such as XML, are not very readable for humans. Moreover, these workflow systems often require wrapper tools, which are called "shims", to handle third-party unsupported existing code or data sources [17,18]. This sometimes obstructs quick deployment of newer tools. In actual bioinformatics projects, we realized that scientific workflow systems often require quick deployment of cutting-edge software to implement new algorithms and data formats, frequent workflow optimization after trial and error and in following changes in computational resources and the environment. The agile software development method considers similar problems in software development projects. Kane *et al.* summarized this by stating that "Agile is an iterative approach to software development on strong collaboration and automation to keep pace with dynamic environment", and "Agile methods are well suited to the exploratory and iterative nature of scientific inquiry" [19]. Therefore, scientific workflow systems require both rigidity in workflow management and agility in workflow development.

One of the traditional solutions for balancing the two aspects of a workflow system is the `make` command, a standard build tool in the Unix system. The `make` command interprets a `Makefile`, which defines dependencies between files in a declarative programming manner, and then generates the final target by resolving dependencies, by only executing out-of-date steps. This approach has been extended to cluster environments such as GXP make [20]. However, the make-based approach has limitations in describing scientific workflows because it is intended for building software. For example, it is difficult to describe the "multiple instances with *a priori* runtime knowledge" pattern, which is one of the workflow patterns defined by Van der Aalst *et al.* [1], in makefiles without external tools. In this pattern, the number of instances is unknown before the workflow is started, but becomes known at some stage during runtime. In other words, this situation requires dynamic workflow definition at runtime. This pattern appears frequently in scientific workflows as well as embarrassingly parallel problems. Introduction of internal domain specific languages (DSLs) to workflow description is an approach to overcome this limitation. Internal DSLs are implemented as libraries of the host languages. Thus, an internal DSL retains the descriptiveness of the host language.

Introduction of the internal DSL into make-like workflow systems has been shown in object-oriented scripting languages including Python [21] and Ruby [22]. An implementation in Python is Ruffus [23], which is a scientific workflow system supporting execution limited to out-of-date stages, dynamic workflow definition, flowchart generation, and parallelism. PaPy [24], another workflow system in Python, was implemented with a modular design and offers parallel and distributed workflow management. On the other hand, the Ruby programming language also has a greater affinity to the internal DSL approach because of its flexible syntax, including omissible parentheses and a code-block grammar [25]. Rake [26] is a 'Ruby Make', which is a build tool with workflow definition implemented as an internal DSL in Ruby and a standard library of Ruby version 1.9 or later. Rake supports execution of workflows limited to out-of-date stages and dynamic workflow definition during workflow execution. The following is a simple example of a workflow definition file, a `Rakefile`:

```
1: CC = "gcc"
2: rule '.o' => '.c' do |t|
3:   sh "#{CC} -c #{t.source}"
4: end
5: file "sample" => ["sample.o"] do |t|
6:       sh  "#{CC}  -o  #{t.name}  #{t.
prerequisites}"
7: end
8: task :default => "sample"
```

This example defines a workflow to generate an executable `sample` from `sample.c` via `sample.o`. If `sample.c` is out-of-date, i.e., older than `sample.o`, Rake skips compiling `sample.c` and just links `sample.o` to generate `sample`. Note that the grammar of the rakefile is fully compatible with that of Ruby.

Recently Tanaka and Tatebe developed Pwrake [27], a parallel workflow extension of Rake. Pwrake has been demonstrated to be a flexible scientific workflow system in the astronomy domain [28]. It interprets rakefiles that are fully compatible with Rake. Pwrake supports parallelism by automatically detecting parallelizable tasks and executing them via SSH connections. Pwrake generates a flowchart as a directed acyclic graph in the DOT language, which is then visualized by software such as Graphviz [29]. Although we focus on workflow management using a local multiprocessor and multicore environment, Pwrake can be used with computer clusters together with the support of a distributed filesystem such as NFS. Pwrake is especially designed for scalable parallel I/O performance using the Gfarm global distributed filesystem [28,30].

In this paper, we show agile workflow management using Pwrake in the bioinformatics domain.

Implementation

Rakefiles

In actual bioinformatics workflow development, we found that the scientific workflow development iterates over two phases, the workflow definition phase and the parameter adjustment phase. The former focuses on the functional combination and order of tasks, while the latter focuses on the optimization of command-line parameters for invoking tools. We therefore, designed

separate rakefiles corresponding to these two phases. Task dependencies are defined in `Rakefile`, while command-line programs and parameters are defined in `Rakefile.invoke`. To simplify the description, we also implemented a file to define helper methods, `Rakefile.helper` (Figure 1).

`Rakefile` is the main and default task definition file. It loads two other rakefiles, sets target filenames in constants, and declares task dependencies. Other rakefiles are loaded by the `Kernel#load` method to enable reloading to reflect changes immediately.

`Rakefile.invoke` defines a class with a unique name in the RakefileInvoke module. In the class, paths to commands and common files, as well as adjustable parameters are set to constants. It also defines methods to invoke command-lines using `FileUtils#sh` methods. These methods are defined as singleton methods (eigenmethods) of the class. This is an internal DSL technique in Ruby to enable invocation in rakefiles as in "`RakefileInvoke::Gatk::command t, opts`", where `t` is an instance of the `Rake::Task` class and `opts` is a hash object containing the optional information to invoke commands. `Rakefile.helper` defines helper methods to simplify the rakefile descriptions. For example, the `suffix` method in the top level allows the replacement of the filename suffix using expressions with arrows. Additionally, Pwrake requires a nodefile to specify hostnames and maximum numbers of processes to be submitted via SSH connections. A nodefile declaring a local machine that can execute 16 processes simultaneously is set as "`localhost 16`".

Command-lines to start the workflow using Rake and Pwrake are "`rake`" and "`pwrake NODEFILE = nodefile`", respectively. By default, Rake and Pwrake load the file called "`Rakefile`" in the current directory. Rakefiles are usually placed in the topmost directory in a project file tree. To simplify provenance management, we recommend that each project file tree has its own copy of the rakefile.

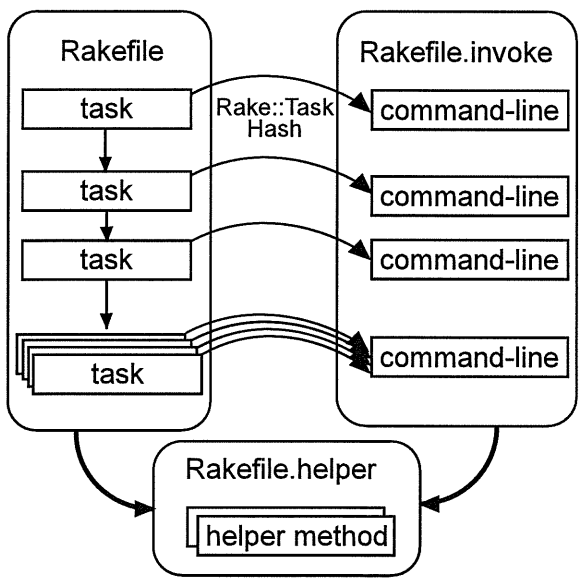


Figure 1 Structure of distinct rakefiles. A `Rakefile` file consists of task dependency descriptions. Tasks may be executed in parallel, if possible automatically. The `rakefile.invoke` file defines a class of the `RakefileInvoke` module. This class defines class methods to invoke command-lines and constants of command paths and parameters. Tasks in the rakefile call methods with an instance of the `Rake::Task` class and a hash containing additional parameters for invoking the command-line. The `Rakefile.helper` file defines helper methods to simplify descriptions in the `Rakefile` and `Rakefile.invoke` files.

Example workflows

To demonstrate the workflows described in Pwrake rakefiles, we implemented two kinds of workflows for the Genome Analysis Toolkit (GATK) [31,32] and Dindel [33] using rakefiles. Both GATK and Dindel have been used in whole genome sequencing projects including the 1000 genomes project [34]. We selected GATK and Dindel as typical examples for sequential and parallel workflows, respectively. Furthermore, we implemented a combined workflow loading externally defined GATK and Dindel workflows to show the modularity thereof.

The GATK workflow

GATK is a program suite written mainly in Java to process mapped reads obtained from massively parallel sequencing data to detect genetic variants including single nucleotide variants (SNVs). The GATK development team offers several recommended workflows depending on the samples and analyses. We implemented their 'better' workflow (Figure 2A). In `Rakefile`, the `Rakefile::Gatk` class defines constants indicating the target files in each step of the workflow. These constants are used to define the `:default` task to obtain the final product of the workflow. In `Rakefile.invoke`, the `RakefileInvoke::Gatk` class defines constants indicating the file paths to executables and downloaded public data files, such as the reference genome sequence and dbSNP data. These help the workflow configuration in other environments and improve readability. The class also defines methods to execute command-lines for each step in the workflow.

The Dindel workflow

Dindel is a suite of tools for detecting small genetic insertions and deletions (indel) from massively parallel sequencing data. The overview of the rakefile structure for GATK and Dindel is the same; however, a Dindel workflow is a good example of a parallel workflow using

