Genetic diagnosis by whole exome capture and massively parallel DNA sequencing

Murim Choi[a], Ute I. Scholl[a], Weizhen Ji[a], Tiewen Liu[a], Irina R. Tikhonova[b], Paul Zumbo[b], Ahmet Nayir[c], Ayşin Bakkaloğlu[d], Seza Özen[d], Sami Sanjad[e], Carol Nelson-Williams[a], Anita Farhi[a], Shrikant Mane[b], and Richard P. Lifton[a,1]

[a]Department of Genetics, Howard Hughes Medical Institute, [b]Keck Foundation for Biotechnology Resources, Yale University School of Medicine, New Haven, CT 06510; [c]Department of Pediatric Nephrology, Istanbul Medical Faculty, Istanbul 34390, Turkey; [d]Department of Pediatric Nephrology and Rheumatology, Hacettepe University Faculty of Medicine, Ankara 06100, Turkey; and [e]American University of Beirut, Beirut 11072020, Lebanon

Contributed by Richard P. Lifton, September 17, 2009 (sent for review September 8, 2009)

Protein coding genes constitute only approximately 1% of the human genome but harbor 85% of the mutations with large effects on disease-related traits. Therefore, efficient strategies for selectively sequencing complete coding regions (i.e., "whole exome") have the potential to contribute to the understanding of rare and common human diseases. Here we report a method for whole-exome sequencing coupling Roche/NimbleGen whole exome arrays to the Illumina DNA sequencing platform. We demonstrate the ability to capture approximately 95% of the targeted coding sequences with high sensitivity and specificity for detection of homozygous and heterozygous variants. We illustrate the utility of this approach by making an unanticipated genetic diagnosis of congenital chloride diarrhea in a patient referred with a suspected diagnosis of Bartter syndrome, a renal salt-wasting disease. The molecular diagnosis was based on the finding of a homozygous missense D652N mutation at a position in *SLC26A3* (the known congenital chloride diarrhea locus) that is virtually completely conserved in orthologues and paralogues from invertebrates to humans, and clinical follow-up confirmed the diagnosis. To our knowledge, whole-exome (or genome) sequencing has not previously been used to make a genetic diagnosis. Five additional patients suspected to have Bartter syndrome but who did not have mutations in known genes for this disease had homozygous deleterious mutations in *SLC26A3*. These results demonstrate the clinical utility of whole-exome sequencing and have implications for disease gene discovery and clinical diagnosis.

Bartter syndrome | congenital chloride diarrhea | next-generation sequencing | whole-exome sequencing | personal genomes

Genetic variation plays a major role in both Mendelian and non-Mendelian diseases. Among the approximately 2,600 Mendelian diseases that have been solved, the overwhelming majority are caused by rare mutations that affect the function of individual proteins; at individual Mendelian loci, approximately 85% of the disease-causing mutations can typically be found in the coding region or in canonical splice sites (1). For complex traits, genome-wide association studies have identified more than 250 common variants associated with risk alleles that contribute to a wide range of diseases (2, 3). To date, most of these impart small effects on disease risk (e.g., odds ratio of 1.2); moreover, even when extremely large studies have been performed, the vast majority of the genetic contribution to disease risk remain unexplained (4–6). These findings suggest that individually rare variants with relatively large effect may account for a large fraction of this missing trait variance. Indeed, studies addressing this question have documented the presence of individually rare variants with relatively large effect (7, 8). Consistent with the Mendelian model, coding variants have proven to be prevalent sources of such rare variants.

These considerations motivate implementation of robust approaches to sequencing complete coding regions of genomes (i.e., the "exome"). This has the potential to play a major role in disease gene discovery and also in clinical use for establishing a genetic diagnosis. As coding regions constitute only approximately 1% of the human genome, this is a potentially efficient strategy for identification of rare functional mutations, the more so given that our current ability to interpret the functional consequences of sequence variation outside coding regions is highly limited. The utility of this approach has been demonstrated in cancer, in which PCR amplification of individual exons has led to identification of new somatic mutations with large effect (9); nonetheless, PCR amplification of each coding sequence is cumbersome. Methods of enriching targeted genomic segments by hybridization have been used for 30 years (10), and have recently been extended to the whole exome scale (11), although their utility has been limited by the large amounts of genomic DNA required and by coupling to sequence platforms of modest throughput. Key considerations in this process will be cost effectiveness and completeness of the information obtained.

We report the adoption of whole-exome capture on single arrays on the Roche/NimbleGen platform to the Illumina sequencing platform. We illustrate the utility of this approach by identification of a rare mutation in a patient that led to an unexpected clinical diagnosis.

Results

Coupling of NimbleGen Whole-Exome Capture to Illumina Sequencing. The Roche/NimbleGen whole-exome array capture protocols were developed for DNA sequencing on the 454 platform (11); because the cost of sequencing on the Illumina platform is potentially considerably lower, we adapted hybrid capture using the NimbleGen 2.1M Human Exome Array to the Illumina DNA sequencing platform (see *Methods*). These arrays tile oligonucleotides from approximately 180,000 exons of 18,673 protein-coding genes and 551 micro-RNAs and comprise 34.0 Mb of genomic sequence. Resulting sequence data were processed using an automated pipeline: quality filtered sequence reads were aligned to the reference human genome (hg18) using Maq software (12). Single nucleotide variants (SNVs) were detected using Samtools (13) and further filtered (see *Methods*). For heterozygote calls we required at least 10× coverage by reads with different start sites, base call with Phred-like (14) quality scores greater than 45, and the probability of the frequency of major and minor alleles deviating from the binomial distribution of at least 10^{-7}. Rare SNVs that cluster within 1 kb were tagged and evaluated for mapping errors. SNVs were annotated for effect on the encoded protein and for conservation by comparison versus sequences of 43 vertebrate species (15) and orthologues in fly and worm (see *Methods*).

