Clinical Chemistry 56:1
62–72 (2010)

DNA Sequence Capture and Enrichment by Microarray Followed by Next-Generation Sequencing for Targeted Resequencing: Neurofibromatosis Type 1 Gene as a Model

Lan-Szu Chou,[1*] C.-S. Jonathan Liu,[2] Benjamin Boese,[3] Xinmin Zhang,[4] and Rong Mao[1,5]

BACKGROUND: The introduction and use of next-generation sequencing (NGS) techniques have taken genomic research into a new era; however, implementing such powerful techniques in diagnostics laboratories for applications such as resequencing of targeted disease genes requires attention to technical issues, including sequencing template enrichment, management of massive data, and high interference by homologous sequences.

METHODS: In this study, we investigated a process for enriching DNA samples that uses a customized high-density oligonucleotide microarray to enrich a targeted 280-kb region of the *NF1* (neurofibromin 1) gene. The captured DNA was sequenced with the Roche/454 GS FLX system. Two *NF1* samples (CN1 and CN2) with known genotypes were tested with this protocol.

RESULTS: Targeted microarray capture may also capture sequences from nontargeted regions in the genome. The capture specificity estimated for the targeted *NF1* region was approximately 60%. The de novo Alu insertion was partially detected in sample CN1 by additional de novo assembly with 50% base-match stringency; the single-base deletion in sample CN2 was successfully detected by reference mapping. Interferences by pseudogene sequences were removed by means of dual-mode reference-mapping analysis, which reduced the risk of generating false-positive data. The risk of generating false-negative data was minimized with higher sequence coverage (>30×).

CONCLUSIONS: We used a clinically relevant complex genomic target to evaluate a microarray-based sample-enrichment process and an NGS instrument for clinical resequencing purposes. The results allowed us to develop a systematic data-analysis strategy and algorithm to fit potential clinical applications.

Sequencing whole genomes for personalized medical treatment may become possible with new-generation sequencing applications, such as targeted resequencing, whole-genome de novo sequencing, transcriptome sequencing, and microRNA profiling *(1–5)*. Targeted resequencing of an extremely large gene or multiple disease-related genes has the unique advantage of allowing the identification of genetic alterations that can be targeted by gene therapy, such as the use of rapamycin for the mTOR pathway for the *NF1* (neurofibromin 1) gene and the use of antisense oligonucleotides for read-through of premature stop codons and exon skipping to treat Duchenne muscular dystrophy *(6–9)*. Research studies that have used these new-generation sequencing applications, although few in number, have emphasized the importance of analyzing human genome sequences and have suggested that sequence information will have a noteworthy impact on molecular medicine *(10)*.

Targeted resequencing by next-generation sequencing (NGS)[6] techniques requires the targeted enrichment of DNA samples. Currently, PCR-based methods are the most widely used for preparing a genomic sample for sequencing purposes *(11–14)*; however, PCR methods fall short of meeting the technological needs for sequencing highly complex genomic regions and in large-scale genomic studies. PCR methods cannot amplify very long sequences reliably, and multiplex enrichment of thousands of sequences is difficult *(14)*. Several technologies have been developed for targeted enrichment by sequence

[1] Institute for Clinical and Experimental Pathology, ARUP Laboratories, Salt Lake City, UT; [2] SoftGenetics, LLC, State College, PA; [3] 454 Life Sciences, A Roche Company, Branford, CT; [4] Roche NimbleGen, Madison, WI; [5] Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT.
[*] Address correspondence to this author at: ARUP Laboratories, 500 Chipeta Way, Salt Lake City, UT 84108-1221. Fax 801-584-5114; e-mail choulb@aruplab.com.

Received July 6, 2009; accepted October 20, 2009.
Previously published online at DOI: 10.1373/clinchem.2009.132639

[6] Nonstandard abbreviations: NGS, next-generation sequencing; SCE, sequence capture and enrichment; emPCR, emulsion PCR; HQ, high-quality; DMRM, dual-mode reference mapping.

capture, including solid phase–based microarrays and solution phase–based methods *(15–18)*. Owing to the limited availability of the SureSelect™ system (Agilent Technologies) and the complexity of synthesizing padlock probes, we used the microarray method in this study *(16)*. This recently developed technology selects targeted sequences by hybridization to an oligonucleotide microarray and demonstrates great potential for the efficient enrichment of specific, large high-complexity genomic regions of interest *(14, 15, 19, 20)*. Combining this technology with NGS produces a powerful sequencing tool that has the potential to be implemented in a clinical diagnostics laboratory. Aside from such issues as test reporting and ethical considerations, 2 main technical hurdles need to be overcome besides targeted DNA enrichment: management of massive amounts of data and potential interference from highly homologous sequences (e.g., pseudogenes) *(10)*. Because the NGS technology is so new, QC of the sequence data (including the accuracy of reads, quality scores for reads, and sequencing-coverage needs) has not yet been well defined, but it will need to be before NGS technology can be used routinely in molecular diagnostics laboratories *(10, 21)*. Therefore, the goal of this study was to benchmark this new sequence capture and enrichment (SCE) NGS technology by comparing it with the current gold standard, Sanger sequencing, in the application of targeted resequencing.

We selected the *NF1* gene as the clinical target in this pilot study. Neurofibromatosis type 1, a common tumor-predisposing disorder that occurs in 1 of 3000 births, is caused by autosomal dominant mutations in a single gene, *NF1* (17q11.2). Neurofibromatosis type 1 presents with a wide range of manifestations that can lead to serious disabilities, including cognitive problems *(22–25)*. In view of the complexity of the premature stop codon mutations in *NF1*, the existence of 7 pseudogenes and pseudogene sequences on different chromosomes, and the fact that this gene consists of 58 exons and harbors complex genomic structures, such as repetitive sequences and Alu sequences, *NF1* is an ideal clinical target for evaluating the effectiveness of the SCE-NGS method for clinical applications *(24, 26–29)*. In the present study, we successfully enriched targeted sequences with a specifically designed *NF1* capture microarray and developed detailed methods for data analysis, including removal of interferences due to highly homologous pseudogenes.

