ORIGINAL ARTICLE

Next generation diagnostics of cystic fibrosis and *CFTR*-related disorders by targeted multiplex high-coverage resequencing of *CFTR*

D Trujillano,[1,2,3,4] M D Ramos,[5] J González,[1,2,3,4] C Tornador,[1,2,3,4] F Sotillo,[5] G Escaramis,[1,2,3,4] S Ossowski,[6,2] L Armengol,[7] T Casals,[5] X Estivill[4,1,2,3]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2013-101602).

[1]Genetic Causes of Disease Group, Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
[3]Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain
[4]CIBER in Epidemiology and Public Health (CIBERESP), Barcelona, Catalonia, Spain
[5]Human Molecular Genetics Group, IDIBELL, L'Hospitalet de Llobregat, Barcelona, Catalonia, Spain
[6]Genomic and Epigenomic Variation in Disease Group, Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain
[7]qGENOMICS, Quantitative Genomic Medicine Laboratories SL, Barcelona, Catalonia, Spain

Correspondence to
Dr X Estivill, Genetic Causes of Disease Group, Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Doctor Aiguader 88, Barcelona, Catalonia 08003, Spain; xavier.estivill@crg.cat

Received 13 February 2013
Revised 28 March 2013
Accepted 16 April 2013
Published Online First
17 May 2013

To cite: Trujillano D, Ramos MD, González J, *et al. J Med Genet* 2013;**50**:455–462.

ABSTRACT

Background Here we have developed a novel and much more efficient strategy for the complete molecular characterisation of the cystic fibrosis (CF) transmembrane regulator (*CFTR*) gene, based on multiplexed targeted resequencing. We have tested this approach in a cohort of 92 samples with previously characterised *CFTR* mutations and polymorphisms.

Methods After enrichment of the pooled barcoded DNA libraries with a custom NimbleGen SeqCap EZ Choice array (Roche) and sequencing with a HiSeq2000 (Illumina) sequencer, we applied several bioinformatics tools to call mutations and polymorphisms in *CFTR*.

Results The combination of several bioinformatics tools allowed us to detect all known pathogenic variants (point mutations, short insertions/deletions, and large genomic rearrangements) and polymorphisms (including the poly-T and poly-thymidine-guanine polymorphic tracts) in the 92 samples. In addition, we report the precise characterisation of the breakpoints of seven genomic rearrangements in *CFTR*, including those of a novel deletion of exon 22 and a complex 85 kb inversion which includes two large deletions affecting exons 4–8 and 12–21, respectively.

Conclusions This work is a proof-of-principle that targeted resequencing is an accurate and cost-effective approach for the genetic testing of CF and *CFTR*-related disorders (ie, male infertility) amenable to the routine clinical practice, and ready to substitute classical molecular methods in medical genetics.

INTRODUCTION

Cystic fibrosis (CF; MIM #219700) is one of the most common, life-threatening, autosomal recessive genetic disorders, with a carrier frequency in the Caucasian population of around 1 in 20–80 people.[1] Mutations in the CF transmembrane conductance regulator (*CFTR/ABCC7*; MIM #602421) gene determine the impairment of chloride transport in epithelial cells, mainly affecting lungs, digestive tract, sweat glands and vas deferens in men.[2] Although a major mutation (deltaF508) accounts for over two-thirds of CF alleles worldwide,[3] a high level of allelic heterogeneity has been described within different CF populations,[4] including single nucleotide variants (SNVs), short insertions and deletions (InDels) and large structural variants (SVs). Since the characterisation of *CFTR* more than 20 years ago,[5–7] 1937 *CFTR* variants have been reported (Cystic Fibrosis Mutation Database, http://www.genet.sickkids.on.ca). In addition to the classical CF phenotype, mild mutations in *CFTR* can cause other *CFTR*-related disorders (*CFTR*-RD), such as male infertility due to congenital bilateral absence of the vas deferens (CBAVD; MIM #277180), idiopathic chronic pancreatitis (MIM #167800), and bronchiectasis (MIM #211400) among others.[8] Some of these mild alleles are common polymorphisms, such as polythymidine (poly-T) and poly-thymidine-guanine (poly-TG) tracts, associated with aberrant splicing of exon 10 of *CFTR*, being the most common mutation in CBAVD.[9] Although *CFTR* is one of the most extensively studied human disease genes, its high allelic heterogeneity makes CF and *CFTR*-RD molecular diagnostics challenging.

The precise diagnosis of CF combines clinical evaluation (clinical features of CF phenotype and sweat test measurements) with *CFTR* molecular genetic studies. To date, the molecular characterisation of *CFTR* mutations in a given sample relies on commercial tests that screen for specific common mutations (reverse dot blot INNO-LIPA CFTR [Innogenetics], Cystic Fibrosis Genotyping Assay/OLA [Abbott], Elucigene CF-EU2 [Zeneca], xTAG Cystic Fibrosis 71 kit v2 [Luminex], among others). The detection rate of these panels varies depending on the mutations included (ranging from 4 to 70 *CFTR* mutations) and the molecular heterogeneity of each population. For many patients with common *CFTR* mutations that are present in these commercial panels, there is no need for additional studies, but the high heterogeneity of *CFTR* mutations in some CF populations and in *CFTR*-RD, often makes necessary the complete molecular screening of the 27 exons and the regulatory regions of *CFTR*, which is a costly and labour-intensive task.

As a first step towards the implementation of next-generation sequencing (NGS) approaches to molecular testing that can replace current low-throughput and time-consuming molecular methods, we assessed the efficacy of targeted resequencing for the molecular diagnosis of CF and *CFTR*-RD in a heterogeneous panel of 92 patients with CF and *CFTR*-RD, and CF carriers with known *CFTR* mutations.

MATERIALS AND METHODS

Detailed protocols are available in online supplementary materials.

Subjects

High-quality genomic DNA from 92 unrelated samples, including patients with CF (n=45), CF carriers (n=27) and patients with *CFTR*-RD (n=20), were extracted from peripheral blood lymphocytes using standard protocols. The group of subjects with *CFTR*-RD included 12 patients with CBAVD, 5 patients with idiopathic bronchiectasis, and 3 patients with *CFTR*-related metabolic syndrome. All samples included in this study had previously undergone conventional *CFTR* screening,[10 11] and all *CFTR* mutations were confirmed by Sanger sequencing, multiplex ligation-dependent probe amplification (MLPA) or quantitative PCR (qPCR). The samples selected for this study were recruited for diagnostic purposes between 1998 and 2011. For obvious reasons, it has been impossible to obtain the corresponding informed consents, although all samples were obtained with the purpose of *CFTR* mutation screening. For that, all samples were anonymised in order to ensure the protection of their identity and the list of confirmed mutations was not provided to the investigators performing the bioinformatics mutation analysis until the end of the variant prioritisation process.

Insolution capture and multiplexed resequencing of *CFTR*

Figure 1 summarises the mutation screening workflow that we have implemented in this study. Briefly, DNA from blood was sonicated to obtain fragments of approximately 200 bp. Then, fragments underwent end repair, A-tailing, and ligation to Illumina paired-end indexed adapters, as outlined in the DNA Truseq protocol (Illumina). Once the DNA libraries were indexed, they were PCR amplified and pooled before in-solution hybridisation to a custom NimbleGen SeqCap EZ Choice Library (Roche) of *CFTR* complementary oligonucleotide DNA baits. After stringent washing, the captured libraries were PCR amplified and sent for sequencing (24 libraries per lane) to generate 2×101 bp paired-end reads with a HiSeq 2000 instrument (Illumina). Finally, the resulting DNA sequences were aligned to the human reference genome and sequence variants were detected and annotated as outlined in online supplementary materials.

Identification of CF and *CFTR*-RD mutations

In order to identify *CFTR* pathogenic mutations that could cause CF and *CFTR*-RD, we applied the following filtering steps[12]:

1. We required all candidate variants on both sequenced DNA strands and to account for ≥15% of total reads at that site.
2. Common polymorphisms (≥5% in the general population) were discarded by comparison with National Center for Biotechnology Information (NCBI) single nucleotide polymorphism (SNP) Database (dbSNP) build 132, the March 2010 release of the 1000 Genomes project (http://www.1000genomes.org), the Exome Variant Server (http://evs.gs.washington.edu) and an inhouse exome variant database to filter out common benign variants and recurrent artefact variant calls. However, since these databases contain known disease-associated mutations, all detected variants were compared with gene-specific mutation databases (http://www.hgmd.cf.ac.uk and http://www.genet.sickkids.on.ca).
3. Then, we screened for mutations that could give rise to premature protein truncating mutations, that is, stop mutations, damaging missense variants, splice sites, exonic deletions/insertions and large SVs.
4. Variants were ranked based upon evolutionary conservation and potential deleteriousness of the affected nucleotide using Sift,[13] Polyphen2,[14] PhyloP,[15] and MutationTaster.[16]

RESULTS

CFTR enrichment

We designed oligonucleotides to target the complete genomic sequence (the 27 exons plus all introns), and 10 kb of 5′ and 3′ flanking genomic regions of *CFTR* covering a total of 208 701 bp. After removal of repetitive sequences, 87% of the

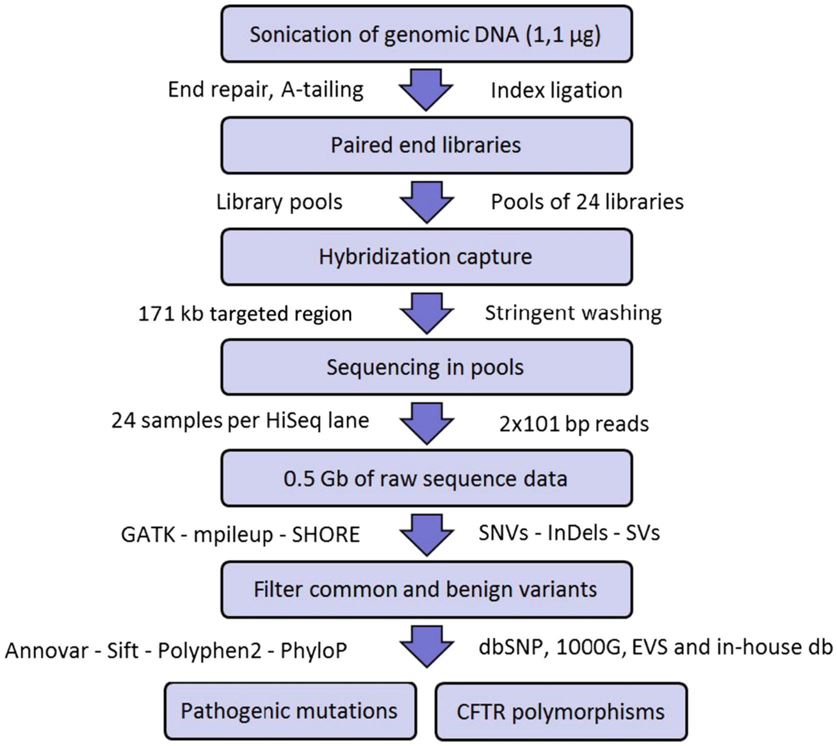

Figure 1 Assay workflow to identify *CFTR* polymorphisms and pathogenic mutations.

targeted bases could be covered with capture baits for a total targeted region of 181 539 bp in 171 individual regions, with lengths ranging from 68 bp to 6689 bp (average 1062 bp). We included the untranslated region of *CFTR* to have a complete definition of the non-coding variability and to favour the detection and sizing of large SVs within the gene.

CFTR sequencing statistics

On average, for each of the four HiSeq2000 (Illumina) lanes, 95.8% of the paired-end 2×101 bp reads could be assigned unambiguously to individual samples, according to their tags, receiving similar proportion of reads for each sample. The remaining 4.2% of unassigned reads were removed because of sequencing errors in their index tags. Therefore, the losses of sequence data associated with high sample multiplexing were minimal. On average, for every sample, 95% of high quality sequencing reads mapped to the reference genome. This resulted in an evenly distributed mean depth of coverage for *CFTR* of 231X (199X if the targeted regions are expanded by 150 bp at each end) with a coefficient of variation of 35%, across samples. In fact, 99.7% of all targeted bases were covered by at least 5 reads (the minimum that we require for variant calling) and 78.53% by at least 100 reads (table 1). For a comprehensive summary of the obtained sequencing results, see also online supplementary table S1.

To determine if coverage was substantially lower for any region, we calculated the proportion of base pairs that were captured by <50 reads. The proportion of these poorly covered regions accounted for 0.069 of *CFTR* targeted bases, and only 0.07% of the targeted bases were not covered by any read (table 1). As expected, these low-covered genomic regions are characterised by low complexity and a high GC content. Sequence targets with these two characteristics are usually refractory to enrichment, resulting in reduced coverage for these sites. However, as shown above, this was the case for only a very small proportion of all bases intended to be captured in this study. From these data we can conclude that all samples, regardless of the pool sizes in the precapture step, were uniformly covered at depths that in all cases exceed by far the minimum coverage required for a reliable variant calling (see online supplementary table S1). The minor differences between samples and pools were neutralised by the excessive overall *CFTR* coverage achieved by our assay. The sequence quality metrics of this data warrant a confident detection of variants in all samples.

Identification of CF and *CFTR*-RD mutations

The selection of the samples for this study was done with the idea to include as many different types of *CFTR* mutations as possible, to simulate a real-world diagnostics scenario, including SNVs, InDels, and large SVs, so that we could test the performance of our approach for all these types of genetic variations. To assess the sensitivity of our assay to detect pathogenic mutations, we blindly inspected all mapped sequence reads from the 92 samples with previously defined mutations in *CFTR*.

By using our multiplexed capture approach and automated variant calling pipeline, we were able to detect, before variant filtering and ranking, 115 SNVs (4 novel) and 28 InDels (19 novel) in *CFTR* per sample on average (table 1). Among these variants we identified several common *CFTR* polymorphisms (see online supplementary table S2). Then, we applied our variant prioritisation strategy to identify *CFTR* pathogenic mutations present in each sample. Using this strategy we detected 122 different pathogenic mutations on *CFTR* in their correct heterozygous/homozygous state across the 92 samples included in the study (some variants were present in more than one sample). We correctly identified 58 missense, 14 nonsense, 23 splice site SNVs, 12 frameshift deletions, 2 frameshift insertions, and 3 inframe deletions (one of 84 nucleotides long), known to cause CF and *CFTR*-RD (see online supplementary table S3). In addition, we were also able to detect three different 5T pathogenic haplotypes, five large deletions, one duplication and one large genomic rearrangement (that includes one inversion and two deletions) involving various *CFTR* exons.

Intron 9 poly-TG and poly-T haplotypes and alternative splicing of CFTR

The 5T variant in intron 9 (c.1210–12T[5] is the most common mutation associated with CBAVD.[9] The penetrance of the 5T variant depends on the neighbouring TG sequence repeat.[17] Thus, the definition of the TG-T (c.1210–34TG[11–13]T[5–9]) haplotype contributes to predict the most likely *CFTR*-RD phenotype of the carrier subject. However, the repetitiveness of its sequence at the nucleotide level makes difficult to determine the TG-T haplotype using standard variant calling algorithms (figure 2). In order to address this issue, we developed an in-house script that scans the very raw sequencing data of each sample for all possible combinations of c.1210-34TG[11–13]T[5–9]. By doing this, we were able to determine the exact TG-T haplotype of each sample, including three T5-TG11, eight T5-TG12 and two T5-TG13 haplotypes (see online supplementary table S4).

