Investigating Copy Number Variation in Very-Early-Onset Inflammatory Bowel Disease using CNV Workshop

Edward Zhao

Aim

In this project, CNV Workshop is used to conduct a copy number variation analysis to identify genes and genetic variants associated with very-early-onset inflammatory bowel disease. Specifically, the mathematics and statistics underlying CNV Workshop are explored and methods to improve CNV Workshop are discussed.

Scientific Background

Inflammatory bowel disease (IBD) is a group of diseases characterized by inflammation of the gastrointestinal tract. The two main forms of IBD are Crohn's disease and ulcerative colitis, which are distinguished by the nature and location of the inflammation (Kelsen and Baldassano 2008). Ulcerative colitis is typically limited to the colon and rectum (which are parts of the large intestine) and involves shallow ulcers. In contrast, Crohn's disease can occur in any part of the gastrointestinal tract and is characterized by deeper but fewer ulcers. Both forms of IBD can lead to symptoms including abdominal pain, internal bleeding, and diarrhea (Podolsky 1991). The incidence of IBD is highest in young adults, though the disease can occur at any age (Kelsen and Baldassano 2008). In particular, a small number of IBD patients are diagnosed at or before five years of age and tend to present with more severe symptoms and disease progression. These patients are termed very-early-onset IBD (VEO-IBD) cases.

The exact causes underlying IBD are not well understood and likely to vary from patient to patient. A combination of environmental and genetic factors are believed to contribute to the etiology of IBD. For example, smoking is associated with increased risk for Crohn's disease and decreased risk for ulcerative colitis (Kelsen and Baldassano 2008). In addition, IBD is more common in developed countries (Benchimol, et al. 2011). This association may be driven by the changes in the diversity of gut microbiota in developed countries due to different lifestyles and diets. Recently, genome-wide association studies have also identified genes associated with increased risk for IBD. In particular, genes involved in immune and inflammatory pathways such as nucleotide binding oligomerization domain containing 2 (*NOD2*) and interleukin 10 (*IL10*) have been implicated IBD (Kelsen, et al. 2015).

Genetic factors play a larger role in VEO-IBD etiology compared to later-onset IBD. In addition, VEO-IBD patients have a higher probability of disease caused by a single gene, or monogenic defect (Uhlig, et al. 2014). However, large scale genome-wide association studies so far have focused on later-onset IBD given the rarity of VEO-IBD (Jostins, et al. 2012). Genome-wide association studies also lack the power to detect rare genetic variants associated with a disease. Recently, Kelsen et al. analyzed whole-exome sequencing data of VEO-IBD patients to identify novel variants in *CD19* and *MSH5*, which are genes in the common variable

immunodeficiency pathway (Kelsen, et al. 2015). Given the importance of genetics in VEO-IBD and the large number of VEO-IBD patients with monogenic defects, it is hypothesized that identification of rare genetic variants associated with VEO-IBD will improve our understanding of VEO-IBD pathoetiology. To this end, copy number variation analysis is used to identify rare copy number variants (CNVs) in a case series of VEO-IBD patients. Variations in the copy number of genes can directly affect gene expression and potentially lead to disease.