Author contributions: M.C., U.I.S., S.M., and R.P.L. designed research; M.C., U.I.S., W.J., T.L., I.R.T., P.Z., A.N., A.B., S.Ö., S.S., C.N.-W., A.F., S.M., and R.P.L. performed research; M.C., U.I.S., W.J., and R.P.L. analyzed data; and M.C., U.I.S., W.J., and R.P.L. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: richard.lifton@yale.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0910672106/DCSupplemental.

Illumina Exhibit 1023, Page 1 of 6

Table 1. Summary statistics for whole exome capture on samples from 5 subjects

Sample	GIT 264*	GIT 683	LMB 01	LMB 03	LMB 04	Mean
Mean per-base coverage (×)	40.1	37.5	36.2	43.5	60.5	43.6
Bases mapped to exome, %	34.8	38.2	36.1	39.5	36.1	36.9
Homozygous SNVs						
Total SNVs (novel)	9,045 (117)	7,395 (37)	8,273 (57)	8,424 (67)	8,860 (66)	8,399 (69)
Sensitivity,† %	99.1	98.4	99.0	98.6	99.3	98.9
Specificity,‡ %	99.4	98.9	98.6	97.0	98.8	98.5
Heterozygous SNVs						
Total SNVs (novel)	10,473 (905)	11,125 (865)	13,153 (1,186)	11,812 (994)	13,222 (979)	11,957 (986)
Sensitivity, %	96.3	95.3	94.9	91.3	96.4	94.8
Specificity, %	99.9	99.8	100.0	100.0	99.9	99.9
cSNVs						
Synonymous (novel)	6,462 (253)	6,377 (212)	7,229 (284)	6,708 (235)	7,299 (365)	6,815 (240)
Missense (novel)	5,091 (357)	4,986 (305)	5,731 (462)	5,226 (352)	5,672 (365)	5,341 (368)
Premature termination (novel)	34 (6)	28 (7)	33 (10)	29 (8)	36 (5)	32 (7)
Splice site (novel)	84 (10)	72 (8)	83 (6)	84 (6)	80 (4)	81 (7)
Indels (all novel, novel frameshift)	123 (21, 17)	149 (35, 25)	86 (15, 13)	83 (15, 13)	74 (12, 10)	103 (20, 16)

*Patient is a product of consanguineous union.
†Sensitivity defined as percentage of variant calls by SNP genotyping that were also called by capture sequencing divided by the number of all SNPs from genotyping.
‡Specificity defined as the percentage of variant calls at known SNP positions by sequencing that were also called by SNP genotyping.

Major changes in the protocol included shortening the length of the genomic DNA used for hybridization from 250 to 150 bases, which significantly increased the percentage of sequenced bases mapping to targeted bases (from 29% to 37%). We varied stringency by increasing the wash temperature from 47.5 °C to 53.5 °C. The higher stringency increased the percentage of on-target bases to 56%, but reduced breadth of coverage: at mean 30× per base coverage, the percentage of bases read fewer than 10 times increased from 7% to 23%.

In our current protocol, we captured and sequenced samples from 5 Caucasian subjects to a mean depth of 44×; these samples were also genotyped on the Illumina 370K SNP platform. The targeted bases constituted approximately 37% of all bases read (Table 1). An additional 12% of bases were within 100 base pairs of targeted sequences. The distribution of reads across all targeted bases is shown in Fig. 1*A*, which demonstrates that nearly all targeted bases are captured. Poorly captured bases, however, are correlated across different capture experiments, with approximately 352 kb (1.0%) read fewer than 10 times in each capture experiment. These underrepresented bases are predominantly from segments with exceptionally high G-C composition [supporting information (SI) Fig. S1]. We have captured 10 additional samples with this protocol with similar enrichment. With a single lane of Illumina sequencing using 75 base paired-end sequence using Illumina pipeline v. 1.4, we were able to produce an approximate mean 30× per-base coverage of the targeted bases, with 93% of targeted bases read at least 10 times; at 60× coverage with 2 lanes, 97% were read at least 10 times.

To assess the effect of coverage depth on sensitivity to detect sequence variants, one of these samples, GIT 264–1, was further sequenced to a mean depth of 99× and the genotype calls at known SNP positions were compared versus the results obtained with genotype calls on the Illumina 370K chip. At this depth of mean coverage, the sensitivity of exome sequencing for detecting homozygous and heterozygous variants was 99.8% and 98.2%, respectively (Fig. 1*C*). Sensitivity increased steeply as coverage increased from 5× to 20×, then more gradually thereafter, reaching a final plateau at approximately 50×. At any mean depth of coverage, there is a range of coverage at each base, and the ability to call heterozygous bases varies with the exact read coverage. Thus, the sensitivity to detect heterozygous variants with 10 reads is 78.6%, but increases to 95.2% at 20× and approximately 100% at 30× and greater (Fig. 1*D*). The overall per-base error rate for this sample was 0.75%, with error rate increasing approximately 10-fold from base 20 to base 75 (Fig. 1*B*).

