excluded from the published set.

10 Validation

Table S2: Validation candidates cross-tabulated by whether we were able to call a variant genotype from the GSNAP frequency trends (rows) and the genotype as callled by Sequenom. The results matched our expectations, with 97% of tracking sites called as variant, and 74% of the non-tracking sites validating as wildtype.

Pattern	Variant	Wildtype	No call	Total
Non-Tracking	10	28	6	44
Tracking	66	2	18	86

When calculating our FDR, we rely on the assumption that variants with frequencies that are concordant with the expected trend across the mixtures are true positives, despite their omission from the published genotypes. To validate this argument, we selected a subsample of FP variant calls, i.e., those

that were detected by VariantTools but were not present in the published genotypes. We stratified by whether we were able to predict a genotype from the GSNAP frequencies, with the expectation that variants with called genotypes would validate as variant, and those for which we were unable to make a call would validate as wildtype.

Our initial attempts at Sequenom-based validation failed, with almost all failures occuring at the primer design phase. Only about 27% of the primers succeeded. We encountered similar issues with a dual-primed Sanger-based approach. To facilitate primer design, we therefore required that all of the sites were located inside of coding exons and outside of self-chain regions. While this introduces bias, the bias is preferable to the priming bias, since we are enriching for sites that are functional.

Table S3: Confusion matrix of the predicted genotypes (rows) vs. Sequenom-validated genotypes (columns) for a subsample of FP calls. Concordance was 91% excluding the no-call sites.

Pred/SQM	0/0	0/0.5	0/1	0.5/0	1/0	0.5/0.5	0.5/1	1/0.5	1/1
NC	28	0	0	5	0	2	1	0	2
0/0.5	0	3	1	0	0	0	0	0	0
0/1	0	0	1	0	0	0	0	0	0
0.5/0	2	0	0	44	1	1	0	0	0
1/0	0	0	0	1	8	0	0	0	0
0.5/0.5	0	0	0	0	0	2	0	0	0
0.5/1	0	0	0	0	0	0	0	0	0
1/0.5	0	0	0	0	0	0	0	1	0
1/1	0	0	0	0	0	0	0	0	3

We interrogated the sites with Sequenom technology, and the results shown in Table S2 are consistent with our expectations. About 18% of the sites were uncallable by Sequenom due to primer design difficulties and thus excluded from this analysis. This filtering was not significantly associated with whether we were able to predict a genotype from the cross-mixture frequency trends. Of the sites called by Sequenom, there were 68 tracking sites and 38 non-tracking sites. Based on the counts in Table S2, the positive predictive value (PPV) was 0.97 (66/68) for tracking sites validating as variant and 0.74 (28/38) for the non-tracking sites validating as wildtype (i.e., sites that are actually FPs). Table S3 cross-tabulates the predicted and validated genotypes. Concordance between the two sets of genotypes, where both made a call, was 91%. There were 10 non-tracking sites that

Sequenom called as variant. Of those, 9 were present at low frequency in all samples were they were detected (and thus are difficult to interpret with our reference-based calling), and the last one is the result of misalignment due to a proximal indel.

Table S4: Confusion matrix of the predicted genotypes (rows) vs. Sequenom-validated genotypes (columns) for a subsample of TP calls with disagreeing predicted and published genotypes. Overall concordance was 75%. Most of the discordance was due to an underestimation of YRI frequency in the predicted genotypes (23 NC sites validated as 0.5/0 and 10 0.5/0 sites validated as 1/0).

Pred/SQM	0/0	0/0.5	0/1	0.5/0	1/0	0.5/0.5	0.5/1	1/0.5	1/1
NC	10	4	0	23	0	6	1	1	0
0/0.5	0	0	0	0	0	0	0	0	0
0/1	0	0	0	0	0	0	0	0	0
0.5/0	0	0	0	0	10	4	0	1	0
1/0	0	0	0	0	0	0	0	0	0
0.5/0.5	0	0	0	0	0	22	0	2	0
0.5/1	0	0	0	0	0	2	10	0	0
1/0.5	0	0	0	0	0	1	0	8	0
1/1	1	0	0	0	0	0	0	0	23

We also considered 137 TP coding, non-self-chained sites with conflicting predicted and published genotypes. Tables S4 and S5 cross-tabulate the Sequenom calls with the predicted and published genotypes, respectively. For the tracking sites, the overall concordance was 75% (63/84), with 8 failing primer design. For the non-tracking, 10/45 were validated as wildtype, suggesting that non-tracking TPs are usually real. The main source of error was that our predicted genotypes underestimated the alt frequency in the YRI (23 NC validated as 0.5/0 and 10 0.5/0 validated 1/0). We suspect that this was due to bias against the YRI in the actual mixture proportions.

