Appendix S1

Patients and DNA sampling. Samples of 16 unrelated NF1 patients were collected through the NF1 clinic, Department of Dermatology, Turku University Hospital, Finland, or through the Finnish NF Patient Association. The study was carried out with the approval of the Ethics Committee of the Hospital District of Southwest Finland and the research permission of Turku University Hospital. The samples were collected with the informed consent of the patients. All patients were of Finnish ancestry and fulfilled the NIH diagnostic criteria for NF1. Of all patients, 4 had an inherited mutation, while 12 patients did not have a known family history of NF1. The *NF1* mutations of 6 patients were previously established for clinical indications in diagnostic laboratories abroad. In addition, one patient had a previously detected *NF1* type 2 microdeletion.

Genomic DNA was isolated from saliva using a saliva sampling kit (OG-575, Oragene, DNA Genotek, Inc., Ottawa, Ontario, Canada). Contaminating RNA was degraded using RNase treatment at 37°C for 30 min (RNase Cocktail Enzyme mix, Ambion Inc., Applied Biosystems, Foster City, CA, USA, AM2286). The DNA was further purified using a gDNA isolation kit (NucleoSpin® Tissue, Macherey-Nagel GmbH & Co., Düren, Germany, 740952). The quality of the DNA was evaluated on agarose gels, the NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA, ND-1000), and 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA, G2939AA). The gDNA was quantified using the Quant-iT Picogreen assay (Life Technologies, Eugene, Oregon, USA, P11496) and Plate Chameleon V fluorometer (Hidex, Turku, Finland, 425-106).

Sample library preparation. A total of 16 indexed Rapid libraries were prepared according to the manufacturer's instructions starting from 500 ng of gDNA per sample (Rapid Library Preparation Method Manual, GS Junior Titanium Series, May 2010 and March 2012, Roche). The gDNA was fragmented by nebulisation and the fragments were end-repaired, leaving a single A overhang in the 5' end of each strand. Subsequently, indexed adaptors were ligated to the fragment ends. The library preparation was completed by removing small DNA fragments with the Agencourt Ampure XP reagent (Ordior Inc, Helsinki, Finland, A63880). The final preparation contained less than 10% of fragments < 350 bp.

Sequence capture and sequencing. The data presented in this article is derived from 2 separate sequence captures and sequencing runs, referred to sets A and B. Set A contained 10 and Set B 6 libraries per set, respectively. The sample libraries were amplified and pooled into 2 sets prior to the sequence capture. The total amount of DNA in both sets was 1 µg. Samples with known mutations were used as control samples for the new sequencing approach. Set A contained 3 controls and 7 samples with unknown mutations, and set B had 4 controls and 2 samples with unknown mutations. The 2 different sets were sequenced to search for the ideal number of samples per sequencing run.

The *NF1* exome was enriched using a sequence capture with custom probe design, and a modified NimbleGen SeqCap EZ Choice Library protocol optimised for the enrichment of target areas < 100 kb (Roche NimbleGen SeqCap EZ Rapid Library Small Target Capture LR, December 2009, Roche Nimblegen Inc., Madison, WI, USA). The targeted regions included 58 *NF1* exons and an additional 50 bp of flanking upstream and downstream intronic sequences, resulting in a total target size of 16 kb. The target region did not include alternatively spliced exon 10a2, central nervous system-specific exon 9b, or exon 48a expressed in muscles. To our knowledge, disease-causing mutations have not been discovered in these regions.

The enrichment of the target region in sequence capture was assessed by qPCR by measuring the relative abundance of 4 control targets in amplified sample library and amplified captured DNA. The control assays have been validated on NimbleGen Sequence Capture arrays, but the control sequences are not revealed to the customer. The captured DNA was quantified using the Picogreen reagent, and the fragment sizes were measured with BioAnalyzer before and after capture. The average lengths of captured fragments were 757 bp and 727 bp, in the sets of A and B, respectively. The emulsion PCR and sequencing were performed separately for each set according to the manufacturer's instructions (emPCR Amplification Method Manual - Lib-L, and Sequencing Method Manual, GS Junior Titanium Series, May 2010 and March 2012, Roche) using 2720 Thermal cycler (Life Technologies, Eugene, Oregon, USA, 435965), and the 454 GS Junior instrument (454 Life Sciences, Branford, CT, USA). Following the sequencing run, the reads were automatically quality filtered using the default settings of GS Run Processor. Only reads that passed all the quality filters were used in the analysis. The number of reads passing the quality-filtering exceeded 100,000 which is the cut-off limit for a successful run (Sequencing Method Manual, GS Junior Titanium Series, May 2010 and March 2012, Roche). Furthermore, over 60% of the control reads had less than 5% errors in the first 400 bp, consistent with the criteria for a successful sequencing run.

