Identifying Variants Associated with Left Ventricular Non-Compaction By Incorporating Known INDELs Into the Human Reference Genome

Rachel L. Goldfeder, B.S., Stanford University Biomedical Informatics Program

Abstract

Little is known about the genetic mutations underlying the development of Left Ventricular Non-Compaction (LVNC), a cardiomyopathy associated with heart failure. I hypothesize that small insertions or deletions (INDELs), which are known to play a significant role in disease and genetic variation, cause this disease. However, INDEL detection from next-generation sequencing data is still a major informatics challenge. There are many software programs available for detecting INDELs, but these programs depend on receiving accurately aligned reads as input, and, unfortunately, sequencing reads that contain insertions or deletions are difficult to properly align to the reference genome.

In this project, I identified INDELs that are known to be present in humans and incorporated these INDELs into the human reference genome, hg19. To test the effectiveness of using the updated reference genome, I aligned whole genome sequence reads from a publically available human genome (NA12878) to this new reference genome and to hg19 separately. I evaluated the variant calls from each set by comparing to a gold standard set of INDELs known to be present in the genome of NA12878. Using the new reference genome provided slightly increased recall at the cost of slightly decreased precision. Finally, I used the INDEL-inclusive reference genome to identify INDELS that are associated with LVNC. Using exome sequencing data from five affected members of a family and one healthy family member, I found 1,162 INDELs that segregate with the disease; 63 are exonic and 50 of those are likely to be deleterious. This project provides an INDEL-inclusive reference genome that other researchers can use to identify INDELs associated with other diseases beyond LVNC.

Introduction

Left Ventricular Non-Compaction (LVNC) is a disease of the heart muscle where the wall of the left-ventricle appears spongy. Symptoms of LVNC include increased risk of blood clots, intolerance to exercise, and increased risk of sudden cardiac death; it is estimated that .014% to 1.3% of the population has this condition [1]. A typical heart has pieces of heart muscle that project from the inner heart wall of the left ventricle towards the inside of the chamber, known as trabeculations; as a normal heart develops, these trabeculations become compacted. LVNC occurs when these trabeculations do not become compacted (Figure 1) [2]. This disease is typically diagnosed by measuring wall-thickness and looking for trabeculations from an echocardiogram [1]. LVNC exhibits an autosomal dominant model of inheritance, but little is known about the precise mutations or genes involved in causing the disease.

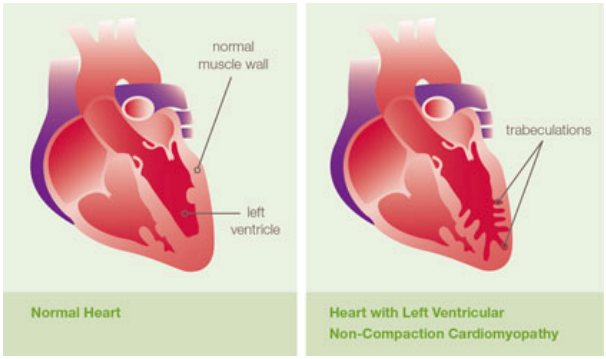

Figure 1: Cartoon showing the trabeculations present in the heart of an individual with LVNC[1]

My lab group studies this disease; we have access to a dataset from the Stanford Center for Inherited Cardiovascular Disease that includes a family with several members affected by LVNC. Each member of the family had an echocardiogram, which is the gold-standard for identifying LVNC, to determine whether or not he/she has LVNC. The dataset includes whole exome sequencing from 5 affected family members, and one unaffected family member.

We have been analyzing this data, but we have not yet found a genetic explanation for LVNC using next-generation sequencing and traditional tools focused on SNPs or using arrayCGH to identify large structural variants. I hypothesize that small insertions or deletions (INDELs), which are known to play a significant role in disease and genetic variation [3], cause this disease.