References

[DePristo et al., 2011] DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M., Hanna, M., McKenna, A., Fennell, T., Kernytsky, A., Sivachenko, A., Cibulskis, K., Gabriel, S., Altshuler, D., and Daly, M. (2011). A framework for

Table S5: Confusion matrix of the published genotypes (rows) vs. Sequenom-validated genotypes (columns) for a subsample of TP calls with disagreeing predicted and published genotypes. Overall concordance was 30%. Most of the discordance was due to an underestimation of YRI frequency in the published genotypes.

Pub/SQM	0/0	0/0.5	0/1	0.5/0	1/0	0.5/0.5	0.5/1	1/0.5	1/1
0/0	0	0	0	0	0	0	0	0	0
0/0.5	**9**	4	0	0	0	**30**	0	**10**	0
0/1	1	0	0	0	0	2	**10**	1	**23**
0.5/0	1	0	0	23	0	1	1	0	0
1/0	0	0	0	0	10	0	0	1	0
0.5/0.5	0	0	0	0	0	2	0	0	0
0.5/1	0	0	0	0	0	0	0	0	0
1/0.5	0	0	0	0	0	0	0	0	0
1/1	0	0	0	0	0	0	0	0	0

variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43:491–498.

[Gerstung et al., 2012] Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature communications*, 3:811.

[Lawrence et al., 2013] Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., and Carey, V. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9.

[Li, 2011] Li, H. (2011). A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25:1754–1760.

[Mills et al., 2011] Mills, R., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel,

M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stutz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A., and Korbel, J. O. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332).

[Picard, 2014] Picard (2014). Picard. `http://picard.sourceforge.net/`.

[Plagnol et al., 2012] Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., et al. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747–2754.

[Sherry et al., 2001] Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

11 Demonstration of diagnostic software toolset

Our toolset facilitates analysis of variant calls in the context of genomic annotations. Much of the functionality is borrowed from Bioconductor infrastructure through the use of common abstract data types. We formally represent a set of variant calls as a *VRanges* object, which extends *GRanges*, a general R container for data on genomic features [Lawrence et al., 2013]. Overlap-based join operations enable us to query for associations between a set of variants and some other set of features stored as a *GRanges* or a compatible data structure.

The following code implements the homology analysis. Assume we have a set of variant calls in a *VRanges* object named `variants`. We begin by annotating each variant by the gene ID of its enclosing coding sequence, if any. We retrieve a *GRanges* of the CDS regions from a Bioconductor transcript database (`txdb`) and rely on overlap detection implemented by the GenomicRanges package.

```
cds <- cds(txdb, columns="gene_id")
variants$gene_id <- cds$gene_id[overlapsAny(variants, cds, ignore.strand=TRUE)]
```

Then, assuming we have a /GRaZZn of genes with paralogs named `paralogs`, we can determine which variants are in a paralog:

```
variants$has_paralog <- variants$gene_id %in% paralogs
```

And compute the FDR by paralog status:

```
fdr <- tapply(variants$fp, variants$has_paralog, mean)
```

The primitives demonstrated here are sufficiently general to support arbitrary annotation types, including homopolymers, repeats, and self-chain regions.