Among the 5 test samples, we found a mean of 20,356 exomic variations from the reference sequence per subject (Table 1). An average of 1,055 per exome are not previously reported to our knowledge [i.e., not in the reference genome or 4 individual published genomes (16–19)] and were classified as novel. The small fraction of novel variants is consistent with a very high specificity for variant detection. These variants included a mean per exome of 12,372 coding SNVs (cSNVs; 642 novel), of which there were 5,341 missense variants (368 novel), 32 premature termination codons (7 novel), 81 canonical splice site variants (7 novel), and 6,815 synonymous substitutions (240 novel). Among the novel missense variants, 147 per exome introduce substitutions that are highly conserved among vertebrates, and among these, 40 are at positions that are also conserved in *Drosophila melanogaster* and/or *Caenorhabditis elegans*. Of the previously unreported exomic variants, 166 were shared among 2 or more subjects.

In subjects such as GIT 264–1 who are the product of consanguineous union (as described later), large segments of the genome are homozygous by descent; in these segments, all variations from the major allele can be considered errors, and heterozygous calls in such segments provide a further estimate of specificity (a small fraction of these could be bona fide de novo mutations). In this subject, 462 Mb of the genome were homozygous by descent as determined by genotyping (defined as segments of ≥100 consecutive SNPs spanning ≥1 Mb with ≥98% homozygosity; Table S1), including 5.3 Mb of the exome, comprising 2,459 genes. In these homozygous segments there was 1 heterozygous call at 99× coverage. This leads to an estimated false heterozygote discovery rate of 6 per exome and specificity for heterozygous calls of 99.9%. The specificity remains 99.9% at mean coverage of 30×. Interestingly, the false-positive call is present in the SNP database, so if one considers only novel variants in these calculations, specificity would be 100% at both levels of coverage.

These findings demonstrate a protocol of sufficient sensitivity and specificity to be highly useful for detecting rare sequence variants across the whole exome.

Case Report, Subject GIT 264–1. Subject GIT 264–1, a Turkish male, presented at age 5 months for evaluation of failure to thrive and dehydration. The history was notable for premature birth at 30 weeks and parental consanguinity, with 2 spontaneous abortions and death of a premature sibling on day 4; the parents were healthy

MEDICAL SCIENCES

Illumina Exhibit 1023, Page 2 of 6

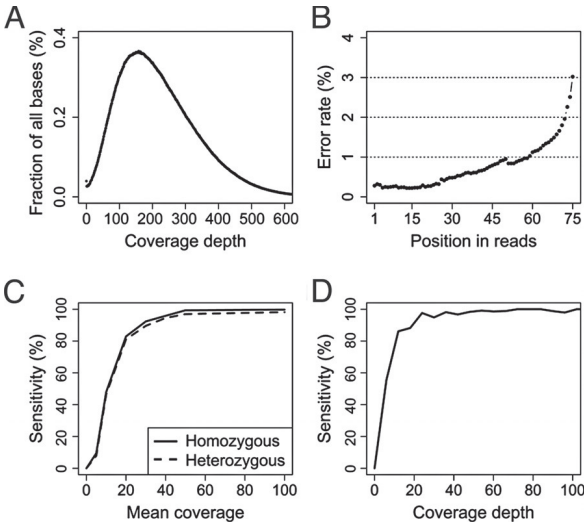

Fig. 1. Coverage of targeted bases, error rate, and sensitivity to detect variants in whole-exome capture data. (*A*) Distribution of per-base read coverage among 5 capture experiments. A small fraction of targeted bases are poorly captured across all experiments. (*B*) The per-base error rate in this data set is shown as a function of read position. (*C*) Subject GIT 264–1 was sequenced to a mean depth of 99×. The sensitivity to detect homozygous (solid line) or heterozygous (dashed line) variants as mean depth of whole-exome sequence coverage increases from 0 to 100× is shown. Sensitivity to detect heterozygous variants increases from 81% to 90% to 95% as mean coverage is increased from 20× to 30× and 40×, and plateaus at 98%. (*D*) Sensitivity of detection of heterozygous variants at exact per-base coverage. Sensitivity is approximately 80% at 10× coverage, and approaches 100% at or greater than 20× per-base coverage.

(Fig. 2). Blood pressure was 90/55 mmHg and laboratory evaluation was notable for serum K^+ level of 2.8 mmol/L, HCO_3^- level of 36 mmol/L; plasma renin activity and serum aldosterone levels were markedly elevated (Table 2). Despite volume depletion, his diapers were noted to be wet. The differential diagnosis of the admitting physician and consultants was broad, including a renal defect such as Bartter syndrome; however, a neurologic defect or an infectious process was not excluded. A venous blood sample was obtained and genomic DNA was prepared.

