Development of a Clinical Next Generation Sequencing Panel for Diagnosis of Cystic Lung Diseases: Evaluation of Diagnostic yield and Optimization of Bioinformatics pipelines

Joowon Oh

Department of Medicine

The Graduate School, Yonsei University

Development of a Clinical Next Generation Sequencing Panel for Diagnosis of Cystic Lung Diseases: Evaluation of Diagnostic yield and Optimization of Bioinformatics pipelines

Directed by Professor Kyung-A Lee

The Doctoral Dissertation
submitted to the Department of Medicine
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy of Medical Science

Joowon Oh

December 2019

This certifies that the Doctoral Dissertation of Joowon Oh is approved.

Thesis Supervisor : Kyung-A Lee

Thesis Committee Member#1 : Seung-Tae Lee

Thesis Committee Member#2 : Hyo Sup Shim

Thesis Committee Member#3: Sangwoo Kim

Thesis Committee Member#4: Saeam Shin

The Graduate School
Yonsei University

December 2019

ACKNOWLEDGEMENTS

First, and most of all, I would like to express my very profound gratitude to my supervisor Prof. Kyung-A Lee for her expertise, guidance, encouragement and patience throughout my years of study and the process of writing this thesis. Without her guidance, this paper and all the related researches would not have been possible.

I would also like to acknowledge the rest of my thesis committee: Prof. Seung-Tae Lee, Prof. Hyo Sup Shim, Prof. Sangwoo Kim and Prof. Saeam Shin, for their insightful comments and supervision.

I also express my special thanks to my dearest colleagues, Dr. Yoonjung Kim and Dr. Dokyun Kim, for sharing their knowledge and sincere encouragement and the gifted laboratory bioinformatician, Seoung Chul Oh for his tremendous technical support.

Last but not least, I would like to give praise to the Lord for my parents, my sister, my dear husband, Sejune Oh, my lovely daughter, Seowon and my adorable son, Jeongmin. None of this would have been possible without their undying support, love and patience. To my darling husband, thank you for weathering the storms of life with me and for being my inspiration and my rock.

Joowon Oh

December, 2019

<TABLE OF CONTENTS>

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Development of a Clinical Next Generation Sequencing Panel for Diagnosis of Cystic Lung Diseases: Evaluation of Diagnostic yield and Optimization of Bioinformatics pipelines

Joowon Oh

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Kyung-A Lee)

Multiple cystic lung disease (CLD) represents a diverse group of uncommon disorders that can present a diagnostic challenge due to the increasing number of diseases associated with this presentation. Among CLD, several diseases have well-defined causative mutations in the relevant genes; e.g. Birt-Hogg-Dube syndrome (BHD), tuberous sclerosis complex (TSC) and cystic fibrosis (CF). Thus far, the molecular diagnosis of CLD is mainly based on Sanger sequencing. As Sanger sequencing of all the candidate genes substantially increase the cost, genetic testing usually starts with the most commonly involved genes and proceeds to less likely genes only when clinical suspicion is very high. In recent years, targeted next-generation sequencing (NGS) platform has been further developed, allowing us to focus specifically on genomic regions of interest for cheaper multiplexed sequencing of more cases. However, inaccuracy in detecting the length of homopolymers repeats and complexity in detecting structural variation became a critical barrier against accurate detection of genomic variations. Herein, we seek to establish an optimal bioinformatics pipeline for processing the NGS data. Through detemination of the optimal parameter settings for detecting mutations in error suspceptible region, we tried to increase the overall diagnostic sensitivity. Also,

we evaluated diagnostic yield and validate the analytical performance of the CLD panel.

62 patients with multiple lung cysts was enrolled. Mutations in *FLCN* gene was characterized through Sanger sequencing, Multiplex Ligation-dependent Probe Amplification (MLPA) and quantitative PCR (qPCR). Evaluation of diagnostic yield of the CLD NGS panel was done using Ion torrent S5 NGS platform. Genomic DNA reference materials obtained from Coriell cell repository and results of Sanger sequencing-confirmed mutations in *FLCN* was used to validate analytical performance of the CLD panel. Three bioinformatics(BI) pipeline for processing NGS data were used; NextGENe v.2.4.2.2 (Softgenetics, PA, USA), Ion Reporter Software 5.10 (Thermo Fisher Scientific, Waltham, MA, USA), Biomedical Genomics Workbench 5.0 (QIAGEN bioinformatics, CA, USA).

Optimization of each BI pipeline increased analytical sensitivities from 98.6% to 98.8% for Ion Reporter, from 99.0% to 100.0% for NextGENe and from 99.8% to 100.0% for Workbench. Overall diagnostic yield using NGS went up from 38.7% to 40.3% compared with *FLCN* Sanger sequencing alone. Even though the pathogenic hotspot of *FLCN* is mostly filtered out using Ion Reporter BI pipeline due to 8 homopolymers repeat sequences, adjusting BI can dramatically improve the overall performance. Optimization of the BI pipeline is essential when designing difficult NGS panel.

Diesease-associated variants anlaysis and haplotype construction of *CFTR* gene showed CLD as a newly addressed phenotype of *CFTR* mutation carriers. When compared with allele frequency in normal control, NM_000492.3:c.374T>C variant classified as VUS according to ACMG guideline, showed Odds ratio of 3.90-5.62 with statistically significant difference. Among 14 haplotypes constructed, p.M470V in combination with p.Q1352H on the backgrounds of wild type of the rest of genetic loci showed

borderline significant difference between the control group and CLD patients from this study (p value=0.0542).

We developed CLD NGS panel and optimized variant calling process in BI pipelines. Adjusting BI improved the overall performances. Diagnostic yield using NGS went up from 38.7% to 40.3% compared with *FLCN* Sanger sequencing alone. Lastly, clinical usefulness of NGS panel is not limited to detecting the pathogenic variants. With the abundant genetic information, clinical laboratory can inform the clinicians about the disease-associated variants, suggest haplotype if needed, and provide informations of low AF variants with possible germline mosaicisim.

Key words: cystic lung disease; Birt-Hogg-Dubé syndrome; next generation sequencing; bioinformatic

Development of a Clinical Next Generation Sequencing Panel for Diagnosis of Cystic Lung Diseases: Evaluation of Diagnostic yield and Optimization of Bioinformatics pipelines

Joowon Oh

Department of Medicine
The Graduate School, Yonsei University

(Directed by Professor Kyung-A Lee)

I. INTRODUCTION

Multiple cystic lung disease (CLD) represents a diverse group of uncommon disorders that can present a diagnostic challenge due to the increasing number of diseases associated with this presentation such as Lymphangioleiomyomatosis (LAM), pulmonary Langerhans cell histiocytosis (PLCH), Birt– Hogg–Dubé syndrome (BHD), lymphocytic interstitial pneumonia (LIP), Pneumocystis jiroveci pneumonia (PJP), pulmonary amyloidosis, light chain deposition disease (LCDD) and Cystic Fibrosis (CF)[1]. Careful review and characterization of the radiographic abnormalities, coupled with assessment of clinical and laboratory features that may point to an underlying pulmonary or systemic disease, are helpful in distinguishing among numerous possibilities to arrive at the correct diagnosis[2]. Algorithmic radiologic approach according to a classification of cysts is widely used in the current field[3]. Although multidisciplinary approach for differential diagnosis of CLD is helpful, correct diagnosis is often obscured by atypical symptoms and environmental factors. This challenging diagnosis can be overcome by molecular testing in several diseases which have well-defined causative

mutations in the relevant genes; e.g. BHD[4], tuberous sclerosis complex (TSC)[5] and CF[6].

Thus far, the molecular diagnosis of CLD is mainly based on Sanger sequencing. As Sanger sequencing of all the candidate genes substantially increase the cost, genetic testing usually starts with the most commonly involved genes and proceeds to less likely genes only when clinical suspicion is high. In this aspect, differential diagnosis based on Sanger sequencing may increase cost and the time involved, where the need for implementing larger gene panels is increasing. In recent years, targeted next-generation sequencing (NGS) platform has been further developed, allowing us to focus specifically on genomic regions of interest for cheaper multiplexed sequencing of more cases. Most NGS applications focus on the detection of single nucleotide variants (SNVs) or small insertions/deletions (indels) and reported to have great sensitivity/specificity detecting these variants[7]. However, inaccuracy in detecting the length of homopolymers repeats[8 9] and complexity in detecting structural variation became a critical barrier against accurate detection of genomic variations.

There are different types of mutations in CLD-related genes[10], comprised of base pair substitutions, small indels, deletions, duplications and rarely copy number variant (CNV) (Table 1). Folliculin (*FLCN*) gene, disease-causing gene in BHD patients, has mutational hotspots in homopolymer region[11, 12]. For this specific reason, despite the fact that NGS can be a cost-effective and time-saving solution for molecular testing of genetically heterogenous CLD, setting up a CLD-associated NGS panel can be challenging.

