Identifying the Causal Variant in an Uncharacterized Rare Bone Disorder

499 Progress Report after 3 Semesters

ABSTRACT

Rare human disorders can be often difficult to study because of the challenge of ascertaining a sufficient number of affected individuals to carry out a successful study. However, successful studies into the nature of rare human disorders often lead to novel and significant insights into fundamental biological processes. Those affected by rare disorders, in the meantime, are often left with few to no options when it comes to specialized diagnosis and treatment. Genomic sequencing and analysis provides a possible solution by attempting to identify, on a genetic level, the cause of the rare disorder. By calling variants of a patient's genome against genomes of unaffected individuals, we can compile and sort variations to locate the causal variant. In this study, consanguineous siblings from first-cousin parents are presented with symptoms including insulin resistance, deformed hands/feet, and short stature. This likely recessive disorder mirrors the symptoms of dominant disorders Albrights Hereditary Dystrophy and Acrodysostosis, suggesting this disorder is expected to affect cAMP, cGMP, or calcium signaling, bone growth, and/or cartilage development. By searching through all possible variants, suspected causal variants were narrowed down and deeply characterized. Variation of the gene *DUSP7*, responsible for dual-specificity phosphatase and interaction with human growth hormone receptor (GHR) is expected to be the cause of this disorder. Correct identification of the causal variant will yield insight into the function of the gene in organismal function.

INTRODUCTION

Whole-genome sequencing (WGS) combined with computational analysis is quickly becoming the most cost-effective and efficient way to identify genetic variation in afflicted individuals[1]. This is most useful when trying to identify the cause of a rare disorder[2]. By using WGS to determine the complete DNA sequence of an individual/individuals affected by a rare disorder and computational analysis to to analyze variation against unaffected genomes, we can facilitate the diagnosis and discovery of new disorders while overcoming the challenge of finding sufficient affected individuals for a traditional study[1]. Studying rare disorders gives new insight into how specific biological processes function.

We performed the whole genome sequencing and computational analysis of two consanguineous siblings of second-cousin parents. The ultimate goal is to use WGS to identify the variant(s) responsible for their rare unnamed disorder, characterized by insulin resistance, deformed hands/feet, and short stature. Both siblings have acanthosis nigricans (characterized by dark patches in the folds of skin), spinal stenosis, hip dysplasia, brachydactyly, and recurrent otitis media (recurring middle ear infection). Proband SL106253 also has developmental delay caused by gliosis of subcortical white matter.

Symptoms mirror those of Albrights Hereditary Dystrophy and Acrodysostosis. Previous studies have identified causal variants of Albrights in *GNAS1* and Acrodysostosis to both the *PRKAR1A* and *PDE4D* genes[3,4]. These genes influence protein kinase signaling and cAMP regulatory activity, yet cannot be confirmed as causal variants of this disorder[3]. By combining WGS / computational analysis with Runs of Homozygosity (RoH) analysis based on consanguinity, we can hypothesize which areas are most likely to be compromised. Therefore, when searching for variants, we can expect the causal variant to follow an autosomal recessive inheritance pattern and compromise cAMP signaling, cGMP bone growth, or cartilage formation.

After WGS, mapping, and variant discovery, the proband genomes are run through Bystro, a program that provides annotation for each variant call with information that includes allele frequency, functional classification, CADD score, and a host of other descriptors[9]. The CADD score is a subjective estimation on how deleterious a mutation is predicted to be. Bystro creates a list of categorized gene information based on search queries relevant to the disorder. Variants that do not match potential criteria are eliminated until a list of potential causal variants are compiled. In parallel, the program PLINK is used to compare runs of homozygosity in the consanguineous siblings' genomes. Genes that overlap the two independent projects are quartered and further analyzed via deep characterization and literature review. By doing this, we can create a concise list of variants more likely to be causing the rare disorder. This will both help those affected by rare disorders find more suitable treatment, and researchers to better understand the biological process compromised.

METHODS

Whole Genome Sequencing

Genomes were first cut to around 400 base pairs. Samples were sequenced using the Illumina HiSeq Ex platform, underwent ligation via AMPure XP beads, and further amplified through multiple rounds of PCR. Illumina software generated files containing base pair reads for the sample.

Sequence Alignment

Generated reads from WGS are stored as fastq.gz files. The shell program PEMapper is used to align reads (in fastq.gz format) to a reference genome. Once launched, the program outputs pileup.gz and indel.txt.gz files. These files contain information on different mutations, such as insertions, deletions, and frameshifts, on the mapped genomes.

