HHS Public Access

Author manuscript

*Genet Med*. Author manuscript; available in PMC 2018 January 10.

Published in final edited form as:

*Genet Med*. 2018 January ; 20(1): 159–163. doi:10.1038/gim.2017.86.

# Long-read genome sequencing identifies causal structural variation in a Mendelian disease Running title: Long-read WGS identifies causal SV in a Mendelian disease

**Jason D. Merker, MD, PhD**[1,2], **Aaron M. Wenger, PhD**[3], **Tam Sneddon, DPhil**[2], **Megan Grove, MS,LCGC**[2], **Zach Zappala, PhD**[1,4], **Laure Fresard, PhD**[1], **Daryl Waggott, MSc**[5,6], **Sowmi Utiramerur, MS**[2], **Yanli Hou, PhD**[1], **Kevin S. Smith, PhD**[1], **Stephen B. Montgomery, PhD**[1,4], **Matthew Wheeler, MD, PhD**[5,6], **Jillian G. Buchan, PhD**[1,2], **Christine C. Lambert, BA**[3], **Kevin S. Eng, MS**[3], **Luke Hickey, BS**[3], **Jonas Korlach, PhD**[3], **James Ford, MD**[4,5,7], and **Euan A. Ashley, MRCP, DPhil**[2,4,5,6]

[1]Department of Pathology, Stanford University, Stanford, California, 94305

[2]Stanford Medicine Clinical Genomics Service, Stanford Health Care, Stanford, California, 94305

[3]Pacific Biosciences, Menlo Park, California, 94025

[4]Department of Genetics, Stanford University, Stanford, California, 94305

[5]Department of Medicine, Stanford University, Stanford, California, 94305

[6]Stanford Center for Inherited Cardiovascular Disease, Stanford University, Stanford, California, 94305

[7]Stanford Cancer Institute, Stanford, California, 94305

## Abstract

**Purpose—**Current clinical genomics assays primarily utilize short-read sequencing (SRS), but SRS has limited ability to evaluate repetitive regions and structural variants. Long-read sequencing (LRS) has complementary strengths, and we aimed to determine if LRS could offer a means to identify overlooked genetic variation in patients undiagnosed by SRS.

**Methods—**We performed low coverage genome LRS to identify structural variants in a patient who presented with multiple neoplasia and cardiac myxomata, in whom targeted clinical testing and genome SRS were negative.

**Results—**This LRS approach yielded 6,971 deletions and 6,821 insertions >50bp. Filtering for variants that are absent in an unrelated control and overlap a disease gene coding exon identified

Correspondence: Euan A. Ashley, Division of Cardiovascular Medicine, Falk Cardiovascular Research Building, 300 Pasteur Drive, Stanford, CA 94304; Tel: 650-736-7878; Fax: 650-498-7452; euan@stanford.edu.

The first two authors contributed equally to this work.

**Conflict of interest statement**

AW, CL, KE, LH, and JK are employees and shareholders of Pacific Biosciences, a company commercializing DNA sequencing technologies.

SUPPLEMENTARY MATERIAL

Supplementary material is available at the Genetics in Medicine website

three deletions and three insertions. One of these, a heterozygous 2,184 bp deletion, overlaps the first coding exon of *PRKAR1A*, which is implicated in autosomal dominant Carney complex. RNA sequencing demonstrated decreased *PRKAR1A* expression. The deletion was classified as pathogenic based on guidelines for interpretation of sequence variants.

**Conclusions—**This first successful application of genome LRS to identify a pathogenic variant in a patient suggests that LRS has significant potential to identify disease-causing structural variation. Larger studies will be ultimately required to evaluate the potential clinical utility of LRS.

### Keywords

Long-read sequencing; Carney complex; PRKAR1A; structural variant; PacBio

## INTRODUCTION

Short-read sequencing (SRS) methods are primarily used in clinical laboratory medicine because of their cost effectiveness and low per-base error rate. However, these methods do not capture the full range of genomic variation.[1] Areas of low complexity, such as repeats, and areas of high polymorphism, such as the HLA region, present challenges to SRS and reference-based genome assembly. Indeed, with 100 base pair (bp) read length, fully 5% of the genome cannot be uniquely mapped.[2] In addition, many diseases are caused by repeats in a range beyond the resolution of SRS. Another challenge comes in the form of structural variation, and although SRS has been very successful in the discovery of single nucleotide and small insertion-deletion variation, recent findings suggest we have greatly underestimated the extent and complexity of structural variation in the genome.[3,4]