To provide accurate results of clinical genetic testings in clinical laboratories, each should meet high standards in NGS process including the wet procedure and bioinformatics analysis. There are several guidelines

encompassing selecting targeted genes, choice of sequencer and sequencing methods, choice of data analysis tools, variant interpretation, dealing with interference such as homologous sequences and result confirmations[13-15]. The guidelines also specifies about test development, platform optimization and test validations. Recently, standards and guidelines for validating NGS bioinformatics (BI) pipelines has been published and stated consensus recommendation for NGS BI pipeline validation[16]. Abiding by the general principles of the guidelines above, clinical laboratories can implement high-quality BI flows for ideal patient care. Our goal in this study is to provide practical guidelines for setting up optimal parameters of multiple variant callers for detecting the length of homopolymers repeats using *FLCN* mutation positive samples. Additionally, we seek to evaluate diagnostic yield of customized CLD panel using NGS platform.

Table 1. Types of mutations in 9 genes of CLD NGS panel

Gene		***EFEMP2***	***ELN***	***FBLN5***	***FLCN***	***LTBP4***	***SERPINA1***	***TSC1***	***TSC2***	***CFTR***
Mutation type	Missense/nonsense	11	22	20	40	9	53	105	382	997
	Splicing substitutions	1	14	0	23	1	3	38	141	230
	Regulatory substitutions	0	0	0	0	0	3	0	0	25
	Small deletions	2	29	0	60	3	15	106	234	253
	Small insertions/duplications	1	12	0	16	2	4	47	130	93
	Small indels	0	1	0	7	1	0	5	11	27
	Gross deletions	0	26	0	22	0	3	20	143	86
	Gross insertions/duplications	0	1	1	2	0	0	2	11	20
	Complex rearrangements	0	3	0	1	0	2	3	13	33
	Repeat variations	0	0	0	0	0	0	0	0	16
	TOTAL	15	108	21	171	16	83	326	1065	1780

Abbreviations: *CFTR*, cystic fibrosis transmembrane conductance regulator; *EFEMP2*, egf-containing fibulin-like extracellular matrix protein 2; *ELN*, elastin; *FBLN5*, fibulin 5; *FLCN*, folliculin; *LTBP4*, latent transforming growth factor-beta-binding protein 4; *SERPINA1*, serpin peptidase inhibitor, clade a, member 1; *TSC1*, tsc1 gene; *TSC2*, tsc2 gene

II. MATERIALS AND METHODS

1. Study design

We enrolled sixty two patients with bilaterally located multiple basal lung cysts, who were suspected with BHD. They were subjected to Sanger sequencing for *FLCN* variants by clinicians. This study was authorized by Institutional Review Board (IRB) of Gangnam Severance Hospital. Informed consent was obtained from each patient. We also obtained genomic DNA reference materials from Coriell cell repository (https://www.coriell.org/). Sequence information of NA12878 was downloaded from GeT-RM Browser (https://www.ncbi.nlm.nih.gov/variation/tools/get-rm/). The manufacturer-identified variants of reference materials are listed in Table 2. Total of 69 samples from patients and manufacturer was divided into two groups for analysis. For initial evaluation and BI optimization, seven reference materials and thirteen patients were grouped as optimization group. The rest forty-nine patients were grouped as validation group.

Table 2. Manufacture indicated variant list of commercial reference material

Reference material	Variant description	ACMG classification
NA04330	NM_000492.3(*CFTR*):c.1680-1G>A	pathogenic
	NM_000492.3(*CFTR*):c.313delA (p.Ile105Serfs)	pathogenic
NA08299	NM_000548.4(*TSC2*):c.2468dupT (p.Pro824Alafs)	Likely pathogenic
NA09374	NM_000548.4(*TSC2*):c.3693_3696delGTCT (p.Ser1232Thrfs)	pathogenic
NA07421	NM_000368.4(*TSC1*):c.994_995insA, p.(Ser332Tyrfs*9)	Likely pathogenic
NA07830	NM_000492.3(*CFTR*):c.429delT, p.(Phe143Leufs*10)	pathogenic
	NM_000492.3(*CFTR*):c.1521_1523delCTT (p.Phe508delPhe)	Likely pathogenic
NA18668	NM_000492.3(*CFTR*):c.1521_1523delCTT (p.Phe508delPhe)	Likely pathogenic

Abbrevation ACMG, American college of medical genetics and genomics

2. Sample preparation and Sequencing with Ion S5 XL and NextSeq 550 system

Reference materials were purchased in the form of genomic DNA. Genomic DNA was extracted from patient's EDTA blood sample using a QIAamp® DNA Blood Mini Kit (Qiagen, Venlo, the Netherlands). We used NanoDrop® 1000 system (Thermo Fisher Scientific, Waltham, MA, U S A) for assessment of DNA purity of each sample. To measure DNA concentration, we used a Qubit® 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). All sixty nine samples went through sequencing with Ion S5 XL (Thermo Fisher Scientific, MA, USA), and sixty two patient samples were also sequenced with NextSeq 550 system (Illumina, San Diego, CA, U S A). For NGS testing on S5 XL, library preparation was carried out according to the manufacturer's instructions. Libraries for S5 XL were constructed using customized CLD panel. The panel was designed to cover all coding exons and flanking introns of targeted nine genes. We evaluated the quality of final libraries using the 4200 TapeStation (Agilent Technologies, Inc, Santa Clara, CA, U S A). The prepared libraries were then sequenced on Ion S5 XL Sequencer using Ion 530 Chip and Ion 530 kit–Chef Kit (Thermo Fisher Scientific, Waltham, MA, USA). For NGS testing on NextSeq 550 system, at least 500 ng of genomic DNA was fragmented into segments approximately 150 bp long. After DNA end-repair and adapter ligation, libraries were hybridized with the capture probes (Celemics, Seoul, Korea). Enriched DNA was then amplified, and clusters were generated and sequenced on NextSeq 550 system.

3. Sequence alignment, variant calling and annotation

Sequence alignment was done differently according to NGS platforms; i) for S5 XL data, reads alignment and mapping to human genomic reference sequences (GRCh37) were performed using the Torrent Mapping Alignment

Program aligner, ii) for NextSeq data, reads were aligned to human genomic reference sequences (GRCh37) using the Burrows-Wheeler Alignment tool (0.7.17). For variant calling, NextGENe v.2.4.2.2 (Softgenetics, PA, USA), Ion Reporter Software 5.10 (Thermo Fisher Scientific, Waltham, MA, USA), Biomedical Genomics Workbench 5.0 (QIAGEN bioinformatics, CA, USA) and HaplotypeCaller in the Genome Analysis Toolkit package (4.1.2.0) were used. Called variants were annotated using ANNOVAR (http://www.openbioinformatics.org/annovar/) and Alamut® v.2.10 (Interactive Biosoftware, Rouen, France). The reference transcript of each targeted genes[17] are listed in Table 3. All identified variants were classified into 5 categories; pathogenic, likely pathogenic, uncertain significance, likely benign and benign. We used criteria for classifying variants according to American college of medical genetics and genomics (ACMG) guidelines[18].

Table 3. Reference cDNA transcript of targeted genes in CLD NGS panel

Gene name	MIM number	Chromosomal Location	Reference cDNA	Related phenotype of lung	Inheritence pattern
FLCN	*607273	17p11.2	NM_144997.5	Birt-Hogg-Dube syndrome	AD
TSC1	*605284	9q34.13	NM_000368.4	Lymphangioleiomyomatosis	AD
TSC2	*191092	16p13.3	NM_000548.3	Lymphangioleiomyomatosis, somatic	AD
CFTR	*602421	7q31.2	NM_000492.3	Congenital bilateral absence of vas deferens	AR
EFEMP2	*604633	11q13.1	NM_016938.4	Cutis laxa, autosomal recessive, type IB	AR
ELN	*130160	7q11.23	NM_001278939.1	Cutis laxa, AD	AD
FBLN5	*604580	14q32.12	NM_006329.3	Cutis laxa, autosomal dominant 2	AR / AD
LTBP4	*604710	19q13.2	ENST00000308370.7	Cutis laxa, autosomal recessive, type IC	AR
SERPINA1	*107400	14q32.13	NM_000295.4	Emphysema due to AAT deficiency	AR

Abbreviations: AAT, Alpha-1 Antitrypsin ; AD, autosomal dominant; AR, autosomal recessive; MIM, Mendelian inheritance in Man

4. Parameter adjustment for variant calling

To achieve high sensitivity and specificity, we tried to set optimal parameters in three different commercial variant callers and compared the results with default settings. For Biomedical Genomics Workbench 5.0, forward/reverse balance and average read quality score were adjusted to find optimal cut-off for variant filtering. For Ion Reporter software, homopolymeric lenghth limitation were adjusted and mutational hotspot bed file was additionally adapted. For NextGENe software, strand bias of homopolymeric indel was adjusted. To determine the cut-offs of the parameters describe above, we used ROC curve analysis and Youden index maximization.

5. Confirmation of detected aberration

All patients went through Sanger sequencing targeting all coding exons and adjacent introns of *FLCN* gene. The primer set of *FLCN* sequencing is listed in Table 4. Sanger sequencing was performed to verify pathogenic variants and variants of uncertain significance (VUS) identified with NGS sequencing. A multiplex ligation-dependent probe amplification (MLPA) assay was applied for all the patients' samples regardless of the results from CNV analysis using NextGENe software. MLPA was performed using P256-B4 *FLCN* probe mixes (MRC-Holland, Amsterdam, the Netherlands) according to the manufacturer's instructions. MLPA results were analyzed using GeneMarker software (Softgenetics, State College, PA, USA).