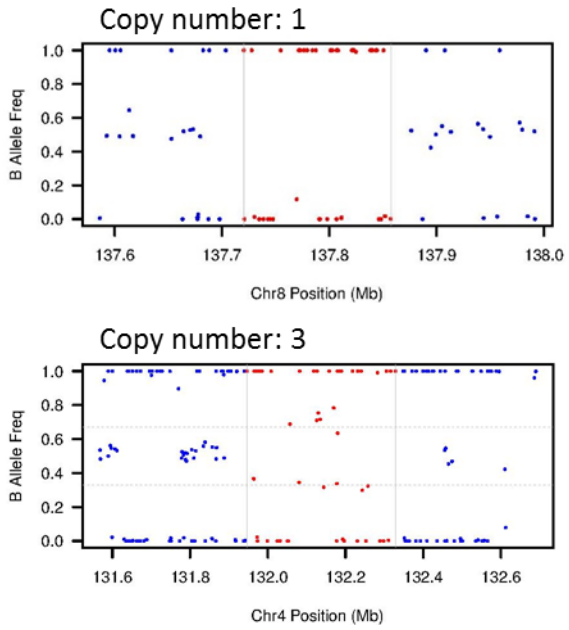
Methods

9 patients with onset of IBD at 5 years of age or younger were included in this analysis. All study subjects were patients at The Children's Hospital of Philadelphia. Diagnosis of IBD occurred through a combination of endoscopy, X-ray exam, laboratory tests, and physical examination. Patients varied in disease presentation and severity. DNA samples were obtained from all patients.

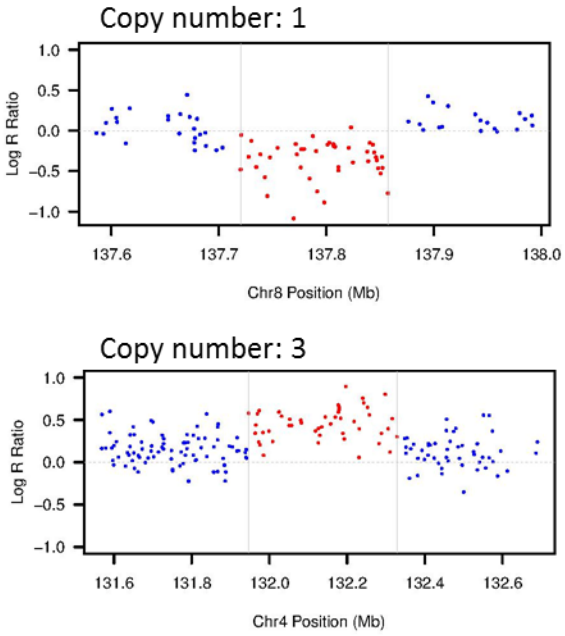
Whole-genome genotyping was accomplished through the Illumina Human610-Quad microarray kit (Illumina, Inc., San Diego, CA). This kit genotypes at 616,794 markers spread throughout the genome with a mean marker spacing of 5.51 kilobases. Each marker captures variation at a location in the genome, termed single nucleotide polymorphism (SNP). Since nearby locations in the genome are genetically linked, markers can be spaced to adequately capture the variation in the genome without the expense of whole-genome sequencing. The markers map to locations in the reference genome. The data for the samples were mapped to human genome version 19 (hg19).

SNP arrays function by having fluorescently labelled sample DNA fragments hybridize to oligonucleotide probes on the microarray. The SNPs targeted in Illumina arrays are bi-allelic, and an allele-specific oligonucleotide probe exists for each allele. The relative fluorescent intensities of the two sets of probes for each SNP can be used to infer allele frequencies (Lin, Naj and Wang 2013). The two alleles at each SNP are termed A and B, and patterns in B allele frequencies (BAF) are shown in Figure 1B. By plotting the A and B allele intensities, the BAF can be taken as the angle (θ) of the polar transformation ($\arctan \frac{Intensity_B}{Intensity_A}$). Loss of one copy results in a BAF of 0% or 100% at each marker (genotype A or B), meaning there is a loss of heterozygosity (genotype AB). A duplication appears with four levels of frequencies, corresponding to genotypes AAA, AAB, ABB, and BBB.

A.



B.



C.

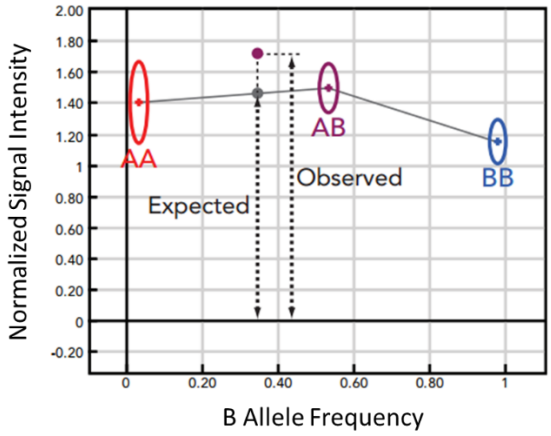


Figure 1. Patterns for B allele frequencies (A.) and log R ratios (B.) are shown. Red points indicate markers within a detected CNV. The horizontal axis shows the position along chromosome 8 in the top figures and along chromosome 4 in the bottom figures. Panel C illustrates how LRR values are generated (adapted from Illumina product manual). The purple point represents the observed values while the grey point represents the expected signal intensity at the BAF value given a normal copy number of 2.

