Whole Exome Association of Rare Deletions in Multiplex Oral Cleft Families

Jack Fu[1] , Terri H. Beaty[2] , Alan F. Scott[3] , Jacqueline Hetmanski[2] , Margaret M. Parker[4] , Joan E. Bailey Wilson[5] , Mary L. Marazita[6] , Elisabeth Mangold[7] , Hasan Albacha-Hejazi[8] , Jeffrey C. Murray[9] , Alexandre Bureau[10] , Jacob Carey[2] , Stephen Cristiano[1] , Ingo Ruczinski[1] , and Robert B. Scharpf[111]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore MD, USA, [2]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore MD, USA, [3]Center for Inherited Disease Research and Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore MD, USA, [4] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston MA, USA, [5]Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore MD, USA, [6]Department of Oral Biology, Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, PA, USA, [7]Institute of Human Genetics, University of Bonn, Bonn, Germany, [8]Dr. Hejazi Clinic, Damascus, Syrian Arab Republic, [9]Department of Pediatrics, School of Medicine, University of Iowa, IA, USA, [10] Centre de Recherche de l'Institut Universitaire en Santé Mentale de Québec and Département de Médecine Sociale et Préventive, Université Laval, Québec, Canada, [11]Department of Oncology, Johns Hopkins School of Medicine, Baltimore MD, USA.

[1]To whom correspondence should be addressed
email: rscharpf@jhu.edu
550 N. Broadway, Suite 1101
Baltimore, MD 21205
phone: (443) 287-9408

Whole Exome Association of Rare Deletions in Multiplex Oral Cleft Families

Contents

Binary tree for global p-value

In a binary tree representation of our test described earlier, each level of the tree corresponds to a specific (deletion, family) pair. Going left at a node represents a deletion event shared by that family and the edge carries the sharing probabilities for that family; going right represents a deletion event not shared and that edge carries the probability of 1 minus the sharing probability. This representation is possible under the assumption of indepedence in sharing between deletion-family pairs. In our dataset, one could do a full expansion of the 86 level tree, accumulating the edge probabilities after each expansion to the bottom leaves. For an exact p-value, we sum the probabilities in the final leaves that were less than or equal to the probability of the observed deletion event pattern.

To expedite this process, recognize that the full expansion is not needed. Let us first define

p_{obs} as the probability of the observed deletion event pattern. Everytime we expand a node on the tree, if the cumulative probability of the expansion is less than or equal to p_{obs}, any further expansion on that branch is unecessary. This is because (a) the leaves that result from further expansion of that branch will be less than p_{obs} and will factor into the summation for the p-value and (b) the sum of those leaves will have the same probability of the cumulative probability up to that point of the parent node.

In the sample code below, `p_current` denotes the cumulative probabilities along the edges and is a placekeeper for when the tree reaches full expansion; `level` is the current level of the tree; `direction` denotes whether expansion is to the left (`direction=1`) or to the right (`direction=2`); `max_del` is the maximum level of the tree (86 in our application); `p_list` is a `max_del` $\times$ 2 matrix, where element [i, 1] is the sharing probability and element [i, 2] is one minus the sharing probability for the i^{th} deletion-family pair.

```
binaryTree <- function(p_current, level, direction, max_del, p_list){
      # Base case
      if(level==max_del){
            if(p_current*p_list[max_del, direction] <= p_obs){
                  return(p_current*p_list[max_del, direction])
            }
            else{return(0)}
      }
      p_current = p_current*p_list[level, direction]
      if(p_current<=p_obs){
            return(p_current)
      }
      else{
            return(expandHelper(p_current, level+1, 1, max_del) +
                  expandHelper(p_current, level+1, 2, max_del))
      }
}
```

Comparison to alternative methodologies for whole exome copy number analysis

To investigate concordance with alternative methodologies for whole exome analysis of copy number, we evaluated XHMM, CoNIFER, and CLAMMS (Fromer et al., 2012; Krumm et al.,

2012; Packer et al., 2016). XHMM and CoNIFER use principal components analysis and singular value decomposition, respectively, to normalize coverage. CLAMMS is most similar to the approach implemented here for preprocessing, differing mainly in scale. The filters used to identify rare deletions can not be straightforwardly adapted to XHMM and CoNIFER as these pipelines do not distinguish between homozygous and hemizygous deletions. The identification of homozygous deletions is a critical aspect of the pipeline proposed here as we assume there is only one deletion allele shared IBD between offspring within a pedigree. Consequently, we exclude regions where any homozygous deletion is detected in any oral cleft subject. As an alternative to directly comparing rare deletions identified by the different methodologies, we evaluated (1) the fraction of rare deletions identified by our approach that are also identified by other methodologies and (2) the signal to noise ratio (SNR) for any deletion identified by our method and another method irrespective of rarity status.

Of the 88 rare hemizygous deletions identified in our manuscript, XHMM recovered 61 (69%), CoNIFER recovered 53 (60%), and CLAMMS recovered 32 (36%). None of the alternative methods identified the rare deletion shared by distantly related offspring on chromosome 6 that was subsequently validated by qPCR (boxed region, Supplementary Figure 9). The lower concordance between RV and CLAMMS reflects a fundamental difference in the two strategies for identifying deletions. CLAMMS uses a mixture model fit at each bin across samples to derive probabilistic estimates of the mixture component labels that are presumed to represent distinct copy number states. Cluster-based identification of copy number works best when deletions are common in the population. The HMM implemented in CLAMMS uses the emission probabilities from the mixture models to segment the exome and identify copy number variants. By contrast, our approach puts the bin-level estimates on the same $\log_2$-based scale so that segmentation can be applied directly to the normalized and $\log_2$-transformed coverage to identify deletions that are private to an individual sample. A constrained mixture model is applied following the segmentation to *exclude* common deletions.