Data analysis. Control samples were analysed blindly in parallel with samples of which the mutations were unknown. Sequencing reads in sff (standard flowgram format) file format obtained from the 454 GS Junior sequencing run were converted to fastq format using a freely available sff-extract script (version 0.3.0) (14). The reads were mapped to the whole human genome (UCSC hg19, NCBI build 37) using the Bowtie 2 alignment program (version 2.0.0-beta7) (15). The mapping was performed as an end-to-end alignment using the "very-sensitive" option in order to minimise the risk of mapping pseudogene sequences to the *NF1* gene. Samtools (version 0.1.18) (16), Picard (version 1.78) (17) and Biopython (version 1.60) (18) software packages were used for file handling and making files compatible with the different programs used in the analysis pipeline. The sequencing coverage across target and non-target areas was calculated using BEDTools (version 2.16.2) (19) and R (version 2.15.1) (20, 21). The alignments were visualised with the Integrative Genomics Viewer (IGV, version 2.0.34 1623) (22).

Identification of mutations. Putative mutations in the region of the *NF1* gene were identified from the sequence alignments using GATK UnifiedGenotyper (version 2.1–11) (23) with settings that allow for both single nucleotide polymorphisms (SNPs) and insertions/deletions to be called. The main parameters influencing the power of detecting heterozygous variants are the coverage and selection of the used sample variant frequency. The variants were further filtered by the following criteria: 1) Minimum coverage of 20× or higher, 2) sample variant frequency between 30–70%. If the combination of values 20× coverage and minimum variant frequency 30% are used, 96.8% of the heterozygous variants are detected (24, 25), 3) Variant in a targeted region (Fig. S2[1]), and 4) variant present in only one sample within a sample set with a frequency of 30% or higher. The fourth criterion was based on the expectation that 16 unrelated NF1 patients were not likely to have shared mutations. This criterion also excluded possible SNPs present in the Finnish population, as well as a majority of the false positive insertions/deletions in homopolymer sequences, which are a common type of error in 454 sequencing data (26, 27). The variants which passed all criteria were compared to the dbSNP database build 135 (28) and Finnish "The Sequencing Initiative Suomi" database (29). Variants were also checked whether they were supported by evidence from reads mapped to both the sense and antisense strands.

Sanger sequencing. Verification of discovered variants was carried out using Sanger sequencing. Fragments with putative mutations or potential mutations in homopolymer regions were amplified by PCR and sequenced with Applied Biosystems 3130xl Genetic Analyzer in DNA Sequencing Service of Turku Centre for Biotechnology, Turku, Finland.

Appendix S2

ACKNOWLEDGEMENTS

We thank Eric Legius for fruitful comments. Miso Immonen is acknowledged for the valuable help with the patient samples, Paula Pennanen for the help with methodology and Linda Nissi for the help with confirmation of the found mutations. The computing resources used in the data analysis were provided by CSC – IT Center for Science, Espoo, Finland. The Sanger sequencing was done by DNA Sequencing Service of Turku Centre for Biotechnology. The Sequencing Initiative Suomi The SISu project is an international collaboration between research groups aiming to build tools for genomic medicine. These groups are generating whole genome and whole exome sequence data from Finnish samples and provide data resources for the research community. Key groups of the project are from UCLA, Michigan University, The Broad Institute of MIT and Harvard, Oxford University, The Wellcome Trust Sanger Institute, Lund University, University of Helsinki, University of Eastern Finland and The Institute of Health and Welfare, Finland.