Long-read sequencing (LRS), typified by PacBio® single molecule, real-time (SMRT®) sequencing, offers complementary strengths to SRS. PacBio LRS produces reads of several thousand base pairs with uniform coverage across sequence contexts.[5] Individual long reads have a lower accuracy (85%) than short reads, but errors are random and are correctable with sufficient coverage, leading to high consensus accuracy.[5,6] Furthermore, long reads are more accurately mapped to the genome and access regions that are beyond the reach of short reads.[1] Of note, recent PacBio LRS *de novo* human genome assemblies have revealed tens of thousands of structural variants per genome, many times more than previously observed with SRS.[3,7] These capabilities, together with continuing progress in throughput and cost, may make LRS an option for broader application in human genomics.

Here, we report the use of low coverage genome LRS to secure a diagnosis of Carney complex in a patient unsolved by clinical single gene testing and genome SRS. This initial application of LRS to identify a pathogenic structural variant in a patient when considered with other prior studies suggests that LRS can identify disease-causing structural variants that are difficult to detect with current technologies. Larger studies are needed to evaluate the molecular diagnostic yield and potential clinical utility of LRS.

## MATERIALS AND METHODS

### Case report

The patient is an Asian/Hispanic male, the product of an uncomplicated term pregnancy who was hospitalized for the first 10 days of life for cardiac and respiratory issues (Figure 1A). He remained well until the age of 7 years, when, following the discovery of a heart murmur, he was found to have a left atrial myxoma that was surgically removed. At 10 years, he was noted to have a testicular mass that, at orchiectomy, was found to be a Sertoli-Leydig cell tumor. At 13 years, a pituitary tumor was found and initial conservative management was adopted. Aged 16, he was noted to have both an adrenal microadenoma and recurrence of the cardiac myxomata in the left ventricle and right atrium. Blue naevi were reported. He underwent a second surgical resection of the myxomata with uncomplicated recovery. Aged 18, recurrent cardiac myxomata including a right ventricular and two left ventricular tumors were once again resected and a goretex patch was placed in the right ventricular wall. In the immediate post-operative period, he suffered ventricular tachycardia (VT) and cardiac arrest with spontaneous return of circulation. At this time, a genetics evaluation suggested the possibility of Carney complex but clinical sequencing of *PRKAR1A* was negative for disease causing variation. At age 19, multiple thyroid nodules were noted on ultrasound, and he was diagnosed with ACTH-independent Cushing's syndrome, secondary to the adrenal microadenoma. At 21, he was found to have a pituitary lesion and acromegaly. He subsequently underwent trans-sphenoidal resection of the pituitary tumor with pathology confirming a growth-hormone producing pituitary adenoma. At this time, he was found to have recurrent myxomata in the left ventricular outflow tract that have subsequently increased in size (Figure 1B–C). To date, these have been treated conservatively with anti-coagulation to reduce the risk of stroke. As of 2016, he is under consideration for heart transplantation, and the transplant team judged molecular confirmation of the clinical diagnosis desirable prior to transplant listing.

### Short read genome sequencing and analysis

A library was generated from genomic DNA using the Illumina® TruSeq® DNA PCR-Free Library Prep Kit and sequencing was performed on the Illumina HiSeq® 2500 System with paired-end 100 bp reads to a 36-fold mean depth of coverage. The Stanford Medicine Clinical Genomics Service performed the data analysis and variant curation. Single nucleotide variants (SNVs) and small insertions and deletions were identified using MedGAP v2.0, a pipeline based on GATK best practices for data pre-processing and variant discovery with GATK HaplotypeCaller v3.1.1.[8] This analysis pipeline did not identify any variants that would explain the clinical findings in the patient. Multiple short-read structural variant callers, including Pindel, Lumpy, BreakDancer, Manta, CNVKit, and CNVnator, were retrospectively used to identify structural variants,[9–14] as described in the Supplementary Materials and Methods.

### Long read genome sequencing and analysis

Following informed consent under a protocol approved by the Stanford University Institutional Review Board, low coverage genome LRS was performed on the PacBio Sequel™ system to evaluate structural variation. The sequencing generated a 9-fold mean

depth of coverage with an average read length of >9 kb. Further details provided in the Supplementary Materials and Methods.

### Other methods

RNA sequencing and parentage studies are described in Supplementary Materials and Methods.

