HHS Public Access

Author manuscript

Sci Transl Med. Author manuscript; available in PMC 2013 August 11.

Published in final edited form as:

Sci Transl Med. 2011 January 12; 3(65): 65ra4. doi:10.1126/scitranslmed.3001756.

Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing

Callum J. Bell[1,*], **Darrell L. Dinwiddie**[1,2,*], **Neil A. Miller**[1,2], **Shannon L. Hateley**[1], **Elena E. Ganusova**[1], **Joann Mudge**[1], **Ray J. Langley**[1], **Lu Zhang**[3], **Clarence C. Lee**[4], **Faye D. Schilkey**[1], **Vrunda Sheth**[4], **Jimmy E. Woodward**[1], **Heather E. Peckham**[4], **Gary P. Schroth**[3], **Ryan W. Kim**[1], and **Stephen F. Kingsmore**[1,2,†]

[1]National Center for Genome Resources, Santa Fe, NM 87505, USA

[2]Children's Mercy Hospital, Kansas City, MO 64108, USA

[3]Illumina Inc., Hayward, CA 94545, USA

[4]Life Technologies, Beverley, MA 01915, USA

Abstract

Of 7028 disorders with suspected Mendelian inheritance, 1139 are recessive and have an established molecular basis. Although individually uncommon, Mendelian diseases collectively account for ~20% of infant mortality and ~10% of pediatric hospitalizations. Preconception screening, together with genetic counseling of carriers, has resulted in remarkable declines in the incidence of several severe recessive diseases including Tay-Sachs disease and cystic fibrosis. However, extension of preconception screening to most severe disease genes has hitherto been impractical. Here, we report a preconception carrier screen for 448 severe recessive childhood diseases. Rather than costly, complete sequencing of the human genome, 7717 regions from 437 target genes were enriched by hybrid capture or microdroplet polymerase chain reaction, sequenced by next-generation sequencing (NGS) to a depth of up to 2.7 gigabases, and assessed with stringent bioinformatic filters. At a resultant 160× average target coverage, 93% of nucleotides had at least 20× coverage, and mutation detection/genotyping had ~95% sensitivity and ~100% specificity for substitution, insertion/deletion, splicing, and gross deletion mutations

[†]To whom correspondence should be addressed. sfk@ncgr.org.
[*]These authors contributed equally to this work.

Author contributions: C.J.B. led the project, contributed computer programming and data analysis, and wrote the manuscript. D.L.D. contributed to study design; performed literature research, target enrichments, sequencing, genotyping, and data analysis; and wrote the manuscript. N.A.M. carried out data pipelining, software development, and bioinformatics. S.L.H. carried out literature research and data analysis and contributed to target enrichment and sequencing. E.E.G. performed literature research, target enrichment, and sequencing. J.M. provided data analysis. R.J.L. provided target enrichment and sequencing and assisted with data analysis. L.Z. performed sequencing. C.C.L. designed the SOLiD sequencing and data analysis. J.E.W. provided sequencing and genotyping. H.E.P. performed SOLiD 3 sequencing. F.D.S. assisted in project management and provision of resources. V.S. performed data pipelining and bioinformatic analysis. G.P.S. designed the HiSeq sequencing. R.W.K. provided oversight of sequencing operations. S.F.K. conceived and designed the study, wrote the manuscript, and carried out data analysis.

Competing interests: L.Z. is an employee of Illumina Inc. At the time the research was performed, G.P.S. was an employee of Illumina Inc. C.C.L., H.E.P., and V.S. are employees of Life Technologies. U.S. patent application 20090183268 entitled "Methods and systems for medical sequencing analysis" was filed by the National Center for Genome Resources on July 16, 2009. This application has claims related to this work. The other authors declare no competing interests.

Accession numbers: Nucleotide sequences are deposited in the NCBI at SRA026957.1. Nucleotide variants may be searched at http://hematite.ncgr.org.

and single-nucleotide polymorphisms. In 104 unrelated DNA samples, the average genomic carrier burden for severe pediatric recessive mutations was 2.8 and ranged from 0 to 7. The distribution of mutations among sequenced samples appeared random. Twenty-seven percent of mutations cited in the literature were found to be common polymorphisms or misannotated, underscoring the need for better mutation databases as part of a comprehensive carrier testing strategy. Given the magnitude of carrier burden and the lower cost of testing compared to treating these conditions, carrier screening by NGS made available to the general population may be an economical way to reduce the incidence of and ameliorate suffering associated with severe recessive childhood disorders.

INTRODUCTION

Preconception testing of motivated populations for recessive disease mutations, together with education and genetic counseling of carriers, can markedly reduce disease incidence within a generation. Tay-Sachs disease [TSD; Online Mendelian Inheritance in Man (OMIM) accession number 272800], for example, is an autosomal recessive neuro-degenerative disorder with onset of symptoms in infancy and death by 2 to 5 years of age. Formerly, the incidence of TSD was 1 per 3600 Ashkenazi births in North America (1, 2). After 40 years of preconception screening in this population, however, the incidence of TSD has been reduced by more than 90% (2–5). Although TSD remains incurable, therapies are available for many severe recessive diseases of childhood. Thus, in addition to disease prevention, preconception testing could enable perinatal diagnosis and treatment, which can profoundly diminish disease severity.

Although individual Mendelian diseases are uncommon in general populations, collectively, they account for ~20% of infant mortality and ~10% of pediatric hospitalizations (6, 7). Over the past 25 years, 1139 genes that cause Mendelian recessive diseases have been identified (8). To date, however, preconception carrier testing has been recommended in the United States only for five of these: fragile × syndrome (OMIM #300624) in selected individuals; cystic fibrosis (OMIM #219700) in Caucasians; and TSD, Canavan disease (OMIM #271900), and familial dysautonomia (OMIM #223900) in individuals of Ashkenazi descent (9–13). A framework for the development of criteria for comprehensive preconception screening can be inferred from an American College of Medical Genetics (ACMG) report on expansion of newborn screening for inherited diseases (14). Criteria included test accuracy and cost, disease severity, highly penetrant recessive inheritance, and whether an intervention was available for those identified. These criteria are also relevant for expansion of preconception carrier screening. Hitherto, important criteria precluding extension of preconception screening to most severe recessive mutations or the general population have been cost [defined in that report as an overall analytical cost requirement of <$1 per test per condition (14)] and the absence of accurate, sensitive, scalable technologies.

Target capture and next-generation sequencing (NGS) have shown efficacy and, recently, scalability for resequencing human genomes and exomes, providing an alternative potential paradigm for comprehensive carrier testing (15–22). In genome research, an average depth of sequence coverage of 30-fold has been accepted as sufficient for single-nucleotide

polymorphism (SNP) and nucleotide insertion or deletion (indel) detection (15–22). However, acceptable false-positive and false-negative rates for routine use in clinical practice are more stringent and are driven by the intended purpose for which the data are to be used. Data demonstrating the sensitivity and specificity of genotyping of disease mutations, particularly polynucleotide indels, gross insertions and deletions, copy number variations (CNVs), and complex rearrangements, are very limited (20–22). In particular, the accuracy of disease mutation genotypes derived from NGS of enriched targets has been uncertain.

A recent workshop provided recommendations for qualification of new methodologies for broader population-based carrier screening (23). These were high analytical validity, concordance in many settings, high throughput, and cost-effectiveness (including sample acquisition and preparation). Here, we report the development of a preconception carrier screen for 448 severe recessive childhood disease genes, based on target enrichment and NGS that meets most of these criteria, and use of the screen to assess carrier burden for severe recessive diseases of childhood.

RESULTS

Disease inclusion

The carrier test reported herein was based on several hypotheses. First, cost-effectiveness was assumed to be critical for test adoption. The incremental cost associated with increasing the degree of multiplexing was assumed to decrease toward an asymptote. Thus, very broad coverage of diseases was assumed to offer optimal cost-benefit. Second, comprehensive mutation sets, allele frequencies in populations, and individual mutation genotype-phenotype relationships have been defined in very few recessive diseases. In addition, some studies of cystic fibrosis carrier screening for a few common alleles have shown decreased prevalence of tested alleles with time, rather than reduced disease incidence (24, 25). These two lines of evidence suggested that very broad coverage of mutations offered the greatest likelihood of substantial reductions in disease incidence with time. Third, physician, patient, and societal adoption of screening was assumed to be optimal for the most severe and highly penetrant childhood diseases, before conception and where the anticipated clinical validity and clinical utility of testing was clear (26). Therefore, diseases were chosen that would almost certainly change family planning by prospective parents or affect antenatal, perinatal, or neonatal care. Milder recessive disorders, such as deafness, and adult-onset diseases, such as inherited cancer syndromes, were omitted, as were conditions lacking strong evidence for causal mutations (26).

Database and literature searches and expert reviews were performed on 1123 diseases with recessive inheritance of known molecular basis (8, 27, 28). In general, diseases were selected to meet ACMG guidelines for genetic testing for rare, highly penetrant disorders (26). Assessment of the clinical validity and utility of testing was primarily based on literature review and was challenging for some disorders because of the paucity of data. Several subordinate requirements were gathered: In view of pleiotropy and variable severity, disease genes were included if mutations caused severe illness in a proportion of affected children. All but six diseases that featured genocopies (including variable inheritance and

mitochondrial mutations) were included. Diseases were not excluded on the basis of low incidence. Diseases for which large population carrier screens exist were included, such as TSD, hemoglobinopathies, and cystic fibrosis. Mental retardation genes were not included in this iteration. Four hundred and forty-eight X-linked recessive and autosomal recessive diseases, encompassing 437 genes, met these criteria (table S1). The disease type was cardiac for 8, cutaneous for 45, developmental for 46, endocrine for 15, gastroenterological for 3, hematological for 15, hepatic for 3, immunological for 29, metabolic for 142, neurological for 122, ocular for 12, renal for 25, respiratory for 8, and skeletal for 28. Note that these genes, although a good representative set, require further assessment of clinical readiness before translation into clinical testing.

Technology selection

Array hybridization with allele-specific primer extension was initially favored for expanded carrier detection because of test simplicity, cost, scalability, and accuracy, as has recently been described (29). To be well suited for array-based screening, however, most carriers must be accounted for by a few mutations, and most disease mutations must be nucleotide substitutions (8, 27, 28). Of 215 autosomal recessive disorders examined, only 87 were assessed to meet these criteria. Most recessive disorders for which a large proportion of burden was attributable to a few disease mutations were limited to specific ethnic groups. Indeed, 286 severe childhood autosomal recessive diseases encompassed 19,640 known disease mutations (8, 27, 28). Given that the Human Gene Mutation Database (HGMD) lists 102,433 disease mutations (27), a number that is steadily increasing, a fixed-content method appeared impractical. Other concerns with array-based screening for recessive disorders were type 1 errors in the absence of confirmatory testing and type 2 errors for disease mutations other than substitutions (complex rearrangements, indels, or gross deletions with uncertain boundaries). A serendipitous discovery (discussed below) that supported this decision was an unexpectedly high number of characterized mutations that are misannotated.