Materials and Methods

DNA SAMPLES

Two neurofibromatosis type 1–positive samples (with known *NF1* genotypes) were obtained from Coriell Cell Repositories and coded as CN1 (NA11601) and CN2 (NA11602). DNA quantity and quality were measured with a NanoDrop spectrophotometer (Thermo Scientific). The DNA concentration was 373 ng/μL for CN1 and 344 ng/μL for CN2. To ensure sample quality for sequence capture of the targeted region, we prepared 80 μL of each sample and shipped it to NimbleGen (Roche NimbleGen) for targeted sequence capture.

DESIGN OF THE OLIGONUCLEOTIDE MICROARRAY FOR *NF1* REGION CAPTURE

The *NF1* region to be captured is at genomic position 26 446 121 to 26 728 821 (UCSC hg18, NCBI Build 36). Long oligonucleotide probes (>60mers) were chosen from the target region, and 15mer frequency and the uniqueness of probes in the genome were evaluated to exclude repetitive regions from probe selection [see *(30)* for details]. We selected 385 000 probes specifically to tile the target region (*NF1*) with high density. We used a uniqueness score of 1 during the design to avoid capturing pseudogene sequences.

TARGETED SEQUENCE CAPTURE BY THE CUSTOMIZED OLIGONUCLEOTIDE MICROARRAY

The methods used for sequence capture and elution followed previously described protocols *(31)*. In brief, DNA was nebulized to yield fragments from 250 bp to approximately 1 kb in size, and a 22mer linker was added to both ends. The DNA fragments were then hybridized to the *NF1* sequence-capture array for 3 days. After hybridization and stringent washing, the DNA fragments remaining on the array (mostly from target regions) were eluted off the array and amplified with 22mer linkers to generate enough DNA template for downstream applications. After amplification, the product was tested for enrichment level by quantitative PCR with 4 QC control loci to determine the degree of capture success. These 4 QC loci, which are conserved in both human and mouse genomes, have been empirically determined to accurately predict enrichment with several different experimental designs. Finally, the amount of captured DNA (after elution) for GS FLX sequencing was measured by spectrophotometry, on the assumption that an A_{260} value of 1.0 is equivalent to approximately 50 μg/mL double-stranded DNA.

GS FLX SEQUENCING AND COMPARATIVE DATA ANALYSIS

The captured DNA was sequenced with the GS FLX instrument at the 454 Sequencing Service Center (Branford, CT). Because of the relatively small size of the sequence-capture fragments (250 bp to approximately 1 kb), the nebulization step was not performed. The Low Molecular Weight DNA Protocol *(32)* was used to prepare the 454 GS FLX sequence-ready libraries. The DNA-sequencing libraries for the 2 samples (CN1 and CN2) were prepared separately before amplification by emulsion PCR (emPCR). The emPCR followed the steps described in the GS FLX emPCR Method Manual *(33)*. The 2 prepared samples (after emPCR amplification) were loaded in a single 2-lane gasket PicoTiterPlate device (70 × 75 mm; Roche/454) and sequenced in a GS FLX system with standard Roche/454 protocols. The 454 pyrosequencing data were collected after a 7-h run on the GS FLX system.

The Roche/454 gsMapper was used initially to analyze all raw sequence reads generated from the GS FLX system. We used NextGENe™ software (version 1.12; SoftGenetics) to analyze the GS FLX data (.sff files) to investigate the possibility of analyzing NGS data with Windows-based software packages (see Supplemental Data Analysis in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol56/issue1). We fully sequenced both samples via Sanger sequencing (with specific primer sets used to avoid amplification of pseudogenes) to confirm any sequence variants detected with the SCE-NGS method.

Results

NF1 SCE

In the present study, we used NimbleGen's long oligonucleotide microarray with 385 000 features to design specific probes to capture entire *NF1* sequences located in 17q11.2 (a 282 700-bp region). Capture efficiencies (-fold enrichment), which were estimated by quantitative PCR before and after capture with 4 QC control loci, were 336-fold and 586-fold increases for samples CN1 and CN2, respectively. Both enrichments passed the QC standard (>100-fold enrichment), although CN1 was less enriched (an approximately 42% reduction) than CN2.

GS FLX SEQUENCING AND REFERENCE-MAPPING ANALYSIS

After SCE with the customized *NF1* capture array, both SCE samples (CN1 and CN2) were sequenced by the GS FLX genome sequencer. The data were analyzed with the gsMapper software. Sample CN1 had a mean read length of 210 bp and 254 743 high-quality (HQ) reads, 52.0% of which were uniquely mapped to the reference genome (UCSC hg18, NCBI Build 36). Of the uniquely mapped HQ reads, 59.3% were on the targeted *NF1* loci (Table 1). Sample CN2 had a mean read length of 222 bp and 271 921 HQ reads, 70.0% of which were uniquely mapped to the reference genome. Of the uniquely mapped HQ reads, 58.8% were on the targeted *NF1* loci (Table 1). The reference-mapping analyses showed similar percentages of HQ reads (59.3% for CN1, 58.8% for CN2) mapped to the targeted *NF1* region, suggesting that the SCE-NGS method had a reasonable within-run reproducibility and capture specificity (approximately 60%), although the 2 samples differed with respect to the degree of enrichment (336-fold vs 586-fold) and the percentage of HQ reads uniquely mapped to the reference genome (52.0% vs 70.0%). In addition, the percentage of reads mapped to chromosomes other than the targeted 17q11.2 was not large (<4%, Table 1) for both samples analyzed, suggesting that the design of the sequence-capture array was unique and specific for the corresponding target; however, this supposition was not confirmed with base pair–level analyses of the entire data set.

Table 1. Reference-mapping analysis of sample CN1 and CN2.