The total fluorescent intensity for a SNP is related to the copy number of the DNA. For example, human somatic cells are diploid and would be expected to have a copy number of two for autosomal DNA. Duplications or deletions can increase or decrease the copy number respectively. Examination of the "log R ratio" (LRR), which is the base two log of the ratio between the observed fluorescent intensity and an expected intensity ($\log_2 \frac{R_{observed}}{R_{expected}}$), can help identify copy number variants (Wang, et al. 2007). Here, R is the radius of the polar transformed data (while BAF is the θ as mentioned earlier). The expected signal intensity is known (provided by Illumina) for different copy number 2 allele combinations (AA, AB, BB). Since expected signal intensity varies depending on the BAF (Figure 1C), the expected intensity for a given BAF can be determined by a linear interpolation (shown as grey lines) between the expected clusters (between AA and AB or between AB and BB). Patterns in the LRR seen in deletions and duplications are shown in Figure 1A. Lower values for LRR are seen in deletions and higher values for LRR are seen in duplications.

CNV Workshop is used to detect, or call, CNVs (Gai, et al. 2010). CNVs were called for each of the 9 samples separately. CNV Workshop is written in Perl though most of the analytical steps (namely segmentation and CNV calling) are done by calling R from Perl. The segmentation step is done using a circular binary segmentation algorithm (Olshen and Venkatraman 2004). The series of LRR values along a chromosome can be represented by $X_1, X_2, \ldots, X_n$ for n markers and can be plotted as done in Figure 1B. The distribution of X_i depends on the copy number. Therefore, the goal of segmentation is to detect change-points in the signal intensities, which correspond to the locations of changes in copy number. First, the data is smoothed to remove any extreme outliers that arise from technical measurement errors and replace them with an imputed value (closer to the median of the surrounding LRR values). Then, to find if there is a change-point along the chromosome, binary segmentation uses a likelihood ratio statistic to test the null hypothesis that there is no change-point against the alternative hypothesis that there is a single change-point at some location i. The likelihood Z_i of a change-point at a location i is given by:

$$Z_i = \left(\frac{1}{i} + \frac{1}{n-i}\right)^{-\frac{1}{2}} \left(\frac{S_i}{i} - \frac{S_n - S_i}{n-i}\right),$$

where partial sums of X_i are represented by $S_i = X_1 + \cdots + X_i, 1 \leq i < n$. Essentially, this compares the average intensity for markers 1 through i with the average intensity for markers $i+1$ to n. The maximum $|Z_i|$ is determined to be a change-point if it exceeds the upper αth quantile of the null distribution of likelihoods (testing at significance level α). This null distribution can be determined using a permutation approach where a large number of $Z^*_{\max}$ are simulated to provide a null distribution which the observed $Z_{\max}$ from the data can be compared against. The binary segmentation procedure is then repeated for each of the two segments separated by an identified change-point until no more change-points can be found. One issue with binary segmentation is the algorithm may fail to detect a small changed segment that lies within a large segment since it only looks for one change-point at a time. This is a common situation within

genomic data since copy number variations (deletions/duplications) tend to be relatively small compared to the size of genome (which is mostly copy number 2). To address this problem, a variation of binary segmentation called circular binary segmentation can identify either 1 or 2 change-points within a segment. By connecting the two ends of a segment, circular binary segmentation tests if the arc from $i+1$ to j is different from the complementary arc:

$$Z_{ij} = \left(\frac{1}{j-i} + \frac{1}{n-j+i}\right)^{-\frac{1}{2}} \left(\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i}\right)$$

