Author's response to reviews

Title: Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci

Authors:

Martin L Buchkovich Mr. (buchkovm@email.unc.edu)
Karl Eklund Mr. (keklund@email.unc.edu)
Qing Duan Ms. (qduan@email.unc.edu)
Yun Li Dr. (yunli@med.unc.edu)
Karen L Mohlke Dr. (karen_mohlke@med.unc.edu)
Terrence S Furey Dr. (tsfurey@email.unc.edu)

Version: 2 Date: 5 June 2015

Author's response to reviews: see over
Dear Editors:

Thank you for considering our manuscript entitled “Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci”. We are encouraged by the positive comments and agree with them that our work highlights many important aspects of allelic imbalance analysis. We value the feedback of the reviewers and have worked to revise our manuscript accordingly, which we believe is now much stronger as a result. Please find below, a detailed, point-by-point response to each of the suggested revisions.

Reviewer 1:

Major Compulsory Revisions

1a) The finding that poor SNP calls are associated with false positives for imbalance is not novel (e.g. Heap et al. 2010; León-Novelo et al. 2014), but the high false positive rate found in this study is significant because it emphasizes the particular importance of this issue with respect to making calls from this type of data in the absence of other genotype information. However, the methodology for calling heterozygous sites from the sequence data needs to be clarified. Are these calls included in the part of the pipeline indicated by the box labeled 1 (Create Custom Reference Genome-described briefly in lines 132-134)? Could poor quality heterozygous site calls be remedied by adjusting the approach in the pipeline or are they a property of the type of sequence data analyzed?

Heterozygous sites that were not known or present in the 1000 Genomes reference set of variants are only included during allelic imbalance detection and are not included during the initial alignment (Box 1). During a second alignment, heterozygous sites, predicted simply from the observation of each of two alleles in a certain number of sequence reads, are then used to update the custom reference genome (Box 1) and included during sequence alignment (Box 3). The high false positive rate in allelic imbalance predictions was still observed after this second round of alignment, and might indeed reflect a weakness of the method for identifying heterozygous SNPs. Increasing the required number of reads containing each allele to consider a site heterozygous to as many as 20 reads increases the accuracy of allelic imbalance detection at heterozygous sites predicted from the sequence data (Table 2. Minimum reads/allele section). This adjustment in our variant calling approach was simple but increased accuracy, though still not as accurate as at known SNPs. More sophisticated variant calling methods could increase accuracy further. To clarify these points, we have made the following changes in the manuscript:

- Modifications to the Results text describing AA-ALIGNER

“When testing for imbalances, AA-ALIGNER includes predicted heterozygous sites not included in the initial custom reference during sequence alignment. New heterozygous sites are predicted based on having a minimum number of reads containing each of two alleles. In addition, a minimum read threshold per allele can be applied to all heterozygous sites during imbalance detection to guard against incorrectly annotated heterozygous sites. While predicted heterozygous sites are not included in the initial reference genome customization (Figure 1, Box 1) or sequence alignment steps (Figure 1, Box 3), they can be added in a second round of reference customization and alignment if desired.”

- Modifications to the Discussion, including citing Heap et al. 2010 and León-Novelo et al. 2014:
“Predicting heterozygous sites in genome sequencing data is an active area of research, and many studies have demonstrated the difficulty of calling variants in sequencing data[25, 30, 45]. In addition to the GM12878 genotype annotation used in this study, other generally more conservative annotations exist. We found that most predicted imbalances were at common variants, and even when all common variants were included in alignments in the case of no genotypes, the true heterozygous variants and imbalances could be predicted well at these common variant sites. In contrast, the accuracy of imbalance detection at predicted heterozygous sites corresponding to rare variants is poor, even when these predicted heterozygous sites were included in a second alignment. Inaccurate imbalance detection can be caused by either i) incorrectly predicted heterozygous sites in the sequencing data (false positives) or ii) correctly predicted heterozygous sites in the sequencing data that were incorrectly annotated in the complete genotype (false negatives). Requiring more evidence to predict heterozygous sites increased the accuracy of imbalance detection, suggesting that false positives in heterozygous site predictions contributed to inaccurate imbalance detection. These incorrect predictions may be partly due to sequencing errors, but as some are still present at high minimum read thresholds, errors in sequence mapping likely contribute to false positives. The inclusion of incorrectly annotated heterozygous sites or absence of true heterozygous sites during sequence alignment can cause erroneous read mappings to highly similar genomic regions leading to incorrect heterozygous site identification.”