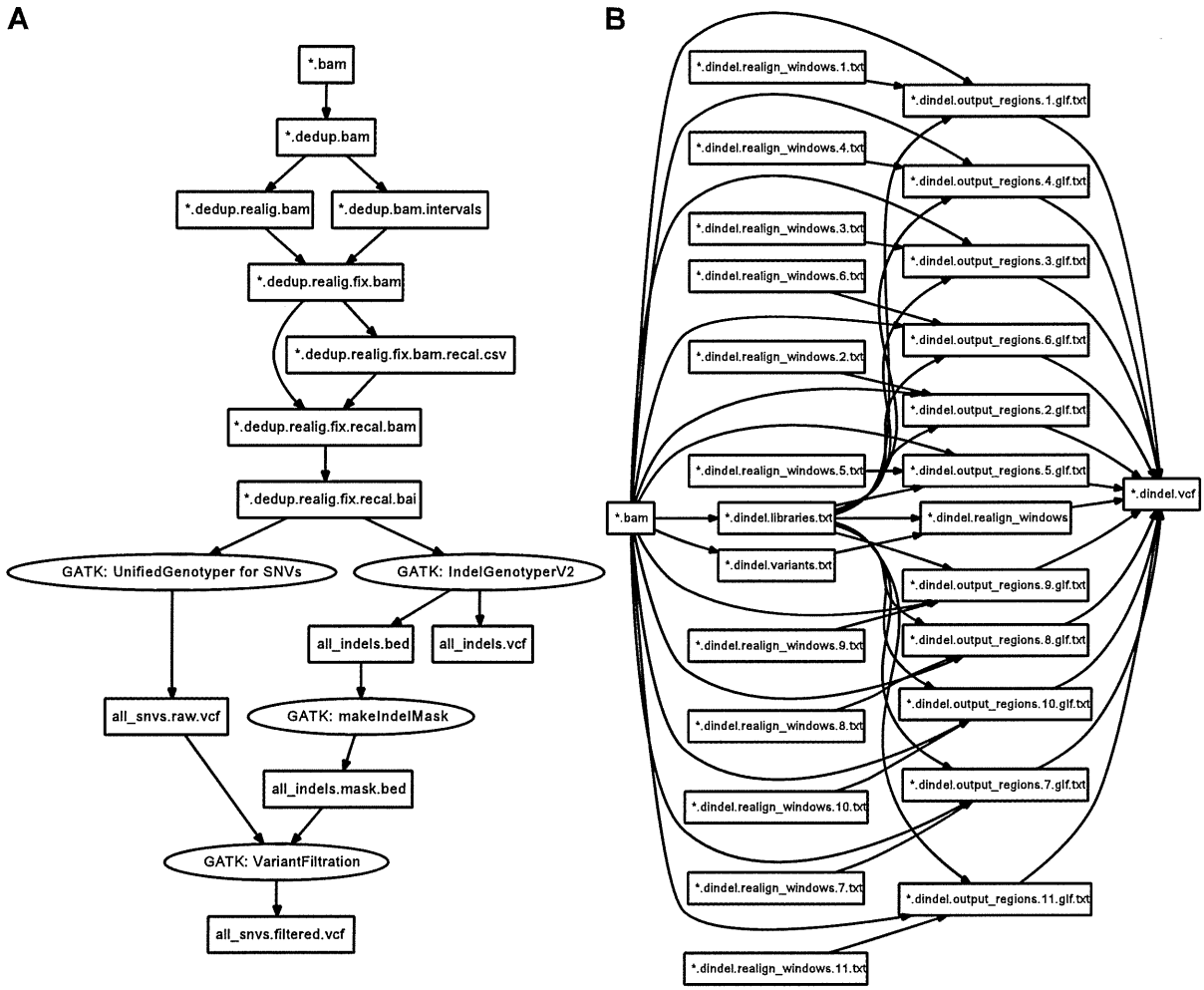


Figure 2 Directed acyclic graphs of GATK and Dindel workflows. Directed acyclic graphs to describe GATK (A) and Dindel (B) workflows are generated by Pwrake, manually adjusted, and visualized using Graphviz. (A) depicts a workflow from a `*.bam` file to an `all_snvs.filtered.vcf`, while (B) depicts a workflow from a `*.bam` file to a `*.dindel.vcf` file. Tasks to process the `*.dindel.realign_windows.N.txt` files are dynamically generated during the embarrassingly parallel stage (stage 3) depending on `*.dindel.libraries.txt` and `*.libralies.txt`. Prior to stage 3, an option "`-numWindowsPerFile 1000`" is applied to `makeWindows.py` to generate 11 files containing windows. File `*.didndel_reagion_window` is a dummy file representing all the `*.dindel.realign_windows.N.txt` files.

the dynamic task definition (Figure 2B). Such a workflow generates many intermediate files. In the authors' experience, one human exome generates more than 300 "window" files, where each window file can contain a maximum of 1000 windows. These intermediate window files are named systematically; however, the number of window files is unknown prior to the workflow execution. A rakefile can describe this situation using a dynamic task definition. Furthermore, Pwrake can automatically detect tasks that can be executed in parallel. The following is an example of dynamic task definition codes based on the stage 3 definition of the Dindel workflow in `Rakefile` and `Rakefile.invoke`.

```
1: # Rakefile
2: task :stage3 = > :stage2 do
3:  Rakefile::Dindel::BAM.each do |bam|
4:    prefix =
5:          bam.sub(/\.bam$/, ".dindel.
realign_windows")
6:    FileList["#{prefix}.*.txt"].each
do |f|
7:      target = f.sub(/\.realign_win-
dows\./,
9:            ".output_regions.").
6:          sub(/\.txt$/, ".glf.txt")
7:     prerequisites =
```

```
 8:       [f,
 9:      f.sub(/\.dindel\.realign_windows
\..*/, ".bam"),
 10:      f.sub(/\.dindel\.realign_windows
\..*/,
 11:          ".dindel.libraries.txt"),]
 12:      file target => prerequisites do
|t|
 13:        RakefileInvoke::Dindel.din-
del_stage3 t
 14:      end
 15:      file :stage3_invoke => target
 16:     end
 17:    end
 18:     (task :stage3_invoke).invoke
 18: end
 1: # Rakefile.invoke
 2: def dindel_stage3(t)
 3:  sh [DINDEL,
 4:    "-analysis indels",
 5:    "-doDiploid",
 6:    "-bamFile #{t.prerequisites[1]}",
 7:    "-ref #{REFERENCE}",
 8:    "-varFile #{t.prerequisites[0]}",
 9:    "-libFile #{t.prerequisites[2]}",
 10:     "-outputFile #{t.name.sub(/\.glf
\.txt$/, "")}",
 11:     "1 > #{t.name.sub(/\.glf\.txt$/,
"")}.log 2 > &1",
 12:    ].join(" ")
 13: end
```

In this sample rakefile, the `:stage3` task expects that the previous task `:stage2` generates files that are named `*.dindel.realign_windows.N.txt`, where N is the serial number of the intermediate file. The maximum value of N is unknown prior to execution of the `:stage2` task. The dependency of the following stages can be defined using the task name `:stage3`.

Pwrake automatically detects that `:stage3` consists of independent `file` tasks and executes them as an embarrassingly parallel stage. In the `:stage2` definition in `Rakefile.invoke`, the granularity of parallelism can be defined by the "`-numWindowsPerFile`" option of `makeWindows.py`. For the exome dataset aligned to chromosome 21, we used 1000 and 1 for this option and obtained 11 and 3381 intermediate `realign_windows` files, respectively.

Combination of rakefiles

Existing rakefiles can be combined by being loaded into another rakefile. Constants and methods defined in `rakefile.invoke` files have independent namespaces. Moreover, a task with the same identifier, such as the `:default` task, can be defined multiple times and thus can be appended. Pwrake and Rake do not overwrite, but append the files. For example, a rakefile to define GATK and Dindel workflows simultaneously simply contains the following:

```
1: load "../GATK/Rakefile"
2: load "../Dindel/Rakefile"
```

Results

Performance

To evaluate the performance of the GATK and Dindel workflows, we analysed publicly available short read sequence data using a Linux system that can execute 16 concurrent threads (2 processors × 4 cores with hyper-threading). Whole genome sequencing data [35] obtained from a HapMap [36] JPT sample NA18943 was used as the test dataset. The dataset was mapped to the GRCh37 referential genome sequence using the Burrows-Wheeler Alignment tool (BWA) [37] to generate a SAM file [38]. The SAM file was converted to a BAM file using Picard [39]. Reads mapped on chromosome 21 were used as initial data for both the GATK and Dindel workflows. We executed both Rake and Pwrake with the same rakefiles to compare the performance with parallelism. The wall-clock times for the GATK workflows executed by Rake and Pwrake were almost identical (approximately 12.0 min). We assume that this is due to the high sequentiality of the workflow. For the Dindel workflow, we assessed different parallelism granularities. When the task was divided into 11 processes in stage 3, the Dindel workflow executed by Pwrake was 2.6 times faster (approximately 6.0 min) than that by Rake (approximately 15.5 min). When the task was divided into 3381 processes in stage 3, the Pwrake execution was 4.6 times faster (approximately 4.0 min) than the Rake execution (approximately 18.3 min). While the ideal parallel acceleration efficiency was 16 times for our computer environment, the actual efficiency differed. These results can be explained by the fact that the required CPU-time to finish each process was uneven, and a few heavy processes were bottlenecks in the workflow execution. This is a limitation of process-based parallelism because of the relatively coarse parallelization granularity.

Agility in workflow development

A characteristic of agile software development is the iterative development process. We introduced an agile scientific workflow development that employed the iteration of two developmental phases, i.e., the workflow definition phase and the parameter adjustment phase. In each phase, our implementation of distinct rakefiles enabled the separate files to be modified. This separation increased efficiency in the iterative development.

Here, we show an example of the iterative development in our GATK workflow. In the workflow definition

phase, we focus on describing a task dependency in a rakefile as shown below:

```
1: rule '.dedup.bam.intervals' => 
2:  [ suffix_proc(".bam.intervals" => ".
bam") ] do |t|
3:     RakefileInvoke::Gatk.gatk_rea-
ligner_target_creater t
4: end
```

Next, in the parameter adjustment phase, we focus on describing command-line parameters for invoking external tools in the rakefile.invoke such as the following:

```
1: def gatk_realigner_target_creater(t)
2: sh [Java,
3:   "-Xmx#{JavaMemory}",
4:           "-Djava.io.tmpdir  =  #
{JavaTempFile}",
5:   "-jar #{GATK_JAR}",
6:   "-T RealignerTargetCreator",
7:   "-R #{REFERENCE}",
8:   "-o #{t.name}",
9:   "-I #{t.source}",
10:   "-D #{DBSNP}",
11:            RakefileInvoke::Gatk::
INTERVAL_OPTION,
12:   " > #{t.name}.log 2 > &1",
13:   ].join(" ")
14: end
```

Note that all constants with names starting with uppercase letters are defined at the top of the file, `rakefile.invoke`. The next iteration starts with the workflow definition phase again to extend the workflow. Modification or optimization after the workflow has completed can be achieved by iterating the same two phases using two distinct files. Separating the rakefiles simplifies finding files and places to be modified.

Procedure to describe new workflows

As a summary of the agile workflow development, the general procedure for describing new workflows in Pwrake is given below.

1) Workflow definition phase. Describe file dependencies in `Rakefile`.

```
1: task "output.dat" => "input.dat" do |
t|
2:  RakefileInvoke::generate_target t
3: end
```

2) Parameter adjustment phase: Define the `RakefileInvoke::generate_target` method in `Rake.invoke`.

```
1: module RakefileInvoke
2:  def generate_target(t)
3:    sh "command-line #{t.prerequisite}
> #{t.name}"
4:  end
5: end
```

3) Iteration of phases. Parameter adjustments require modifications to `Rakefile.invoke` only. Similarly, changes in file dependencies require modification to `Rakefile` only.

