nature genetics

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng[1,10], Kati J Buckingham[2,10], Choli Lee[1], Abigail W Bigham[2], Holly K Tabor[2,3], Karin M Dent[4], Chad D Huff[5], Paul T Shannon[6], Ethylin Wang Jabs[7,8], Deborah A Nickerson[1], Jay Shendure[1] & Michael J Bamshad[1,2,9]

We demonstrate the first successful application of exome sequencing to discover the gene for a rare mendelian disorder of unknown cause, Miller syndrome (MIM%263750). For four affected individuals in three independent kindreds, we captured and sequenced coding regions to a mean coverage of 40× and sufficient depth to call variants at ~97% of each targeted exome. Filtering against public SNP databases and eight HapMap exomes for genes with two previously unknown variants in each of the four individuals identified a single candidate gene, *DHODH*, which encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway. Sanger sequencing confirmed the presence of *DHODH* mutations in three additional families with Miller syndrome. Exome sequencing of a small number of unrelated affected individuals is a powerful, efficient strategy for identifying the genes underlying rare mendelian disorders and will likely transform the genetic analysis of monogenic traits.

Rare monogenic diseases are of substantial interest because identification of their genetic bases provides important knowledge about disease mechanisms, biological pathways and potential therapeutic targets. However, to date, allelic variants underlying fewer than half of all monogenic disorders have been discovered. This is because the identification of allelic variants for many rare disorders is fundamentally limited by factors such as the availability of only a small number of affected individuals (cases) or families, locus heterogeneity, or substantially reduced reproductive fitness; each of these factors lessens the power of traditional positional cloning strategies and often restricts the analysis to a priori–identified candidate genes. In contrast, deep resequencing of all human genes for discovery of allelic variants could potentially identify the gene underlying any given rare monogenic disease. Massively parallel DNA sequencing technologies[1] have rendered the whole-genome resequencing of individual humans increasingly practical, but cost remains a key consideration. An alternative approach involves the targeted resequencing of all protein-coding subsequences (that is, the exome)[2–4], which requires ~5% as much sequencing as a whole human genome[2].

Sequencing of the exome, rather than the entire human genome, is well justified as an efficient strategy to search for alleles underlying rare mendelian disorders. First, positional cloning studies focused on protein-coding sequences have, when adequately powered, proven highly successful at identification of variants underlying monogenic diseases. Second, the clear majority of allelic variants known to underlie mendelian disorders disrupt protein-coding sequences[5]. Splice acceptor and donor sites represent an additional class of sequences that are enriched for highly functional variation and are therefore targeted here as well. Third, a large fraction of rare nonsynonymous variants in the human genome are predicted to be deleterious[6]. This contrasts with noncoding sequences, where variants are more likely to have neutral or weak effects on phenotypes, even in well-conserved noncoding sequences[7,8]. The exome therefore represents a highly enriched subset of the genome in which to search for variants with large effect sizes.

We recently showed how exome sequencing of a small number of affected, unrelated individuals could potentially be used to identify a causal gene underlying a monogenic disorder[2]. Specifically, we performed targeted enrichment of the exome by hybridization to programmable microarrays and then sequenced each enriched shotgun genomic library on an Illumina Genome Analyzer II. The exome was conservatively defined using the NCBI Consensus Coding Sequence (CCDS) database[9] (version 20080902), which covers approximately 164,000 noncontiguous regions over 27.9 Mb, of which 26.6 Mb were 'mappable' using 76-bp single-end reads. Approximately 96% of targeted, mappable bases comprising the exomes of eight HapMap individuals and four individuals with Freeman-Sheldon syndrome (FSS; MIM#193700) were successfully sequenced to high quality[2]. Using both dbSNP and HapMap exomes as filters to remove common variants, we showed that we could accurately identify the causal gene for FSS by exome sequencing alone. This effort demonstrated that low-cost, high-throughput technologies for deep resequencing have the potential to rapidly accelerate the discovery of allelic variants for rare

[1]Departments of Genome Sciences and [2]Pediatrics, University of Washington, Seattle, Washington, USA. [3]Treuman Katz Center for Pediatric Bioethics, Seattle Children's Hospital, Seattle, Washington, USA. [4]Departments of Pediatrics and [5]Human Genetics, University of Utah, Salt Lake City, Utah, USA. [6]Institute of Systems Biology, Seattle, Washington, USA. [7]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA. [8]Department of Pediatrics, Johns Hopkins University, Baltimore, Maryland, USA. [9]Seattle Children's Hospital, Seattle, Washington, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to M.J.B. (mbamshad@u.washington.edu) or J.S. (shendure@u.washington.edu).

Received 2 October; accepted 9 November; published online 13 November; corrected online 22 November 2009 (details online); doi:10.1038/ng.499

diseases. However, it provided only a proof of concept, as the causal gene for FSS had previously been identified[10]. A more recent report describes the application of exome sequencing to make an unanticipated genetic diagnosis of congenital chloride diarrhea[3].