## RESULTS

The resulting call set from LRS consisted of 6,971 deletions and 6,821 insertions >50 bp (Table S1). To prioritize candidate pathogenic variants, the call set was filtered to exclude variants within a segmental duplication or present in the unrelated control individual NA12878. This left 2,476 deletions and 3,171 insertions. Focusing on variants that overlap a RefSeq coding exon resulted in 39 deletions and 16 insertions, with 3 deletions and 3 insertions in genes linked to a genetic disease in OMIM. The three OMIM genes, as well as phenotype and mode of inheritance, included in deletions are: *CASP8* (autoimmune lymphoproliferative syndrome type IIB, autosomal recessive), *CD209* (susceptibility to or protection from certain pathogens), and *PRKAR1A* (Carney complex, autosomal dominant); and the three OMIM genes included in insertions are: *KALRN* (susceptibility to coronary heart disease), *PAPSS2* (brachyolmia, autosomal recessive), and *PCDH15* (Usher syndrome, autosomal recessive). Manual review of the 6 candidate variants and correlation with phenotype identified a heterozygous deletion that removes the first coding exon of *PRKAR1A* (NM_212472.2). Germline variants in *PRKAR1A* cause Carney complex, type 1 (MIM #160980), an autosomal dominant multiple neoplasia syndrome.

Two of four reads at the locus unambiguously support the presence of a deletion (Figure 2A). Because of the random errors in LRS, individual reads from the same allele can have slight disagreements, and two reads can be insufficient to define exact deletion breakpoints with full confidence. Here, the higher quality read supports a 2,184 bp deletion of GRCh37/hg19 chr17:66,510,475–66,512,658 (NC_000017.10:g.66510475_66512658del). This heterozygous deletion variant was validated by Sanger sequencing, confirming the precise breakpoints identified by LRS (Figure 2B). Sanger sequencing of the parental specimens did not detect this deletion, and SNV-based identity testing was consistent with both parental samples being from the biological parents of the proband, indicating a de novo variant.

RNA sequencing of peripheral blood mononuclear cells from the proband demonstrates that the observed genomic deletion has an effect at the RNA level. The overall *PRKAR1A* expression level in the proband is significantly lower than in equivalently processed controls (Figure S1A). When relative expression is examined at the exon level, exon 2, which is deleted in the genomic DNA, demonstrates the largest observed reduction, but 10 of 11 exons demonstrate a trend toward reduced expression (Figure S1B). Splicing analysis identifies an isoform that skips exon 2 in the proband that is not detected in any of the controls (Figure S1C). Overall, this splice isoform in the proband that skips exon 2 is observed at an approximately 4-fold lower level than the canonical isoform. The genomic DNA encoding the transcribed exons of *PRKAR1A* did not contain any heterozygous sites, so we were unable to analyze allele specific expression.

Using the ACMG Standards and Guidelines for the interpretation of sequence variants[15], this variant was categorized as pathogenic based on: 1) identification of a null variant in a gene where loss of function is a known mechanism of disease (PVS1), and 2) de novo variant in a patient with disease and no family history (where both maternity and paternity confirmed, PS2).

It is difficult to call structural variants in SRS data with simultaneously high sensitivity and specificity that is necessary for clinical laboratory testing. Nevertheless, once a small candidate gene list or approximate breakpoints are known, many variants can be identified retrospectively.[5] In such cases, SRS often provides exact breakpoints to refine the variant discovered by LRS.[16] Manual inspection of SRS data from the *PRKAR1A* locus shows support for the heterozygous deletion through a drop in read depth and alignment clipping at the deletion breakpoints (Figure 2C). Multiple short-read structural variant callers, including Pindel, Lumpy, BreakDancer, Manta, CNVKit, and CNVnator, were retrospectively used to identify structural variants.[9–14] Pindel, Lumpy, BreakDancer, and Manta all identify a deletion in the locus. Pindel and Manta approximate the breakpoints identified from LRS and Sanger sequencing. Comparisons of the variant filtering results and overlap for LRS and SRS for Pindel and Manta are provided in Tables S1 and S2.

## DISCUSSION

Carney complex is a rare, autosomal dominant disease diagnosed by clinical criteria, including pigmented skin abnormalities, myxomas, endocrine tumors and dysfunction, and schwannomas.[17] Two or more major diagnostic criteria are required for a definitive diagnosis of Carney complex[18], and this patient meets three: 1) cardiac myxomas, 2) large-cell calcifying Sertoli cell tumor, and 3) acromegaly as a result of a growth hormone-producing pituitary adenoma (all histologically confirmed). Additional signs suggestive of Carney complex include skin findings (a few lentigines and multiple blue nevi) and multiple thyroid nodules detected by ultrasound in an individual older than 18 years. Genome and RNA sequencing identified a de novo pathogenic deletion in *PRKAR1A*, providing molecular confirmation of the diagnosis.