Characterisation of large structural changes in *CFTR*

Several of the unknown CF and *CFTR*-RD mutations in affected individuals may not have been identified yet because of the intrinsic low sensitivity of traditional PCR-based *CFTR* screening approaches for large SVs. It has been estimated that large genomic rearrangements of *CFTR*, which exhibit extensive allelic heterogeneity and are mainly caused by non-homologous recombination events, may account for up to 20% of the unidentified *CFTR* alleles in patients with CF and *CFTR*-RD.[18] A major step-forward of NGS technologies with respect to classical molecular approaches is the possibility to detect large genomic rearrangements at the same time than SNVs and InDels, without the need for additional assays specific for large SVs, such as array-comparative genomic hybridisation, semi qPCR based methods, MLPA or quantitative multiplex PCR of short fluorescent fragments. In our study, the combination of paired-end mapping, split-read analysis, and normalised depth of coverage strategies allowed the blind identification of 7/7 (100% sensitivity) large SVs (5 deletions, one duplication and one complex rearrangement) in *CFTR* (figure 3). We were able to accurately identify the breakpoints of all of them, with a perfect concordance between the prediction of the algorithms and the validations for each of them (table 2).

Among the seven SVs analysed in this study we have also characterised in silico and validated by Sanger sequencing the breakpoints of a novel (ie, not previously reported to the public databases) *CFTR* 1899 bp deletion (chr7:117267155-117269054, hg19) that includes the loss of exon 22 (c.3469-420_3717+1230del1899), and all the breakpoints of a large genomic rearrangement previously reported as CFTR50kbdel (legacy name).[21] These two SVs were previously identified in their respective samples by means of MLPA and qPCR, but their breakpoints were not known. Thanks to the results of this study now we know that CFTR50kbdel consists of a 85 kb inversion, with breakpoints chr7:117169862/117169876-117255003; containing two large deletions:

Table 1 Sequencing quality control parameters, coverage and detected variants by targeted resequencing of the *CFTR* gene using pools of 8, 12, 16 and 24 samples

Samples	Pool name Precapture pooling (number of samples)	All samples All	8A 8	8B 8	12A 12	12B 12	16A 16	16B 16	24A 24
Sequencing	QC-passed reads±%CV	11701689±35	19187383±14	17639.073±20	8430967±15	11654345±15	8614519±10	10449749±30	11779102±26
	% Mapped	94.9	97.13	96.36	94.53	96.36	96.4	94.23	92.58
	% Properly paired	92.98	96.03	95.05	92.49	95.06	95.04	92.07	89.73
Coverage	Mean coverage (X)±%CV	231±43	425±16	358±22	101±12	223±16	173±12	212±29	244±22
	Mean coverage extended 150 bp (X)±%CV	199±43	367±17	313±22	88±11	192±16	150±13	181±29	209±22
	% Enrichment	55.31	58.54	57.12	43.29	55.8	56.42	56.67	57.75
	% target bases covered=0×	0.07	0.03	0.01	0.12	0.1	0.09	0.07	0.06
	% target bases covered≥1×	99.93	99.97	99.99	99.88	99.9	99.91	99.93	99.94
	% target bases covered≥5×	99.7	99.84	99.86	99.41	99.72	99.76	99.67	99.72
	% target bases covered≥10×	99.39	99.76	99.78	98.55	99.49	99.56	99.29	99.45
	% target bases covered≥20×	98.48	99.63	99.58	95.92	98.82	98.92	98.23	98.69
	% target bases covered≥50×	93.13	98.86	98.38	79.25	95.1	94.74	92.49	94.78
	% target bases covered≥100×	78.53	96.34	93.79	43.14	83.23	77.56	78.29	83.64
CFTR variants	SNVs	115	124	89	121	119	124	121	104
	Novel SNVs	4	5	5	3	3	4	3	4
	Exonic SNVs	3	2	2	3	3	3	3	2
	Missense, nonsense and splice site SNPs	2	2	1	2	2	2	2	2
	InDels	28	29	29	28	30	31	28	25
	Novel InDels	19	19	19	19	20	21	17	16
	Frameshift and non-Frameshift InDels	1	1	1	1	1	1	1	1

CFTR, cystic fibrosis transmembrane regulator; CV, Coefficient of variation; InDels, insertions and deletions; SNV, single nucleotide variants; QC, Coefficient of variation.

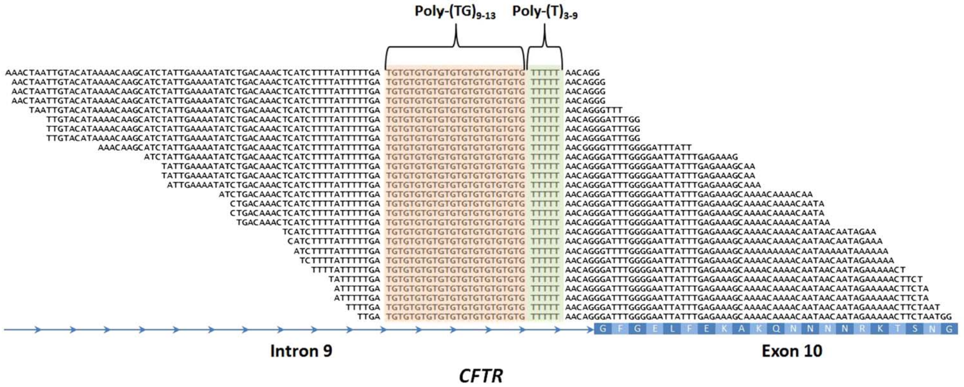

Figure 2 Detection of the intron 9 poly-TG-T haplotype involved in male infertility and other *CFTR*-RD. Example of a patient with *CFTR*-RD with the c.1210-34TG[12]T[5] haplotype. The centre of the alignment of the 100 nt NGS reads shows the poly-TG (in orange) and poly-T (in green) tracts. The *CFTR* intron 9 and exon 10 (with the amino acid sequence in white) are represented in the bottom in blue. poly-TG, poly-thymidine-guanine.

chr7:117169908-117180511(10.6 kb deletion of exons 4–8), and chr7:117216401-117254987(38.6 kb deletion of exons 12–21), 49.2 kb in total, which is remarkably close to the 50 kb deletion originally estimated by classical molecular methods[21] (figure 4A,B). In addition, cDNA analysis evidenced an aberrant transcript showing a unique junction exon 3/22 indicating the loss of the entire inverted region (figure 4C). This is the first time that a *CFTR* large inversion is reported, and, to our knowledge, it is the most complex rearrangement ever characterised in *CFTR* (c.[274-1091_3468+236inv85141ins38; 274-1044_1116+111del10602insTATAT; 1585-11 392_3468+219del38585]).

Sensitivity and specificity of targeted resequencing of *CFTR*

The molecular diagnostic strategy for CF and *CFTR*-RD that we present here has blindly identified all previously known pathogenic *CFTR* variants in the 92 CF samples studied. This represents a mutation detection rate of 100% (122/122), with zero false-positive calls, and would have resulted in a positive molecular diagnosis in 91 of the 92 patients with CF and *CFTR*-RD and CF carriers (diagnostic rate of 98.9%), since for one of the patients with CF (sample 80) we were unable to identify his previously unknown second CF allele. As expected, for patients with *CFTR*-RD with only one previously known *CFTR* mutation our NGS strategy hasn't identified a second *CFTR* allele, as it is the case of three idiopathic bronchiectasis and four patients with CBAVD. It is known that other genetic and environmental factors may contribute to these phenotypes,[23–25] so the apparently missing *CFTR* alleles in these samples cannot be solely attributed to issues with the specificity or sensibility of our approach. Overall, the high success rate achieved in this study highlights the accuracy of this strategy as a molecular diagnostics tool for CF and *CFTR*-RD.

Precapture pooling and multiplexed sequencing reproducibility

Precapture pooling reduces substantially the library preparation time and, in combination with multiplexed sequencing, allows to exploit the full potential of NGS for clinical diagnostics. In order to assess how precapture multiplexing affects coverage and accuracy, we tested different pool sizes: two captures of 8, 12 and 16 samples each and one capture of 24 samples. All samples were marked with a specific index/tag, so that their individual identification was warranted at the end of the sequencing run. The sequence quality data and the variant calling results indicate that there were no sensitivity or specificity problems associated with the use of precapture pools of high number of samples (table 1). Thus, the major technical consequences of precapture pooling, which are the reduction in the input amount of the individual libraries and the addition of multiple barcodes, which may lead to less efficient blocking and

Figure 3 Detection of large structural variants in the *CFTR* gene by normalised depth of coverage analysis. Representation of the SVD-ZRPKM Values calculated by Conifer[29] for the 92 samples. Coloured peaks indicate the five largest structural variants identified in this study.

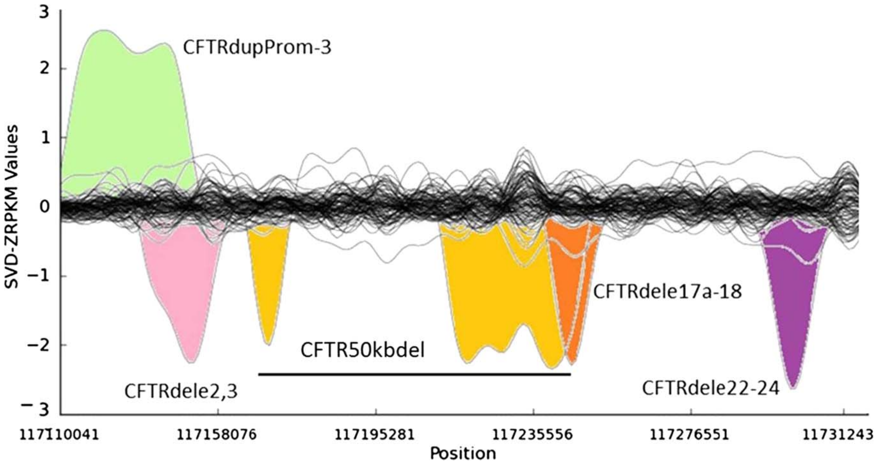

Table 2 Large structural variants identified in the *CFTR* by targeted resequencing

Sample	SV	Predicted breakpoints	Validated breakpoints	Reference
47FQ	CFTRdele20	chr7:117282464-117283245	chr7:117282468-117283248	[18]
69FQ	CFTRdele2,3	chr7:117138362-117159442	chr7:117138367-117159446	[19]
70FQ	CFTRdupProm-3	chr7:117113959-117149700	chr7:117113985-117149644	[10]
78FQ	CFTRdele22-24	chr7:117300851-117310305	chr7:117300852-117310305	[20]
83FQ	CFTR50kbdel	INV chr7:117169861-117254986+DELs chr7:117170000-117182000+chr7:117217000-117255000	INV chr7:117169862/117169876-117255003+DELs chr7:117169908-117180511+chr7:117216401-117254987	[21] and this study
88FQ	CFTRdele17a-18	chr7:117247975-117256874	chr7:117247980-117256878	[22]
93FQ	CFTRdele22	chr7:117267155-117269054	chr7:117267155-117269054	This study

CFTR, cystic fibrosis transmembrane regulator; SV, structural variant.

favour unspecific hybridisation,[26] have minor effects in the final variant calling process.

Reproducibility was determined by running four samples (two patients with CF and two patients with *CFTR*-RD) in duplicate on the same run, but captured in pools of different sample sizes (in two different precapture pools of 8 and 24 samples, respectively), and sequenced in independent HiSeq2000 lanes. We detected eight of eight pathogenic mutations in the replicated samples, yielding 100% reproducibility of mutation detection. We next assessed the reproducibility for all variant calls in the entire *CFTR* captured region (mean=138±47 SNVs and 32±11 InDels per sample). Across the four samples, reproducibility was 96.09% for SNVs and 71.62% for InDels, with an overall reproducibility of 91% for all variants in all four samples (table 3). Variant calls that did not replicate were all intronic or in intergenic regions, and almost all of them were located very close to the ends of the targeted regions of *CFTR* or in regions not covered by capture baits. This explains the observed low coverage of these unreplicated variants (mean=33.47X vs 144.58X of the replicated variants), and highlights the impact of the depth of coverage on the assay reproducibility.

Twenty-one out of the 122 pathogenic variants detected by our analysis were present in two or more individuals (see online supplementary table S3). This means a reproducibility of 100% for pathogenic variant calls between two or more samples (based on the results of 17.21% of the mutations included in this study). Since most of the samples bearing these mutations were multiplexed in independent precapture pools of different sample sizes, and also were run in different sequencer lanes, we can conclude that our approach offers great robustness and reproducibility in the detection of *CFTR* pathogenic variants. Although the coverage for a given mutation can vary significantly between samples, the proportion of reads supporting the non-reference allele was always maintained (see online supplementary table S3). Altogether, these results highlight the sensitivity and reproducibility of our assay, and support the use of a larger number of samples in precapture pools in future studies, when more index tags are available (24 when we planned this study).

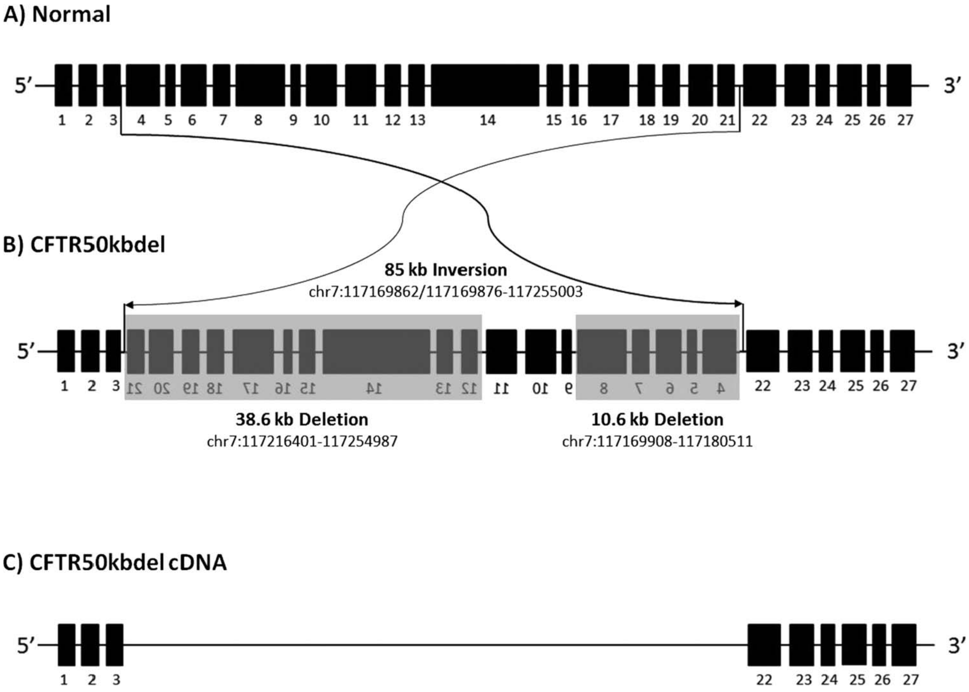

Figure 4 Schematic representation of the complex CFTR50kbdel. (A) Normal structure of *CFTR*. Black boxes represent each of the 27 *CFTR* exons. (B) Diagram shows the complex architecture of the CFTR50kbdel mutation. Arrows indicate the breakpoints of the 85 kb inversion. Grey areas indicate the two deleted regions. (C) cDNA sequence of CFTR50kbdel, showing the loss of exons 4–21.