Discussion

Advantages in workflow execution

Workflows involving actively developed software packages, such as GATK, require frequent updates of details, such as combinations of data and programs, recommended parameters, and command-line options. Thus, well-organized workflow management helps GATK users to follow updates and process their data in improved workflows. A GATK workflow consists of multiple steps and takes a relatively longer time to finish. Pwrake has advantages of continuous execution of workflow tasks and selective task execution to ignore already executed tasks. Such ignorable tasks can be obtained from unexpected workflow suspension. Thus far, Pwrake cannot automatically remove output files containing partial results; such files have to be removed manually prior to restarting the workflow.

For the Dindel workflows, the parallelism offered by Pwrake improved performance. The parallelization model of Pwrake is process-based. Parallel programs based on technologies such as message passing interface (MPI) [40] enable efficient parallelization with fine granularity. However, scientists implementing bioinformatics software often focus not on parallelization, but on the novel implementation methodology. Therefore, process-based parallelization using non-parallel programs is a realistic solution and still has the advantage [41]. Furthermore, process-based parallelization can be efficient enough for embarrassingly parallel problems that can easily be separated into independent tasks and executed in parallel. For example, a stage in the Dindel workflow creates multiple intermediate files. Processes using these files as input are independent and do not need to communicate with each other. This stage is a typical embarrassingly parallel problem. Although the GATK framework supports the functional programming concept of MapReduce [42] and parallelism in the GATK framework is expected to improve its performance, it has only been supported to a limited extent by GATK components to date. Therefore, Pwrake still has the advantage with respect to parallelism.

Workflow description flexibility

One of the advantages of using an internal DSL is that the power of the host language is also available in the DSL scripts. The rakefile description is an internal DSL in Ruby, which is a programming language with a shallow learning curve for biologists [43]. Thus, rakefiles can make full use of the control flow features of Ruby, as well as the rich libraries for text processing, file manipulation,

network access, and so on. In particular, the BioRuby [44] library offers highly abstracted data processing methods for bioinformatics.

Sharing workflows

One of the key characteristics of agile software development is strong collaboration among all the people involved in the project. This can be accomplished naturally in projects in small laboratories. However, the nature of science is a global collaboration. Indeed, efforts to share and reuse workflows in the science community, such as the myExperiment project [45] and Wf4Ever [46], have already been started. From this point of view, the simplicity and readability of the rakefile DSL are advantageous, and improvement of helper methods to standardize the scripting style on the "Do not Repeat Yourself (DRY)" principle may enhance the advantages.

Conclusions

We have shown an appreciation of Pwrake as an agile parallel workflow system suitable for the bioinformatics domain using examples of GATK and Dindel workflows. Pwrake is able to invoke command-line tools without any "shims", define tasks dynamically during the workflow execution, and invoke tasks automatically in parallel. Separating a rakefile into two files for the workflow definition phase and the parameter adjustment phase increases the efficiency of the iterative workflow development. The nature of scientific projects is explorative and iterative. This is also a characteristic of agile software development. Another aspect of agile development, the reliance on the strong collaboration, may be enhanced by sharing and reusing workflows among the scientific community by taking advantage of the simplicity, readability and maintainability of rakefiles.

Availability and requirements

Project name: Workflows
Project home page: http://github.com/misshie/Workflows
Operating system(s): Platform independent
Programming language: Ruby 1.9.1 or higher
Other requirement: Pwrake or Rake
License: the MIT license
Any restrictions for use by non-academics: none

Availability of supporting data

Sample short read data for workflow evaluation: http://trace.ddbj.nig.ac.jp/DRASearch/experiment?acc=DRX000358

List of abbreviations used

HPC: high-performance computing; DSL: domain specific language; GATK: Genome Analysis Toolkit; SNV: single nucleotide variant; BWA: Burrows-Wheeler Alignment tool; MPI: message passing interface; DRY: do not repeat yourself.

Acknowledgements

The authors would like to thank members of the BioRuby mailing list for their informative discussions. HM is supported by the MEXT Grant-in-Aid for Young Scientists (B) 21791566 and 23791230. OT is supported by the MEXT Grant-in-Aid for Scientific Research on Priority Areas 21013005. OT and MT are supported by the MEXT Promotion of Research for Next Generation IT Infrastructure "Resources Linkage for e-Science (RENKEI)", and JST CREST "Development of System Software Technologies for post-Peta Scale High Performance Computing". KY is supported by grants from the Ministry of Health, Labour and Welfare, Grant-in-Aid for Scientific Research (B) 21390100 and the Takeda Scientific Foundation.

Author details

[1]Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, 1-12-4 Sakamoto, Nagasaki, Nagasaki, Japan. [2]Nagasaki University Global Center of Excellence Program, 1-12-4 Sakamoto, Nagasaki, Nagasaki, Japan. [3]Center for Computational Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, Japan. [4]Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama, Japan. [5]Departmentent of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, Japan.

Authors' contributions

HM conceived the study, implemented the workflows, and co-authored the manuscript. KS implemented the workflows. MT and OT developed Pwrake and evaluated the details of the workflows and the computational performance. KY conceived the study and co-authored the manuscript. All authors read and approved the final manuscript

Competing interests

The authors declare that they have no competing interests.

Received: 27 May 2011 Accepted: 8 September 2011
Published: 8 September 2011

References

1. Van der Aalst WMP, Ter Hofstede AHM, Kiepuszewski B, Barros AP: **Workflow patterns.** *Distrib Parallel Dat* 2003, **14**:5-51.
2. Zhao Y, Raicu I, Foster I: **Scientific Workflow Systems for 21st Century, New Bottle or New Wine?** *2008 IEEE Congress on Services - Part I* Honolulu, HI, USA; 2008, 467-471.
3. Hoon S, Ratnapu KK, J-ming Chia, Kumarasamy B, Juguang X, Clamp M, Stabenau A, Potter S, Clarke L, Stupka E: **Biopipe: A Flexible Framework for Protocol-Based Bioinformatics Analysis.** *Genome Res* 2003, **13**:1904-1915.
4. Deelman E, Blythe J, Gil Y, Baker C, Mehta G, Vahi K, Blackburn K, Lazzarini A, Arbree A, Cavanaugh R: **Mapping complex scientific workflows onto distributed systems.** *J Grid Comp* 2003, **1**:25-39.
5. Eker J, Janneck JW, Lee EA, Liu J, Liu X, Lidvig J, Neuendorffer S, Sachs S, Xiong Y: **Taming heterogeneity - the Ptolemy approach.** *Proc IEEE* 2003, **91**:127-144.
6. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P: **Taverna: a tool for the composition and enactment of bioinformatics workflows.** *Bioinformatics* 2004, **20**:3045-3054.
7. Shah S, He D, Sawkins J, Druce J, Quon G, Lett D, Zheng G, Xu T, Ouellette BF: **Pegasys: software for executing and integrating analyses of biological sequences.** *BMC Bioinformatics* 2004, **5**:40.
8. Ludäscher B, Altintas I, Berkley C, Higgins D, Jaeger E, Jones M, Lee EA, Tao J, Zhao Y: **Scientific workflow management and the Kepler system.** *Concurrency Computat Pract Exper* 2006, **18**:1039-1065.
9. Churches D, Gombas G, Harrison A, Maassen J, Robinson C, Shields M, Taylor I, Wang I: **Programming scientific and distributed workflow with Triana services.** *Concurrency Computat Pract Exper* 2006, **18**:1021-1037.
10. Romano P, Bartocci E, Bertolini G, De Paoli F, Marra D, Mauri G, Merelli E, Milanesi L: **Biowep: a workflow enactment portal for bioinformatics applications.** *BMC Bioinformatics* 2007, **8**:S19.

11. Zhao Y, Hategan M, Clifford B, Foster I, Von Laszewski G, Nefedova V, Raicu I, Stef-Praun T, Wilde M: **Swift: Fast, reliable, loosely coupled parallel computation.** *Proceedings - 2007 IEEE Congress on Services, SERVICES 2007* 2007, 199-206.
12. Bartocci E, Corradini F, Merelli E, Scortichini L: **BioWMS: a web-based Workflow Management System for bioinformatics.** *BMC Bioinformatics* 2007, **8**:S2.
13. Fiers M, van der Burgt A, Datema E, de Groot J, van Ham R: **High-throughput bioinformatics with the Cyrille2 pipeline system.** *BMC Bioinformatics* 2008, **9**:96.
14. Berthold MR, Cebron N, Dill F, Gabriel TR, Kotter T, Meinl T, Thiel K, Wiswedel B: **KNIME - The Konstanz Information Miner.** *SIGKDD Explorations* 2009, **11**:26-31.
15. Orvis J, Crabtree J, Galens K, Gussman A, Inman JM, Lee E, Nampally S, Riley D, Sundaram JP, Felix V, Whitty B, Mahurkar A, Wortman J, White O, Angiuoli SV: **Ergatis: a web interface and scalable software system for bioinformatics workflows.** *Bioinformatics* 2010, **26**:1488-1492.
16. Goecks J, Nekrutenko A, Taylor J, Galaxy Team T: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**:R86.
17. Radetzki U, Leser U, Schulze-Rauschenbach SC, Zimmermann J, Lüssem J, Bode T, Cremers AB: **Adapters, shims, and glue–service interoperability for in silico experiments.** *Bioinformatics* 2006, **22**:1137-1143.
18. Lin C, Lu S, Fei X, Pai D, Hua J: **A Task Abstraction and Mapping Approach to the Shimming Problem in Scientific Workflows.** In *Services Computing, IEEE International Conference on. Volume 0.* Los Alamitos, CA, USA: IEEE Computer Society; 2009:284-291.
19. Kane D, Hohman M, Cerami E, McCormick M, Kuhlman K, Byrd J: **Agile methods in biomedical software development: a multi-site experience report.** *BMC Bioinformatics* 2006, **7**:273.
20. Taura K: **Grid Explorer: A Tool for Discovering, Selecting, and Using Distributed Resources Efficiently.** *IPSJ SIG Technical Report* 2004, **2004-HPC-099**:235-240.
21. **Python Programming Language.** [http://www.python.org/].
22. **Ruby Programming Language.** [http://www.ruby-lang.org/].
23. Goodstadt L: **Ruffus: a lightweight Python library for computational pipelines.** *Bioinformatics* 2010, **26**:2778-2779.
24. Cieslik M, Mura C: **A lightweight, flow-based toolkit for parallel and distributed bioinformatics pipelines.** *BMC Bioinformatics* 2011, **12**:61.
25. Cunningham HC: **A little language for surveys: Constructing an internal DSL in Ruby.** *Proceedings of the 46th Annual Southeast Regional Conference on XX, ACM-SE 46* 2008, 282-287.
26. **Rake.** [http://rake.rubyforge.org/].
27. **Pwrake.** [https://github.com/masa16/pwrake].
28. Tanaka M, Tatebe O: **Pwrake: a parallel and distributed flexible workflow management tool for wide-area data intensive computing.** *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing* New York, NY, USA: ACM; 2010, 356-359.
29. **Graphviz.** [http://graphviz.org/].
30. Tatebe O, Hiraga K: **Gfarm Grid File System.** *New Generat Comput* 2010, **28**:257-275.
31. 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**:1297-1303.
32. 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.
33. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R: **Dindel: Accurate indel calls from short-read data.** *Genome Res* 2010.
34. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061-1073.
35. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T: **Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing.** *Nat Genet* 2010, **42**:931-936.
36. The International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437**:1299-1320.
37. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-1760.
38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078-2079.
39. **Picard.** [http://picard.sourceforge.net/].
40. Gropp W, Lusk E, Doss N, Skjellum A: **A high-performance, portable implementation of the MPI message passing interface standard.** *Parallel Comput* 1996, **22**:789-828.
41. Mishima H, Lidral AC, Ni J: **Application of the Linux cluster for exhaustive window haplotype analysis using the FBAT and Unphased programs.** *BMC Bioinformatics* 2008, **9(Suppl 6)**:S10.
42. Dean J, Ghemawat S: **MapReduce: simplified data processing on large clusters.** *Commun ACM* 2008, **51**:107-113.
43. Aerts J, Law A: **An introduction to scripting in Ruby for biologists.** *BMC Bioinformatics* 2009, **10**:221.
44. Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T: **BioRuby: Bioinformatics software for the Ruby programming language.** *Bioinformatics* 2010, btq475.
45. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P, De Roure D: **myExperiment: a repository and social network for the sharing of bioinformatics workflows.** *Nucleic Acids Res* 2010, **38**:W677-W682.
46. **Wf4ever.** [http://www.wf4ever-project.org/].