Detecting INDELs from next-generation short sequence reads presents many challenges. For instance, the sequencing technologies have difficulty correctly determining the DNA sequence in long homopolymer runs, which can lead to errors that look like INDELs. Additionally, alignment algorithms have difficulty correctly aligning short reads that contain an INDEL to the appropriate genomic positions, and even if the reads are correctly placed, there can still be alignment errors due to the repetitive nature of the human genome [4].

There are many software programs available for detecting INDELs, but, in general, they rely on receiving accurately aligned reads as input. For example, Dindel, the INDEL caller used in the 1000 Genomes Project, examines all of the reads that align within a pre-specified window, and then uses these reads to determine the most likely haplotype within the window [5]. However, if the sequence reads that contain information about the INDEL do not align properly, there will not be evidence within the window to detect an INDEL; this can create both false negatives (in the location the reads should have aligned to) and false positives (in the location the reads mistakenly align to).

The best way to get around this issue is to use genome-wide de novo assembly—leveraging the overlap of the reads to determine the sequence of nucleotides in the genome of interest – instead of alignment to a reference genome. However, this is extremely computationally expensive. In a recent study, assembling an entire human genome with 30x coverage (using 100 base-pair paired end reads) took nearly four days on a cluster with 150 cores [6]. The same study reports that approximately 80Gb of memory was required for one human chromosome, and the authors note that an entire genome would require "significantly more memory." This requires more resources than what is available in most research laboratories.

There are some efforts to correct for errors made in alignment – namely local realignment. In local realignment, all sequence reads are aligned to the human reference genome, and then initial alignment is refined by local realignment within candidate windows (typically these candidate regions are regions with known INDELs). Though local realignment is an excellent approach, greatly improving sensitivity and accuracy [7], it still does not prevent reads from being aligned to the wrong place. If the "wrong place" is not nearby (for instance, on a different chromosome), local realignment will not be able to correct for this error.

Previous studies have shown that a reference genome that includes common SNPs can improve genotype accuracy for disease-associated SNPs [8]. I hypothesize that a similar effect on the accuracy of calling disease-associated INDELs will be seen with a reference genome that includes INDELs. Therefore, in this project, I present an alternate reference genome that includes known INDELs (found by Mills et al and 1000 Genomes Project) [3, 9, 10]. I show how this reference genome, which I will refer to as RG1, increases the INDEL detection recall using sequence reads from public sample NA12878 and gold standard INDEL calls on this individual from the Genome in a Bottle Consortium [11]. Finally, I use RG1 to detect INDELs that are associated with LVNC.

Methods

Incorporation of INDELs into the human reference genome:
I downloaded INDELs that are known to exist in the population from the GATK bundle, which contains INDELs discovered in the 1000 Genomes Project and from Mills et al in 2006 [10]. Using custom perl scripts, I modified the human reference genome, hg19, to include the appropriate extra bases at genomic positions where insertions are found and remove bases at genomic positions where deletions are found. For the purposes of this project, I only incorporated insertions or deletions at locations where there was only one known INDEL as an alternative to the reference genome. I also kept track of the relative change in genomic coordinates so that it is easy to translate the coordinates of an INDEL found using RG1 to hg19 coordinates (Figure 2).