This case demonstrates the ability of genome LRS to detect causal structural variation in a rare disease, and to our knowledge, this is the first reported application of genome LRS to identify a pathogenic variant in a patient. Although manual inspection of the aligned read data and short-read structural variant callers are able to identify this 2,184 bp deletion, these approaches are not practical to apply genome wide due to limited throughput and high false-positive call rates, respectively. Looking forward, clinical-grade genomics ideally would provide strong precision and recall across the full spectrum of genetic variation.

SRS has decreased sensitivity for insertion and deletion variant detection as the size of the event increases, and it can miss up to 80% of the structural variants in an individual genome.[3] Current cytogenomic arrays have a maximum resolution >5–10 kb.[19] This leaves an opportunity for a technology that can detect insertions and deletions too large for SRS, e.g., >50 bp, and too small for cytogenomic arrays. LRS appears to be capable of identifying much of the missed variation, and manifests high recall of structural variants even at low

depths of coverage.[16] This initial proof-of-concept case demonstrates that this variation can be clinically relevant. We suggest that larger studies on the molecular diagnostic yield of LRS will be required to fully evaluate the relative performance of LRS versus SRS for the identification of intermediate size insertions and deletions and to determine the ultimate clinical utility of this approach. Likewise, cost reductions in LRS technologies will be required prior to any clinical implementation.

In the current manuscript, we describe the use of long-read genome sequencing to identify a ~2.2 kb deletion in *PRKAR1A* in a patient with Carney complex, providing a molecular explanation for disease. This first successful application of genome long-read sequencing to identify a pathogenic variant in a patient, when considered in the context of prior studies, suggests that LRS may be one approach to identify disease-causing structural variants that are difficult to detect with current technologies.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

The authors thank the research subject and clinical care teams for their participation in this research study; Chen-Shan (Jason) Chin for helpful discussions; and Primo Baybayan and Matt Boitano for PacBio library preparation and sequencing.

## References

1. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016; 17(9):507–522. [PubMed: 27528417]
2. Goldfeder RL, Priest JR, Zook JM, et al. Medical implications of technical accuracy in genome sequencing. Genome Med. 2016; 8(1):24. [PubMed: 26932475]
3. Chaisson MJ, Huddleston J, Dennis MY, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015; 517(7536):608–611. [PubMed: 25383537]
4. Huddleston J, Chaisson MJ, Meltz Steinberg K, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2016; Advance online publication. doi: 10.1101/gr.214007.116
5. Chaisson MJ, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of human genomes. Nat Rev Genet. 2015; 16(11):627–640. [PubMed: 26442640]
6. Chin CS, Alexander DH, Marks P, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013; 10(6):563–569. [PubMed: 23644548]
7. Seo JS, Rhie A, Kim J, et al. De novo assembly and phasing of a Korean human genome. Nature. 2016; 538(7624):243–247. [PubMed: 27706134]
8. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43:11.10.11–33. [PubMed: 25431634]
9. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25(21):2865–2871. [PubMed: 19561018]
10. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. Genome Biol. 2014; 15(6):R84. [PubMed: 24970577]
11. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009; 6(9):677–681. [PubMed: 19668202]

12. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016; 32(8):1220–1222. [PubMed: 26647377]
13. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016; 12(4):e1004873. [PubMed: 27100738]
14. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011; 21(6):974–984. [PubMed: 21324876]
15. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015; 17:405–424. [PubMed: 25741868]
16. English AC, Salerno WJ, Hampton OA, et al. Assessing structural variation in a personal genome-towards a human reference diploid genome. BMC Genomics. 2015; 16:286. [PubMed: 25886820]
17. Stratakis, CA., Salpea, P., Raygada, M. Carney Complex. In: Pagon, RA.Adam, MP.Ardinger, HH., et al., editors. Gene Reviews. Seattle: University of Washington; 1993–2017.
18. Mateus C, Palangie A, Franck N, et al. Heterogeneity of skin manifestations in patients with Carney complex. J Am Acad Dermatol. 2008; 59(5):801–810. [PubMed: 18804312]
19. Le Scouarnec S, Gribble SM. Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics. Heredity. 2012; 108(1):75–85. [PubMed: 22086080]