doi:10.1186/1756-0500-4-331
Cite this article as: Mishima *et al.*: **Agile parallel bioinformatics workflow management using Pwrake.** *BMC Research Notes* 2011 **4**:331.

Eur Respir J 2011; 38: 861–869
DOI: 10.1183/09031936.00143610

Surfactant protein C G100S mutation causes familial pulmonary fibrosis in Japanese kindred

S. Ono*,#,§§, T. Tanaka¶,§§, M. Ishida¶, A. Kinoshita*, J. Fukuoka+, M. Takaki¶, N. Sakamoto§, Y. Ishimatsu§, S. Kohno§, T. Hayashif, M. Senba, M. Yasunami##, Y. Kubo¶¶, L.M. Yoshida¶, H. Kubo++, K. Ariyoshi¶, K. Yoshiura* and K. Morimoto¶**

ABSTRACT: Several mutations in the surfactant protein C (SP-C) gene (*SFTPC*) have been reported as causing familial pulmonary fibrosis (FPF). However, the genetic background and clinical features of FPF are still not fully understood.

We identified one Japanese kindred, in which at least six individuals over three generations were diagnosed with pulmonary fibrosis. We examined the patients radiologically and histopathologically and sequenced their *SFTPC* and *ABCA3* genes. We also established a cell line stably expressing the mutant gene.

All the patients had similar radiological and histopathological characteristics. Their histopathological pattern was that of usual interstitial pneumonia, showing numerous fibroblastic foci even in areas without abnormal radiological findings on chest high-resolution computed tomography. No child had respiratory symptoms in the kindred. Sequencing of *SFTPC* showed a novel heterozygous mutation, c.298G>A (G100S), in the BRICHOS domain of proSP-C, which co-segregated with the disease. However, in the *ABCA3* gene, no mutation was found. *In vitro* expression of the mutant gene revealed that several endoplasmic reticulum stress-related proteins were strongly expressed.

The mutation increases endoplasmic reticulum stress and induces apoptotic cell death compared with wild-type SP-C in alveolar type II cells, supporting the significance of this mutation in the pathogenesis of pulmonary fibrosis.

KEYWORDS: Endoplasmic reticulum stress, familial pulmonary fibrosis, mutation, surfactant protein C

Familial pulmonary fibrosis (FPF) is characterised by cases of idiopathic interstitial pneumonia in two or more first-degree relatives [1]. MARSHALL *et al.* [1] estimated that familial cases account for 0.5–2.2% of all individuals with idiopathic pulmonary fibrosis (IPF). Several kindreds with FPF have been reported, and the familial form is likely to be transmitted in an autosomal dominant inheritance mode [1–3]. Recent studies have revealed that several cases of FPF are associated with mutations in *SFTPC* [4, 5]. *SFTPC*, located at 8p21.3, has six exons and encodes the hydrophobic peptide surfactant protein C (SP-C). The first reported *SFTPC* mutation, IVS4+1G>A, located at the first base of intron 4, disrupted the donor splice site and resulted in the skipping of exon 4 and the deletion of 37 amino acids from the C-terminal region of the proprotein of SP-C (proSP-C) [6]. 26 *SFTPC* mutations have since been identified [6–19] (online supplementary table 1), all of which are heterozygous mutations in affected individuals. However, only a few reports have described familial cases including several affected individuals (table 1) [6–11].

SP-C is synthesised as a 197-amino acid proSP-C, which undergoes multiple processing steps to form mature SP-C. It is finally released into the alveoli in association with other surfactant proteins and phospholipids [4, 5, 20]. Mature SP-C, consisting of 35 amino acids corresponding to Phe24–Leu58 of proSP-C, is encoded within exon 2 of *SFTPC* and is stored in the lamellar body, from where it is secreted into the alveolar space. In the

AFFILIATIONS
*Depts of Human Genetics,
#Psychiatry, Nagasaki University Graduate School of Biomedical Sciences,
¶Depts of Clinical Medicine,
**Pathology,
##Immunogenetics, and
¶¶Preventive Medicine and AIDS Research, Institute of Tropical Medicine, Nagasaki University,
§Second Dept of Internal Medicine, Nagasaki University School of Medicine,
fDept of Pathology, Nagasaki University Hospital, Nagasaki,
+Dept of Surgical Pathology, Toyama University Hospital, Toyama, and
++Dept of Advanced Preventive Medicine for Infectious Disease, Tohoku University Graduate School of Medicine, Sendai, Japan.
§§These authors contributed equally to this study.

CORRESPONDENCE
K. Morimoto
1-12-4 Sakamoto
Nagasaki 852-8523
Japan
E-mail: komorimo@nagasaki-u.ac.jp

Received:
Sept 09 2010
Accepted after revision:
March 10 2011
First published online:
Aug 04 2011

European Respiratory Journal
Print ISSN 0903-1936
Online ISSN 1399-3003

This article has supplementary material accessible from www.erj.ersjournals.com

TABLE 1 Published surfactant protein C mutations found in large families

Mutation	First author [ref.]	Families with ILD n	Pathology
Met71Val	VAN MOORSEL [11]	1	UIP/DIP adults
Ile73Thr	CAMERON [8]	3	NSIP
	ABOU TAAM [10]	1	Unspecified
	VAN MOORSEL [11]	3	UIP adults NSIP/DIP child
IVS4+1,G>A	NOGEE [6]	1	NSIP child DIP/UIP adults
IVS4+2,T>C	VAN MOORSEL [11]	1	NSIP/DIP children UIP adults
Leu188Gln	THOMAS [7]	1	NSIP children DIP/UIP adults
Cys189Tyr	GUILLOT [9]	1	NSIP
Leu194Pro	GUILLOT [9]	1	NSIP

ILD: interstitial lung disease; UIP: usual interstitial pneumonia; DIP: desquamative interstitial pneumonia; NSIP: nonspecific interstitial pneumonia.

lung, proSP-C is expressed only in alveolar type II epithelial cells. The N-terminus of proSP-C is in the cytosol, with the mature SP-C domain anchoring it in the membrane [4, 5, 20]. Furthermore, proSP-C contains a domain known as BRICHOS, which is thought to be involved in proteolytic processing and protecting the peptide from aggregation [21], corresponding to residues Phe94–Ile197 in the C-terminal domain of proSP-C. About three-quarters of all mutations that have been reported in *SFPTC* from interstitial lung diseases are in the BRICHOS domain. It has been reported that a BRICHOS mutant protein increased the amount of insoluble aggregates and resulted in apoptosis following an ER stress response [22].

The current study investigated the clinical features of one Japanese FPF kindred with a heterozygous mutation, G100S, in the BRICHOS domain of proSP-C (SP-C^{G100S}).

MATERIALS AND METHODS

Subjects

Pedigree and DNA samples

The family we encountered is shown in figure 1a. Patient IV-1 was the proband, a Japanese female who was referred to our hospital at the age of 18 yrs for further assessment of an abnormal shadow in the lung field that was noticed at a school medical health check. Patient IV-2 is a younger brother of the proband (fig. 1a). In a routine preoperative chest radiological examination (orthopaedic surgery for congenital dysplasia of the femur), abnormal chest shadows were noticed, and further analysis was performed in our department after the operation at the age of 16 yrs. Patient IV-3 is a younger sister of the proband. After the family history of the proband was taken, we assessed patient IV-3's chest by radiographic examination at the age of 14 yrs, in accordance with her and her father's requests. The three individuals in generation IV were delivered without any problem and showed normal development. None of them had histories of coughing, shortness of breath or environmental exposures, and all were free from other respiratory symptoms. Patient III-3 was the proband's mother. She had been diagnosed with IPF at age 34 yrs; she died at age 41 yrs from lung fibrosis. Patient II-4 is a grandmother of the proband who developed a cough at age 63 yrs and was diagnosed with interstitial pneumonia. Patient III-1 is an aunt of the proband. She had no respiratory symptoms. After the family history was taken, we performed a chest radiological examination on patient III-1, at her request. It was supposed that three more family members, I-1, II-1 and II-2, died from lung disease at ages between 35 and 45 yrs following a few years of illness.

Written informed consent was obtained from the patients and their family members before they participated in this study. Genetic counselling was given to patients before and after genetic analyses. Genomic DNA was extracted from individuals' peripheral blood (II-3, II-4, III-1, III-2, IV-1, IV-2 and IV-3) or from formalin-fixed paraffin-embedded lung tissue (III-3) using a QIAamp DNA Mini Kit (Qiagen, Hilden, Germany). This study was approved by the institutional review boards of Nagasaki University (Nagasaki, Japan).