	Percentage of the uniquely mapped reads on the targeted *NF1* region	
Chromosome	Sample CN1	Sample CN2
Chr01	3.40%	3.50%
Chr02	3.90%	3.80%
Chr03	4.10%	3.80%
Chr04	2.60%	2.40%
Chr05	2.20%	2.20%
Chr06	2.30%	2.40%
Chr07	1.90%	2.00%
Chr08	2.00%	2.00%
Chr09	1.90%	1.80%
Chr10	1.70%	1.90%
Chr11	1.90%	1.90%
Chr12	2.10%	2.00%
Chr13	1.30%	1.40%
Chr14	1.30%	1.30%
Chr15	1.30%	1.40%
Chr16	1.40%	1.40%
Chr17	59.30%	58.80%
Chr18	1.10%	1.20%
Chr19	0.70%	0.80%
Chr20	1.00%	1.00%
Chr21	0.50%	0.50%
Chr22	0.70%	0.70%
ChrM[a]	0.20%	0.20%
ChrX	1.40%	1.30%
ChrY	0.10%	0.10%

[a] ChrM, mitochondrial DNA.

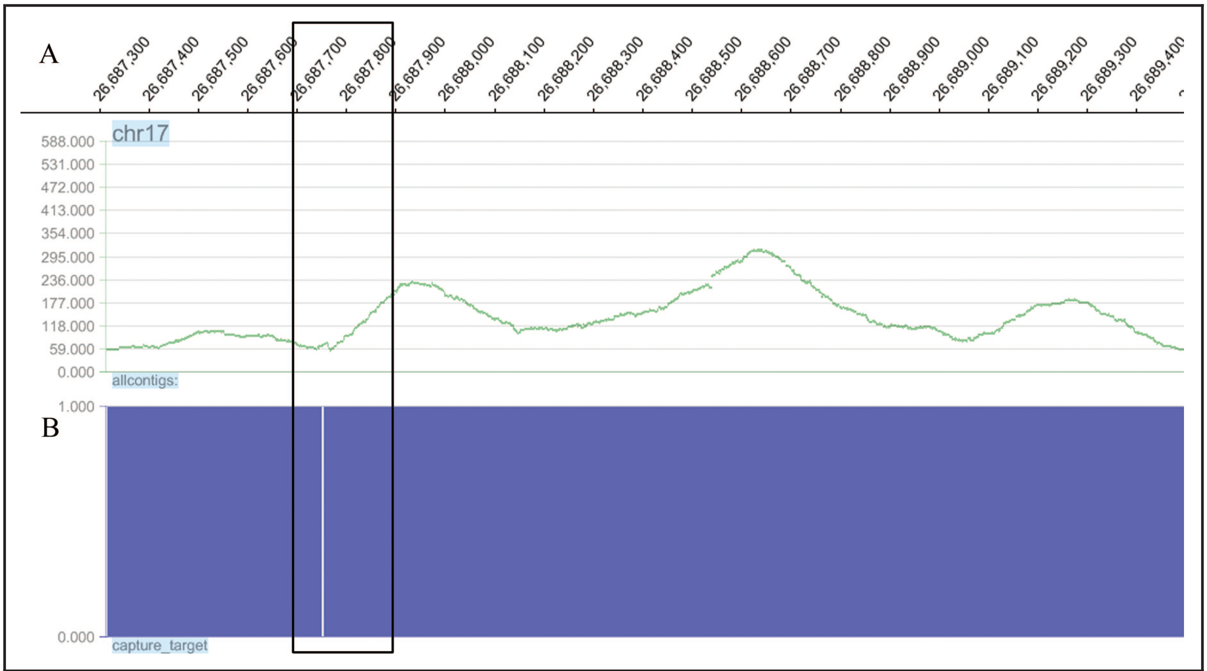


Fig. 1. Sequence coverage (after array capture and GS FLX sequencing) of sample CN1, as plotted by SignalMap (Roche NimbleGen).

(A), The green line indicates coverage of each base sequenced. The numbers across the top are base positions; the *y* axis indicates sequence coverage depth, from 0× to 588×. (B), The corresponding coverage map of the designed capture probes for the targeted region (*NF1* gene). The black box indicates the genomic location of the inserted de novo Alu sequence, which is between base-pair positions 26 687 700 and 26 687 800.

SEQUENCE ANALYSIS, SANGER SEQUENCING CONFIRMATION, AND PSEUDOGENE ALIGNMENT

To further analyze all captured sequences in both intron and exon regions, we used the Windows-based NextGENe software package to confirm the results obtained from the Linux-based 454 gsMapper software and to generate lists of all sequence variants found by the SCE-NGS method. To ensure result accuracy, we used a high stringency of alignment identity (≥95%) to align HQ reads to the reference genome. In general, the gsMapper and NextGENe software produced similar results. Sample CN1 had 58.4% of the HQ reads uniquely mapped to the reference genome, and 59.4% of them were mapped on the targeted *NF1* loci. Sample CN2 had 68.5% of the HQ reads uniquely mapped to the reference genome, and 64.0% of them were mapped on the targeted *NF1* loci. The unmapped reads to the genome (with a mean of 110 bp) are much shorter than the mean of all HQ reads, indicating these short reads are possible artifacts. All variants detected in exons and at intron/exon boundaries were confirmed by bidirectional Sanger sequencing with a capillary sequencer (ABI 3130 Genetic Analyzer; Applied Biosystems).