To evaluate the effectiveness of this strategy with a mendelian condition of unknown cause, we sought to find the gene for a rare multiple malformation disorder named Miller syndrome[11], the cause of which has been intractable to standard approaches of discovery[12]. The clinical characteristics of Miller syndrome include severe micrognathia, cleft lip and/or palate, hypoplasia or aplasia of the posterior elements of the limbs, coloboma of the eyelids and supernumerary nipples (**Fig. 1a**,**b**). Miller syndrome has been hypothesized to be an autosomal recessive disorder. However, only three multiplex families, each consisting of two affected siblings born to unaffected, nonconsanguineous parents, have been described among a total of ~30 reported cases of Miller syndrome for which substantial clinical information is available[11,13–17]. Accordingly, there has been speculation that Miller syndrome is an autosomal dominant disorder[18] and the rare occurrence of affected siblings is the result of germline mosaicism. Although we thought it likely that Miller syndrome is recessive, we also considered a dominant model of inheritance.

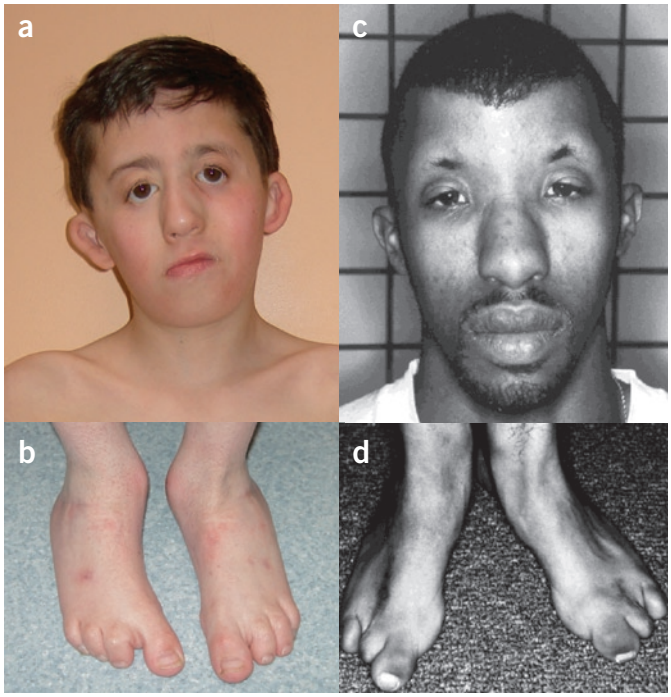

Figure 1 Clinical characteristics of an individual with Miller syndrome and an individual with methotrexate embryopathy. (**a**,**b**) A 9-year-old boy with Miller syndrome caused by mutations in *DHODH*. Facial anomalies (**a**) include cupped ears, coloboma of the lower eyelids, prominent nose, micrognathia and absence of the fifth digits of the feet (**b**). (**c**,**d**) A 26-year-old man with methotrexate embryopathy. Note the cupped ears, hypertelorism, sparse eyebrows and prominent nose (**c**) accompanied by absence of the fourth and fifth digits of the feet (**d**). **c** and **d** are reprinted with permission from ref. 30.

RESULTS

Exome sequencing identifies a candidate gene for Miller syndrome

We sequenced exomes in a total of two siblings with Miller syndrome (kindred 1 in **Table 1**) and two additional unrelated affected individuals (kindreds 2 and 3 in **Table 1**), totaling four affected individuals in three independent kindreds. An average of 5.1 Gb of sequence was generated per affected individual as single-end, 76-bp reads. After discarding reads that had duplicated start sites, we achieved ~40-fold coverage of the 26.6-Mb mappable, targeted exome defined by Ng *et al.*[2] (**Table 2**). About 97% of targeted bases were sufficiently covered to pass our thresholds for variant calling. To distinguish potentially pathogenic mutations from other variants, we focused only on nonsynonymous (NS) variants, splice acceptor and donor site mutations (SS), and short coding insertions or deletions (indels; I), anticipating that synonymous variants would be far less likely to be pathogenic. We also predicted that the variants responsible for Miller syndrome would be rare and therefore likely to be previously unidentified. A new variant was defined as one that did not exist in the datasets used for comparison, namely dbSNP129, exome data from eight HapMap individuals sequenced in our previous study[2], and both groups combined (**Table 1**).

Each sibling (A and B) in kindred 1 was found to have at least a single NS/SS/I variant in ~4,600 genes and two or more NS/SS/I variants in ~2,800 genes. In our dominant model, each sibling was required to have at least one new NS/SS/I variant in the same gene, and filtering these variants against dbSNP129 and eight HapMap exomes reduced the candidate gene pool ~40-fold compared to the full CCDS gene set. In our recessive model, each sibling was required to have at least two new NS/SS/I variants in the same gene, and the candidate pool was reduced >500-fold compared to the full CCDS gene set. Both siblings were predicted to share the causal variant for Miller syndrome, so we next considered candidate genes shared between them. Under our dominant model, this reduced the pool of candidate genes to 228, and under our recessive model, the number of candidate genes was reduced to 9.