The max likelihood is given by $\max_{1 \le i < j \le n} |Z_{ij}|$ which detects a single change at i if $j = n$ and two changes at i, j if $j < n$. After segmentation is completed, LRR and BAF statistics of the markers in the segment are used to determine the copy number of the segment (Gai, et al. 2010). Specifically, CNV determination is based on the mean LRR of markers in the segment and two standard deviation statistics for BAF and the percentage of markers with a BAF between 0.6 and 0.4. The two standard deviation statistics are shown below:

$$b2.sd = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (\min(B_i - 0, 1 - B_i, |B_i - 0.5|))^2}$$

$$b3.sd = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (\min(B_i - 0, 1 - B_i, |B_i - 0.33|, |B_i - 0.67|))^2},$$

where B_i is the BAF at marker i out of n in a segment. We see that b2.sd is the standard deviation from expected BAF values of copy number two (diploid): 0, 1, 0.5. Similarly, b3.sd is the standard deviation from expected BAF values of copy number three: 0, 0.33, 0.67, 1. Therefore, a segment that is duplicated would be expected to have $b2.sd > b3.sd$. Two distinguish a deletion from copy number two, we would expect the percentage of markers with a BAF between 0.6 and 0.4 to be low for a deletion segment. The exact criteria for CNV calling are shown in Table 2.

Table 2. CNV Determination criteria.

CNV	Mean LRR	BAF
Homozygous deletion	$Mean < -2$	—
Heterozygous deletion	$-2 < Mean < -0.3$	$Percent\ [0.4,0.6] \le 4\%$
Duplication	$Mean > 0.2$	$b2.sd > b3.sd$

To put the results from CNV Workshop into context, CNVs are also compared to previous results done using PennCNV software. Rather than using binary segmentation, PennCNV uses a hidden Markov model to determine CNVs. In addition to LRR and BAF values, PennCNV also needs a file containing the population frequency of B allele for each SNP and a hidden Markov model file as input. The population frequency of B allele refers to the average BAF estimated from a large reference population. This provides a reference to which each sample BAF can be compared and allows for greater accuracy in CNV calls. Then, a hidden Markov model determines the most likely hidden states that existed to generate the observed data. In this case, the hidden states of interest are the copy numbers, while the observed data consists of the LRR and BAF values. Given a sequence of LRR and BAF values along the chromosome, PennCNV uses the Viterbi algorithm to infer the most likely sequence of copy numbers along the chromosome that would generate the data (Wang, et al. 2007). The hidden Markov model input file specifies expected LRR values for each copy number and expected transition probabilities between different copy numbers along the chromosome. Both the population frequency of B allele and hidden Markov model file are provided in the PennCNV package.

CNV Workshop was also used to annotate CNVs by checking for genes that overlap with CNVs. The refGene database contains the locations of human genes used to annotate CNVs. Additionally, the Database of Genomic Variants (DGV) maintains a catalogue of structural variants found in peer-reviewed original research (MacDonald, et al. 2014). Each CNV detected by CNV Workshop is annotated with the number of overlaps it has with known structural variants found in DGV to provide an indicator of the commonness of each CNV in a general population. All summary statistics were computed using Microsoft Excel or R version 3.2.0 (R Core Team 2015).

Results

A total of 9 patients with VEO-IBD were included in the analysis. The summary statistics of CNV calls from CNV Workshop are shown in Table 3. On average, approximately 31 CNVs were detected per individual, with the number per individual ranging from 19 to 55. Overall, many more deletions than duplications were detected. In comparison, while PennCNV found slightly more CNVs per individual (around 40), there were only slightly more deletions than duplications. This suggests that CNV Workshop criteria for identifying duplications may be too exclusive (or PennCNV's criteria may be too inclusive). Each CNV found by CNV Workshop contained approximately 11 markers on average. CNV Workshop uses a minimum of two markers to detect a CNV, and the CNV with the largest number of markers contained over 100. Duplications were typically larger and contained more markers (34.6) than deletions (9.0).