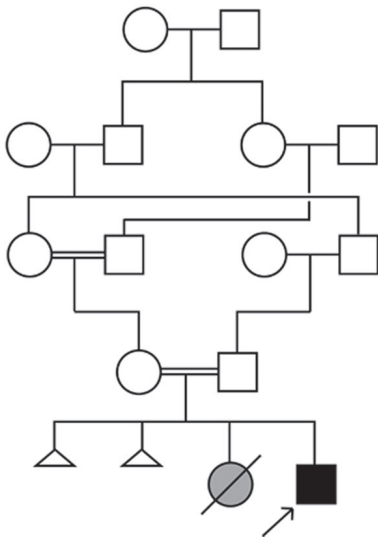

Fig. 2. Kindred GIT 264. The affected subject is indicated by the arrow, and the pedigree structure demonstrates parental consanguinity. A presumably affected sister died 4 d after premature birth (gray symbol), and there were 2 other spontaneous abortions (small triangles, disease status unknown). The parents and other relatives were free from clinical syndromes like that seen in the index case. For simplification, the mother's 10 siblings and the father's 8 siblings are not depicted in the pedigree.

Molecular Genetic Analysis. From the history, it was inferred that the patient likely had a recessive genetic trait. Genome-wide SNP genotyping was performed as described earlier. To search for copy number variants, the SNP data were interrogated by using a likelihood ratio-based copy number variant detection algorithm (see *Methods*). Within the homozygous-by-descent segments, 20 homozygous deletions were identified. All these were previously reported in the Database of Genomic Variants (20); none of these alter protein coding sequences (Table S2). No homozygous duplications were identified. Interestingly, none of the known loci for Bartter syndrome were within segments of homozygosity.

Analysis of Whole-Exome Sequence of Subject GIT 264–1. Given the diagnostic uncertainty, this patient's genomic DNA was subjected to whole-exome sequencing as described earlier. A total of 143 million reads passed quality assessment and 95.0% aligned to the human reference sequence; 38.4% of the total bases mapped to the targeted bases with mean coverage of 99×. At this depth of coverage, 99.3% of the bases were read at least 5 times and 98.4% of the bases were read at least 10 times (Table S3).

Mutation in *SLC26A3*. We anticipated finding a homozygous disease-causing mutation within a segment that is homozygous by consanguineous descent. We consequently sought novel and rare mutations that alter the encoded protein within the 5.3M base pairs and 2,495 genes in these segments. We identified 2,405 homozygous variations, including 1,493 homozygous cSNVs, from the reference sequence in these segments. These cSNVs included 668 nonsynonymous substitutions (29 of which were novel), 791 synonymous coding variants (16 novel), 12 canonical splice site variants (0 novel) and 19 coding region indels (0 novel; Table S4). There were 3 premature termination codons in these segments (0 novel). The remaining 931 variants were in introns, UTRs, or intergenic regions. All of the novel missense variants were confirmed by Sanger sequencing of PCR-amplified segments, confirming the high specificity for detection of rare variation.

Substitutions at positions that are completely conserved from invertebrates to humans are highly likely to disrupt normal protein function (8), and we ranked the novel missense variants by conservation scores to identify the most likely functional mutations using the phyloP conservation score (Table S4). Ten of the homozygous variants were at highly conserved positions (i.e., score ≥2).

Among the novel homozygous mutations at extremely conserved positions, one strongly stood out: a single-base substitution that introduced a missense variant in *SLC26A3* (Fig. 3). This mutation was identified in 172 of 173 reads covering this base; reads of this base had divergent start and end points and were read from both strands; the remaining read of this base was neither WT nor this mutant (Fig. 3*A*). The identity and homozygosity of this mutation was confirmed by PCR amplification and direct Sanger sequencing (Fig. 3*B*). This mutation introduces a D652N substitution at a position that is completely conserved among all invertebrates and vertebrates studied (Fig. 3*C*). This residue is also identical in 8 of 9 human paralogues (Fig. 3*D*); the only exception, SLC26A11, has glutamate rather than aspartate at the corresponding position. This mutation is noteworthy because *SLC26A3* encodes an epithelial Cl^-/HCO_3^- exchanger comprising 764 aa (Fig. 3*E*); recessive loss of function mutations in this gene are known to cause congenital chloride-losing diarrhea [CLD; i.e., OMIM 214700 (21)] in humans and mice (22). The D652N mutation lies in a β-pleated sheet in the STAS domain (Sulfate Transporters and bacterial Anti-Sigma factor antagonists) of the protein (Fig. 3*E*), which is known to be required for both the activity and biosynthesis/stability of the

Illumina Exhibit 1023, Page 3 of 6

Table 2. Clinical data of patients with mutations in *SLC26A3*

Kindred	Mutation	Presenting age	Ancestry	K^+, mmol/L	HCO_3^-, mmol/L	PRA, ng/mL/h	Aldosterone, pg/mL
264	D652N	5 months	Turkey	2.8	36	20.9	677.1*
409	Y520C	10 months	Turkey	2.5	32	17.1	52.3*
366	Q454FS	6 months	Turkey	2.3	32	NA	1,800†
396	D652N	NA	Turkey	<3.0	>30	NA	NA
178	G187X	6 years	Saudi Arabia	2.0	28	NA	NA
507	G187X	8 months	Turkey	1.8	43	NA	NA

K^+, serum potassium, normal level 3.5–5.1 mmol/L; HCO_3^-, serum bicarbonate, normal level 23–26 mmol/L; PRA, plasma renin activity; normal level 0.5–1.9 ng/mL/h. NA, not applicable.
*Normal level, 10–160 pg/mL.
†Normal level, 38.1–313.3 pg/mL.

transporter. Indeed, other mutations at conserved positions in the STAS domain have been shown to cause congenital chloride diarrhea (23). This mutation was absent among 190 control chromosomes. We identified no other compelling homozygous or heterozygous candidates for disease-causing mutations outside homozygous chromosome segments.