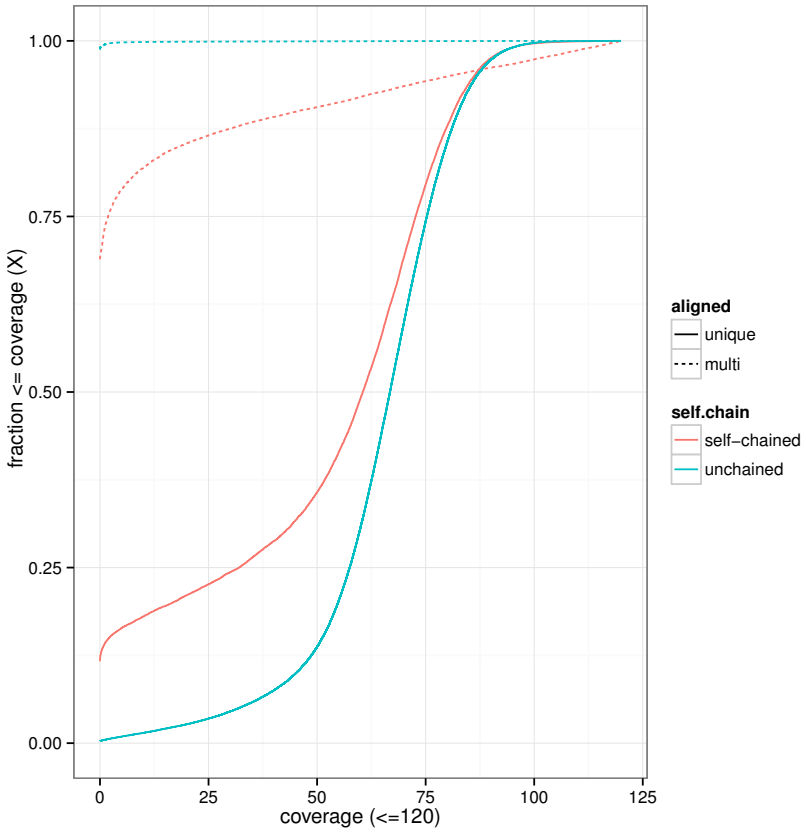

Figure S4: Comparison of the coverage distributions from the 50Y/50C sample, conditioned on self-chained status and whether the reads were multi-mapping or uniquely mapped by GSNAP. We find that the multi-mapping coverage (dashed lines) is generally much lower than expected, especially in unchained regions. Self-chained regions tend to be more prone to multi-mapping. The trend is the opposite for the uniquely mapped coverage: the average coverage matches expectations, and the self-chained regions have lower coverage. It appears that GSNAP has more difficulty uniquely aligning to self-chained regions.

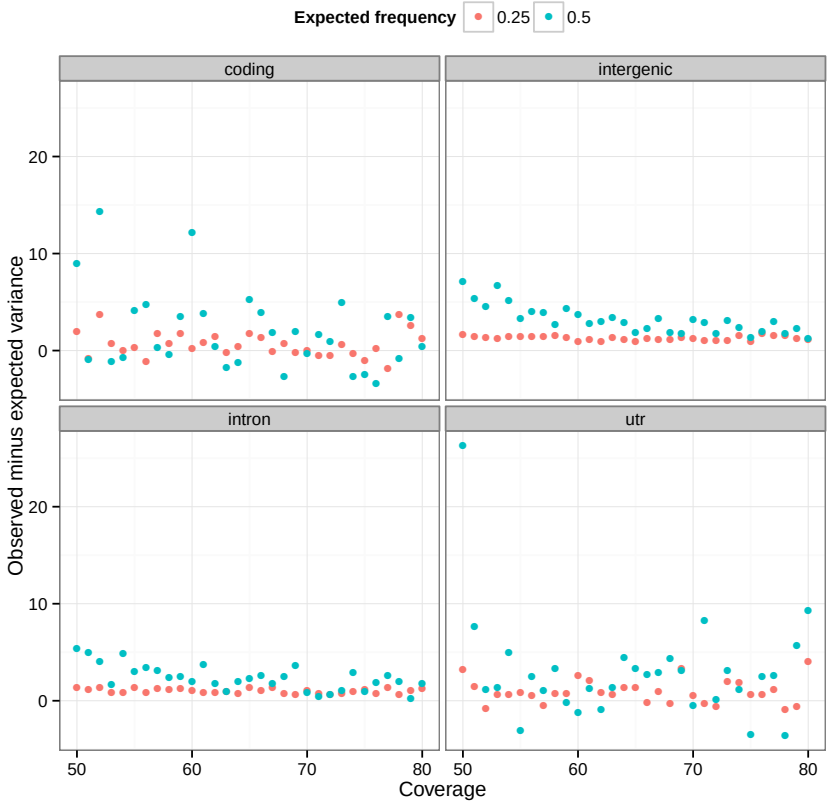

Figure S5: The difference between the observed and expected variance for each coverage value, from 50 to 80, conditional on genomic context and restricted to the YRI-specific variants in 50Y50C. There is some overdispersion at low coverage, with the strongest evidence in the intergenic homozygote variants. Overall the effect is minor.