- **Figure 1 legend:**

  “Allelic imbalance is tested at heterozygous sites included in the customized reference genome and at predicted heterozygous sites, identified based on a minimum number of mapped reads containing each of two alleles. If desired, predicted heterozygous sites can be used to update the custom reference and be included in a second alignment repeating steps 3-6.”

**1b) It was also unclear if these are false positives in the new variant calls or false negatives in the reference sequence?**

Our results, using more stringent criteria for predicting heterozygous sites, suggest that false positives in predicted variant calls contribute to decreased accuracy. However, 61% of predicted heterozygous sites found to be imbalanced using the partial genotype alignment were not annotated as heterozygous in the complete genotype data, but based on our criteria would have been considered imbalanced in the complete genotype alignment if tested. Likewise, 41% of predicted heterozygous sites found to be imbalanced in the no genotype alignment were not annotated as heterozygous in the complete genotype data, but would have been considered imbalanced in the complete genotype alignment. These results suggest that the complete genotype annotation may be incorrect. This possibility is addressed in the following updated text.

- **Discussion:**

  “Interestingly, many imbalances at sites not annotated as heterozygous in the complete genotype would have been considered imbalanced in the complete genotype alignment using our criteria. This suggests that errors may exist in the complete genotype data leading to false negative imbalance predictions. Further study is needed, but these data suggest that both false positives and false negatives contribute to decreased detection accuracy at predicted variants. Thus, AA-ALIGNER outputs three sets of detected imbalance sites: i) a complete set of all imbalances identified; ii) imbalances at known or common heterozygous variants (higher confidence); and iii) imbalances at predicted rare variants (lower confidence).”
2) The validation study is appropriate for determining if there is differential binding at a particular site, but it is incomplete with respect to testing the analytical approach. Only sites with inferred AI were tested, thus the frequency with which differences in protein binding would be detected when no imbalance is inferred is unknown. To conclude that AA-aligner is detecting AI correctly, it would need to be shown that protein binding is concordant with detecting/not detecting AI.

We investigated this issue by performing an additional EMSAs for 5 sites that (i) contained a heterozygous variant in a detected CREB1 binding motif; (ii) had a substantial number of reads mapped containing each of the two alleles; but (iii) were not predicted as allelic imbalanced ($P>.3$) (Supplemental Figure 4). Three sites showed no allelic differences in protein binding, as predicted, while two sites (Supplemental Fig 4A and 4B) showed significant differences. The imperfect prediction of allelic differences may indicate that AA-ALIGNER is limited in its ability to detect all instances of allelic imbalance, but may also show that the in vitro EMSAs are limited in their ability to accurately detect in vivo imbalances. EMSAs analyze DNA binding in an environment lacking chromatin structure, and in the case of our experiments, other nuclear proteins. Yet, some allelic differences in protein binding may be dependent on that chromatin/protein context.

We examined other possible characteristics of the genomic loci that may contribute to prediction accuracy. We used MEME-ChIP [39] to define a CREB1 binding motif from the 10,000 strongest ChIP-seq peaks. We then calculated motif scores using FIMO considering each of the two alleles to predict whether allelic differences altered the strength of CREB1 binding, based on this motif score. Of the 5 sites examined by EMSA above for which no allelic difference was predicted, the two that showed differences in protein binding also had significant differences in motif scores (difference in FIMO score >5). Of the 9 original sites for which AA-ALIGNER predicted allelic imbalance in protein binding, 3 also have differences in motif scores, and all 3 showed protein binding differences in EMSAs consistent with the allelic imbalance and motif predictions. At one other site, the allele with stronger protein binding by EMSA and with the higher motif score was the allele predicted to have less binding by imbalance prediction. The CREB1 binding motif was only predicted at one of the 3 remaining sites with detected differences in protein binding by EMSA, and there were no significant differences in motif score by FIMO. Together, these experimental results and analyses of motifs led us to revise the description and discussion of the EMSA experiments.