Variant Calling

The shell program PECaller pulls together pileup.gz and indel.txt.gz files of the cases' genomes. It 'joint-calls' these files with 59 other unaffected genomes to give every location in the mapped genome where variation exists. PECaller outputs a shell replicate of the call, along with an SNP file containing all genomes (59 plus the genomes of interest). This file contains every case of variance from every genome; however, the file only outputs the counts for certain insertions/deletions/substitutions, without base identification. This step collects everything that could possibly be calling the disorder.

Post-Call Processing

Post-call processing, which involves cleaning up base-pair duplicates, helps reveal the identity of base-pair insertions and deletions. The process of Merging 'merges' the SNP file output of PECaller and the indel.txt.gz file of PEMapper. The process of Filtering 'filters' the genome(s) of interest from the 59 they were called against. Filtering is an optional process that is only sometimes utilized; for instance, if we were interested in finding homozygous variants that are *only* present in the genome(s) of interest, we would continue with the process so we do not lose that information.

Quality Control

Quality Control is used to identify and remove sites that are likely to be sources of error from the analysis. Quality Control (QC) was used first used when calling alongside 59 high quality genomes. The genome is checked for a consistent transversion/transition ratio, as well as a silent/replacement ratio. The transition/transversion ratio should fall between 2.01 and 2.04, while silent/replacement should fall around 1.15. Deviance from these ranges raises alarm for some error in the calling process or genome quality. The program PLINK is used to further quality control

based on identity by descent (IBD). Calls with more than 10% missing data are automatically removed.

Analysis

After calling, post-calling, and quality control, the genome(s) is/are run through the program *Bystro*. This program creates search queries with all variants for affected sites near known genes. Based on information regarding the familial inheritance and phenotypes of the disorder, we create different search queries. Filtering by exonic variation, frequency in populations (minor allele frequency), and proposed inheritance pattern (autosomal dominant, autosomal recessive) narrows down the list of variants. The resource gnomAD (Genomic Aggregate Database) provides information on allele frequency in the general population – thus, any homozygous recessive set of alleles with a frequency greater than 0.001 (1 in 1000) were automatically eliminated. Variants that fit these profiles are researched for phenotypic relevance to the disorder at hand.

Runs of Homozygosity

In tandem with variant analysis, we analyzed the runs of homozygosity via PLINK. The consanguineous nature of the parents means approximately 1/64 of the siblings' genomes will be homozygous. Using that, we identified runs of homozygosity that were allelically identical and cross-referenced the results with the original analysis. Variants that appeared on both lists were regarded as promising, and underwent further deep characterization.

RESULTS

Our first hypothesis was to analyze the known causes of similar disorders to see if similar genes were affected. So, the causal variants of Albright's Hereditary Dystrophy and Acrodysostosis were analyzed in the probands to see if our disorder follows these patterns. As explained in the Introduction, Albright's is associated with *GNAS1* while Acrodysostosis is associated with *PRKAR1A* and *PDE4D*. However, none of the mutations on these genes were homozygous only in the probands (SL106253 and SL106254) relative to the 59 unaffected genomes. Furthermore, none showed a minor allele frequency less than 0.001. We can effectively rule out these genes and, therefore, these disorders.

Our second hypothesis was to analyze homozygous variants shared by both siblings. The sequenced genome yielded 2.3 million variants; of these, only 4,375 variants were homozygous in both probands (see Fig. 1). We filtered out variants that had minor allele frequencies greater than 0.001 and CADD scores less than 15.

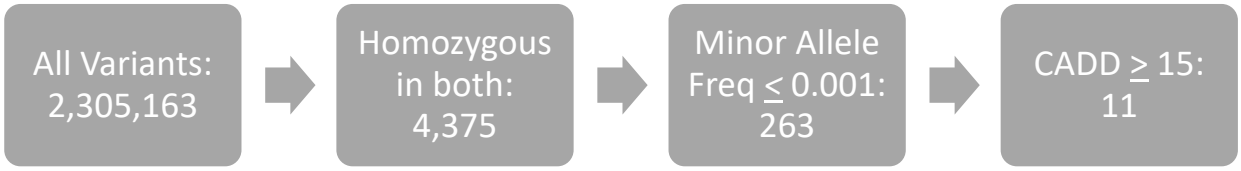

Figure 1.1 shows the process of narrowing down causal variants via suspected inheritance pattern, rarity, and deleteriousness.