According to the genotyped result *(34)*, sample CN1 has a heterozygous de novo Alu sequence (approximately 483 bp) inserted 44 bp upstream of exon 42 and immediately adjacent to a poly(A/T) stretch, which causes deletion of the downstream exons during splicing. Although the probe design for the *NF1* sequence-capture microarray ignores the repetitive sequences, it does not rule out possible de novo insertions. Therefore, it is reasonable to expect partial detection of the inserted Alu sequences, either from the 3′ sequences before the insertion or the 5′ sequences after the insertion, owing to the tiling design of the *NF1* capture microarray. Both the capture probe design and the sequencing coverage map showed sequence reads in this particular region (Fig. 1), suggesting that our expectation was valid; however, we initially did not identify the heterozygous de novo Alu insertion after performing the targeted reference-mapping analysis. A possible explanation for this scenario is that the large size of the insertion might prevent alignment with the reference sequence. Therefore, in addition to reference-mapping analysis, we assembled contigs de novo for the HQ unmatched reads (remaining from the targeted reference-mapping analysis) and aligned them with the genome with low stringency (50% identity). We then partially observed the inserted Alu sequence, as we had expected (see Fig. 1 in the online Data Supplement). In addition, the number of the inserted-sequence reads (coverage) is low, which further challenges the detection of the inserted Alu sequence. We found 16 reads of the inserted alleles and 32 reads of wild-type alleles that crossed the insertion boundary. Although the inserted Alu sequence could be detected when extra care was taken, the low stringency used in the analysis generated too many nonunique mappings (structure variations) and thus led to too many false-positive calls.

In addition to the inserted Alu sequence, we also conducted a detailed investigation of all sequence variants found in sample CN1 by the SCE-NGS method (Table 2). To consider clinically important variants, we first focused on and reported changes in all exons and within ±25 bp of each exon. After filtering out all known nonpathogenic single-nucleotide polymorphisms with the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) and eliminating clinically unrelated homopolymers, we identified 17 sequence variants in sample CN1. Bidirectional Sanger sequencing of the entire *NF1* coding region confirmed 3 sequence variants, leaving 14 variants that were discrepant with the Sanger sequencing data (Table 2). Because of the inherent variation of the SCE and the pyrosequencing-based NGS platform used in this study, our second data-analysis strategy was to apply 2 parameters to determine true sequence variants. The first parameter was sequence coverage depth, and the second was the percentage of the heterozygous-allele calls. In accordance with Smith et al. *(35)* and our current data set, we decided to set the parameters to a sequence coverage depth of >30× and the percentage of heterozygous-allele calls to within 40%–60% to obtain reliable sequence data. Our data set showed that approximately 208.3 kb of 282 kb of the *NF1* gene had 30× sequence coverage. The reason that some positions either were not covered or had low coverage may be related to low capture efficiency, the exclusion of repetitive regions from the probe design, or potential variations caused by sample-processing steps, such as fragmentation and addition of linkers.

Use of these settings allowed us to filter out discrepant sequence variants detected by the SCE-NGS method that were potentially due to pyrosequencing errors. This step eliminated 10 discrepant results and left 4 undeterminable variants (Table 2). Interestingly, these 4 discrepant sequence variants were all located in the largest exon, exon 21 (441 bp). Analysis of the similarity of the exon 21 sequence of the functional *NF1* gene (on chromosome 17) to its pseudogenes (on chromosomes 2, 12, 14, 15, 18, 21, and 22) explained the remaining 4 sequencing discrepancies between the SCE-NGS and the Sanger sequencing results. Alignment analysis showed a high homology (>97%) between *NF1* exon 21 and its pseudogenes on chromosomes 12 and 15 (Table 2, Fig. 2; see Fig. 2 in the online Data Supplement). These alignment results resolved all discrepant sequence variants found in exon 21, whereas the 2 primary parameters (sequencing coverage and base change percentage) failed to filter out SCE-NGS data for potential false-positive sequence changes, even with >100× sequencing coverage (Table 2).

Sample CN2 had a single-base deletion in the coding region (exon 28) that caused a frameshift mutation *(36)*. In our SCE-NGS study, we detected the same single-base deletion, but it was in exon 37. This difference in exon location was due to differences between the current nomenclature and the nomenclature used in early 1990s. In addition to this deletion, we also detected 9 more sequence variants with the SCE-NGS method (Table 3). Bidirectional Sanger sequencing confirmed only 3 sequence variants. Of the remaining 6 discrepant sequence variants, 2 were in exon 13, and 4 were in exon 21 (Table 3). Because sequencing coverage of all 6 of these variants was >30× and because the rule of heterozygous-allele call failed to explain discrepancies between the SCE-NGS data and the Sanger sequencing data, we assumed pseudogene interference. An alignment analysis showed that sequence variants detected in exon 13 by SCE-NGS matched with pseudogenes located on chromosomes 2 and 22 (Table 3, Fig. 3). Similarly, sequence variants detected in exon 21 by SCE-NGS matched with pseudogenes located on chromosomes 12 and 15 (Table 3; Figs. 3 and 4 in the online Data Supplement). Interestingly, only 2 of 4 sequence variants (reference positions 134 285 and 134 329) detected in exon 21 of sample CN2 were the same as those detected in exon 21 of sample CN1 (Tables 2 and 3).

In the present study, all pseudogene sequences were manually checked and were successfully identified and removed from both of the tested samples (Tables 2 and 3). When this information was combined with that from the analysis that used defined parameters, all discrepancies between the SCE-NGS results and the Sanger sequencing results were resolved (100% concordance), leaving 3 true sequence variants for CN1 (in addition to the undetected Alu insertion) and 3 true sequence variants for CN2 (in addition to the single-base deletion).

