Published in final edited form as:
Genet Med. 2011 March ; 13(3): 218–229. doi:10.1097/GIM.0b013e318203cff2.

Next generation massively parallel sequencing of targeted exomes to identify genetic mutations in primary ciliary dyskinesia: implications for application to clinical testing

Jonathan S. Berg, MD, PhD[1,2], **James P. Evans, MD, PhD**[1,2], **Margaret W. Leigh, MD**[3], **Heymut Omran, MD**[4], **Chris Bizon, PhD**[5], **Ketan Mane, PhD**[5], **Michael R. Knowles, MD**[2], **Karen E. Weck, MD**[1,6], and **Maimoona A. Zariwala, PhD**[6]

[1]Department of Genetics, UNC Chapel Hill

[2]Department of Medicine, UNC Chapel Hill

[3]Department of Pediatrics, UNC Chapel Hill

[4]Department of Paediatrics and Adolescent Medicine, University Hospital, Freiburg, Germany

[5]The Renaissance Computing Institute, UNC Chapel Hill

[6]Department of Pathology and Laboratory Medicine, UNC Chapel Hill

Abstract

PURPOSE—Advances in genetic sequencing technology have the potential to enhance testing for genes associated with genetically heterogeneous clinical syndromes, such as primary ciliary dyskinesia (PCD). The objective of this study was to investigate the performance characteristics of exon-capture technology coupled with massively parallel sequencing for clinical diagnostic evaluation.

METHODS—We performed a pilot study of four individuals with a variety of previously identified PCD mutations. We designed a custom array (NimbleGen) to capture 2089 exons from 79 genes associated with PCD or ciliary function and sequenced the enriched material using the GS FLX Titanium (Roche 454) platform. Bioinformatics analysis was performed in a blinded fashion in an attempt to detect the previously identified mutations and validate the process.

RESULTS—Three of three substitution mutations and one of three small insertion/deletion mutations were readily identified using this methodology. One small insertion mutation was clearly observed after adjusting the bioinformatics handling of previously described SNPs. This process failed to detect two known mutations: one single nucleotide insertion and a whole exon deletion. Additional retrospective bioinformatics analysis revealed strong sequence-based evidence for the insertion but failed to detect the whole exon deletion. Numerous other variants were also detected, which may represent potential genetic modifiers of the PCD phenotype.

CONCLUSIONS—We conclude that massively parallel sequencing has considerable potential for both research and clinical diagnostics, but further development is required before widespread adoption in a clinical setting.

Correspondence: Jonathan S. Berg MD/PhD, Assistant Professor, Department of Genetics, The University of North Carolina at Chapel Hill, Room 5092 Genetics Medicine Building, CB# 7264, Chapel Hill, NC 27599-7264, jsberg@med.unc.edu, 919-966-7043 (phone)/ 919-843-0291 (fax).

Keywords

Next-generation sequencing; exon-capture; molecular diagnostic testing; primary ciliary dyskinesia; clinical genetics

INTRODUCTION

Primary ciliary dyskinesia (PCD) is an autosomal recessive disorder involving abnormalities of motile cilia, resulting in a range of manifestations including situs inversus, neonatal respiratory distress at full term birth, recurrent otitis media, chronic sinusitis, chronic bronchitis that may result in bronchiectasis, and male infertility.[1,2,3] The disorder is genetically heterogeneous, rendering molecular diagnosis challenging given that mutations in nine different genes (*DNAH5*, *DNAH11*, *DNAI1*, *DNAI2*, *KTU*, *LRRC50, RSPH9*, *RSPH4A* and *TXNDC3*) account for only ~1/3 of PCD cases.[4] *DNAH5* and *DNAI1* account for the majority of known mutations, and the other genes each account for a small number of the remaining cases.[5,6] Electron microscopy (EM) can reveal the presence of defective dynein arms or other axonemal components, and immunohistochemistry can suggest the loss of specific proteins,[4] but in most cases it is impossible to distinguish between patients with different genetic etiologies. Thus, from a diagnostic standpoint it would be advantageous to engage in multiplex testing of multiple genes for causative mutations. Such an approach could also be readily adapted for gene discovery since the known involvement of ciliary genes in PCD suggests numerous "candidate genes" which are likely to play a role in cases of PCD without identifiable mutations.

Recent advancements in massively parallel sequencing (so-called "next-generation sequencing") are revolutionizing genetic research[7–12] and demonstrate potential in clinical diagnostics.[8, 10–11, 13–17] Because of the genetic heterogeneity in PCD, we investigated the performance characteristics of NimbleGen targeted exon capture followed by massively parallel sequencing using Roche 454 technology[18] for detection of genetic variants in known and candidate PCD genes. We envision this technology as a hybrid platform capable of being used in the clinical diagnostic setting as well as in the research setting for those patients with no mutations in known PCD genes.

In this pilot study, we analyzed four PCD patients in whom disease-causing mutations (three substitution mutations, three small insertion/deletion mutations, and one whole exon deletion) were known. One of the patients had only one mutation identified previously, and we hoped to detect a second deleterious mutation. We expected a high detection rate for nonsense, small insertion/deletion, and missense mutations that were previously identified in the patient samples, but we recognized that this approach might fail to detect whole exon deletions. We anticipated the discovery of "variants of uncertain significance" and "false positive" results, and we were interested in exploring the reasons for "false negative" results. The systematic analysis and troubleshooting of such results is a necessary prerequisite to the implementation of robust genomic analysis in the clinical arena.

MATERIALS AND METHODS

Patients and family members

The patients are enrolled in a study of clinical and molecular aspects of PCD[1] approved by the institutional review board at the University of North Carolina at Chapel Hill (study # 05-2979). DNA was prepared by salt extraction from lymphoblastoid cell lines (patient 475 and patient 1205) or blood samples (patient 998 and patient 1072). DNA quality was measured by UV spectrometry (A260/280 ratios between 1.84 and 1.89 for all samples) and

gel electrophoresis. The sample from patient 475 had a small amount of degradation but was deemed acceptable for use. None of the patients came from consanguineous families.

Exon capture design, enrichment, and sequencing

We designed a custom oligonucleotide microarray (NimbleGen) to capture 2089 exons of 79 genes known to be associated with PCD or candidate genes based on function/expression in cilia/flagella of human or model organisms (Supplemental Digital Content 1 and 2). The final targeted region included 510,558 base pairs (95.6%) with an offset of 0 bases, or 520,838 base pairs (97.5%) with an offset of 100 bases. DNA samples from each of the patients were sent to Roche and subjected to capture and 454 sequencing according to standard operating procedures. Roche performed mapping, alignment, and variant detection against the human NCBI Build 36 reference genome using the GS Reference Mapper software package.[19] Samples from patients 475 and 998 underwent an additional round of sequencing due to concerns about read length in the initial sequencing runs.

Variant analysis

One investigator, blinded to the identity of the previously determined disease-causing mutations, analyzed all variants within the targeted region for the four patients. The *HYDIN* gene on chromosome 16q22 was excluded from our analysis due to the existence of a paralog mapping to chromosome 1q21 that is not accounted for in the NCBI 36 reference sequence. Variants identified by GS Reference Mapper are annotated with genomic coordinates, reference nucleotide, variant nucleotide, number of reads containing the variant nucleotide, and percentage of reads containing the variant nucleotide. Additional annotations (if applicable) include the reference amino acid, variant amino acid, reading frame, gene name, and overlap with a known SNP from dbSNP 130. For identification of possible splice site mutations, we calculated the distance of each variant to the nearest intron-exon junction and then selected variants occurring within the last two nucleotides of an exon and first five nucleotides of the donor intron, or within the last 5 nucleotides of the intron and the first nucleotide of the acceptor exon. Further characterization of missense mutations was performed on selected variants using evolutionary conservation and prediction of pathogenicity with PolyPhen.[20,21] PCR amplification and follow-up Sanger sequencing of newly identified candidate mutations was performed.

Statistical analysis

Tables of variants were organized using Microsoft Excel. Statistical analysis and graphing was performed using GraphPad Prism (version 5.0b).

RESULTS

Previous characterization of disease-causing mutations in PCD patients

Our group has performed extensive analysis of selected genes in PCD patients and family members. For the current study, four previously characterized patients were selected based on qualitatively different types of mutations in three PCD genes (Summarized in Table 1).