Table 1 Direct identification of the gene for a mendelian disorder by exome resequencing

	Kindred 1-A		Kindred 1-B		Kindred 1 (A+B)		Kindreds 1+2		Kindreds 1+2+3	
Filter	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive	Dominant	Recessive
NS/SS/I	4,670	2,863	4,687	2,859	3,940	2,362	3,099	1,810	2,654	1,525
Not in dbSNP129	641	102	647	114	369	53	105	25	63	21
Not in HapMap 8	898	123	923	128	506	46	117	7	38	4
Not in either	456	31	464	33	228	9	26	1*	8	1*
Predicted damaging	204	6	204	12	83	1	5	0	2	0

Each cell indicates the number of genes with nonsynonymous (NS) variants, splice acceptor and donor site mutations (SS) and coding indels (I). Filtering either by requiring the presence of NS/SS/I variants in siblings (kindred 1 (A+B)) or of multiple unrelated individuals (columns) or by excluding annotated variants (rows) identifies 26 and 8 candidate genes under a dominant model and only a single candidate gene, *DHODH*, under a recessive model (light gray cells). Exclusion of mutations predicted to be benign using PolyPhen (row 5) increases sensitivity under a dominant model but excludes *DHODH* under a recessive model because a variant in kindred 1 is predicted to be benign. A single candidate gene is identified in kindred 1 under a recessive model and excluding benign mutations (dark gray cell), but this candidate is excluded in comparisons with unrelated cases of Miller syndrome. Mutations in this candidate, *DNAH5*, were found to cause a primary ciliary dyskinesia in kindred 1. The asterisk indicates that a second gene, *CDC27*, was also identified as a candidate gene, but this is due to the presence of multiple copies of a processed pseudogene that recurrently gave rise to a false positive signal in exome analyses.

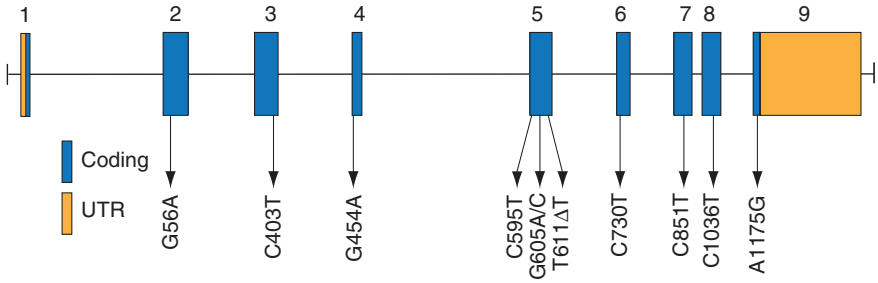

Figure 2 Genomic structure of the exons encoding the open reading frame of *DHODH*. *DHODH* is composed of nine exons that encode untranslated regions (UTR) (orange) and protein coding sequence (blue). Arrows indicate the locations of 11 different mutations found in 6 families with Miller syndrome.

To further exclude candidate genes containing nonpathogenic variants, we next compared the candidate genes from both siblings in kindred 1 to those in two unrelated individuals with Miller syndrome (kindreds 2 and 3). Using both dbSNP129 and the eight HapMap exomes as filters, comparison between the affected siblings in kindred 1 and the unrelated case in kindred 2 reduced the number of candidate genes to 26 under our dominant model. Under the autosomal recessive model, this comparison revealed that only a single gene, *DHODH*, which encodes the enzyme dihydroorotate dehydrogenase, was a shared candidate. Thus, comparison of exome data from two affected siblings and one unrelated affected individual was sufficient to identify *DHODH* as the sole candidate gene for Miller syndrome under our recessive model. Comparison between the siblings in kindred 1 and the unrelated cases in kindreds 2 and 3 reduced the number of candidate genes to eight under a dominant model, while retaining *DHODH* as the sole candidate under the recessive model.

We calculated a Bonferroni-corrected *P* value for the null hypothesis of seeing no deviation from the expected frequency of two variants in the same gene in three out of three unrelated, affected individuals over the ~17,000 genes in CCDS2008. Assuming all genes are of the same length and have the same mutation rate, the rate of new NS/SS/I variants per gene was 0.0309 (~526 new NS/SS/I variants per 17,000 genes). If the variants occur independently of one another, two variants occur in the same gene at a rate of $(0.0309)^2$, or 9.57×10^{-4}, so the *P* value is calculated as $((9.57 \times 10^{-4})^3 \times 17{,}000)$, or ~0.000015. Hence, even after correcting for searching across all genes, the result remains highly significant.

We also (i) examined the effect on the size of the candidate gene list when analyzing the exomes of affected individuals in various pairwise or three-way combinations and (ii) examined the potential consequences of genetic heterogeneity by relaxing selection criteria such that only a subset of the exomes of affected individuals were required to contain new variants in a given gene for it to be considered as a candidate gene (**Table 3**). Heterogeneity clearly increases the number of candidate genes that must be considered under any fixed number of exomes analyzed. However, this can likely be overcome by the inclusion of a greater number of cases with mutations in the same gene.