Table 4. Primer information for Sanger sequencing of *FLCN* gene

Target gene Exon number	Primer sequence	Amplicon size (bp)
FLCN Exon4 F	TCATGGAGTCAATAGGCATTGGCA	564
FLCN Exon4 R	TGCAGTGAGCCATGATCACACCAT	
FLCN Exon5 F	GTTACCTACTTCGTAAGTGCTCAGC	479
FLCN Exon5 R	CCTGTGCAATGCTGGCTCCGAGC	

FLCN Exon6 F	AGAGTACAGTCTTCGGCTCTCATGG	515
FLCN Exon6 R	ACAATTCACACAGTGCACTGGCTG	
FLCN Exon7 F	TCCAGGAGTCAGGTCCTGGAGTT	421
FLCN Exon7 R	CAGATCTGTGCTCACTGACAAGTG	
FLCN Exon8 F	GTTGACTTGTGGAACTGCCTGCAT	381
FLCN Exon8 R	CTCGTTCTGGGCTGATTCAGAGC	
FLCN Exon9 F	CCAGGAATCTACACTGACCGGCT	420
FLCN Exon9 R	GAGGCTGTCAGTCACTTCCTGCA	
FLCN Exon10 F	GCCTCCCTGAGAAGATAAGTGTCTT	519
FLCN Exon10 R	GGTGCACAGCGGTTCTGTGCT	
FLCN Exon11 F	TGGGTAGTAGAGCATGGATG	254
FLCN Exon11 R	TCTCCACAACCCATGACAGAGATCT	
FLCN Exon12-13 F	ACTGACCTGGGATGAGCGGAGT	592
FLCN Exon12-13 R	ACCTGAGCTTTGCAGTGGCGGA	
FLCN Exon14 F	GCTGGTGCCAAAGCCGTGTCA	404
FLCN Exon14 R	ACAGCTCCTTCCAGCAGTTGAGA	

Abbreviations: F, forward; bp, basepair; R, reverse

6. Data analysis

Evaluation of analytic sensitivity, analytic specificity, and accuracy was performed with the candidate variants in the 'region of interest' which spanned all protein-coding regions and intron-exon boundaries (± 20 bp). In the analytical performance analysis, 'positive' indicates the case where the variants were detected in at least two different variant callers from NGS data of NextSeq platform or confirmed by Sanger sequencing. 'Negative' means that the variant-free regions were confirmed as negative by Sanger sequencing of *FLCN* coding regions. All statistical analyses were performed using Analyse-it® v.3.90.7 (Analyse-it Software, Ltd. Leeds, United Kingdom), MedCalc Software (https://www.medcalc.org/) and R 3.5.3. P-values less than 0.05 were regarded as significant.

III. RESULTS

1. Analytical performance and BI optimization

A. Quality of NGS sequencing

Quality metrics of sixty nine specimen sequenced on S5 XL and those of sixty two samples sequenced on Nextseq 550 are described in Table 5. All the indices met quality criteria acceptable for laboratory strategy. The average gDNA concentration was 30.35 ng/ul and 29.7 ng/ul respectively. The number of mapped reads were 411,690 and 676,637 in average on each platforms. The average on-target reads was 96.7% on S5 and 62.4% on Nextseq 550, and uniformity of base coverage was 96.6% and 97.8% respectively. Depth of on-target regions was 1,801× with minimum depth of 1,138× on S5 and average of 1,119× with minimum depth of 657× on Nextseq 550 system.

Table 5. Quality Metrics in two NGS platform

	S5 XL (n = 69)	Nextseq (n=62)
Average of gDNA (Interquartile Range), ng/ul	30.35 (23.2-35.0)	29.7 (21.9 - 34.7)
Number of mapped reads (min, max)	411,690 (309,203 - 1,249,676)	676,637 (485,849 -927,615)
On-target reads, %	96.7	62.4
Uniformity of base coverage at 0.2, %	96.6	97.8
Average depth of on-target regions (min, max)	1,801× (1,138×, 4,578×)	1,119× (657×, 1,576×)
Target base coverage at 20×, %	100	99.4
Target base coverage at 100×, %	98.7	98.7
a coverage of 100% at 20x	71.0 (49/69)	50.0 (31/62)

B. Optimization by adjusting variant filtering parameters

For initial optimization of three BI pipelines described in the Method

section, 165 Sanger confirmed or manufacture provided known variants were used. Decision process of cutoff level of adjustable parameters using ROC curve and Youden index is visualized in Figure 1. All adjustments were made to achieve zero false negative calls. For Ion Reporter Software, 'homopolymer maximum length' was changed from 8bp to 12bp and extra hotspot regions were adapted for mendatory variant calling. For NextGENe software, filtering cutoff of 'balance ratio in homopolymeric indels' was adjusted from 0.8 to 0.4. Lastly, for Workbench software, after not adapting 'minimum central quality filter', cut off of 'forward/reverse balance' was newly adapted. Additionally to reduce false positive rate in Workbench software, 'variant allele frequency' more than 30% was applied. After adjusting thresholds for variant filtering parameters (Not all data shown), each BI pipelines showed improved performance. Detailed sensitivity, specificities are described in Table 6. In analyzing the optimizing group, Ion Reporter showed the highest sensitivity (99.4%, 96.2-100.0) and specificity (100.0%, 100.0-100.0) in default settings and the lowest false positive rates (0.6%, 0.0-3.8) in all three variant callers.

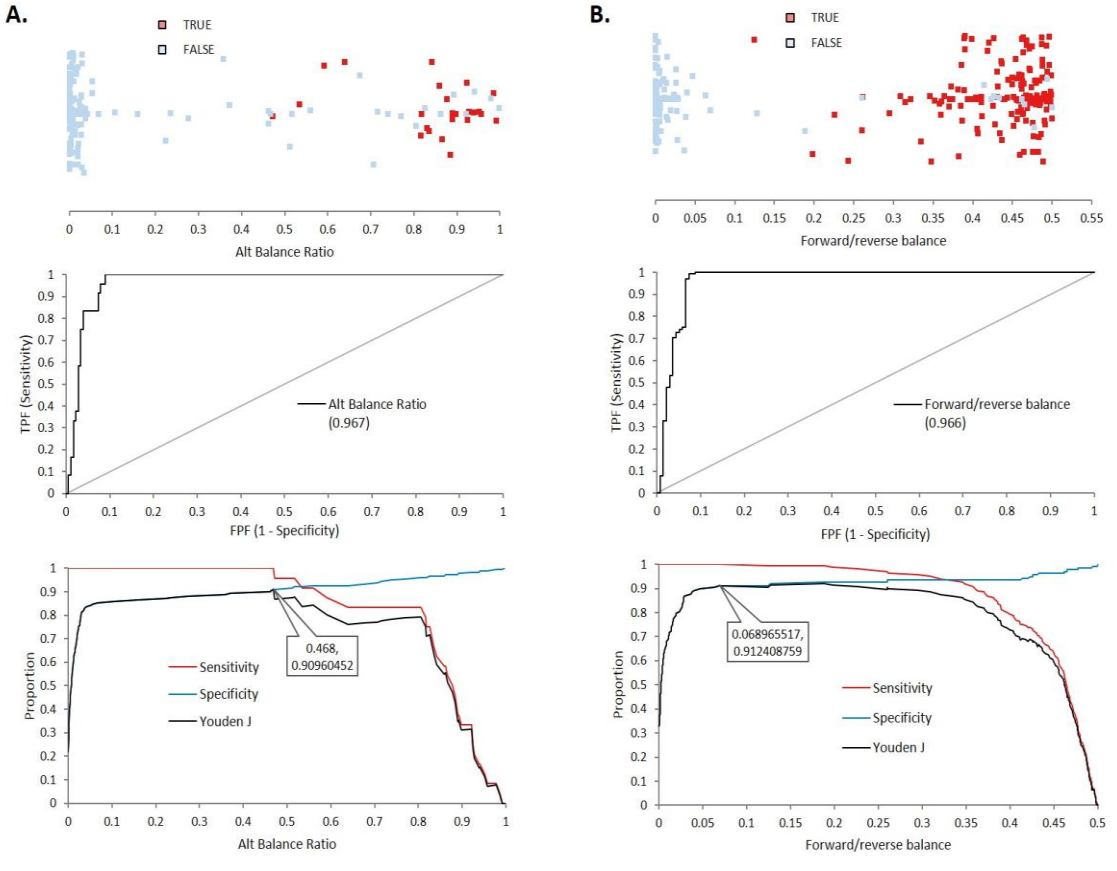

Figure 1. Examples of cutoff evaluation of forward/reverse balance ratio achieving maximum analytical performance. A. Applying 0.47 cutoff of 'reference and alteration read balance' in homopolymeric indel variants on NextGENe v.2.4.2.2 shows maximum analytical performance (Youden index 0.910, Sensitivity 100.0%, Specificity 91.0%). B. When filtered with 0.07 cutoff of 'forward/reverse balance' in overall variants on Biomedical Genomics Workbench 5.0, the highest analytical performance can be achieved (Youden index 0.912, Sensitivity 100.0%, Specificity 91.2%).