Table 1. Summary statistics of CNV calls.

Summary Statistic	Total	Deletion	Duplication
Number of CNVs			
Total	281	256	25
Per sample	31.2	28.4	2.8
Average markers per CNV	11.1	9.0	34.6
Average length of CNV (kb)*	46.4	25.9	257.0
Size distribution (kb)*			
25th Percentile	0.5	0.4	49.6
Median	4.0	3.2	136.2
75th Percentile	29.2	20.0	313.7
X chromosome CNVs	14 (5.0%)	5 (2.0%)	9 (25%)
Overlapping with gene(s)			
Overlapping CNVs	126 (44.8%)	106 (41.4%)	20 (80.0%)
Non-overlapping CNVs	155 (55.2%)	150 (58.6%)	5 (20.0%)

*kb: 1000 bases (1 kilobase)

The lengths of CNV calls varied widely. The median CNV length was 4.0 kilobases (kb) while the mean CNV length was 46.4 kb, indicating that the distribution of lengths is very skewed to the right. While the CNV distribution in PennCNV is also skewed to the right, the mean length is only twice as high as the median length (rather than over 10 times higher). In CNV Workshop, the longest CNV was a duplication over 1400 kb long found on chromosome 15. In general, duplications were larger in size than deletions.

Nearly half of CNVs overlapped with one or more genes in the refGene database. While more duplications overlapped with a gene, the number of duplications is small so this discrepancy may not be meaningful. CNVs overlapping with genes are important since deletions and duplications in genes are most relevant to disease.

Given the large number of CNVs detected in the population, high-quality calls would be prioritized in a downstream analysis for the identification of novel IBD genes. CNVs containing few SNPs are often false positives and are not reliable. Typically, the minimum number of markers for a reliable CNV call is 10. In addition, out of the 281 CNVs identified in CNV Workshop, 130 corresponded to a CNV found by PennCNV, so slightly less than half of the variants were reproducible. An attempt to identify disease-causing genes from this data would begin from this subset of high quality and reproducible CNV calls.

In addition, annotation using known DGV variants provides information on whether variants in the region are common in a broader non-IBD population. While most CNVs detected in this analysis overlapped with a known variant in the DGV database, notably, a few variants were novel in that they did not overlap with any DGV variant, and can be examined for potential functional significance.

Discussion and Future Direction

In this analysis, over 200 CNVs were detected in 9 VEO-IBD patients using CNV Workshop. CNVs are important contributors to overall genetic variation, and functionally, gene duplications and deletions have profound effects on gene expression. These changes translate into altered protein levels that may ultimately lead to pathogenesis of diseases like IBD. VEO-IBD is rare, affecting 14 per 100,000 children on average, but rates of the disease are rising (Benchimol, et al. 2011). There is no cure for IBD and treatment focuses on managing the symptoms, so better therapy strategies need to be developed.

One limitation to CNV Workshop and the current analysis in general is the low degree of correspondence between PennCNV and CNV Workshop results. The lack of reproducibility suggests that these CNV analysis tools may be producing a large number of false positives. To improve the overall quality of CNV calls, a first step can be to allow the researcher to set a minimum number of markers or length in kilobases for a segment during the segmentation procedure. This can reduce the number of spurious and meaningless CNVs in the results. For example, the current minimum of two markers for a segment has little meaning, and would likely not be detected as (part of) a CNV if the minimum segment length was stricter.

Additionally, PennCNV uses a probabilistic model and is able to compute a quality score (Bayes factor) for a CNV based on the probability of the data given the null model (copy number two) versus the alternative model (deletion or duplication). Development of a similar measure for CNV Workshop to assess the confidence of each CNV call would be useful. One possible way to do this would be to determine the distribution of mean LRR and the various BAF statistics for deletions and duplications and then compute a likelihood statistic to show the likelihood of a given segment being a duplication/deletion. Currently, the CNV determination procedure uses hard limits for mean LRR to find duplications/deletions. The criteria were set by the developers of CNV Workshop presumably through testing with various data sets. A way to evaluate the quality of the call would be helpful for researchers who want to filter for potential disease-causing variants.