Most variants underlying rare mendelian diseases either affect highly conserved sequence and/or are predicted to be deleterious. Accordingly, we also sought to investigate to what extent the pool of candidate genes could be reduced by combining variant filtering with predictions of whether NS/SS/I variants were damaging. This strategy further reduced the pool of candidate genes for each of the comparisons made previously (**Table 1**). However, *DHODH* was not identified as a candidate under a recessive model in any of these comparisons. Review of predicted biophysical consequences of *DHODH* variants revealed that the effect of one variant, G605A, found in both siblings in kindred 1, was classified as benign. As a result, *DHODH* was eliminated from further consideration as a candidate under a recessive model in kindred 1 and in all subsequent comparisons. However, because the other variant found in kindred 1, G454A, was predicted to be damaging, as was every other new *DHODH* variant identified in this study, *DHODH* was still one of only two candidate genes for Miller syndrome under a dominant model in a comparison of kindreds 1, 2 and 3 (**Table 1**). Nevertheless, the recessive model was favored over the dominant model for Miller syndrome based on the observation that each case was a compound heterozygote for new *DHODH* mutations and five of six mutations were predicted to be damaging.

Combinatorial filtering supplemented by PolyPhen predictions initially identified a second candidate gene, *DNAH5*, in kindred 1 under a recessive model (**Table 1**). However, this candidate was excluded in subsequent comparisons to kindreds 2 and 3. *DNAH5* encodes a dynein heavy chain found in cilia, and mutations in *DNAH5* are a well-known cause of primary ciliary dyskinesia (PCD; MIM#608644), a disorder characterized by recurrent sinopulmonary infections, bronchiectasis and chronic lung disease. This was of particular interest because some of the clinical findings in the siblings in kindred 1 are unique among reported cases of Miller syndrome. Specifically, both siblings have recurrent lung infections, bronchiectasis and chronic obstructive pulmonary disease for which they have received medical management in a specialty clinic for individuals with cystic fibrosis. Accordingly, exome analysis revealed that both siblings in kindred 1 have, in addition to Miller syndrome, PCD due to mutations in *DNAH5*.

Sanger sequencing of implicated gene

To confirm that mutations in *DHODH* are responsible for Miller syndrome, we screened three additional unrelated kindreds (three simplex cases) and an affected sibling in kindred 2 by directed Sanger sequenc-

Table 2 Summary statistics for exome sequencing of four individuals with Miller syndrome

	Sequencing reads				Called coverage		
Kindred-sibling	Total	Uniquely mapping	Overlapping target	Nonduplicated	Mean coverage	Called bases	% of CCDS
1-A	62,974,440	52,854,115	25,267,592	17,872,660	36.85	25,720,216	97
1-B	72,539,306	61,940,123	40,335,280	21,971,509	44.24	25,825,104	97
2	63,839,828	55,022,098	29,987,198	19,686,779	40.31	25,790,427	97
3	68,690,600	57,970,901	36,180,596	19,649,281	39.81	25,617,361	96

The total number of unpaired 76-bp sequencing reads per individual is reported (total), along with the number that map uniquely to the human genome (uniquely mapping, Maq map score > 0), the number that overlap at least one base of the target space (overlapping target) and the number left after removing reads with duplicate start sites (nonduplicated). Mean coverage over the whole of CCDS2008 is also given. Called bases refer to bases passing quality and coverage thresholds (Maq consensus quality ≥20 and read depth ≥8×). % of CCDS refers to the fraction of the mappable 26.6 Mb of CCDS2008 (that is, masked for poorly mappable coordinates, as described in Ng *et al.*[2]) that is called in each exome.

Table 3 Number of candidate genes identified based on different filtering strategies

	Number of affected exomes			Subsets of 3 exomes		Subsets of all 4 exomes		
Dominant model	1	2	3	Any 1	Any 2	Any 1	Any 2	Any 3
NS/SS/I	4,645–4,687	3,358–3,940	2,850–3,099	6,658	4,489	6,943	5,167	3,920
Not in dbSNP129	634–695	136–369	72–105	1,617	274	1,829	553	172
Not in HapMap 8	898–979	161–506	55–117	2,336	409	2,628	835	222
Not in either	453–528	40–228	10–26	1,317	109	1,516	333	44
Predicted damaging	204–284	10–83	3–6	682	37	787	126	11
Recessive model								
NS/SS/I	2,780–2,863	1,993–2,362	1,646–1,810	4,097	2,713	4,293	3,172	2,329
Not in dbSNP129	92–115	30–53	22–31	226	61	270	90	42
Not in HapMap 8	111–133	13–46	5–13	329	32	397	75	19
Not in either	31–45	2–9	2–3	100	6	121	14	4
Predicted damaging	6–16	0–2	0–1	35	2	44	4	1

Under the dominant model, at least one nonsynonymous variant, splice acceptor or donor site variant or coding indel (NS/SS/I) in a gene was required in the gene. Under the recessive model, at least two novel variants were required, and these could be either at the same position (a homozygous variant) or at two different positions in the same gene (a potential compound heterozygote, though we are unable to ascertain phase at this stage). In each column are the range for the number of candidate genes for exomes considered individually (column 1) and all combinations of 2–4 exomes (columns 2–4). Note that the upper bound on the ranges may be inflated relative to what would be the case if four unrelated affected individuals had been used because the comparisons in which the two siblings were included provided reduced power compared to unrelated individuals. Columns 5–9 show the number of candidate genes when at least 1, 2 or 3 individuals is required to have one variant in a gene (dominant model) or two or more variants in a gene (recessive model). This is a simple model of genetic heterogeneity or incomplete data. For example, the total number of candidate genes common to any 3 of all 4 exomes is shown in column 9. For columns 5–6, one of the siblings (kindred 1-B) was not included in the analysis as siblings share 50% of variants.