Table 6. Comparison of analytical perfomance adjusting parameters in three different variant callers

Variant caller	Parameter settings	FP	FN	TP	TN	Sensitivity, % (95% CI)	Specificity, % (95% CI)	FPR, % (95% CI)	FNR, % (95% CI)
Ion Reporter Software 5.10	Default: homopolymer maximum length: 8bp	1	1	164	28174	99.4% (96.2-100.0)	100.0% (100.0-100.0)	0.6% (0.0-3.8)	0.0% (0.0-0.0)
	Adjusted: homopolymer maximum length: 12bp, Hotspot regions adapted	1	0	165	28174	100.0% (97.2-100.0)	100.0% (100.0-100.0)	0.6% (0.0-3.8)	0.0% (0.0-0.0)
NextGENe v.2.4.2.2	Default: balance ratio in homopolymer indels≥0.8, VAF≥20%	53	2	163	28122	98.8% (95.2-99.8)	99.9% (99.8-99.9)	24.5% (19.1-30.9)	0.0% (0.0-0.0)
	Optimized: balance ratio in homopolymer indels≥0.4, VAF≥20%	61	0	165	28114	100.0% (97.2-100.0)	99.8% (99.7-99.8)	27.0% (21.4-33.4)	0.0% (0.0-0.0)
Biomedical Genomics Workbench 5.0	Default: minimum central quality ≥20, VAF≥20%	248	8	157	27927	95.2% (90.3-97.7)	99.1% (99.0-99.2)	61.2% (56.3-66.0)	0.0% (0.0-0.0)
	Optimized: quality filter not adjusted, forward/reverse balance ≥ 0.07, VAF ≥30%	137	0	165	28038	100.0% (97.2-100.0)	99.5% (99.4-99.6)	45.4% (39.7-51.2)	0.0% (0.0-0.0)

Abbreviations: FN, false negative; FNR, false negative rate; FP, false positive; FPR, false positive rate; TN, true negative; TP, true positive

C. Validation of the optimized BI pipeline and comparison between two NGS platform

With total of forty-nine CLD patients' samples, we validated the previously optimized BI pipelines. After adapting hotspot regions in Ion Reporter software, the variant caller failed to call six pathogenic hotspot variants. The sensitivity of adjusted Ion reporter was 98.8% (95% CI: 97.4-99.5). With NextGENe software, the optimized result showed 100.0% sensitivity (95% CI: 99.2-100.0), 99.7% specificity (95% CI: 99.6-99.7) and 38.8% false positive rate (95% CI: 35.7-42.0). Workbench software showed 100.0% sensitivity (95% CI: 99.2-100.0), 99.9% specificity (95% CI: 99.9-100.0) and 15.2% false positive rate (95% CI: 12.6-18.1). Analytical performances of three variant callers on NextSeq 550 system showed 99.8%/~100.0% of sensitivity with relatively low false positive rate than those of S5 XL based variant callings. The parameter settings and the absolute number of false positives, false negatives, true positives and true negatives are detailed in Table 7.

Table 7. Analytical performance of next-generation sequencing compared with Sanger sequencing (n=49)

Sequencer	Variant caller	Management detail in parameter settings or newly adapted strategies	FP	FN	TP	TN	Sensitivity, % (95% CI)	Specificity, % (95% CI)	FPR, % (95% CI)	FNR, % (95% CI)
S5 XL	Ion Reporter Software 5.10	Default: homopolymer maximum length: 8bp	13	8	574	106225	98.6% (97.2-99.4)	100.0% (100.0-100.0)	2.2% (1.2-3.9)	0.0% (0.0-0.0)
		Adjusted: homopolymer maximum length: 12bp, Hotspot regions adapted	12	7	575	106226	98.8% (97.4-99.5)	100.0% (100.0-100.0)	2.0% (1.1-3.6)	0.0% (0.0-0.0)
	NextGENe v.2.4.2.2	Default: balance ratio in homopolymer indels≥0.8, VAF≥20%	208	6	576	106030	99.0% (97.7-99.6)	99.8% (99.8-99.8)	26.5% (23.5-29.8)	0.0% (0.0-0.0)
		Optimized: balance ratio in homopolymer indels≥0.4, VAF≥20%	369	0	582	105869	100.0% (99.2-100.0)	99.7% (99.6-99.7)	38.8% (35.7-42.0)	0.0% (0.0-0.0)
	Biomedical Genomics Workbench 5.0	Default: minimum central quality ≥20, VAF≥20%	596	1	581	105642	99.8% (98.9-100.0)	99.4% (99.4-99.5)	50.6% (47.7-53.5)	0.0% (0.0-0.0)
		Optimized: quality filter not adjusted, forward/reverse balance ≥ 0.07, VAF ≥30%	104	0	582	106134	100.0% (99.2-100.0)	99.9% (99.9-100.0)	15.2% (12.6-18.1)	0.0% (0.0-0.0)
	GATK 4.1.2.0		51	0	582	106187	100.0% (99.2-100.0)	100.0% (99.9-100.0)	8.1% (6.1-10.5)	0.0% (0.0-0.0)
NextSeq 550	NextGENe v.2.4.2.2		89	0	582	106149	100.0% (99.2-100.0)	99.9% (99.9-100.0)	13.3% (10.8-16.1)	0.0% (0.0-0.0)
	Biomedical Genomics Workbench 5.0		1	0	582	106237	100.0% (99.2-100.0)	100.0% (100.0-100.0)	0.2% (0.0-1.1)	0.0% (0.0-0.0)

Abbreviation: FN, false negative; FNR, false negative rate; FP, false positive; FPR, false positive rate; GATK, Genome Analysis Toolkit package
TN, true negative; TP, true positive

2. Clinical performance of CLD NGS panel

A. Overall diagnostic yield of CLD NGS panel

Through initial Sanger sequencing of *FLCN* coding exon and adjacent introns, twenty four out of sixty two patients (38.7%) revealed pathogenic variants. Using CLD NGS panel, one CNV was newly detected and fifteen variants of uncertain significance (VUS) in *CFTR*, two VUS in *EFEMP2*, one VUS in *ELN*, eight VUS in *LTBP4*, two VUS in *SERPINA1*, three VUS in *TSC1* and one VUS in *TSC2* were detected. Diagnostic yield went up to 40% using CLD NGS panel (Figure 2). Patients' characteristics between BHD (*FLCN*-pathogenic-variant-positive patients) group and non-BHD group are detailed in Table 8. Clinical manifestation of spontaneous pneumothorax was increased in number in BHD group (p value=0.001). Other than that, no statistically significant difference was observed between the two groups.

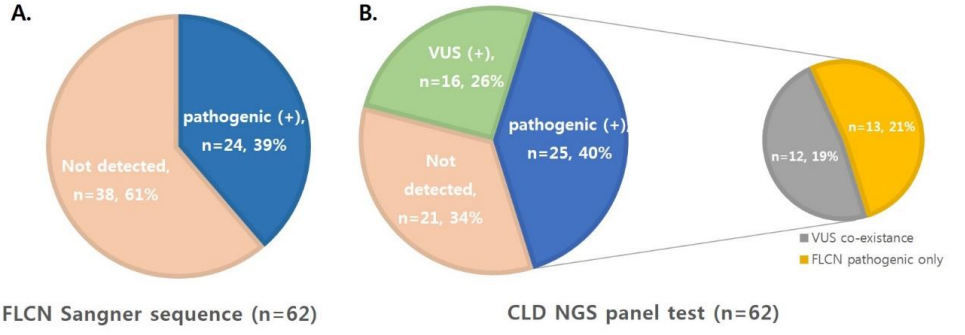

Figure 2. Diagnostic yields of *FLCN* sequencing and CLD NGS panel. (A) Through *FLCN* Sanger sequencing, 39% (24/62) of pathogenic mutations were detected. (B) Using CLD NGS panel test, 40% of pathogenic aberrations were revealed and 26% of the patients had extra VUS.

Table 8. Patient characteristics of the study population (n=62)

	All (n=62)	BHD (n=25)	non-BHD (n=37)	P value
Age, yr	50.6	48.3 ± 13.6	52.1 ± 17.3	0.36
Female sex	34 (54.8%)	15 (60.0%)	19 (51.4%)	0.681
Smoking history (n=40)	7 (17.5%)	1 (7.7%)	6 (22.2%)	0.491
Family history of CLD (n=33)	5 (15.2%)	2 (20.0%)	3 (13.0%)	1.000
Pneumothorax (n=40)	**17 (42.5%)**	**11 (84.6%)**	**6 (22.2%)**	**0.001**
Renal lesion (RCC or angiomyolipoma) (n=36)	2(5.6%)	1 (9.1%)	1 (4.0%)	1.000

Abbreviations: BHD, Birt– Hogg–Dubé syndrome; CLD, cystic lung disease; RCC, renal cell carcinoma
Bold values show statistical data with significant difference.

B. Haplotype construction of *CFTR* in Korean CLD patients

Out of sixty two patients in this group, one Russian patient showed difference in ethnicity. Identified SNVs in *CFTR* gene of sixty one patients were compared with the allele frequency in control group. Detailed odds ratio of allele frequencies of *CFTR* variants compared with those of normal population are listed in Table 9. With all the variants combined, the Odds ratio (OR) compared with gnomAD control group showed 2.31 with statistical significance. NM_000492.3:c.374T>C variant classified as VUS according to ACMG guideline[18], showed OR of 3.90-5.62 with statistically significant difference. NM_000492.3: c.3468G>T variant showed OR of 3.81 compared with gnomAD database (https://gnomad.broadinstitute.org/). However it showed no significant OR compared with allele frequency from Korean Reference Genome databse (http://coda.nih.go.kr/coda/KRGDB). We constructed *CFTR* haplotypes of sixty one Korean CLD patients using 10 polymorphisms and VUSs. For control group and other patients group with different phenotypes, we used data from the previous literature[19] and reconstructed the haplotypes according to our results. All haplotypes and the frequencies are described in Table 10. Among 14

haplotypes constructed, p.M470V in combination with p.Q1352H on the backgrounds of wild type of the rest of genetic loci showed significant difference between the control group and bronchiectasis patients (p value=0.0281) and pancreatitis patients (p value=0.0062) and showed borderline significant difference between the control group and CLD patients from this study (p value=0.0542).