Variant gene	**Chr#**	**Site Type**	**CADD Score**	**MAF (minor allele frequency)**	**Amino Acid Alteration**
USP19	3	Exonic	28.8	0.000743951	Arginine → Histidine
RNR2	M	Intergenic	19.3	0	N/A
RBM6	3	Exonic/Intronic/5'UTR	19	0.000549273	Alanine → Valine
DUSP7	3	Exonic	18.8	0	Tyrosine → Cysteine
GTF2IP7	7	Intronic	18.3	0	N/A

CTNNB1	3	Intergenic	17.7	0.000387622	N/A
ITA9	3	Intronic	17.3	0	N/A
BSN	3	Exonic	16.7	0.000355435	Valine → Methionine
POC1A	3	Intronic	16.3	0	N/A
MAP4	3	Intronic	16.2	0.000129374	N/A
ULK4	3	Intronic	15	0	N/A

Table 1.2 shows the list of all homozygous variants filtered by MAF and CADD.

Both parents are unaffected, but are expected to be heterozygous for the disease-causing mutation. Following that, the consanguineous nature of the family allowed us to narrow down potential variants to several large runs of homozygosity. For this family, approximately 1/64 of the genome is expected to be homozygous.

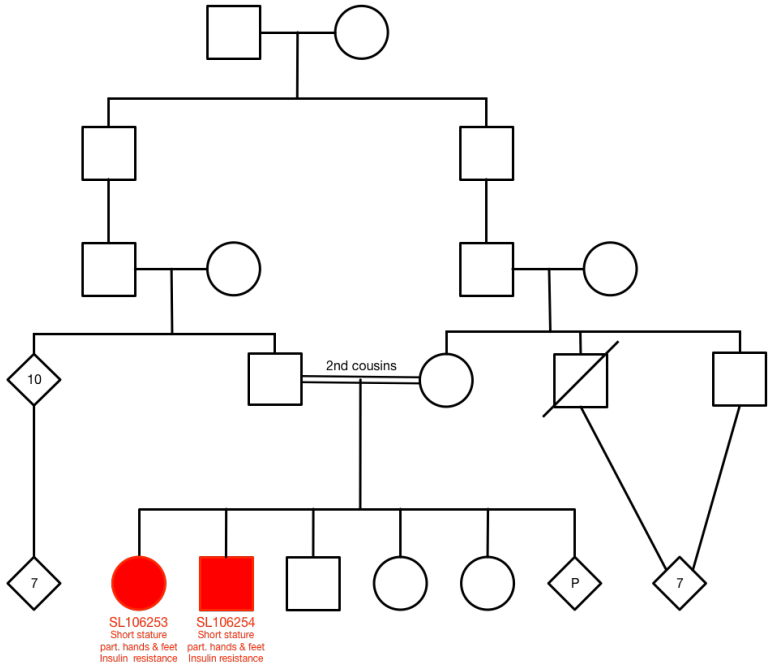

Figure 1.3 shows the heritage of our consanguineous parents and the resulting probands, along major physical phenotypes.

An analysis of runs of homozygosity was conducted using PLINK, and the variants present were cross-referenced with the *Bystro* analysis. The two samples shared 13 runs of homozygosity. Of those 13 runs, only 7 runs contain the same homozygous variants, *i.e.* shared allelically, exclusively on chromosome 3. This is an indication that our causal variant could exist somewhere on chromosome 3 of the affected siblings.

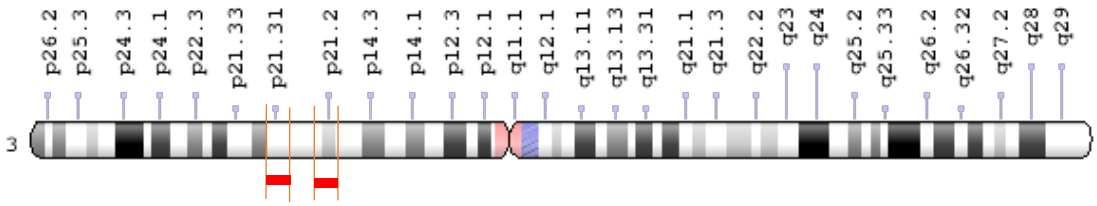

Figure 1.4 shows Chromosome 3, along with the location where the allelically identical Runs of Homozygosity were found using Plink analysis.