Our genetic findings strongly suggested that this patient had congenital chloride diarrhea; the clinical findings of volume depletion, hypokalemia, and metabolic alkalosis are all consistent with this diagnosis. This prompted clinical follow-up with the referring physician to determine whether the volume loss was from renal or gastrointestinal sources. Follow-up revealed that the patient indeed had diarrhea that occurred every 1 to 2 h, which had been persistent. Urinary electrolyte measurements with the patient volume replete did not reveal hypercalciuria or other features of Bartter syndrome (urinary Ca^{2+}:Cre ratio, 0.0076 mg:mg), indicating a primary diagnosis of CLD.

Additional Patients with *SLC26A3* Mutations. These findings led us to consider whether additional subjects referred with a presumptive diagnosis of Bartter syndrome might instead have CLD caused by mutation in *SLC26A3*. This mistaken diagnosis is particularly likely at the acute presentation: after volume and electrolyte resuscitation, urinary electrolyte levels can be appropriately high (when in balance, net intake and output are equal) but mistakenly interpreted as inappropriate, and volume loss from the urinary tract versus the gastrointestinal tract may not be immediately discriminated in infants.

We consequently screened 39 subjects referred with a suspected diagnosis of Bartter syndrome in whom mutations in NKCC2, ROMK, ClC-Kb, and Barttin had not been identified. We found 5 additional patients who had homozygous mutations in *SLC26A3* (Table 2 and Fig. 4). These mutations included 2 unrelated subjects with a G187X premature termination codon, one with a frameshift mutation at codon 454 resulting in premature termination at codon 458, one with the same D652N mutation found in the index case, and one with a Y520C mutation (Y520 is completely conserved among all vertebrates and invertebrates studied; Fig. 4*F*). All these mutations are considered novel except for the G187X mutation,