Table 9. Frequency of *CFTR* gene variants in Korean patients with cystic lung disease compared with those of healthy Korean population

Genomic position	Reference transcript	Nucleotide change	Predicted amino acid change	Minor allele frequency (allele count/allele number)						
				Patients	Control 1[a]			Control 2[b]		
				(n=61)	Korean (n=1909)	Odds ratio (95% CI)	P value	KRGDB 622 individuals	Odds ratio (95% CI)	P value
CFTR variants combined				0.1311 (16/122)	0.0614 (234/3808)	**2.31 (1.34-4.00)**	**0.015**[c]	0.0667 (83/1244)	**2.11 (1.19-3.74)**	0.0618[c]
7:117171053	NM_000492.3	c.374T>C	p.Ile125Thr	0.0242 (3/122)	0.0045 (17/3808)	**5.62 (1.63-19.44)**	**0.0064**	0.0064 (8/1244)	**3.90 (1.02-14.88)**	**0.0468**
7:117304834	NM_000492.3	c.4056G>C	p.Gln1352His	0.0323 (5/122)	0.0176 (67/3806)	2.38 (0.94-6.03)	0.0661	0.0209 (26/1244)	2.00 (0.75-5.31)	0.1632
7:117254767	NM_000492.3	c.3468G>T	p.Leu1156Phe	0.0242 (3/122)	0.0066 (25/3802)	**3.81 (1.13-12.79)**	**0.0305**	0.0096 (12/1244)	2.59 (0.72-9.30)	0.1451
7:117307076	NM_000492.3	c.4357C>T	p.Arg1453Trp	0.0081 (1/122)	0.0037 (14/3760)	2.21 (0.29-16.95)	0.4451	0.0064 (8/1244)	1.28 (0.16-10.30)	0.8185
7:117175372	NM_000492.3	c.650A>G	p.Glu217Gly	0.0242 (3/122)	0.0273 (104/3816)	0.90 (0.28-2.88)	0.8587	0.0217 (27/1244)	1.14 (0.34-3.80)	0.8357
7:117144344	NM_000492.3	c.91C>T	p.Arg31Cys	0.0081 (1/122)	0.0018 (7/3796)	4.47 (0.55-36.65)	0.1627	0.0016 (2/1244)	5.13 (0.46-57.01)	0.1831

One Russian patient from this study were excluded from *CFTR* analysis. P value was calculated from the Fisher's exact test, comparison of values between patients group and each control group. P value lower than 0.05 was considered significant. Bold values show statistical data with significant difference.

a: gnomAD Exome database (https://gnomad.broadinstitute.org)

b: Korean Reference Genome Database (http://coda.nih.go.kr/coda/KRGDB)

c: Bonferroni correctiona was made.

Table 10. Haplotype assembly *CFTR* gene

Allele ID	-G>C	I125T	E217G	M470V	I556V	2562T>G	L1156F	Q1352H	4389G>A	R1453W	In this study		From previous literature (Lee *et al.*)		
											Cystic lung disease	P-value	Control	Bronchi-ectasis	Pancrea-titis
1	G	I	E	V	I	T	L	Q	G	R	56(45.9)	0.1161	123 (52.6)	52 (55.3)	25 (44.6)
2	G	I	E	M	I	G	L	Q	G	R	40(32.8)	0.6369	80 (34.2)	26 (27.7)	19 (33.9)
3	C	I	E	M	I	G	L	Q	G	R	7(5.7)	0.8017	11 (4.7)	3 (3.2)	6 (10.7)
4	G	I	E	V	I	T	L	H	G	R	4(3.3)	0.0542	1 (0.4)	4 (4.3)*	4 (7.1)*
5	G	I	E	V	V	T	L	Q	G	R	3(2.5)	1.0000	6 (2.6)	2 (2.1)	0 (0.0)
6	G	I	G	M	I	G	L	Q	G	R	2(1.6)	1.0000	3 (1.3)	4 (4.3)	1 (1.8)
7	G	T	E	V	I	T	L	Q	G	R	2(1.6)	0.2847	1 (0.4)	1 (1.1)	1 (1.8)
8	G	I	E	V	I	T	F	Q	G	R	2(1.6)	0.1237	0 (0.0)	0 (0.0)	0 (0.0)
9	G	I	E	V	I	G	L	Q	G	R	1(0.8)	0.6602	4 (1.7)	1 (1.1)	0 (0.0)
10	G	I	E	M	I	T	L	Q	A	R	1(0.8)	1.0000	2 (0.9)	0 (0.0)	0 (0.0)
11	G	I	E	M	I	G	L	Q	G	W	1(0.8)	1.0000	2 (0.9)	1 (1.1)	0 (0.0)
12	G	I	E	M	I	G	F	Q	A	R	1(0.8)	0.3526	0 (0.0)	0 (0.0)	0 (0.0)
13	G	T	E	V	I	T	L	H	G	R	1(0.8)	0.3526	0 (0.0)	0 (0.0)	0 (0.0)
14	C	I	G	M	I	G	L	Q	G	R	1(0.8)	0.3526	0 (0.0)	0 (0.0)	0 (0.0)
										Total	122		223[a]	94	56

Haplotypes were assembled using R package 'hapassoc' based on the Expectation-Maximization algorithm (Reference). Differences between control and disease groups were analyzed by the Fisher's exact test. * : P value <0.05

IV. DISCUSSION

1. Optimizing BI pipeline process for detecting error-prone pathogenic variants

Developing error-prone NGS panel with homopolymeric hotspot pathogenic mutations on semiconductor-based sequencing platform was challenging yet achievable. It is well known that semiconductor-based sequencing platforms suffer from the inaccuracy in detecting the length of homopolymers repeats of the same nucleotide[8,22] and many attempts are already made to improve the accuracy of alignment around homopolymeric regions[8,23]. Most of these papers are based on the data from Ion Personal Genome Machine (PGM™) or Ion Proton™, the earlier model of semiconductor-based NGS sequencer, and as it has been pointed out, the flow-call inaccuracy is systematic[22] and can be improved. Based on our experience, sequenced data from S5 XL have captured true homopolymeric indels judging from the visual inspection of the aligned BAM files on IGV software. Therefore, we focused on optimizing the variant calling process in the overall BI pipelines.

When optimizing the BI pipelines, we used three commercial variant callers. Schematic workflow for optimizing BI pipelines are described in Figure 3. In routine laboratory work, lowering false negative calls and achieving 100% sensitivity must come in the first priority to avoid patients from misdiagnosis, missing the opportunity of appropriate treatment or genetic counseling of the family members. With that in mind, one should start with preparing at least 20 samples[14] with known variants or at least 60 unique variants to provide a maximum sensitivity of 95% (within a CI of 95%), when all 60 variants are detected by the new technology for evaluation[16,24]. Then, after running the variant callers in default parameter settings[25], if there's any false negatives, get no-filtered vcf files. Next step is to make the list of true and false variants from the vcf files. Find the adjustable parameters and draw ROC curve and find ideal

cutoff of selected parameter to maximize sensitivity. In this study, depth or VAF were not considered as adjustable parameter because filtering with the cutoff of depth or VAF made high false positive rate than other parameters (data not shown). After achieving 100% sensitivity, Supplemental validation of a bioinformatics pipeline is required when components of the pipeline are modified[16].