To better understand the relative sensitivity for deletion detection, we estimated a SNR for each deletion identified by multiple methods. Specifically, we estimated the numerator of the SNR as the absolute difference between median normalized coverage within the deletion and the median normalized coverage across all autosomal bins. The denominator is given by the median absolute deviation (MAD) of the autosomal normalized coverage. We found that the SNR for RV (SNR_{RV}) is greater than the SNR for CoNIFER ($\text{SNR}_{\text{CoNIFER}}$) for 85% of the deletions called by both RV and CoNIFER. For homozygous deletions called by RV, SNR_{RV} is at least two-fold larger than $\text{SNR}_{\text{CoNIFER}}$. Compared to XHMM, 40% of the deletions called as hemizygous by RV have a larger SNR than XHMM and all but one homozygous deletion called by RV has an SNR two-fold greater than SNR_{XHMM} (Supplementary Figure 10). We cannot calculate the SNR for CLAMMS using the above approach as the substantial bin-to-bin heterogeneity in scale (accomodated by their mixture model) artificially inflates the noise estimate in the denominator. However, the CLAMMS normalized coverage is qualitatively similar to the RV normalized coverage following $\log_2$ transformation and recentering (Supplementary Figure 11). As discussed above, the discordance between RV and CLAMMS reflects, in part, diametrically opposed uses of mixture models rather than differences in preprocessing.

Version of R and supporting packages

- R version 3.3.0 (2016-05-03), `x86_64-apple-darwin15.5.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: Biobase 2.32.0, BiocGenerics 0.18.0, Biostrings 2.40.2, CleftExome 0.0.37, CNPBayes 1.2.2, devtools 1.12.0, foreach 1.4.3, GenomeInfoDb 1.8.3, GenomicRanges 1.24.2, ggplot2 2.1.0, IRanges 2.6.1, Rsamtools 1.24.0, RUnit 0.4.31, S4Vectors 0.10.2, SummarizedExperiment 1.2.3, XVector 0.12.0

- Loaded via a namespace (and not attached): acepack 1.3-3.3, AnnotationDbi 1.34.4, AnnotationHub 2.4.2, BiocInstaller 1.22.3, BiocParallel 1.6.2, biomaRt 2.28.0, biovizBase 1.20.0, bitops 1.0-6, BSgenome 1.40.1, chron 2.3-47, cluster 2.0.4, coda 0.18-1, codetools 0.2-14, colorspace 1.2-6, combinat 0.0-8, data.table 1.9.6, DBI 0.4-1, dichromat 2.0-0, digest 0.6.9, DNAcopy 1.46.0, ensembldb 1.4.7, foreign 0.8-66, Formula 1.2-1, GenomicAlignments 1.8.4, GenomicFeatures 1.24.4, grid 3.3.0, gridExtra 2.2.1, gtable 0.2.0, gtools 3.5.0, Hmisc 3.17-4, htmltools 0.3.5, httpuv 1.3.3, httr 1.2.1, interactiveDisplayBase 1.10.3, iterators 1.0.8, lattice 0.20-33, latticeExtra 0.6-28, magrittr 1.5, Matrix 1.2-6, matrixStats 0.50.2, memoise 1.0.0, mime 0.5, munsell 0.4.3, nnet 7.3-12, plyr 1.8.4, R6 2.1.2, RColorBrewer 1.1-2, Rcpp 0.12.5, RCurl 1.95-4.8, reshape2 1.4.1, rpart 4.1-10, RSQLite 1.0.0, rtracklayer 1.32.1, scales 0.4.0, shiny 0.13.2, splines 3.3.0, stringi 1.1.1, stringr 1.0.0, survival 2.39-5, tools 3.3.0, VariantAnnotation 1.18.5, withr 1.0.2, XML 3.98-1.4, xtable 1.8-2, zlibbioc 1.18.0

References

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., ... Purcell, S. M. (2012, Oct). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet*, *91*(4), 597–607. Retrieved from http://dx.doi.org/10.1016/j.ajhg.2012.08.005

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., ... Eichler, E. E. (2012, Aug). Copy number variation detection and genotyping from exome sequence data. *Genome Res*, *22*(8), 1525–1532. Retrieved from http://dx.doi.org/10.1101/gr.138115.112

Packer, J. S., Maxwell, E. K., O'Dushlaine, C., Lopez, A. E., Dewey, F. E., Chernomorsky, R., ... Reid, J. G. (2016, Jan). Clamms: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics*, *32*(1), 133–135. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btv547