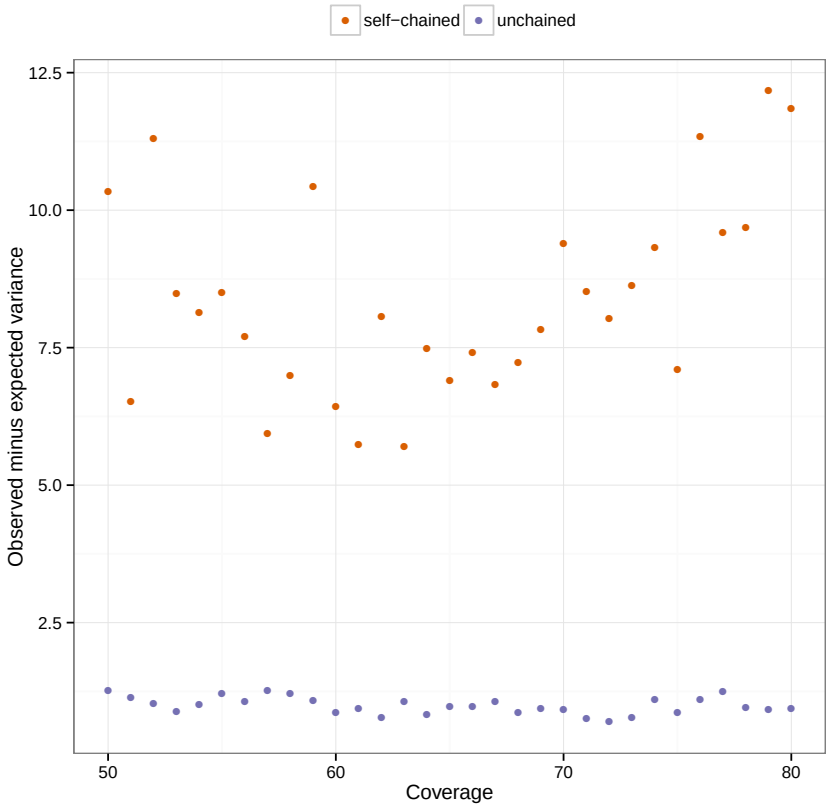

Figure S6: The difference between the observed and expected variance for each coverage value, from 50 to 80, conditional on self-chained status and restricted to the YRI-specific hets in 50Y50C. There is strong overdispersion in the self-chained regions, suggesting it is likely to be associated with mapping issues.

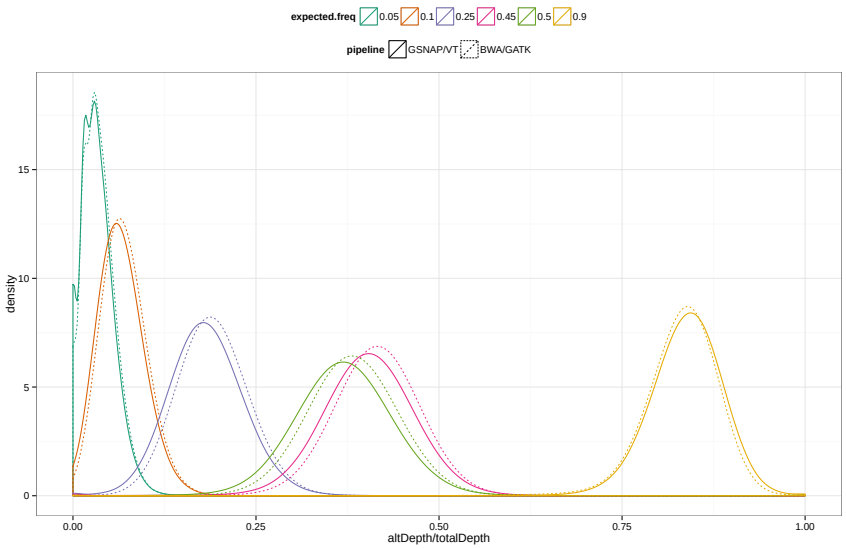

Figure S7: Distribution of observed variant frequencies for alt positions unique to the YRI, conditioned on the expected frequency and the aligner (GSNAP vs BWA).