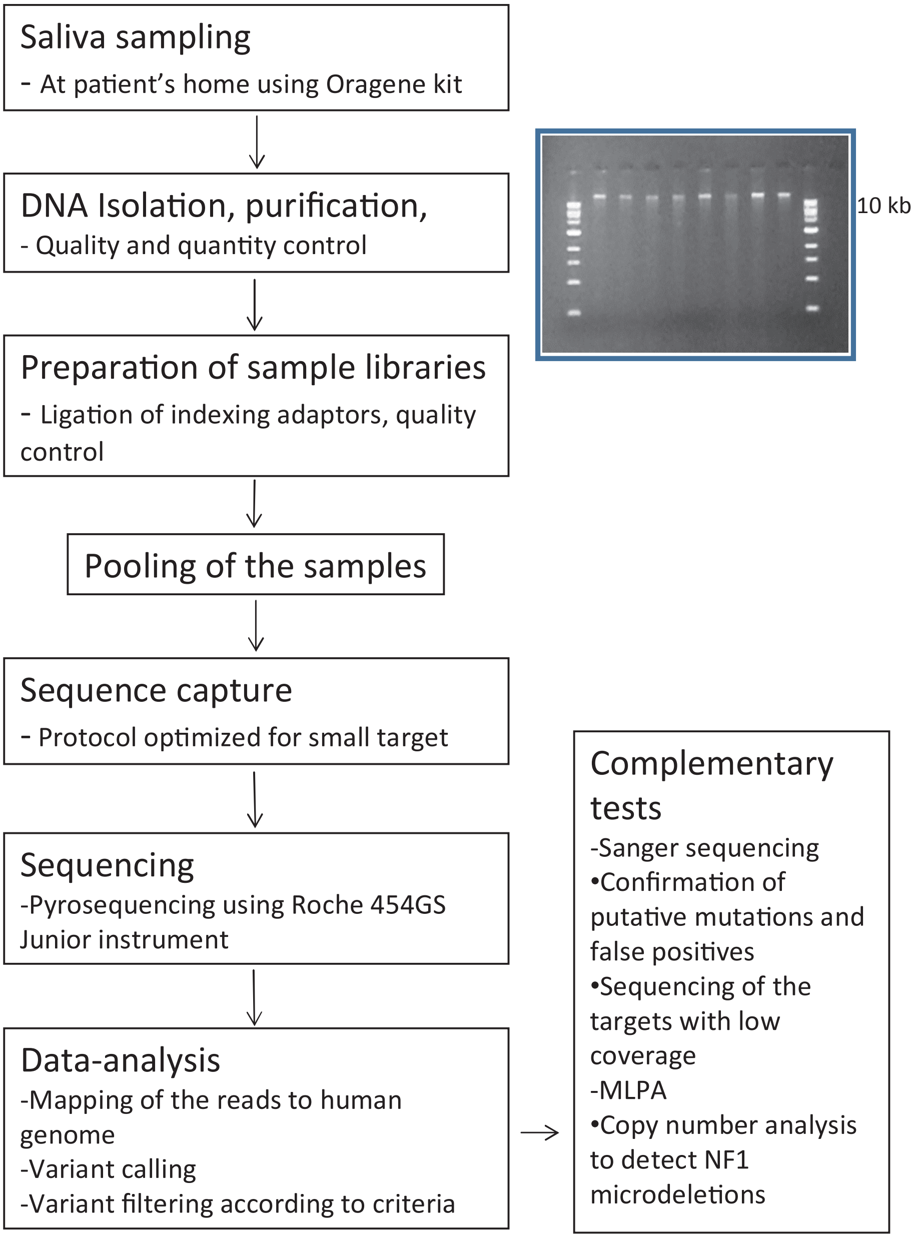
Saliva sampling
- At patient's home using Oragene kit
DNA Isolation, purification,
- Quality and quantity control
Preparation of sample libraries
- Ligation of indexing adaptors, quality control
Pooling of the samples
Sequence capture
- Protocol optimized for small target
Sequencing
-Pyrosequencing using Roche 454GS Junior instrument
Data-analysis
-Mapping of the reads to human genome
-Variant calling
-Variant filtering according to criteria
Complementary tests
-Sanger sequencing
•Confirmation of putative mutations and false positives
•Sequencing of the targets with low coverage
-MLPA
•Copy number analysis to detect NF1 microdeletions
10 kb

Fig. S1. Pipeline for neurofibromatosis type 1 (NF1) mutation analysis. Quality of the isolated saliva DNA is shown in electrophoretic analysis on agarose gel.