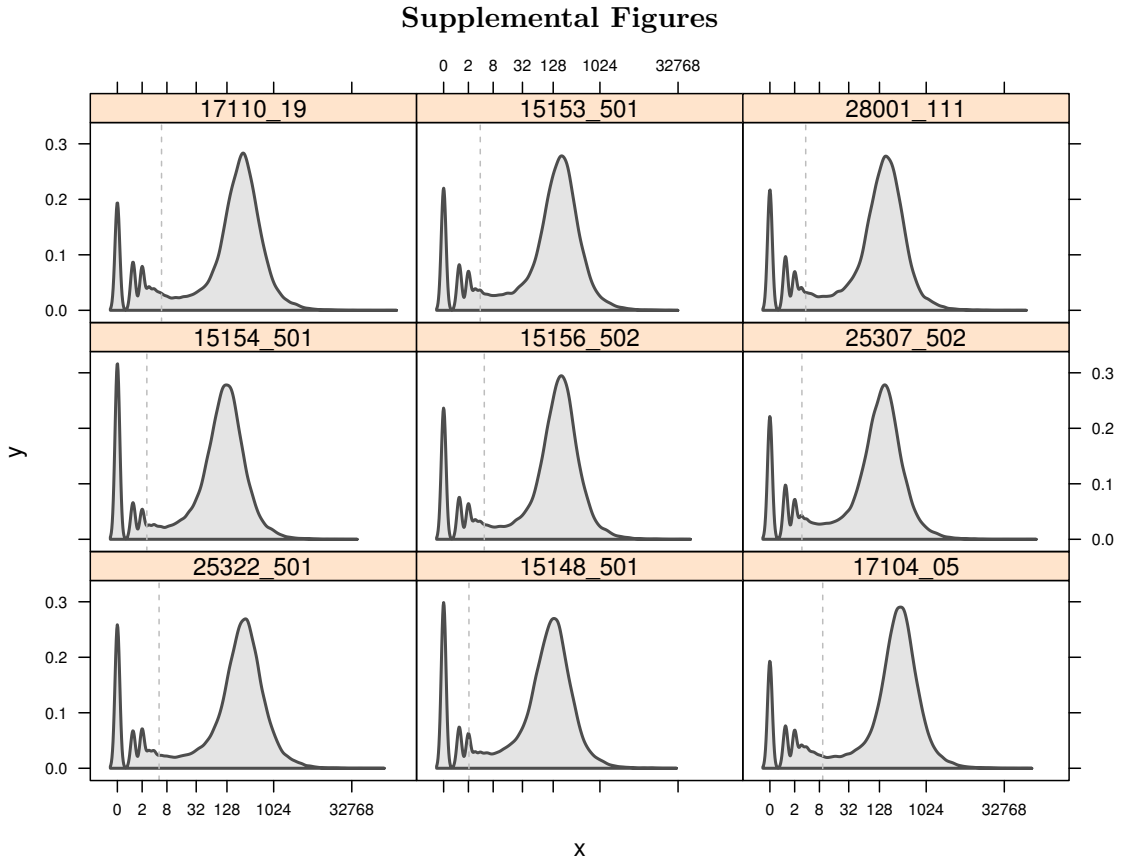

Figure 1. The density of $\log_2$ counts for nine randomly selected samples.

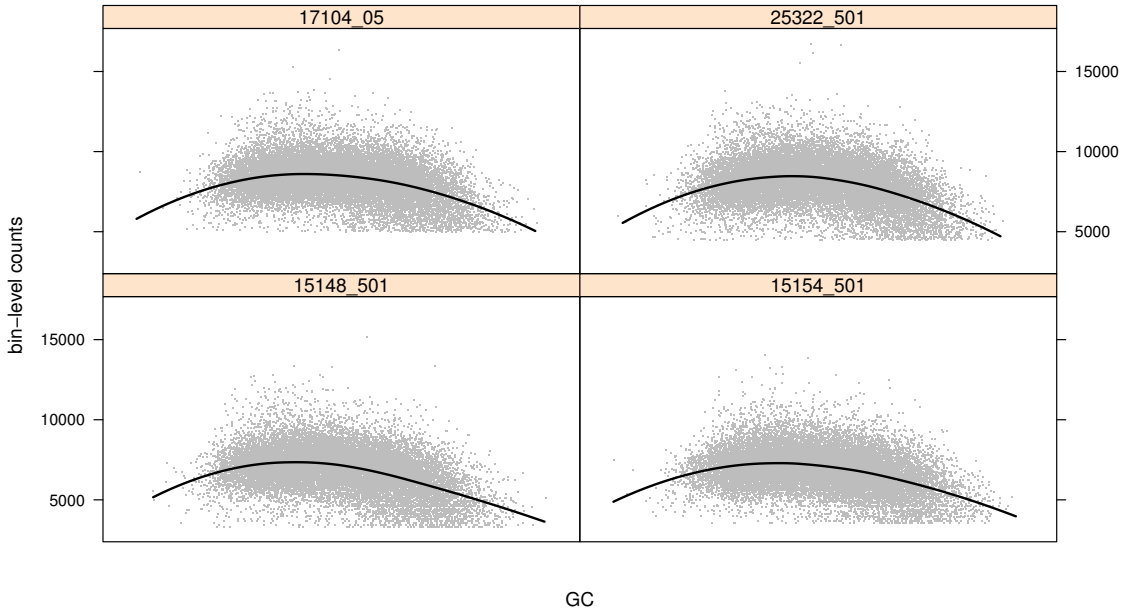

Figure 2. GC content versus bin-level counts. A loess scatterplot smoother with span 0.75 (black line) was used to model the non-linear relationship of GC and bin-counts.