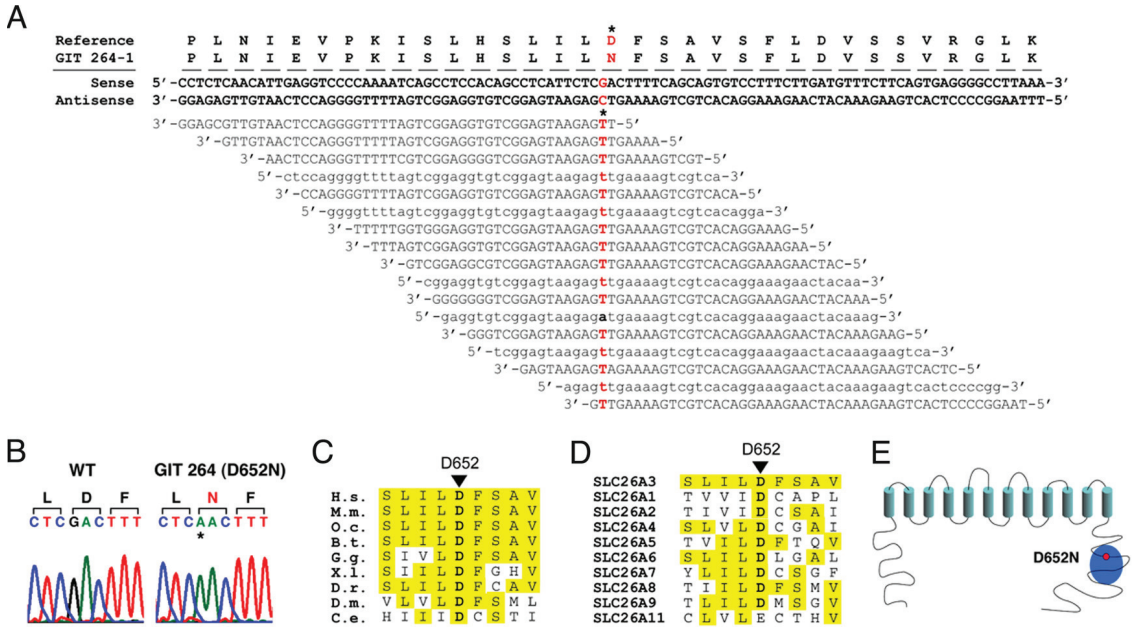

Fig. 3. Homozygous missense mutation at highly conserved position in SLC26A3 in GIT 264–1. (*A*) *Top*: Reference sequence of aa 636–668 of SLC26A3 and the corresponding DNA sequences are shown. *Below*: Independent Illumina DNA sequence reads from GIT 264–1 are shown. Forward and reverse reads are shown in capital and lowercase letters, respectively. The results demonstrate a homozygous missense mutation, D652N. (*B*) Sanger sequence of codons 651–653 of *SLC26A3* in a WT subject and GIT 264–1 are shown and confirm the D652N mutation. The mutated residue is indicated by an asterisk, and the encoded amino acids are shown in single-letter code. (*C*) Conservation of D652 across species. The amino acid sequence of segment 648–656 of human SLC26A3 is shown and compared to the corresponding sequence of 6 (of 39) vertebrates and identified in *D. melanogaster* paralogue and *C. elegans* orthologue. Positions identical to *Homo sapiens* (*H.s.*) are highlighted in yellow. D652 is completely conserved among all species examined. M.m., *Mus musculus*; O.c., *Oryctolagus cuniculus*; B.t., *Bos taurus*; G.g., *Gallus gallus*; X.l., *Xenopus laevis*; D.r., *Danio rerio*; D.m., *D. melanogaster*; C.e., *C. elegans*. (*D*) The sequence of the same segment of human SLC26A3 is compared to 9 paralogues of the human SLC26A gene family (*SLC26A1-A11*; *SLC26A10* is a pseudogene). (*E*) Structure of SLC26A3. The protein has 12 transmembrane domains and a C-terminal STAS domain (highlighted in blue). The D652N mutation (red circle) lies in the STAS domain.

MEDICAL SCIENCES

Illumina Exhibit 1023, Page 4 of 6

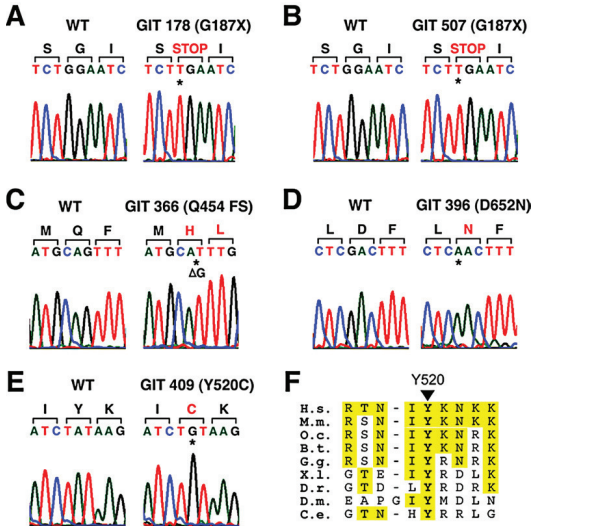

Fig. 4. Additional patients with *SLC26A3* mutations. Sanger sequence traces of WT and 5 subjects with homozygous mutations in *SLC26A3* are shown. These include two subjects with premature termination at codon 187 (*A* and *B*); a frameshift at codon 454 resulting from a single base deletion that leads to termination at codon 458 (*C*); a second patient with the D652N mutation (*D*); and a patient with a Y520C mutation (*E*). (*F*) Conservation of Y520 across species: Y520 is conserved among all orthologues and paralogues examined.

which is a known founder mutation for congenital chloride diarrhea in Saudi Arabia (24). None of these mutations were found in sequencing of 190 control chromosomes from unrelated Caucasian subjects. All these patients had hypokalemia, metabolic alkalosis, and evidence of salt wasting. Follow-up with the referring physicians in 3 cases documented that volume depletion was caused by gastrointestinal losses from watery diarrhea and not renal losses; in 2 cases, high stool chloride level was documented (GIT 366–1, 112 mM; GIT 409–1, 148 mM; value should be >90 mM in CLD), a finding that is pathognomonic for congenital chloride diarrhea. In each of the 2 remaining cases, the patient was not available for further evaluation.

Discussion

Our results demonstrate the utility of whole-exome sequencing, coupling sequence capture on the NimbleGen platform and sequencing on the Illumina platform, for identification of rare variants and disease-causing mutations. The protocol used provided sufficient enrichment to drastically reduce the amount of DNA sequencing required to identify sequence variation in coding regions. Virtually all the targeted bases can be captured by using the current protocol albeit with varying efficiencies, and additional efforts might further balance the coverage of poorly sequenced bases. We estimate that exon capture reduces the cost of detection of exonic mutations by a factor of 10 to 20 compared with whole-genome sequencing, depending on the sensitivity pursued. Using the current protocol, 2 lanes of capture sequence can produce approximately 60× read depth per base, resulting in detection of approximately 97.2% of the heterozygous bases; it is expected that, among the approximate average of 402 protein-altering rare variants per exome, there will be a very small number of false-positive variants. Increased reads per lane are expected to occur in the near future, likely permitting virtually complete capture of variation in a single lane of sequence, and further improvements in base calling will likely occur as well, resulting in improvements in both sensitivity and specificity.

During preparation of this manuscript, a related paper was published that demonstrated development of a different platform for whole-exome capture and sequencing on the Illumina instrument (25). The targeted exons were similar, both based on the consensus coding sequence database. The overall results are similar, with both platforms showing high sensitivity and specificity. We anticipate that it will be useful to have multiple platforms available for performing whole-exome analyses.

We have demonstrated the utility of this platform by making an unexpected genetic diagnosis in a patient with an undiagnosed illness; to our knowledge, this has not been previously reported. The diagnosis of congenital chloride diarrhea was not considered in the referral differential diagnosis, but after identification of mutation in *SLC26A3* suggested the diagnosis, it was confirmed by follow-up clinical evaluation. This example provides proof of concept of the use of whole-exome sequencing as a clinical tool in evaluation of patients with undiagnosed genetic illnesses. These findings further underscore the ability to parse large quantities of sequence data to produce clinically useful information that combines clues from the clinical condition in conjunction with the genetic data to arrive at a correct diagnosis. We can envision a future in which such information will become part of the routine clinical evaluation of patients with suspected genetic diseases in whom the diagnosis is uncertain.