Table 3 Assay reproducibility of the identification of *CFTR* mutations by targeted resequencing

	1		2		3		4		4 Samples	
Sample	Reproducibility	Per cent	Reproducibility	Per cent	Reproducibility	Per cent	Reproducibility	Per cent	Reproducibility	Per cent
SNPs	200/208	96.15	129/131	98.47	134/142	94.37	78/82	95.12	541/563	96.09
InDels	38/55	69.09	21/33	63.64	31/37	83.78	16/23	69.57	106/148	71.62
All Variants	238/263	90.49	150/164	91.46	165/179	92.18	94/105	89.52	647/711	91.00
Coverage	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Matched	128.14	107.16	242.77	174.01	106.61	89.59	99,19	76.98	144.58	130.48
Unmatched	41.94	74.56	80.86	132.04	10.15	7.58	15,22	13.65	33.47	70.24

CFTR, cystic fibrosis transmembrane regulator; InDels, insertions and deletions.

DISCUSSION

Here we have implemented and tested a novel strategy for the molecular analysis of CF and *CFTR*-RD, based on pooled target enrichment and multiplexed NGS of *CFTR*. We have validated this new approach in a cohort of 92 samples with previously known pathogenic *CFTR* mutations. The different pools of simultaneously enriched *CFTR* samples were multiplexed in groups of 24 samples in four sequencer lanes. After mapping the sequencing reads to the reference genome and performing blind variant calling and filtering, our bioinformatics pipeline successfully retrieved all known pathogenic mutations in their correct heterozygous/homozygous state. With this approach we were able to identify a heterogeneous panel of *CFTR* mutations, including SNVs, InDels and large SVs. Our results (mutation detection rate of 100% and diagnostic rate of 98.91%) demonstrate the suitability of targeted resequencing for the routine clinical diagnosis of CF and *CFTR*-RD.

Clinical diagnostic tools must meet very stringent sensitivity and specificity parameters, while keeping their cost-effectiveness and time-effectiveness. The approach that we describe here represents a change in the paradigm for the molecular diagnostics of CF and *CFTR*-RD. Until now, the ideal strategy for *CFTR* screening consisted of three sequential steps:[10] (1) genotyping by commercially available kits a small subset (30–50) of common *CFTR* mutations; (2) in case of not having identified the two *CFTR* alleles, complete screening of the coding portion and flanking regions of *CFTR* by scanning techniques, like denaturing gradient gel electrophoresis or single strand conformation polymorphism/heteroduplex among others, and subsequent Sanger sequencing; and if still insufficient, (3) screening by MLPA and/or array-comparative genomic hybridisation for large genomic rearrangements. The average cost per sample of this strategy is around €400 with an estimated turnaround time of 2–3 months for samples that have to undergo all three steps described above. We estimate that the approach that we present here has an overall cost of less than €200 per sample, which represents a 50% of cost savings per sample and makes the whole process eight times faster when compared with the techniques currently used for the molecular diagnosis of CF and *CFTR*-RD. In addition, our strategy offers a complete definition of the captured *CFTR*, without the need for stepwise testing anymore. We foresee that these differences will become even more significant because of the constantly dropping sequencing costs[27] and optimised library preparation and sequencing protocols. The complete process of library preparation, sequence enrichment, NGS and bioinformatics analysis could be completed within 14 days after reception of the DNA sample. The most time-consuming step was sequencing the *CFTR*-enriched DNA libraries on the HiSeq2000 (Illumina), which took approximately 10 days. In addition to saving time in the process of library preparation with new capture strategies, using the most recent enrichment technologies such as Haloplex (Agilent), and optimising the bioinformatics, major time savings could be made by using the new generation of HiSeqs (Illumina) sequencers (series 2500), which have been recently reported to be able to generate up to 140 GB of sequence (2×100 bp) in approximately 24 h.[28] As an alternative that would reduce NGS costs, we propose the use of smaller, benchtop, personal sequencers such as the MiSeq (Illumina) or Ion Torrent (Ion Torrent Systems). The amount of sequence output of these instruments is approximately 10 times smaller than its bigger siblings, so they would be ideal for the analysis of batches of reduced numbers of samples (up to 10 samples per run).

The major drawback of capturing the complete genomic sequence of *CFTR* instead of focusing only on the coding regions is that more sequencing is needed to achieve similar coverage. However, the benefits of this approach are that no deep intronic mutations are missed, nor variants in the promoters or in the Untranslated regions (UTRs). In addition, this strategy has also proven its utility to detect large deletions, duplications and inversions, involving various *CFTR* exons, as well as to detect their breakpoints. The detection of variation in the untranslated regions of *CFTR* can also be used for the identification of alleles of clinical relevance, such as the 5T variant, which has variable penetrance and accounts for part of the phenotypic variability of *CFTR*-RD.[17]

In addition to the technical limitations inherent to hybrid capture, such as selection bias and uneven capture efficiency, the main limitation of the targeted resequencing approach is the impossibility to efficiently capture and sequence the repetitive and low-complexity, and GC-rich genomic segments of *CFTR* that are refractory to enrichment. However, the constant optimisation of the capture probes and NGS chemistries will gradually close the capture gaps (mainly due to uniqueness constraints, homopolymer runs, ambiguous bases or other factors that are known to cause issues in either oligonucleotide synthesis or hybridisation), and reduce enrichment variability between samples. But until then, this will require backup methods to assess the variability in these 'dark' regions, in the case of samples with clear CF or *CFTR*-RD phenotypes, but with no identified mutations in the captured fraction of *CFTR*, as in the case of sample 80 for which we were not able to find its previously unknown second CF allele. However, the major sources variability that could potentially affect the sensitivity and specificity in our study (such as variations in Guanine-cytosine (GC) content or differential hybridisation efficiency of the two alleles in a diploid genome) are neutralised by the high level of sequencing depth achieved.

The transition of NGS technologies from basic research to routine molecular diagnostics over the next years, will take advantage of the constant improvements in the reliability and robustness of these technologies, and of simplified bioinformatics analyses able to generate medical report-like outputs adapted to clinical laboratories. We are still in the process of defining the methods and guidelines for the application of NGS to clinical genetic diagnostics. In this initial phase, we still recommend that novel mutations are validated by Sanger sequencing before informing the patient.

In conclusion, this represents, to the best of our knowledge, the first study successfully using targeted NGS to detect pathogenic lesions in the CFTR gene. With the approach reported here we have been able to describe for the first time the breakpoints of a novel deletion and the most complex genomic rearrangement in CFTR. We have only had one false positive and zero spurious calls. Altogether, our assay shows a clear superiority with respect to traditional methods for *CFTR* screening and overcomes their technical limitations, making it their natural replacement in the diagnostic laboratories.

Acknowledgements We thank the subjects and referring physicians who participated in this study.

Contributors This study was conceived and designed by DT, TC, LA, and XE. Selection of samples was performed by TC. NGS libraries were prepared by JG. The bioinformatics pipeline and the NGS analysis was performed by DT, CT, GE, and SO. Validation of SVs was performed by FS, MDR, and TC. The manuscript was written by DT, TC, and XE. All aspects of the study were supervised by TC and XE.

Funding This project was funded by the Spanish Plan Nacional SAF2008–00357 (NOVADIS); the Generalitat de Catalunya AGAUR 2009 SGR-1502; the Instituto de Salud Carlos III (FIS/FEDER PI11/00733); and the European Commission 7th Framework Program, Project N. 261123 (GEUVADIS), and Project N. 262055 (ESGI). DT is a PhD student supported by the Spanish Ministry of Economy and Competiveness; FS is a PhD student from Associació Catalana de Fibrosi Quística.

Competing interests None.

Ethics approval PRBB Ethics Committee.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Farrell PM. The prevalence of cystic fibrosis in the European Union. *J Cyst Fibros* 2008;7:450–3.

2 O'Sullivan BP, Freedman SD. Cystic fibrosis. *Lancet* 2009;373:1891–904.

3 Bobadilla JL, Macek M Jr, Fine JP, Farrell PM. Cystic fibrosis: a worldwide analysis of CFTR mutations–correlation with incidence data and application to screening. *Hum Mutat* 2002;19:575–606.

4 Estivill X, Bancells C, Ramos C. Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. The Biomed CF Mutation Analysis Consortium. *Hum Mutat* 1997;10:135–54.

5 Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. Identification of the cystic fibrosis gene: genetic analysis. *Science* 1989;245:1073–80.

6 Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, Drumm ML, Iannuzzi MC, Collins FS, Tsui LC. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 1989;245:1066–73.

7 Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N, Zsiga M, Buchwald M, Riordan JR, Tsui LC, Collin FS. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* 1989;245:1059–65.

8 Dequeker E, Stuhrmann M, Morris MA, Casals T, Castellani C, Claustres M, Cuppens H, des Georges M, Ferec C, Macek M, Pignatti PF, Scheffer H, Schwartz M, Witt M, Schwarz M, Girodon E. Best practice guidelines for molecular genetic diagnosis of cystic fibrosis and CFTR-related disorders–updated European recommendations. *Eur J Hum Genet* 2009;17:51–65.

9 Chillon M, Casals T, Mercier B, Bassas L, Lissens W, Silber S, Romey MC, Ruiz-Romero J, Verlingue C, Claustres M, Nunes V, Férec C, Estivill X. Mutations in the cystic fibrosis gene in patients with congenital absence of the vas deferens. *N Engl J Med* 1995;332:1475–80.

10 Ramos MD, Masvidal L, Gimenez J, Bieth E, Seia M, des Georges M, Armengol L, Casals T. CFTR rearrangements in Spanish cystic fibrosis patients: first new duplication (35kb) characterised in the Mediterranean countries. *Ann Hum Genet* 2010;74:463–9.

11 Alonso MJ, Heine-Suner D, Calvo M, Rosell J, Gimenez J, Ramos MD, Telleria JJ, Palacio A, Estivill X, Casals T. Spectrum of mutations in the CFTR gene in cystic fibrosis patients of Spanish ancestry. *Ann Hum Genet* 2007;71(Pt 2):194–201.

12 Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, Nord AS, Mandell JB, Swisher EM, King MC. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc Natl Acad Sci USA* 2010;107:12629–33.

13 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073–81.

14 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–9.

15 Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005;15:901–13.

16 Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;7:575–6.

17 Groman JD, Hefferon TW, Casals T, Bassas L, Estivill X, Des Georges M, Guittard C, Koudova M, Fallin MD, Nemeth K, Fekete G, Kadasi L, Friedman K, Schwarz M, Bombieri C, Pignatti PF, Kanavakis E, Tzetis M, Schwartz M, Novelli G, D'Apice MR, Sobczynska-Tomaszewska A, Bal J, Stuhrmann M, Macek M Jr., Claustres M, Cutting GR. Variation in a repeat sequence determines whether a common variant of the cystic fibrosis transmembrane conductance regulator gene is pathogenic or benign. *Am J Hum Genet* 2004;74:176–9.

18 Ferec C, Casals T, Chuzhanova N, Macek M Jr., Bienvenu T, Holubova A, King C, McDevitt T, Castellani C, Farrell PM, Sheridan M, Pantaleo SJ, Loumi O, Messaoud T, Cuppens H, Torricelli F, Cutting GR, Williamson R, Ramos MJ, Pignatti PF, Raguenes O, Cooper DN, Audrezet MP, Chen JM. Gross genomic rearrangements involving deletions in the CFTR gene: characterization of six new events from a large cohort of hitherto unidentified cystic fibrosis chromosomes and meta-analysis of the underlying mechanisms. *Eur J Hum Genet* 2006;14:567–76.

19 Dork T, Macek M Jr., Mekus F, Tummler B, Tzountzouris J, Casals T, Krebsova A, Koudova M, Sakmaryova I, Macek M Sr, Vavrova V, Zemkova D, Ginter E, Petrova NV, Ivaschenko T, Baranov V, Witt M, Pogorzelski A, Bal J, Zekanowsky C, Wagner K, Stuhrmann M, Bauer I, Seydewitz HH, Neumann T, Jakubiczka S. Characterization of a novel 21-kb deletion, CFTRdele2,3(21 kb), in the CFTR gene: a cystic fibrosis mutation of Slavic origin common in Central and East Europe. *Hum Genet* 2000;106:259–68.

20 Chevalier-Porst F, Souche G, Bozon D. Identification and characterization of three large deletions and a deletion/polymorphism in the CFTR gene. *Hum Mutat* 2005;25:504.

21 Morral N, Nunes V, Casals T, Cobos N, Asensio O, Dapena J, Estivill X. Uniparental inheritance of microsatellite alleles of the cystic fibrosis gene (CFTR): identification of a 50 kilobase deletion. *Hum Mol Genet* 1993;2:677–81.

22 Lerer I, Laufer-Cahana A, Rivlin JR, Augarten A, Abeliovich D. A large deletion mutation in the CFTR gene (3120+1Kbdel8.6Kb): a founder mutation in the Palestinian Arabs. Mutation in brief no. 231. Online. *Hum Mutat* 1999;13:337.

23 Casals T, Bassas L, Egozcue S, Ramos MD, Gimenez J, Segura A, Garcia F, Carrera M, Larriba S, Sarquella J, Estivill X. Heterogeneity for mutations in the CFTR gene and clinical correlations in patients with congenital absence of the vas deferens. *Hum Reprod* 2000;15:1476–83.

24 Casals T, De-Gracia J, Gallego M, Dorca J, Rodriguez-Sanchon B, Ramos MD, Gimenez J, Cistero-Bahima A, Olveira C, Estivill X. Bronchiectasis in adult patients: an expression of heterozygosity for CFTR gene mutations? *Clin Genet* 2004;65:490–5.

25 Casals T, Aparisi L, Martinez-Costa C, Gimenez J, Ramos MD, Mora J, Diaz J, Boadas J, Estivill X, Farre A. Different CFTR mutational spectrum in alcoholic and idiopathic chronic pancreatitis? *Pancreas* 2004;28:374–9.

26 Redin C, Le Gras S, Mhamdi O, Geoffroy V, Stoetzel C, Vincent MC, Chiurazzi P, Lacombe D, Ouertani I, Petit F, Till M, Verloes A, Jost B, Chaabouni HB, Dollfus H, Mandel JL, Muller J. Targeted high-throughput sequencing for diagnosis of genetically heterogeneous diseases: efficient mutation detection in Bardet-Biedl and Alstrom Syndromes. *J Med Genet* 2012;49:502–12.

27 Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. http://www.genome.gov/sequencingcosts (accessed Nov 2012).

28 Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Alnadi NA, Andraws N, Patterson ML, Krivohlavek LA, Fellis J, Humphray S, Saffrey P, Kingsbury Z, Weir JC, Betley J, Grocock RJ, Margulies EH, Farrow EG, Artman M, Safina NP, Petrikin JE, Hall KP, Kingsmore SF. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med* 2012;4:154ra35.

29 Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Quinlan AR, Nickerson DA, Eichler EE. Copy number variation detection and genotyping from exome sequence data. *Genome Res* 2012;22:1525–32.

SUPPLEMETARY METHODS

In-solution Capture and Multiplexed Resequencing of CFTR

To carry out the DNA capture, we designed through Roche's Customer Services a custom NimbleGen SeqCap EZ Choice Library to target the complete genomic sequence of *CFTR* (chr7:117120017-117308718, hg19). Baits were designed to cover the coding regions, the noncoding intronic sequences, and 10 kb of genomic sequence flanking each end of *CFTR*. Thus, the coordinates of the final target region were chr7:117110017-117318718 (hg19), which account for 208,701 bp. After masking repetitive DNA elements, the total targeted DNA represented 181,539 bp (87% of target bases covered) distributed in 171 individual regions, which were selected using the most stringent settings for probe design (uniqueness tested by Sequence Search and Alignment by Hashing Algorithm [SSAHA]) [1]. No probe redundancy was allowed in the final capture design. The BED file of probe sequences is available on request to the authors.