- Patient 475: Sanger sequencing of *DNAH11*, *WDR63*, *RSHL1*, *RSPH3*, *RSPH4A*, *DNAH5*, and selected exons of *DNAI1* revealed compound heterozygous mutations in *DNAH5*: a missense mutation in the last base of an exon leading to a splicing defect, and a 21 bp deletion causing an in-frame deletion of 7 amino acid residues.

- Patient 998: Sanger sequencing of the entire coding regions of *DNAH11, DNAH5* and *DNAI1* identified only one mutation in *DNAH11*, a 4 bp deletion leading to a frameshift.
- Patient 1072: Sanger sequencing of *DNAH5* and *DNAI1* revealed compound heterozygous nonsense mutations in *DNAI1* resulting in premature truncation.
- Patient 1205: Analysis of *DNAH11* and selected exons of *DNAH5* revealed compound heterozygous mutations in *DNAH5*: one single base insertion leading to a frameshift detected by Sanger sequencing and a deletion of exon 62 detected by multiplex ligation-dependent probe amplification (MLPA).

Exon Capture, High-throughput Sequencing, Mapping, and Variant Identification

Patient DNA was enriched for 2089 exons from 79 known or predicted PCD genes, using a custom NimbleGen oligonucleotide capture array. Roche GS FLX Titanium massively parallel sequencing of the enriched material yielded 760,738 – 1,174,723 high quality reads encompassing 249,757,548 – 363,712,243 high quality bases per patient (see Tables, Supplemental Digital Content 3 and 4). After mapping, >99.8% of the targeted region was covered in each patient and >97.7% of the targeted region was covered at >10-fold depth. Weighted unique mean depth of coverage for the targeted region was 39x–89.2x. Thus, the coverage should have been adequate to reliably detect germline variants within the vast majority of targeted regions. There was an average of 538 variants per patient within the 510,558 bp targeted exome region (~1 variant per kb of targeted genome). Roche considers variants observed in 15–85% of unique reads to be heterozygous, and the GS Reference Mapper algorithm matches variants to dbSNP version 130. Using these criteria, there were 32–53 novel variants detected per patient sample (Table 2).

Identification of Candidate Disease-Causing Mutations

Since a platform such as this one could eventually be employed in a clinical diagnostic setting, we attempted to develop an analytic process compatible with the workflow of a typical molecular diagnostic laboratory, in which variants meeting the following criteria would be considered potentially causative: (i) detected in >15% of reads, (ii) not overlapping with an annotated SNP, (iii) predicted to be truncating (premature stop/frameshift/splice-site disruption) or altering an amino acid, and (iv) presence of two mutations in one gene, consistent with autosomal recessive inheritance (summarized in Table 2). While such a scheme would miss rare mutations in regulatory regions or splice enhancer sequences, it would be expected to identify the majority of clinically relevant mutations. Variants identified in the initial workflow are further detailed in Table 3.

Patient 475—The initial blinded approach identified both of the previously characterized disease-causing mutations in *DNAH5*. The heterozygous 21-bp deletion in the *DNAH5* gene was present in 57% of the reads. The second mutation in *DNAH5* was observed in 52% of sequence reads and consisted of a substitution that changes a methionine residue to leucine and involves the last nucleotide of the exon, thus disrupting a canonical splice donor sequence.

This individual had three other genes (*RSPH4A*, *DNAH10*, and *LRRC50*) each with two novel variants. Both *RSPH4A* and *LRRC50* have been associated with PCD,[22,23] while *DNAH10* is an inner arm dynein heavy chain[24] but has not been extensively analyzed in PCD patients. The two *RSPH4A* variants were both single base insertions that would cause a frameshift, but were observed in a low percentage of reads. Interestingly, previous analysis of the *RSPH4A* gene by Sanger sequencing had identified three common polymorphisms in this patient but no evidence for the putative novel insertions, indicating that they are likely

false positive results of 454 sequencing. One of the variants in *DNAH10* changed a highly conserved arginine to tryptophan. The other variant altered an evolutionarily non-conserved histidine to tyrosine and was also observed in one other patient, suggesting that it is likely benign. The two variants in *LRRC50* were both observed in a low percentage of reads and were located in a short region of low complexity sequence (CCCCACCACCCCCGCCACC) within a simple tandem repeat, suggesting that they are sequencing artifacts.

Heterozygous variants of uncertain significance were identified in nine other genes (*DNAH14, DNAH6, SPEF2, DNAH11, DYNC2H1, DYNC1H1, CCDC114, RSPH1*, and *RPGR*). One of these (*DNAH6*) was an insertion detected in 16% of sequence reads that was not detected on subsequent Sanger sequencing.

Patient 998—This patient was the only one in whom only a single disease-causing mutation had been previously identified – a 4-bp deletion in the *DNAH11* gene associated with PCD and normal axonemal ultrastructure.[25] It was therefore hoped that two deleterious mutations would be identified in *DNAH11*, or in another known or candidate PCD gene.

The initial blinded analysis failed to identify the known 4-bp deletion mutation in *DNAH11* previously detected by Sanger sequencing. The only gene identified as having possible heterozygous variants was *DNAH9*, a candidate gene for PCD.[26] One *DNAH9* variant was a single base insertion observed in 18% of sequence reads, but it was not detected on Sanger sequencing. The other *DNAH9* variant changes an arginine residue to glutamine, but glutamine is observed at this site in other species, suggesting that the variant is benign. Together, these results argue strongly against *DNAH9* as the cause of PCD in this individual.

Two dynein heavy chain genes (*DNAH3* and *DNAH6*) demonstrated novel homozygous variants. The *DNAH3* variant changes an arginine residue to glutamine, but several species have a glutamine at this position indicating that it is unlikely to be deleterious. The *DNAH6* variant changes valine to alanine, but is again poorly conserved; in addition, this variant was also detected in the homozygous state in patient 1205. Therefore, both of these homozygous variants most likely represent benign SNPs.

Since some known disease-causing mutations are present in the dbSNP database,[27] we reconsidered all coding variants in this individual, regardless of whether they were annotated as overlapping with a previously described SNP. In doing so, we recognized the previously detected 4-bp frameshift deletion in *DNAH11*, clearly present in 53% of the reads. This variant had been annotated as overlapping with SNP rs72657376 and was therefore eliminated from consideration in the initial analytic scheme. This false negative represents a bioinformatics failure, since the use of dbSNP data as an analytic filter caused us initially to miss this mutation. Further analysis of *DNAH11* revealed several other known SNPs but no nonsense or likely splice site mutations. Thus, after adjusting the analytic scheme (albeit in a *post hoc* fashion) we were ultimately able to identify the 4-bp deletion but no other likely disease-causing mutations in *DNAH11*.

Four additional genes (*CCDC63*, *DNAH2*, *DNAH10*, and *RSPH1*) were found to contain novel heterozygous missense variants. In *DNAH10* and *RSPH1* the affected amino acids were poorly conserved among other species and the variants were both present in patient 475, suggesting that they are likely to be benign. The variant in the *CCDC63* gene also affects a poorly conserved amino acid. However, we were intrigued by a novel heterozygous missense variant in *DNAH2* (chr17:7677205 T>A), altering a highly conserved valine residue to aspartic acid predicted by PolyPhen to be possibly damaging. This variant was not observed in any of the three other patients. We therefore considered whether the patient's

phenotype could be due to digenic inheritance in this case (Figure 1). Testing of family members revealed that the *DNAH11* mutation was inherited from the patient's father, and the *DNAH2* variant was inherited from the patient's mother. Among the three unaffected siblings, the sister carries only the *DNAH11* mutation, a brother carries only the *DNAH2* variant, and another brother is wild-type at both loci. The familial segregation is consistent with the possibility of digenic inheritance, although further studies will be required to establish this with certainty.

Patient 1072—The initial blinded analysis revealed two heterozygous nonsense mutations in the *DNAI1* gene, which has been previously associated with PCD.[28] These were the precise disease-causing mutations that had previously been identified by Sanger sequencing.