In addition to clinical genetics, we anticipate applications for whole-exome sequencing that include discovery of genes and alleles contributing to Mendelian and complex traits, and to somatic mutations in cancers. For dominant traits, we expect that many whose causes have not yet been identified are explained by alleles that have been difficult to map by linkage analysis. Likely explanations include traits with reduced penetrance, traits that show substantial locus heterogeneity, and alleles that impair reproductive fitness sufficiently that many affected subjects harbor de novo mutations. The finding of independent de novo mutations in the same gene among a small number of affected subjects would constitute compelling evidence of disease causation. Similarly, with locus heterogeneity, identification of a significant excess in the number of independent mutations in the same gene versus that expected by chance will constitute evidence that a disease gene has been identified. Mapping data that constrain the location of the disease locus, animal models that produce similar phenotypes, and compelling biology can all contribute to identification of such loci; however, in the absence of such information, inference from the large number of variants in a single subject will be challenging.

For recessive traits, affected subjects arising from consanguineous union contain substantial mapping information: in this setting, the disease locus is expected to be homozygous within a segment that is homozygous by descent from a recent ancestor. For first cousins, this constrains the search on average to approximately 10% of the exome. Detection of homozygous novel variants in these segments is highly efficient at even low levels of per-base coverage (e.g., 5×); these considerations simplify the problem as there are very few (approximately 29) rare homozygous protein-altering variants in these segments. As described earlier, the finding of an excess of independent mutations in the same locus will provide compelling evidence that a disease locus has been identified, and KO animal models may be particularly helpful as recessive mutations so typically impart loss of function.

Finally, for complex traits, sequencing large numbers of cases/controls or subjects at the tails of a quantitative distribution will enable identification of genes preferentially harboring an excess of rare functional mutations at one end of the distribution. The sample sizes required to identify such loci are expected to be large for alleles that have odds ratios of disease of 2 to 3 (26–28). Such studies can be achieved either by whole-exome sequencing of all subjects or by sequencing the exome of a subset of subjects followed by selective sequencing of the most promising genes in follow-up studies. It is of course likely that non-coding sequences will harbor some rare variants with relatively large effect that would be missed by selective sequencing of coding regions.

Illumina Exhibit 1023, Page 5 of 6

At present, the sequence of each new individual identifies approximately 402 novel variants introducing changes in the encoded protein. Recognition of functional mutations among these poses a challenge. Investigation of consanguineous kindreds for recessive traits can reduce the search complexity by an order of magnitude. Mutations that introduce truncations of the encoded protein, or that occur at highly conserved positions, are more likely to have functional effects and can be prioritized. In addition, as databases are filled with exome sequences of large numbers of subjects, the ability to distinguish low-frequency alleles descendent from ancient ancestors from de novo or extremely rare mutations recently introduced into the population will markedly improve, permitting focusing attention on this smaller set of alleles. We anticipate that whole-exome sequencing will make broad contributions to understanding the genes and pathways that contribute to rare and common human diseases, as well as clinical practice.

Materials and Methods

Human Subjects. The study protocol was approved by the Yale Human Investigation Committee. Consent for participation was obtained in accordance with institutional review board standards. Patients were referred for studies of Bartter syndrome, and consanguineous kindred subjects in whom no mutation in genes causing Bartter syndrome had been identified were chosen for further analysis. Genomic DNA was prepared from venous blood by standard procedures.

Targeted Sequence Capture. Genomic DNA was captured on a NimbleGen 2.1M human exome array following the manufacturer's protocols (Roche/NimbleGen) with modifications at the W. M. Keck Facility at Yale University. DNA was sheared by sonication and adaptors were ligated to the resulting fragments. The adaptor-ligated templates were fractionated by agarose gel electrophoresis and fragments of the desired size were excised. Extracted DNA was amplified by ligation-mediated PCR, purified, and hybridized to the capture array at 42.0 °C using the manufacturer's buffer. The array was washed twice at 47.5 °C and 3 more times at room temperature using the manufacturer's buffers. Bound genomic DNA was eluted using 125 mM NaOH for 10 min at room temperature, purified, and amplified by ligation-mediated PCR. The resulting fragments were purified and subjected to DNA sequencing on the Illumina platform. Captured and non-captured amplified samples were subjected to quantitative PCR to measure the relative fold enrichment of the targeted sequence.

Sequencing. Captured libraries were sequenced on the Illumina genome analyzer as single-end 50-, 74-, and 75-bp reads, or 75-bp paired-end reads, following the manufacturer's protocols. Image analysis and base calling was performed by Illumina pipeline versions 1.3 and 1.4 with default parameters.

Analysis. Sequence reads were mapped to the reference genome (hg18) using the Maq program (12). Reads outside the targeted sequences were discarded and statistics on coverage were collected from the remaining reads using perl scripts. For indel detection, BWA was used to allow gapped alignment to the reference genome (29). SAMtools was used to call targeted bases, and any base call that deviates from reference base was regarded as a potential variation (13). Additional filters were applied as described in the SI text. Identified variants were annotated based on novelty, impact on the encoded protein, conservation, and expression using an automated pipeline (SI Text).