In this study we took advantage of the capabilities of the NimbleGen SeqCap EZ Choice Library (Roche) to enrich multiple DNAs (pools of 8 to 24 samples) in one single capture reaction. Thus, capture of *CFTR* was carried out following the instructions of the NimbleGen SeqCap EZ Library SR User's Guide v3.0, that is available for free download, to perform sequence capture from pooled libraries prepared with the TruSeq DNA Sample Preparation Kits (Illumina). Briefly, for each sample, one microgram of genomic DNA was sheared by sonication to generate double stranded DNA fragments of 200-300 bp with 3′ and 5′ overhangs on a Covaris S2 instrument (Covaris) in 52.5 µl of 1X low TE (10 mmol/L Tris/0.1 mmol/L EDTA) for 2 min using frequency sweeping mode with duty cycle, 10%; intensity, 5.0; bursts per second, 200 at a temperature of 5.5°C to 6°C; and power, 23 W. After sonication, DNAs were subjected to three enzymatic steps: end repair, A-tailing, and ligation to Illumina paired-end indexed adapters. All purification steps between the enzymatic reactions were carried out with AMPure XP beads (Beckman Coulter). The adapter-ligated library was amplified by polymerase chain reaction (PCR) for seven cycles with the TS-PCR oligos 1 and 2. PCR products were cleaned using the QIAquick PCR Purification Kit (QIAGEN) and quantified by a DNA1000 chip on a Bioanalyzer 2100 instrument (Agilent Technologies). Then, the DNA libraries were multiplexed in pools of 8, 12, 16 and 24 samples, for a final combined mass of 1.1 µg, and the resulting library pools were hybridized to the custom design of complementary DNA biotinylated oligonucleotides described above. After 72 hours of hybridization at 47°C, the library-bait hybrids were purified by incubation with streptavidin-bound T1 Dynabeads (LifeTechnologies) and washed with increasing stringency to remove nonspecific binding. After capture, each library pool was amplified by PCR for 17 cycles with the TS-PCR oligos 1 and 2. After PCR amplification, the library pools were cleaned using QIAquick PCR Purification Kit (QIAGEN), and quantified by a high-sensitivity chip on a Bioanalyzer 2100 instrument (Agilent Technologies). Then, different combinations of pools of different sample sizes were arranged to multiplex a total of 24 enriched samples per sequencer lane (four lanes in total). Sequencing was performed with 2×100 bp paired-end reads and a 6-bp index read using SBS v3 chemistry on a HiSeq2000 (Illumina). The resulting fastq files were analyzed with an in-house developed pipeline described below.

Bioinformatics Analysis of DNA Variants

The authors involved in the analysis of the mutations performed the study blindly, i.e. they had no information about the *CFTR* pathogenic variants known to be present in each sample. Image analyses, base calling and de-multiplexing on each lane of data were performed using Illumina's Sequencing Analysis Pipeline version 1.7.0 (Illumina). Reads were aligned to the human reference genome hg19 using the Burrows-Wheeler Aligner (bwa aln) version 0.5.9 [2] allowing for maximally six mismatches and one gap of up to 20bp. Alignments in SAM format were generated using bwa sampe, sorted by genome position and converted to BAM format using samtools version 0.1.16 [3]. We also performed local re-alignment around potential insertions/deletions and SNP clusters, base-quality recalibration and duplication marking using the GATK pipeline [4] and picard-tools (http://picard.sourceforge.net). The resulting alignments were used as input for three different variant prediction tools, namely GATK Unified Genotyper [5], samtools mpileup [3] and SHORE (http://1001genomes.org). The three independent SNV and InDel predictions were subsequently quality filtered using GATK VariantFiltration (parameters MQ < 30.0 || QUAL <25.0 || QD <4.0 || DP <5 || DP >2000|| GQ <15) and intersected using GATK CombineVariants. Only SNVs and InDels of up to 30 bp, called by at least two approaches, and found within 150 bp of the ends of the enriched targets, were considered for subsequent analysis. Functional annotation of high quality variants was performed using Annovar [6], providing a comparison of predicted variants to the National Center for Biotechnology Information (NCBI) SNP Database (dbSNP) build 132, the March 2010 pilot release of the 1000 Genomes project (www.1000genomes.org), conservation around variants based on phastCons [7], segmental duplication filter, gene annotation (exon/intron/UTR), amino-acid substitutions and splice variants based on UCSC Genome Browser [8] tracks as well as multiple estimates of the impact of amino acid substitution on the structure and function of proteins (tools: Sift [9], Polyphen2 [10], PhyloP [11] and MutationTaster [12]).

To identify large InDels and SVs we used Pindel [13], Conifer [14], and PeSV-Fisher (http://gd.crg.eu/tools), which is a module-based tool for the detection of deletions, gains, intra- and inter-chromosomal translocations, and inversions. This algorithm provides comprehensive information on co-localization of SVs in the genome. The tool uses methods based on paired-end mapping and read-depth analysis corrected for oligonucleotide probe coverage and local guanine and cytosine (GC) content for 100-bp windows. The compendium of modules of PeSV-Fisher toolkit include: definition of anomalous read-pairs, clustering procedure and breakpoint prediction, read depth analysis, definition and interpretation of SVs, and filtering SV calls (FinalCountDown).

To characterize the TG-T haplotypes of the samples we developed an in-house script that scans the raw sequences in the fastq files and counts the number of reads that contain each possible c.1210-34TG(11-13)T(5-9) haplotype.

Characterization of Newly Identified Structural Variants

Different set of primers were designed using Primer3 v0.4.0 (http://frodo.wi.mit.edu) to corroborate the breakpoints of two large structural variants characterized for the first time in this study; a novel deletion involving exon 22, and CFTR50kbdel, which is a previously known complex genomic rearrangement [15]. PCR amplifications were performed using 1X Takara *Taq* polymerase (Clontech Laboratories), and PCR products were separated in a 1% low melting agarose gel and were subjected to Sanger sequencing.

Deletion of exon 22. Forward 5’-tgcagatgtatgccaatgact-3’ (chr7:117267046-117267066) and Reverse 3’-gcatattacttctttctcattttgc-5’ (chr7:117269178-117269202) primers were used to amplify and sequence the 1899 bp deleted region in the affected CF sample.

*CFTR50kbdel (legacy name)***.** First, primers Fw-A1 5'-cttggcaataccaggggtag-3' (chr7:117169863-117169882) and Rv-A1 3'-ctcctcccagaaggctgtta-5' (chr7:117182143-117182162) were used to amplify and sequence the 1600 bp PCR product, which led to the identification of the first deleted region (exons 4-8) comprising 10.601 kb between chr7:117169909-117180509 with a TATAT inserted sequence in the junction. Then, primers Rv-A1 3'-ctcctcccagaaggctgtta-5' (chr7:117182143-117182162) and Rv-B1 3'-aggacaatttggcaccactc-5' (chr7:117255073-117255092) were used to amplify a 1700 bp fragment. Again, by Sanger sequencing we were able to determine the 3' breakpoint of the deletion. In order to also corroborate the *CFTR* 5' region of this complex rearrangement, primers Fw-7.11 5'-cttgcattcagagccttggt-3' (chr7:117169599-117169620) and Fw-B4 5'-tcttgcacgtgaatgaatgag-3' (chr7:117215595-117215615) were designed. Sanger sequencing of the resulting 1400 bp PCR product evidenced the complexity of this mutation (Fig. S1). To evaluate the effect of this complex rearrangement at the transcript level, we analyzed the cDNA sample from a CF carrier of this complex rearrangement by PCR using primers Fw-e3: 5'-gctggcttcaaagaaaaatcc-3' (chr7:117149103-117149123) and Rv-e22: 5'-tctgttggcatgtcaatgaac-3' (chr7:117267602-117267622).