In addition, two novel heterozygous variants in the dynein heavy chain gene *DNAH8* were identified. One of the *DNAH8* variants predicts a change of a histidine residue to arginine, which is observed at this position in numerous other species and was predicted by PolyPhen to be benign. The other *DNAH8* variant changes a threonine residue to methionine and was predicted by PolyPhen to be probably damaging. Both variants were confirmed by Sanger sequencing but were found to have been inherited in *cis* from the mother, whereas the father was wild-type at both positions. Thus, only one allele of *DNAH8* is affected by these changes. This individual also had heterozygous missense mutations of highly conserved or invariant amino acids in several other dynein heavy chain genes (*DNAH2*, *DNAH7*, *DNAH9*) and Tektin-2 (*TEKT2*), which encodes a protein that is coassembled with ciliary and flagellar microtubules and may play a role in asthenozoospermia.[29]

Patient 1205—Neither deleterious mutation previously identified in this patient (a large exonic deletion and a single nucleotide insertion in *DNAH5*) was detected by massively parallel sequencing in our established analytic scheme.

The initial blinded analysis identified only one gene, *SPAG17*, as having two candidate variants. The first of these variants changes a methionine to isoleucine and was predicted by PolyPhen to be benign, while the second variant changes a threonine residue to asparagine and was predicted by PolyPhen to be damaging. The *SPAG17* gene encodes an orthologue of *Chlamydomonas* PF6, a component of the flagellar central apparatus.[30] Mutations of the *Chlamydomonas* ortholog cause flagellar paralysis, but phenotypes have not been reported in humans or mice. Sanger sequencing confirmed the presence of both variants, but their clinical significance is uncertain.

This individual had eleven other genes with a single heterozygous novel variant, including four dynein heavy chain genes (*DNAH2, DNAH3, DNAH8,* and *DNAH10*) and a gene associated with intraflagellar transport (*IFT172*). In addition, the *DNAH6* variant that was detected in patient 998 was observed in 100% of the reads in this individual. Of the eleven novel heterozygous variants, five were single base insertions that would be predicted to cause a frameshift, but all five were present in <25% of the sequence reads. We selected four of these variants (*IFT172*, *SPEF2*, *DNAH8*, and *LRRC50*) for confirmation by Sanger sequencing and none of them were confirmed. The clinical significance of the remaining heterozygous missense variants remains uncertain at this time.

Having failed to identify likely disease-causing mutations in this individual, we unblinded the analysis and attempted to find the two mutations in *DNAH5* previously discovered through Sanger sequencing and MLPA. Given that exon capture is non-quantitative, we did not necessarily expect to find the exon deletion using this platform, and indeed, upon examining sequence-based data and depth of coverage across all exons, we could not detect any evidence of the exon deletion (see Text and Figure, Supplemental Digital Content 6).

However, we were surprised that the insertion mutation in *DNAH5* (c.13458_59 insT) was not identified, even among low-quality variants that failed to meet the "high confidence" standards. This mutation has been observed in multiple families and may represent a founder mutation.[5] The *DNAH5* cDNA is transcribed from the minus strand, and the c.13458_59 insT mutation localizes to a stretch of 7 adenine (7A) nucleotides on the plus strand of the reference genomic sequence (chr5:13,754,426-13,754,432). Thus, the c.13458_59 insT mutation should result in 8 adenines (8A) at this position.

Based on previous reports regarding the deficiency of 454 sequencing in regions containing short homopolymers,[31] we hypothesized that the number of adenine residues in different reads at this position should follow a normal distribution around the "true" number. We therefore made local alignments of sequence reads (Figure 2A) mapping to this location in all four patients, tabulated the number of adenine nucleotides in each read, and compared the proportion of reads containing 5, 6, 7, 8, 9, or 10 reads between each patient (Figure 2B). The sequence data for patients 475, 998, and 1072 were consistent with a homozygous 7A genotype, while the sequence data for patient 1205 suggested a heterozygous 7A/8A genotype (Figure 2C), consistent with the prior Sanger sequencing results. Thus, the massively parallel sequencing data supports the presence of a mutant 8A allele at this position when analyzed in a *post hoc* manner, but the GS Reference Mapper variant detection algorithm did not identify it prospectively.

Raca and colleagues reported a similar problem with 454 sequencing in a pilot study of massively parallel sequencing for ocular birth defects.[32] These authors reported an initial failure, also using the GS Reference Mapper software, to detect a single base deletion within a stretch of 7 guanine nucleotides in the *PAX2* gene. However, they were able to demonstrate the presence of the deletion using two different software packages (CLC Genomic Workbench and NextGENe). We likewise utilized NextGENe to independently assemble the raw sequence reads and call variants, and indeed the heterozygous insertion of adenine was recognized at this location (data not shown). However, this increased sensitivity came at the cost of greatly increasing the number of putative single nucleotide insertions and deletions. For instance, using default settings, NextGENe called 5 or more deletion variants in the coding regions of several genes (observed in >25% of reads and in some cases ~50% of reads) that were not called by 454's suite and which occurred in the immediate context of homopolymer stretches, strongly suggesting that they are false positives.

Analytical sensitivity and specificity of exon capture followed by massively parallel sequencing

An important potential limitation of this approach is that it may be prone to both false positive and false negative results. This was clearly the case with the four patient samples we analyzed. In the course of evaluating novel variants, we used Sanger sequencing to follow up 16 variants identified by 454 sequencing. All 8 of the putative novel single base insertion variants, observed in 15–24% of sequence reads, were determined to be false positives of 454 sequencing based on the negative results of confirmatory Sanger sequencing. In contrast, Sanger sequencing confirmed all 8 novel missense variants, observed in 41–56% of sequence reads.

As we inspected the next-generation sequencing data as a whole, it became clear that a large number of putative novel small insertion/deletion variants were observed in fewer than 25% of the 454 sequence reads and that this might be a useful signal for flagging possible false positive results. To further explore this phenomenon, we analyzed the percentage of sequence reads in which different types of variants were identified (Figure 3). We noted that insertions and deletions were likely to be detected in a lower percentage of reads than single base substitutions, which clustered near 50% and 100% of reads, as would be expected for

heterozygous or homozygous variants, respectively. Putative “novel” (not annotated in dbSNP) insertion/deletion variants were much more frequently observed in fewer than 25% of sequence reads compared to “known” (previously annotated in dbSNP) insertion/deletion variants or “novel” substitution variants (Figure 3A). Although there were no dramatic differences between the types of variants in terms of total number of sequence reads (Figure 3B), higher read depth was associated with tighter clustering of “known” dbSNP variants near 50% or 100% of the reads (Figures 3C and 3D), demonstrating the greater precision that can be achieved with higher depth of coverage. Similarly, “novel” substitution variants tended to cluster near 50% or 100% of the reads, with the exception of four variants observed in fewer than 15% of reads that were thus called homozygous wild-type (Figure 3E). In contrast, the putative “novel” insertion/deletion variants displayed a completely different pattern in which there was no direct relationship between the read depth and the percentage of reads containing the variant (Figure 3F). As noted above, Sanger sequencing did not confirm any of the 8 putative “novel” insertion/deletion variants detected by 454 sequencing. Thus, the accuracy of 454 sequencing for detection of small insertion/deletion variants is not ideal, especially for “novel” insertion/deletion variants identified in fewer than 25% of sequence reads.

Since the four patient samples were previously subjected to extensive Sanger sequencing of PCD-associated genes (*DNAH5*, *DNAH11*, *DNAI1*, *WDR63*, *RSPH3/RSHL2*, *RSPH4A/RSHL3*, *RSPH6A*/*RSHL1*), many benign sequence variants were known to be present in addition to the presumptive disease-causing mutations (see Table, Supplemental Digital Content 7). We used this additional information to analyze the sensitivity of our targeted exome massively parallel sequencing approach for detection of sequence variants. Of the 119 benign variants that had previously been found within the exon-targeted region, 114 were also identified with the expected zygosity using massively parallel sequencing. Four of the discrepancies (three single base substitutions and one 2-bp deletion) were found to be homozygous wild-type when reanalyzed by Sanger sequencing and thus were determined to be false positive calls by the previous automated Sanger sequencing. The remaining variant was a complex alteration in the region of an oligo-dT tract near the end of an intron of *WDR63* (hg18; chr1:85,346,294-85,346,301). Prior Sanger sequencing identified one allele with a single C insertion and another allele with a CT insertion, immediately preceding the oligo-dT tract. Massively parallel sequencing correctly detected the homozygous C insertion (rs11427716) but not the additional heterozygous T insertion. This result further demonstrates the difficulties posed by homopolymer tracts for 454 sequencing.