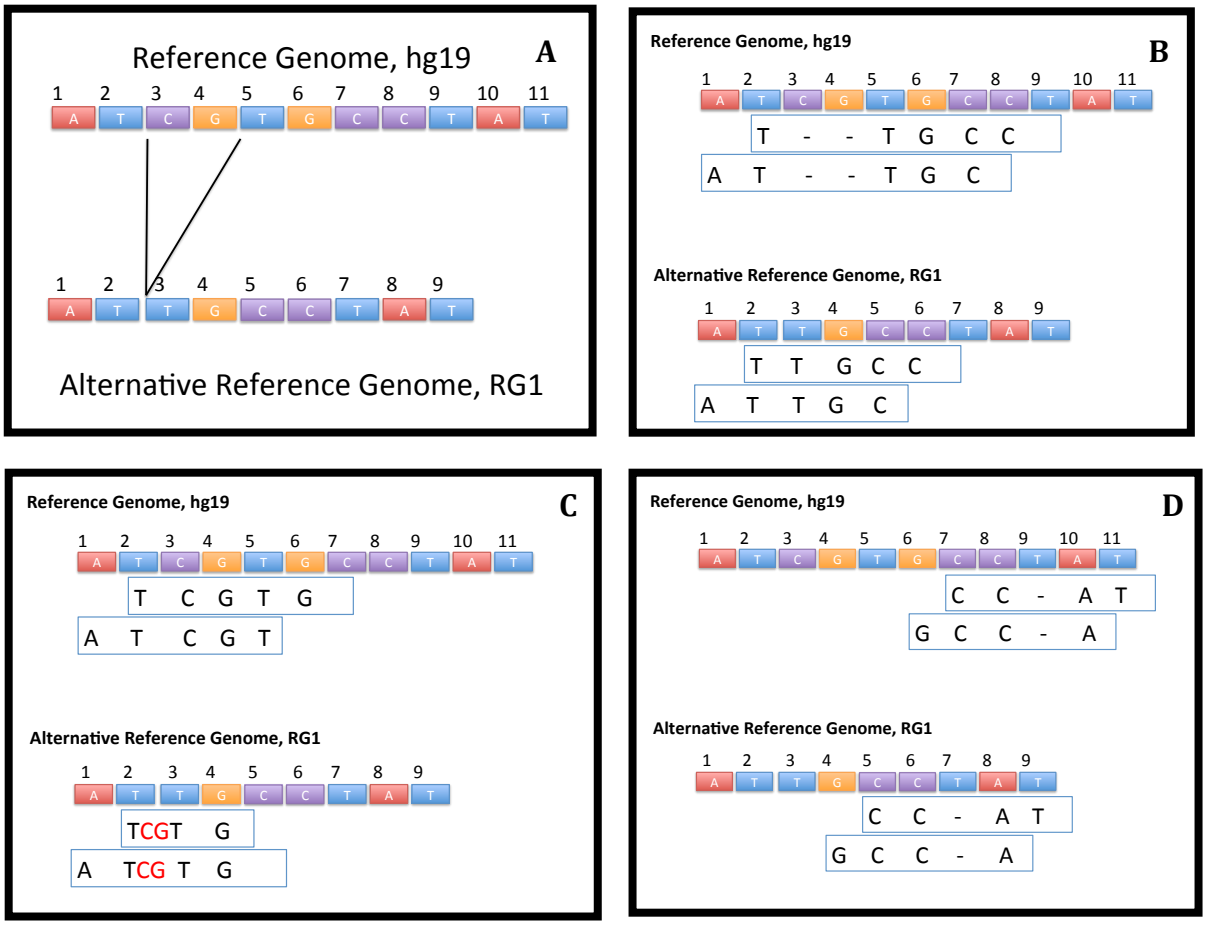

Figure 2: Overview of approach. Figure 2A shows how RG1 was created by adding and deleting bases to and from hg19. The top genome represents the reference genome, hg19, and the bottom genome represents the alternative reference genome, RG1; the genomic positions are labeled above each reference genome. In Figure 2A, an example deletion is shown: a deletion was made that removes nucleotides seven and eight from hg19. This deletion changes the coordinates of all nucleotides that follow the deletion to new coordinates in RG1. For instance, the "T" at position 5 in hg19, is at position 3 in RG1. In Figure 2B-D, the sequence reads within each panel contain the same bases, and their alignment to each reference genome is shown. Panel 2B shows the advantage of using RG1; if an INDEL that was incorporated into RG1 is seen in the person of interest, the reads will align better to RG1 than hg19. The INDEL will be reported during the coordinate translation from RG1 to hg19. Figure 2C shows the disadvantage to this approach; if an INDEL that was incorporated into RG1 is not seen in the person of interest, the reads will align better to hg19 than RG1. After coordinate translation, this will be reported as a reference call. Figure 2D shows the case where an INDEL is seen in the person of interest, but not in RG1 or hg19. In this case, coordinates of the INDEL are simply translated to correct coordinates in hg19.