- **Modifications to abstract:**
  
  “We used electrophoretic mobility shift assays (EMSAs) to experimentally test …”

- **Modifications to background:**

  “We experimentally detected differential protein binding at six of nine tested imbalance predictions …”

- **Revisions to results:**

  “Electrophoretic mobility shift assays (EMSAs) using purified CREB1, conducted in the absence of chromatin and other nuclear proteins, can experimentally test for differential binding of CREB1 to a specific DNA sequence. Multiple, independently performed EMSAs supported allelic differences in binding at rs2382818 (Figure 2D).”
“Using EMSA, we detected evidence of allelic differences in protein binding in the same direction as our predicted imbalance at 4 of the 7 sites (Figure 2D), for a total of 6 of 9 supported imbalances. Surprisingly, at rs1695359, we consistently detected increased protein binding for the allele predicted by our imbalance analysis to have decreased binding. Of the 6 EMSA-supported sites, only 3 were predicted to have allelic differences based on the FIMO-calculated motif score (difference>5). Of the 3 imbalance sites that were not supported by EMSA, only one (rs1695359) had a significant difference in motif binding score, and the allele with the stronger motif score demonstrated increased binding in the EMSA result, rather than the allele predicted to be enriched by imbalance detection. For comparison, we used EMSA to test 5 additional CREB1 binding locations with a heterozygous variant that fell within a CREB1 binding motif, but were not predicted as sites of allelic imbalance ($P$>.3). We found evidence of allelic differences in protein binding at two of these sites (Supplemental Figure 5). For these two sites, a CREB1 motif was only predicted when the allele with stronger protein binding was present.

These data provide strong supporting evidence of allelic differences in protein binding at 6 of the 9 predicted imbalanced sites and suggest that the sequence-specific binding preferences of CREB1 influence binding at these sites. It is unclear whether the remaining three sites not supported by EMSA indicate errors in AA-ALIGNER imbalance detection, or whether these show limitations of EMSA in detecting in vivo allelic differences in protein binding that are dependent on chromatin context or the presence of other nuclear proteins. Likewise, it is unclear whether AA-ALIGNER failed to detect allelic imbalance at two sites with allelic differences in protein binding based on EMSA, or whether chromatin and/or other proteins compensate for reduced sequence specificity in vivo resulting in similar binding regardless of allele present. Overall, these EMSA results provide evidence supporting allelic differences in protein binding at individual imbalance sites detected by AA-ALIGNER.”

- Revisions to discussion:

  “We experimentally tested for allelic differences in CREB1 binding using EMSA at nine sites with predicted allelic imbalance and five sites with no predicted imbalance. In general, EMSA results matched predicted differences in FIMO-calculated motif scores based on the presence of each of the two alleles, though we note that we were able to detect allelic imbalance and observe differential protein binding at three sites without predicted allelic differences in motif scores. EMSAs were performed in the absence of chromatin context and other nuclear proteins, and so are limited to detecting differences in the sequence binding specificity of a protein. Despite this limitation, we detected allelic differences in CREB1 binding at 6 of 9 predicted imbalanced sites providing strong supporting evidence of allelic differences in protein binding. Further testing is required to understand the cases when EMSA results do not support predicted allelic imbalances. For example, it is unknown whether any of the 3 sites not supported by EMSA were falsely detected as imbalanced by AA-ALIGNER, or whether they failed to validate because of the limitations inherent to EMSA. Likewise, further study is needed to determine whether the two sites that AA-ALIGNER did not predict as imbalanced but that EMSA showed allelic differences in protein binding are due to limitations in AA-ALIGNER or EMSA. These results highlight the need for better experimental assays to validate allelic imbalances, and underscore the difficulty of creating comprehensive catalogs of sites with experimental evidence of differences in protein binding.”

- Revisions to conclusions:
“With AA-ALIGNER, we were able to detect allelic imbalance in ChIP-seq data for a single transcription factor from a single cell line and provide supporting experimental evidence of differential protein binding at a small subset of imbalanced sites. These sites with experimental evidence included variants at two inflammatory bowel disease-associated loci.”

Minor Essential Revisions

3) Only mapping bias is substantially discussed in the background, other sources of bias are not. Other sources of bias could contribute to false positives, as potentially indicated by cases where AI is detected in the analysis, but no protein binding differences are detected. This should be discussed.