DISCUSSION

Targeted exome massively parallel sequencing has great potential for both research and clinical use. Based on work by Ng and colleagues with whole exome sequencing, each individual likely has 6–12 genes with two predicted damaging mutations that are not present in control populations (dbSNP, HapMap).[9] In a clinically distinctive autosomal recessive disorder with only one disease locus, discovery of the disease gene by whole exome sequencing could require as few as three or four unrelated individuals.[7] However, due to the genetic heterogeneity of PCD (and many other genetic disorders), a much larger number of individuals might need to be analyzed in order to identify genes in common among subsets of such patients. We are currently engaged in whole-exome sequencing of additional PCD patients whose mutations remain unidentified (M.A.Z. and M.R.K.). However, while the research implications of next-generation sequencing are clear, use of this emerging technology for clinical diagnostic testing still requires careful calibration.

Detection of known mutations associated with PCD

In this work we conducted a pilot study of four patients with clinical features consistent with PCD and with previously identified mutations in genes associated with PCD. We specifically chose samples with qualitatively different types of mutations occurring in different genes in order to determine whether targeted exome capture combined with next-generation sequencing could identify different classes of mutations. The exome capture of 79 known or candidate PCD genes, followed by massively parallel sequencing using 454 technology and structured analysis, identified five of the seven known mutations associated with PCD in three of the four patients, including one patient (#475) with two previously identified mutations in the dynein heavy chain gene *DNAH5* (a nucleotide substitution resulting in a splice site alteration and a 21 bp deletion) and another patient (#1072) with two different single nucleotide nonsense mutations in the dynein intermediate chain gene *DNAI1*. In the third patient (#998), previous analysis had identified only one disease-causing mutation in *DNAH11* (a 4-bp frameshift deletion). This mutation was incorrectly excluded by the initial analytic scheme due to its annotation as a known SNP, but was correctly identified after a simple adjustment of the handling of annotated SNPs. In the fourth patient (#1205), the single base frameshift insertion could only be demonstrated retrospectively with analysis of sequence reads, but the exonic deletion was not detected.

Variants of Uncertain Clinical Significance

A number of other heterozygous and homozygous novel variants were identified in the four patients. Although these variants are of unknown clinical significance, it is tempting to speculate that some could play a role in the PCD phenotype, acting as genetic modifiers. Further massively parallel sequencing of a large number of PCD patients with known deleterious mutations could lend further insight into the roles of putative disease-modifying variants.

The patient in whom a single disease-causing mutation in *DNAH11* had previously been identified (#988) was found in this study to have a missense variant in *DNAH2*. Mutation profiling of *DNAH2* has not been carried out in PCD patients, but mutations of the *Chlamydomonas* ortholog (Dhc10) alter flagellar motility.[33] The *DNAH2* variant affected a highly conserved amino acid residue that was predicted to be damaging, and family segregation analysis was consistent with the possibility of digenic inheritance. Although this finding does not prove that digenic variants of *DNAH11* and *DNAH2* can result in pathogenicity, digenic inheritance has been previously hypothesized in PCD[34] and there is precedent for triallelic inheritance in a different ciliopathy, Bardet-Biedel syndrome.[35] It is worth noting that *DNAH11* encodes an outer dynein arm protein while *DNAH2* encodes an inner dynein arm protein; since both DNAH11 and DNAH2 proteins are components of the ciliary dynein arms, which require complex protein interaction for motility, it is thus possible that haploinsufficiency of these two different genes could result in reduced ciliary function without a structural defect discernable by EM. We speculate that this patient's phenotype could be due to the combinatorial effects of mutations in *DNAH11* and *DNAH2*. However, it is also possible that there simply remains an undetected mutation in a regulatory element, promoter, or intron of *DNAH11* or *DNAH2*, or biallelic mutations in a gene not present among the 79 genes analyzed here.

Finally, it is worth noting in this context that massively parallel approaches such as whole-exome and whole-genome sequencing will result in the identification thousands of protein coding variants in each individual, any of which could potentially impact disease (either directly as a disease-causing mutation or as a genetic modifier), rendering the determination of clinical significance exceedingly complex with respect to various combinations of mutations. For example, if alternative inheritance patterns such as digenic inheritance are

considered, the number of different possibilities quickly becomes astronomical when variants are simultaneously detected in hundreds or thousands of genes. This will be a major challenge faced by molecular diagnosticians and clinicians in the future as multiplex gene sequencing is more broadly applied.

Performance and limitations of targeted exome massively parallel sequencing

In addition to the previously identified disease-causing mutations, we analyzed 119 benign variants previously identified in the four patients by automated Sanger sequencing. Of these, targeted exome massively parallel sequencing correctly identified 114 variants and revealed 4 of these putative variants to be false positive results of the Sanger sequencing. Thus, the final 454 and Sanger sequencing results were in agreement for 114 of 115 benign polymorphic variants and >99% of wild-type reference calls. Including the six previously identified small deleterious mutations, we calculate a sensitivity of 98.3% (119/121) for detection of sequence variants by exon capture followed by 454 sequencing. The vast majority (>90%) of the polymorphisms studied here were single nucleotide substitution variants, so this result primarily confirms that 454 sequencing has high sensitivity and specificity for the detection of substitution mutations.

All of the variants in which there was discordance between the final 454 and Sanger sequencing results were small insertions or deletions. This drawback is inconvenient but not devastating for researchers, who expect and tolerate a certain level of error, but it is of particular importance for clinical diagnostic testing. Small insertion and deletion mutations make up ~22% of the ~100,000 known mutations in the human gene mutation database (http://www.hgmd.cf.ac.uk/), and inability to reliably identify this class of mutations would be a major limitation of adopting 454 sequencing for clinical diagnostic testing.

In an empirical study using a library of reference templates, Huse and colleagues found that insertions and deletions were the most common errors observed with 454 pyrosequencing,[31] comprising 63% of all sequence errors. These calculations were based on individual reads, whereas variant detection is based on the percentage of mapped reads that contain the variant, so it is unclear how this error rate might translate directly to inaccurate detection of insertion/deletion variants. In our experience, the majority of putative novel (not previously reported in dbSNP) insertion/deletion variants were observed in fewer than 25% of the sequence reads and this phenomenon appeared to be independent of the depth of coverage. Sanger sequencing did not confirm any of the eight putative novel single nucleotide insertions that we analyzed. Restricting analysis only to variants detected in greater than 25% of sequence reads greatly improved the specificity of this approach without affecting sensitivity (data not shown).

One explanation for the detection of sequence variants in fewer than 25% of reads could be the existence of pseudogenes not currently annotated in the reference genome but nevertheless contributing to the material being sequenced (eg. *HYDIN*; see Text and Figure, Supplemental Digital Content 8). However, in this case one might expect recurrent substitution and insertion/deletion variants affecting the same few genes (those with previously undiscovered pseudogenes or paralogous copies), and for the same variants to be present in more than one individual. None of these were the case in our experience, suggesting that false positive insertion/deletion identification is largely inherent to 454 sequencing technology. Thus, many (perhaps most) of the novel insertion/deletion variants detected in <25% of 454 sequence reads appear to be artifacts of the biochemistry, bioinformatics, or both.

In our study, a single base insertion mutation in patient #1205 was only identified retrospectively by demonstrating that the sequence data was consistent with the presence of

the heterozygous insertion in comparison to the three other patients. This false negative result thus represents a combined platform and bioinformatics failure. Indeed, the 454 bioinformatics pipeline appears to employ proprietary models to compensate for the imprecision of pyrosequencing with regard to homopolymers, perhaps explaining why the true deleterious insertion mutation in patient #1205 was not called by GS Reference Mapper but was called by NextGENe. As noted above, the increased sensitivity of the NextGENe analysis came at the unacceptable cost of lower apparent specificity, since a much higher number of putative small insertion/deletion variants were detected in the vicinity of homopolymer stretches. Since all possibly deleterious variants will require confirmation with clinical Sanger sequencing, at least in the short term, follow-up sequencing could be very costly or impractical with a high false positive rate for novel insertion/deletion variants.

It was not entirely unexpected that the approach described here was unable to detect a whole exon deletion in the *DNAH5* gene in patient 1205, since targeted exon capture is not designed to detect large deletions. We attempted to uncover evidence for the whole exon deletion in structural variant alignments and analysis of sequence read depth, but were unable to find such confirmation. Variability in the depth of coverage between exons and between samples in our study appears to have been too high to reliably detect changes in exon copy number using non-quantitative exon capture methods. Interestingly, while this manuscript was under review, another group published a report of a similar approach to detection of inherited mutations for breast and ovarian cancer, in which deletions or duplications encompassing single or multiple exons were, in fact, accurately detected.[36] Possible reasons for Walsh and colleagues' success include differences in platforms or technical aspects of the exon capture process, a greater than 10-fold higher mean depth of coverage (~1200x), and a larger number of samples between which comparisons could be made.