ing. All four individuals were found to be compound heterozygotes for missense mutations in *DHODH* that are predicted to be deleterious. Collectively, 11 different mutations in 6 kindreds with Miller syndrome were identified in *DHODH* by a combination of exome and targeted resequencing (**Table 4** and **Fig. 2**). Each parent of an affected individual who was tested was found to be a heterozygous carrier, and none of the mutations appeared to have arisen *de novo*. In the kindreds with affected siblings, none of the unaffected siblings were compound heterozygotes. None of these mutations were found in 200 control chromosomes from unaffected individuals of matched geographical ancestry that were genotyped. Ten of these mutations were missense mutations, two of which affected the same amino acid codon, and one was a 1-bp indel that is predicted to cause a frameshift resulting in a termination codon seven amino acids downstream. One mutation, C1036T, was shared between two unrelated individuals with Miller syndrome who are of different self-identified geographical ancestry. Each of the amino acid residues affected by a *DHODH* mutation is highly conserved among homologs studied to date (**Supplementary Fig. 1**). A single, validated nonsynonymous polymorphism in human *DHODH* has been studied previously[19]. This polymorphism causes a lysine-to-glutamine substitution in the relatively diverse N-terminal extension of dihydroorotate dehydrogenase that is responsible for the association of the enzyme with the inner mitochondrial membrane.

DISCUSSION

We show that the sequencing of the exomes of affected individuals from a few unrelated kindreds, with appropriate filtering against public SNP databases and a small number of HapMap exomes, is sufficient to identify a single candidate gene for a monogenic disorder whose cause had previously been unknown, Miller syndrome. Several factors were important to the success of this study. First, Miller syndrome is a very rare disorder that is inherited in an autosomal recessive pattern. Therefore, the causal variants were unlikely to be found in public SNP databases or in control exomes. Second, genes for recessive diseases will, in general, be easier to find than genes for dominant disorders because fewer genes in any single individual have two or more new or rare nonsynonymous variants. Third, we were fortunate that there was no genetic heterogeneity in our sample of individuals with Miller syndrome. In the presence of heterogeneity, it is possible to relax stringency by allowing genes common to subsets of all affected individuals to be considered candidates, although this method will reduce power (**Table 3**). Fourth, all of the individuals with Miller syndrome for whom exomes were sequenced were of European ancestry. Sequencing exomes of affected individuals sampled from populations with a different geographical ancestry who have a higher number of novel and/or rare variants (for example, individuals with sub-Saharan African or East Asian ancestry) will make the identification of candidate genes more difficult. This will become less of an issue as databases of human polymorphisms become increasingly comprehensive.

Additional factors could facilitate the future application of this strategy. Mapping information, such as blocks of homozygosity, could focus the search to a smaller pool of candidates. The number of candidate variants can also be reduced further by comparison between variants in an affected individual to those found in each parent. For autosomal dominant disorders, this strategy can discover *de novo* coding variants, as neither parent is predicted to have a mutation that causes a fully penetrant dominant disorder; by contrast, in recessive disorders, parents are predicted to be carriers of the disease-causing variants.

There are at least three aspects of this approach where we see substantial scope for improvement. The first relates to missed variant calls, either due to low coverage or because some variants are not identified easily with current sequencing platforms—for example, those within repeat tracts in coding sequences. The second is that our filtering relied on a public SNP database (dbSNP) that has a highly uneven ascertainment of variation across the genome. It would be better to rely on catalogs of common variation that are ascertained in a single study either exome wide (as with the eight HapMap exomes[2]) or genome wide (for example, as with the 1,000 genomes project) and where estimates of allele frequency are available. Increasing the number of control exomes progressively reduces the relevance of dbSNP to this analysis (**Supplementary Fig. 2**). Furthermore, as increasingly deep catalogs of polymorphism become available, it may be necessary to establish frequency-based thresholds for defining common variation that is unlikely to be causal for disease. A third concern is that the

specificity of this approach is currently reduced by a subset of genes that recurrently appear to be enriched for new variants. These include long genes, but also genes that are subject to systematic technical artifacts (for example, mismapped reads due to duplicated or highly similar sequences in the genome). For sequences that are known to be duplicated or have paralogs (for example, genes from large gene families, or pseudogenes), these artifacts are mostly removed during read alignment (as reads with nonunique placements are removed from consideration). However, duplicated sequences not represented in the reference genome are not removed and spuriously appear as enriched for new variants (for example, *CDC27*).