This is an important point. When building AA-ALIGNER, our main goal was to correct for reference mapping biases, but we also strived to correct of other sources of biases whenever possible. Some of these biases include allelic differences in sequence read mappability, incorrect heterozygous site identification, and PCR duplication. To emphasize this point, we have added the following text to the manuscript:

- **Background:**

  “AA-ALIGNER also attempts to correct other biases that can influence imbalance detection, such as incorrect heterozygous site annotations in reference genome sequences and incorrectly detected imbalances due to differences in mappability between reads containing each of the alleles or due to PCR duplications introduced during sequencing [25].”

- **Results, Overview of AA-ALIGNER:**

  “AA-ALIGNER is designed to correct for multiple sources of bias in the data whenever possible. Increasing the minimum read threshold required to test for an imbalance can guard against incorrect heterozygous site identification. Mappability biases, where reads containing one allele map uniquely while reads containing the other allele map to multiple locations and are filtered, may result in an artificial imbalance. AA-ALIGNER only considers reads that map uniquely to the same position in the genome regardless of the allele present. Post alignment filtering of duplicate reads corrects for biases that can arise from PCR duplication during library preparation.”

Discretionary Revisions and Comments

4) The bioinformatics overall are well described and executed, the statistical approach was not a focus of the paper and was very basic. While caveats are addressed in the discussion, I found the lack of consideration of statistical concerns distracting. It might be worthwhile to prominently point out that the software can be used with whatever statistical approach is most appropriate for the user.

To emphasize this point, we have added the following text to the manuscript:

- **Results-Overview of AA-ALIGNER:**

  “By default, significance of allelic imbalances is determined using a standard binomial test, but the AA-ALIGNER pipeline can be easily modified to incorporate alternative statistical methods of detecting imbalance.”
• **Discussion:**

“AA-ALIGNER was designed to be modular allowing for allowing for the incorporation of alternative methods for variant identification and tests for significance of imbalances.”

5) Are the input control sequence data available? Perhaps there is some reason not to, but it seems to me that with appropriate filtering detection of AI in these data would be an appropriate negative control for the overall analysis. This would be similar to the approach used for peak calls themselves.

We agree that control data can be useful in identifying sites of allelic imbalance likely caused by factors other than differential protein binding, such as uncorrected biases or copy number variation. Unfortunately, these data are not always available, and when available, overall signal intensity is often much lower than for ChIP-seq data due to greater signal dispersion and low sequencing depth. Input control data is available for GM12878 ChIP-seq and we investigated its use for imbalance detection by aligning reads using AA-ALIGNER and complete genotypes. In total, sufficient reads were present to test for allelic imbalance at 51% (2588/5046) of sites that were tested in the complete genotype CREB1 ChIP-seq alignment. We detected allelic imbalance at 17 of these 2588 sites (0.66%) in the input control data, and at 26 additional predicted heterozygous sites not tested for imbalance in the CREB1 data. The 17 sites, which were also predicted as imbalanced in the CREB1 data, are clustered within five genomic regions, suggesting that imbalance detection in these regions may not be accurate. Because the control data were only informative for a small portion of the imbalanced sites (17/200, 8.5%) in the CREB1 ChIP-seq analyses and because this data is not always available, in this study we were most interested in the performance of AA-ALIGNER independent of control data. For these reasons, we have excluded control data from our analyses even though we agree that it can provide additional insight into imbalance accuracy at a subset of sites. We have emphasized the utility of using control data by modifying the following text:

> Alternatively, sequence data from non-ChIP genomic input or other control experiments, when sequenced with sufficient read depth in the same sample, could be used to estimate an expected proportion of aligned reads per allele and to adjust for copy number variation within the binomial test. These control sequences could also correct for other biases that cause incorrect allelic imbalance detection in both the control and ChIP-seq data.

**Reviewer 2:**

**Major points:**

1a) I am not really clear on how the SPP identified “peaks” were used in the analyses. Is allelic imbalance test only carried out for reads that are assigned to “peaks”?

SPP peaks were included as a metric to show that reference mapping biases negatively impact multiple facets of ChIP-seq data analysis and not simply allelic imbalance detection. These peaks were not used in conjunction with allelic imbalance detection. We have modified the text to clarify use of peaks:
• **Results:**

“We examined, separately, the effect of biased alignments at heterozygous sites on peak and allelic imbalance detection”

• **Methods:**

“To investigate the effects of reference mapping biases on peak calling, peaks were called using SPP within an Irreproducible Discovery Rate (IDR) analysis [54] as outlined by the ENCODE Consortium [13, 55]. Overlaps were determined between the 10,000 peaks with the strongest signal and heterozygous sites identified by genomic sequencing (complete genotypes).”