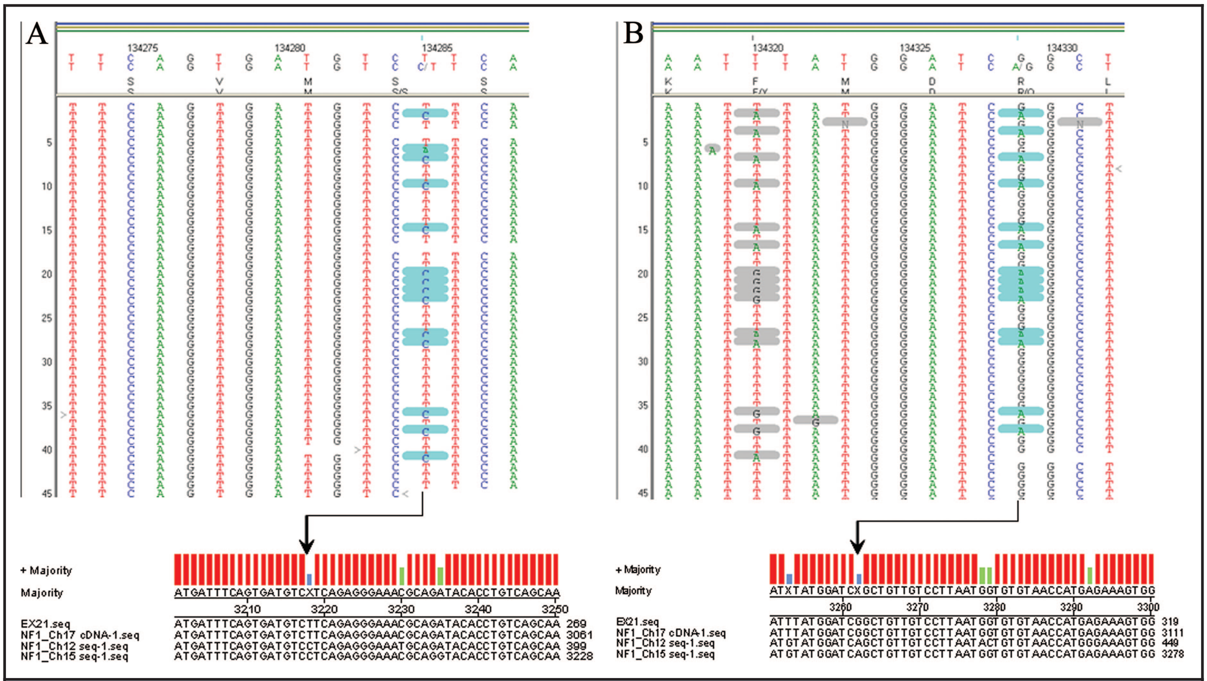
Discussion

The primary concern and challenge in applying the SCE-NGS method as a diagnostic tool are having the

Table 2. Sequence variants identified by the SCE-NGS method in sample CN1.

Reference position	CDS[a]	Reference nucleotide	Coverage	A, %	C, %	G, %	T, %	Ins, %	Del, %	Amino acid change	Sanger sequencing confirmation	Explanation
86 781	7	G	15×	100	0	0	0	0	0	234L>L	Confirmed	
132 588	20	G	51×	35	0	65	0	0	0	789K>KK	G>G homozygous (discrepant)	Base change <40%
134 285	21	T	109×	1	45	0	54	0	0	882S>SS	T>T homozygous (discrepant)	Aligned to Chr12 and Chr15 pseudogene sequences
134 329	21	G	123×	43	2	54	1	0	1	897R>QR	G>G homozygous (discrepant)	Aligned to Chr12 and Chr15 pseudogene sequences
134 431	21	C	122×	0	64	0	36	0	0	931P>PL	C>C homozygous (discrepant)	Aligned to Chr12 and Chr15 pseudogene sequences
134 432	21	A	122×	64	0	36	0	0	0	931P>PP	A>A homozygous (discrepant)	Aligned to Chr12 and Chr15 pseudogene sequences
134 959	22	G	82×	37	0	63	0	0	0	984G>EG	G>G homozygous (discrepant)	Base change <40%
135 005		A	77×	60	35	5	0	0	0		A>A homozygous (discrepant)	Base change <40%
135 854		A	23×	61	0	39	0	0	0		A>A homozygous (discrepant)	Coverage <30
135 862		A	25×	64	0	0	36	0	0		A>A homozygous (discrepant)	Coverage <30
135 875	24	G	24×	0	0	58	42	0	0	1041M>IM	G>G homozygous (discrepant)	Coverage <30
135 959		G	13×	15	0	38	0	0	46		G>G homozygous (discrepant)	Coverage <30
135 960		A	14×	57	0	0	0	0	43		A>A homozygous (discrepant)	Coverage <30
135 966		C	14×	0	36	0	64	0	0		C>C homozygous (discrepant)	Coverage <30
135 972		G	12×	75	0	25	0	0	0		G>G homozygous (discrepant)	Coverage <30
231 299		T	202×	0	51	0	49	0	0		Confirmed	
232 882		T	250×	41	1	0	58	0	0		Confirmed	

[a] CDS, coding sequence; Ins, insertion; Del, deletion; Chr, chromosome.



```
+ Majority
Majority               ATGATTTCAGTGATGTCXTCAGAGGGAAACGCAGATACACCTGTCAGCAA
                                 3210      3220      3230      3240      3250
EX21.seq               ATGATTTCAGTGATGTCTTCAGAGGGAAACGCAGATACACCTGTCAGCAA 269
NF1_Ch17 cDNA-1.seq    ATGATTTCAGTGATGTCTTCAGAGGGAAACGCAGATACACCTGTCAGCAA 3061
NF1_Ch12 seq-1.seq     ATGATTTCAGTGATGTCCTCAGAGGGAAATGCAGATACACCTGTCAGCAA 399
NF1_Ch15 seq-1.seq     ATGATTTCAGTGATGTCCTCAGAGGGAAACGCAGGTACACCTGTCAGCAA 3228
```

```
+ Majority
Majority               ATXTATGGATCXGCTGTTGTCCTTAATGGTGTGTAACCATGAGAAAGTGG
                                 3260      3270      3280      3290      3300
EX21.seq               ATTTATGGATCGGCTGTTGTCCTTAATGGTGTGTAACCATGAGAAAGTGG 319
NF1_Ch17 cDNA-1.seq    ATTTATGGATCGGCTGTTGTCCTTAATGGTGTGTAACCATGAGAAAGTGG 3111
NF1_Ch12 seq-1.seq     ATGTATGGATCAGCTGTTGTCCTTAATACTGTGTAACCATGGGAAAGTGG 449
NF1_Ch15 seq-1.seq     ATGTATGGATCAGCTGTTGTCCTTAATGGTGTGTAACCATGAGAAAGTGG 3278
```

Fig. 2. Sequence-alignment analysis of exon 21 of sample CN1 for the NF1functional gene and the pseudogenes on chromosomes 12 and 15.