Conclusions

Massively parallel sequencing has already proven successful in research settings and its adoption for clinical diagnostic testing appears to be inevitable, particularly as the costs per base continue to decrease. Nevertheless, clinical tests based on next-generation sequencing present special challenges not germane to research applications and will need careful validation to assess the performance characteristics of this new technology and to optimize the analytical sensitivity and specificity. Next-generation sequencing platforms may have inherent differences in detection of diverse types of variants,[37,38] so any clinical applications based on next-generation technologies will need to be validated on different types of known reference mutations to determine their performance characteristics. Analytic software for clinical diagnostic approaches will also need to be optimized for detection of different types of mutations. Reliance on dbSNP as a bioinformatics filter appears to be problematic, given that some rare deleterious mutations are contained among the many benign variants. We expect that as whole-exome and whole-genome sequencing are carried out in large numbers of patients and control subjects, an approach will evolve towards using information about the population frequency of a variant, rather than its presence or absence in a database, as a key filter for evaluating the likely clinical implications of a given variant. We conclude that exon-capture followed by massively parallel sequencing has considerable potential in clinical diagnostics, but deficiencies will need to be quantified and rectified before widespread adoption in a clinical setting.

SDC 1. Table containing genes targeted for exon capture (.pdf)

SDC 2. File containing genomic coordinates of targeted exons and exon capture probes used in this study (.bed file)

SDC 3. Table containing sequence read characteristics for each patient sample (.pdf)

SDC 4. Table containing summary of GS FLX Titanium sequencing and mapping (.pdf)

SDC 5. Text and Figure detailing attempts to detect a whole exon deletion (.pdf)

SDC 6. Table containing benign polymorphisms previously identified by Sanger sequencing (.pdf)

SDC 7. Text and Figure showing that variants detected in the *HYDIN* gene have a distinct pattern with respect to the percentage of sequence reads containing the variants (.pdf)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to the patients and their families for their participation in this study. We thank the German patient support group “Kartagener Syndrom und Primaere Ciliaere Dyskinesie e.V.”

We appreciate technical assistance from L. Huang, A. Cutting, A. Heer, J. Wallmeier, H. Olbrich and Dr. K. Chao.

- JSB is supported by the University Cancer Research Fund (UCRF; http://ucrf.unc.edu/).
- JPE is supported by the UNC Center for Genomics and Society (NHGRI 5-P50-HG004488-03) and a UNC Clinical Translational Science Award (1-UL1 -RR025747-01).
- MRK, MAZ, and MWL are supported by NIH research grant 5-U54-HL096458-06, funded by the office of the Director, and supported by ORDR and NHLBI, NIH.
- MRK and MAZ are supported by NIH grant 5-R01HL071798.
- This work was also supported in part by grant M0O1RR00046 and UL-1-RR025747 from the National Center of Research resources, NIH.
- HO is supported by a grant of the “Deutsche Forschungsgemeinschaft” (DFG Om 6/4 to H.O.).
- MLPA was supported by a Multidisciplinary Research Grant (MRG), North Carolina Biotechnology Center

References

1. Noone PG, Leigh MW, Sannuti A, et al. Primary ciliary dyskinesia: Diagnostic and phenotypic features. Am J Respir Crit Care Med. 2004; 169:459–467. [PubMed: 14656747]
2. Leigh MW, Pittman JE, Carson JL, et al. Clinical and genetic aspects of primary ciliary dyskinesia/Kartagener syndrome. Genet Med. 2009; 11:473–487. [PubMed: 19606528]
3. Kennedy MP, Omran H, Leigh MW, et al. Congenital Heart Disease and other Heterotaxic Defects in a Large Cohort of Patients with Primary Ciliary Dyskinesia. Circulation. 2007; 115:2814–2821. [PubMed: 17515466]
4. Zariwala, MA.; Knowles, MR.; Leigh, MW. Primary Ciliary Dyskinesia. In: Pagon, RA.; Bird, TC.; Dolan, CR.; Stephens, K., editors. GeneReviews [Internet]. Seattle (WA): University of Washington, Seattle; 1993–2007 Jan 24. [updated 2009 Oct 6]
5. Hornef N, Olbrich H, Horvath J, et al. *DNAH5* mutations are a common cause of primary ciliary dyskinesia with outer dynein arm defects. Am J Respir Crit Care Med. 2006; 174:120–126. [PubMed: 16627867]
6. Zariwala MA, Leigh MW, Ceppa F, et al. Mutations of *DNAI1* in primary ciliary dyskinesia: evidence of founder effect in a common mutation. Am J Respir Crit Care Med. 2006; 174:858–866. [PubMed: 16858015]
7. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–276. [PubMed: 19684571]

8. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009; 106:19096–19101. [PubMed: 19861545]
9. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010; 42:30–35. [PubMed: 19915526]
10. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med. 2010; 362:1181–1191. [PubMed: 20220177]
11. Sobreira NL, Cirulli ET, Avramopoulos D, et al. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet. 2010; 6:e1000991. [PubMed: 20577567]
12. Biesecker LG. Exome sequencing makes medical genomics a reality. Nat Gen. 2010; 42:13–14.
13. ten Bosch JR, Grody WW. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. J Mol Diagn. 2008; 10:484–492. [PubMed: 18832462]
14. Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet. 2009; 85:142–154. [PubMed: 19679224]
15. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009; 55:641–658. [PubMed: 19246620]
16. Vasta V, Ng SB, Turner EH, Shendure J, Hahn SH. Next generation sequence analysis for mitochondrial disorders. Genome Med. 2009; 1:100. [PubMed: 19852779]
17. Ashley EA, Butte AJ, Wheeler MT, et al. Clinical assessment incorporating a personal genome. Lancet. 2010; 375:1525–1535. [PubMed: 20435227]
18. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008; 452:872–876. [PubMed: 18421352]
19. Droege M, Hill B. The Genome Sequencer FLX^{TM} system – longer reads, more applications, straight forward bioinformatics and more complete data sets. J Biotechnol. 2008; 136:3–10. [PubMed: 18616967]
20. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet. 2001; 10:591–597. [PubMed: 11230178]
21. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002; 30:3894–3900. [PubMed: 12202775]
22. Castleman VH, Romio L, Chodhari R, et al. Mutations in radial spoke head protein genes RSPH9 and RSPH4A cause primary ciliary dyskinesia with central-microtubular-pair abnormalities. Am J Hum Genet. 2009; 84:197–209. [PubMed: 19200523]
23. Loges NT, Olbrich H, Becker-Heck A, et al. Deletions and point mutations of LRRC50 cause primary ciliary dyskinesia due to dynein arm defects. Am J Hum Genet. 2009; 85:883–889. [PubMed: 19944400]
24. Maiti AK, Mattéi MG, Jorissen M, Volz A, Zeigler A, Bouvagnet P. Identification, tissue specific expression, and chromosomal localisation of several human dynein heavy chain genes. Eur J Hum Genet. 2000; 8:923–932. [PubMed: 11175280]
25. Schwabe GC, Hoffmann K, Loges NT, et al. Primary ciliary dyskinesia associated with normal axoneme ultrastructure is caused by DNAH11 mutations. Hum Mutat. 2008; 29:289–298. [PubMed: 18022865]
26. Bartoloni L, Blouin JL, Maiti AK, et al. Axonemal beta heavy chain dynein DNAH9: cDNA sequence, genomic structure, and investigation of its role in primary ciliary dyskinesia. Genomics. 2001; 72:21–33. [PubMed: 11247663]
27. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311. [PubMed: 11125122]
28. Zariwala M, Noone PG, Sannuti A, et al. Germline mutations in an intermediate chain dynein cause primary ciliary dyskinesia. Am J Respir Cell Mol Biol. 2001; 25:577–583. [PubMed: 11713099]
29. Zuccarello D, Ferlin A, Garolla A, et al. A possible association of a human tektin-t gene mutation (A229V) with isolated non-syndromic asthenozoospermia: case report. Hum Reprod. 2008; 23:996–1001. [PubMed: 18227105]