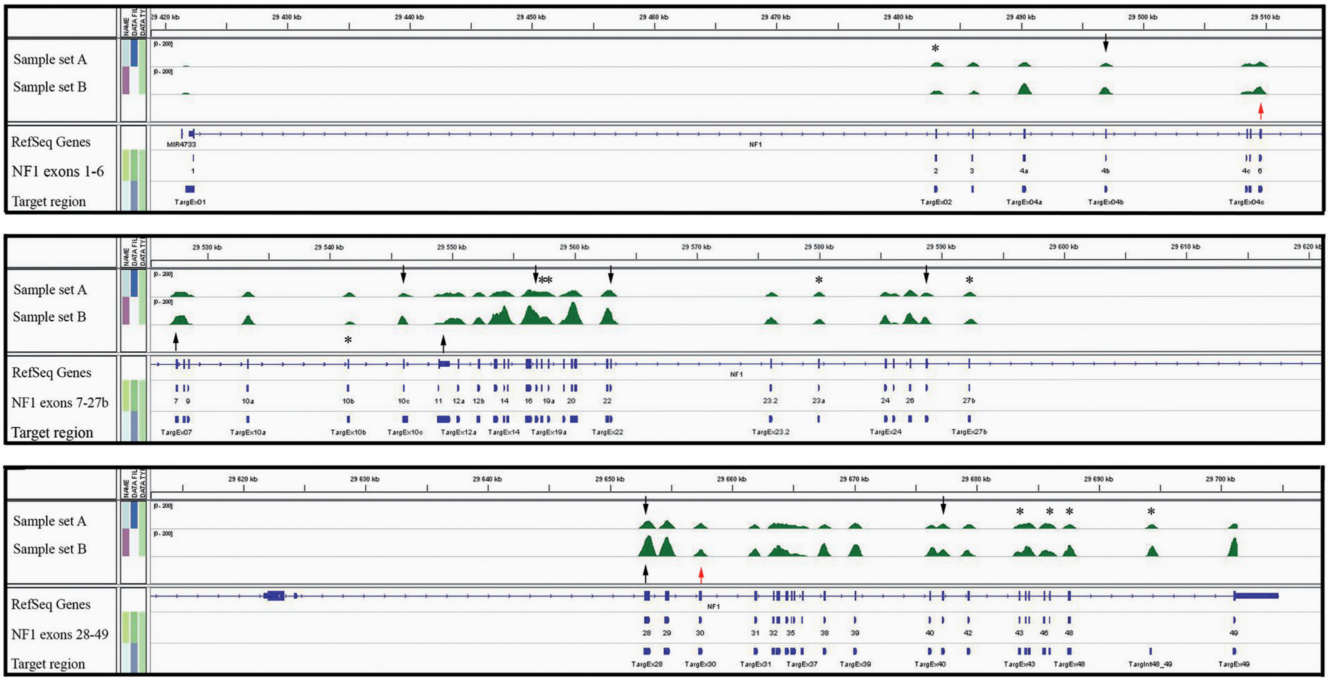

Fig. S2. The coverage across target regions. The coverage in the Y axes (values between 0–200) and the target region (exons 1–6, 7–27b, 28–49) in the X axes for both sample sets A and B. In addition to the exon 1, there were no other regions with constantly low coverage. Black arrows indicate positions of confirmed mutations, red arrows are positions of false negative and asterisk indicate positions of false positive mutations.