Shell script

```
my @qsub = ("#!/bin/bash

#\$ -e $workdir/tmp/$sample
#\$ -o $workdir/tmp/$sample

#\$ -pe smp 4
#\$ -hard
#\$ -l h_vmem=5G

export TMPDIR=$workdir/tmp/$sample
export PATH=/users/GD/tools/annovar/annovar_2011Nov20/:\$PATH


NAME=$name
READ1=$read1
READ2=$read2
OUTF=$outfolder/$name
EXOME=/users/xe/dtrujillano/Exomes/custom_capture/qgenomics
KNOWNINDEL=/users/GD/resource/human/hg19/databases/dbSNP/dbIndel132_20101103.vcf
BWA=/users/GD/tools/bwa/bwa-0.5.10/bwa
GATK=/users/GD/tools/GATK/GATK_src_1.4-15-gcd43f01/dist/GenomeAnalysisTK.jar
SAMTOOLS=/soft/bin
ANNOVAR=/users/GD/tools/annovar/annovar_2011Nov20
SHORE=/users/GD/tools/shore/shore
NGSBOX=/users/GD/tools/ngsbox
RSCRIPT=/soft/bin/Rscript


### Align reads with bwa
\$BWA aln -k 2 -i 5 -q -1 -t 2 -R 0 -n 6 -o 1 -e 20 -l 28 -f \$OUTF/\$NAME.r1.sai
\$REF \$READ1
\$BWA aln -k 2 -i 5 -q -1 -t 2 -R 0 -n 6 -o 1 -e 20 -l 28 -f \$OUTF/\$NAME.r2.sai
\$REF \$READ2


### Correct paired end files
\$BWA sampe -r \"\@RG\\tID:\$NAME\\tSM:\$NAME\" -a 500 -o 100000 -n 10 -N 10 -f
\$OUTF/\$NAME.sam \$REF \$OUTF/\$NAME.r1.sai \$OUTF/\$NAME.r2.sai \$READ1 \$READ2


### Convert SAM to BAM
\$SAMTOOLS/samtools view -b -S -o \$OUTF/\$NAME.bam \$OUTF/\$NAME.sam


### Sort BAM file
\$SAMTOOLS/samtools sort \$OUTF/\$NAME.bam \$OUTF/\$NAME.sort


### Index sorted BAM file
\$SAMTOOLS/samtools index \$OUTF/\$NAME.sort.bam


### Local Re-alignment

java -Xmx4g -jar \$GATK -T RealignerTargetCreator -R \$REF -I
\$OUTF/\$NAME.sort.bam -o \$OUTF/\$NAME.intervals -known $KNOWNINDEL --
minReadsAtLocus 6 --maxIntervalSize 200
java -Xmx4g -jar \$GATK -T IndelRealigner -R \$REF -I \$OUTF/\$NAME.sort.bam -
targetIntervals \$OUTF/\$NAME.intervals -o \$OUTF/\$NAME.realigned.bam -known
$KNOWNINDEL --maxReadsForRealignment 10000 --consensusDeterminationModel USE_SW
```

```
### Duplicate marking

java -Xmx4g -jar /users/GD/tools/picard/picard-tools-1.40/MarkDuplicates.jar
INPUT=\$OUTF/\$NAME.realigned.bam OUTPUT=\$OUTF/\$NAME.realigned.dm.bam
METRICS_FILE=\$OUTF/duplication_metrics.txt ASSUME_SORTED=true
VALIDATION_STRINGENCY=LENIENT
\$SAMTOOLS/samtools index \$OUTF/\$NAME.realigned.dm.bam

### Base quality recallibration

java -Xmx4g -jar \$GATK -T CountCovariates -nt 2 --default_platform illumina -cov
ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate -cov
DinucCovariate -recalFile \$OUTF/recal_data.csv -R \$REF -I
\$OUTF/\$NAME.realigned.dm.bam -knownSites $KNOWNINDEL
java -Xmx4g -jar \$GATK -T TableRecalibration --default_platform illumina -R
\$REF -I \$OUTF/\$NAME.realigned.dm.bam -recalFile \$OUTF/recal_data.csv --out
\$OUTF/\$NAME.realigned.dm.recalibrated.bam



### Cleanup
rm \$OUTF/\$NAME.sam
rm \$OUTF/\$NAME.bam
rm \$OUTF/\$NAME.realigned.bam
rm \$OUTF/\$NAME.realigned.bai
rm \$OUTF/\$NAME.realigned.dm.bam
rm \$OUTF/\$NAME.realigned.dm.bai
rm \$OUTF/\$NAME.realigned.dm.bam.bai


### GATK: Call SNPs and Indels with the GATK Unified Genotyper

java -Xmx4g -jar \$GATK -T UnifiedGenotyper -nt 2 -R \$REF -I
\$OUTF/\$NAME.realigned.dm.recalibrated.bam -o \$OUTF/GATK.snps.raw.vcf -glm SNP
java -Xmx4g -jar \$GATK -T UnifiedGenotyper -nt 2 -R \$REF -I
\$OUTF/\$NAME.realigned.dm.recalibrated.bam -o \$OUTF/GATK.indel.raw.vcf -glm
INDEL

### MPILEUP: Call SNPs and Indels

\$SAMTOOLS/samtools mpileup -uf \$REF \$OUTF/\$NAME.realigned.dm.recalibrated.bam
| \$SAMTOOLS/bcftools view -bcg - > \$OUTF/MPILEUP.variant.raw.bcf
\$SAMTOOLS/bcftools view \$OUTF/MPILEUP.variant.raw.bcf | \$SAMTOOLS/vcfutils.pl
varFilter -d5 -D$max_cov -W 20 > \$OUTF/MPILEUP.variant.raw.vcf
egrep \"INDEL|#\" \$OUTF/MPILEUP.variant.raw.vcf > \$OUTF/MPILEUP.indel.raw.vcf
grep -v INDEL \$OUTF/MPILEUP.variant.raw.vcf > \$OUTF/MPILEUP.snps.raw.vcf

### SHORE: Prepare format map.list
mkdir \$OUTF/shore

\$SHORE convert --sort -r \$REF -n 6 -g 1 -e 20 -s Alignment2Maplist
\$OUTF/\$NAME.realigned.dm.recalibrated.bam \$OUTF/shore/map.list.gz


### SHORE: compute coverage plot in GFF format for browsers
\$SHORE coverage -m \$OUTF/shore/map.list.gz -o \$OUTF/shore/CoverageAnalysis
```

```
### SHORE: Enrichment plots

grep enriched \$OUTF/shore/Count_SureSelect_plus150/meancov.txt | cut -f5 >
\$OUTF/shore/Count_SureSelect_plus150/exome_enriched.txt
grep depleted \$OUTF/shore/Count_SureSelect_plus150/meancov.txt | cut -f5 >
\$OUTF/shore/Count_SureSelect_plus150/exome_depleted.txt
grep enriched \$OUTF/shore/Count_SureSelect_plus150/readcount.txt | cut -f5 >
\$OUTF/shore/Count_SureSelect_plus150/exome_count_enriched.txt
grep depleted \$OUTF/shore/Count_SureSelect_plus150/readcount.txt | cut -f5 >
\$OUTF/shore/Count_SureSelect_plus150/exome_count_depleted.txt

### plot data

\$RSCRIPT \$NGSBOX/Statistics/R_examples/OCDpools_enrichment_stats.R
\$OUTF/shore/Count_SureSelect_plus150/


### SHORE: Call SNPs and Indels
\$SHORE qVar -n \$NAME -f /users/GD/resource/human/hg19/shore/hg19.fasta.shore -o
\$OUTF/shore/Variants -i \$OUTF/shore/map.list.gz -s
/users/GD/tools/shore/scoring_matrices/scoring_matrix_het.txt -E
\$OUTF/shore/Count_SureSelect_plus150/meancov.txt -e -c 4 -d 4 -C $max_cov -r 3 -
q 10 -Q 15 -a 0.25 -b 6 -y -v



### SHORE: Compute enrichment + plot CFTR
\$SHORE count -m \$OUTF/shore/map.list.gz -o \$OUTF/shore/CFTR -f
\$EXOME/CFTR_150.bed -H 1,1 -k

grep enriched \$OUTF/shore/CFTR/meancov.txt | cut -f5 >
\$OUTF/shore/CFTR/exome_enriched.txt
grep depleted \$OUTF/shore/CFTR/meancov.txt | cut -f5 >
\$OUTF/shore/CFTR/exome_depleted.txt
grep enriched \$OUTF/shore/CFTR/readcount.txt | cut -f5 >
\$OUTF/shore/CFTR/exome_count_enriched.txt
grep depleted \$OUTF/shore/CFTR/readcount.txt | cut -f5 >
\$OUTF/shore/CFTR/exome_count_depleted.txt

\$RSCRIPT \$NGSBOX/Statistics/R_examples/OCDpools_enrichment_stats.R
\$OUTF/shore/CFTR/


### Clean up
rm -r \$OUTF/shore/Variants/ConsensusAnalysis/supplementary_data
gzip -9 \$OUTF/shore/Variants/ConsensusAnalysis/reference.shore



### Filter and compare SNP calls from 3 different pipelines
# Filtering
mkdir \$OUTF/SNP_Intersection

java -jar -Xmx4g \$GATK -T VariantFiltration -R \$REF -o
\$OUTF/SNP_Intersection/GATK.snps.filtered.vcf --variant \$OUTF/GATK.snps.raw.vcf
--mask \$OUTF/GATK.indel.raw.vcf --clusterWindowSize 10 --filterExpression \"MQ <
30.0 || QUAL < 25.0 || QD < 4.0 || HRun > 9\" --filterName CRG --
genotypeFilterExpression \"DP < 5 || DP > $max_cov || GQ < 15\" --
genotypeFilterName CRGg

java -jar -Xmx4g \$GATK -T VariantFiltration -R \$REF -o
\$OUTF/SNP_Intersection/MPILEUP.snps.filtered.vcf --variant
\$OUTF/MPILEUP.snps.raw.vcf --mask \$OUTF/GATK.indel.raw.vcf --clusterWindowSize
```

```
10 --filterExpression \"MQ < 30.0 || QUAL < 15.0 || DP < 5 || DP > $max_cov\" --
filterName CRG --genotypeFilterExpression \"DP < 5 || DP > $max_cov || GQ < 15\"
--genotypeFilterName CRGg

perl \$NGSBOX/Parser/VCF/vcf_filter/vcf_filter.pl
\$OUTF/shore/Variants/ConsensusAnalysis/snp.vcf $max_cov >
\$OUTF/SHORE.snps.raw.vcf

java -jar -Xmx4g \$GATK -T VariantFiltration -R \$REF -o
\$OUTF/SNP_Intersection/SHORE.snps.filtered.vcf --variant
\$OUTF/SHORE.snps.raw.vcf --mask \$OUTF/GATK.indel.raw.vcf --clusterWindowSize 10
--filterExpression \"QUAL < 20.0 || DP < 5 || DP > $max_cov\" --filterName CRG

# greping
grep -v \"CRG\" \$OUTF/SNP_Intersection/GATK.snps.filtered.vcf | grep -v
\"SnpCluster\" > \$OUTF/SNP_Intersection/GATK.snps.filtered.cleaned.vcf
grep -v \"CRG\" \$OUTF/SNP_Intersection/MPILEUP.snps.filtered.vcf | grep -v
\"SnpCluster\" > \$OUTF/SNP_Intersection/MPILEUP.snps.filtered.cleaned.vcf
grep -v \"CRG\" \$OUTF/SNP_Intersection/SHORE.snps.filtered.vcf | grep -v
\"SnpCluster\" > \$OUTF/SNP_Intersection/SHORE.snps.filtered.cleaned.vcf

# Correct sample names in VFC files
sed -i -e \"s/FORMAT\\t\$NAME/FORMAT\\t\$NAME-GATK/\"
\$OUTF/SNP_Intersection/GATK.snps.filtered.cleaned.vcf
sed -i -e \"s/FORMAT\\t\$NAME/FORMAT\\t\$NAME-MPILEUP/\"
\$OUTF/SNP_Intersection/MPILEUP.snps.filtered.cleaned.vcf
sed -i -e \"s/FORMAT\\t\$NAME/FORMAT\\t\$NAME-SHORE/\"
\$OUTF/SNP_Intersection/SHORE.snps.filtered.cleaned.vcf

# Intersecting
java -jar -Xmx4g \$GATK -T CombineVariants -R \$REF -genotypeMergeOptions
PRIORITIZE -V:SHORE \$OUTF/SNP_Intersection/SHORE.snps.filtered.cleaned.vcf -
V:GATK \$OUTF/SNP_Intersection/GATK.snps.filtered.cleaned.vcf -V:MPILEUP
\$OUTF/SNP_Intersection/MPILEUP.snps.filtered.cleaned.vcf -priority
GATK,MPILEUP,SHORE -o \$OUTF/SNP_Intersection/merged.vcf

# Evaluation
java -jar -Xmx4g \$GATK -T VariantEval -R \$REF --dbsnp $KNOWNINDEL -select
'set==\"Intersection\"' -selectName Intersection -select 'set==\"SHORE\"' -
selectName SHORE -select 'set==\"MPILEUP\"' -selectName MPILEUP -select
'set==\"GATK\"' -selectName GATK -select 'set==\"GATK-MPILEUP\"' -selectName
GATK_MPILEUP -select 'set==\"GATK-SHORE\"' -selectName GATK_SHORE -select
'set==\"MPILEUP-SHORE\"' -selectName MPILEUP_SHORE -o
\$OUTF/SNP_Intersection/report.all.txt --eval \$OUTF/SNP_Intersection/merged.vcf
-l INFO

# Annotate Enrichment
perl \$NGSBOX/Parser/VCF/vcf_filter/vcf_filter_enriched.pl
\$EXOME/NimbleGen_Tiled_Regions_150.bed \$OUTF/SNP_Intersection/merged.vcf >
\$OUTF/SNP_Intersection/merged.all.vcf

# Evaluate calls on enriched regions
grep -v \"NOTENRICHED\" \$OUTF/SNP_Intersection/merged.all.vcf >
\$OUTF/SNP_Intersection/merged.enriched.vcf

java -jar -Xmx4g \$GATK -T VariantEval -R \$REF --dbsnp $KNOWNINDEL -select
'set==\"Intersection\"' -selectName Intersection -select 'set==\"SHORE\"' -
selectName SHORE -select 'set==\"MPILEUP\"' -selectName MPILEUP -select
'set==\"GATK\"' -selectName GATK -select 'set==\"GATK-MPILEUP\"' -selectName
GATK_MPILEUP -select 'set==\"GATK-SHORE\"' -selectName GATK_SHORE -select
'set==\"MPILEUP-SHORE\"' -selectName MPILEUP_SHORE -o
\$OUTF/SNP_Intersection/report.enriched.txt --eval
\$OUTF/SNP_Intersection/merged.enriched.vcf -l INFO
```

```
### Filter and compare indel calls from 3 different pipelines
# Filtering
mkdir \$OUTF/Indel_Intersection

java -jar -Xmx4g \$GATK -T VariantFiltration -R \$REF -o
\$OUTF/Indel_Intersection/GATK.indel.filtered.vcf --variant
\$OUTF/GATK.indel.raw.vcf --filterExpression \"MQ < 30.0 || QUAL < 20.0 || MQ0 >
5 || QD < 4.0 || HRun > 9\" --filterName CRG --genotypeFilterExpression \"DP < 5
|| DP > $max_cov || GQ < 15\" --genotypeFilterName CRGg

java -jar -Xmx4g \$GATK -T VariantFiltration -R \$REF -o
\$OUTF/Indel_Intersection/MPILEUP.indel.filtered.vcf --variant
\$OUTF/MPILEUP.indel.raw.vcf --filterExpression \"MQ < 30.0 || QUAL < 10.0 || DP
< 5 || DP > $max_cov\" --filterName CRG --genotypeFilterExpression \"DP < 5 || DP
> $max_cov || GQ < 15\" --genotypeFilterName CRGg

java -jar -Xmx4g \$GATK -T VariantFiltration -R \$REF -o
\$OUTF/Indel_Intersection/SHORE.indel.filtered.vcf --variant
\$OUTF/shore/Variants/ConsensusAnalysis/indels.vcf --filterExpression \"QUAL <
2.0 || DP < 4 || DP > $max_cov || RE > 1.3\" --filterName CRG

# greping
grep -v \"CRG\" \$OUTF/Indel_Intersection/GATK.indel.filtered.vcf >
\$OUTF/Indel_Intersection/GATK.indel.filtered.cleaned.vcf
grep -v \"CRG\" \$OUTF/Indel_Intersection/MPILEUP.indel.filtered.vcf >
\$OUTF/Indel_Intersection/MPILEUP.indel.filtered.cleaned.vcf
grep  -v \"CRG\" \$OUTF/Indel_Intersection/SHORE.indel.filtered.vcf | grep -v
\"SHOREFILTER\" > \$OUTF/Indel_Intersection/SHORE.indel.filtered.cleaned.vcf

# Correct sample names in VFC files
sed -i -e \"s/FORMAT\\t\$NAME/FORMAT\\t\$NAME-GATK/\"
\$OUTF/Indel_Intersection/GATK.indel.filtered.cleaned.vcf
sed -i -e \"s/FORMAT\\t\$NAME/FORMAT\\t\$NAME-MPILEUP/\"
\$OUTF/Indel_Intersection/MPILEUP.indel.filtered.cleaned.vcf
sed -i -e \"s/FORMAT\\t\$NAME/FORMAT\\t\$NAME-SHORE/\"
\$OUTF/Indel_Intersection/SHORE.indel.filtered.cleaned.vcf



# Intersecting
java -jar -Xmx4g \$GATK -T CombineVariants -R \$REF -genotypeMergeOptions
PRIORITIZE -V:SHORE \$OUTF/Indel_Intersection/SHORE.indel.filtered.cleaned.vcf -
V:GATK \$OUTF/Indel_Intersection/GATK.indel.filtered.cleaned.vcf -V:MPILEUP
\$OUTF/Indel_Intersection/MPILEUP.indel.filtered.cleaned.vcf -priority
GATK,MPILEUP,SHORE -o \$OUTF/Indel_Intersection/merged.vcf

# Evaluation
java -jar -Xmx4g \$GATK -T VariantEval -R \$REF --dbsnp $KNOWNINDEL -select
'set==\"Intersection\"' -selectName Intersection -select 'set==\"SHORE\"' -
selectName SHORE -select 'set==\"MPILEUP\"' -selectName MPILEUP -select
'set==\"GATK\"' -selectName GATK -select 'set==\"GATK-MPILEUP\"' -selectName
GATK_MPILEUP -select 'set==\"GATK-SHORE\"' -selectName GATK_SHORE -select
'set==\"MPILEUP-SHORE\"' -selectName MPILEUP_SHORE -o
\$OUTF/Indel_Intersection/report.all.txt --eval
\$OUTF/Indel_Intersection/merged.vcf -l INFO

# Annotate Enrichment
perl \$NGSBOX/Parser/VCF/vcf_filter/vcf_filter_enriched.pl
\$EXOME/NimbleGen_Tiled_Regions_150.bed \$OUTF/Indel_Intersection/merged.vcf >
\$OUTF/Indel_Intersection/merged.all.vcf
```

```
# Evaluate calls on enriched regions
grep -v \"NOTENRICHED\" \$OUTF/Indel_Intersection/merged.all.vcf >
\$OUTF/Indel_Intersection/merged.enriched.vcf

java -jar -Xmx4g \$GATK -T VariantEval -R \$REF --dbsnp $KNOWNINDEL -select
'set==\"Intersection\"' -selectName Intersection -select 'set==\"SHORE\"' -
selectName SHORE -select 'set==\"MPILEUP\"' -selectName MPILEUP -select
'set==\"GATK\"' -selectName GATK -select 'set==\"GATK-MPILEUP\"' -selectName
GATK_MPILEUP -select 'set==\"GATK-SHORE\"' -selectName GATK_SHORE -select
'set==\"MPILEUP-SHORE\"' -selectName MPILEUP_SHORE -o
\$OUTF/Indel_Intersection/report.enriched.txt --eval
\$OUTF/Indel_Intersection/merged.enriched.vcf -l INFO


### Annotate SNPs with ANNOVAR: Union, any tool predicts SNP
mkdir \$OUTF/SNP_Intersection/AnnovarUnion
 \$ANNOVAR/convert2annovar.pl -includeinfo -format vcf4
\$OUTF/SNP_Intersection/merged.all.vcf >
\$OUTF/SNP_Intersection/AnnovarUnion/snps.ann
 \$ANNOVAR/custom_summarize_annovar.pl --buildver hg19 --outfile
\$OUTF/SNP_Intersection/AnnovarUnion/sum
\$OUTF/SNP_Intersection/AnnovarUnion/snps.ann --ver1000g 1000g2011may
\$ANNOVAR/hg19/


### Annotate Indels with ANNOVAR: Union, indels predicted by any tool
mkdir \$OUTF/Indel_Intersection/AnnovarUnion
 \$ANNOVAR/convert2annovar.pl -includeinfo -format vcf4
\$OUTF/Indel_Intersection/merged.all.vcf >
\$OUTF/Indel_Intersection/AnnovarUnion/indels.ann
 \$ANNOVAR/custom_summarize_annovar.pl --buildver hg19 --outfile
\$OUTF/Indel_Intersection/AnnovarUnion/sum
\$OUTF/Indel_Intersection/AnnovarUnion/indels.ann --ver1000g 1000g2011may
\$ANNOVAR/hg19/


### Clean up
rm \$OUTF/MPILEUP.variant.raw.bcf

\n");


print OUT @qsub;
close OUT;
```