30. Zhang Z, Jones BH, Tang W, et al. Dissecting the axoneme interactome: the mammalian orthologue of Chlamydomonas PF6 interacts with sperm-associated antigen 6, the mammalian orthologue of Chlamydomonas PF16. Mol Cell Proteomics. 2005; 4:914–923. [PubMed: 15827353]

31. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007; 8:R143. [PubMed: 17659080]

32. Raca G, Jackson C, Warman B, Bair T, Schimmenti LA. Next generation sequencing in research and diagnostics of ocular birth defects. Mol Genet Metab. 2010; 100:184–192. [PubMed: 20359920]

33. Perrone CA, Myster SH, Bower R, O'Toole ET, Porter ME. Insights into the structural organization of the I1 inner arm dynein from a domain analysis of the 1beta dynein heavy chain. Mol Biol Cell. 2000; 11:2297–2313. [PubMed: 10888669]

34. Blouin J-L, Meeks M, Radhakrishna U, et al. Primary ciliary dyskinesia: a genome-wide linkage analysis reveals extensive locus heterogeneity. Eur J Hum Genet. 2000; 8:109–118. [PubMed: 10757642]

35. Katsanis N, Ansley SJ, Badano JL, et al. Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. Science. 2001; 293:2256–2259. [PubMed: 11567139]

36. Walsh T, Lee MK, Casadei S, et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci. 2010; 107:12629–12633. [PubMed: 20616022]

37. Dames S, Durtschi J, Geiersbach K, Stephens J, Voelkerding KV. Comparison of the Illumina Genome Analyzer and Roche 454 GS FLX for resequencing of hypertrophic cardiomyopathy-associated genes. J Biomol Tech. 2010; 21:73–80. [PubMed: 20592870]

38. Hoppman-Chaney N, Peterson LM, Klee EW, Middha S, Courteau LK, Ferber MJ. Evaluation of Oligonucleotide Sequence Capture Arrays and Comparison of Next-Generation Sequencing Platforms for Use in Molecular Diagnostics. Clin Chem. 2010 Jun 18. [Epub ahead of print].

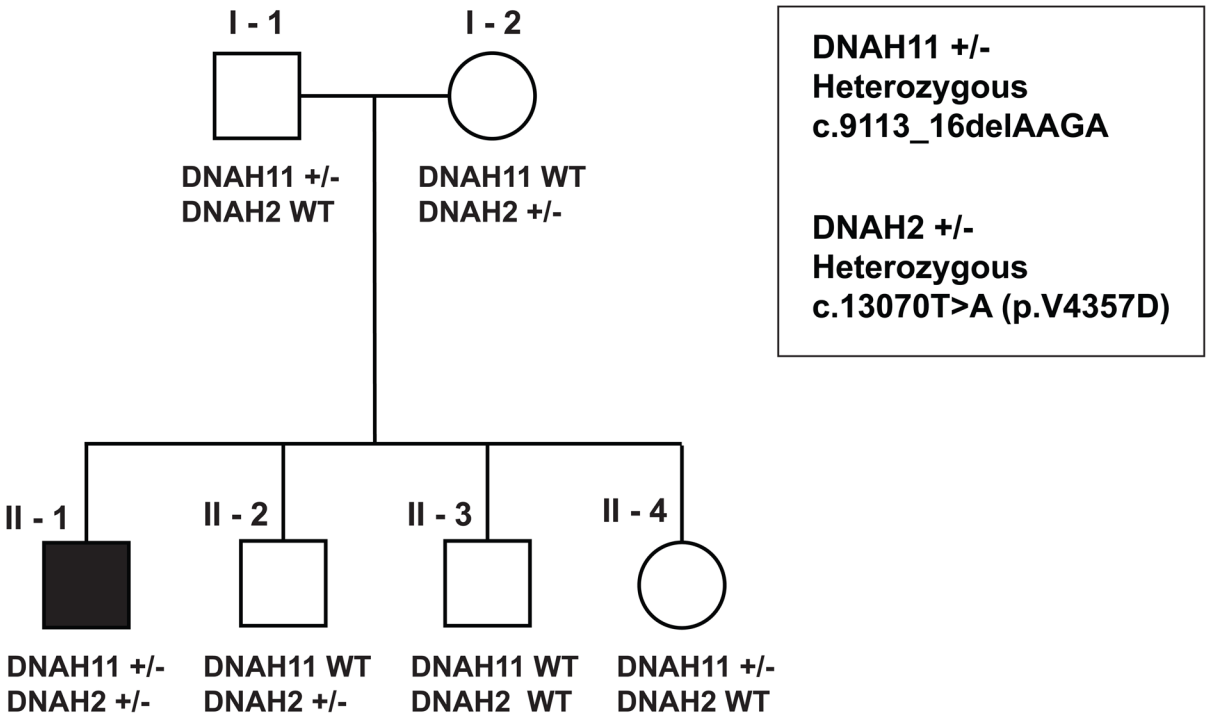

Figure 1. Family segregation analysis supports the possibility of digenic inheritance of *DNAH11* and *DNAH2* mutations in patient 998

A patient with a phenotype consistent with PCD (II-1) was previously found to have a single mutation in *DNAH11*, which encodes a component of the outer dynein arm. Targeted exome sequencing found no additional possible disease-causing mutations in *DNAH11* but identified a heterozygous missense mutation of a highly conserved residue of *DNAH2*, which encodes a component of the inner dynein arm. Family segregation analysis revealed that the *DNAH11* mutation was inherited from the patient's father (I-1), while the *DNAH2* mutation was inherited from the patient's mother (I-2). Furthermore, the three unaffected siblings were heterozygous for only the *DNAH2* allele (brother II-2), heterozygous for only the *DNAH11* allele (sister II-4), or wild-type for both alleles (brother II-3).