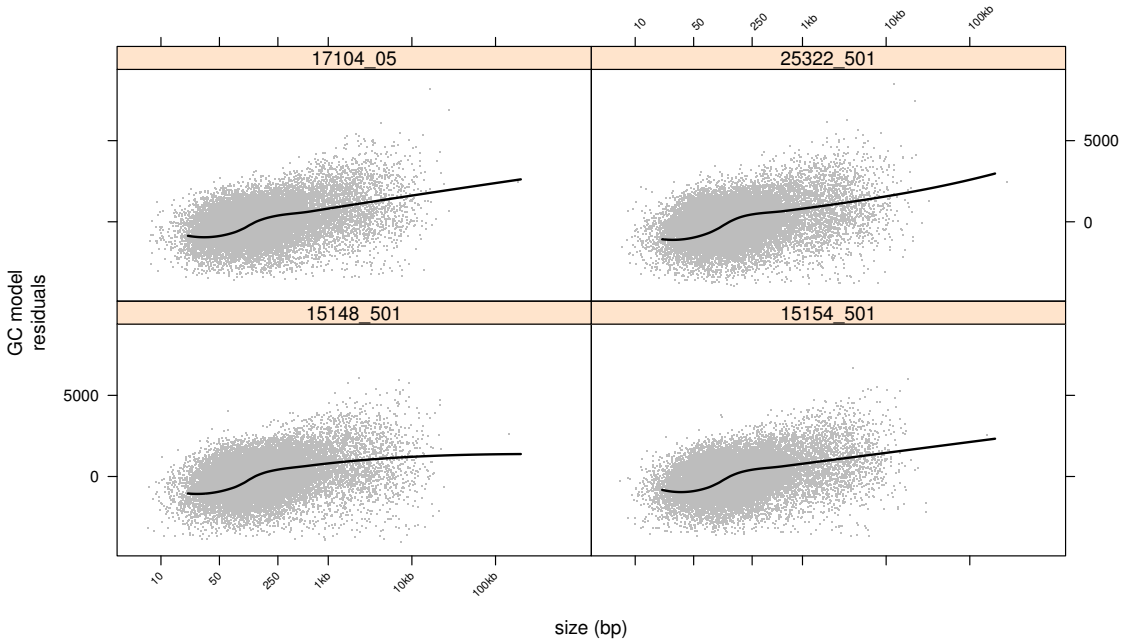

Figure 3. Bin size versus GC-adjusted counts. A loess scatterplot smoother with span 0.75 (black line) was used to model the non-linear relationship of the GC residuals and bin-counts.

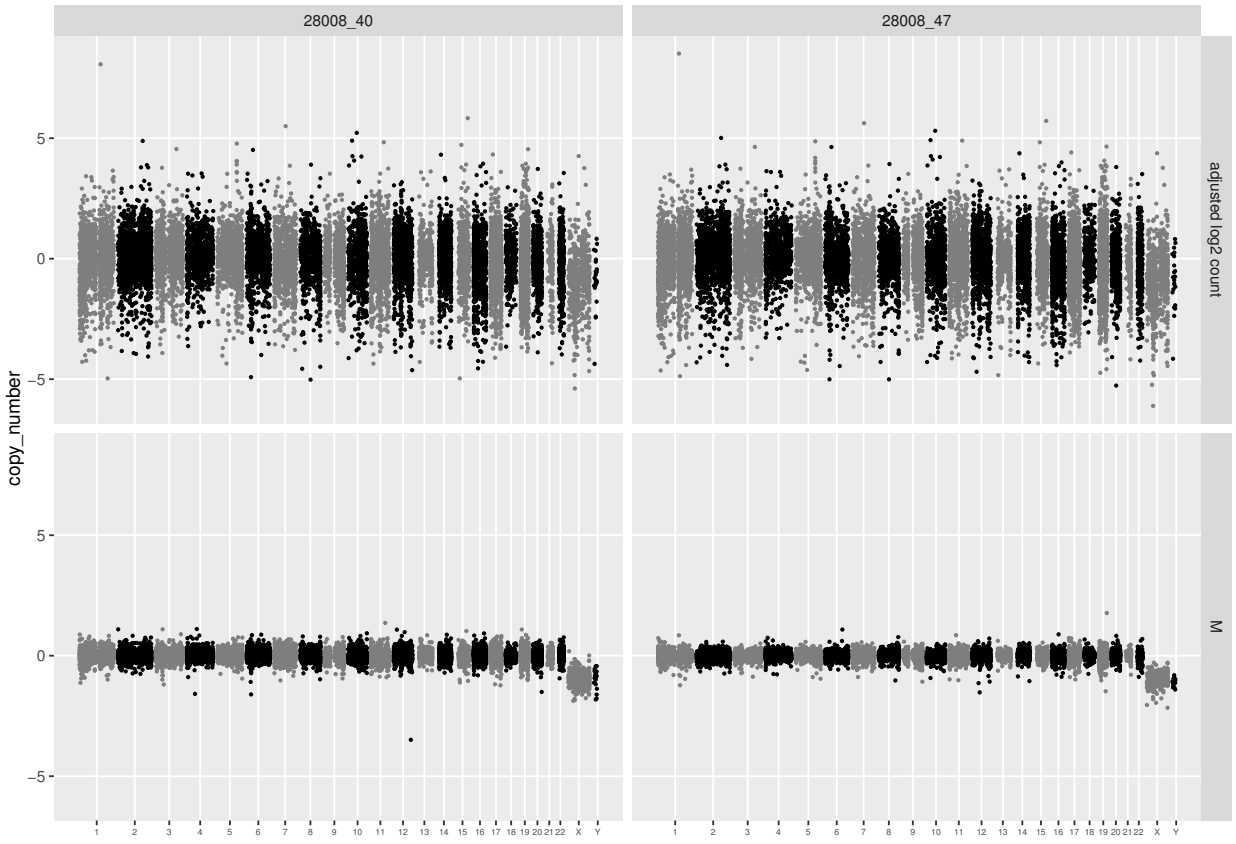

Figure 4. We preprocessed the number of single end tags aligned to 176,912 autosomal bins and 6,409 chromosome X and Y bins for samples '40' (column 1) and '47' (column 2) in family 28008. Here, we have plotted every tenth bin. Bin-level summaries of copy number calculated as $\log_2(\text{count} + 1)$ are adjusted for GC-content and bin size (top). As our interest is in rare deletions effecting a small fraction of all oral cleft patients, we centered each bin at its median across all of the oral cleft samples to remove common effects whether technical or biological in origin (bottom). The resulting bin-level estimates, referred to as M values, correspond to the log fold-change from the standard diploid genome which generally have low MAD and ACF_{10}.