Evaluation of new reference genome:

I evaluated RG1 in terms of its ability to aid in detecting INDELs. To do this, I downloaded whole genome sequence data from a publically available genome, HapMap sample NA12878. This whole genome was sequenced on an Illumina HiSeq2000 using 100 base-pair paired end reads [12]. I aligned these sequence reads to hg19 and RG1 separately using BWA with default parameters [13]. Then, I used Dindel with default parameters to identify INDELs from each set of alignments and converted the genomic coordinates to hg19 coordinates to allow for comparison. I performed a filtering step and only retained variants with passed Dindel's quality filter (quality must be greater than 20 on Dindel's quality scale) and homopolymer filter (homopolymer length must be less than or equal to 10 bases). To assess accuracy, I compared to the "gold standard" set of INDELs found by Genome In A Bottle [11] using vcf-compare from vcftools to compare the positions of INDELs called. Exact genotypes (heterozygous versus homozygous) were not compared.

Translation of coordinates from RG1 to hg19:
When RG1 was created, I tracked the relative change in genomic position that came about with the incorporation of each INDEL. When RG1 is used for alignment and variants are identified using that alignment, the coordinates of those variants must be translated back to hg19 for comparison with other studies. To do this, I wrote custom perl scripts that use the information about each position in RG1's difference in position relative to hg19 to convert back to hg19 coordinates. Before reporting final INDEL calls, I intersect the hg19 coordinates of the INDELs found using RG1 with the list of INDELs that were incorporated to create RG1. If an insertion is seen in the person of interest, but RG1 had a deletion incorporated at that position, the genotype at that position is reported as reference. Similarly, if a deletion is seen in the person of interest, but RG1 had an insertion incorporated at that position, the genotype at that position is reported as reference. If no variant is detected in RG1 coordinates at a location where an INDEL was incorporated, an INDEL is reported at that location. If there is an INDEL detected at a location where no INDELs were incorporated into RG1, the INDEL is reported in hg19 coordinates.

Discovery of causative variants for Left Ventricular Non-Compaction:
I obtained exome sequencing data for 5 family members with LVNC, and one healthy family member from labmates. Each exome was sequenced using 100 base-pair paired end reads on an Illumina HiSeq2000; the exome capture was performed using a Nimblegen capture methodology. For each individual in the dataset, I aligned sequence reads to RG1 and then translated coordinates to hg19. I leveraged the family structure of the dataset to create a list of INDELs that appropriately segregate with LVNC in the family. Custom perl and shell scripts were used for filtering and segregation analyses.

Results

I incorporated 728,592 deletions and 450,770 insertions into the human reference genome, hg19. The size distribution of these INDELs is shown in Figure 3. The vast majority, 1,061,452 of 1,179,362 (90 %), of incorporated INDELs were less than 5 bases long.

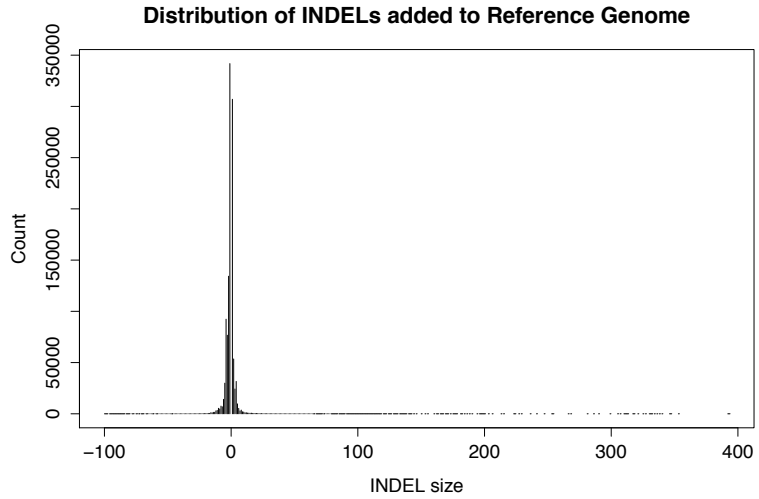

Figure 3: Size distribution of INDELs from Mills and 1000G. INDELs with a negative size are deletions, and those with a positive size are insertions. Most incorporated INDELs are small (less than 5 bases long), but some were several hundred bases long.