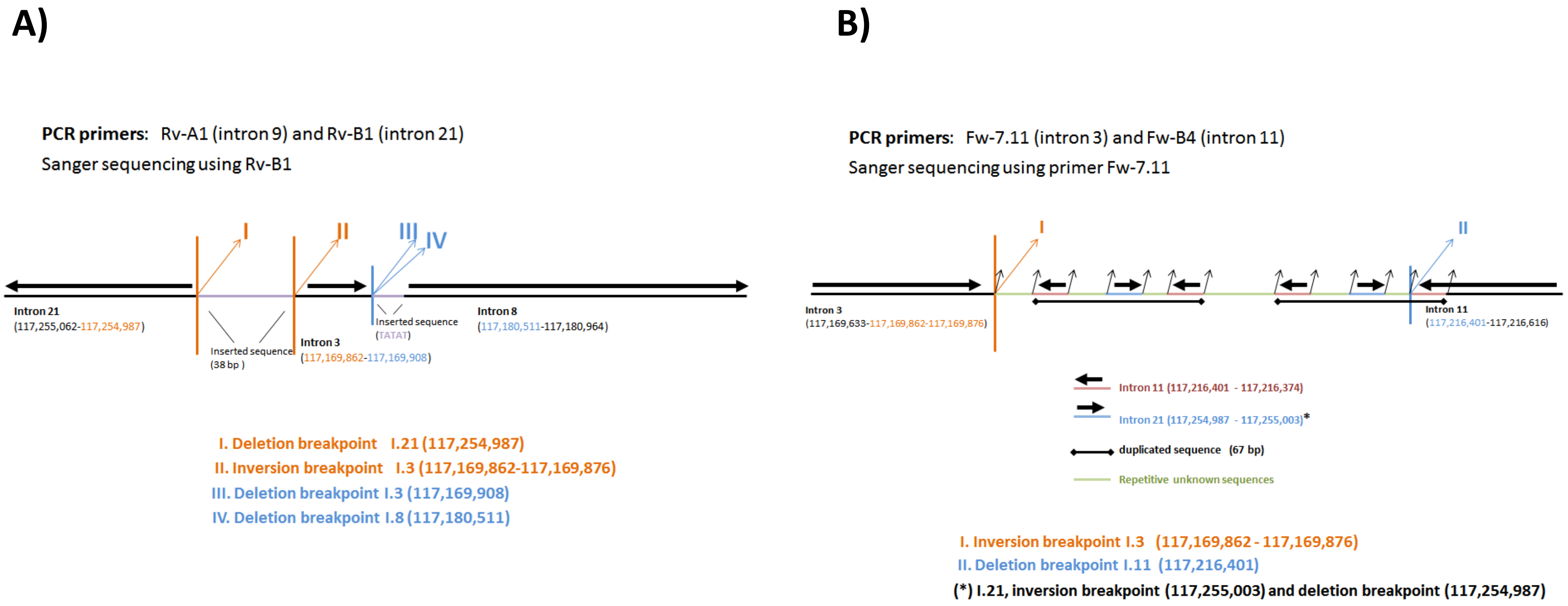
A)
PCR primers: Rv-A1 (intron 9) and Rv-B1 (intron 21)
Sanger sequencing using Rv-B1
I
II
III
IV
Intron 21
(117,255,062-117,254,987)
Inserted sequence
(38 bp)
Intron 3
(117,169,862-117,169,908)
Inserted sequence
(TATAT)
Intron 8
(117,180,511-117,180,964)
I. Deletion breakpoint I.21 (117,254,987)
II. Inversion breakpoint I.3 (117,169,862-117,169,876)
III. Deletion breakpoint I.3 (117,169,908)
IV. Deletion breakpoint I.8 (117,180,511)
B)
PCR primers: Fw-7.11 (intron 3) and Fw-B4 (intron 11)
Sanger sequencing using primer Fw-7.11
I
II
Intron 3
(117,169,633-117,169,862-117,169,876)
Intron 11
(117,216,401-117,216,616)
Intron 11 (117,216,401 - 117,216,374)
Intron 21 (117,254,987 - 117,255,003)*
duplicated sequence (67 bp)
Repetitive unknown sequences
I. Inversion breakpoint I.3 (117,169,862 - 117,169,876)
II. Deletion breakpoint I.11 (117,216,401)
(*) I.21, inversion breakpoint (117,255,003) and deletion breakpoint (117,254,987)

Fig. S1. Diagram showing the position of the primers used for the validation of the breakpoints of CFTR50kbdel. (A) Regions amplified with PCR primers Rv-A1 and Rv-B1. (B) Regions amplified with PCR primers Fw-7.11 and Fw-B4.

Table S1. Sequencing quality control parameters, coverage and detected variants by targeted resequencing of the *CFTR* gene by sample

	Sample	1FQ	2FQ	3FQ	4FQ	5FQ	6FQ	7FQ
Sequencing	QC-passed reads	17.255.934	19.355.484	17.214.012	15.075.448	22.476.580	18.270.438	21.955.770
	% Mapped	97,27	97,20	97,03	97,04	97,02	97,22	97,16
	% Properly paired	96,22	96,07	95,98	95,94	95,87	96,05	96,04
Coverage	Mean coverage (X)	375	451	377	315	503	398	484
	Mean coverage extended 150 bp (X)	324	390	325	269	436	342	420
	Max coverage (X)	1188	1263	1137	1087	1455	1309	1489
	% Enrichment	58,23	59,65	58,29	57,40	59,03	58,43	58,45
	% target bases covered = 0X	0,102	0,077	0,007	0,052	0,000	0,000	0,000
	% target bases covered >= 1X	99,90	99,92	99,99	99,95	100,00	100,00	100,00
	% target bases covered >= 5X	99,81	99,84	99,83	99,75	99,87	99,92	99,85
	% target bases covered >= 10X	99,74	99,79	99,78	99,65	99,80	99,79	99,78
	% target bases covered >= 20X	99,61	99,70	99,64	99,33	99,70	99,62	99,69
	% target bases covered >= 50X	98,78	99,19	98,79	97,66	99,22	98,77	99,17
	% target bases covered >= 100X	95,64	97,43	96,01	92,74	97,71	96,05	97,41
Variants	SNVs	205	130	141	82	131	135	45
	Novel SNVs	2	4	9	6	1	6	4
	Exonic SNVs	3	1	3	2	2	2	1
	Missense, nonsense and splice site SNPs	1	2	1	2	2	2	2
	SNVs in CFTR database	3	2	3	2	2	2	2
	InDels	48	24	36	20	24	26	13
	Novel InDels	32	16	23	13	16	18	7
	Frameshift and non-Frameshift InDels	2	1	1	0	0	1	0
	InDels annotated in CFTR database	2	1	1	0	0	1	0

8FQ	9FQ	10FQ	11FQ	12FQ	14FQ	15FQ	16FQ	17FQ	18FQ	19FQ
21.895.400	17.133.058	19.670.588	23.977.106	13.535.418	13.022.662	17.755.884	19.176.270	16.841.600	7.730.556	9.645.318
97,11	96,40	96,18	95,83	95,12	96,71	96,53	97,08	96,98	95,92	95,34
96,04	95,15	94,72	94,00	93,42	95,56	95,52	96,12	95,93	94,42	93,60
499	375	407	482	260	236	361	391	356	99	114
430	357	418	225	206	317	344	310	328	87	99
1503	1284	1382	2074	1100	973	1225	1299	1463	408	496
58,84	57,50	56,93	56,19	55,20	57,25	57,30	58,22	58,36	44,35	42,79
0,004	0,000	0,000	0,000	0,051	0,003	0,000	0,028	0,000	0,094	0,065
100,00	100,00	100,00	100,00	99,95	100,00	100,00	99,97	100,00	99,91	99,94
99,84	99,96	99,96	99,82	99,82	99,79	99,92	99,82	99,81	99,55	99,58
99,79	99,84	99,82	99,76	99,70	99,72	99,80	99,79	99,77	98,89	98,98
99,71	99,74	99,71	99,64	99,23	99,26	99,70	99,72	99,65	96,70	97,15
99,29	99,21	99,18	98,92	96,37	96,58	99,11	99,16	98,54	81,12	85,12
97,72	96,62	96,64	96,71	86,86	85,47	96,41	96,64	95,02	42,05	51,91
126	125	204	76	17	41	154	22	70	63	120
7	3	8	4	3	4	6	9	6	3	1
3	4	2	3	0	0	3	0	1	1	5
1	2	1	2	0	0	2	0	0	0	2
3	4	3	3	0	0	3	0	1	1	5
40	36	58	29	10	19	46	10	21	22	38
26	22	39	21	8	13	28	6	14	15	29
0	0	0	0	1	1	0	0	1	1	0
0	0	0	0	1	1	0	0	1	1	0

20FQ	21FQ	22FQ	23FQ	24FQ	25FQ	26FQ	27FQ	29FQ	30FQ	31FQ
9.823.268	6.484.528	7.110.480	7.678.796	9.785.126	9.332.386	10.205.974	8.380.250	7.333.946	7.660.972	8.108.704
94,43	95,25	94,88	94,71	91,10	95,22	93,67	95,05	93,63	95,13	96,55
92,69	93,60	92,94	92,38	87,21	93,23	91,56	93,11	91,79	93,29	95,25
113	77	85	89	93	114	106	105	107	105	151
98	68	74	78	82	99	93	90	92	92	130
472	310	370	427	583	445	430	401	364	380	640
42,44	42,88	43,10	42,97	39,31	44,08	40,16	43,96	47,30	46,17	55,51
0,167	0,183	0,099	0,055	0,157	0,100	0,127	0,175	0,164	0,074	0,151
99,83	99,82	99,90	99,95	99,84	99,90	99,87	99,82	99,84	99,93	99,85
99,52	99,31	99,32	99,16	98,91	99,53	99,52	99,45	99,46	99,61	99,46
98,95	98,22	98,00	97,89	97,46	98,88	98,80	98,78	98,79	98,95	98,79
96,97	94,12	94,50	94,42	92,78	97,12	96,64	96,66	96,94	97,08	97,15
83,53	68,31	72,49	73,02	70,89	84,62	82,36	82,21	83,76	83,64	88,72
50,32	26,15	32,60	35,57	37,83	52,14	46,71	46,00	48,91	47,49	65,60
122	127	201	67	116	157	130	16	208	129	173
6	1	1	2	5	1	2	3	2	5	2
4	2	3	4	7	2	2	1	4	2	3
2	2	2	3	4	2	2	0	3	3	2
4	2	4	4	7	2	2	1	5	3	3
35	27	36	22	30	29	23	12	42	25	33
22	18	24	14	21	15	14	9	29	17	20
0	1	1	0	0	2	0	1	0	0	0
0	1	1	0	0	2	0	1	0	0	0

32FQ	33FQ	34FQ	35FQ	38FQ	39FQ	40FQ	41FQ	42FQ	43FQ	44FQ
12.267.092	11.960.280	12.673.174	8.834.572	12.403.370	11.188.176	12.761.496	13.361.576	11.251.326	14.216.384	10.825.984
96,41	96,52	96,49	95,63	96,14	95,59	96,46	96,58	96,83	96,54	96,63
95,07	95,25	95,17	94,17	94,99	93,99	95,18	95,31	95,68	95,29	95,40
239	220	240	162	231	208	255	250	244	265	206
205	190	208	139	202	182	218	217	211	231	179
922	861	944	623	819	710	1810	964	794	1074	782
56,53	55,44	55,45	55,48	55,47	55,41	56,40	55,07	57,59	55,23	56,02
0,158	0,171	0,068	0,148	0,064	0,118	0,034	0,080	0,082	0,073	0,084
99,84	99,83	99,93	99,85	99,94	99,88	99,97	99,92	99,92	99,93	99,92
99,77	99,72	99,74	99,62	99,76	99,78	99,72	99,70	99,79	99,80	99,74
99,65	99,46	99,61	99,12	99,58	99,59	99,63	99,59	99,72	99,59	99,61
99,21	98,92	99,13	97,87	98,93	99,11	99,06	98,98	99,47	99,12	98,93
96,65	95,52	96,24	90,56	95,83	95,82	96,62	96,07	97,28	96,62	95,27
87,69	83,88	85,91	69,87	85,63	84,64	88,27	86,36	90,07	88,56	82,30
128	132	125	200	14	133	145	132	16	105	126
4	2	2	2	2	10	2	5	2	4	1
5	2	2	4	1	4	3	6	0	3	2
3	2	2	2	1	2	2	4	1	2	2
6	2	2	4	1	4	3	6	1	2	2
49	29	24	41	13	38	35	38	12	28	23
37	21	16	27	11	25	20	24	9	16	14
0	1	1	0	0	0	0	1	0	0	0
0	1	1	0	0	0	0	1	0	0	0

45FQ	46FQ	47FQ	48FQ	49FQ	50FQ	51FQ	52FQ	53FQ	54FQ	55FQ
6.637.998	8.190.110	8.038.462	8.284.024	9.080.058	8.159.248	7.664.812	9.987.790	10.079.020	8.358.090	8.962.716
93,46	96,74	96,96	96,83	96,50	96,55	96,90	96,13	96,64	96,70	96,54
91,05	95,48	95,80	95,68	95,09	95,17	95,71	94,70	95,43	95,38	95,24
141	155	157	160	183	167	150	216	205	167	176
119	134	136	138	157	144	130	187	179	144	152
459	610	568	554	661	624	531	643	737	553	594
57,16	55,14	55,64	55,33	56,38	56,62	56,09	57,72	56,36	56,52	56,49
0,048	0,113	0,161	0,131	0,094	0,118	0,050	0,034	0,044	0,078	0,077
99,95	99,89	99,84	99,87	99,91	99,88	99,95	99,97	99,96	99,92	99,92
99,68	99,74	99,71	99,75	99,75	99,77	99,72	99,80	99,82	99,75	99,75
99,21	99,47	99,47	99,51	99,58	99,62	99,51	99,68	99,68	99,55	99,65
98,10	98,57	98,61	98,81	98,90	99,00	98,58	99,41	99,42	98,90	99,18
90,82	93,01	93,85	93,89	95,50	94,75	93,47	96,86	96,95	94,55	95,39
67,29	71,05	72,47	73,27	79,77	76,81	71,04	88,04	86,83	75,95	78,97
158	130	130	126	197	124	73	139	190	152	131
4	1	0	1	4	4	3	4	4	5	10
3	3	2	2	6	4	3	3	4	1	5
3	2	2	2	3	2	3	1	3	2	3
4	3	2	2	6	4	3	3	5	2	5
33	25	23	26	56	44	25	39	48	28	37
24	16	15	18	46	32	17	27	33	20	23
0	1	1	1	0	0	0	1	0	1	0
0	1	1	1	0	0	0	1	0	1	0

56FQ	57FQ	58FQ	59FQ	60FQ	61FQ	62FQ	63FQ	64FQ	65FQ	66FQ
8.455.604	8.841.018	9.609.066	8.256.002	9.228.290	10.798.998	11.362.794	9.386.230	14.580.506	10.231.428	11.384.616
96,61	96,45	96,61	96,70	96,07	93,40	86,52	95,78	95,30	95,92	96,17
95,30	95,20	95,35	95,41	94,71	91,20	81,72	94,30	93,58	94,30	94,64
164	175	209	164	185	222	254	182	305	209	225
143	152	182	144	161	189	215	156	262	180	195
625	663	734	599	682	718	702	708	1005	2431	857
56,23	56,55	57,69	56,34	56,46	56,77	59,41	55,66	57,30	56,59	55,49
0,092	0,042	0,128	0,155	0,007	0,021	0,012	0,115	0,100	0,068	0,017
99,91	99,96	99,87	99,85	99,99	99,98	99,99	99,89	99,90	99,93	99,98
99,75	99,79	99,79	99,75	99,82	99,76	99,89	99,76	99,79	99,73	99,84
99,58	99,59	99,69	99,59	99,61	99,57	99,75	99,59	99,74	99,40	99,69
98,86	99,07	99,38	98,81	99,16	99,06	99,55	98,71	99,55	98,72	99,17
94,20	95,34	97,27	94,33	95,70	95,98	97,66	94,28	98,13	94,63	95,96
75,49	79,55	88,06	75,22	81,23	86,10	92,28	77,74	93,68	80,19	85,34
132	128	13	33	129	195	175	43	133	146	152
3	5	2	1	6	2	3	5	1	4	4
3	1	0	1	5	4	3	1	2	4	5
3	2	0	1	4	1	2	1	2	3	4
3	2	0	1	5	4	3	1	2	4	5
21	26	12	14	36	45	35	14	23	34	35
13	18	9	7	22	30	20	9	14	18	18
0	0	1	0	0	1	0	1	1	0	0
0	0	1	0	0	1	1	1	1	0	0