1b) If not, it may be informative to only consider very stringently defined peaks when calling allelic imbalance.

Testing in only stringently defined peaks is an interesting suggestion, and we have investigated this further. While we did not directly use peaks to restrict imbalance detection, we did use a similar metric, minimum read threshold. For most analyses, we required at least 10 reads (5 reads/allele) to be mapped to each heterozygous site. At this minimum read threshold, we tested for imbalance in 1,284 of the 1,618 SPP peaks with heterozygous sites (79%). The highest minimum read threshold examined in the manuscript required at least 40 total reads mapped (20 reads/allele) for a site to be tested. Using this threshold, we tested imbalances at only 25% (401/1618) of SPP peaks with heterozygous sites. Peak scores, or signal strength, were higher for peaks tested for imbalance when using both the 5 reads/allele (Mann Whitney U P=7 x 10^{-39}) and 20 reads/allele thresholds (Mann Whitney U P=3 x 10^{-129}). Additionally, considering the 1,284 heterozygous sites tested for allelic imbalance that were also in SPP peaks, we found that the number of reads at the heterozygous site was highly correlated with the peak score (Pearson R^2=.7), as shown in the figure below. Considering these data together, we conclude that using a minimum read threshold to restrict imbalance detection is very similar to restricting imbalance detection to stringently defined peaks.
2) I appreciate the authors’ discussion on p-value thresholds, but still think that the authors should additionally report the adjusted p-values (given that correcting for multiple testing is standard practice in genome-wide studies).

We added the unadjusted and Benjamini-Hochberg and Bonferroni-adjusted p-values for each site tested for allelic imbalance in the alignments of CREB1 50-bp ChIP-seq reads using complete, partial, and no genotype data (Supplemental Tables 6-8). Additionally, in Supplemental Table 2, we now report the allelic imbalance statistics when using an unadjusted P-value threshold of $1.0\times10^{-5}$, which corresponds closely to the Bonferroni-corrected threshold for these alignments (alpha=0.05 and N~5000). We added text in the table for this threshold and for a threshold of $1.0\times10^{-6}$, which corresponds to a Bonferroni correction with alpha=0.005 and N~5000.

3) The finding summarized on lines 230-237 about incorrect prediction of heterozygous sites sounds very surprising to me. Given the strict threshold of 5 reads assigned to each allele, and that duplicate reads are removed, does this imply that 5 independent reads have erroneous sequence error at the same genomic location? Otherwise, how many independent reads support falsely predicted heterozygous sites?

This point is important to clarify. While it may be possible that sequencing errors in 5 independent reads are responsible for falsely predicted heterozygous sites, we think it is more likely that mapping errors caused by highly similar genomic regions and/or errors in genotyping or the reference genome are responsible. In particular, errors in the reference genome at sites where the sample genotype differs from the reference genome may cause reads to map incorrectly. In this case, reads do not need to contain sequencing errors to map incorrectly, although such errors compound the problem. Allowing an increased number of mismatches during sequence alignment can magnify the effect of similar regions on sequence alignment, as shown by decreased imbalance detection at predicted heterozygous sites with increased mismatch thresholds. Additionally, correcting for allelic imbalance at sites incorrectly identified as heterozygous can introduce errors into the reference sequence and cause reads to map incorrectly. To clarify the possible sources of these errors in heterozygous site prediction we have included the following text:

- **Discussion:**

  “These incorrect predictions may be partly due to sequencing errors, but as some are still present at high minimum read thresholds, errors in sequence mapping likely contribute to false positives. The inclusion of incorrectly annotated heterozygous sites or absence of true heterozygous sites during sequence alignment can cause erroneous read mappings to highly similar genomic regions leading to incorrect heterozygous site identification.”

4) Previous studies have shown that over dispersion in allele specific reads, which are not correctly modeled by a binomial model. Can the authors comment (or perform experiments) to show that over dispersion does not impact their results (or if so, to what extent)?