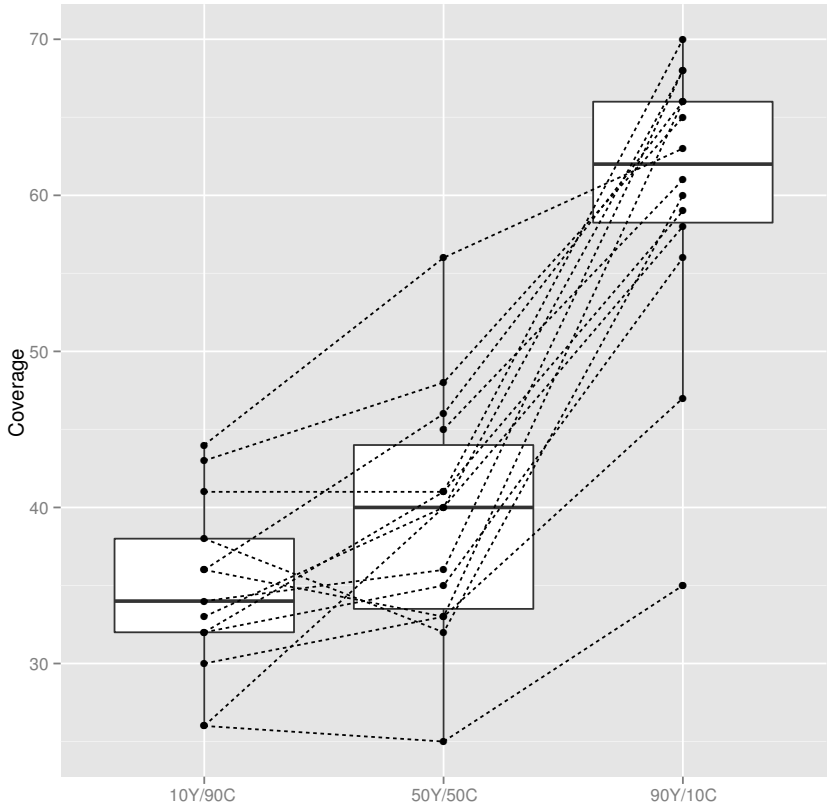

Figure S8: Coverage for 14 YRI specific variants in a region of a CEU deletion at chr14:80106289-80115049 (8.8kb). Coverage increases with increasing YRI proportion.