To test the new genome, I downloaded and aligned whole genome sequence data from NA12878, as described in the methods section, to hg19 and RG1 separately and then Dindel was used to detect INDELs. This whole genome had a median coverage of 28x. I found 371,726 INDELs that passed quality filters when I aligned to hg19 and 394,621 INDELs that passed quality filters when I aligned to RG1 (Table 1). Note that all INDELs found in RG1 were

translated to hg19 coordinates, as described in the methods. The Genome in a Bottle Consortium has published gold standard INDEL calls for NA12878 [11]; using these 174,883 gold standard INDELs as ground truth, the majority of INDELs that I detected are false positives (222,892 when aligning with hg19 and 245,307 INDELs when aligning with RG1). Precision is slightly lowered when aligning with RG1, but recall is slightly increased (Table 1).

Table 1: Summary of INDELs detected using hg19 or RG1 for NA12878.

	Number of INDELs detected	INDELs Passing filters	Number of False Positives	Precision	Recall
hg19	559,697	371,726	222,892	0.4004	0.8510
RG1	582,063	394,621	245,307	0.3784	0.8538

With the error profile from Table 1 in mind, I aligned exome sequencing data from 6 members of a family affected by LVNC to RG1, including 5 affected individuals and 1 healthy control family member. The coverage of these exomes was quite high; ranging from 85x to 112x median coverage. After aligning each individual's exome to RG1 and calling INDELs with Dindel, I found an average of 65,849 INDELs per individual (ranging from 58,670 to 72,374 per person). After filtering INDELs that had a low variant quality score and INDELs within homopolymer runs, as recommended by Dindel's "best practices" manual, there were an average of 37,611 INDELs per individual remaining (ranging from 33,932 to 41,674 per person).

There were a total of 77,002 unique INDELs across all individuals. Of those INDELs, only 4,274 segregated with the disease in this family -- meaning that the INDEL was present in all individuals with LVNC and not in the healthy individual. 63 INDELs were within exons, and 50 of those caused a frameshifting mutation (28 deletions and 22 insertions), or a change in the open reading frame, which are hypothesized to be extremely damaging (see Discussion). All 50 INDELs are reported in 1000 Genomes with a frequency less than 1% or were not seen in the 1000 Genomes project dataset. These 50 INDELs were found within 20 genes, listed in the supplementary material.

Table 2: INDELs detected using RG1 for a family with LVNC. Note LVNC10 is a healthy family member; all other participants are diagnosed with LVNC.

	LVNC10 (healthy)	LVNC04	LVNC12	LVNC13	LVNC14	LVNC15
Median Coverage	86	85	88	112	104	112
Number of INDELs	65,730	60,918	58,670	67,138	72,374	70,266
Number of INDELs Passing Quality Filters	38,634	35,688	33,932	37,260	41,674	38,480
Number of INDELs that Segregate with the Disease	4,274					
Number of INDELs within Exons	63					
Number of frameshifting INDELs	50					

Discussion

In this project, I accomplished the following: 1) created an INDEL-inclusive reference genome, RG1, by incorporating INDELs that are known to exist in the human population into hg19, 2) evaluated RG1's effectiveness and error profile, and 3) used RG1 to align whole exomes from individuals with LVNC in order to discover INDELs associated with LVNC. A key aspect of this project is the ability to convert from RG1 coordinates to hg19 coordinates after aligning with RG1. This is important because it is crucial for investigators to have a standard set of coordinates to use to report and discuss variations.