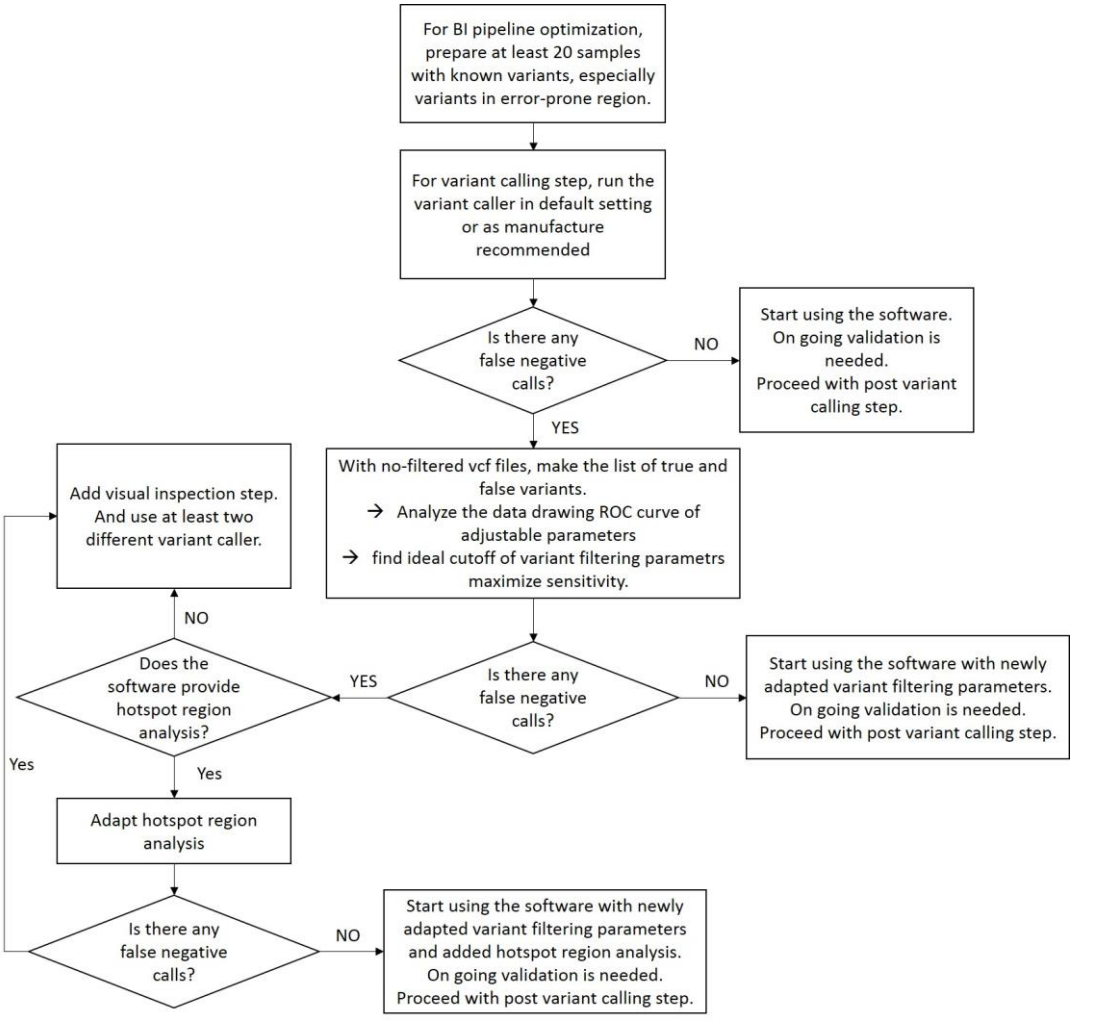

Figure 3. Schematic workflow of NGS optimization

2. Post variant-calling process

To achieve 100% sensitivity, we observed the increasing false positive calls (Table 6-7). To effectively sort out false positive calls, we built an intra-laboratory database to compare noise and true variants (Figure 4). In the annotation process, by showing the number of total samples analyzed, the cumulated incidence of calling the same variant in each variant caller and the minimum and maximum variant allele frequency (VAF), we could decide which variant should be confirmed by Sanger-sequencing. Accumulation of intra-laboratory database makes the interpretation process shorter and more precise as the NGS panel gets mature and stabilized through on going validation and variant confirmations.

A	B			C	D			E			
TPN	VN(IOT:NG	IOT_MinAF	IOT_MaxAF	NG_MinAF	NG_MaxAF	AF_IOT	DP_IOT	AF_NG	DP_NG	SB_NG	Zygosity
140	0:40			0.051	0.106			0.062	1637	0.656	?
140	0:87			0.057	0.25			0.071	1728	0.694	?
140	0:138			0.086	0.334			0.182	949	0.442	?
140	0:104			0.07	0.229			0.086	949	0.64	?
140	0:36			0.05	0.103			0.066	1030	0.789	?
140	0:8			0.051	0.073			0.057	1894	0.256	?
140	0:8			0.05	0.057			0.052	1066	0.618	?
140	0:13			0.051	0.196			0.088	465	0.864	?
140	0:28			0.167	0.396			0.325	461	0.765	?
140	0:14			0.198	0.414			0.252	456	0.438	?
140	0:92			0.054	0.116			0.109	1066	0.871	?
140	0:71			0.05	0.152			0.07	1016	0.614	?
140	0:89			0.053	0.087			0.075	1561	0.603	?
140	0:58			0.052	0.119			0.069	1652	0.39	?
140	0:32			0.05	0.114			0.07	715	0.923	?
140	0:45			0.05	0.073			0.071	1389	0.868	?
140	0:87			0.055	0.26			0.096	1826	0.853	?
140	0:92			0.054	0.148			0.097	1339	0.605	?

Figure 4. Example of usage of internal database to sort out false positive variants. (A) Number of total samples sequenced with the panel (B) cumulative number of variant called in each variant caller (Ion reporter: NextGENe) (C) Identified minimum VAF in previous analyses (D) Identified maximum VAF in previous analyses (E) VAF and read depth information of the variant on the current analysis

In the post variant calling process, we tried to implement a strategy to avoid

allele dropout, which is a known limitation of PCR-based enrichment technology[26]. Carefully designing primer sets avoiding SNPs is essential in the first place. However, as a recent case reported, rare variant in primer site associated with pathogenic variant in *cis* form[27] cannot be detected in normal process. To detect 5% VAF pathogenic variant whose VAF cutoff was based on the paper by Jeong *et al*.[27], we showed if detected variants are on the primer sites. By indicating that, low-AF variants on the affected amplicon can go through extra careful inspection. For example (Figure 5), rare variant on the primer site affecting amplicon 'OBRA_BRCA2_83' was detected and low-AF pathogenic variant was detected on the specific amplicon. However, we can assume that the variant has a high chance of being a false positive call by the fact that it's been called 25 times previously and has not been reported in the literature. When using NGS panels commercially purchasable, information on primer sets are usually confidential making the previously described step impossible. However, by analyzing CNV, we can have tips on which amplicon has been dropped out. These steps are not mandatory for the practicing laboratories, however, we should put efforts to detect these low VAF true variants from allele drop out phenomenon. Also, addressing the limitation of the platform in the clinical report is essential. We did not find this allele drop out phenomenon during the research of this paper.

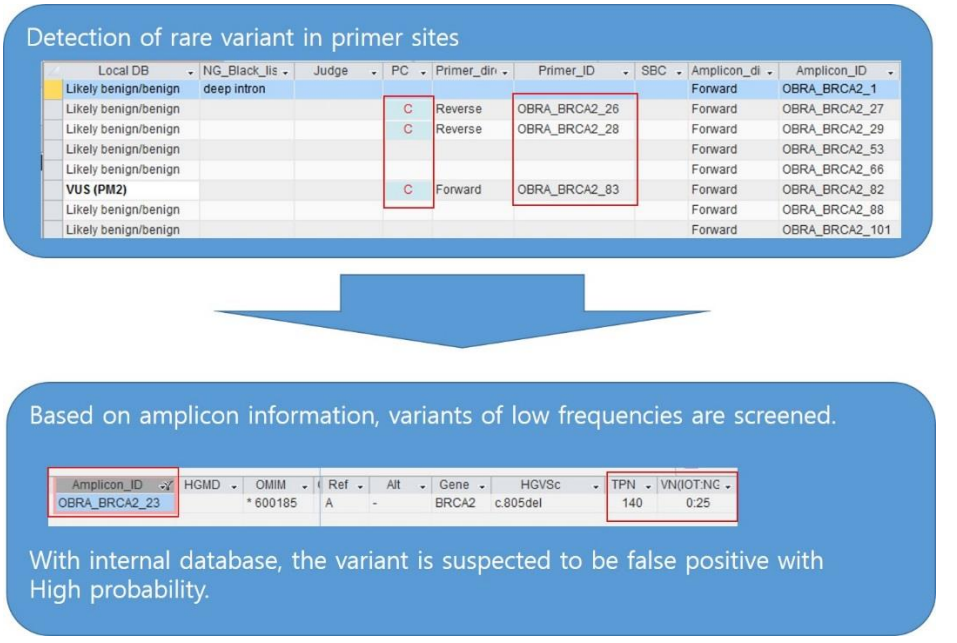

Figure 5. Example of strategy to detect allele drop out phenomenon. Pathogenic variants in *cis* position with rare SNV may be reported to be much lower than the VAF of normal heterozygous variants (exampled case indicates the variant of low AF is noise)

3. Diagnostic yield and clinical feature of BHD patients

Diagnostic yield of CLD NGS panel was 40.3% (25/62) and pathogenic aberrations were all in *FLCN* gene. Eleven patient (44.0%, 11/25) had hotspot mutation of either c.1285dup or c.1285del, which is consistent with the results from previous study[28]. One patient (4.0%, 1/25) had exon 7 deletion in *FLCN* and the incidence of CNV in *FLCN* is consistent with data in GeneReviews® (3-5%)[29]. Since germline mutations in *FLCN* were first identified in BHD patients in 2002[28,30], clinical features of BHD patients were intensively described in the literature; skin fibrofolliculomas, pulmonary cysts, spontaneous pneumothorax and renal cancer[31]. About 84% of BHD cases have cystic lung disease and 38% have spontaneous pneumothorax[32]. The risk of

renal cell carcinoma (RCC) is increased in BHD and in study investigating 312 BHD patients, the incidence of RCCs in *FLCN* mutation carriers over the age of 40 was 34.8%[4]. Interestingly, while the majority of Caucasian BHD families have fibrofolliculomas[31], only 3.8% of Japanese BHD patients had typical fibrofolliculomas[4]. As shown in our results on Table 8, out of thirteen BHD patients whose clinical data were available, 11 patients (84.6%) had spontaneous pneumothorax distinctive from non-BHD patients with multiple lung cysts (p value=0.001). And cutaneous skin lesions were not featured in any BHD patients (data not shown). Even though long-term follow-up of the BHD patients should be added to truly estimate the incidence of renal cancer in this group, only one patient had RCC when diagnosed with BHD. Our clinical data stands in line with the study that investigated Japanese BHD patients that characteristic features of lung and kidney lesions may be more informative than fibrofolliculomas as diagnostic criteria for BHD in the Japanese Asian population[4].