The effectiveness and remarkable decline in cost of exome capture and NGS for variant detection in genomes and exomes suggested an alternative potential paradigm for comprehensive carrier testing. Four target enrichment and three NGS methods were preliminarily evaluated for multiplexed carrier testing. Preliminary experiments suggested that existing protocols for Agilent SureSelect hybrid capture (15) and RainDance microdroplet polymerase chain reaction (PCR) (16) but not Febit HybSelect microarray-based biochip capture (30) or Olink padlock probe ligation and PCR (31) yielded consistent target enrichment. Therefore, workflows and software pipelines were developed for comprehensive carrier testing by hybrid capture or microdroplet PCR, followed by NGS (Fig. 1). Baits or primers were designed to capture or amplify 1,978,041 nucleotides (nt), corresponding to 7717 segments of 437 recessive disease genes by hybrid capture and microdroplet PCR, respectively. Targeted were all coding exons and splice site junctions, and intronic, regulatory, and untranslated regions known to contain disease mutations (table S2). In general, baits for hybrid capture or PCR primers were designed to encompass or flank disease mutations, respectively. Primers were also designed to avoid known polymorphisms and to minimize nontarget nucleotides. To capture or amplify both the normal and the disease mutation alleles, we also designed custom baits or primers for 11

gross deletion disease mutations for which boundaries had been defined (table S3). A total of 29,891 120-mer RNA baits were designed to capture 98.7% of targets. Fifty-five percent of 101 exons that failed bait design contained repeat sequences (table S4). Primer pairs (10,280) were designed to amplify 99% of targets (table S5). Twenty exons failed primer design by falling outside the amplicon size range of 200 to 600 nt.

Analytic metrics

An ideal target enrichment protocol would inexpensively result in at least 30% of nucleotides being on target, which corresponded to ~500-fold enrichment with ~2-million-nucleotide target size. This was achieved with hybrid capture after one round of bait redesign for underrepresented exons and decreased bait representation in over-represented exons (Table 1). An ideal target enrichment protocol would also give a narrow distribution of target coverage and without tails or skewness (indicative of minimal enrichment-associated bias). After hybrid capture, the sequencing library size distribution was narrow (Fig. 2A). The aligned sequence coverage distribution was unimodal but flat (platykurtic) and right-skewed (Fig. 2B). This implied that hybrid capture would require oversequencing of most targets to recruit a minority of poorly selected targets to adequate coverage. As expected, median coverage increased linearly with sequence depth. The proportion of bases with greater than zero and >20× coverage increased toward asymptotes at ~99 and ~96%, respectively (Table 1 and Fig. 2C). Targets with low (<3×) coverage were highly reproducible and had high GC content (table S6). This suggested that targets failing hybrid capture could be predicted and, perhaps, rescued by individual PCRs.

Given the need for highly accurate carrier detection, we required >10 uniquely aligned reads of quality score >20 and >14% of reads to call a variant (20, 21). The requirement for >10 reads was highly effective for nucleotides with moderate coverage. For heterozygote detection, for example, this was equivalent to ~20× coverage, which was achieved in ~96% of exons with ~2.6 gigabases (Gb) of sequence (Fig. 2C). The proportion of targets with at least 20× coverage appeared to be useful for quality assessment. The requirement for ≥14% of reads to call a variant was highly effective for nucleotides with very high coverage and was derived from the genotype data discussed below. A quality score requirement was important when NGS started, but is now largely redundant.

In theory, microdroplet PCR should result in all cognate amplicons being on target and should induce minimal bias. In practice, the coverage distribution was narrower than hybrid capture but with similar right skewing (Fig. 2D). However, these results were complicated by ~11% recurrent primer synthesis failures. This resulted in linear amplification of a subset of targets, ~5% of target nucleotides with zero coverage and a similar proportion of nucleotides on target to that obtained in the best hybrid capture experiments (~30%; Table 1). Hybrid capture was used for subsequent studies for reasons of cost.

Multiplexing of samples during hybrid selection and NGS had not previously been reported. Six- and 12-fold multiplexing was achieved by adding molecular bar codes to adaptor sequences. Interference of bar code nucleotides with hybrid selection did not occur appreciably: The stoichiometry of multiplexed pools was essentially unchanged before and after hybrid selection. Multiplexed hybrid selection was found to be ~10% less effective

than singleton selection, as assessed by median fold enrichment. Less than 1% of sequences were discarded at alignment because of bar code sequence ambiguity. Therefore, up to 12-fold multiplexing at hybrid selection and per sequencing lane (equivalent to 96-plex per sequencing flow cell) was used in subsequent studies to achieve the targeted cost of <$1 per test per sample.

Several NGS technologies are currently available. Of these, the Illumina sequencing-by-synthesis (SBS) and SOLiD sequencing-by-ligation (SBL) platforms are widely disseminated and have throughput of at least 50 Gb per run and read lengths of at least 50 nt. Therefore, the quality and quantity of sequences from multiplexed, target-enriched libraries were compared with SBS (GAIIx singleton 50-mer) and SBL (SOLiD3 singleton 50-mer; Table 1). SBS- and SBL-derived 50-mer sequences (and alignment algorithms) gave similar alignment metrics (Table 1). When compared with Infinium array results, specificity of SNP genotypes by SBS and SBL was very similar (SBS, 99.69%; SBL, 99.66%), reflecting both target enrichment and multiplexed sequencing (Fig. 3).

Given approximate parity of throughput and accuracy, consideration was given to optimal read length. Unambiguous alignment of short-read sequences is typically confounded by repetitive sequences, but was not relevant for carrier testing, because targets overwhelmingly contained unique sequences. The number of mismatches tolerated for unique alignment of short-read sequences is highly constrained but increases with read length. The vast majority of disease mutations are single-nucleotide substitutions or small indels. However, comprehensive carrier testing also requires detection of polynucleotide indels, gross insertions, gross deletions, and complex rearrangements. A combination of bioinformatic approaches was used to overcome short-read alignment shortcomings (Fig. 4). First, with the Illumina HiSeq SBS platform, we used the novel approach of read pair assembly before alignment (99% efficiency) to generate longer reads with high-quality scores (148.6 ± 3.8 nt combined read length and increase in nucleotides with quality score >30 from 75 to 83%). This was combined with generation of 150-nt sequencing libraries without gel purification by optimization of DNA shearing procedures and use of silica membrane columns. Omission of gel purification was critical for scalability of library generation. Second, we reduced the penalty on polynucleotide variants, rewarding identities (+1) and penalizing mismatches (−1) and indels [−1–log(indel – length)]. Third, gross deletions were detected both by perfect alignment to mutant junction reference sequences and by local decreases in normalized coverage (normalized to total sequence generated; C. H. Hu, personal communication). Previous studies have identified CNVs on the basis of changes in regional coverage along a chromosome in an individual sample (20, 21). However, concomitant analysis of normalized coverage in batches of samples appears to circumvent the need for adjustment for GC content (32), allowing more accurate detection of segmental losses. This was illustrated by identification of eight known gross deletion disease mutations (Fig. 5). Furthermore, seeking perfect alignment to mutant junction reference sequences obviates low alignment scores when short reads containing polynucleotide variants are mapped to a normal reference. This was illustrated by identification of 11 gross deletion mutations for which boundaries had been defined (table S3). It is anticipated that these approaches could be extended to gross insertions and complex rearrangements but will require additional analytical validation.

Clinical metrics

On the basis of these strategies and our previous experience of genotyping variants identified in next-generation genome and chromosome sequences (20, 21, 33, 34), a bioinformatic decision tree for genotyping disease mutations was developed (Fig. 4). Clinical utility of target enrichment, SBS sequencing, and this decision tree for genotyping disease mutations was assessed. SNPs in 26 samples were genotyped by both high-density arrays and sequencing. The distribution of read count–based allele frequencies of 92,106 SNP calls was tri-modal, with peaks corresponding to homozygous reference alleles, heterozygotes, and homozygous variant alleles, as ascertained by array hybridization (Fig. 6B). Optimal genotyping cutoffs were 14 and 86% (Fig. 6B). With these cutoffs and a requirement for 20× coverage and 10 reads of quality ≥20 to call a variant, the accuracy of sequence-based SNP genotyping was 98.8%, sensitivity was 94.9%, and specificity was 99.99%. The positive predictive value (PPV) of sequence-based SNP genotypes was 99.96% and negative predictive value (NPV) was 98.5%, as ascertained by array hybridization. As sequence depth increased from 0.7 to 2.7 Gb, sensitivity increased from 93.9 to 95.6%, whereas PPV remained ~100% (Fig. 6A). Areas under the curve (AUCs) of the receiver operating characteristic (ROC) for SNP calls by hybrid capture and SBS were calculated. When genotypes in 26 samples were compared with genome-wide SNP array hybridization, the AUC was 0.97 when either the number or the percent reads calling a SNP were varied (Fig. 6, C and D). When the parameters were combined, the AUC was 0.99. For known substitution, indel, splicing, gross deletion, and regulatory alleles in 76 samples, sensitivity was 100% (113 of 113 known alleles; table S7). The higher sensitivity for detection of known mutations reflected manual curation. The 20 known indels were confirmed by PCR and Sanger sequencing. Notably, substitutions, indels, splicing mutations, and gross deletions account for the vast majority (96%) of annotated mutations (27).

Unexpectedly, 14 of 113 literature-annotated disease mutations were either incorrect or incomplete (table S7) (35–39). PCR and Sanger sequencing confirmed that the 14 variants and genotypes called by NGS were correct. For example, sample NA07092, from a male with X-linked recessive Lesch-Nyhan syndrome (OMIM #300322), was characterized as a deletion of HPRT1 exon 8 by complementary DNA (cDNA) sequencing (40), but had an explanatory splicing mutation (intron 8, IVS8+1_4delGTAA, chrX: 133460381_133460384delGTAA; Fig. 7A). NA09545, from a male with XLR Pelizaeus-Merzbacher disease (PMD; OMIM #312080), characterized as a substitution disease mutation [PLP1 exon 5, c.767C>T, P215S (41)], was found to also feature PLP1 gene duplication [which is reported in 62% of sporadic PMD (42); Fig. 7B]. NA02057, from a female with aspartylglucosaminuria (OMIM #208400), characterized as a compound heterozygote, was homozygous for two adjacent substitutions (AGA exon 4, c.482G>A, R161Q, chr4:178596918G>A and exon 4, c.488G>C, C163S, chr4:178596912G>C in 38 of 39 reads; Fig. 8), of which C163S had been shown to be the disease mutation (43). Although one allele of NA01712, a CHT with Cockayne syndrome type B (OMIM #133540), had been characterized by cDNA analysis as a deletion of ERCC6 exon 9 [c.1993_2169del, p. 665_723del, exon 9 del, chr10:50360915_50360739del (44)], no decrease in normalized exon 9 read number was observed despite more than 300× coverage (Fig. 5G). Instead, however, 64 of 138 NA01712 reads contained a nucleotide substitution that created a

premature stop codon (Q664X, chr10:50360741C>T). Both ERCC4 mutations described in CHT NA03542 were absent in at least 130 aligning reads (44). However, the current study used DNA from Epstein-Barr virus (EBV)–transformed cell lines in which somatic hypermutation has been noted (45). In particular, ERCC4, a DNA repair gene, is a likely candidate for somatic mutation. Including these results, the specificity of sequence-based genotyping of substitution, indel, gross deletion, and splicing disease mutations was 100% (97 of 97).