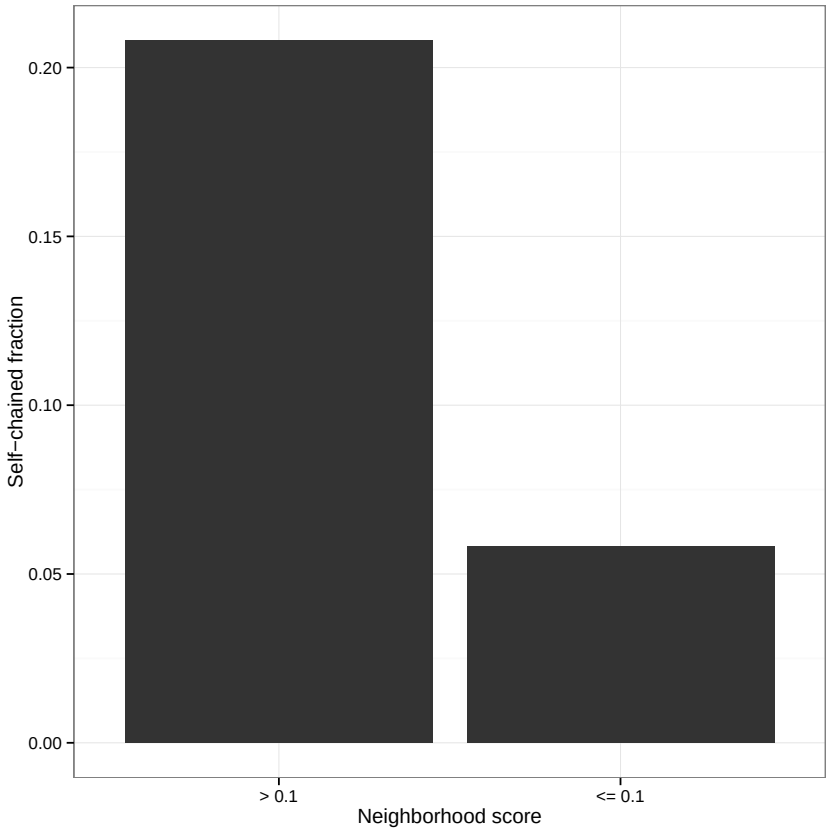

Figure S9: Self-chained fraction, separately for variants in dense neighborhoods (score > 0.1) and the rest.

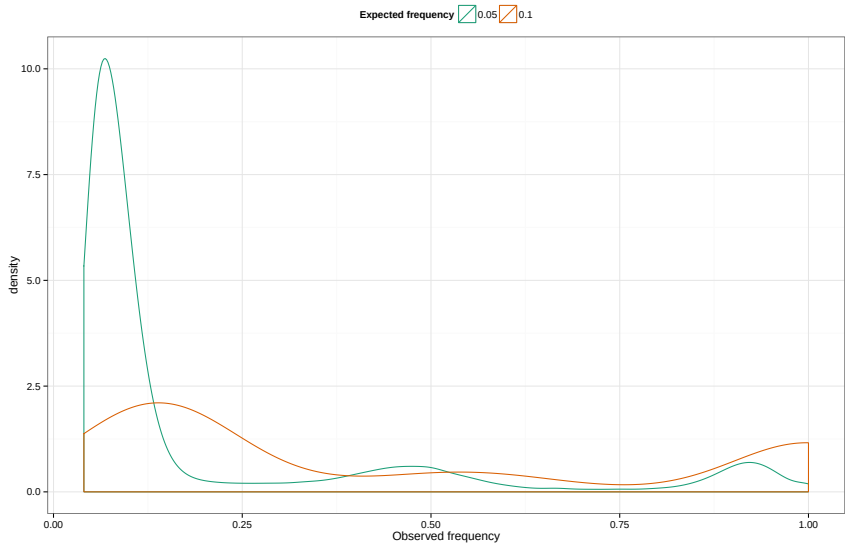

Figure S10: Distribution of alt frequencies for the variants expected at 5% and 10% in the 90/10C mixture. These variants should be CEU het or hom only (0/0.5 or 0/1), but the densities have peaks at frequencines that are consistent with their being YRI variants, 0.5/0.5 and 0.5/1, 1/0.5 and 1/1. This is consistent with substantial undercalling of the YRI genome in the published genotypes.

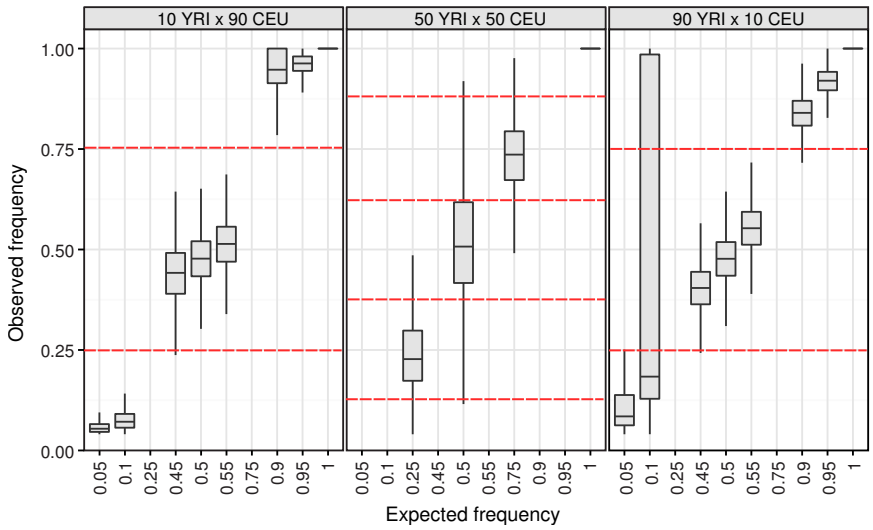

Figure S11: The data as in Figure 1B, except with red dashed lines indicating the boundaries of the frequency bins used to classify variants into predicted genotypes.