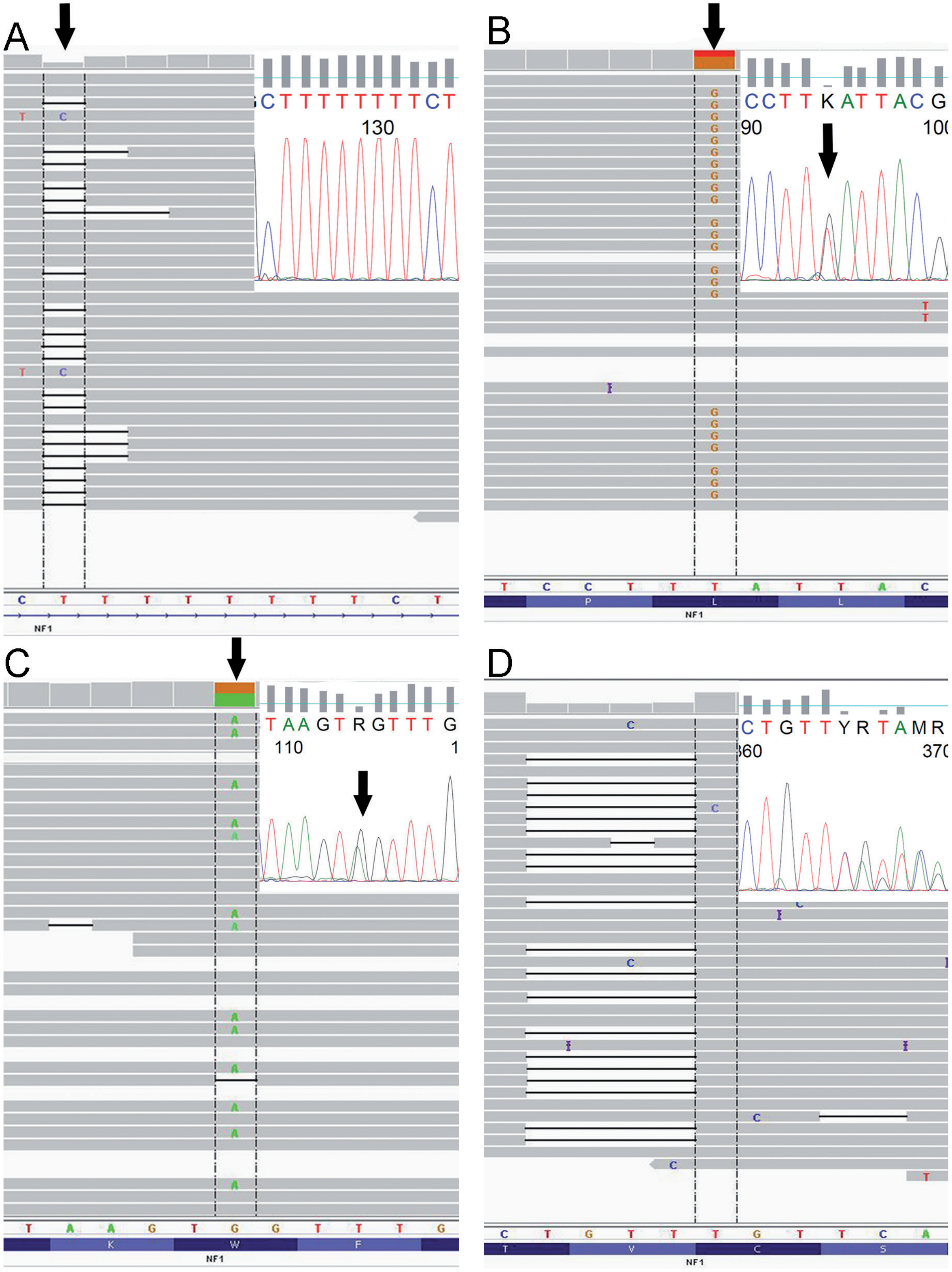

Fig. S3. Mutations found in sequencing with Roche 454 GS Junior as illustrated in Integrative Genomics Viewer (IGV). Inserts: Sanger sequencing of the corresponding area. (A) E579 intronic delT, a homopolymer-related false positive. Confirmed mutations: (B) E396 c.3911T>G, (C) S47 c.4922G>A, (D) E71 c.499_502delTGTT.

Table SI. *The number of reads for each sample*

Sample	Total number of reads	Number of mapped reads	Percentage of mapped reads
Sample set A			
S96	8,816	8,604	97.60
E46	16,090	15,842	98.46
E396	8,241	8,025	97.38
S65	13,672	13,484	98.63
S594	8,023	7,800	97.22
E579	12,498	12,293	98.36
E13	10,668	10,498	98.41
S47	10,870	10,671	98.17
E71	16,783	13,132	78.25
E66	9,475	9,326	98.43
Sample set B			
S49	17,990	17,721	98.50
S97	18,933	18,671	98.62
E27	29,886	28,274	94.61
S2122	25,944	25,629	98.78
E39	13,984	13,800	98.68
E38	22,803	22,306	97.82

Table SII. *Overview of the sequencing results: size of reads, numbers of passed and mapped reads, mean coverage, coverage SD and number of bases <20×*

	Average size of reads (bp)	Passed reads	Mapped reads	Mean coverage	Coverage SD	Bases <20× *n* (%)
Set A (10 samples)	403	115,519	109,675	41	23	19,399 (12.07)
Set B (6 samples)	408	130,293	126,401	74	59	19,014 (19.73)