Carrier burden

The average carrier burden of severe recessive disease mutations for severe childhood recessive diseases was assessed in 104 DNA samples. All variants meeting the filtering criteria described above and flagged as disease mutations in HGMD were enumerated. Seventy-four percent of these, however, were accounted for by 47 substitutions each with an incidence of ≥5%, of which 20 were homozygous in samples unaffected by the corresponding disease (table S8). These were omitted. Literature support for pathogenicity was evaluated for the remaining variants flagged as disease mutations in HGMD. Variants were retained as disease mutations if they had been shown to result in loss of activity in a functional assay, were the only variants detected in affected individuals and absent in controls, and/or were predicted to result in a premature stop codon or loss of a substantial portion of the protein (Fig. 4). In total, 27% (122 of 460) of literature-cited disease mutations were omitted, because they were adjudged to be common polymorphisms or sequencing errors or because of a lack of evidence of pathogenicity. New, putatively deleterious variants (variants in severe pediatric disease genes that create premature stop codons or coding domain frameshifts) were quantified: 26 heterozygous or hemizygous new nonsense variants were identified in 104 samples (table S9). Including the latter, 336 variants were retained as likely disease mutations.

The average carrier burden of severe recessive substitutions, indels, and gross deletion disease mutations, after exclusion of one allele in compound heterozygotes, was 2.8 per genome (291 in 104 samples). The carrier burden frequency distribution was unimodal with slight right skewing (Fig. 7C). The range in carrier burden was surprisingly narrow (zero to seven per genome, with a mode of two; Fig. 7C).

As exemplified by cystic fibrosis, the carrier incidence and mutation spectrum of individual recessive disorders vary widely among populations (46). However, whereas group sizes were small, no significant differences in total carrier burden were found between Caucasians and other ethnicities, between males and females, nor between affected and unaffected individuals (after correction for compound heterozygosity in those affected). Hierarchical clustering of samples and disease mutations revealed an apparently random topology, suggesting that targeted population testing is likely to be ineffective (Fig. 7D). Adequacy of hierarchical clustering was attested to by samples from identical twins being nearest neighbors, as were two disease mutations in linkage disequilibrium.

DISCUSSION

We have described a screening test for carriers of 448 severe childhood recessive illnesses consisting of target enrichment, NGS, and bioinformatic analyses, which worked well in a research setting. Specificity was 99.96%, and a sensitivity of ~95% was attained with hybrid capture at a sequence depth of 2.5 Gb per sample. Because enrichment failures with hybrid capture were reproducible, they may be amenable to rescue by individual PCR or probe redesign. Alternatively, microdroplet PCR should theoretically achieve a sensitivity of ~99%, albeit at higher cost (16, 47). The test was scalable, modular, and amenable to automation, with batches of 192 samples and a turnaround of 2 weeks. The time to first result could be reduced substantially with microdroplet PCR and third-generation sequencing. At high volume, the overall analytical cost of the hybrid enrichment-based test was $378, achieving the requirement of <$1 per test per condition and approximating that expended on treatment of severe recessive childhood disorders per U.S. live birth (14, 29). Although the analytical cost will decrease as the throughput of NGS improves, test interpretation, reporting, genetic counseling, and stewardship of mutation databases will confer considerable additional costs.

Having established technical feasibility in a research setting, the next phases of carrier test development will be refinement of the list of diseases, automation, software implementation, report development, and, most important, validation in a realistic testing situation featuring investigator blinding and less manual review. For example, genes associated with severe cognitive developmental disorders may merit inclusion. Although technical standards and guidelines have been established for laboratory-developed genetic testing for rare disorders in accredited laboratories (26), there are several challenges in their adoption for NGS and bioinformatic-based testing of ~500 conditions. For example, specific national standards for quality assurance, quality control, test accessioning and reporting, and proficiency evaluation do not currently exist. Addressing crucial issues such as specificity and false positives is complex when hundreds of genes are being sequenced simultaneously. For certain diseases, such as cystic fibrosis, reference sample panels and metrics have been established. For diseases without such materials, it is prudent to test as many samples containing known mutations as possible. In setting up and validating the screen, it would also be necessary to test examples of all classes of mutations and situations that are anticipated to be potentially problematic, such as mutations within high GC content regions, simple sequence repeats, and repetitive elements.

The ethical, legal, and social implications of comprehensive carrier testing warrant much discussion. These issues, in turn, are influenced by the scope and setting in which testing is proposed. The ideal age for recessive disease screening is in early adulthood and before pregnancy (48, 49). One possibility would be voluntary community-based population testing. This would have an advantage over testing in a hospital setting, where information about carrier testing often is communicated during pregnancy or after the birth of an affected child (50). Community-based carrier testing has had high uptake, without apparent stigma or discrimination and with substantial reductions in the frequencies of tested disorders (3, 48, 49, 51–54). After stakeholder discussions, the cost-effectiveness and clinical utility of offering community-based carrier testing would require detailed assessment. Examination of

the results of existing population-based carrier screening programs for TSD and cystic fibrosis could provide templates for such analyses.

Rapid adoption of comprehensive carrier testing is likely by in vitro fertilization clinics, where screening of sperm and oocyte donors has high clinical utility, lower counseling burden, and small incremental cost (55). Early adoption is also likely in medical genetics clinics, where counseling resources already exist, to screen individuals with a family history of inherited disease. Although the data reported herein are preliminary, the apparent random distribution of mutations in individuals argues against screening different populations for different diseases. The most significant hurdles to implementing comprehensive carrier screening will be facile interpretation of results, reporting in a manner comprehensible by physicians and patients, education of the public of the benefits and limitations of screening, and provision of genetic counselors.

Currently, a two-stage approach is used for preconception carrier screening of couples, with confirmatory testing of all positive results. However, this has been in a setting of testing individual genes for specific mutations where positive results are rare. The requirement for at least 10 high-quality reads to substantiate a variant call resulted in a specificity of 99.96% for single-nucleotide substitutions (which is the limit of accuracy for the gold standard method used) and 100% for about 200 known mutations and new indels in our screening method. It appeared, therefore, that confirmatory testing of all single-nucleotide substitutions and indels was unnecessary. Obviously, inclusion of controls in each test run and random sample retesting will be required. Experience with polynucleotide indels, copy number variants, gross insertions and deletions, and complex rearrangements is as yet insufficient to draw firm conclusions. However, detection of perfect alignments to mutant reference sequences appeared to be robust for identification of gross insertions and deletions. We noted, however, that identification of larger polynucleotide indels was influenced in some sequences by the particular alignment seed, suggesting that additional refinement of alignment parameters is needed.

We found an unexpectedly high proportion of literature-annotated disease mutations that were incorrect, incomplete, or common polymorphisms. Differentiation of common polymorphisms from disease mutations requires genotyping a large number of unaffected individuals. Severe, orphan disease mutations should be uncommon (<1% incidence) and should not be found in the homozygous state in unaffected individuals. Unexpectedly, we found that 74% of “disease mutation” calls were accounted for by substitutions with incidences of ≥5%, of which almost one-half were homozygous in samples unaffected by the corresponding disease. Also unexpected was the finding that 14 of 113 literature-annotated disease mutations were incorrect. Thus, for many recessive diseases, HGMD, dbSNP, OMIM, and the literature are insufficient arbiters of whether variants are disease mutations. We have shown NGS of samples from affected individuals to be a powerful method for error correction: More than three-quarters of errors in mutation identification were Sanger sequencing interpretation errors or incorrect imputation of genomic mutations from cDNA sequencing. Key advantages of NGS are clonal derivation (facilitating unambiguous detection of heterozygous and indel variants), maintenance of phase information (allowing haplotype derivation for adjacent variants), and highly redundant

coverage (resulting in extremely low consensus error rates). Thus, although we have shown that it is technically feasible to undertake comprehensive analysis of recessive gene sequences, sequencing of many unaffected and affected samples will be required to establish an authoritative disease mutation database. Specifically, current reference resources contain common polymorphisms that are annotated as disease mutations and erroneous disease mutations. Without reference database improvements, the clinical utility of comprehensive carrier testing will be limited. Aside from nonsense mutations and premature stop codons in known disease genes and the study of affected individuals, additional bioinformatic approaches will be needed to distinguish rare benign variants from pathogenic variants: Amino acid substitution characteristics such as physicochemical and evolutionary conservation and location (where tertiary structure is known) are useful but not definitive. For many rare variants, functional assays will need to be developed to assess pathogenicity rigorously. Establishment of an authoritative database of disease mutations is clearly needed and represents a nascent bottleneck in progress toward prevention, diagnosis, and treatment of recessive diseases. In the interim, clinical interpretation of the functional importance or pathogenicity of variants will be challenging for many recessive diseases.

A first estimate of the average carrier burden of disease mutations (substitutions, indels, and gross deletions) causing severe childhood recessive diseases was determined: In 104 unrelated individuals, it was 2.8 per genome. Several qualifications of this burden estimate should be noted. First, as discussed, an adequate compilation of pathogenic mutations does not currently exist, and strong evidence of pathogenicity was absent for some of the variants referred to as disease mutations. Second, the burden estimate excluded new, rare, missense variants of unknown significance (VUSs), some of which are likely to be pathogenic. The burden of nonconservative, nonsynonymous, uncommon (<5% incidence) VUS was ~11 per sample. Additional strategies are needed to triage these variants. Third, many individuals in our cohort were affected by one of these diseases. Although a correction was made for compound heterozygote and homozygote alleles, the burden estimate did not correct for other potential selection biases. Fourth, we did not assess gross deletions or other copy number variants beyond limited CNV array hybridization and examination of coverage changes in a small number of known deletions. Nevertheless, a burden of 2.8 per genome agreed with theoretical estimates of reproductive lethal allele burden (56). It also concurred with severe childhood recessive carrier burdens that we obtained by analyzing published individual genomes [2 substitution disease mutations in the Quake genome and a monozygotic twin pair (21, 57), 5 each in the YH and Watson genomes (58, 59), 4 in the NA07022 genome (60, 61), and 10 in the AK1 genome (20)]. The range in carrier burden was surprisingly narrow (zero to seven per genome). Given the large variations in SNP burden and incidence of individual disease alleles among populations, it will be of great interest to evaluate variation in the burden of severe recessive disease mutations among human populations and how this has been influenced by population bottlenecks.