NextGENe software indicated the positions (arrows) of the sequence variants detected by the SCE-NGS method: 134 285 T>C (A) and 134 329 G>A (B). The detected variants are from the pseudogenes, not from the functional gene. For each panel, the top graph (produced by the NextGENe software) shows sequences of each read, and sequence coverage is indicated on the *y* axis. The blue boxes highlight the detected sequence changes, and the gray boxes indicate detected sequence changes in which the percentage change is not sufficiently high to make a call. The sequence-alignment analysis below each graph is produced by Lasergene software (version 8.0; DNASTAR) and shows alignment of the sequences of interest in exon 21 [134 285 T>C (A) and 134 329 G>A (B)] for the functional gene (on chromosome 17) and the pseudogenes (on chromosomes 12 and 15).

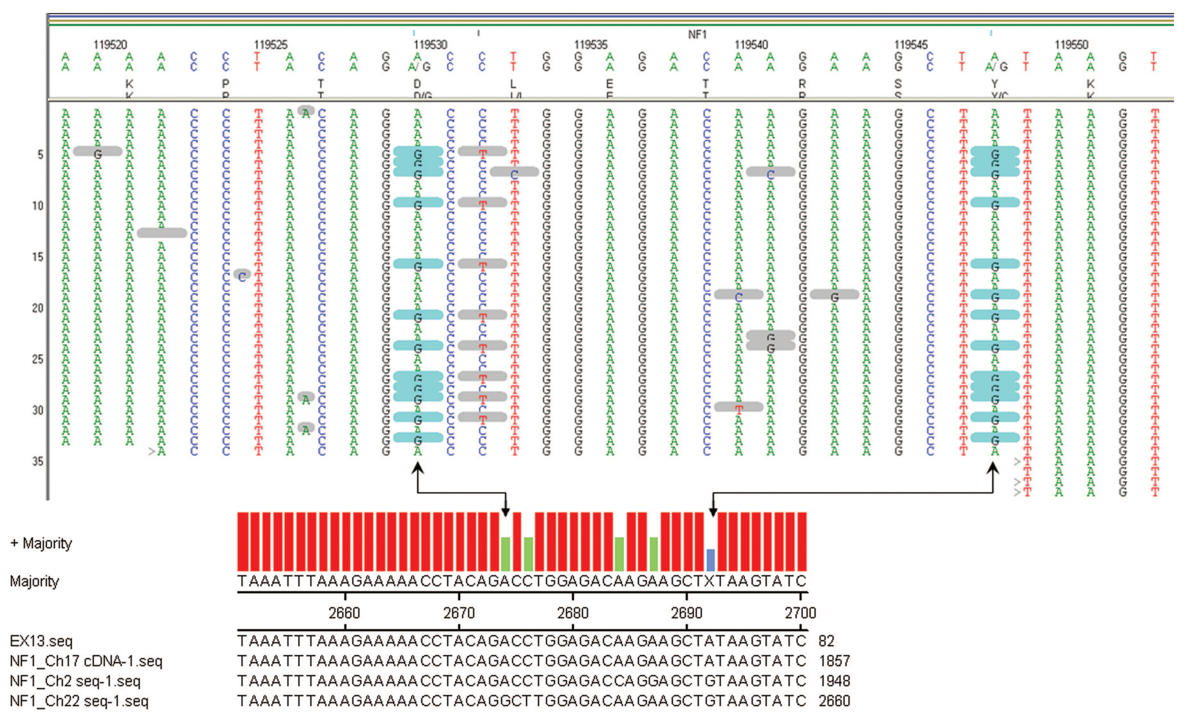
capability to detect large de novo insertions, especially direct-repeat sequences (i.e., Alu) in a heterozygous format. Although insertions are not commonly present in all genetic diseases, insertions are present in relatively high percentages in patients with neurofibromatosis type 1 *(37)*. In our study, we were able to detect the heterozygous inserted Alu sequence by first performing reference-mapping with all HQ reads, followed by de novo assembly with low alignment stringency (50%) with the remaining unmatched reads. In other words, the inserted Alu sequence could have been missed if only the reference-mapping analysis had been used. This circumstance could have led to a potential false-negative clinical interpretation. Use of de novo assembly with low alignment stringency helped to detect the inserted Alu sequence, but it also generated too many false-positive calls.

A second concern with the SCE-NGS method is the lack of definitive parameters for QC of the sequence data. For clinical applications, the major hurdle in the use of the SCE-NGS method is how to set a reliable cutoff threshold for determining true sequence variants when multiple variants from the reference sequence are detected. Although the requirement for sequencing coverage to obtain reliable results has previously been investigated *(35)*, we found no recommendations regarding a threshold for confidently establishing a heterozygous-allele call when heterozygous sequence variants are detected. Such a threshold value needs to be established to rule out potential pyrosequencing errors, and we have established a range of 40%–60% for heterozygous-allele calls based on our data set. For the discrepant cases with heterozygous-allele calls less than but close to 40%, 2 possible factors might cause this scenario. First, pyrosequencing may exhibit different systematic errors compared with other NGS technologies. Second, sequence artifacts, such as short reads, homopolymers, and the presence of pseudogenes, can interfere with the percentage of heterozygous-allele calls. A large sample set needs to be studied before specific QC guidelines can be recommended for the use of this new technology in routine

Table 3. Sequence variants identified by the SCE-NGS method in sample CN2.