Variant gene	Chr#	Intronic / Exonic	CADD Score	MAF (minor allele frequency)	Amino Acid Alteration
USP19	3	Exonic	28.8	0.000743951	Arginine → Hystidine
RBM6	3	Exonic/Intronic/5'UTR	19	0.000549273	Alanine → Valine
DUSP7	3	Exonic	18.8	0	Tyrosine → Cysteine
CTNNB1	3	Intergenic	17.7	0.000387622	N/A
ITA9	3	Intronic	17.3	0	N/A
BSN	3	Exonic	16.7	0.000355435	Valine → Methionine
POC1A	3	Intronic	16.3	0	N/A
MAP4	3	Intronic	16.2	0.000129374	N/A
ULK4	3	Intronic	15	0	N/A

Table 1.5 shows a list of all variants revealed by RoH analysis. Note that the location of all revealed variants is exclusively on chromosome 3.

Variant gene	Chr#	Intronic / Exonic	CADD Score	MAF (minor allele frequency)	Amino Acid Alteration
DUSP7	3	Exonic	18.8	(no recorded frequency)	Tyrosine → Cysteine
MAP4	3	Intronic	16.2	0.000129374	Isoleucine → Serine
ITGA9	3	Intronic	17.3	(no recorded frequency)	N/A
POC1A	3	Intronic	16.3	(no recorded frequency)	N/A
ULK4	3	Intronic	15.0	(no recorded frequency)	N/A
NCKIPSD	3	Exonic	14.4	(no recorded frequency)	N/A

PCBP4	3	Exonic	14.5	0.0002	N/A

Table 1.6 shows the overlap between *Table 1.2* and *Table 1.5* – these variants are homozygous, CADD > 15, MAF < 0.001, and present in allelically identical RoHs.

These two sets of analyses working in tandem revealed variants that were indicated by gnomAD to be extremely rare, *and* exist within the 7 allelically identical runs of homozygosity. These variants are promising, since they were identified by two independent processes; they underwent further deep characterization to identify corresponding gene function. The disorder is characterized by insulin resistance, deformed hands/feet, and short stature.

Although the two-step process revealed this list of variants, deep characterization was necessary to identify what processes these genes were associated with. Two variants that underwent deep characterization stood out as potential causal variants due to high CADD Score, low frequency, and interesting interactors: *DUSP7* and *MAP4*. However, it should be noted there are more variant candidates that need to be looked into.

Discussion and Future Results

Rare bone disorders, though often not fatal, often render the affected unable to perform many facets of everyday life. Marble Bone Disease, which affects approximately 1 in 100,000 to 500,000 people, leave the affected susceptible to osteomyelitis and osteosclerosis[6]. Hypophosphatasia is a highly variable rare disorder that can lead to craniosysostosis, intercranial hypertension, and painful skeletal irregularities[7]. Albright's Hereditary Dystrophy and Acrodysostosis, as mentioned earlier, lead to cartilage deformation and short stature[3,4]. In all these cases, treatment is almost completely limited to bone marrow transplants. Further treatment only comes with a stronger understanding of the pathways and genes behind bone and cartilage development.

We have demonstrated a way to use WGS and computational analysis to narrow down variants associated with an uncharacterized rare bone disorder. This method bypasses the problem of ascertaining enough affected individuals to constitute a traditional study, and opens the door for exploring variants on an individual genetic level. In relation to our study, we subjected two top variants to deep characterization in order to explore compromised pathways and links to our phenotype. The first, our top candidate, is a missense mutation on the 3rd exon of the *DUSP7* gene. This gene is shown to interact with a host of MAPK kinases, GHR Human Growth Hormone receptor, *DUSP6* dual-specificity phosphatase, and PDE12 phosphodiesterase 12(NCBI). Interaction with the MAPK pathway has known physiological implications in skeletogenesis and postnatal bone maintenance[8]. These interactions could account for two associations, including cAMP via phosphodiesterase/phosphatase loss-of-function and short stature via GHR loss-of-function. The amino acid alteration introduces a sulfur-containing amino acid; this could cause a disulfide bridge that alters protein shape and therefore function. MAPK has been implicated in insulin pathways in tandem with GHR as well. MAPK can act as negative feedback to insulin; phosphorylation of the receptor deactivates insulin production. GHR mutations have been observed in Laron's Syndrome, characterized by short stature similar to that seen in our probands.

Figure 1.7 shows the location of the *DUSP7* gene on Chromosome 3.

The second variant is a missense mutation on the MAP4 gene. This gene is known to associate heavily with cartilage formation, via chondrocyte formation and development. Although an intronic mutation, its association with possible phenotypic expression prevents it from being ruled out. The missense mutation alters a hydrophobic amino acid (isoleucine) to a hydrophilic amino acid (serine), which suggests alteration of protein shape.

Figure 1.8 shows the location of the MAP4 gene on Chromosome 3.