67FQ	68FQ	69FQ	70FQ	71FQ	72FQ	73FQ	74FQ	75FQ	76FQ	77FQ
15.272.850	10.516.996	10.842.778	8.913.236	4.497.442	8.974.522	14.370.752	3.974.106	9.772.282	12.316.454	12.962.294
95,23	96,05	95,80	92,32	95,23	94,85	93,58	93,64	95,39	92,43	87,14
93,25	94,66	94,28	89,66	93,11	92,95	91,09	91,42	93,62	89,37	83,54
283	211	199	204	89	187	289	79	198	253	292
245	184	168	170	77	161	247	68	170	216	251
1110	747	778	768	391	697	965	325	712	875	900
54,97	56,12	54,52	59,79	55,55	57,13	57,15	56,28	56,52	57,45	60,03
0,002	0,142	0,126	0,138	0,047	0,053	0,010	0,150	0,147	0,014	0,013
100,00	99,86	99,87	99,86	99,95	99,95	99,99	99,85	99,85	99,99	99,99
99,80	99,78	99,74	99,72	99,08	99,65	99,87	98,69	99,75	99,87	99,85
99,65	99,68	99,52	99,45	97,77	99,24	99,71	96,68	99,61	99,65	99,77
99,25	99,24	98,53	98,67	93,78	98,17	99,45	91,48	99,00	99,31	99,55
96,79	96,00	93,74	94,82	72,62	93,38	97,65	65,58	95,61	96,97	97,89
89,63	84,32	77,82	81,38	35,30	77,06	92,31	28,32	82,22	88,98	93,12
42	21	180	14	193	43	129	124	215	125	142
6	5	4	3	1	3	3	3	3	1	3
2	1	3	1	4	0	5	1	4	4	2
2	1	2	1	2	1	3	1	2	2	2
2	1	3	1	5	1	5	2	4	5	2
10	9	36	13	40	16	40	23	39	33	34
4	5	22	11	26	9	26	15	25	20	26
0	0	1	0	1	1	0	1	1	0	1
0	0	1	0	1	1	0	1	1	0	1

78FQ	**79FQ**	**80FQ**	**81FQ**	**82FQ**	**83FQ**	**84FQ**	**85FQ**	**86FQ**	**87FQ**	**88FQ**
11.402.816	13.845.360	10.442.674	9.996.370	13.482.678	13.020.756	22.079.312	12.873.084	15.079.064	9.773.980	11.544.038
91,20	90,14	92,99	93,06	93,22	92,12	90,27	91,21	85,60	93,21	91,61
87,46	86,47	90,38	90,81	90,43	89,52	86,59	87,57	80,68	90,55	88,40
228	315	242	242	279	257	332	276	326	209	261
196	272	207	208	237	220	278	235	279	178	224
894	998	784	813	1032	970	1585	1010	1072	793	957
57,46	60,00	59,81	60,28	58,17	56,40	50,53	58,44	59,54	58,15	59,05
0,066	0,061	0,015	0,145	0,057	0,060	0,121	0,026	0,009	0,066	0,023
99,93	99,94	99,99	99,85	99,94	99,94	99,88	99,97	99,99	99,93	99,98
99,73	99,83	99,84	99,76	99,73	99,75	99,66	99,83	99,87	99,73	99,80
99,48	99,74	99,71	99,69	99,58	99,59	99,43	99,65	99,77	99,47	99,61
98,79	99,56	99,28	99,37	98,99	99,00	98,71	99,13	99,48	98,57	99,07
95,18	98,03	97,08	97,13	96,31	96,26	95,36	96,63	97,76	94,27	96,14
83,15	94,15	89,47	90,85	88,65	87,33	86,94	89,19	93,57	80,78	88,11
19	124	150	131	87	136	21	127	24	26	131
4	4	8	4	5	2	7	6	5	7	5
0	3	2	6	3	2	1	4	1	4	1
0	3	2	4	2	1	1	2	2	4	1
0	4	2	6	3	2	2	5	2	4	1
9	32	29	38	24	23	7	35	9	10	23
7	20	17	24	16	15	5	21	7	7	15
0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0

89FQ	90FQ	91FQ	92FQ	93FQ	94FQ	95FQ	96FQ	1FQ_bis	2FQ_bis	3FQ_bis
13.582.002	12.632.482	13.398.820	7.956.772	12.087.950	11.894.622	11.334.266	9.813.002	8.723.572	9.721.670	7.842.304
93,20	93,97	92,37	95,36	90,30	89,96	94,74	95,78	96,33	96,08	96,02
90,19	91,15	88,85	93,50	86,08	85,93	92,27	94,18	95,00	94,68	94,67
290	276	286	150	254	241	235	199	171	204	154
248	240	245	128	216	205	201	173	148	175	135
1088	1016	1045	664	997	1001	947	735	733	707	602
58,42	58,45	58,52	55,75	58,20	57,81	57,22	56,85	56,63	57,86	56,50
0,023	0,027	0,031	0,196	0,018	0,027	0,017	0,020	0,139	0,131	0,069
99,98	99,97	99,97	99,80	99,98	99,97	99,98	99,98	99,86	99,87	99,93
99,80	99,79	99,81	99,42	99,79	99,79	99,67	99,66	99,63	99,72	99,56
99,66	99,64	99,65	98,76	99,45	99,47	99,39	99,31	99,13	99,52	98,95
99,14	99,23	99,23	96,79	98,68	98,59	98,66	98,21	97,95	98,80	97,48
96,64	96,80	96,72	86,89	95,11	95,31	95,07	93,11	91,29	94,98	89,70
89,65	89,98	89,70	63,50	85,15	84,56	83,22	77,37	70,67	81,49	67,30
18	131	129	199	16	122	129	80	203	130	135
4	4	3	3	2	2	3	3	4	4	9
1	2	2	4	1	4	3	3	3	1	3
1	2	2	2	1	2	1	2	1	2	1
1	2	2	4	1	4	3	3	3	2	3
9	20	26	44	10	33	36	21	45	30	35
7	11	17	30	7	21	22	13	30	23	23
0	0	1	0	0	0	1	1	2	1	1
0	0	1	0	0	0	1	1	2	1	1

4FQ_bis
7.208.564
95,93
94,53
137
118
574
55,95
0,112
99,89
99,36
98,48
96,39
85,10
59,45
78
4
2
2
2
19
11
0
0

Table S2. Common *CFTR* polymorphisms identified in the 92 samples

Legacy name	HGVS name	hg19 position	nt change	dbSNP
-102T>A	c.-234T/A	117119915	T>A	-
125G>C	c.-8G>C	117120141	G>C	rs1800501
875+40A>G	c.743+40A>G	117175505	A>G	rs1800502
1001+11C>T	c.869+11C>T	117176738	C>T	rs1800503
1525-61A>G	c.1393-61A>G	117199457	A>G	rs34855237
1898+152T>A	c.1766+152T>A	117230645	T>A	rs4148711
3601-65C>A	c.3469-65C>A	117267511	C>A	rs213989
R74W	c.220C>T	117149143	C>T	rs115545701
R74Q	c.221G>A	117149144	G>A	-
I148T	c.443T>C	117171122	T>C	rs35516286
T351S	c.1052C>G	117180336	C>G	rs1800086
I506V	c.1516A>G	117199641	A>G	rs1800091
F508C	c.1523T>G	117199648	T>G	rs74571530
1540A>G (M470V)	c.1408A>G	117199533	G>A	rs213950
2694T>G (T854T)	c.2562T>G	117235055	T>G	rs1042077
2736A>G (V868V)	c.2604A>G	117235097	A>G	rs1800105
3030G>A (T966T)	c.2898G>A	117243826	G>A	rs1800109
3041-71G>C	c.2909-71G>C	117246657	G>C	rs34830471
I1027T	c.3080T>C	117250664	T>C	rs1800112
4002A>G (P1290P)	c.3870A>G	117282644	A>G	rs1800130
4029A>G (T1299T)	c.3897A>G	117292919	A>G	rs1800131
4404C>T (Y1424Y)	c.4272C>T	117306991	C>T	rs1800135
4521G>A (Q1463Q)	c.4389G>A	117307108	G>A	rs1800136
GATT repeat	-	117176569	GATT	rs67140043
TG repeat	c.1210-34[TG]	117188660	TG	rs3832534
TG repeat	-	117188661	TG	rs67408451
TG repeat	-	117188662	TG	-
1342-13G>T	c.1210-13G>T	117188682	G>T	rs10229820
3041-92G>A	c.2909-92G>A	117246636	G>A	rs35050470
3500-140A>G	c.3368-140A>G	117254527	A>G	rs213981
4269-139G>A	c.4137-139G>A	117305374	G>A	rs4727855
186-155G>C	c.54-155G>C	117144152	G>C	.
712-159G>A	c.580-159G>A	117175143	G>A	rs34159932
2751+106T>A	c.2619+106T>A	117235218	T>A	rs4148713
2751+85_2751+86delAT	-	117235197	AT	-

Table S3. Pathogenic *CFTR* mutations identified in the 92 samples

Sample	Phenotype	Allele	Region	HGVS Name	Legacy Name	total counts	% mutation counts
1	CF	allele 1	exon 11	c.1519_1521delATC	[delta]I507	330	50,61
		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	334	38,62
1_bis	CF	allele 1	exon 11	c.1519_1521delATC	[delta]I507	146	58,21
		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	149	36,24
2	CF	allele 1	exon 11	c.1521_1523delCTT	[delta]F508	464	39,22
		allele 2	intron 4	c.489+1G>T	621+ 1G- >T	510	49,02
2_bis	CF	allele 1	exon 11	c.1521_1523delCTT	[delta]F508	184	47,82
		allele 2	intron 4	c.489+1G>T	621+ 1G- >T	272	44,85
3	CFTR-RD	allele 1	exon 14	c.2051_2052delAAinsG	2183AA- >G	243	47,33
		allele 2	intron 9	c.1210-34TG(12)T(5)	5T-12TG	.	.
3_bis	CFTR-RD	allele 1	exon 14	c.2051_2052delAAinsG	2183AA- >G	73	56,16
		allele 2	intron 9	c.1210-34TG(12)T(5)	5T-12TG	.	.
4	CFTR-RD	allele 1	exon 22	c.3484C>T	R1162X	262	54,58
		allele 2	intron 9	c.1210-34TG(12)T(5)	5T-12TG	.	.
4_bis	CFTR-RD	allele 1	exon 22	c.3484C>T	R1162X	109	41,28
		allele 2	intron 9	c.1210-34TG(12)T(5)	5T-12TG	.	.
5	CF CARRIER	allele 1	exon 24	c.3909C>G	N1303K	378	50
6	CF	allele 1	exon 21	c.3454G>C	D1152H	417	47,48
		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	339	35,1
7	CFTR-RD	allele 1	exon 19	c.2991G>C	L997F	612	50,98
		allele 2	intron 5	c.579+1G>T	711+ 1G- >T	301	42,19
8	CF CARRIER	allele 1	exon 14	c.1817_1900delAAATGGAACATTTAAAGAAAGCTGACAAAATATTAATTTTGCATGAAGGTAGCAGCTATTTTTATGGGACATTTTCAGAACTCC	1949del84	.	.
9	CF CARRIER	allele 1	exon 8	c.1040G>A	R347H	429	51,28
10	CF CARRIER	allele 1	intron 11	c.1585-1G>A	1717- 1G- >A	316	42,09
11	CF CARRIER	allele 1	exon 4	c.350G>A	R117H	595	51,43
12	CF CARRIER	allele 1	exon 23	c.3773_3774insT	3905insT	293	36,52
14	CF CARRIER	allele 1	exon 22	c.3527delC:p.T1176fs	3659delC	219	40,18
15	CF CARRIER	allele 1a	exon 22	c.3484C>T	R1162X (in cis)	312	48,08
		allele 1b	intron 9	c.1210-34TG(13)T(5)	5T-13TG (in cis)	.	.
16	CF CARRIER	allele 1	intron 22	c.3717+12191C>T	3849+ 10kbC- >T	367	47,14
17	CF CARRIER	allele 1	exon 14	c.2471delT:p.I824fs	2603delT	383	48,04
18	CF CARRIER	allele 1	exon 14	c.1919_1920delTT:p.640_640del	2051delTT	47	48,94
19	CF	allele 1	exon 20	c.3196C>T	R1066C	133	95,49
		allele 2	exon 20	c.3196C>T	R1066C	133	95,49
20	CFTR-RD	allele 1	exon 11	c.1523T>G	F508C	92	52,17
21	CFTR-RD	allele 1	exon 11	c.1523T>G	F508C	49	63,27

		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	59	32,2
22	CF	allele 1	exon 11	c.1521_1523delCTT	[delta]F508	81	43,21
		allele 2	intron 18	c.2988+1G>A	3120+ 1G- >A	29	37,93
23	CF	allele 1	exon 22	c.3472C>T	R1158X	69	43,48
		allele 2	exon 8	c.1040G>C	R347P	85	54,12
24	CF	allele 1	exon 14	c.2128A>T	K710X	75	68
		allele 2a	exon 13	c.1684G>A	V562I (in cis)	60	55
		allele 2b	exon 19	c.3017C>A	A1006E (in cis)	116	39,66
		allele 2c	intron 9	c.1210-34TG(11)T(5)	5T-11TG (in cis)	.	.
25	CF	allele 1	exon 11	c.1477_1478delCA:p.493_493del	1609delCA	105	49,52
		allele 2a	exon 11	c.1521_1523delCTT	[delta]F508 (in cis)	95	33,68
		allele 2b	exon 19	c.3080T>C	I1027T (in cis)	156	47,44
26	CF	allele 1	exon 6	c.695T>A	V232D	107	44,86
		allele 2	intron 12	c.1679+1.6kbA>G	1811+ 1.6kbA- >G	35	57,14
27	CF CARRIER	allele 1	exon 4	c.387delT:p.L129fs	519delT	151	45,7
29	CF	allele 1	exon 17	c.2668C>T	Q890X	151	43,05
		allele 2	intron 5	c.580-1G>T	712- 1G- >T	154	51,3
30	CF	allele 1	exon 20	c.3302T>A	M1101K	120	43,33
		allele 2	intron 4	c.489+1G>T	621+ 1G- >T	148	34,46
31	CFTR-RD	allele 1	exon 14	c.2260G>A	V754M	235	49,79
32	CF	allele 1	intron 16	c.2657+5G>A	2789+5G>A	537	49,53
		allele 2a	exon 13	c.1684G>A	V562I (in cis)	186	57,53
		allele 2b	exon 19	c.3017C>A	A1006E (in cis)	270	50
		allele 2c	intron 9	c.1210-34TG(11)T(5)	5T-11TG (in cis)	.	.
33	CF	allele 1	exon 11	c.1521_1523delCTT	[delta]F508	149	39,6
		allele 2	exon 13	c.1763A>T	E588V	141	48,23
34	CFTR-RD	allele 1	exon 11	c.1521_1523delCTT	[delta]F508	202	38,61
		allele 2	exon 7	c.772A>G	R258G	145	55,86
35	CF CARRIER	allele 1	exon 12	c.1652G>A	G551D	146	49,32
38	CF	allele 1	exon 23	c.3731G>A	G1244E	246	95,93
		allele 2	exon 23	c.3731G>A	G1244E	246	95,93
39	CF CARRIER	allele 1	exon 12	c.1673T>C	L558S	156	52,56
40	CFTR-RD	allele 1	exon 17	c.2758G>A	V920M	416	46,15
		allele 2	intron 9	c.1210-34TG(13)T(5)	5T-13TG	.	.
41	CFTR-RD	allele 1	exon 11	c.1545_1546delTA	1677delTA	181	53,04
		allele 2a	exon 10	c.1327G>T	D443Y (in cis)	200	32,5
		allele 2b	exon 13	c.1727G>C	G576A (in cis)	214	50,47
		allele 2c	exon 14	c.2002C>T	R668C (in cis)	153	49,02
42	CF CARRIER	allele 1	intron 3	c.273+1G>A	405+ 1G- >A	127	50,39
43	CFTR-RD	allele 1	exon 14	c.1934T>A	M645K	132	43,94
44	CF CARRIER	allele 1	exon 8	c.1094T>C	L365P	207	50,24
45	CF	allele 1	exon 9	c.1196C>A	A399D	91	54,95