Reference position	CDS[a]	Reference nucleotide	Coverage	A, %	C, %	G, %	T, %	Ins, %	Del, %	Amino acid change	Sanger sequencing confirmation	Explanation
86 781	7	G	52×	100	0	0	0	0	0	234L>L	Confirmed	
119 530	13	A	38×	58	0	42	0	0	0	483D>DG	A>A homozygous (discrepant)	Aligned to Chr22 pseudogene sequence
119 548	13	A	37×	57	0	43	0	0	0	489Y>YC	A>A homozygous (discrepant)	Aligned to Chr2 and Chr22 pseudogene sequences
134 267	21	T	46×	0	39	0	61	0	0	876S>SS	T>T homozygous (discrepant)	Aligned to Chr15 pseudogene sequence
134 285	21	T	43×	0	49	0	51	0	0	882S>SS	T>T homozygous (discrepant)	Aligned to Chr12 and Chr15 pseudogene sequences
134 329	21	G	56×	41	0	59	0	0	0	897R>QR	G>G homozygous (discrepant)	Aligned to Chr12 and Chr15 pseudogene sequences
134 396	21	T	60×	0	35	0	65	0	0	919D>DD	T>T homozygous (discrepant)	Aligned to Chr15 pseudogene sequence
231 081	37	G	184×	0	0	55	0	0	45	FS	Confirmed	
231 299		T	106×	0	98	0	2	0	0		Confirmed	
232 882		T	169×	99	0	0	1	0	0		Confirmed	

[a] CDS, coding sequence; Ins, insertion; Del, deletion; Chr, chromosome.



```
Majority                TAAATTTAAAGAAAAACCTACAGACCTGGAGACAAGAAGCTXTAAGTATC
                                  2660      2670      2680      2690      2700
EX13.seq                TAAATTTAAAGAAAAACCTACAGACCTGGAGACAAGAAGCTATAAGTATC  82
NF1_Ch17 cDNA-1.seq     TAAATTTAAAGAAAAACCTACAGACCTGGAGACAAGAAGCTATAAGTATC  1857
NF1_Ch2 seq-1.seq       TAAATTTAAAGAAAAACCTACAGACCTGGAGACCAGGAGCTGTAAGTATC  1948
NF1_Ch22 seq-1.seq      TAAATTTAAAGAAAAACCTACAGGCTTGGAGACAAGAAGCTGTAAGTATC  2660
```

Fig. 3. Sequence-alignment analysis of exon 13 of sample CN2 for the NF1functional gene and the pseudogenes on chromosomes 2 and 22.

NextGENe software indicate the positions (arrows) of sequence variants detected by the SCE-NGS method (119 530 A>G; 119 548 A>G). The detected variants are from the pseudogenes, not from the functional gene.

diagnostics laboratories. At this time, routine Sanger sequencing is still required to confirm data generated by NGS.

When multiple pseudogenes (or highly homologous sequences) exist, the algorithm used for data analysis is particularly important in studies that use a sample-enrichment process with capture-based technologies and highly sensitive NGS for gene sequence analysis. In our study, the initial whole-chromosome mapping analysis of both CN1 and CN2 showed statistically nonsignificant sequences captured from *NF1* pseudogenes; however, we found that all of the discrepancies between SCE-NGS and Sanger sequencing data were due to pseudogene interference. Such interference suggests that NGS substantially amplified the signals from trace nonspecific sequences captured by SCE and thereby decreased the clinical specificity of the SCE-NGS method compared with Sanger sequencing. Before spending efforts to redesign capture probes and to reoptimize target–probe hybridization conditions, we explored the use of a software approach to eliminate interference by pseudogenes (or highly homologous sequences). We applied a data-analysis algorithm, dual-mode reference-mapping (DMRM), to the CN1 and CN2 data sets to filter out all detected sequence variants originating from pseudogenes. Basically, the first mode of this analysis algorithm is to align all 454 HQ reads (from both samples) to the reference genome, one chromosome at a time. Reads that match perfectly (≥50 bp with 95% identity) to multiple locations are allowed to match to all possible locations in the entire genome. This mode generates 25 separated reference-sequence files (with much reduced data sizes) containing sequences specific to each chromosome (including X and Y) and chromosome 17 for *NF1* sequences only. The second mode is then to realign all 454 HQ reads simultaneously with all newly generated reference-sequence files (with reduced sizes). Taking advantage of the relatively long sequence read length produced by the GS FLX system, we identified the genomic signature of each highly homologous sequence captured (from different chromosomes) and removed the sequences. Once these sequences are removed, the true sequence variants (if any) detected by the SCE-NGS method can be determined. Our initial data analysis of CN1 and CN2 showed that the DMRM algorithm successfully identified and removed all pseudogene sequences. This procedure resolved all remaining discrepancies between the SCE-NGS data and the Sanger sequencing confirmation data. In addition to reducing the risk of reporting false positives due to interference by highly homologous sequences, the reference files (with reduced sizes) generated from the first mode can be reused to analyze other *NF1* SCE-NGS data in much less time, which is another advantage of the DMRM. Further testing of the DMRM analysis is currently underway with more SCE-NGS data to reveal its utility.

In conclusion, by avoiding massive PCR preparation and sequencing of multiple large regions of interest simultaneously, the SCE-NGS method has potential for use in clinical settings. Compared with other short-read NGS platforms (i.e., the Applied Biosystems SOLiD and Illumina GAII systems), the long sequence reads provided by the GS FLX system not only allows better assembly for detecting large insertions (especially repetitive sequences) but also allows easier identification of genomic signatures from highly homologous sequences. This study evaluated the utility of microarrays for targeted sequence capture and addressed the issues regarding data management and interference by highly homologous sequences (e.g., pseudogenes) through development of a step-by-step data-analysis strategy and the prototype of DMRM analysis. In addition, we recommend further studies to compare the microarray capture method with solution phase–based enrichment methods, such as Agilent's SureSelect.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures of Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:*

Employment or Leadership: C.-S.J. Liu, SoftGenetics; B. Boese, Roche/454 Life Sciences.
Consultant or Advisory Role: None declared.
Stock Ownership: C.-S.J. Liu, SoftGenetics.
Honoraria: None declared.
Research Funding: We thank the Institute for Clinical and Experimental Pathology, ARUP Laboratories (Salt Lake City, UT) for funding this work.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Acknowledgments: We thank Jacquelyn McCowen-Rose and Alison Millson for critical review of this manuscript.