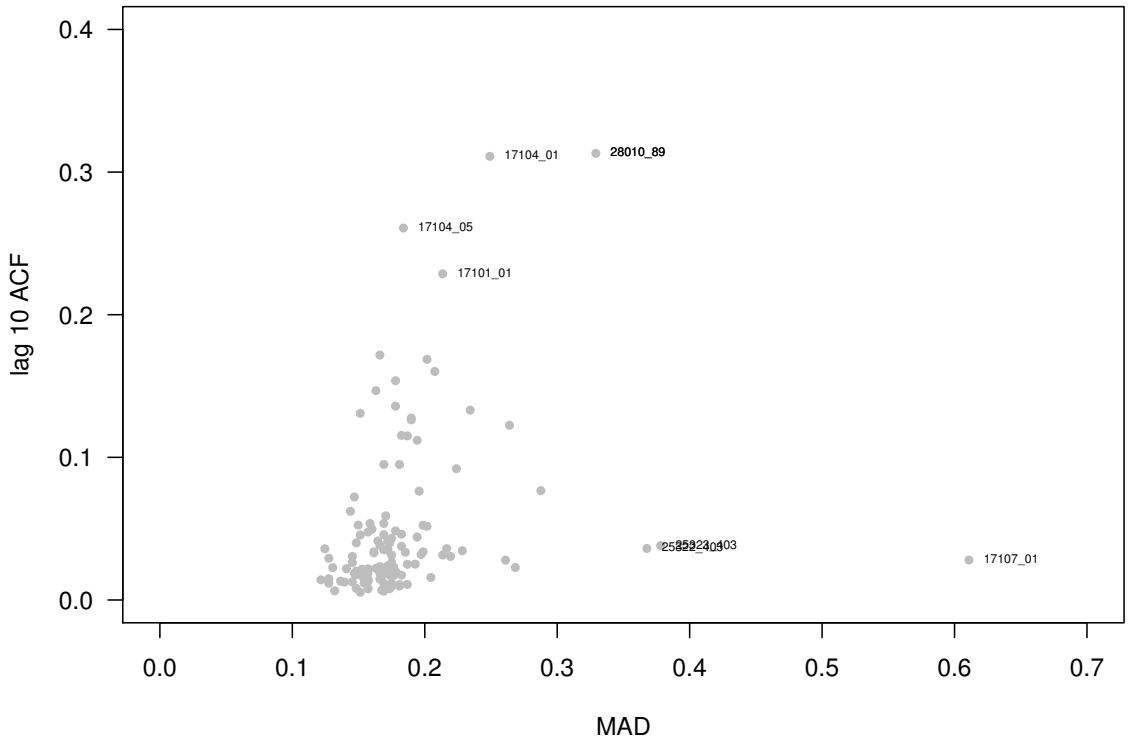

Figure 5. For each sample, we calculated the median absolute deviation (MAD) and the lag 10 autocorrelation of the autosomal M values.