From the data collected, a definite causal variant cannot be determined. Further deep characterization of the *DUSP7* and *MAP4* genes will grant more information. More genes in chromosome 3 must also be characterized to expand our pool of possible variants. Undergoing Sanger sequencing with a third, unaffected sibling from the consanguineous parents will also be of use; if the homozygous recessive mutation for *DUSP7* is not present in this sibling, there is a higher chance it must be the causal variant.

However, deep characterization has a ceiling. Without interacting with the pathway itself, *DUSP7* cannot be fully concluded as the causal variant. Experiments involving *DUSP7* knockout models could be utilized in order to observe what happens when the pathway loses function. Protein modeling could help show how the tyrosine to cysteine substitution affects protein structure. These experiments will either prove or disprove the *DUSP7* theory, allowing us to undergo deep characterization with other genes revealed by computational analysis and runs of homozygosity analysis.

References

1. Bainbridge, M. N., W. Wiszniewski, D. R. Murdock, J. Friedman, C. Gonzaga-Jauregui, I. Newsham, J. G. Reid, J. K. Fink, M. B. Morgan, M. C. Gingras, D. M. Muzny, L. D. Hoang, S. Yousaf, J. R. Lupski and R. A. Gibbs (2011). "Whole-genome sequencing for optimized patient management." *Sci Transl Med* **3**(87): 87re83.

2. Johnston, H. R., P. Chopra, T. S. Wingo, V. Patel, B. International Consortium on, S. Behavior in 22q11.2 Deletion, M. P. Epstein, J. G. Mulle, S. T. Warren, M. E. Zwick and D. J. Cutler (2017). "PEMapper and PECaller provide a simplified approach to whole-genome sequencing." *Proc Natl Acad Sci U S A* **114**(10): E1923-E1932.

3. Cutler, D. J., M. E. Zwick, D. T. Okou, S. Prahalad, T. Walters, S. L. Guthery, M. Dubinsky, R. Baldassano, W. V. Crandall, J. Rosh, J. Markowitz, M. Stephens, R. Kellermayer, M. Pfefferkorn, M. B. Heyman, N. LeLeiko, D. Mack, D. Moulton, M. D. Kappelman, A. Kumar, J. Prince, P. Bose, K. Mondal, D. Ramachandran, J. F. Bohnsack, A. M. Griffiths, Y. Haberman, J. Essers, S. D. Thompson, B. Aronow, D. J. Keljo, J. S. Hyams, L. A. Denson, P.-K. R. Group and S. Kugathasan (2015). "Dissecting Allele Architecture of Early Onset IBD Using High-Density Genotyping." *PLoS One* **10**(6): e0128074.

4. Linglart, A., H. Fryssira, O. Hiort, P. M. Holterhus, G. Perez de Nanclares, J. Argente, C. Heinrichs, A. Kuechler, G. Mantovani, B. Leheup, P. Wicart, V. Chassot, D. Schmidt, O. Rubio-Cabezas, A. Richter-Unruh, S. Berrade, A. Pereda, E. Boros, M. T. Munoz-Calvo, M. Castori, Y. Gunes, G. Bertrand, P. Bougneres, E. Clauser and C. Silve (2012). "PRKAR1A and PDE4D mutations cause acrodysostosis but two distinct syndromes with or without GPCR-signaling hormone resistance." *J Clin Endocrinol Metab* **97**(12): E2328-2338.

5. Wilson, L. C. and C. M. Hall (2002). "Albright's hereditary osteodystrophy and pseudohypoparathyroidism." *Semin Musculoskelet Radiol* **6**(4): 273-283.

6. Arumugam, E., Harinathbabu, M., Thillaigovindan, R., & Prabhu, G. (2015). Marble Bone Disease: A Rare Bone Disorder. *Cureus, 7*(10), e339. doi:10.7759/cureus.339

7. Wenkert, D., McAlister, W. H., Coburn, S. P., Zerega, J. A., Ryan, L. M., Ericson, K. L., . . . Whyte,

M. P. (2011). Hypophosphatasia: nonlethal disease despite skeletal presentation in utero (17 new cases and literature review). *J Bone Miner Res, 26*(10), 2389-2398. doi:10.1002/jbmr.454

8. Thouverey, C., & Caverzasio, J. (2015). Focus on the p38 MAPK signaling pathway in bone development and maintenance. *Bonekey Rep, 4*, 711. doi:10.1038/bonekey.2015.80

9. Kotlar, A. V., Trevino, C. E., Zwick, M. E., Cutler, D. J., & Wingo, T. S. (2017). Bystro: Rapid online variant annotation and natural-language filtering at whole-genome scale. *bioRxiv*. doi:10.1101/146514