Genotyping and Identification of Loss-of-Heterozygosity Intervals. Samples were genotyped on the Illumina HumanCNV370-Duo BeadChips. Sample processing and labeling were performed following the manufacturer's protocols. Plink v. 1.06 was used to identify homozygous intervals from subject GIT 264–1. A sliding window of 50 SNPs was used on the tag SNPs and included no more than one possible heterozygous genotype. Resulting intervals have to meet the limit of at least 100 SNPs and 1 Mb in size.

Copy Number Variation Prediction Algorithm. To predict possible deletion or duplication events in the genome, LogR and B allele frequency data were extracted from SNP array and normalized to the larger sample pool to remove sample-specific noise. Based on the empirical distribution of LogR and B allele frequency values from the validated deletion and duplication SNPs, for a given SNP, the likelihoods of being 0, 1, 2, and 3 or more copies were calculated, respectively. If more than one SNP supporting deletion or duplication arose consecutively and the likelihood ratio of the SNPs exceeded threshold, they were called a deletion or duplication.

DNA Sequencing. The ExonPrimer script (http://ihg2.helmholtz-muenchen.de/ihg/ExonPrimer.html) was used to generate primers for amplification of *SLC26A3* coding exons using as a template genomic DNA of patients or controls (Table S5). Products were analyzed via gel electrophoresis, and amplicons sequenced using the forward or reverse primers. Disease-causing mutations were confirmed by at least 2 independent sequences from different primers.

Orthologues and Paralogues. Full-length orthologous and paralogous protein sequences from vertebrate and invertebrate species and human paralogous protein sequences were extracted from GenBank. Orthologues were confirmed based on database identity of annotation. If an orthologue could not be identified, the closest paralogue (i.e., top ''hit'' of a BLAST search of the respective proteome) was studied. Protein sequences were aligned using the ClustalW algorithm. GenBank accession numbers are listed in the SI Text.

ACKNOWLEDGMENTS. We thank the patients studied and their families for their irreplaceable contribution to this study. We thank Tom Albert of Roche NimbleGen, Ali Gharavi, Lynn Boyden, George Farr, Murat Gunel, and Matt State for helpful discussions. Supported in part by the Leducq Transatlantic Network in Hypertension, a National Center for Research Resources High End Instrumentation grant and the Yale National Institutes of Health O'Brien Center for Kidney Research.

1. Cooper DN, Krawczak M, Antonorakis SE (1995) The nature and mechanisms of human gene mutation. *The Metabolic and Molecular Bases of Inherited Disease*, eds Scriver C, Beaudet al., Sly WS, Valle D (McGraw-Hill, New York), 7th Ed, pp. 259–291.
2. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322:881–888.
3. Hirschhorn JN (2009) Genomewide association studies–illuminating biologic pathways. *N Engl J Med* 360:1699–1701.
4. Aulchenko YS, et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* 41:47–55.
5. Kathiresan S, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 41:56–65.
6. Zeggini E, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 40:638–645.
7. Cohen JC, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305:869–872.
8. Ji W, et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* 40:592–599.
9. Parsons DW, et al. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* 321:1807–1812.
10. Goldberg ML, Lifton RP, Stark GR, Williams JG (1979) Isolation of specific RNAs using DNA covalently linked to diazobenzyloxymethyl cellulose or paper. *Methods Enzymol* 68:206–220.
11. Albert TJ, et al. (2007) Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 4:903–905.
12. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
13. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
14. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–194.
15. Siepel A, Pollard KS, Haussler D (2006) New methods for detecting lineage-specific selection. *Proceedings of the 10th International Conference on Research in Computational Molecular* Biology (*RECOMB 2006*) (Springer, Berlin), pp. 190–205.
16. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.
17. Ng PC, et al. (2008) Genetic variation in an individual human exome. *PLoS Genet* 4:e1000160.
18. Wang J, et al. (2008) The diploid genome sequence of an Asian individual. *Nature* 456:60–65.
19. Wheeler DA, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–876.
20. Iafrate AJ, et al. (2004) Detection of large-scale variation in the human genome. *Nat Genet* 36:949–951.
21. Hoglund P, et al. (1996) Mutations of the Down-regulated in adenoma (DRA) gene cause congenital chloride diarrhoea. *Nat Genet* 14:316–319.
22. Schweinfest CW, et al. (2006) slc26a3 (dra)-deficient mice display chloride-losing diarrhea, enhanced colonic proliferation, and distinct up-regulation of ion transporters in the colon. *J Biol Chem* 281:37962–37971.
23. Dorwart MR, et al. (2008) Congenital chloride-losing diarrhea causing mutations in the STAS domain result in misfolding and mistrafficking of SLC26A3. *J Biol Chem* 283:8711–8722.
24. Hoglund P, et al. (1998) Clustering of private mutations in the congenital chloride diarrhea/down-regulated in adenoma gene. *Hum Mutat* 11:321–327.
25. Ng SB, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–276.
26. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 40:695–701.
27. Li B, Leal SM (2009) Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet* 5:e1000481.
28. Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 6:109–118.
29. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Illumina Exhibit 1023, Page 6 of 6