To evaluate RG1's effectiveness, I aligned sequence reads from NA12878 to each reference genome separately, and ran Dindel software to detect INDELs from each alignment. In both cases, I found many more INDELs than the gold standard set provided by the Genome in a Bottle Consortium. This result makes sense given that the gold standard INDEL calls only include calls that have been agreed upon by multiple sequencing platforms and algorithms [14] and given the fact that INDEL agreement across multiple platforms is low [15, 16].

When aligning with RG1, there were 22,366 more INDELs found than when aligning with hg19. I found more true positives and more false positives when aligning with RG1. The precision of using RG1 is slightly lower than hg19, but the recall when using RG1 is slightly increased. It is hard to say if the slight increase in recall when using RG1 is worth the decrease in precision; by providing this alternative reference genome, each researcher has the opportunity to make this decision based on their own experimental design. In the case of LVNC, where we have exhausted other methods for finding a causative variant, having this increased recall is an advantage even if it means more false positives. Note that specificity measures are not reported due to the vast number (approximately 3 billion) of "true negative," or non-INDEL calls.

To identify variants associated with LVNC, I aligned whole exomes from 5 family members affected with LVNC, and one healthy family member to RG1. After aligning the sequence reads, calling INDELs with Dindel, and performing appropriate quality filters, I filtered out any INDELs that did not segregate with the disease. In other words, I filtered out any INDEL that did not follow a pattern where the INDEL was present in all individuals with LVNC and the INDEL was not present in the healthy individual. Even though these individuals had exome sequencing performed, there are still some non-exonic regions (ie: intergenic, intronic, etc) captured and sequenced. Only 63 of the segregating variants were found in exons, and, of those, 50 caused a frameshifting insertion or deletion. Frameshifting INDELs (ie: an INDEL with size that is not a multiple of 3) are hypothesized to be particularly deleterious because the majority of the amino acids following the mutation will be changed, which causes a significant impact on the resulting protein.

These 50 INDELs associated with LVNC were all rare variants, meaning that they were seen in the 1000 Genomes dataset at a frequency of less than or equal to 1% or not seen in the 1000 Genomes dataset at all. On the one hand, this provides more support for the idea that these candidate variants could indeed be causing LVNC -- since the disease is so rare, I would not expect to see a variant that is common in the population causing the disease. On the other hand, in the evaluation of RG1, I found many false positives and low precision, so it is also possible that these INDELs are errors. My lab will need to perform validation assays to ensure that these INDELs are true variants before pursuing experimental follow-up studies to assess causality and confirm function.

Challenges and Future Directions

This project contributes an alternative INDEL-inclusive reference genome for use in the alignment of next-generation DNA sequence data, code to convert from RG1 coordinates to hg19 coordinates, and INDELs associated with LVNC. Future directions include downloading more whole genome sequence data for NA12878 to have greater depth of coverage. In this case, it would be possible to titrate the number of sequence reads included to see if the accuracy of INDEL calling changes with either hg19 or RG1 with a higher or lower depth of coverage. Additionally, there are a vast number of parameters that could be varied in this analysis.

One of the challenges of this work is choosing the best parameters for alignment and INDEL calling. I used BWA and Dindel with default parameters, but parameters for alignment with BWA or variant calling with Dindel could be altered to find the combination of parameters that provides the highest accuracy. Of course, different aligners (Novoalign, RTG, etc) and different INDEL callers (SOAPindel, Pindel, etc) could also be used; it would be

interesting to see if INDEL calls have particularly high accuracy with either RG1 or hg19 in any of these combinations. Future work could also include evaluating the accuracy of SNP calls when using RG1 instead of hg19.

Another challenge in this analysis is that each INDEL that is incorporated into RG1 that is not present in the genome being analyzed creates an opportunity for a false positive since alignment will be challenging – the exact problem that I am trying to mitigate by using RG1. It would be interesting to create and evaluate ethnicity specific INDEL-inclusive reference genomes, using population information about the linkage disequilibrium of INDELs. This would allow a smaller set of INDELs to be incorporated into the reference genome that will be more similar to the genome being analyzed. This would be beneficial since alignment will improve as the similarity between the reference genome and the genome being analyzed increase.