Table 4 Summary of *DHODH* mutations in kindreds with Miller syndrome

Kindred	Mutation	Exon	Amino acid change	Location[b]
1[a]	G454A	4	G152R	chr16: 70608443
	G605C	5	G202A	chr16: 70612611
2[a]	C403T	3	R135C	chr16: 70606041
	C1036T	8	R346W	chr16: 70614936
3[a]	C595T	5	R199C	chr16: 70612601
	611ΔT	5	L204PfsX8	chr16: 70612617
4	G605A	5	G202D	chr16: 70612611
	C730T	6	R244W	chr16: 70613786
5	G56A	2	G19E	chr16: 70603484
	C1036T	8	R346W	chr16: 70614936
6	C851T	7	T284I	chr16: 70614596
	A1175G	9	D392G	chr16: 70615586

[a]Kindreds in which mutations were originally identified by exome resequencing. [b]Chromosomal position was determined using the March 2006 assembly from UCSC (hg18).

The mechanism by which mutations in *DHODH* cause Miller syndrome is unclear. The primary known function of dihydroorotate dehydrogenase is to catalyze the conversion of dihydroorotate to orotic acid, an intermediate in the pyrimidine *de novo* biosynthesis pathway (**Supplementary Fig. 3**)[20]. Orotic acid is subsequently converted to uridine monophosphate (UMP) by UMP synthase. Pyrimidine biosynthesis might be particularly sensitive to the step mediated by dihydroorotate dehydrogenase[21], and the classical rudimentary phenotype in *Drosophila melanogaster*, reported by T.H. Morgan in 1910 and characterized by wing anomalies, defective oogenesis and malformed posterior legs, is caused by mutations affecting the same pathway[22–24]. However, the clinical characteristics of the other inborn errors of pyrimidine biosynthesis—such as orotic aciduria, caused by mutations in UMP synthase—do not include malformations. Indeed, inborn errors of metabolism are, in general, a rare cause of birth defects, so *DHODH* would be given little weight a priori as a candidate for a multiple malformation disorder. Thus, the discovery that mutations in *DHODH* cause Miller syndrome reveals both a new role for pyrimidine metabolism in craniofacial and limb development as well as a newly discovered function of dihydroorotate dehydrogenase that remains to be explored.

Selective inhibition of pyrimidine or purine biosynthesis has long been used as a therapeutic option to treat various cancers and autoimmune disorders. Leflunomide, a prodrug that is converted in the gastrointestinal tract to the active metabolite, A771726, reduces *de novo* pyrimidine biosynthesis by selectively inhibiting dihydroorotate dehydrogenase[21]. In mice, use of leflunomide during pregnancy causes a wide range of limb and craniofacial defects in the offspring, the most common of which are exencephaly, cleft palate and 'open eye' or failure of the eyelid to close[25]. These phenotypic characteristics recapitulate some of the malformations observed in individuals with Miller syndrome, providing further evidence that it is caused by mutations in *DHODH*.

The developmental pathways disrupted by leflunomide are unknown, but their elucidation could help us understand the mechanism by which *DHODH* mutations cause malformations. In the liver of mice treated with leflunomide, TNF-α production is repressed by the direct inhibition of NF-κB activity[26]. Interruption of NF-κB signaling during development can result in disrupted cell migration, diminished cellular proliferation and increased apoptosis[27]. Indeed, open-eye is a defect observed in mice with targeted disruption of *TNFA*[28] . Furthermore, NF-κB has an important role in limb morphogenesis, specifically as a transducer of signals that regulate *Shh* (encoding the sonic hedgehog homolog) expression. *Shh* controls, in part, anterior-posterior patterning of the digits, and *Shh*$^{-/-}$ knockout mice fail to form digits 2–5 (ref. 29). These observations suggest that the malformations observed in individuals with Miller syndrome could be caused by perturbed NF-κB signaling due to loss of *DHODH* function.

The pattern of malformations observed in individuals with Miller syndrome is similar to those in individuals with fetal exposure to methotrexate (**Fig. 1c**,**d**). Methotrexate is a well-established inhibitor of *de novo* purine biosynthesis, and its antiproliferative actions are thought to be due to its inhibition of dihydrofolate reductase and folate-dependent transmethylations. Accordingly, defects of both purine and pyrimidine biosynthesis appear to be capable of causing a similar pattern of birth defects. However, at low doses, methotrexate also decreases plasma levels of pyrimidines as well as purines. This observation raises the possibility that methotrexate embryopathy might indeed be caused by the drug's effects on pyrimidine rather than purine metabolism. Given that not all embryos exposed to methotrexate manifest birth defects, functional polymorphisms in *DHODH* or other genes encoding proteins in the *de novo* pyrimidine biosynthesis pathway could influence susceptibility to methotrexate embryopathy.

Individuals with Miller syndrome have similar phenotypic characteristics to those with Nager syndrome (MIM%154400), another rare monogenic disorder that primarily affects the craniofacial skeleton. In contrast to Miller syndrome, the limb defects observed in individuals with Nager syndrome affect the anterior elements of the upper limb. Nevertheless, it has been hypothesized that Miller and Nager syndromes are caused by mutations in the same gene. We resequenced *DHODH* in 12 unrelated individuals diagnosed with Nager syndrome but found no pathogenic mutations (data not shown). Accordingly, either Nager syndrome and Miller syndrome are not allelic or Nager syndrome is caused exclusively by mutations in regulatory elements that alter the expression of *DHODH*.