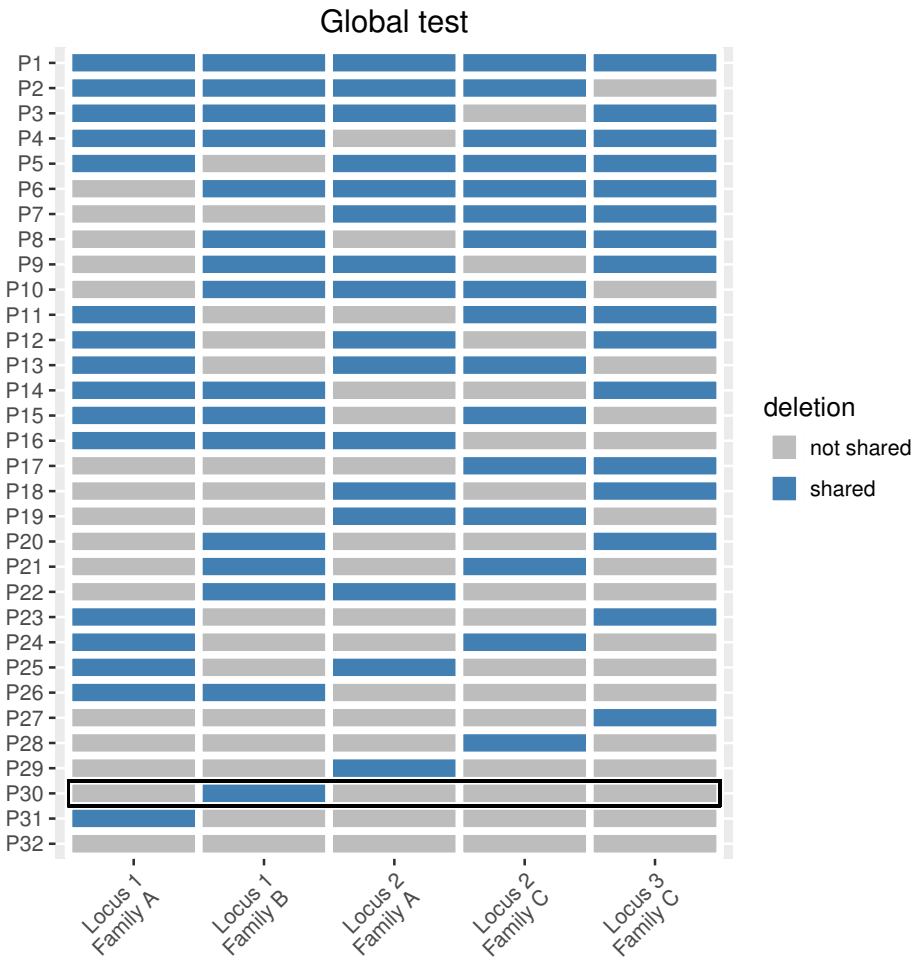

Figure 6. A toy example of the permutation scheme implemented to estimate the global sharing probability. Our simulated dataset is comprised of 3 families (A, B, and C) and three loci. The observed data is a single shared deletion in Family B (boxed by bold black rectangle) that is not shared in any of the other families. Note, locus 2 for Family B and locus 3 for families A and B are not included in the grid because none of these families had deletions at these loci. The rows of the table indicate all 32 theoretically possible observations for this toy dataset (including the observed data) and are ordered from top-to-bottom by the sharing probability ($P1 \leq P2 \leq \ldots \leq P32$) calculated as previously described. The p-value is simply the sum of the P's that are less than or equal to the observed sharing probablity, $P30$.

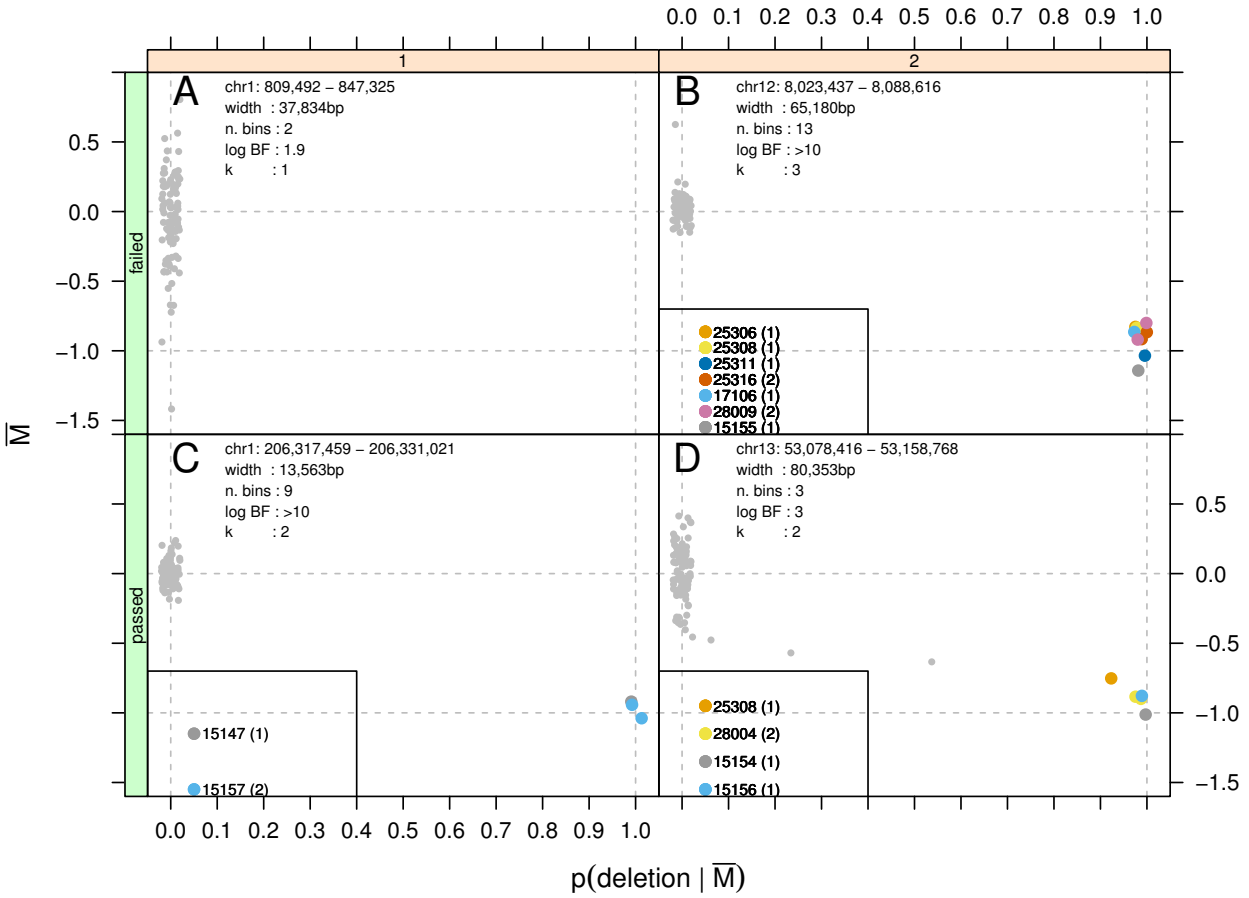

Figure 7. Regions with high variance are filtered by the mixture models (A), as well as regions in which additional hemizygous samples were identified implicating common rather than rare deletions (B). Regions identified as rare deletions by the mixture model had 5 or fewer families harboring a deletion allele and a log Bayes factor (log BF) comparing the hemizygous model to the normal model of at least 2 (C and D).

Figure 8. Each panel represents a rare deletion identified in the oral cleft study. At the top left (panel [1,1])is the rare deletion with the highest potential and at bottom right (panel [7, 5]) is the rare deletion with lowest potential. Populations represented in the 1000G study are ordered along the y-axis. Regions that have high CNV frequencies in the 1000G subpopulations tend to span less than 80% of the deletion (gray) identified in our oral cleft study (e.g., panel [3,4]). Such regions may be more prone to structural alterations, though these CNVs are at least 20% smaller than the deletions we identified. Regions such as panels [3, 7] and [3, 8] indicate the presence of several subpopulations with high overlap (black) and CNV frequencies near the 2 percent cutoff.