Lung biopsy and lung histopathology

For histopathological diagnosis, a lung biopsy was performed by video-assisted thoracic surgery (VATS) under general anaesthesia. Tissue sections were prepared from formalin-fixed paraffin-embedded samples. Haematoxylin and eosin-stained sections were prepared following conventional procedures. Pathology slides were observed by two trained pulmonary pathologists.

Mutation analysis

We performed PCR-based mutation analysis of *SFTPC* (National Center for Biotechnology Information (NCBI) Reference Sequence NM_003018.3) from eight specimens, composed of six affected individuals (II-4, III-1, III-3, IV-1, IV-2 and IV-3) and two unaffected individuals (II-3 and III-2). Subsequently, we also sequenced ATP-binding cassette, sub-family A member 3 (*ABCA3*; NM_001089), which was postulated to be a gene that modifies the disease severity of FPF caused by *SFTPC* mutations [23]. All exons and intron–exon boundaries of the two genes were sequenced on a 3130xl automated sequencer (Applied Biosystems, Foster City, CA, USA) using BigDye Terminator version 3.1 (Applied Biosystems). DNA sequences were analysed using Variant Reporter and Sequencing Analysis (Applied Biosystems). Genomic sequences were obtained from the University of California, Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu/; assembly: March 2006; NCBI36/hg18). PCR primers were designed with the assistance of Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3.cgi). Primer sequences are available from the authors on request.

In silico *analysis*

The determination of whether an amino acid substitution is a recognised polymorphism was carried out using the dbSNP database (www.ncbi.nlm.nih.gov/SNP/). Predicted protein functions caused by an amino acid substitution were examined using PolyPhen (http://genetics.bwh.harvard.edu/pph/) and SIFT (http://sift.jcvi.org/). Comparisons of genomic alignments of human and other species were accessed using online software, the Evolutionary Conserved Regions (ECR) browser (http://ecrbrowser.dcode.org/) and the UCSC genome browser.

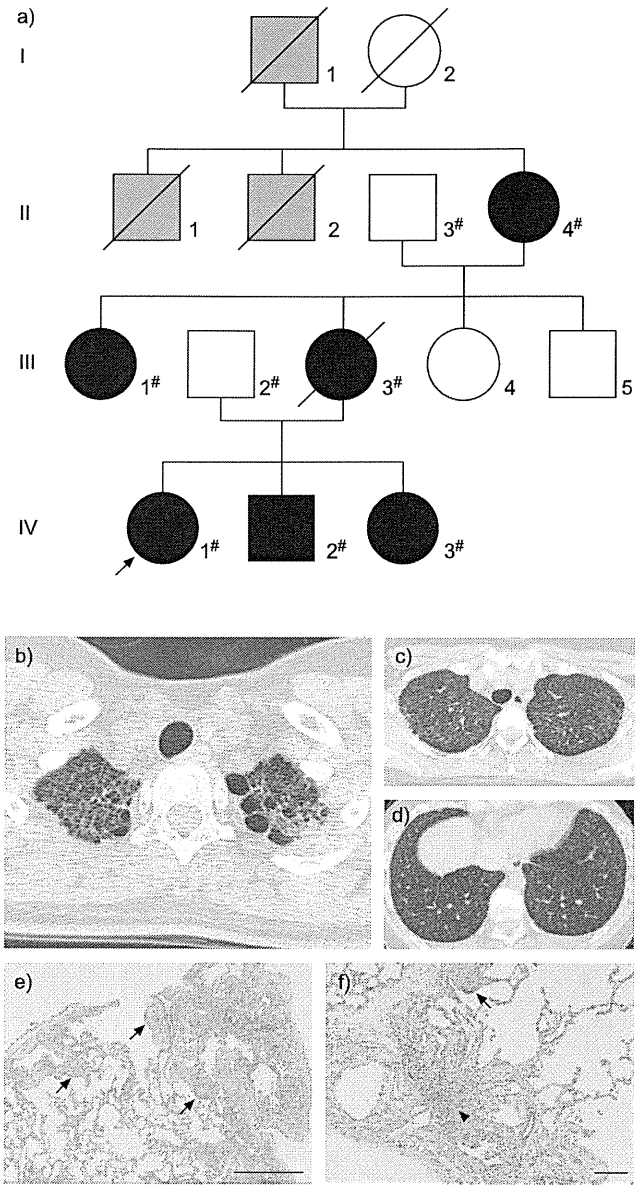


FIGURE 1. Pedigree of a family with familial pulmonary fibrosis (FPF), and radiological and histopathological findings of the proband. a) Pedigree of the family with FPF. Squares: males; circles: females; black: individuals diagnosed with pulmonary fibrosis; grey: individuals who had died due to respiratory failure, but about whom detailed information was not available; #: individuals whose DNA was available and used in direct sequencing; arrow: proband. b) High-resolution computed tomography (HRCT) image of the proband. Reticulonodular opacity, predominantly in both upper lung fields, and intralobar opacity in the subpleural area were observed in the HRCT image from the proband. No honeycombing lesions could be seen. c, d) Magnified CT images of the proband. e and f) Haematoxylin and eosin-stained tissue samples from the proband (right lung S8). Haematoxylin and eosin staining revealed a usual interstitial pneumonia pattern, including patchy peripheral accentuated fibrosis, marked fibroblastic foci (arrows), smooth muscle hyperplasia (arrow head) and abrupt changes to adjacent normal lung areas. Scale bars: e) 1 mm; f) 200 μm. Biopsies were performed from right lung S2 and S8. Pathological findings were similar in S2 and S8.

Functional analysis of mutant protein

SP-C cDNA constructs

A cDNA encoding the full-length human SP-C (SP-C^{1-197}) was cloned into the pcDNA3.1 vector (Invitrogen, Carlsbad, CA) to generate SP-C^{1-197}/pcDNA3.1. A QuikChange® II Site-Directed Mutagenesis Kit (Stratagene, Santa Clara, CA, USA) was used to generate mutant SP-C^{G100S} in a single PCR with two primers: 5′-ATCGGCTCCACTAGCCTCGTGGTGT-3′ (forward) and 5′-ACACCACGAGGCTAGTGGAGCCGAT-3′ (reverse). The mutation site is underlined.

Cell culture and transfection

A human embryonal kidney (HEK) 293T cell line was obtained from the American Type Culture Collection (ATCC; Manassas, VA, USA) and cultured in Dulbecco's modified Eagle's medium (Gibco, Carlsbad, CA, USA) at 37 °C in 5% CO_2. The culture media were supplemented with 10% fetal bovine serum (Biofluids, Rockville, MD, USA). A549 cells (ATCC) overexpressing wild-type (SP-C^{WT}) or mutant proSP-C were constructed as follows. HEK293T cells were transiently transfected with murine leukaemia virus gag–pol (2 μg) (TaKaRa Bio, Shiga, Japan), proSP-C-encoding retroviral vector (2 μg), and VSV-G expression plasmids (2 μg), which were obtained from L. Chang through the AIDS Research and Reference Reagent Program (Division of AIDS, National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA) [24] using the FuGene HD reagent (30 μL) (Roche Applied Science, Mannheim, Germany). The cells were washed 24 h after transfection and cultured for 24 h in fresh medium. Culture supernatant of the transfected cells was inoculated into A549 cells. The inoculated cells were selected by puromycin (2.5 $\mu g \cdot mL^{-1}$). The puromycin-resistant cell pool was utilised in this study.

RNA isolation and real-time RT-PCR

Total RNA from the stably transfected A549 cells was isolated using a FastPure RNA Kit (TaKaRa Bio) and reverse transcribed into cDNA using a PrimeScript RT Reagent Kit with gDNA Eraser (TaKaRa Bio). We performed real-time quantitative RT-PCR using Thunderbird SYBR qPCR Mix reagent (Toyobo, Osaka, Japan). PCR amplification was run on a LightCycler 480 Real-Time PCR system (Roche Diagnostics, Mannheim, Germany). All samples were measured in triplicate.

Western blot analysis

Cells were solubilised in RIPA buffer with PhosSTOP Phosphatase inhibitor cocktail (Roche Applied Science). Cells treated with proteasome inhibitor MG-132 (Merck Ltd, Lutterworth, UK) for 16 h were also solubilised in the same manner. Total protein extracts were separated by 5–15% Tris-HCl gel (BioRad Laboratories, Richmond, CA, USA) electrophoresis and transferred to polyvinylidene fluoride membranes. The membranes were blocked in blocking buffer (1 × PBS, 0.1% Tween-20 with 5% weight/volume nonfat dry milk) for 1 h at room temperature and incubated with primary antibodies at 4°C overnight. After washing in 1 × PBS with 0.1% w/v Tween-20, membranes were incubated with horseradish peroxidase-linked secondary antibodies for 1 h at room temperature. Detection was performed by enhanced chemiluminescence with ECL-Plus (GE Health Care, Little Chalfont, UK). Primary antibodies to BiP, IRE1α and cleaved caspase-3 were purchased from Cell Signaling Technology (Danvers, MA, USA). Anti-phosphorylated PERK

(phospho-PERK) antibody was purchased from Santa Cruz Biotechnology (Santa Cruz, CA, USA). β-actin was measured as a loading control for each sample using anti-β-actin antibodies (Santa Cruz Biotechnology).

Statistics

Data are presented as mean±SE. The *t*-statistic was used to determine significant differences between two groups. One-way ANOVA was used to determine significant differences among groups.

RESULTS

Clinical presentation of patients

High-resolution computed tomography (HRCT) findings of the proband, patient IV-1, revealed a reticulonodular shadow and intralobular fine linear opacity predominantly in both upper lung fields. Centrilobular micronodule lesions were observed mainly in subpleural lesions (fig. 1b–d). The HRCT findings of patients II-4, III-1, III-3, IV-2 and IV-3 are presented in online supplementary figure 1a and are similar to those found in the proband. All affected individuals showed similar radiological findings, *i.e.* upper lung field dominant shadow. Additionally, IV-1, IV-2 and IV-3 showed moderate cystic changes, mainly in the upper lobes, as shown in a previous report of adult FPF [11].

VATS lung biopsy was performed for diagnosis and pathological assessment. Haematoxylin and eosin-stained samples from the proband showed features of the usual interstitial pneumonia (UIP) pattern with marked fibroblastic foci and mild infiltration of lymphoid cells (fig. 1e and f). In addition, mild-to-moderate airway-centred fibrosis/inflammation, along with peribronchiolar metaplasia, were observed. No granulomas were seen. Interestingly, all histological samples from the patients (III-3, IV-2 and IV-3) showed a similar UIP pattern (online supplementary fig. 2).