Still, CNV Workshop was able to identify a large number of genes that contain copy number variants in VEO-IBD patients. The next immediate step would be to try to narrow down the list of genes by filtering out low-quality CNVs. One initial method for filtering could involve only keeping copy number variants that are identified by both PennCNV and CNV Workshop. Then, candidate genes can be identified through the analysis results for possible rare or novel causative variants for VEO-IBD. This will include searching online databases such as Online Mendelian Inheritance in Man and the Jackson Laboratory's Human-Mouse: Disease Connection database for potentially functionally relevant genes, which can then be visualized in the Protein Data Bank for homology or functional domains.

References

Benchimol, Eric I., Kyle J. Fortinsky, Peter Gozdyra, Meta Van den Heuvel, Johan Van Limbergen, and Anne M. Griffiths. 2011. "Epidemiology of Pediatric Inflammatory Bowel Disease: A Systematic Review of International Trends." *Inflammatory Bowel Diseases* 17: 423-439.

Gai, Xiaowu, Juan C. Perin, Kevin Murphy, Ryan O'Hara, Monica D'arcy, Adam Wenocur, Hongbo M. Xie, Eric F. Rappaport, Tamim H. Shaikh, and Peter S. White. 2010. "CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics." *BMC Bioinformatics* 11: 1-9.

Goodman, Steven N. 1999. "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor." *Annals of Internal Medicine* 130: 1005-1013.

Jostins, Luke, Stephan Ripke, Rinse K. Weersma, Richard H. Duerr, Dermot P. McGovern, Ken Y. Hui, James C. Lee, L. Philip Schumm, and Yashoda Sharma. 2012. "Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease." *Nature* 491: 119-124.

Kelsen, Judith R., Noor Dawany, Christopher J. Moran, Britt-Sabina Petersen, Mahdi Sarmady, Ariella Sasson, Helen Pauly-Hubbard, et al. 2015. "Exome sequencing analysis reveals variants in primary immunodeficiency genes in patients with very early onset inflammatory bowel disease." *Gastroenterology* 149: 1415-1424.

Kelsen, Judith, and Robert N. Baldassano. 2008. "Inflammatory bowel disease: The difference between children and adults." *Inflammatory Bowel Diseases* 14: S9-S11.

Lin, Chiao-Feng, Adam C. Naj, and Li-San Wang. 2013. "Analyzing Copy Number Variation using SNP Array Data: Protocols for Calling CNV and Association Tests." *Current Protocols in Human Genetics* 1.27: 1-15.

MacDonald, Jeffrey R., Robert Ziman, Ryan K. C. Yuen, Lars Feuk, and Stephen W. Scherer. 2014. "The Database of Genomic Variants: a curated collection of structural variation in the human genome." *Nucleic Acids Research* 42: D986-D992.

Olshen, Adam B., and E. S. Venkatraman. 2004. "Circular binary segmentation for the analysis of array-based DNA copy number data." *Biostatistics* 5: 557-572.

Podolsky, Daniel K. 1991. "Inflammatory bowel disease." *New England Journal of Medicine* 325: 928-937.

R Core Team. 2015. "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ramasundara, Malith, Steven T. Leach, Daniel A. Lemberg, and Andrew S. Day. 2009. "Defensins and inflammation: The role of defensins in inflammatory bowel disease." *Journal fo Gastroenterology and Hepatology* 24: 202-208.

Uhlig, Holm H., Tobias Schwerd, Sibylle Koletzko, Neil Shah, Jochen Kammermeier, Abdul Elkadri, Jodie Ouahed, David C. Wilson, Simon P. Travis, and Dan Turner. 2014. "The Diagnostic Approach to Monogenic Very Early Onset Inflammatory Bowel Disease." *Gastroenterology* 147: 990-1007.

Wang, Kai, Mingyao Li, Rui Liu, Joseph Glessner, Struan F. A. Grant, Hakon Hakonarson, and Maja Bucan. 2007. "PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data." *Genome Research* 17: 1665-1674.