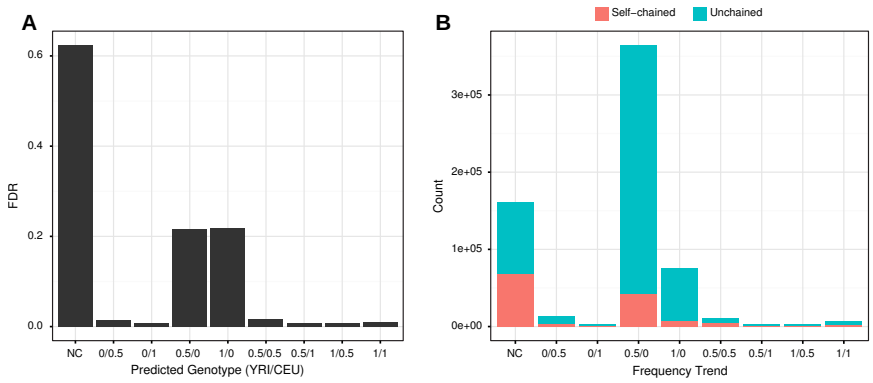

Figure S12: (A) The FDR conditioned on the genotypes (YRI/CEU) derived from the observed frequencies (see Methods). (B) Tabulation of the FP calls by predicted genotype, conditional on the self-chain status (see Methods).

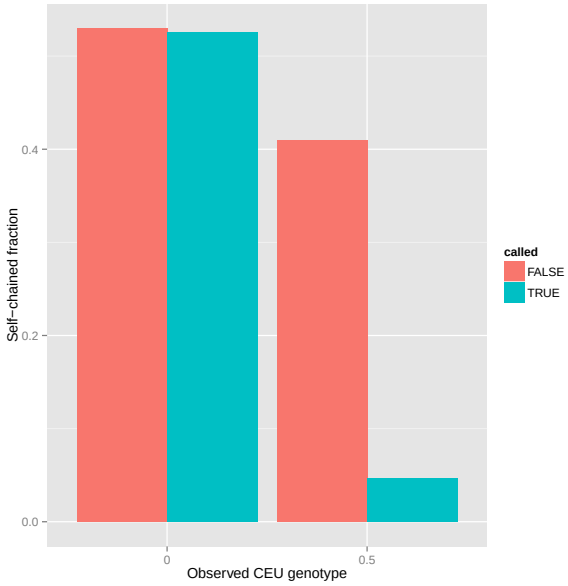

Figure S13: The self-chain fraction for positions called het in the published CEU genotypes, conditioned on whether the frequency trend indicated WT or het and whether VariantTools called the variant.

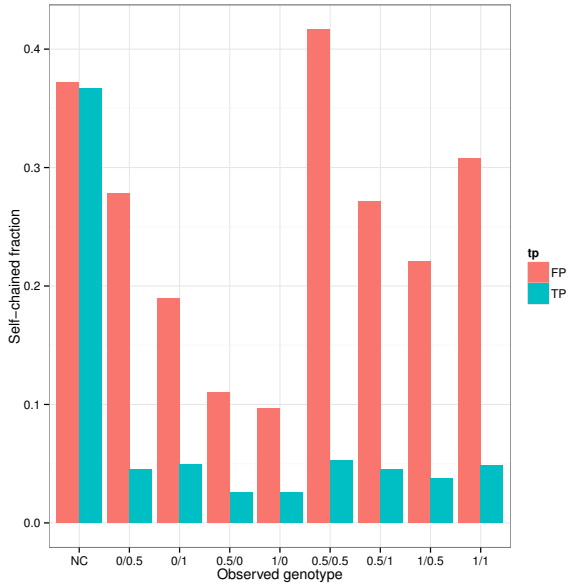

Figure S14: The self-chain fraction by predicted genotype for VariantTools calls from the 50Y/50C sample, restricted to coding regions. The color indicates whether the variant was present in the published genotypes.