The clinical findings and information are summarised in table 2. Briefly, for the proband and her siblings, serum biomarkers, pulmonary function and respiratory condition were almost normal, and no airway inflammation was observed in their bronchoalveolar lavage fluid (BALF). Because they had kept pet birds in their home, we measured serum antibodies to avian antigen, but those were negative for these siblings. Thus, chronic hypersensitivity pneumonitis was clinically ruled out. Based on radiopathological findings and family history, familial interstitial pneumonia was diagnosed.

Mutation analysis and in silico *analysis*

Two genes, *SFTPC* and *ABCA3*, were analysed. We detected a base alteration, c.298G>A, in exon 3 of *SFTPC* causing a GGC-to-AGC change that results in a glycine-to-serine change at codon 100 (fig. 2a). This variant segregated with the disease in this family (fig. 2b) and was not present among 576 ethnically matched control alleles. The ECR browser and the UCSC genome browser indicated that codon 100 of *SFTPC* is conserved among mammals (fig. 2c). Furthermore, *in silico* analysis using SIFT and Polyphen predicted a damaging effect on the protein by this one amino acid change (position-specific independent counts score 1.722; SIFT score 0.03).

Expression of proSP-C in A549 cells

To prove comparable expression of proSP-C, we performed western blotting of cell lysates of SP-C^{WT} and SP-C^{G100S} stably expressed A549 cells. The amount of proSP-C was increased in A549 cells stably expressing SP-C^{G100S} compared with those stable expressing SP-C^{WT} (fig. 3a, b). However, the expression levels of SP-C mRNA from these two cell pools assessed by real-time quantitative RT-PCR were equivalent (fig. 3d).

SP-C^{G100S} causes endoplasmic reticulum stress, resulting in apoptosis

We performed western blotting analysis to detect the expression of proSP-C, BiP, phospho-PERK, IRE1α and cleaved caspase-3 to determine whether the expression of the SP-C^{G100S} induces endoplasmic reticulum (ER) stress in epithelial cells compared with SP-C^{WT}. The activations of BiP, IRE1α and cleaved caspase-3 were increased in mutant cells compared with wild-type cells (fig. 3c and e). After MG-132 treatment, A549 cells stably expressing SP-C^{G100S} showed increases in expression BiP, phospho-PERK, IRE1 and cleaved caspase-3 that significantly exceeded the increases seen in A549 cells stably expressing SP-C^{WT} (fig. 3c and f).

DISCUSSION

In the present study, we have described a novel pathogenic *SFTPC* variant, which is associated with FPF in a Japanese kindred who had abnormal HRCT findings at ages ranging from the mid-second to the fifth decade of life. This pedigree included six individuals with similar radiological findings and histopathological characteristics of the UIP pattern. Notably, all the patients were asymptomatic until they were age at least 15 yrs, and there was no child with respiratory symptoms. Furthermore, we also verified that expression of the mutant protein, SP-C^{G100S}, resulted in caspase-3 activation following the induction of ER stress.

Glycine at codon 100 of *SFTPC*, which was mutated to serine in this kindred, is in the BRICHOS domain of proSP-C. This mutation is novel and is the first reported pathogenic mutation of *SFTPC* in an Asian kindred, proving that pulmonary fibrosis caused by *SFTPC* mutations is a worldwide phenomenon. Recent reports showed that the BRICHOS domain of proSP-C has chaperone-like properties that prevent the transmembrane region of proSP-C from aggregating. Mutations of this region in proSP-C triggered induction of intracellular aggregate formation, ER stress and accumulation in endosomal–lysosomal compartments [22, 25, 26]. To further characterise the mutant protein SP-C^{G100S}, we showed that unfolded protein response (UPR) proteins, including BiP (chaperone proteins), phospho-PERK and IRE1α (proximal sensor for UPR), were upregulated in A549 cells stably transformed with SP-C^{G100S}, eventually resulting in apoptotic cell death. These results are consistent with previous observations in several studies of other mutations in the BRICHOS domain, including SP-$C^{\Delta exon4}$ and SP-C^{L188Q} [22, 26, 27, 28]. Recently, SISSON *et al.* [29] reported that targeted injury of type II alveolar epithelial cells induced pulmonary fibrosis in mice [29]. Collectively, these observations lead us to conclude that SP-C^{G100S} is a pathogenic mutation leading to cell death, which leads to pulmonary fibrosis. Categorising *SFTPC* mutations inducing lung fibrosis by functional analysis of the mutant protein might help in tailoring treatment for IPF patients. ROSEN and WALTZ [16] have reported that hydroxychloroquine was

TABLE 2 Patient profile and laboratory data

	Patient					
	II-4	III-1	III-3#	IV-1	IV-2	IV-3
Sex	Female	Female	Female	Female	Male	Female
Current age yrs	68	46	41(died)	18	16	14
Age of diagnosis yrs	66	46	34	18	16	14
Age of first evidence yrs	57	44	34	18	16	14
Age of first symptoms yrs	63	None	34	None	None	None
Serum biomarker						
KL-6 $U \cdot mL^{-1}$ (normal range <500)	1560	386	NA	245	309	332
LDH $IU \cdot L^{-1}$ (normal range 119–229)	303	147	166	161	141	144
Pulmonary function tests						
VC % pred (normal range >80)	42.5	101.9	65.3	72.2	85	96.6
FEV1 % pred (normal range >70)	92.9	92.8	83.3	84.1	90.3	85
*D*LCO % pred (normal range >80)	38.5	72.2	NA	69.3	NA	65.2
Blood gas analysis (room air)						
P_{a,O_2} Torr (normal range 75–100)	68.3¶	83.7	90.6	113	109	111
$P_{(A\text{-}a),O_2}$ Torr (normal range <10)	24.2¶	16.2	6.1	-8.5	-14	-10.8
BAL						
Cell count $10^5 \cdot mL^{-1}$	1.21	NA	3.85	2.4	2	1.4
Alveolar macrophages %	54.2	NA	80	90	86	91
Lymphocytes %	10.1	NA	17.3	7.5	12	5.8
Neutrophils %	34.5	NA	1.1	2.5	1	2.4
Eosinophils %	1.2	NA	1.6	0	1	0.8
CD4/CD8 ratio	0.25	NA	0.6	1.7	1.6	1.5
Histological pattern	UIP	NA	UIP	UIP	UIP	UIP

LDH: lactate dehydrogenase; VC: vital capacity; % pred: % predicted; FEV1: forced expiratory volume in 1 s; *D*LCO: diffusing capacity of the lung for carbon monoxide; P_{a,O_2}: arterial oxygen tension; $P_{(A\text{-}a),O2}$: alveolar–arterial oxygen tension difference; BAL: bronchoalveolar lavage; NA: not available: UIP: usual interstitial pneumonia. #: data from patient III-3 are based on those from the first diagnosis; ¶: data from the time of first diagnosis.

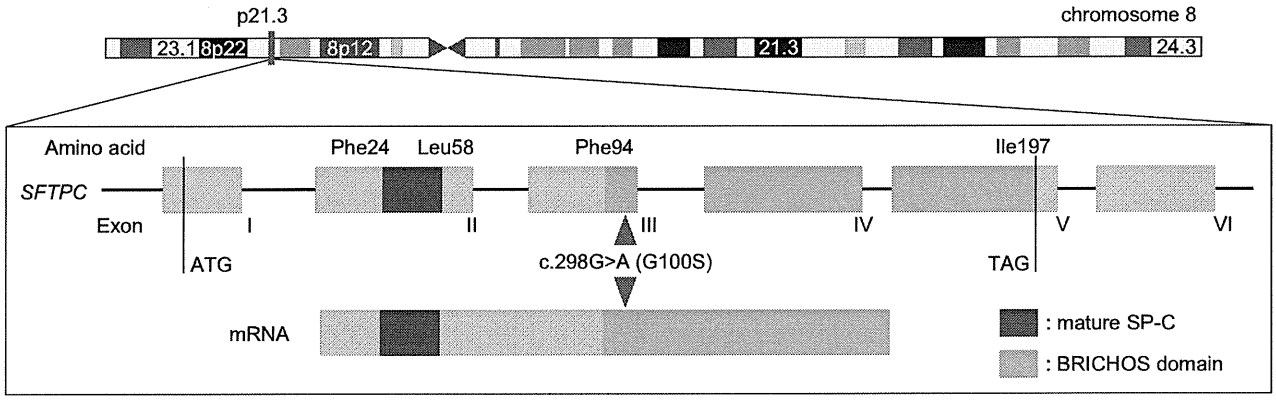
useful in treating a case of interstitial lung disease in a child with an *SFTPC* mutation in the BRICHOS domain [16]. They predicted that hydroxychloroquine caused inhibition of the intracellular processing of proSP-C, thereby reducing the dominant negative effect elicited by mutant proSP-C. It is possible that the suitability of a treatment for interstitial lung diseases with *SFTPC* mutations depends upon the location of the mutation. Hydroxychloroquine might be a suitable treatment for our cases with SP-C^{G100S}.

Intriguingly, A549 cells transfected with SP-C^{G100S} contained more proSP-C protein than cells expressing SP-C^{WT}, despite the SP-C mRNA levels being equivalent. This result was inconsistent with the previous report by BRIDGES *et al.* [27], which showed that the mutant protein of SP-$C^{\Delta exon4}$ was barely detectable in contrast to the wild-type protein in the stably expressing HEK293 cell lines [27]. We also confirmed the minimal accumulation of proSP-C^{G100S} when HEK293 cells were transfected with SP-C^{G100S} (data not shown). Therefore, the observed difference is likely to be due to the difference in cell origins, not the features of the mutations. Our experiments also showed that the expression of the 26-kDa isoform of the mutant SP-C^{G100S} was weaker than that of wild type in A549 cells. Formation of the 26-kDa isoform requires palmitoylation of proSP-C [30] and a 21-kDa isoform is considered to be the proprotein of pre-proteolytic processing [5, 25]. Taken together, we speculate that the palmitoylation process in the mutant proSP-C^{G100S} was impaired and unpalmitoylated proprotein accumulated in human alveolar epithelial cells (A549). We believe that the slow degradation of unpalmitoylated proprotein in A549 cells is a better reflection of the process actually taking place in the patients presented in this report.