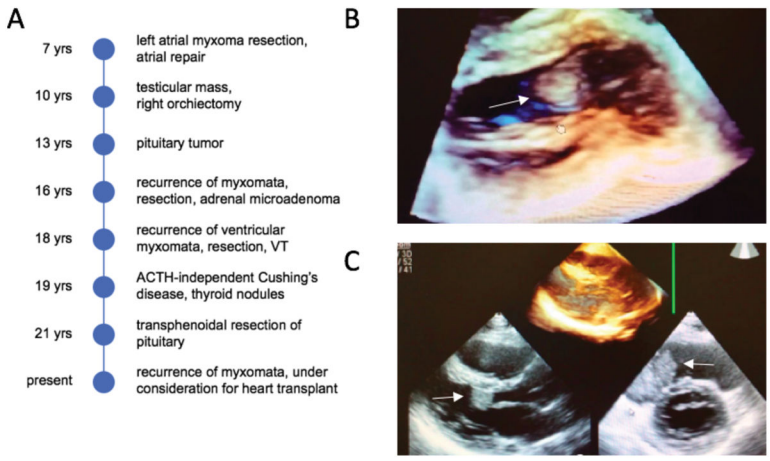


**Figure 1. Clinical history and three-dimensional transthoracic echocardiography of patient with multiple neoplasia including cardiac myxomata**

(A) Patient narrative. VT = ventricular tachycardia (B) A 2 × 3 cm myxoma is seen in the left ventricular outflow tract (white arrow). (C) The 2 × 3 cm myxoma is seen from another perspective (lower left, white arrow). A 5 × 4 cm myxoma is seen in the right atrium (lower right, white arrow).

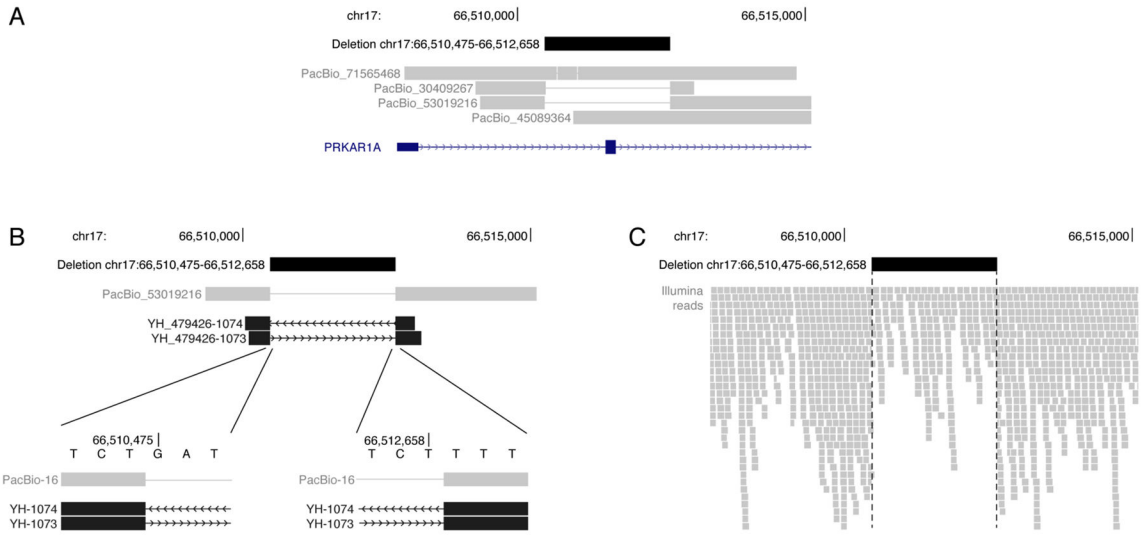


**Figure 2. Heterozygous deletion in *PRKAR1A***

(A) PacBio long reads identify a heterozygous 2,184 bp deletion that includes the first coding exon of *PRKAR1A*. Two of four reads at the locus support the deletion. (B) Sanger sequencing confirms the deletion. The forward (YH_479426-1073) and reverse (YH_479426-1074) sequences from a representative amplicon agree to the base pair with the higher quality PacBio read, PacBio_53019216. (C) Illumina short reads support the heterozygous deletion variant through a drop in read coverage and clipped reads at the deletion breakpoints.