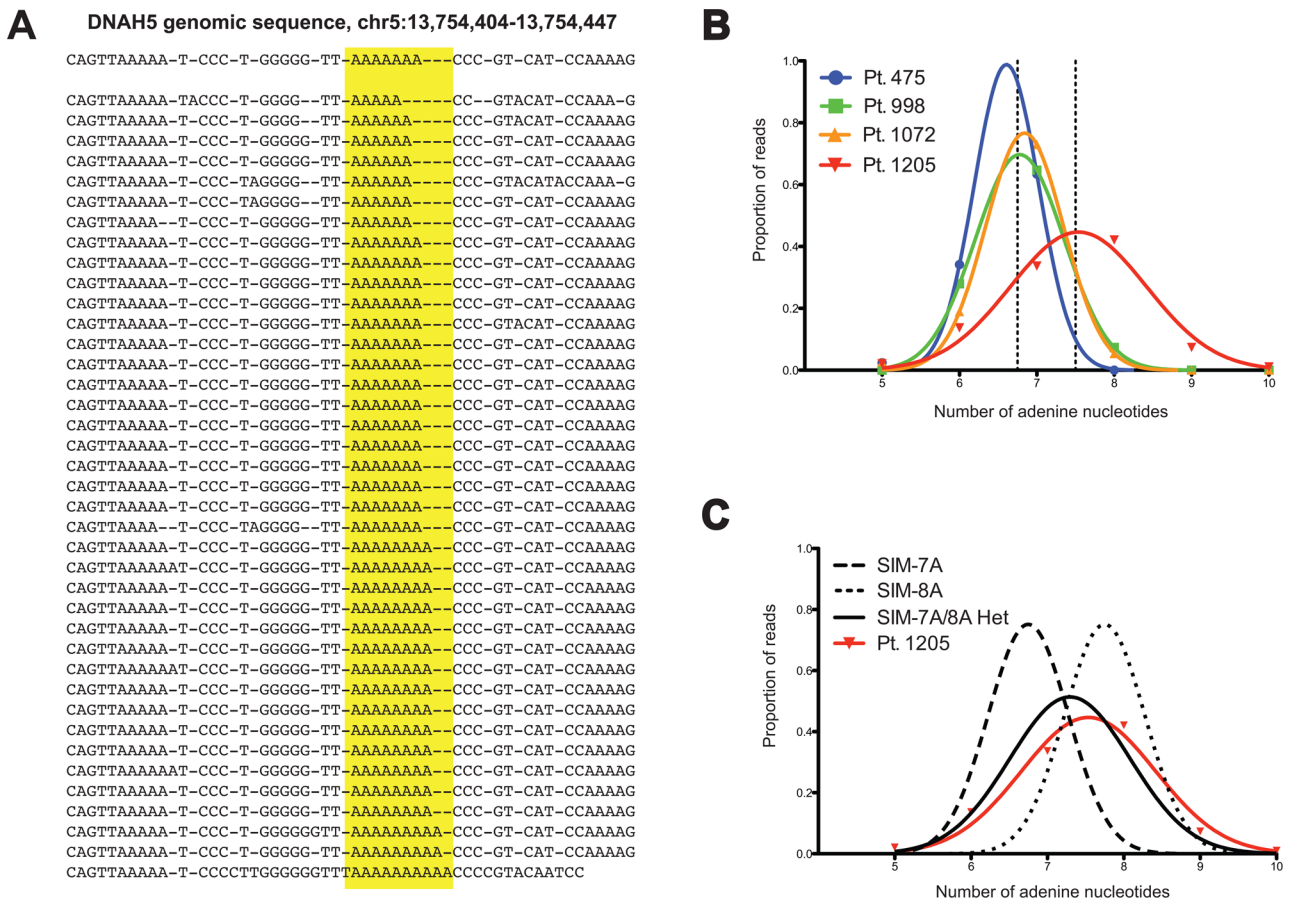

Figure 2. Sequence analysis favors the existence of a heterozygous insertion in patient 1205

A. Local alignment of selected sequence reads from patient 1205 in the region containing a single nucleotide insertion (hg18; chr5:13,754,404 – 13,754,447). The top line of the alignment represents the NCBI 36 reference sequence at this location. Gaps introduced into the alignment are represented by a "-". Yellow highlights the stretch of 7 adenine (7A) nucleotides where the insertion should result in 8 adenine (8A) nucleotides. B. The proportion of reads containing 5, 6, 7, 8, 9, or 10 adenine nucleotides reveal a clear difference in patient 1205. Patients 475, 998, and 1072 act as "controls" since they are homozygous for reference 7A alleles. The highest proportions of reads in these samples were 7A, with Gaussian distributions ($R^2 > 0.998$) having peaks at approximately 6.75 (6.607, 6.778, and 6.842), consistent with a homozygous 7A genotype at this position. In contrast, the number of adenine residues in the sample from patient 1205 ranged from 5 to 10, with a Gaussian distribution ($R^2 = 0.964$) that was broader than the first three patients and had a peak at 7.54. C. When compared to simulated genotypes, the distribution from patient 1205 most closely resembles the heterozygous distribution. The distributions of the three "control" samples were averaged to obtain a simulated homozygous 7A distribution (SIM-7A). This distribution was shifted to the right to generate a simulated homozygous 8A distribution (SIM-8A). The SIM-7A and SIM-8A distributions were averaged to simulate the expected distribution in an individual heterozygous for 7A and 8A (SIM-7A/8A Het).

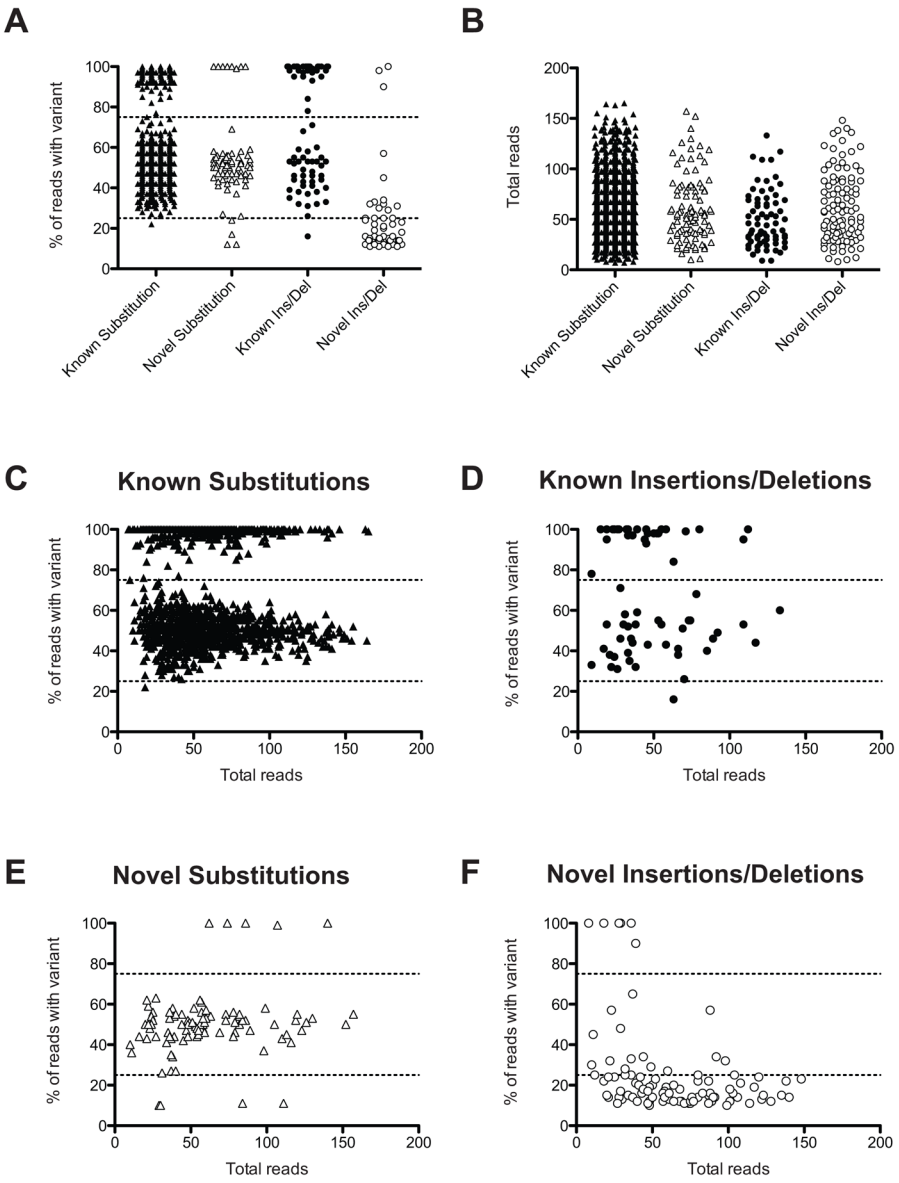

Figure 3. Small insertion/deletion variants are observed in distinct proportions of sequence reads

A. Variants were separated by class and whether they were novel or previously reported in dbSNP, and plotted according to the percentage of sequence reads in which the variant was observed. Dashed lines represent 25% and 75% of sequence reads. Substitution variants largely clustered near 50% or 100% of sequence reads as would be expected for heterozygous or homozygous variants, respectively. Known insertion/deletion variants (Ins/Del) also tended to cluster near 50% or 100% of sequence reads. However, novel small insertion/deletion variants had a much wider distribution with a significant proportion of variants detected in fewer than 25% of the sequence reads. B. The depth of coverage, defined by the number of sequence reads in the alignment for a given variant, is shown for the different types of sequence variants. No obvious differences were observed between the classes of variants in terms of the total depth of coverage. C-F. Known substitutions, known insertions/deletions, novel substitutions, and novel insertions/deletions were plotted according to the total read depth and the percentage of reads in which the variant was observed. The novel insertion/deletion variants clearly follow a distinct pattern, with a large

proportion of the variants detected in fewer than 25% of sequence reads but no clear relationship between depth of coverage and the percentage of reads in which the variant was observed.