Finally, the technology platform described herein is agnostic with regard to target genes or clinical setting. A variety of medical applications for this technology exist beyond use in preconception carrier screening. For example, comprehensive newborn screening for treatable or preventable Mendelian diseases would allow early diagnosis and institution of treatment while neonates are asymptomatic. Early treatment can have a profound impact on

the clinical severity of conditions and could provide a framework for centralized assessment of investigational new treatments before organ failure. In some cases, such as Duarte variant galactosemia, molecular testing would be superior to conventional biochemical testing. Organ or symptom menu-based diagnostic testing, with masking of nonselected conditions, is anticipated to assist clinical geneticists and pediatric neurologists, because current practice often involves costly, sequential testing of numerous candidate genes. Given impending identification of new disease genes by exome and genome resequencing, the number of disease genes is likely to increase substantially over the next several years, requiring incremental expansion of the target gene sets.

In summary, a technology platform for comprehensive preconception carrier screening for 448 recessive childhood diseases is described. Combining this technology with genetic counseling could reduce the incidence of severe recessive pediatric diseases and may help to expedite diagnosis of these disorders in newborns.

MATERIALS AND METHODS

Disease selection

Criteria for disease inclusion for preconception screening were broadly based on those for expansion of newborn screening, but with omission of treatment criteria (14). Thus, very broad coverage of severe childhood diseases and mutations was sought to maximize cost-benefit, potential reduction in disease incidence, and adoption. A Perl parser identified severe childhood recessive disorders with known molecular basis in OMIM (8). Database and literature searches and expert reviews were performed on resultant diseases (8, 27, 28). Six diseases with extreme locus heterogeneity were omitted (OMIM #209900, #209950, Fanconi anemia, #256000, #266510, #214100). Diseases were included if mutations caused severe illness in a proportion of affected children and despite variable inheritance, mitochondrial mutations, or low incidence. Mental retardation and mitochondrial genes were excluded. Four hundred and thirty-seven genes, representing 507 recessive diseases, met these criteria, of which 448 diseases were severe (table S3).

DNA samples

Target enrichment was performed with 104 DNA samples obtained from the Coriell Institute (Camden, NJ) (table S7). Seventy-six of these were known to be carriers or affected by 37 severe, childhood recessive disorders. The latter samples contained 120 known disease mutations in 34 genes (63 substitutions, 20 indels, 13 gross deletions, 19 splicing, 2 regulatory, and 3 complex disease mutations). They also represented homozygous, heterozygous, compound heterozygous, and hemizygous disease mutation states. Twenty-six samples were well characterized, from “normal” individuals, and two had previously undergone genome sequencing (21).

Target enrichment and SBS

For Illumina GAIIx SBS, 3 μg of DNA was sonicated by Covaris S2 to ~250 nt with 20% duty cycle, 5 intensity, and 200 cycles per burst for 180 s. For Illumina HiSeq SBS, shearing to ~150 nt was by 10% duty cycle, 5 intensity, and 200 cycles per burst for 660 s. Bar-coded

sequencing libraries were made per the manufacturer's protocols. After adaptor ligation, Illumina libraries were prepared with AMPure bead (Beckman Coulter) rather than with gel purification. Library quality was assessed by optical density and electrophoresis (Agilent 2100).

SureSelect enrichment of 6-, 8-, or 12-plex pooled libraries was per Agilent protocols (15), with 100 ng of custom bait library, blocking oligonucleotides specific for paired-end sequencing libraries and 60-hour hybridization. Biotinylated RNA library hybrids were recovered with streptavidin beads. Enrichment was assessed by quantitative PCR (Life Technologies; CLN3, exon 15, Hs00041388_cn; HPRT1, exon 9, Hs02699975_cn; LYST, exon 5, Hs02929596_cn; PLP1, exon 4, Hs01638246_cn) and a nontargeted locus (chrX: 77082157, Hs05637993_cn) before and after enrichment.

RainDance RDT1000 target enrichment was as described and used a custom primer library (16, 46): Genomic DNA samples were fragmented by nebulization to 2 to 4 kb and 1 μg mixed with all PCR reagents but primers. Microdroplets containing three primer pairs were fused with PCR reagent droplets and amplified. After emulsion breaking and purification by MinElute column (Qiagen), amplicons were concatenated overnight at 16°C and sequencing libraries were prepared. Sequencing was performed on Illumina GAIIx and HiSeq2000 instruments per the manufacturer's protocols, as described (20, 21).

Hybrid capture and SBL

DNA (3 μg) was sheared by Covaris to ~150 nt with 10% duty cycle, 5 intensity, and 100 cycles per burst for 60 s. Bar-coded fragment sequencing libraries were made with Life Technologies protocols and reagents. Taqman quantitative PCR was used to assess each library, and an equimolar six-plex pool was produced for enrichment with Agilent SureSelect and a modified protocol. Before enrichment, the six-plex pool was single-stranded. Furthermore, 1.2 μg of pooled DNA with 5 μl (100 ng) of custom baits was used for enrichment, with blocking oligonucleotides specific for SOLiD sequencing libraries and 24-hour hybridization. This was the first targeted capture of a multiplex library for SOLiD sequencing, and this protocol has not been subsequently pursued. Alternative methods have been demonstrated to reduce the noise associated with bar coding and enrichment. Sequencing was performed on a SOLiD 3 instrument with one quadrant on a single sequencing slide, generating singleton 50-mer reads.

Sequence analysis

The bioinformatic decision tree for detecting and genotyping disease mutations was predicated on experience with detection and genotyping of variants in next-generation genome and chromosome sequences (20, 21, 33, 34) (Fig. 4). Briefly, SBS sequences were aligned to the National Center for Biotechnology Information (NCBI) reference human genome sequence (version 36.3) with GSNAP and scored by rewarding identities (+1) and penalizing mismatches (–1) and indels [−1–log(indel – length)]. Alignments were retained if covering >95% of the read and scoring >78% of maximum. Variants were detected with Alpheus with stringent filters (>14% and >10 reads calling variants and average quality score >20). Allele frequencies of 14 to 86% were designated heterozygous and >86%

homozygous. Reference genotypes of SNPs and CNVs mapping within targets were obtained with Illumina Omni1-Quad arrays and GenomeStudio 2010.1. Indel genotypes were confirmed by genomic PCR of <600-bp flanking variants and Sanger sequencing.

SBL sequence data analysis was performed with BioScope v1.2. Fifty nucleotide reads were aligned to NCBI genome build 36.3 with a seed and extend approach (max-mapping). A 25-nt seed with up to two mismatches is first aligned to the reference. Extension can proceed in both directions, depending on the footprint of the seed within the read. During extension, each base match receives a score of +1, whereas mismatches get a default score of −2. The alignment with the highest mapping quality value is chosen as the primary alignment. If two or more alignments have the same score, then one of them is randomly chosen as the primary alignment. SNPs were called with the BioScope diBayes algorithm at medium stringency setting (61). diBayes is a Bayesian algorithm that incorporates position and probe errors, as well as color quality value information for SNP calling. Reads with mapping quality of <8 were discarded by diBayes. A position must have at least 2× or 3× coverage to call a homozygous or heterozygous SNP, respectively. The BioScope small indel pipeline was used with default settings and calls insertions of size ≤3 nt and deletions of size ≤11 nt. In comparisons with SBS, SNP and indel calls were further restricted to positions where at least 4 or 10 reads called a variant.

Indel confirmation

PCR primers were designed to amplify 100 to 300 nt upstream and downstream of each variant or indel with PrimerQuest (Integrated DNA Technologies). Targeted regions were amplified from 100 ng of genomic DNA, and resultant PCR amplicons were analyzed for predicted size by LCGX (Caliper Life Sciences). Amplicons of appropriate size were Sanger-sequenced in both the forward and the reverse directions with the same primers used for PCR amplification. Analysis was performed with the Mutation Surveyor (SoftGenetics) software package.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank M. Chandler and M. Spain, who envisioned universal preconception screening, and the many physicians and geneticists who refined the concept and candidate disease list, particularly H. H. Ropers and C. J. Saunders. This work is dedicated to Christiane. *A deo lumen, ab amicis auxilium*.

Funding: This work was funded by grants from the Beyond Batten Disease Foundation and NIH (RR016480 to F.D.S.), and by in-kind support from Illumina Inc., Life Technologies, and British Airways PLC.

REFERENCES AND NOTES

1. Myrianthopoulos NC, Aronson SM. Population dynamics of Tay-Sachs disease. I. Reproductive fitness and selection. Am J Hum Genet. 1966; 18:313–327. [PubMed: 5945951]
2. Kaback, MM. Hexosaminidase A deficiency. In: Pagon, RA.; Bird, TC.; Dolan, CR.; Stephens, K., editors. GeneReviews. University of Washington; Seattle: 1993.

3. Mitchell JJ, Capua A, Clow C, Scriver CR. Twenty-year outcome analysis of genetic screening programs for Tay-Sachs and β-thalassemia disease carriers in high schools. Am J Hum Genet. 1996; 59:793–798. [PubMed: 8808593]

4. Kronn D, Jansen V, Ostrer H. Carrier screening for cystic fibrosis, Gaucher disease, and Tay-Sachs disease in the Ashkenazi Jewish population: The first 1000 cases at New York University Medical Center, New York NY. Arch Intern Med. 1998; 158:777–781. [PubMed: 9554684]

5. Kaback MM. Population-based genetic screening for reproductive counseling: The Tay-Sachs disease model. Eur J Pediatr. 2000; 159:S192–S195. [PubMed: 11216898]

6. Costa T, Scriver CR, Childs B. The effect of Mendelian disease on human health: A measurement. Am J Med Genet. 1985; 21:231–242. [PubMed: 4014310]

7. Kumar P, Radhakrishnan J, Chowdhary MA, Giampietro PF. Prevalence and patterns of presentation of genetic disorders in a pediatric emergency department. Mayo Clin Proc. 2001; 76:777–783. [PubMed: 11499815]

8. Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine. Johns Hopkins University; Baltimore, MD: http://www.ncbi.nlm.nih.gov/omim [accessed 11 December 2010]

9. ACOG Committee on Genetics. ACOG Committee Opinion No. 442: Preconception and prenatal carrier screening for genetic diseases in individuals of Eastern European Jewish descent. Obstet Gynecol. 2009; 114:950–953. [PubMed: 19888064]

10. ACOG Committee on Genetics. ACOG Committee Opinion No. 338: Screening for fragile X syndrome. Obstet Gynecol. 2006; 107:1483–1485. [PubMed: 16738187]

11. ACOG Committee on Genetics, ACOG Committee. Opinion No. 325 December 2005. Update on carrier screening for cystic fibrosis. Obstet Gynecol. 2005; 106:1465–1468. [PubMed: 16319281]

12. Grody WW, Cutting GR, Klinger KW, Richards CS, Watson MS, Desnick RJ. Laboratory standards and guidelines for population-based cystic fibrosis carrier screening. Genet Med. 2001; 3:149–154. [PubMed: 11280952]

13. Board of Directors of the American College of Medical Genetics. Position Statement on Carrier Testing for Canavan Disease. Jan 10. 1998 http://www.acmg.net/StaticContent/StaticPages/Canavan.pdf