Rare diseases are arbitrarily defined as those that affect fewer than 200,000 individuals in the United States. According to this definition, more than 7,000 rare diseases have been delineated, and in the aggregate, these affect more than 25 million people (Rare Diseases Act of 2002, Section 2, Findings; see URL section.). The majority of these diseases are considered genetic disorders, and many of them are thought to be monogenic. The bulk of genes underlying these rare monogenic diseases remain unknown. Lack of information about the genes and pathways that underlie rare monogenic diseases is a major gap in our scientific knowledge. Discovery of the genetic basis of a

large collection of rare disorders that have, to date, been unyielding to analysis will substantially expand our understanding of the biology of rare diseases, facilitate accurate diagnosis and improved management, and provide initiative for further investigation of new therapeutics.

We have demonstrated that exome sequencing of a small number of affected family members or affected unrelated individuals is a powerful, efficient and cost-effective strategy for markedly reducing the pool of candidate genes for rare monogenic disorders and may even identify the responsible gene(s) specifically. This approach is likely to become a standard tool for the discovery of genes underlying rare monogenic diseases and to provide important guidance for developing an analytical framework for finding rare variants influencing risk of common disease.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the families for their participation and the Foundation of Nager and Miller Syndrome for their support. We thank M. McMillin for assistance with project coordination. We thank R. Scott, T. Cox, L. Cox, R. Jack, E. Eichler, M. Emond, G. Cooper, J. Kidd, R. Waterston and E. Wijsman for discussions. Our work was supported in part by grants from the National Heart, Lung, and Blood Institute, National Human Genome Research Institute and National Institute of Child Health and Human Development of the US National Institutes of Health, the Life Sciences Discovery Fund and the Washington Research Foundation. S.B.N. is supported by the Agency for Science Technology and Research, Singapore. A.W.B. is supported by a training fellowship from the National Human Genome Research Institute.

AUTHOR CONTRIBUTIONS

The project was conceived and experiments planned by M.J.B., D.A.N. and J.S. Review of phenotypes and sample collection were performed by E.W.J. and M.J.B. Experiments were performed by S.B.N., K.J.B. and C.L. Genetic counseling and ethical consultation were provided by K.M.D. and H.K.T. Data analysis were performed by S.B.N., K.J.B., A.W.B., C.D.H., P.T.S. and J.S. The manuscript was written by M.J.B., J.S., S.B.N. and K.J.B. All aspects of the study were supervised by M.J.B., D.A.N. and J.S.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
2. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
3. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. USA* published online, doi:10.1073/pnas.0910672106 (27 October 2009).
4. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
5. Stenson, P.D. *et al.* The human gene mutation database: 2008 update. *Genome Med* **1**, 13 (2009).
6. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
7. Chen, C.T., Wang, J.C. & Cohen, B.A. The strength of selection on ultraconserved elements in the human genome. *Am. J. Hum. Genet.* **80**, 692–704 (2007).
8. Ahituv, N. *et al.* Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5**, e234 (2007).
9. Pruitt, K.D. *et al.* The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316–1323 (2009).
10. Toydemir, R.M. *et al.* Mutations in embryonic myosin heavy chain (MYH3) cause Freeman-Sheldon syndrome and Sheldon-Hall syndrome. *Nat. Genet.* **38**, 561–565 (2006).
11. Miller. M., Fineman, R. & Smith, D.W. Postaxial acrofacial dysostosis syndrome. *J. Pediatr.* **95**, 970–975 (1979).
12. Splendore, A., Passos-Bueno, M.R., Jabs, E.W., Van Maldergem, L. & Wulfsberg, E.A. TCOF1 mutations excluded from a role in other first and second branchial arch-related disorders. *Am. J. Med. Genet.* **111**, 324–327 (2002).
13. Fineman, R.M. Recurrence of the postaxial acrofacial dysostosis syndrome in a sibship: implications for genetic counseling. *J. Pediatr.* **98**, 87–88 (1981).
14. Oglivy-Stuart, A.L. & Parsons, A.C. Miller syndrome (postaxial acrofacial dysostosis): further evidence for autosomal recessive inheritance and expansion of the phenotype. *J. Med. Genet.* **28**, 695–700 (1991).
15. Donnai, D., Hughes, H.E. & Winter, R.M. Postaxial acrofacial dysostosis (Miller) syndrome. *J. Med. Genet.* **24**, 422–425 (1987).
16. Genée, E. Une forme extensive de dysostose mandibulo-faciale. *J. Genet. Hum.* **17**, 45–52 (1969).
17. Pereira, S.C.S., Rocha, C.M.G., Guion-Almeida, M.L. & Richieri-Costa, A. Postaxial acrofacial dysostosis: report on two patients. *Am. J. Med. Genet.* **44**, 274–279 (1992).
18. Robinow, M., Johnson, G.F. & Apesos, J. Robin sequence and oligodactyly in mother and son. *Am. J. Med. Genet.* **25**, 293–297 (1986).
19. Grabar, P.B., Rozman, B., Logar, D., Praprotnik, S. & Dolzan, V. Dihydroorotate dehydrogenase polymorphism influences the toxicity of leflunomide treatment in patients with rheumatoid arthritis. *Ann. Rheum. Dis.* **68**, 1367–1368 (2009).
20. Brosnan, M.E. & Brosnan, J.T. Orotic acid excretion and arginine metabolism. *J. Nutr.* **137**, 1656S–1661S (2007).
21. Breedveld, F.C. & Dayer, J.-M. Leflunomide: mode of action in the treatment of rheumatoid arthritis. *Ann. Rheum. Dis.* **59**, 841–849 (2000).
22. Morgan, T.H. Sex limited inheritance in *Drosophila*. *Science* **32**, 120–122 (1910).
23. Jarry, B. & Falk, D. Functional diversity within the rudimentary locus of *Drosophila melanogaster*. *Mol. Gen. Genet.* **135**, 113–122 (1974).
24. Conner, T.W. & Rawls, J.M. Jr. Analysis of the phenotypes exhibited by rudimentary-like mutants of *Drosophila melanogaster*. *Biochem. Genet.* **20**, 607–619 (1982).
25. Fukushima, R. *et al.* Teratogenicity study of the dihydroorotate-dehydrogenase inhibitor and protein tyrosine kinase inhibitor Lefunomide in mice. *Reprod. Toxicol.* **24**, 310–316 (2007).
26. Imose, M. *et al.* Lefunomide protects From T-cell–mediated liver injury in mice through inhibition of nuclear factor κB. *Hepatology* **40**, 1160–1169 (2004).
27. Bushdid, P.B., Brantley, D.M. & Yull, F.E. Inhibition of NF-κB activity results in disruption of the apical ectodermal ridge and aberrant limb morphogenesis. *Nature* **392**, 615–618 (1998).
28. Luetteke, N.C. *et al.* TGF alpha deficiency results in hair follicle and eye abnormalities in targeted and waved-1 mice. *Cell* **73**, 263–278 (1993).
29. Chiang, C. *et al.* Manifestation of the limb prepattern: limb development in the absence of sonic hedgehog function. *Dev. Biol.* **236**, 421–435 (2001).
30. Bawle E.V. *et al. Teratology* **57**, 51–55 (1978).