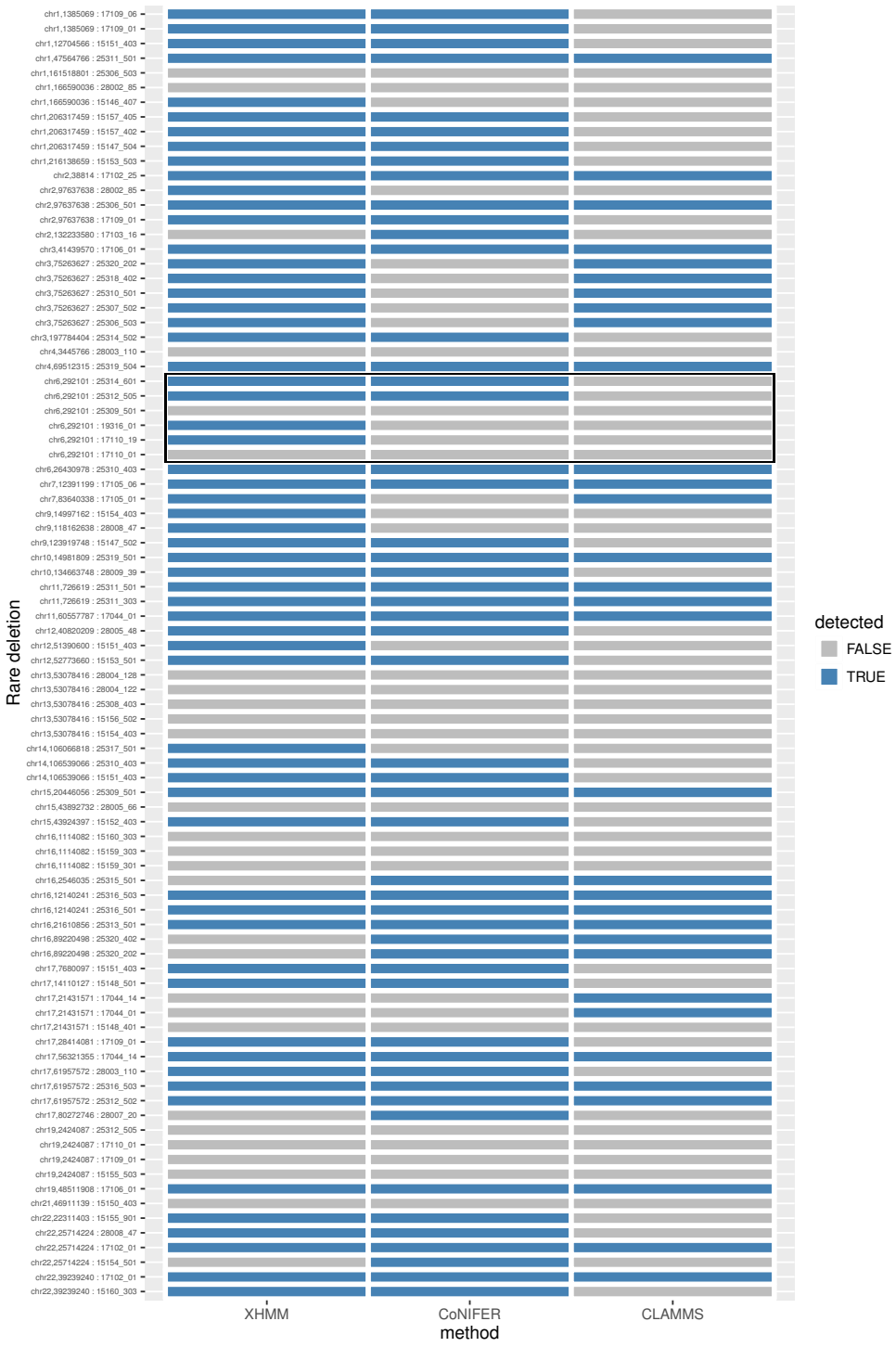

Figure 9. For each of the 88 rare variants identified, we assessed the fraction recovered by other whole exome methodologies for CNV detection. XHMM and CoNIFER were very similar to each other and recovered 69% and 60%, respectively, of the rare deletions. Only 36% of the rare deletions were recovered by CLAMMs. The boxed region highlights samples having a deletion on chr6 that were also evaluated by qPCR. Of the 6 samples identified as hemizygous by RV, 5 were validated by qPCR including the first cousins that share the rare deletion in family 17110. Of the whole exome analysis methods, only RV identified the shared deletion.