14. Watson MS, Lloyd-Puryear MA, Mann MY, Rinaldo P, Howell RR. Main report. Genet Med. 2006; 8:12S–252S.

15. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009; 27:182–189. [PubMed: 19182786]

16. Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, Kotsopoulos SK, Samuels ML, Hutchison JB, Larson JW, Topol EJ, Weiner MP, Harismendy O, Olson J, Link DR, Frazer KA. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol. 2009; 27:1025–1031. [PubMed: 19881494]

17. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–276. [PubMed: 19684571]

18. Hedges DJ, Burges D, Powell E, Almonte C, Huang J, Young S, Boese B, Schmidt M, Pericak-Vance MA, Martin E, Zhang X, Harkins TT, Züchner S. Exome sequencing of a multigenerational human pedigree. PLoS One. 2009; 4:e8232. [PubMed: 20011588]

19. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet. 2010; 42:30–35. [PubMed: 19915526]

20. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Lee S, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park

WY, Kim H, Church GM, Lee C, Kingsmore SF, Seo JS. A highly annotated whole-genome sequence of a Korean individual. Nature. 2009; 460:1011–1015. [PubMed: 19587683]

21. Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, Miller NA, Zhang L, Farmer AD, Bell CJ, Kim RW, May GD, Woodward JE, Caillier SJ, McElroy JP, Gomez R, Pando MJ, Clendenen LE, Ganusova EE, Schilkey FD, Ramaraj T, Khan OA, Huntley JJ, Luo S, Kwok PY, Wu TD, Schroth GP, Oksenberg JR, Hauser SL, Kingsmore SF. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature. 2010; 464:1351–1356. [PubMed: 20428171]
22. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010; 328:636–639. [PubMed: 20220176]
23. Population-based carrier screening for single gene disorders: Lessons learned and new opportunities. Feb 6–7. 2008 http://www.genome.gov/27026048
24. Castellani C, Picci L, Tamanini A, Girardi P, Rizzotti P, Assael BM. Association between carrier screening and incidence of cystic fibrosis. JAMA. 2009; 302:2573–2579. [PubMed: 20009057]
25. Hale JE, Parad RB, Comeau AM. Newborn screening showing decreasing incidence of cystic fibrosis. N Engl J Med. 2008; 358:973–974. [PubMed: 18305279]
26. Maddalena, A.; Bale, S.; Das, S.; Grody, W.; Richards, S. American College of Medical Genetics Standards and Guidelines for Clinical Genetics Laboratories, Technical standards and guidelines: Molecular genetic testing for ultra-rare disorders. http://www.acmg.net/Pages/ACMG_Activities/stds-2002/URD.htm
27. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. The Human Gene Mutation Database: 2008 update. Genome Med. 2009; 1:13. [PubMed: 19348700]
28. GeneTests: Medical Genetics Information Resource (database online). University of Washington; Seattle: 1993–2010. http://www.genetests.org [[accessed 11 August 2010].]
29. Srinivasan BS, Evans EA, Flannick J, Patterson AS, Chang CC, Pham T, Young S, Kaushal A, Lee J, Jacobson JL, Patrizio P. A universal carrier test for the long tail of Mendelian disease. Reprod Biomed Online. 2010; 21:537–551. [PubMed: 20729146]
30. Summerer D, Hevroni D, Jain A, Oldenburger O, Parker J, Caruso A, Stähler CF, Stähler PF, Beier M. A flexible and fully integrated system for amplification, detection and genotyping of genomic DNA targets based on microfluidic oligonucleotide arrays. N Biotechnol. 2010; 27:149–155. [PubMed: 20359559]
31. Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M, Bicknell D, Bodmer WF, Davis RW, Ji H. Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci USA. 2007; 104:9387–9392. [PubMed: 17517648]
32. Fan HC, Quake SR. Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. PLoS One. 2010; 5:e10439. [PubMed: 20454671]
33. Sugarbaker DJ, Richards WG, Gordon GJ, Dong L, De Rienzo A, Maulik G, Glickman JN, Chirieac LR, Hartman ML, Taillon BE, Du L, Bouffard P, Kingsmore SF, Miller NA, Farmer AD, Jensen RV, Gullans SR, Bueno R. Transcriptome sequencing of malignant pleural mesothelioma tumors. Proc Natl Acad Sci USA. 2008; 105:3521–3526. [PubMed: 18303113]
34. Miller NA, Kingsmore SF, Farmer A, Langley RJ, Mudge J, Crow JA, Gonzalez AJ, Schilkey FD, Kim RJ, van Velkinburgh J, May GD, Black CF, Myers MK, Utsey JP, Frost NS, Sugarbaker DJ, Bueno R, Gullans SR, Baxter SM, Day SW, Retzel EF. Management of high-throughput DNA sequencing projects. Alpheus J Comput Sci Syst Biol. 2008; 1:132–148. [PubMed: 20151039]
35. Blanch L, Weber B, Guo XH, Scott HS, Hopwood JJ. Molecular defects in Sanfilippo syndrome type A. Hum Mol Genet. 1997; 6:787–791. [PubMed: 9158154]
36. Zhong N, Martiniuk F, Tzall S, Hirschhorn R. Identification of a missense mutation in one allele of a patient with Pompe disease, and use of endonuclease digestion of PCR-amplified RNA to demonstrate lack of mRNA expression from the second allele. Am J Hum Genet. 1991; 49:635–645. [PubMed: 1652892]

37. Zhong N, Wisniewski KE, Kaczmarski AL, Ju W, Xu WM, Xu WW, Mclendon L, Liu B, Kaczmarski W, Sklower Brooks SS, Brown WT. Molecular screening of Batten disease: Identification of a missense mutation (E295K) in the CLN3 gene. Hum Genet. 1998; 102:57–62. [PubMed: 9490299]
38. Wigderson M, Firon N, Horowitz Z, Wilder S, Frishberg Y, Reiner O, Horowitz M. Characterization of mutations in Gaucher patients by cDNA cloning. Am J Hum Genet. 1989; 44:365–377. [PubMed: 2464926]
39. Charache S, Jacobson R, Brimhall B, Murphy EA, Hathaway P, Winslow R, Jones R, Rath C, Simkovich J. Hb Potomac (101 Glu replaced by Asp): Speculations on placental oxygen transport in carriers of high-affinity hemoglobins. Blood. 1978; 51:331–338. [PubMed: 563749]
40. Gibbs RA, Nguyen PN, McBride LJ, Koepf SM, Caskey CT. Identification of mutations leading to the Lesch–Nyhan syndrome by automated direct DNA sequencing of *in vitro* amplified cDNA. Proc Natl Acad Sci USA. 1989; 86:1919–1923. [PubMed: 2928313]
41. Gencic S, Abuelo D, Ambler M, Hudson LD. Pelizaeus-Merzbacher disease: An X-linked neurologic disorder of myelin metabolism with a novel mutation in the gene encoding proteolipid protein. Am J Hum Genet. 1989; 45:435–442. [PubMed: 2773936]
42. Mimault C, Giraud G, Courtois V, Cailloux F, Boire JY, Dastugue B, Boespflug-Tanguy O. Proteolipoprotein gene analysis in 82 patients with sporadic Pelizaeus-Merzbacher disease: Duplications, the major cause of the disease, originate more frequently in male germ cells but point mutations do not. The Clinical European Network on Brain Dysmyelinating Disease. Am J Hum Genet. 1999; 65:360–369. [PubMed: 10417279]
43. Fisher KJ, Aronson NN Jr. Characterization of the mutation responsible for aspartylglucosaminuria in three Finnish patients. Amino acid substitution $Cys^{163}\rightarrow$Ser abolishes the activity of lysosomal glycosylasparaginase and its conversion into subunits. J Biol Chem. 1991; 266:12105–12113. [PubMed: 1904874]
44. Cleaver JE, Thompson LH, Richardson AS, States JC. A summary of mutations in the UV-sensitive disorders: Xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy. Hum Mutat. 1999; 14:9–22. [PubMed: 10447254]
45. Epeldegui M, Hung YP, McQuay A, Ambinder RF, Martínez-Maza O. Infection of human B cells with Epstein-Barr virus results in the expression of somatic hypermutation-inducing molecules and in the accrual of oncogene mutations. Mol Immunol. 2007; 44:934–942. [PubMed: 16730063]
46. Bobadilla JL, Macek M Jr, Fine JP, Farrell PM. Cystic fibrosis: A worldwide analysis of *CFTR* mutations—correlation with incidence data and application to screening. Hum Mutat. 2002; 19:575–606. [PubMed: 12007216]
47. Hu H, Wrogemann K, Kalscheuer V, Tzschach A, Richard H, Haas SA, Menzel C, Bienek M, Froyen G, Raynaud M, Von Bokhoven H, Chelly J, Ropers H, Chen W. Mutation screening in 86 known X-linked mental retardation genes by droplet-based multiplex PCR and massive parallel sequencing. Hugo J. 2009; 3:41–49. [PubMed: 21836662]
48. Motulsky AG. Screening for genetic diseases. N Engl J Med. 1997; 336:1314–1316. [PubMed: 9113938]
49. Barlow-Stewart K, Burnett L, Proos A, Howell V, Huq F, Lazarus R, Aizenberg H. A genetic screening programme for Tay-Sachs disease and cystic fibrosis for Australian Jewish high school students. J Med Genet. 2003; 40:e45. [PubMed: 12676918]
50. D'Souza G, McCann C, Hiedrick J, Fairley CL, Nagel HL, Kushner JD, Kessel R. Tay-Sachs disease carrier screening: A 21-year experience. Genet Test. 2000; 4:257–263. [PubMed: 11142756]
51. McCabe L. Efficacy of a targeted genetic screening program for adolescents. Am J Hum Genet. 1996; 59:762–763. [PubMed: 8808589]
52. Lau YL, Chan LC, Chan YY, Ha SY, Yeung CY, Waye JS, Chui DH. Prevalence and genotypes of α and β-thalassemia carriers in Hong Kong—implications for population screening. N Engl J Med. 1997; 336:1298–1301. [PubMed: 9113933]
53. Barlow-Stewart K, Keays D. Genetic discrimination in Australia. J Law Med. 2001; 8:250–262.