ONLINE METHODS

Patients and samples. For exome resequencing, we selected four individuals of self-reported European ancestry with Miller syndrome from three unrelated families. In two families, two siblings were affected (kindreds 1 and 2 in **Table 1**), and in one family a single individual had been diagnosed with Miller syndrome (kindreds 3 in **Table 1**). For validation, we selected samples from a sibling from kindred 2 and three simplex cases. All participants provided written consent, and the Institutional Review Boards of Seattle Children's Hospital and the University of Washington approved all studies. Separate informed consent was obtained from the individuals or their guardians to publish the photographs in **Figure 1**.

A referral diagnosis of Miller syndrome made by a clinical geneticist was required for inclusion. The clinical characteristics of several of the individuals who had been diagnosed with Miller syndrome have been reported previously[11,13,14]. Phenotypic data were collected from review of medical records, phone interviews and photographs.

Targeted capture and massive parallel sequencing. Genomic DNA was extracted from peripheral blood lymphocytes, using Gentra Systems PUREGENE DNA purification kit and 10 μg of DNA from each of the four individuals with Miller syndrome in kindreds 1, 2 and 3 was used for construction of a shotgun sequencing library as described previously[2] using adaptors for single-end sequencing on an Illumina Genome Analyzer II (GAII). Each shotgun library was hybridized to two Agilent 244K microarrays for target enrichment, followed by washing, elution and additional amplification[2]. The first array targeted CCDS2007, whereas the second was designed against targets poorly captured by the first array plus updates to CCDS in 2008. Enriched libraries were then sequenced on a GAII.

Read mapping and variant analysis. Reads were mapped to the reference human genome (UCSC hg18), initially with efficient large-scale alignment of nucleotide databases (ELAND) software (Illumina) for quality recalibration and then again with Maq[31]. Sequence calls were also performed by Maq and filtered to coordinates with ≥8× coverage and a Phred-like[31] consensus quality ≥20. Indels affecting coding sequence were identified as described previously[2]. Sequence calls were compared against eight HapMap individuals for whom we had previously reported exome data[2]. Annotations of variants were made using SeattleSeq Annotation based on NCBI and UCSC databases, supplemented with PolyPhen Grid Gateway predictions generated for nearly all nonsynonymous SNPs. Any nonsynonymous variant that was not assigned a 'benign' PolyPhen prediction was considered to be damaging, as were all splice acceptor and donor site mutations and all coding indels.

Mutation validation. Sanger sequencing of PCR amplicons from genomic DNA was used to confirm the presence and identity of variants in the candidate gene identified via exome sequencing and to screen the candidate gene in additional cases of Miller syndrome.

URLs. OMIM, http://www.ncbi.nlm.nih.gov/omim/; Rare Disease Act, http://history.nih.gov/research/downloads/PL107-280.pdf; PolyPhen, http://genetics.bwh.harvard.edu/pph/; SeattleSeq Annotation, http://gvs.gs.washington.edu/SeattleSeqAnnotation/

31. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).

doi:10.1038/ng.499