To date, more than 20 mutations have been described in *SFTPC*. Although studies of *SFTPC* mutations have focused on cases of children with interstitial lung diseases, there have been a few studies focusing on pedigrees with adult FPF [11, 31]. They found five kindreds with *SFTPC* mutations, including two new mutations, M71V and IVS4+2T>C, in adult FPF patients. They showed histopathological patterns of UIP and non-classifiable HRCT patterns with reticulonodular opacity and multiple lung cysts in combination with ground-glass opacities or diffuse lung involvement on chest HRCT. The present study, similarly focusing on a pedigree with adult FPF, highlighted some outstanding characteristics of this kindred with SP-C^{G100S}. Our patients presented with a histopathological pattern of UIP and the HRCT findings had features of reticulonodular opacity and multiple lung cysts. These findings, however, were seen predominantly in the upper lobes. In particular a small number of lung cysts were present only at the apex, a feature that was inconsistent with the above report.

Interestingly, the age of phenotypic appearance (*i.e.* the appearance of positive radiological and histopathological findings, even

a)



b)

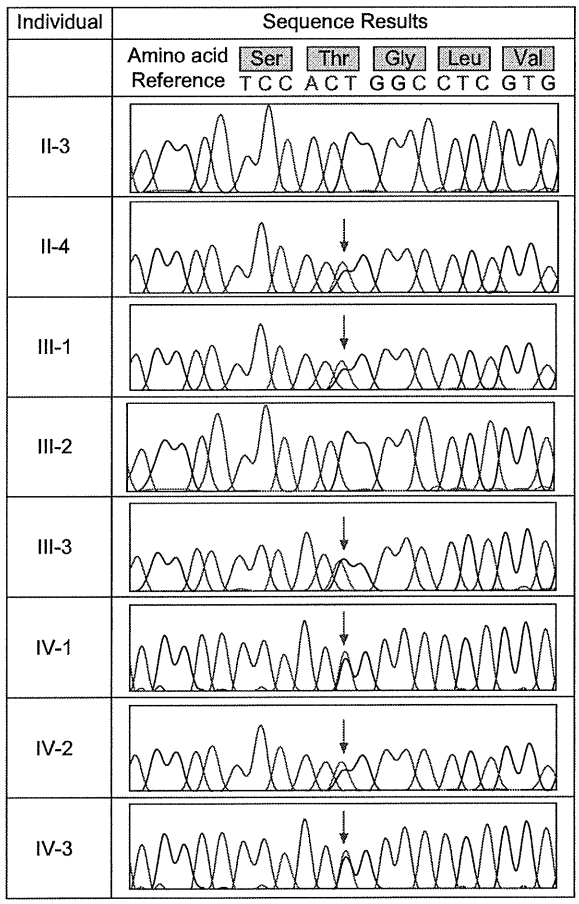


FIGURE 2. Amino acid substitution identified in surfactant protein (SP)-C in individuals with familial pulmonary fibrosis. a) Location of *SFTPC*, the gene encoding SP-C. Red triangle indicates the location of the c.298G>A (G100S) mutation of *SFTPC*, which is in the BRICHOS domain. b) Results of direct DNA sequencing in eight individuals. Red arrows indicate the location of the nonsynonymous substitution (c.298 G>A). c) The highly conserved orthologous protein sequences of SP-C across eight species of mammal. The area surrounded by the red line indicates the location of codon 100 of *SFTPC*.

c)

DNA sequence (human)	A T C	G G C	T C C	A C T	G G C	C T C	G T G	G T G	T A T
Human	I	G	S	T	G	L	V	V	Y
Chimpanzee	I	G	S	T	G	L	V	V	Y
Gorilla	I	G	S	T	G	L	V	V	Y
Mouse	I	G	S	T	G	I	V	V	Y
Rat	I	G	S	T	G	I	V	L	Y
Cow	I	G	S	T	G	T	V	V	Y
Dog	I	G	S	T	G	I	V	V	Y
Opossum	I	G	S	S	G	T	V	V	Y

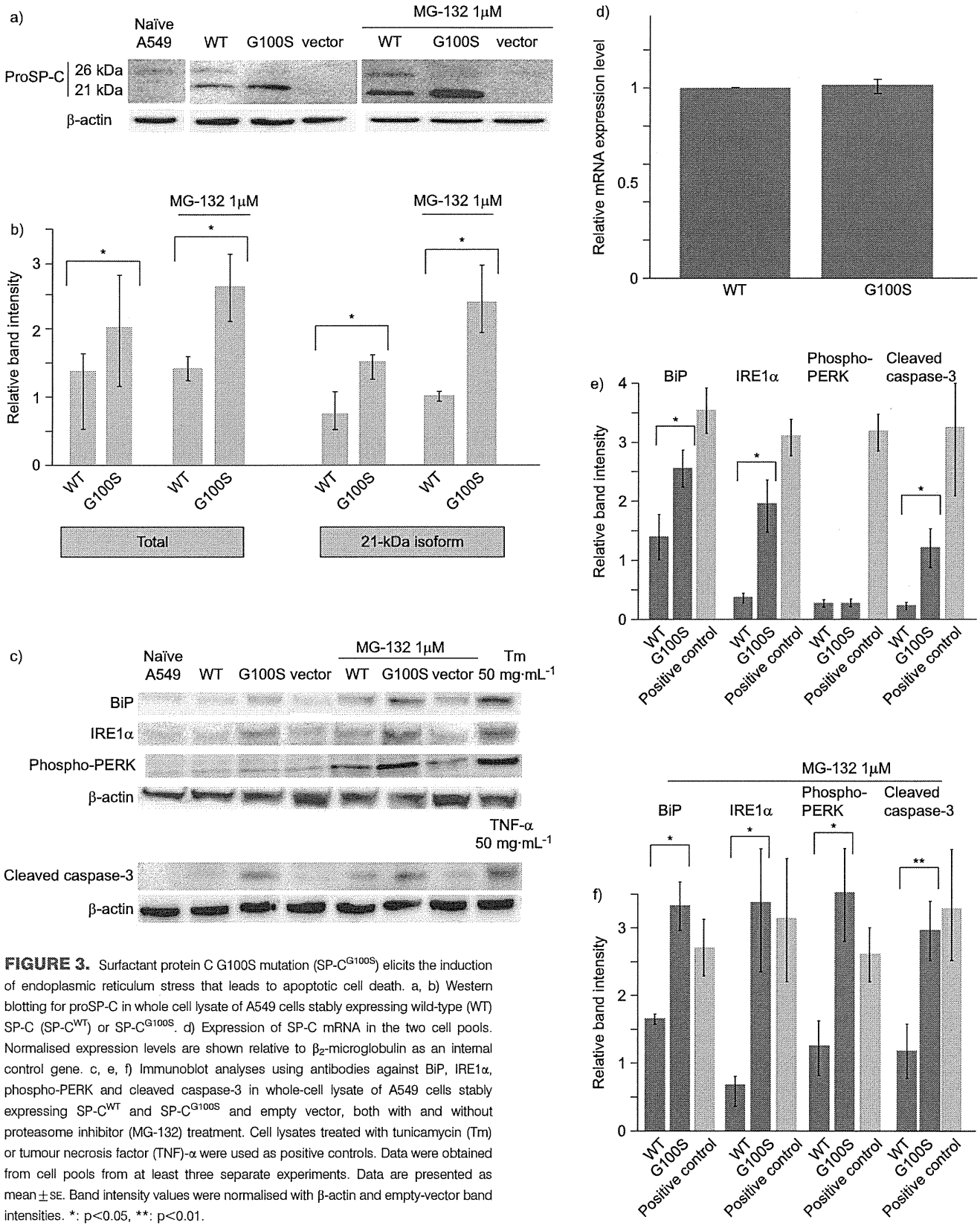


FIGURE 3. Surfactant protein C G100S mutation (SP-C^{G100S}) elicits the induction of endoplasmic reticulum stress that leads to apoptotic cell death. a, b) Western blotting for proSP-C in whole cell lysate of A549 cells stably expressing wild-type (WT) SP-C (SP-C^{WT}) or SP-C^{G100S}. d) Expression of SP-C mRNA in the two cell pools. Normalised expression levels are shown relative to β_2-microglobulin as an internal control gene. c, e, f) Immunoblot analyses using antibodies against BiP, IRE1α, phospho-PERK and cleaved caspase-3 in whole-cell lysate of A549 cells stably expressing SP-C^{WT} and SP-C^{G100S} and empty vector, both with and without proteasome inhibitor (MG-132) treatment. Cell lysates treated with tunicamycin (Tm) or tumour necrosis factor (TNF)-α were used as positive controls. Data were obtained from cell pools from at least three separate experiments. Data are presented as mean±SE. Band intensity values were normalised with β-actin and empty-vector band intensities. *: $p<0.05$, **: $p<0.01$.

if presymptomatic) of all six patients was school-age or older, not at the neonatal or infancy stage, as is commonly reported for other *SFTPC* mutations. These observations caused us to make two speculations. Firstly, SP-C^{G100S} is directly involved in the severity of the disease; the late onset and slow progress of respiratory symptoms might be unique to this mutation. However, the *SFTPC* mutation may have pleiotropic effects across different families, so other families with SP-C^{G100S} mutation need to be investigated carefully to confirm more characteristics of this mutation. IL-8 production in IPF patients is increased [32], but BALF findings of the proband and her siblings showed no inflammatory cell response (table 2) and no IL-8 response (data not shown). This could be related to the relatively modest radiological change and late onset. SP-C^{G100S} might lead to chronic cell death, but it does not induce acute inflammation, eventually resulting in respiratory symptoms and progression to lung fibrosis. Second, any genetic modifier shared with their patients might suppress the progression of the disease caused by SP-C^{G100S}, indicating an indirect involvement of SP-C^{G100S} in the severity of the disease. BULLARD *et al.* [23] implied that *ABCA3* mutations modified the severity of lung disease associated with *SFTPC* mutations. In the present study, we detected no mutations in *ABCA3*. However, considering the late onset of our patients through three generations, it is likely that inherited genetic and epigenetic factors might have homogeneously and moderately suppressed the cytotoxicity induced by SP-C^{G100S}.

Despite the fact that intralobular reticular opacities were barely observed in the lower lobe on chest CT, histopathological findings of fibrotic changes were found, similar to the findings in the upper lobe, where both radiological and histopathological abnormalities were seen. Recent studies have suggested that fibrotic changes might be present in family members with *SFTPC* mutations who have little evidence of disease [31, 33]. In the present study, histopathological examination revealed a UIP pattern in the lower lobe in which no radiological finding was observed. Several reports have shown that individuals carrying other *SFTPC* mutations, including I73T, have not developed symptoms even in adulthood [8, 34]. These observations suggest that individuals with no clinical symptoms, no radiological findings and no phenotypic appearance, but who carry *SFTPC* mutations, might have pathologically recognisable fibrosis, and their lesions might be progressing slowly.