54. Kaback MM. The control of genetic disease by carrier screening and antenatal diagnosis: Social, ethical, and medicolegal issues. Birth Defects Orig Artic Ser. 1982; 18:243–254. [PubMed: 7159734]

55. Baker VL, Rone HM, Adamson GD. Genetic evaluation of oocyte donors: Recipient couple preferences and outcome of testing. Fertil Steril. 2008; 90:2091–2098. [PubMed: 18249390]

56. McConkey, E. Human Genetics: The Molecular Revolution. Sudbury, MA., editor. Vol. 1. Jones & Bartlett; 1993.

57. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M, Sagreiya H, Whaley R, Knowles JW, Chou MF, Thakuria JV, Rosenbaum AM, Zaranek AW, Church GM, Greely HT, Quake SR, Altman RB. Clinical assessment incorporating a personal genome. Lancet. 2010; 375:1525–1535. [PubMed: 20435227]

58. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J. The diploid genome sequence of an Asian individual. Nature. 2008; 456:60–65. [PubMed: 18987735]

59. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008; 452:872–876. [PubMed: 18421352]

60. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, Dahl F, Fernandez A, Staker B, Pant KP, Baccash J, Borcherding AP, Brownley A, Cedeno R, Chen L, Chernikoff D, Cheung A, Chirita R, Curson B, Ebert JC, Hacker CR, Hartlage R, Hauser B, Huang S, Jiang Y, Karpinchyk V, Koenig M, Kong C, Landers T, Le C, Liu J, McBride CE, Morenzoni M, Morey RE, Mutch K, Perazich H, Perry K, Peters BA, Peterson J, Pethiyagoda CL, Pothuraju K, Richter C, Rosenbaum AM, Roy S, Shafto J, Sharanhovich U, Shannon KW, Sheppy CG, Sun M, Thakuria JV, Tran A, Vu D, Zaranek AW, Wu X, Drmanac S, Oliphant AR, Banyai WC, Martin B, Ballinger DG, Church GM, Reid CA. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010; 327:78–81. [PubMed: 19892942]

61. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De La Vega FM, Blanchard AP. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 2009; 19:1527–1541. [PubMed: 19546169]

62. Emery, AEH. Duchenne muscular dystrophy. No 15. In: Motulsky, AG.; Harper, PS.; Bobrow, M.; Scriver, C., editors. Oxford Monographs on Medical Genetics. Oxford Univ. Press; Oxford: 1988.

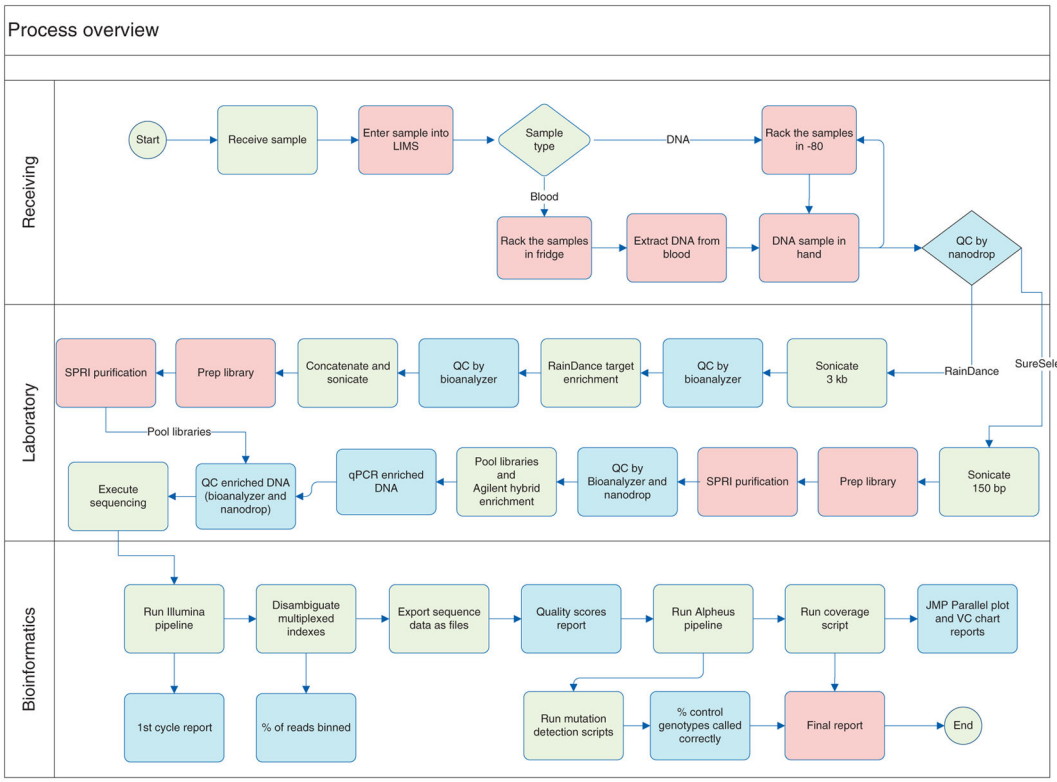

Fig. 1.

Workflow of the comprehensive carrier screening test. Workflow shows receiving samples and DNA extraction, target enrichment from DNA samples, multiplexed sequencing library preparation, NGS, and bioinformatic analysis. (The bioinformatic decision tree is shown in fig. S4.)

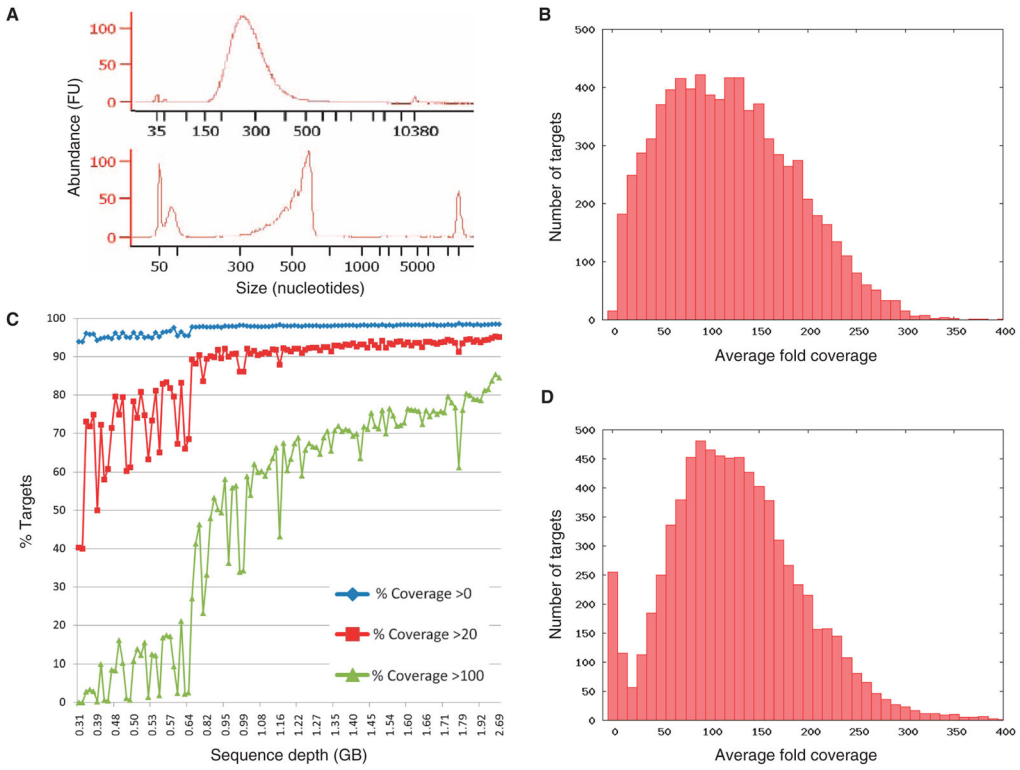

Fig. 2.

Analytic metrics of multiplexed carrier testing by NGS. (**A**) Chromatograms of size distributions of sequencing libraries after target enrichment. Top: Target enrichment by hybrid capture. Bottom: Target enrichment by microdroplet PCR. Size markers are shown at 40 and 8000 nt. FU, fluorescence units. (**B**) Frequency distribution of target coverage after hybrid selection and 1.75 Gb of singleton 50-mer Illumina GAIIx SBS of sample NA13675. Aligned sequences had a quality score of >25. (**C**) Target coverage as a function of depth of sequencing across 104 samples and six experiments. (**D**) Frequency distribution of target coverage after microdroplet PCR and 1.49 Gb of singleton 50-mer SBS of sample NA20379. Aligned sequences had a quality score of >25.

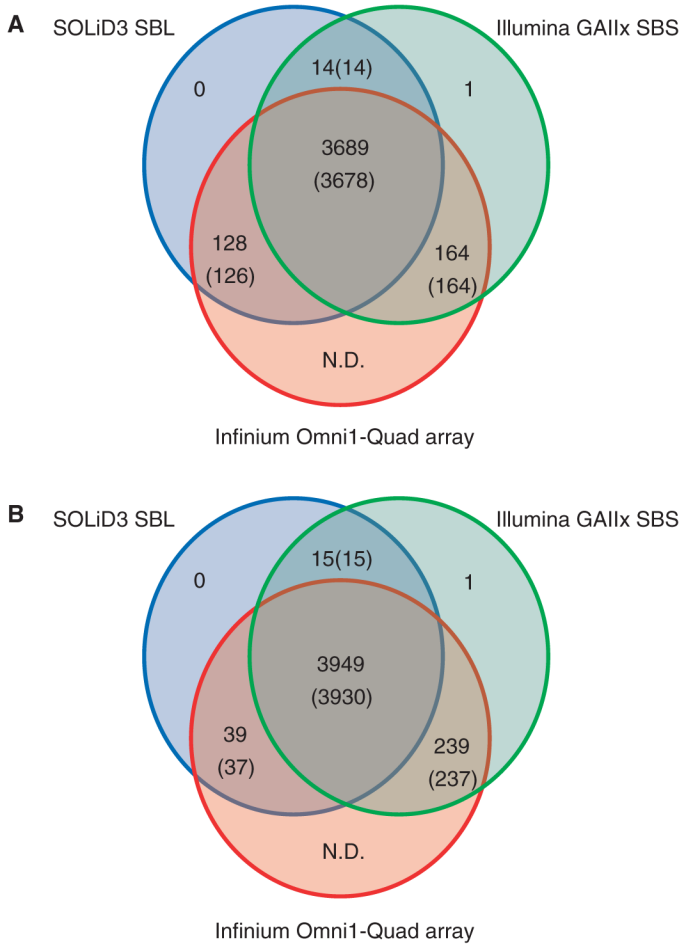

Fig. 3.

Venn diagrams of specificity of on-target SNP calls and genotypes in six samples. Target nucleotides were enriched by hybrid selection and sequenced by Illumina GAIIx SBS and SOLiD3 SBL at sixfold multiplexing. The samples were also genotyped with Infinium Omni1-Quad SNP arrays. (**A**) Comparison of SNP calls and genotypes obtained by SBS, SBL, and arrays at nucleotides surveyed by all three methods. SNPs were called if present in >10 uniquely aligning SBS reads, >14% of reads, and with average quality score of >20. Heterozygotes were identified if present in 14 to 86% of reads. Numbers refer to SNP calls. Numbers in brackets refer to SNP genotypes. (**B**) Comparison of SNP calls and genotypes obtained by SBS, SBL, and arrays. SNPs were called if present in more than four uniquely aligning SBS reads, >14% of reads, and with average quality score of >20. Heterozygotes were identified if present in 14 to 86% of reads.