Table 1

Patient characteristics and known mutations

Patient (age, sex)	Clinical presentation	nNO	EM features	Gene	Mutation 1	Mutation 2
475 (14y, F)	NRD, SI, Bx, Sx, OM.	10.7 nl/min	ODA defect	*DNAH5*	c.6249G>A (p.M2083I, splice site)	c.7468_7488 del (p.2490_2496 del WSAGAAL)
998 (29y, M)	Bx, Sx, OM	70.4 nl/min	Normal dynein arms and central apparatus	*DNAH11*	c.9113_9116delAAGA (p.K3038TfsX13)	Not identified
1072 (12y, M)	NRD, Sx, OM	10.1 nl/min	ODA defect	*DNAI1*	c.1212T>G (p.Y404X)	c.1644G>A (p.W548X)
1205 (16y, F)	NRD, SI, Bx, Sx, OM	18.9 nl/min	ODA defect	*DNAH5*	Del exon 62 by MLPA (junctions not known)	c.13458_59insT (p.N4487fsX1)

nNO = nasal nitric oxide (normal range 376±124 nl/min, mean±SD)[1]

EM = electron microscopy

NRD = neonatal respiratory distress

SI = situs inversus

Bx = bronchiectasis

Sx = sinusitis

OM = frequent otitis media

ODA = outer dynein arm

MLPA = multiplex ligation-dependent probe amplification

Table 2

Candidate mutations identified through an initial blinded analysis of variant data

	Pt. 475	Pt. 998	Pt. 1072	Pt. 1205
Total HC Variants	18,042	19,877	12,332	10,560
- In Region (-*HYDIN*)	521	435	449	456
- >15% of reads	515	423	443	435
- Novel	51	31	28	49
- Intronic	29	21	16	31
- Synonymous	4	2	3	5
- Missense	10	7	7	8
- Nonsense	0	0	2	0
- Small ins/del	7	1	0	4
- Near splice site	1	0	0	4
- Candidate mutations	17	8	9	16
Genes with two candidate mutations	*DNAH5* *RSPH4A*[*] *DNAH10*[**] *LRRC50*[***]	*DNAH9*[*]	*DNAI1* *DNAH8*[****]	*SPAG17*[**]

[*] False positive insertion mutation

[**] Variants of uncertain significance

[***] Likely sequencing artifact

[****] Determined to be in *cis* by Sanger sequencing

Table 3

Summary of candidate mutations identified or missed by exon-capture and massively parallel sequencing

	Mutation	Gene	AA change	% of reads	Comments
Patient 475	**Disease-causing mutations**				
chr5:13,863,923-13,863,313	Del 21-bp	*DNAH5*	del WSAGAAL	57%	Deletion of 7 amino acids
chr5:13,883,135	C>T	*DNAH5*	M – I	52%	Last nucleotide of exon, likely splice donor site mutation
	Possible compound heterozygous variants				
chr6:117,045,024	Ins T	*RSPH4A*	Q – L, fs	17%	False positive, not detected by Sanger sequencing
chr6:117,058,303	Ins C	*RSPH4A*	Y – L, fs	24%	False positive, not detected by Sanger sequencing
chr12:122863856	C>T	*DNAH10*	R – W	52%	Highly conserved except G in canine
chr12:122864030	C>T	*DNAH10*	H – Y	50%	Poorly conserved; Also present in pt. 998
chr16:82,761,275–82,761,277	CAC>A	*LRRC50*	PP – HT, fs	17%	Located within simple tandem repeat, likely artifact
chr16:82,761,284–82,761,286	CGC>G	*LRRC50*	PP – RT, fs	28%	Located within simple tandem repeat, likely artifact
	Heterozygous variants				
chr1:223335050	A>T	*DNAH14*	H – L	53%	Highly conserved
chr2:84865382	Ins T	*DNAH6*	R – R, fs	16%	False positive, not detected by Sanger sequencing
chr5:35763614	G>A	*SPEF2*	A – T	34%	Poorly conserved AA
chr7:21907348	A>G	*DNAH11*	K – R	57%	Not highly conserved (R in mouse)
chr11:102596607	G>A	*DYNC2H1*	E – K	58%	Invariant amino acid
chr14:101519552	Ins A	*DYNC1H1*	V – V, fs	20%	Possible frameshift but likely false positive
chr19:53492145	G>A	*CCDC114*	S – L	48%	Not highly conserved (L in mouse)
chr21:42769212	C>T	*RSPH1*	G – R	44%	Poorly conserved; Also present in pt. 998
chrX:38041528	T>C	*RPGR*	Q – R	55%	Poorly conserved
Patient 998	**Disease-causing mutations**				
chr7:2177919-21779922	Del AAGA	*DNAH11*	KD – TL, fs	53%	Bioinformatics false negative: matched with rs72657376
	Possible compound heterozygous variants				
chr17:11491146	G>A	*DNAH9*	R – Q	46%	Poorly conserved (Q is common); Confirmed by Sanger sequencing
chr17:11549161	Ins T	*DNAH9*	Q – L, fs	18%	False positive, not detected by Sanger sequencing
	Heterozygous variants				
chr12:109829602	G>T	*CCDC63*	R – L	44%	Poorly conserved

	Mutation	Gene	AA change	% of reads	Comments
Patient 475	**Disease-causing mutations**				
chr12:122864030	C>T	*DNAH10*	H – Y	51%	Poorly conserved; Also present in pt. 475
chr17:7677205	T>A	*DNAH2*	V – D	47%	Invariant amino acid; Confirmed by Sanger sequencing
chr21:42769212	C>T	*RSPH1*	G – R	61%	Poorly conserved; Also present in pt. 475
	Homozygous variants				
chr2:84778334	T>C	*DNAH6*	V – A	100%	Poorly conserved; Also present in pt. 1205
chr16:20882241	C>T	*DNAH3*	R – Q	100%	Poorly conserved (Q is common)
Patient 1072	**Disease-causing mutations**				
chr9:34496773	T>G	*DNAI1*	Y – Stop	53%	Nonsense mutation
chr9:34504466	G>A	*DNAI1*	W – Stop	44%	Nonsense mutation
	Possible compound heterozygous variants				
chr6:39005350	A>G	*DNAH8*	H – R	50%	Poorly conserved (R is common); Confirmed by Sanger sequencing but in *cis* with other DNAH8 variant
chr6:39007668	C>T	*DNAH8*	T – M	56%	Highly conserved except S in lizard; Confirmed by Sanger sequencing but in *cis* with other DNAH8 variant
	Heterozygous variants				
chr1:36323224	C>T	*TEKT2*	R – C	55%	Invariant amino acid; Confirmed by Sanger sequencing
chr2:196446610	T>C	*DNAH7*	T – A	63%	Invariant amino acid
chr16:82746850	G>A	*LRRC50*	D – N	50%	Not highly conserved (N in chicken); Confirmed by Sanger sequencing
chr17:7641273	A>G	*DNAH2*	I – V	35%	Highly conserved; conservative substitution
chr17:11495120	G>A	*DNAH9*	V – M	50%	Invariant amino acid; conservative substitution
Patient 1205	**Disease- causing mutations**				
Exon 62 deletion	N/A	*DNAH5*	N/A	X	Not identified
chr5:13754432	Ins A	*DNAH5*	N – Stop, fs	X	Not identified; evidence from sequence alignment
	Possible compound heterozygous variants				
chr1:118349567	G>T	*SPAG17*	T – N	41%	Predicted damaging; Confirmed by Sanger sequencing
chr1:118425275	C>T	*SPAG17*	M – I	45%	Not highly conserved, predicted benign; Confirmed by Sanger sequencing
	Heterozygous variants				
chr2:27530434	Ins C	*IFT172*	A – P, fs	16%	False positive, not detected by Sanger sequencing
chr2:214502993	G>A	*SPAG16*	D – N	44%	Poorly conserved (N in dog)

	Mutation	Gene	AA change	% of reads	Comments
Patient 475	**Disease-causing mutations**				
chr5:35736480	Ins A	*SPEF2*	A – D, fs	22%	False positive, not detected by Sanger sequencing
chr6:38993126	Ins A	*DNAH8*	S – Y, fs	15%	False positive, not detected by Sanger sequencing
chr11:102532339	C>T	*DYNC2H1*	R – W	56%	Invariant amino acid
chr12:109796040	Ins A	*CCDC63*	M – I, fs	15%	Possible frameshift but likely false positive
chr12:122854217	G>A	*DNAH10*	V – I	53%	Highly conserved; conservative substitution
chr16:21030816	G>A	*DNAH3*	T – M	55%	Not highly conserved (M in Rhesus)
chr16:82746855	Ins A	*LRRC50*	L – L, fs	18%	False positive, not detected by Sanger sequencing
chr17:7603162	C>T	*DNAH2*	P – S	55%	Not highly conserved (S in elephant)
chr17:73988359	A>T	*DNAH17*	N/A	50%	Possible splice donor mutation (intron +2 position)
	Homozygous variants				
chr2:84778334	T>C	*DNAH6*	V – A	100%	Poorly conserved; Also present in pt. 998

Genet Med. Author manuscript; available in PMC 2013 August 27.