To address over dispersion, we have used a beta-binomial p-value at the same threshold to identify allelic imbalance in the alignments using complete, partial and no genotypes. We have compared the imbalances detected with this threshold and added these p-values to Supplemental Table 2. Correcting for over dispersion decreased the number of sites of allelic
imbalance detected, and reduced imbalance detection sensitivity and precision by varying amounts in the different alignments. When using a beta-binomial p-value, we still observed much higher sensitivity and precision for imputed and common variants than predicted and uncommon variants, as before. We have discussed these results with the following text:

- **Results:**
  “In addition to a binomial test, other statistical methods of detecting allelic imbalance have been used to measure the significance of allelic imbalance [21, 22, 25]. For example, a beta-binomial test is commonly used to correct for inaccurate imbalance detection caused by over dispersion of the data. Using a beta-binomial test (\(P<.01\)) for the 50bp pair CREB1 ChIP-seq data reduced the number of sites of allelic imbalance identified by 82-83% using complete, partial or no genotype information. Overall sensitivity and precision of imbalance detection using partial or no genotypes declined to \(~50\%\). Sensitivity and precision remained higher at imputed heterozygous sites (partial genotype alignment) and common variants (no genotype alignment) than predicted and uncommon variants as before. This reduction in the sensitivity and precision of imbalance detection is similar to the reduction seen when using a stricter binomial p-value threshold and is likely related to the increased p-values of false positive sites reported above.”

- **Discussion:**
  “Likewise, using beta-binomial p-values to correct for over dispersion and setting the same uncorrected p-value cut-off greatly reduced our power to detect allelic imbalance. Using the beta-binomial p-value, imbalance detection accuracy and precision remain significantly higher for imputed and common variants than for predicted rare variants.”

- **Methods-Identifying allelic imbalance:**
  “To calculate beta-binomial p-values, we first estimated parameter \(\alpha\) of the beta distribution using reference allele proportions across all sites. A Z-statistic for each tested site was calculated the following equation: \(Z = \frac{\hat{P} - 0.5}{\sqrt{\frac{1}{4N(2\alpha+1)}}}\), where \(\hat{P}\) is the proportion of reads containing the reference allele and \(N\) is the total number of reads at the site.

Minor points:

5) Table 2 labeling is confusing: under “complete” columns Np, and Nn are not technically computed using “complete” sequencing data. I would suggest relabeling/reorganizing this table for clarity (perhaps this can be broken up to two tables).

We apologize for the confusing terminology. The Np column under complete sequence data was meant to represent the number of imbalances identified using complete genotypes at sites identified as heterozygous by imputation. Likewise, Nh was meant to represent the number of imbalanced sites using complete genotypes that have a common allele frequency. We have now re-labeled Np to N\textsubscript{imp} (imputed) for partial genotypes and Nh to N\textsubscript{com} (common) when using no genotypes. Additionally we have clarified these labels in the table footnotes for all affected tables.

6) maybe informative to include a figure that graphically demonstrate the problem of allele-specific detection problem.
We have clarified our description of the problems related to reference mapping biases in allelic imbalance detection in the introduction and added Supplemental Figure 1 (see below). We have additionally cited another paper with a figure and accompanying text thoroughly describing other biases influence allelic imbalance detection.

- **Introduction:**

  “Sequence reads containing the allele not represented in the reference genome are penalized as an additional mismatch compared to reads containing the reference allele [15], and are less likely to map to the correct genomic location (Supplemental Figure 1). This can result in false detection of allelic imbalance favoring the reference allele, or failure to detect imbalance favoring the non-reference allele.”

- **Supplemental Figure 1 legend:**

  “Supplemental Figure 1 Reference mapping biases influence sequence alignment. When mapping reads to a reference genome containing a single allele at heterozygous sites (left), sequence reads containing the reference allele (C, blue) are more likely to map correctly than reads containing the non-reference allele (A, red), especially in the presence of sequencing errors (orange). Reads containing each allele are equally as likely to map correctly when mapping reads to sequence containing both alleles at heterozygous sites (right).”

- **New Supplemental Figure 1:**

  ![Supplemental Figure 1]

  We hope that you and the reviewers find these responses address all of these concerns and that you find the manuscript is acceptable for publication.

  On behalf of the authors,
Terrence Furey, PhD

Associate Professor
Departments of Genetics and Biology
Carolina Center for Genome Sciences

120 Mason Farm Road CB#7264
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7264
tsfurey@email.unc.edu
Phone: 919-966-7033
Fax: 919-843-4682