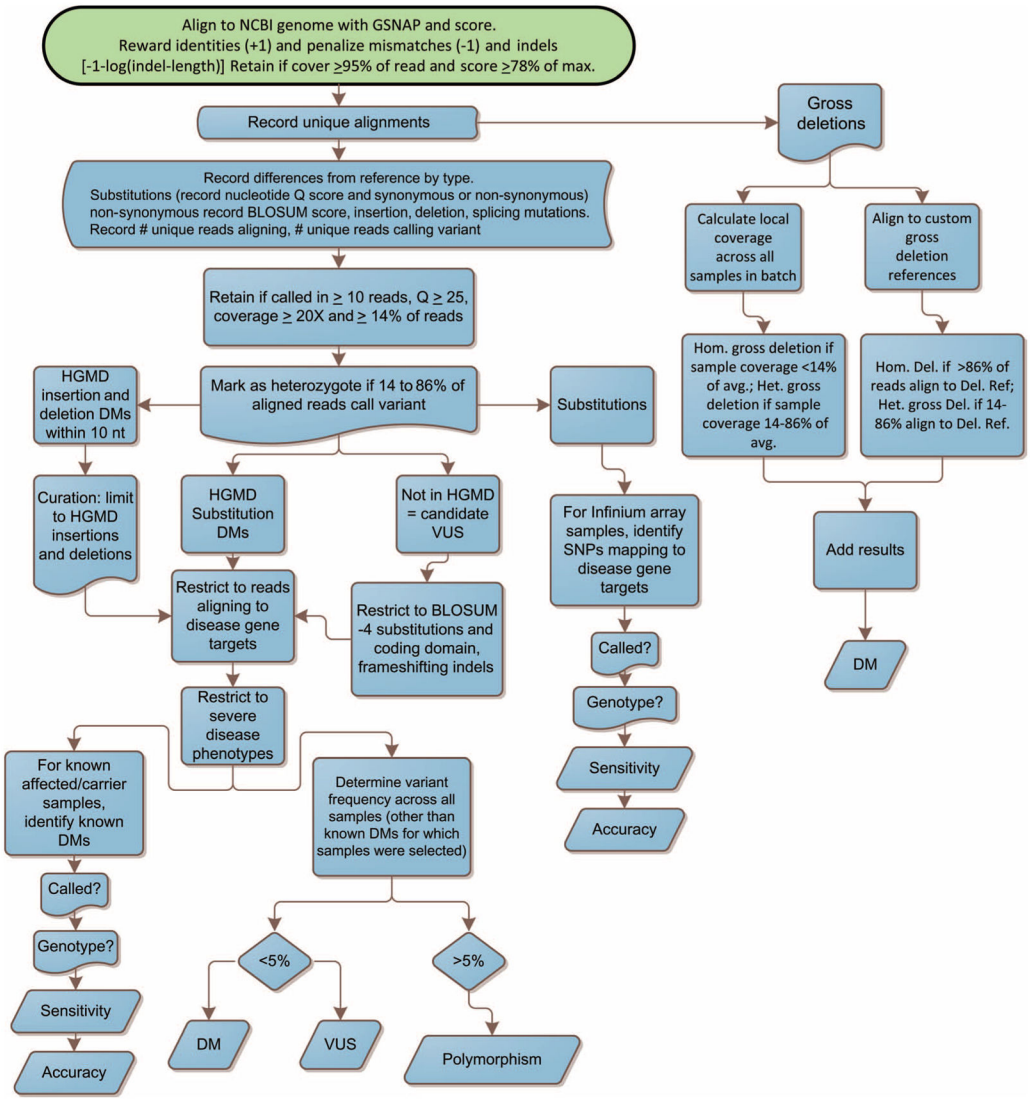

Fig. 4.

Decision tree to classify sequence variation and evaluate carrier status. After reads were aligned to references, substitution, insertion, and deletion events and their associated quality metrics were recorded. Variants were classified as heterozygous or homozygous and annotated by comparison with mutation databases. Variants not in the mutation databases were evaluated for putative functional consequence and were retained as disease mutations if predicted to result in protein truncation. Variants with a frequency of <5% among all samples and that were known to cause a disease phenotype or loss of protein function and that were only found as homozygous in affected individuals were retained and reported.

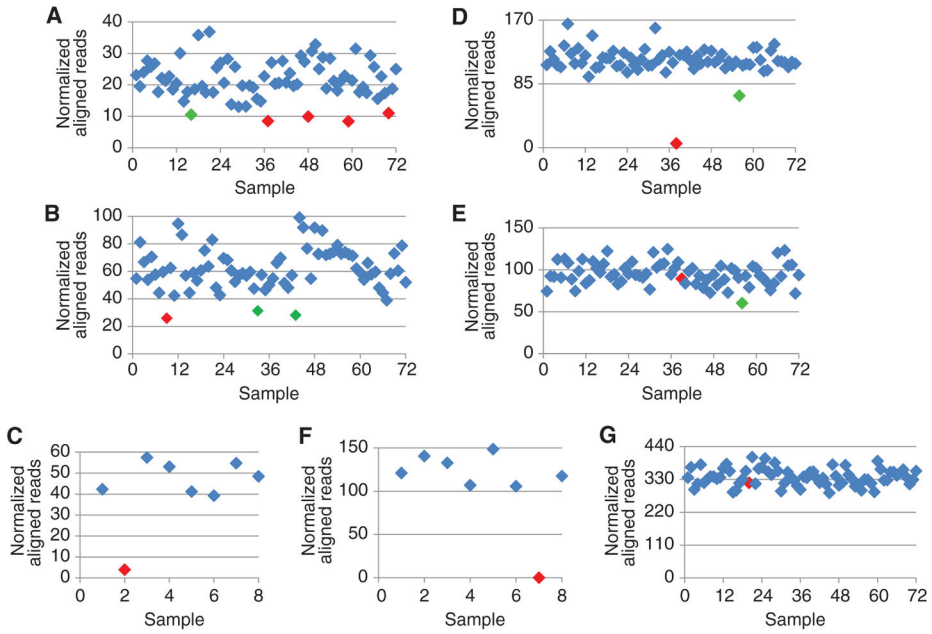

Fig. 5.

Detection of gross deletion mutations by local reduction in normalized aligned reads. (**A**) Deletion of CLN3 introns 6 to 8, 966bpdel, exons7-8del and fs, chr16:28405752_28404787del in four known compound heterozygotes (NA20381, NA20382, NA20383, and NA20384; red diamonds) and one undescribed carrier (NA00006; green diamond) among 72 samples sequenced. (**B**) Heterozygous deletion in HBA1 (chr16:141620_172294del, 30,676-bp deletion from 5′ of ζ2 to 3′ of θ1 in ALU regions) in one known (NA10798; red diamond; normalized coverage, 26; mean normalized coverage, 61.9 ± 15.2) and two undescribed carriers [NA19193 (normalized coverage, 28) and NA01982 (normalized coverage, 31); green diamonds] among 72 samples. Heterozygous deletion in NA10798 was confirmed by array hybridization. (**C**) Known homozygous deletion of exons 7 and 8 of SMN1 in one of eight samples (NA03813; red diamond). (**D**) Detection of a gross deletion that is a cause of Duchenne muscular dystrophy (OMIM #310200, DMD exons 51 to 55 del, chrX:31702000_31555711del) by reduction in normalized aligned reads at chrX:31586112. Among 72 samples, one (NA04364; red diamond) was from an affected male, and another (NA18540, a female JPT/HAN HapMap sample) was determined to carry a deletion that extends to at least chrX:31860199 [see (E)]. (**E**) An undescribed heterozygous deletion of DMD 3′ exon 44–3′ exon 50 (chrX: 32144956-31702228del) in NA18540 (green diamond), a JPT/HAN HapMap sample. This deletion extends from at least chrX:31586112 to chrX:31860199 [see (D)]. Sample NA05022 (red diamond) is the uncharacterized mother of an affected son with 3′ exon 44–3′ exon 50 del, chrX:32144956-31702228del. Given the absence of the mutation in the mother, it likely occurred de novo in the son, as observed in one-third of DMD patients (62). (**F**) Hemizygous deletion in PLP1 exons3_4, c.del349_495del, chrX:102928207_102929424del in one (NA13434; red diamond) of eight samples. (**G**) Absence of gross deletion CG984340 (ERCC6 exon 9, c.1993_2169del, 665_723del, exon 9 del, chr10:50360915_50360739del)

in 72 DNA samples. The sample in red (NA01712) was incorrectly annotated to be a compound heterozygote with CG984340 on the basis of cDNA sequencing.

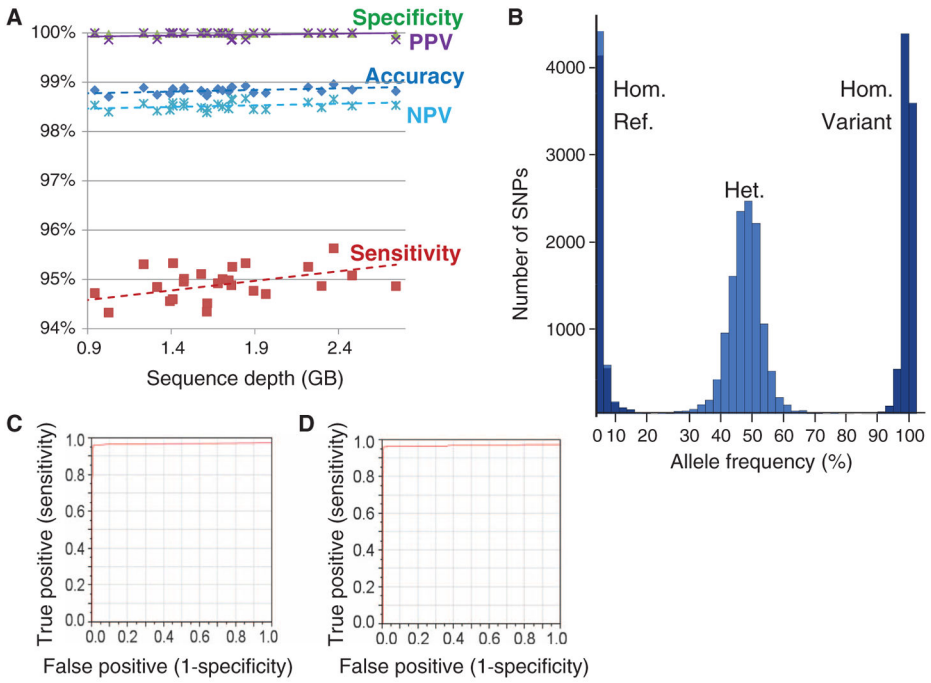

Fig. 6.