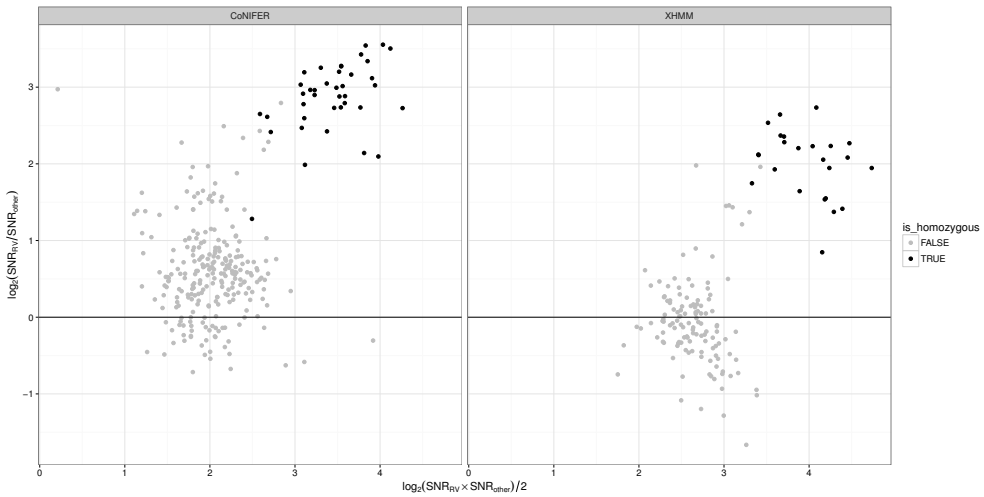

Figure 10. A comparison of the signal to noise ratio (SNR) of deletions identified in RV and CoNIFER (left) or RV and XHMM (right) irrespective of rarity status. Neither CoNIFER nor XHMM distinguish between hemizygous and homozygous deletions. Homozygous deletions called by RV (black) are more than 2-fold the SNR from XHMM or CoNIFER for all but one deletion.

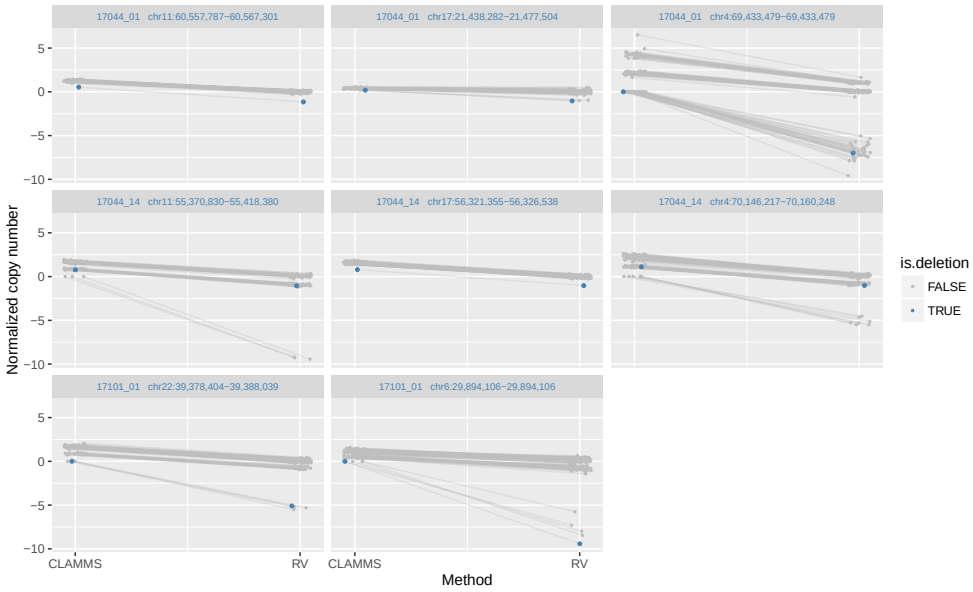

Figure 11. Normalized coverage is comparable for CLAMMS and RV, differing mainly in scale. Dashed lines correspond to theoretical copy numbers on the RV scale, for which RV is nearly unbiased. Panels 1 and 2 are hemizygous deletions private to sample 171044_01 (blue). Panel 3 is an obvious copy number polymorphism that is subsequently excluded in the RV pipeline.

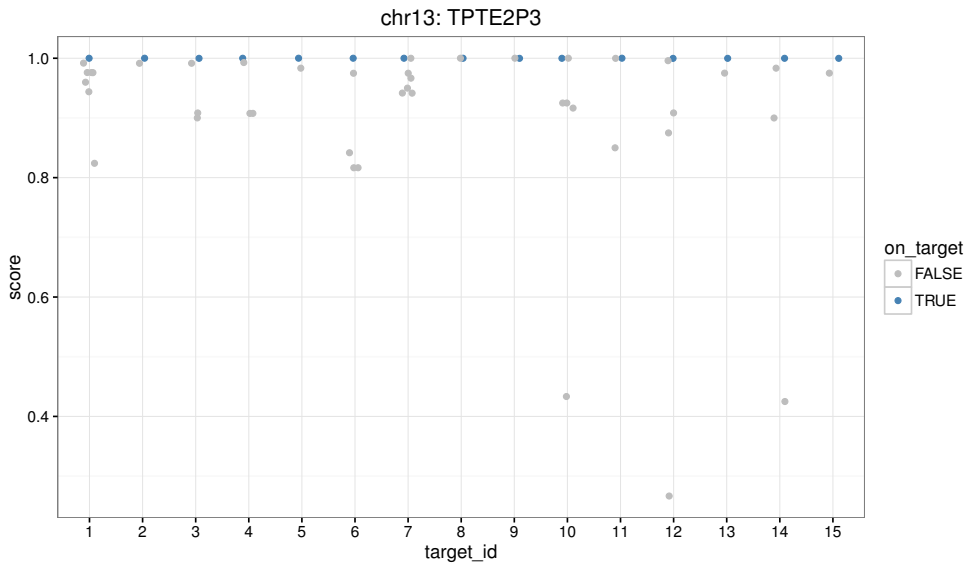

Figure 12. The BLAT score rescaled to 0-1 (1=perfect match) for 15 targets in the chr13 rare deletion (x-axis). All 15 targets have multiple alignments and nearly all targets have an off-target alignment (gray) as good as the on-target alignment (blue).