In the future, it will be important to look for INDELs, SNPs, and other types of variation together. Though it is a challenge to integrate different data types, diseases are often caused by complex combinations of variants – and those variants may not be the same type (ie: a SNP and an INDEL may have additive effects that cause a disease). This analysis identifying INDELs could be incorporated into a much larger analysis that looks for combinations of many types of variants.

In terms of finding variants that cause LVNC, future directions include validation of the genotype calls, both computationally and experimentally, to ensure the variant call was correct and to assess causality. Since LVNC is a severe disease, we hypothesize that the causative variant would likely be in a protein-coding region, so the family was sequenced using exome sequencing. However, future directions could include whole genome sequencing to search for variants in regulatory regions or other non-coding regions.

Conclusions

Identifying INDELs that may cause LVNC is extremely important for understanding the disease and creating potential drug targets or other treatments. If one of the candidate INDELs is a marker for the disease and/or causes the disease, the variant could be used in genetic screening to aid in early diagnosis of LVNC, potentially saving many lives. In addition to the important impact these findings could have for individuals affected by LVNC, this project also provides an INDEL-inclusive reference genome that other researchers can use to identify INDELs associated with other diseases.

References

[1] Feltes-Guzmán G, Núñez-Gil I.J. Left ventricular noncompaction. E-journal of Cardiology Practice. 2012. Available from: http://www.escardio.org/communities/councils/ccp/e-journal/volume10/Pages/Echocardiographiccriteria.aspx#.UyE5VfSwIv4

[2] Left ventricular non-compaction cardiomyopathy (LVNC). Available from: http://www.cincinnatichildrens.org/service/c/cardiomyopathy/types/left-ventricular-non-compaction-cardiomyopathy/

[3] Mills RE, Luttig, CT, Larkins CE, et al. An intitial map of insertion and deletion (INDEL) variation in the human genome. Genome Research. 2006; 16(9): 1182–1190.

[4] Montgomery SB, Goode DL, Kvikstad E, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Research. 2013; 23(5): 749–761.

[5] Albers CA, Lunter G, MacArthur DG, MvVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Research. 2010

[6] De Novo Assembly Using Illumina Reads. Available from: http://res.illumina.com/documents/products/technotes/technote_denovo_assembly_ecoli.pdf

[7] DePristo MA, Banks E, Poplin RE, et al. A framework for variation discovery and genotyping uing next-generation DNA sequencing data. Nature Genetics. 2011; 43(5):491-498.

[8] Dewey FE, Chen R, Cordero SP, et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS Genetics. 2011. 7(9): e1002280.

[9] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010. 467;1061-1073.

[10] GATK: what's in the resource bundle and how can I get it?. Available from: http://gatkforums.broadinstitute.org/discussion/1213/what-s-in-the-resource-bundle-and-how-can-i-get-it

[11] Genome in a bottle FTP site now live at NCBI. Available from: http://genomeinabottle.org/blog-entry/genome-bottle-ftp-site-now-live-ncbi

[12] DRASearch: ERS179577. Available from: https://trace.ddbj.nig.ac.jp/DRASearch/sample?acc=ERS179577

[13] Li H. and Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics, 2009. 25:1754-60.

[14] Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide Winston, Salt M. Integrating human sequence datasets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology. 2014. 32;245-251.

[15] Dewey FE, Grove ME, Pan C, et al. Clinical interpretation and implications of whole-genome sequencing. JAMA.2014;311(10):1035-1045.

[16] O'Rawe J, Jiang T, Sun G, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Medicine. 2013:5:28

Supplemental Materials

LVNC Candidate Genes:

DIP2C
IDI2
PFKP
AKR1C4
CALML5
A2M
NID2
DAAM1
ABCC6
RBBP6
RBL2
DFFB
EPHA8
APH1A
IL17RC
TSC22D2
RGNEF
ACOT12
ADAM32
KANK1