		allele 2	intron 5	c.580-1G>T	712- 1G- >T	182	56,04
46	CF	allele 1	exon 4	c.476T>C	L159S	198	53,54
		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	135	50,37
47	CF	allele 1	introns 22-23	c.3718-24_c.3873+601del781	CFTRdele20	.	.
		allele 2a	exon 19	c.3080T>C	I1027T (in cis)	207	56,52
		allele 2b	exon 11	c.1521_1523delCTT	[delta]F508 (in cis)	137	39,42
48	CF	allele 1	exon 4	c.293A>G	Q98R	145	51,03
		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	143	48,25
49	CFTR-RD	allele 1a	exon 15	c.2552G>T	R851L (in cis)	159	95,6
		allele 1b	exon 8	c.1052C>G	T351S (in cis)	187	98,4
		allele 2a	exon 15	c.2552G>T	R851L (in cis)	159	95,6
		allele 2b	exon 8	c.1052C>G	T351S (in cis)	187	98,4
50	CF	allele 1	exon 26	c.4143C>A	Y1381X	69	98,55
		allele 2	exon 26	c.4143C>A	Y1381X	69	98,55
51	CF	allele 1	exon 14	c.2017G>T	G673X	75	42,67
		allele 2	exon 20	c.3302T>A	M1101K	172	50,58
52	CF	allele 1	exon 17	c.2737_2738insG	2869insG	404	43,32
		allele 2	intron 19	c.3140-26A>G	3272- 26A- >G	199	51,26
53	CF CARRIER	allele 1	intron 22	c.3718-1G>A	3850- 1G- >A	254	42,91
54	CF	allele 1	exon 11	c.1521_1523delCTT	[delta]F508	132	40,91
		allele 2	intron 3	c.274-1G>A	406- 1G- >A	105	54,29
55	CFTR-RD	allele 1	exon 11	c.1516A>G	I506V (1648A/G)	157	51,59
		allele 2	exon 20	c.3275A>G	Y1092C	223	47,98
56	CF	allele 1	exon 12	c.1624G>T	G542X	134	50
		allele 2	exon 3	c.178G>A	E60K	99	56,57
57	CF CARRIER	allele 1	intron 22	c.3717+1G>A	3849+ 1G- >A	234	48,29
58	CF CARRIER	allele 1	promoter	c.-9_14del23	124del23bp	150	28
59	CF CARRIER	allele 1	exon 8	c.1076A>G	Q359R	163	40,49
60	CF	allele 1	exon 4	c.274G>A	E92K	150	54,67
		allele 2	exon 6	c.617T>G	L206W	245	53,88
61	CF CARRIER	allele 1	exon 7	c.805_806delAT	936delTA	164	39,02
		allele 2	.	.	.	.	.
62	CFTR-RD	allele 1a	exon 14	c.2260G>A	V754M (in cis)	225	50,22
		allele 1b	intron 2	c.164+2_164+3insT	296+ 3insT (in cis)	227	39,65
63	CF	allele 1	exon 1	c.44T>C	L15P	150	47,33
		allele 2	exon 14	c.2052_2053insA	2184insA	120	28,33
64	CF	allele 1	exon 13	c.1682C>A	A561E	220	53,64
		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	266	40,23
65	CF CARRIER	allele 1	exon 12	c.1657C>T	R553X	191	49,21
66	CF	allele 1	exon 20	c.3266G>A	W1089X	296	48,99
		allele 2a	exon 14	c.1882G>C	G628R (in cis)	72	50
		allele 2b	exon 22	c.3705T>G	S1235R (in cis)	271	47,97

67	CF	allele 1	exon 9	c.1209G>A	1341G- >A	165	43,03
		allele 2	exon 3	c.254G>A	G85E	148	44,59
68	CFTR-RD	allele 1	exon 19	c.3041A>G	Y1014C	239	39,33
69	CF	allele 1	introns 1-3	c.54-5940_273+10250del21kb	CFTRdele2,3	.	.
		allele 2a	exon 19	c.3080T>C	I1027T (in cis)	292	51,71
		allele 2b	exon 11	c.1521 _1523delCTT	[delta]F508 (in cis)	205	48,29
70	CF CARRIER	allele 1	Prom - intron 3	c.1-6186_273+507dup35741	CFTRdupProm-3	.	.
71	CF	allele 1	intron 5	c.579+3A>T	711+3A>T	50	64
		allele 2a	exon 19	c.3080T>C	I1027T (in cis)	122	36,07
		allele 2b	exon 11	c.1521_1523delCTT	[delta]F508 (in cis)	94	41,49
72	CF	allele 1	exon 11	c.1477_1478delCA	1609delCA	167	34,13
		allele 2	intron 6	c.743+1G>A	875+ 1G- >A	135	47,41
73	CF	allele 1	exon 17	c.2668C>T	Q890X	395	37,47
		allele 2	exon 19	c.3083T>G	M1028R	333	50,75
74	CF	allele 1	exon 3	c.262_263delTT	394delTT	41	29,27
		allele 2	intron 13	c.1766+3A>C	1898+3A>C	50	52
75	CF	allele 1	exon 15	c.2551C>T	R851X	169	48,52
		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	176	38,07
76	CFTR-RD	allele 1	intron 16	c.2657+5G>A	2789+5G>A	539	50,46
		allele 2a	exon 18	c.2930C>T	S977F (in cis)	112	43,75
		allele 2b	intron 9	c.1210-34TG(12)T(5)	5T-12TG (in cis)	.	.
77	CF	allele 1	exon 14	c.2125C>T	R709X	136	50,74
		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	266	44,74
78	CFTR-RD	allele 1	exons 25-27	c.3964-3890_Stop+3143del9454 insTAACT	CFTRdele22-24	.	.
		allele 2	intron 9	c.1210-34TG(12)T(5)	5T-12TG	.	.
79	CF	allele 1	exon 23	c.3731G>A	G1244E	378	41,01
		allele 2	intron 7	c.870-2A>G	1002- 2A>G	209	54,07
80	CF	allele 1	exon 12	c.1647T>G	S549R	226	47,79
		allele 2	.	.	.	.	.
81	CFTR-RD	allele 1	exon 19	c.2991G>C	L997F	287	43,21
		allele 2a	exon 23	c.3846G>A	W1282X (in cis)	209	43,06
		allele 2b	exon 3	c.221G>A	R74Q (in cis)	157	49,68
82	CF	allele 1	exon 14	c.2051_2052delAAinsG	2183AA- >G	109	50,46
		allele 2	exon 1	c.4C>T	Q2X	265	47,17
83	CF CARRIER	allele 1	exons 4-8, 12-21	c.[274-1091_3468+236inv85141ins38; 274-1044_1116+111del10602insTATAT; 1585-11392_3468+219del385	CFTR50kbdel	.	.

				85]			
84	CF	allele 1	intron 18	c.2989-1G>A	3121- 1G- >A	323	97,52
		allele 2	intron 18	c.2989-1G>A	3121- 1G- >A	323	97,52
85	CF	allele 1	exon 3	c.254G>T	G85V	165	42,42
		allele 2a	intron 9	c.1210-34TG(12)T(5)	5T-12TG (in cis)	.	.
		allele 2b	intron 15	c.2619+3A>G	2751+3A>G (in cis)	236	41,95
86	CF	allele 1	exon 8	c.1000C>T	R334W	400	49,25
		allele 2	intron 12	c.1680-1G>A	1812- 1G- >A	240	59,58
87	CF	allele 1	exon 14	c.1826A>G	H609R	60	53,33
		allele 2a	exon 23	c.3808G>A	D1270N (in cis)	238	47,9
		allele 2b	exon 3	c.220C>T	R74W (in cis)	132	46,21
		allele 2c	exon 6	c.601G>A	V201M (in cis)	216	44,91
88	CF CARRIER	allele 1	exons 19-21	c.2989-2634_3468+1508del8600	CFTRdele17a-18	.	.
89	CF CARRIER	allele 1	exon 20	c.3299A>C	Q1100P	349	47,56
90	CFTR-RD	allele 1	exon 3	c.224G>A	R75Q	170	49,41
91	CF	allele 1	exon 6	c.613C>T	P205S	294	52,04
		allele 2	exon 11	c.1521_1523delCTT	[delta]F508	249	35,34
92	CF	allele 1	exon 6	c.595C>T	H199Y	165	55,76
		allele 2	intron 19	c.3140-26A>G	3272- 26A- >G	127	51,97
93	CF	allele 1	exon 20	c.3274T>C	Y1092H	361	43,49
		allele 2	exon 22	c.3469-420_3717+1230del1899	.	.	.
94	CFTR-RD	allele 1	exon 11	c.1584G>A	1716G/A	159	50,94
95	CF CARRIER	allele 1	exon 14	c.2215delG:p.V739fs	2347delG	331	46,22
96	CFTR-RD	allele 1	exon 8	c.948delT	1078delT	266	49,62
		allele 2a	exon 20	c.3222T>A	F1074L (in cis)	230	46,52
		allele 2b	intron 9	c.1210-34TG(12)T(5)	5T-12TG (in cis)	.	.

Table S4. CFTR intron 9 c.1210-34TG(11-13)T(5-9) Haplotypes

Sample	T-TG Haplotype	Haplotype Counts
1FQ	T7-TG10/T9-TG10	33/23
1FQ_bis	T7-TG10/T9-TG10	10/9
2FQ	T9-TG10	59
2FQ_bis	T9-TG10	20
3FQ	T7-TG10/T5-TG12	29/28
3FQ_bis	T5-TG12/T7-TG10	10/9
4FQ	T5-TG12/T7-TG10	27/20
4FQ_bis	T5-TG12/T7-TG10	12/6
5FQ	T7-TG11/T9-TG10	31/27
6FQ	T7-TG11/T9-TG10	27/25
7FQ	T7-TG11/T7-TG10	50/7
8FQ	T7-TG11/T7-TG10	31/2
9FQ	T7-TG10/T7-TG11	27/21
10FQ	T7-TG10/T7-TG12	31/26
11FQ	T7-TG10/T7-TG11	34/22
12FQ	T7-TG11/T7-TG10	23/2
14FQ	T7-TG11/T7-TG10	39/2
15FQ	T7-TG12/T5-TG13	25/19
16FQ	T7-TG12/T7-TG11	28/27
17FQ	T7-TG11/T7-TG10	59/2
18FQ	T7-TG11	9
19FQ	T7-TG10	23
20FQ	T7-TG11/T7-TG10	8/1
21FQ	T9-TG10/T7-TG12	5/4
22FQ	T7-TG11/T9-TG10	6/5
23FQ	T7-TG10/T7-TG11	4/1
24FQ	T5-TG11/T7-TG11	7/2
25FQ	T9-TG10/T7-TG11	10/7
26FQ	T7-TG11/T9-TG10	7/6
27FQ	T7-TG11/T7-TG12	6/5
29FQ	T7-TG10/T9-TG10	6/6
30FQ	T9-TG10/T7-TG11	8/4
31FQ	T9-TG10/T7-TG11	12/11
32FQ	T5-TG11/T7-TG10	19/15
33FQ	T9-TG10/T7-TG11	33/1
34FQ	T7-TG11/T9-TG10	17/15
35FQ	T7-TG10/T9-TG10	12/10
38FQ	T7-TG11/T7-TG10	26/4
39FQ	T7-TG10/T7-TG11	23/16
40FQ	T5-TG13/T7-TG11	18/18
41FQ	T7-TG11/T7-TG10	21/14
42FQ	T7-TG11/T7-TG10	32/2
43FQ	T7-TG10/T7-TG11	30/16

44FQ	T9-TG10/T7-TG11	19/18
45FQ	T7-TG10/T9-TG10	18/1
46FQ	T7-TG12/T9-TG10	11/5
47FQ	T7-TG11/T9-TG10	12/12
48FQ	T9-TG10/T7-TG11	13/6
49FQ	T7-TG11/T9-TG10	19/2
50FQ	T7-TG10	21
51FQ	T7-TG11/T7-TG10	12/6
52FQ	T7-TG10/T7-TG11	17/14
53FQ	T7-TG10/T9-TG10	24/13
54FQ	T7-TG10/T9-TG10	17/7
55FQ	T7-TG10/T7-TG11	13/12
56FQ	T9-TG10/T7-TG11	11/9
57FQ	T9-TG10/T7-TG11	11/10
58FQ	T7-TG11	23
59FQ	T7-TG11/T7-TG10	25/3
60FQ	T7-TG11/T7-TG10	6/1
61FQ	T7-TG11/T7-TG11	12/4
62FQ	T9-TG10/T7-TG11	17/12
63FQ	T7-TG11/T7-TG10	17/1
64FQ	T7-TG11/T9-TG10	22/22
65FQ	T7-TG11/T7-TG12	14/14
66FQ	T7-TG11/T7-TG12	12/6
67FQ	T7-TG11/T7-TG10	35/1
68FQ	T7-TG11/T7-TG12	12/3
69FQ	T9-TG10/T7-TG11	19/8
70FQ	T7-TG11	17
71FQ	T7-TG10/T9-TG10	3/3
72FQ	T7-TG11	19
73FQ	T7-TG10/T7-TG11	16/14
74FQ	T7-TG11/T9-TG10	8/3
75FQ	T9-TG10	16
76FQ	T7-TG10/T5-TG12	16/11
77FQ	T7-TG10/T9-TG10	17/3
78FQ	T5-TG12/T7-TG11	13/6
79FQ	T7-TG12/T7-TG11	24/21
80FQ	T7-TG11/T7-TG10	36/1
81FQ	T7-TG10/T9-TG10	19/15
82FQ	T7-TG10/T7-TG11	12/8
83FQ	T7-TG11/T9-TG10	17/12
84FQ	T7-TG11/T5-TG12	26/1
85FQ	T7-TG10/T5-TG12	19/15
86FQ	T7-TG11/T7-TG10	26/2
87FQ	T7-TG11/T7-TG10	18/2
88FQ	T7-TG11/T9-TG10	16/16

89FQ	T7-TG11/T9-TG10	20/1
90FQ	T9-TG10/T7-TG11	15/11
91FQ	T9-TG10/T7-TG11	12/8
92FQ	T9-TG10/T7-TG10	9/8
93FQ	T7-TG12/T7-TG11	19/18
94FQ	T7-TG11/T7-TG10	12/9
95FQ	T7-TG10/T7-TG11	14/10
96FQ	T7-TG11/T5-TG11	16/10

REFERENCES

1 Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res 2001;11(10):1725-9.

2 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754-60.

3 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078-9.

4 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297-303.

5 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43(5):491-8.

6 Amstutz U, Andrey-Zurcher G, Suciu D, Jaggi R, Haberle J, Largiader CR. Sequence capture and next-generation resequencing of multiple tagged nucleic acid samples for mutation screening of urea cycle disorders. Clin Chem 2011;57(1):102-11.

7 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15(8):1034-50.

8 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res 2002;12(6):996-1006.

9 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4(7):1073-81.

10 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 2010;7(4):248-9.

11 Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005;15(7):901-13.

12 Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010;7(8):575-6.

13 Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;25(21):2865-71.

14 Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Quinlan AR, Nickerson DA, Eichler EE. Copy number variation detection and genotyping from exome sequence data. Genome Res 2012;22(8):1525-32.

15 Morral N, Nunes V, Casals T, Cobos N, Asensio O, Dapena J, Estivill X. Uniparental inheritance of microsatellite alleles of the cystic fibrosis gene (CFTR): identification of a 50 kilobase deletion. Hum Mol Genet 1993;2(6):677-81.