4. Negative findings in *TSC1* and *TSC2* genes

Given the fact that all sixty two patients had bilateral basally located lung cysts and had features that need differential diagnosis with BHD and LAM, we expected to detect pathogenic mutations in other genes than *FLCN*. As all patients had bilateral lung cysts and overlapping features of LAM on chest computed tomographies, we took extra closer look at *TSC1* and *TSC2* mutations to diagnose possible TSC. LAM and TSC are caused by mutations in either of the tuberous sclerosis genes, *TSC1* or *TSC2*[33]. Patients who meet standard clinical criteria for TSC, pathogenic mutations in *TSC1* or *TSC2* are found in 75–90% of cases[34]. 10 to 15% of TSC patients have no mutation identified (NMI) and in these NMI TSC patients, some studies found mosaic or subclonal somatic changes[34,35]. We didn't observe any heterozygous pathogenic mutations

in *TSC1* or *TSC2*. We searched for possible mosaic mutations and looked through all the variants with AF of 1%. However, we didn't find any significant variants distinctive from commonly detected background noises.

5. Disease-associated variants in *CFTR* gene and *CFTR* haplotyping

CF is a common autosomal recessive disorder among the Caucasian population, which is caused by mutations in the *CFTR* gene[36]. The disease affects multiple organs, including lung, pancreas, liver, sweat gland and intestine with variable degrees of severities among patients[37]. While CF incidence is one in 2000 to 4000 live births in US population and carrier (heterozygote) frequency is approximately one in 28 in the North American white population[36], CF in the Asian populations are very rare[38]. Given these facts, even though the patients in our study group have low probability to suffer from CF, carrier status of *CFTR* polymorphism or VUS may give some explanation on the pathogenicity of CLD in this study. Indeed, there have been several studies that investigated carriers of mild *CFTR* mutations have increased risks of chronic pancreatitis, chronic bronchitis or bronchiectatis[19,38,39].

After construction of *CFTR* haplotypes, we concluded that among 14 haplotypes constructed, p.M470V in combination with p.Q1352H on the backgrounds of wild type of the rest of genetic loci showed showed borderline significant difference between the control group and CLD patients from this study (p value=0.0542). Also, NM_000492.3:c.374T>C variant classified as VUS according to ACMG guideline[18], showed OR of 3.90-5.62. NM_000492.3: c.3468G>T variant showed OR of 3.81 compared with gnomAD database (https://gnomad.broadinstitute.org/). Interestingly, the haplotype '4' in Table 10 which indicates p.M470V in combination p.L1156F was only observed in this study group (not significantly different due to small number of analyzed sample). Kondo *et al*. demonstrated that this combinations of variants showed

reduced expression of CFTR protein[40]. CFTR protein acts as an ATP-gated anion channel[41] and there has been the notion that abnormal electrolyte transport is a key component of CF lung pathogenesis[42]. Abolished efficiency of ciliary-dependent mucus clearance in CF patient results in chronic infection and airway obstruction that leads to bronchiectasis[43]. However, assuming that *CFTR*-mutation carrier has gone through a structural deformation of the alveolar as consequences of inflammation or infections lacks evidence in our study because no carriers had chronic bronchitis or any other respiratory disease other than multiple cysts in lungs. Evidence and more functional studies must be conducted to hypothesize the pathogenesis of CLD in CF patients. Most importantly, further investigations and proper control groups are needed to conclude that this haplotype or variants are related to the novel phenotype of CLD in carriers of *CFTR* mid mutations.

V. CONCLUSION

We developed CLD NGS panel and optimized variant calling process in BI pipelines. Adjusting BI dramatically improved the overall performances. We also suggested post annotation process to sort out false positives and reduce misdiagnosis in case of allele dropout. Diagnostic yield using NGS went up from 38.7% to 40.3% compared with *FLCN* Sanger sequencing alone. Lastly, clinical usefulness of NGS panel is not limited to detecting the pathogenic variants. With the abundant genetic information, clinical laboratory can inform the clinicians about the disease-associated variants, suggest haplotype if needed, and provide informations of low AF variants with possible germline mosaicisim.

REFERENCES

1. Francisco FAF, Jr ASS, Zanetti G, Marchiori E. Multiple cystic lung disease. European Respiratory Review 2015;24:552-64.
2. Trotman-Dickenson B. Cystic lung disease: achieving a radiologic diagnosis. Eur J Radiol 2014;83:39-46.
3. Raoof S, Bondalapati P, Vydyula R, Ryu JH, Gupta N, Raoof S, *et al*. Cystic Lung Diseases: Algorithmic Approach. Chest 2016;150:945-65.
4. Furuya M, Yao M, Tanaka R, Nagashima Y, Kuroda N, Hasumi H, *et al*. Genetic, epidemiologic and clinicopathologic studies of Japanese Asian patients with Birt-Hogg-Dube syndrome. Clin Genet 2016;90:403-12.
5. Sato T, Seyama K, Fujii H, Maruyama H, Setoguchi Y, Iwakami S, *et al*. Mutation analysis of the *TSC1* and *TSC2* genes in Japanese patients with pulmonary lymphangioleiomyomatosis. J Hum Genet 2002;47:20-8.
6. Strandvik B, Bjorck E, Fallstrom M, Gronowitz E, Thountzouris J, Lindblad A, *et al*. Spectrum of mutations in the *CFTR* gene of patients with classical and atypical forms of cystic fibrosis from southwestern Sweden: identification of 12 novel mutations. Genet Test 2001;5:235-42.
7. Chin EL, da Silva C, Hegde M. Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. BMC Genet 2013;14:6.
8. Feng W, Zhao S, Xue D, Song F, Li Z, Chen D, *et al*. Improving alignment accuracy on homopolymer regions for semiconductor-based sequencing technologies. BMC Genomics 2016;17 Suppl 7:521.
9. Zeng F, Jiang R, Chen T. PyroHMMsnp: an SNP caller for Ion Torrent and 454 sequencing data. Nucleic Acids Res 2013;41:e136.

10. The Human Gene Mutation Database (HGMD) professional. In. Feb ed: BIOBASE GmbH.
11. Lim DH, Rehal PK, Nahorski MS, Macdonald F, Claessens T, Van Geel M, *et al*. A new locus-specific database (LSDB) for mutations in the folliculin (*FLCN*) gene. Hum Mutat 2010;31:E1043-51.
12. Jensen DK, Villumsen A, Skytte AB, Madsen MG, Sommerlund M, Bendstrup E. Birt-Hogg-Dube syndrome: a case report and a review of the literature. Eur Clin Respir J 2017;4:1292378.
13. HL R, SJ B, P B-T, JS B, KK B, JL D, *et al*. ACMG clinical laboratory standards for next-generation sequencing. Genet Med 2013;15:733–47.
14. Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, *et al*. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Arch Pathol Lab Med 2015;139:481-93.
15. Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, *et al*. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet 2016;24:2-5.
16. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, *et al*. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn 2018;20:4-27.
17. Online Mendelian Inheritance in Man (OMIM®). In: Johns Hopkins University.
18. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, *et al*. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.

Genet Med 2015;17:405-24.

19. Lee JH, Choi JH, Namkung W, Hanrahan JW, Chang J, Song SY, *et al*. A haplotype-based molecular analysis of *CFTR* mutations associated with respiratory and pancreatic diseases. Hum Mol Genet 2003;12:2321-32.
20. Burkett K, McNeney B, Graham J. A note on inference of trait associations with SNP haplotypes and other attributes in generalized linear models. Hum Hered 2004;57:200-6.
21. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, *et al*. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 2003;55:56-65.
22. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol 2013;9:e1003031.
23. Zhu P, He L, Li Y, Huang W, Xi F, Lin L, *et al*. OTG-snpcaller: an optimized pipeline based on TMAP and GATK for SNP calling from ion torrent data. PLoS One 2014;9:e97507.
24. Hume S, Nelson TN, Speevak M, McCready E, Agatep R, Feilotter H, *et al*. CCMG practice guideline: laboratory guidelines for next-generation sequencing. J Med Genet 2019. doi:10.1136/jmedgenet-2019-106152.
25. Gargis AS, Kalman L, Bick DP, da Silva C, Dimmock DP, Funke BH, *et al*. Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nat Biotechnol 2015;33:689-93.
26. Silva FC, Torrezan GT, Brianese RC, Stabellini R, Carraro DM. Pitfalls in genetic testing: a case of a SNP in primer-annealing region leading to allele dropout in BRCA1. Mol Genet Genomic Med 2017;5:443-7.
27. Jeong TD, Cho SY, Kim MW, Huh J. Significant allelic dropout

phenomenon of Oncomine BRCA Research Assay on Ion Torrent S5. Clin Chem Lab Med 2019;57:e124-e7.