Clinical metrics of multiplexed carrier testing by NGS. (**A**) Comparison of 92,128 SNP genotypes by array hybridization with those obtained by target enrichment, SBS, and a bioinformatic decision tree in 26 samples. SNPs were called if present in >10 uniquely aligning reads, >14% of reads, and average quality score of >20. Heterozygotes were identified if present in 14 to 86% of reads. TP = SNP called and genotyped correctly. TN = reference genotype called correctly. FN = SNP genotype undercall. FP = SNP genotype overcall. Accuracy = (TP + TN)/(TP + FN + TN + FP). Sensitivity = TP/(TP + FN). Specificity = TN/(TN + FP). Positive predictive value (PPV) = TP/(TP + FP). Negative predictive value (NPV) = TN/(TN + FN). (**B**) Distribution of allele frequencies of SNP calls by hybrid capture and SBS in 26 samples. Light blue, heterozygotes by array hybridization. (**C**) Receiver operating characteristic (ROC) curve of sensitivity and specificity of SNP genotypes by hybrid capture and SBS in 26 samples (when compared with array-based genotypes). Genomic regions with less than 20× coverage were excluded. Upon varying the number of reads calling the SNP, the area under the curve (AUC) was 0.97. (**D**) ROC curve of SNP genotypes by hybrid capture and SBS in 26 samples. Genomic regions with less than 20× coverage were excluded. Upon varying the percent reads calling the SNP, AUC was 0.97.

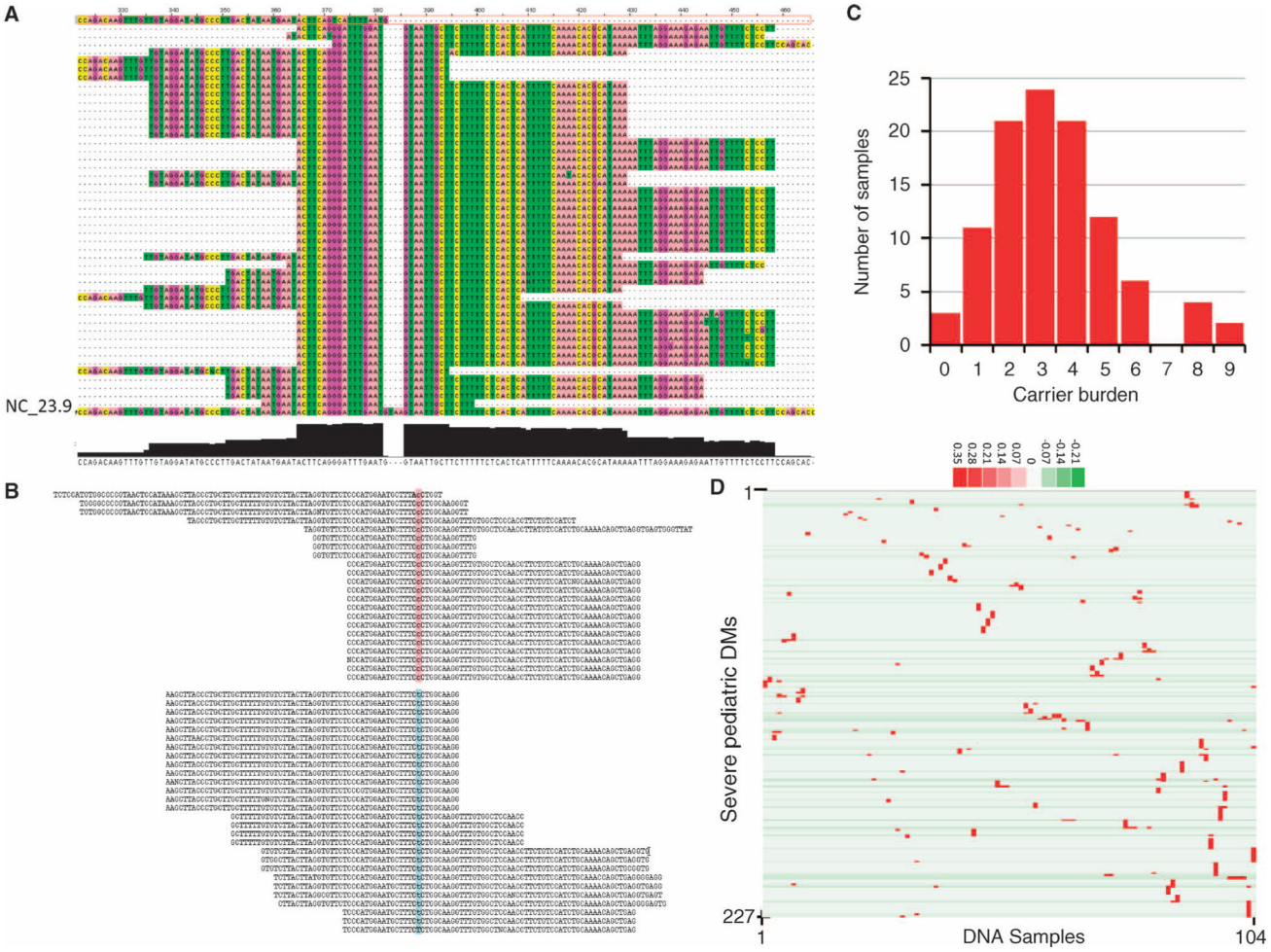

Fig. 7.

Disease mutations and estimated carrier burden in 104 DNA samples. (**A**) Sample NA07092, from an affected male with X-linked recessive Lesch-Nyhan syndrome (OMIM #300322), had been characterized as a deletion of HPRT1 exon 8 by cDNA sequencing (19), but has an explanatory splicing mutation (intron 8, IVS8+1_4delGTAA, chrX: 133460381_133460384delGTAA). (**B**) Sample NA09545, from an affected male with X-linked recessive Pelizaeus-Merzbacher disease (PMD; OMIM #312080), characterized as a substitution disease mutation [PLP1 exon 5, c.767C>T, P215S (20)], also featured PLP1 gene duplication [which is reported in 62% of sporadic PMD (21)]. (**C**) Distribution of carrier burden of severe pediatric diseases among 104 DNA samples. (**D**) Ward hierarchical clustering of 227 severe pediatric disease mutations in 104 DNA samples.

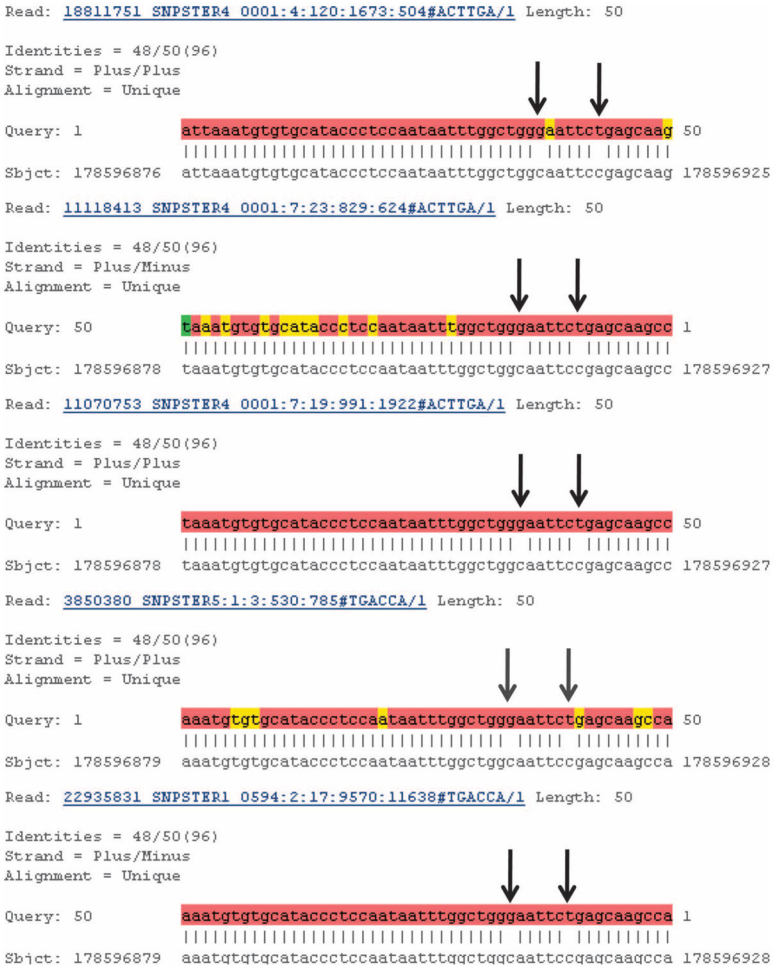

Fig. 8.

Five reads from NA202057 showing AGA exon 4, c.488G>C, C163S, chr4:178596912G>C and exon 4, c.482G>A, R161Q, chr4:178596918G>A (black arrows). One hundred and ninety-three of 400 reads contained these substitution disease mutations (CM910010 and CM910011). The top lines of doublets are Illumina GAIIx 50-nt reads. The bottom lines are NCBI reference genome, build 36.3. Colors represent quality (*Q*) scores of each nucleotide: red, >30; orange, 20 to 29; green, 10 to 19. Reads aligned uniquely to these coordinates.

Table 1

Sequencing, alignment, and coverage statistics for target enrichment and sequencing platforms.

Sample set	Enrichment method	Sequencing method	Multiplexing	Read length (nt)	Quality score*	Total reads ± %CV*†	% uniquely aligning reads*	Total nucleotides*	Aligning depth*	% nt on target ± %CV*	Fold enrichment*	% 0× coverage*	% ≥20× coverage*	Coverage ± %CV*	Pearson's coefficient‡
1 (*n* = 12)	SureSelect	GAIIx	12	50	30	9,952,972.5 ± 21	94	497,648,625	225	13.7 ± 3	214	4.83	61	27 ± 21	0.28
2 (*n* = 12)	SureSelect	GAIIx	12	50	30	10,127,721 ± 16	95	506,386,025	234	23.0 ± 2	358	3.66	80	50 ± 16	0.19
1 + 2 (*n* = 24)	RainDance	GAIIx	12	50	36	9,412,698 ± 30	97	470,634,900	196	29.6 ± 5	462	5.46	86	52.5 ± 33	0.23
1 + 2 (*n* = 12)	RainDance	GAIIx	12	50	31	12,807,392 ± 17	96	640,369,600	277	22.2 ± 7	346	4.62	88	56 ± 12	0.27
3 (*n* = 6)	SureSelect	GAIIx	6	50	30	19,711,735 ± 34	95	985,586,750	463	17.4 ± 3	273	1.80	86	76 ± 30	0.14
3 (*n* = 6)	SureSelect	SOLiD 3	6	50	24	16,506,076 ± 5	82	825,303,800	310	19.5 ± 7	304	6.08	79	58 ± 7	0.24
4 (*n* = 72)	SureSelect 2	HiSeq	8	149§	42§	9,273,596 ± 24	98	1,390,464,487	495	31.7 ± 4	494	2.33	92	152 ± 26	0.02
5 (*n* = 8)	SureSelect	HiSeq	8	149§	41§	9,861,765 ± 35	97	1,493,946,141	517	28.4 ± 4	442	2.25	93	139 ± 40	0.06

*Median value.

†Coefficient of variation (%).

‡Pearson's median skewness coefficient [3(mean – median)/SD].

§After assembly of forward and reverse 130-bp paired reads.