References

1. Asmann YW, Wallace MB, Thompson EA. Transcriptome profiling using next-generation sequencing. Gastroenterology 2008;135:1466–8.

2. Holt RA, Jones SJ. The new paradigm of flow cell sequencing. Genome Res 2008;18:839–46.

3. Marguerat S, Wilhelm BT, Bahler J. Next-generation sequencing: applications beyond genomes. Biochem Soc Trans 2008;36:1091–6.

4. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics 2008;92:255–64.

5. Wheat CW. Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica [Epub ahead of print 2008 Oct 18].

6. Aartsma-Rus A, van Ommen GJ. Antisense-mediated exon skipping: a versatile tool with therapeutic and research applications. RNA 2007; 13:1609–24.

7. Madden HR, Fletcher S, Davis MR, Wilton SD. Characterization of a complex Duchenne muscular dystrophy-causing dystrophin gene inversion and restoration of the reading frame by induced exon skipping. Hum Mutat 2009;30:22–8.

8. Wee KB, Pramono ZA, Wang JL, MacDorman KF, Lai PS, Yee WC. Dynamics of co-transcriptional pre-mRNA folding influences the induction of dystrophin exon skipping by antisense oligonucleotides. PLoS One 2008;3:e1844.

9. Wilton SD, Fletcher S. Redirecting splicing to address dystrophin mutations: Molecular by-pass surgery. Prog Mol Subcell Biol 2006;44:161–97.

10. ten Bosch JR, Grody WW. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. J Mol Diagn 2008;10: 484–92.

11. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations. Science 2005;307:1072–9.

12. Mullis KB, Faloona FA. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 1987;155:335–50.

13. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268–74.

14. Bau S, Schracke N, Kränzle M, Wu H, Stähler PF, Hoheisel JD, et al. Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. Anal Bioanal Chem 2009;393:171–5.

15. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007;4:903–5.

16. Antson DO, Isaksson A, Landegren U, Nilsson M. PCR-generated padlock probes detect single nucleotide variation in genomic DNA. Nucleic Acids Res 2000;28:E58.

17. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009;27:182–9.

18. Larsson C, Koch J, Nygren A, Janssen G, Raap AK, Landegren U, Nilsson M. In situ genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes. Nat Methods 2004;1:227–32.

19. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007; 39:1522–7.

20. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, et al. Multiplex amplification of large sets of human exons. Nat Methods 2007;4: 931–6.

21. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet 2008;24:133–41.

22. Viskochil D, Buchberg AM, Xu G, Cawthon RM, Stevens J, Wolff RK, et al. Deletions and a translocation interrupt a cloned gene at the neurofibromatosis type 1 locus. Cell 1990;62:187–92.

23. Xu GF, O'Connell P, Viskochil D, Cawthon R, Robertson M, Culver M, et al. The neurofibromatosis type 1 gene encodes a protein related to GAP. Cell 1990;62:599–608.

24. Lee MJ, Stephenson DA. Recent developments in neurofibromatosis type 1. Curr Opin Neurol 2007; 20:135–41.

25. Suzuki Y, Suzuki H, Kayama T, Yoshimoto T, Shibahara S. Brain tumors predominantly express the neurofibromatosis type 1 gene transcripts containing the 63 base insert in the region coding for GTPase activating protein-related domain. Biochem Biophys Res Commun 1991;181:955–61.

26. Li Y, O'Connell P, Breidenbach HH, Cawthon R, Stevens J, Xu G, et al. Genomic organization of the neurofibromatosis 1 gene (NF1). Genomics 1995;25:9–18.

27. Kehrer-Sawatzki H, Schwickardt T, Assum G, Rocchi M, Krone W. A third neurofibromatosis type 1 (NF1) pseudogene at chromosome 15q11.2. Hum Genet 1997;100:595–600.

28. Luijten M, Redeker S, Minoshima S, Shimizu N, Westerveld A, Hulsebos TJ. Duplication and transposition of the NF1 pseudogene regions on chromosomes 2, 14, and 22. Hum Genet 2001;109: 109–16.

29. Suzuki H, Ozawa N, Taga C, Kano T, Hattori M, Sakaki Y. Genomic analysis of a NF1-related pseudogene on human chromosome 21. Gene 1994;147:277–80.

30. Roche NimbleGen. Technical note: Roche NimbleGen probe design fundamentals. http://www.nimblegen.com/products/lit/probe_design_2008_06_04.pdf (Accessed November 2009).

31. Roche NimbleGen. NimbleGen services user's guides: sequence capture service. http://www.nimblegen.com/products/lit/SeqCap_UsersGuide_Service_v3p0.pdf (Accessed November 2009).

32. Roche Diagnostics, Roche Applied Science. GS FLX shotgun DNA library preparation method manual: December 2007. http://www.unc.edu/htsf/files/GSFLXShotgunDNA_Library_Preparation_Method_Manual_CD.pdf (Accessed November 2009).

33. Roche Diagnostics, Roche Applied Science. GS FLX emPCR method manual: December 2007. ftp://ftp.genome.ou.edu/pub/for_broe/FLX/9.%20GS%20FLX%20emPCR%20Method%20Manual.pdf (Accessed November 2009).

34. Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. A de novo Alu insertion results in neurofibromatosis type 1. Nature 1991; 353:864–6.

35. Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 2008;18: 1638–42.

36. Colman SD, Collins FS, Wallace MR. Characterization of a single base-pair deletion in neurofibromatosis type 1. Hum Mol Genet 1993;2: 1709–11.

37. Wimmer K, Yao S, Claes K, Kehrer-Sawatzki H, Tinschert S, De Raedt T, et al. Spectrum of single- and multiexon NF1 copy number changes in a cohort of 1,100 unselected NF1 patients. Genes Chromosomes Cancer 2006;45: 265–76.