28. Nickerson ML, Warren MB, Toro JR, Matrosova V, Glenn G, Turner ML, *et al*. Mutations in a novel gene lead to kidney tumors, lung wall defects, and benign tumors of the hair follicle in patients with the Birt-Hogg-Dube syndrome. Cancer Cell 2002;2:157-64.
29. Toro JR AM, Ardinger HH, Pagon RA, *et al*. Birt-Hogg-Dubé Syndrome. In: GeneReviews®, ed., 2014.
30. Schmidt LS, Warren MB, Nickerson ML, Weirich G, Matrosova V, Toro JR, *et al*. Birt-Hogg-Dube syndrome, a genodermatosis associated with spontaneous pneumothorax and kidney neoplasia, maps to chromosome 17p11.2. Am J Hum Genet 2001;69:876-82.
31. Menko FH, van Steensel MA, Giraud S, Friis-Hansen L, Richard S, Ungari S, *et al*. Birt-Hogg-Dube syndrome: diagnosis and management. Lancet Oncol 2009;10:1199-206.
32. Hayashi M, Takayanagi N, Ishiguro T, Sugita Y, Kawabata Y, Fukuda Y. Birt-Hogg-Dube syndrome with multiple cysts and recurrent pneumothorax: pathological findings. Intern Med 2010;49:2137-42.
33. McCormack FX. Lymphangioleiomyomatosis: a clinical update. Chest 2008;133:507-16.
34. Tyburczy ME, Dies KA, Glass J, Camposano S, Chekaluk Y, Thorner AR, *et al*. Mosaic and Intronic Mutations in *TSC1*/*TSC2* Explain the Majority of TSC Patients with No Mutation Identified by Conventional Testing. PLoS Genet 2015;11:e1005637.
35. Nellist M, Brouwer RW, Kockx CE, van Veghel-Plandsoen M, Withagen-Hermans C, Prins-Bakker L, *et al*. Targeted Next Generation Sequencing reveals previously unidentified *TSC1* and *TSC2* mutations. BMC Med Genet 2015;16:10.

36. Moskowitz SM, Chmiel JF, Sternen DL, Cheng E, Gibson RL, Marshall SG, *et al*. Clinical practice and genetic counseling for cystic fibrosis and *CFTR*-related disorders. Genet Med 2008;10:851-68.
37. Chou JL, Rozmahel R, Tsui LC. Characterization of the promoter region of the cystic fibrosis transmembrane conductance regulator gene. J Biol Chem 1991;266:24471-6.
38. Wang P, Naruse S, Yin H, Yu Z, Zhuang T, Ding W, *et al*. The susceptibility of T5-TG12 of the *CFTR* gene in chronic bronchitis occurrence in a Chinese population in Jiangsu province, China. J Biomed Res 2012;26:410-7.
39. Cho SM, Shin S, Lee KA. PRSS1, SPINK1, *CFTR*, and CTRC Pathogenic Variants in Korean Patients With Idiopathic Pancreatitis. Ann Lab Med 2016;36:555-60.
40. Kondo S, Fujiki K, Ko SB, Yamamoto A, Nakakuki M, Ito Y, *et al*. Functional characteristics of L1156F-*CFTR* associated with alcoholic chronic pancreatitis in Japanese. Am J Physiol Gastrointest Liver Physiol 2015;309:G260-9.
41. Hwang TC, Kirk KL. The *CFTR* ion channel: gating, regulation, and anion permeation. Cold Spring Harb Perspect Med 2013;3:a009498.
42. Boucher RC. New concepts of the pathogenesis of cystic fibrosis lung disease. Eur Respir J 2004;23:146-58.
43. Schafer J, Griese M, Chandrasekaran R, Chotirmall SH, Hartl D. Pathogenesis, imaging and clinical characteristics of CF and non-CF bronchiectasis. BMC Pulm Med 2018;18:79.

ABSTRACT(IN KOREAN)

낭성폐질환 진단을 위한 차세대 염기서열 분석 패널의 개발 : 진단률 평가 및 바이오인포매틱스 파이프라인의 최적화

<지도교수 이 경 아>

연세대학교 대학원 의학과

오 주 원

낭성폐질환(cystic lung disease, CLD)은 '낭(cyst)'이라는 공통된 특징을 가지는 다양한 질환을 대표하는 질환군이다. 이러한 공통된 특징을 보이는 질환이 증가함에 따라 그 진단이 더욱 어려워지고 있다. CLD 중 연관된 유전자의 변이와 해당 질환과의 연관성이 잘 정의된 질환이 여러 개가 있다; 림프관평활근종증 (lymphangioleiomyomatosis, LAM, 빌트-호그-두베 증후군(Birt-Hogg-Dube syndrome, BHD), 결절성경화증(Tuberous sclerosis complex, TSC), 낭성섬유증(cystic fibrosis, CF) 등이 그 예이다. 지금까지는 CLD에 대한 분자유전학적 진단이 주로 생거시퀀싱(Sanger sequencing)에 의한 염기서열분석으로 이루어졌다. 질환과의 연관성이 밝혀진 모든 유전자에 대한 생거시퀀싱을 하는 경우 비용이 급격히 증가하기 때문에, 가장 높은 빈도로 변이가 보고되어 있는 유전자부터 먼저 검사를 한 뒤, 임상양상으로 판단했을 때 가장 의심되는 유전자를 순차적으로 검사하는 방식으로 유전학적 검사가 이루어져왔다. 이러한 관점에서 볼 때, 생거시퀀싱 방식으로 CLD 감별진단을 하는 것은 많은 비용과 시간이 필요하다는 문제점이 있기 때문에, 자연스럽게 많은 유전자를 한꺼번에 검사할 수 있는 검사법에 대한

필요성이 대두되었다. 최근, 표적 차세대염기서열분석(targeted next-generation sequencing, NGS) 방법이 개발되어 더 낮은 비용으로 더 많은 유전자들을 한꺼번에 표적하여 검사할 수 있게 되었다. 그러나, NGS 검사장비는 동일염기 반복서열을 발견하는데 부정확하거나 구조적 변이(structural variant, SV)를 찾아내는 것이 어려운 것이 단점이다. 이 연구에서 우리는 NGS 플랫폼을 이용한 맞춤형 낭성폐질환 패널 (CLD panel)의 진단율을 평가하고, 동일염기 반복서열을 정확히 기술하기 위한 최적의 바이오인포매틱스 변수 (parameter)를 정하였다.

총 62명의 낭성폐질환 환자를 대상으로 연구를 진행하였고, 모든 환자 검체는 FLCN 유전자의 Sanger 시퀀싱과 Multiplex Ligation-dependent Probe Amplification 검사를 시행하였으며 Ion torrent S5 NGS platform으로 NGS 시퀀싱을 진행하였다. 3개 종류의 바이오인포매틱스 파이프라인을 비교 하였다; NextGENe v.2.4.2.2, Ion Reporter Software 5.10, Biomedical Genomics Workbench 5.0.

최적화 후의 분석적 민감도는 Ion Reporter는 98.6% 에서 98.8%로 증가하였고, NextGENe은 99.0% 에서 100.0%로, Workbench는 99.8%에서 100.0%로 민감도가 모두 증가하였다. 전반적인 임상적 진단율은 *FLCN* Sanger 시퀀싱과 비교하였을 때, 38.7% 에서 40.3%로 증가하였다. 민감도 측면에서, Ion Reporter의 경우 최적화 후에도, 반복서열로 이루어진 hotspot 변이를 발견하지 못하였으나 일련의 최적화 과정과 여러 개의 독립적인 바이오인포매틱스 파이프라인을 종합적으로 고려함으로서, 전반적으로 NGS panel의 성능을 향상시킬 수 있었다.

그 외에도, CFTR 유전자에서 질병연관 변이를 발견하고, 하플로타입을 조합해보았다. NM_000492.3:c.374T>C 변이는 ACMG 가이드라인에 따르면 VUS로 분류되는 변이지만, CLD 환자군과 정상 환자군의 발현빈도 차에 따른 오즈비를 구했을

때, 3.90-5.62로 높게 나와 질병과의 연관성을 시사하였다. 또한 하플로타입 분석 결과, p.M470V과 p.Q1352H 변이가 있는 하플로타입이 CLD와 통계학적으로 경계적 (borderline) 중요성을 가지는 것으로 나타났다 (p value=0.0542).

우리는 이 연구에서 낭성폐질환 NGS 패널을 개발하고, 바이오인포매틱스 파이프라인을 최적화 시켰다. 진단율은 38.7%에서 40.3%로 증가하였다. 끝으로 하고자 하는 말은 NGS panel의 임상적 유용성은 질병원인 유전자를 밝히는데만 국한되지 않는다는 것이다. NGS를 통해 얻어진 많은 유전적 정보를 통해, 우리는 질병연관성 변이를 보고할 수 있고, 필요에 따라 하플로타입을 보고할 수 있으며, 낮은 빈도로 발견된 임상적으로 중요한 변이의 모자이시즘의 가능성도 언급해줄 수 있다는 점에서 더욱 그 임상적 유용성을 증가시킬 수 있다.

핵심되는 말: 낭성폐질환; 빌트-호그-두베 증후군; 차세대 염기서열 검사; 바이오인포매틱스