In conclusion, we have detected a new pathogenic mutation in *SFTPC*. The functional analyses in this study suggest that this mutant protein, SP-C^{G100S}, elicits ER stress leading to apoptotic cell death. Our study indicates that this mutation is pathogenic and caused slow progression of pulmonary fibrosis in this kindred. We could not confirm the reason for this slow progression; it might be a characteristic of SP-C^{G100S} or it might be due to the influence of other genes or epigenetic modifications. Functional understanding of the misfolded SP-C protein is important to determine treatment approaches for FPF, which might help in tailor-made treatment based on genotype.

STATEMENT OF INTEREST

None declared.

ACKNOWLEDGEMENTS

We are grateful to the family members for their participation in this study. We especially thank M. Kitaichi (Dept of Pathology, National Hospital Organisation Kinki-Chuo Chest Medical Centre, Osaka, Japan) for reviewing the pathological diagnosis.

REFERENCES

1 Marshall RP, Puddincombe A, Cookson WO, *et al.* Adult familial cryptogenic fibrosing alveolitis in the UK. *Thorax* 2000; 55: 143–146.
2 Bitterman PB, Rennard SI, Keogh BA, *et al.* Familial idiopathic pulmonary fibrosis. Evidence of lung inflammation in unaffected family members. *N Engl J Med* 1986; 314: 1343–1347.
3 Marney A, Lane KB, Phillips JA, *et al.* Idiopathic pulmonary fibrosis can be an autosomal dominant trait in some families. *Chest* 2001; 120: 1 Suppl., 56S.
4 Nogee LM. Alterations in SP-B and SP-C expression in neonatal lung disease. *Annu Rev Physiol* 2004; 66: 601–623.
5 Beers MF, Mulgeta S. Surfactant protein C biosynthesis and its emerging role in conformational lung disease. *Annu Rev Physiol* 2005; 67: 663–696.
6 Nogee LM, Dunbar AE, Wert SE, *et al.* A mutation in the surfactant protein C gene associated with familial interstitial lung disease. *N Engl J Med* 2001; 344: 573–579.
7 Thomas AQ, Lane K, Phillips J, *et al.* Heterozygosity for a surfactant protein C gene mutation associated with usual interstitial pneumonitis and cellular nonspecific interstitial pneumonitis in one kindred. *Am J Respir Crit Care Med* 2002; 165: 1322–1328.
8 Cameron HS, Somaschini M, Carrera P, *et al.* A common mutation in the surfactant protein C gene associated with lung disease. *J Pediatr* 2005; 146: 370–375.
9 Guillot L, Epaud R, Thouvenin G, *et al.* New surfactant protein C gene mutations associated with diffuse lung disease. *J Med Genet* 2009; 46: 490–494.
10 Abou Taam R, Jaubert F, Emond S, *et al.* Familial interstitial disease with I73T mutation: A mid- and long-term study. *Pediatr Pulmonol* 2009; 44: 167–175.
11 Van Moorsel CH, van Oosterhout MF, Barlo NP, *et al.* SFTPC mutations are the basis of a significant portion of adult familial pulmonary fibrosis in a Dutch cohort. *Am J Respir Crit Care Med* 2010; 182: 1419–1425.
12 Nogee LM, Dunbar AE, Wert S, *et al.* Mutations in the surfactant protein C gene associated with interstitial lung disease. *Chest* 2002; 121: Suppl. 3, 20S–21S.
13 Brasch F, Griese M, Tredano M, *et al.* Interstitial lung disease in a baby with a *de novo* mutation in the SFTPC gene. *Eur Respir J* 2004; 24: 30–39.
14 Hamvas A, Nogee LM, White FV, *et al.* Progressive lung disease and surfactant dysfunction with a deletion in surfactant protein C gene. *Am J Respir Cell Mol Biol* 2004; 30: 771–776.
15 Tredano M, Griese M, Brasch F, *et al.* Mutation of SFTPC in infantile pulmonary alveolar proteinosis with or without fibrosing lung disease. *Am J Med Genet A* 2004; 126A: 18–26.
16 Rosen DM, Waltz DA Hydroxychloroquine, surfactant protein C deficiency. *N Engl J Med* 2005; 352: 207–208.
17 Stevens PA, Pettenazzo A, Brasch F, *et al.* Nonspecific interstitial pneumonia, alveolar proteinosis, and abnormal proprotein trafficking resulting from a spontaneous mutation in the surfactant protein C gene. *Pediatr Res* 2005; 57: 89–98.
18 Poterjoy BS, Vibert Y, Sola-Visner M, *et al.* Neonatal respiratory failure due to a novel mutation in the surfactant protein C gene. *Perinatol* 2010; 30: 151–153.
19 Soraisham AS, Tierney AJ, Amin HJ. Neonatal respiratory failure associated with mutation in the surfactant protein C gene. *Perinatol* 2006; 26: 67–70.
20 Keller A, Eistetter HR, Voss T, *et al.* The pulmonary surfactant protein C (SP-C) precursor is a type II transmembrane protein. *Biochem J* 1991; 277: 493–499.

21 Sanchez-Pulido L, Devos D, Valencia A. BRICHOS: a conserved domain in proteins associated with dementia, respiratory distress and cancer. *Trends Biochem Sci* 2002; 27: 329–332.
22 Mulugeta S, Nguyen V, Russo SJ, *et al.* A surfactant protein C precursor protein BRICHOS domain mutation causes endoplasmic reticulum stress, proteasome dysfunction, and caspase 3 activation. *Am J Respir Cell Mol Biol* 2005; 32: 521–530.
23 Bullard JE, Nogee LM. Heterozygosity for ABCA3 mutations modifies the severity of lung disease associated with a surfactant protein C gene (SFTPC) mutation. *Pediatr Res* 2007; 62: 176–179.
24 Chang LJ, Urlacher V, Iwakuma T, *et al.* Efficacy and safety analyses of a recombinant human immunodeficiency virus type 1 derived vector system. *Gene Ther* 1999; 6: 715–728.
25 Ten Brinke A, van Golde LM, Batenburg JJ. Palmitoylation and processing of the lipopeptide surfactant protein C. *Biochim Biophys Acta* 2002; 1583: 253–265.
26 Nerelius C, Martin E, Peng S, *et al.* Mutations linked to interstitial lung disease can abrogate anti-amyloid function of prosurfactant protein C. *Biochem J* 2008; 416: 201–209.
27 Bridges JP, Xu Y, Na CL, *et al.* Adaptation and increased susceptibility to infection associated with constitutive expression of misfolded SP-C. *J Cell Biol* 2006; 172: 395–407.
28 Mulugeta S, Maguire JA, Newitt JL, *et al.* Misfolded BRICHOS SP-C mutant proteins induce apoptosis *via* caspase-4- and cytochrome c-related mechanisms. *Am J Physiol Lung Cell Mol Physiol* 2007; 293: L720–L729.
29 Sisson TH, Mendez M, Choi K, *et al.* Targeted injury of type II alveolar epithelial cells induces pulmonary fibrosis. *Am J Respir Crit Care Med* 2010; 181: 254–263.
30 Vorbroker DK, Dey C, Weaver TE, *et al.* Surfactant protein C precursor is palmitoylated and associates with subcellular membranes. *Biochim Biophys Acta* 1992; 1105: 161–169.
31 Crossno PF, Polosukhin VV, Blackwell TS, *et al.* Identification of early interstitial lung disease in an individual with genetic variations in ABCA3 and SFTPC. *Chest* 2010; 137: 969–973.
32 Keane MP, Arenberg DA, Lynch JP, *et al.* The CXC chemokines, IL-8 and IP-10, regulate angiogenic activity in idiopathic pulmonary fibrosis. *J Immunol* 1997; 159: 1437–1443.
33 Rosas IO, Ren P, Avila NA, *et al.* Early interstitial lung disease in familial pulmonary fibrosis. *Am J Respir Crit Care Med* 2007; 176: 698–705.
34 Percopo S, Cameron HS, Nogee LM, *et al.* Variable phenotype associated with SP-C gene mutations: fatal case with the I73T mutation. *Eur Respir J* 2004; 24: 1072–1073.

Acknowledgments:s We are greatly indebted to E. J. M. Ladan-Eygenraam, M. van Dijk-Besling, and H. G. Haasnoot-van der Bent for technical support during the study. N.A. Aziz and R.A.C. Roos had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis.

N. Ahmad Aziz, PhD,[1] Hanno Pijl, MD, PhD,[2]
Marijke Frölich, PhD,[3] Ferdinand Roelfsema, MD, PhD,[2]
Raymund A.C. Roos, MD, PhD[1]

[1]Department of Neurology,
[2]Department of Endocrinology and Metabolic Diseases,
[3]Department of Clinical Chemistry,
Leiden University Medical Center,
Leiden, The Netherlands

References

1. Aziz NA, van der Marck MA, Pijl H, Olde Rikkert MG, Bloem BR, Roos RA. Weight loss in neurodegenerative disorders. J Neurol 2008;255:1872–1880.
2. Trujillo ME, Scherer PE. Adipose tissue-derived factors: impact on health and disease. Endocr Rev 2006;27:762–778.
3. Evidente VG, Caviness JN, Adler CH, Gwinn-Hardy KA, Pratley RE. Serum leptin concentrations and satiety in Parkinson's disease patients with and without weight loss. Mov Disord 2001;16:924–927.
4. Fiszer U, Michalowska M, Baranowska B, et al. Leptin and ghrelin concentrations and weight loss in Parkinson's disease. Acta Neurol Scand 2010;121:230–236.
5. Lorefalt B, Toss G, Granerus AK. Weight loss, body fat mass, and leptin in Parkinson's disease. Mov Disord 2009;24:885–890.
6. Signore AP, Zhang F, Weng Z, Gao Y, Chen J. Leptin neuroprotection in the CNS: mechanisms and therapeutic potentials. J Neurochem 2008;106:1977–1990.
7. Aziz NA, Pijl H, Frolich M, van der Graaf AW, Roelfsema F, Roos RA. Leptin secretion rate increases with higher CAG repeat number in Huntington's disease patients. Clin Endocrinol (Oxf) 2010;73:206–211.

Mutation and Copy Number Analysis in Paroxysmal Kinesigenic Dyskinesia Families

Paroxysmal kinesigenic dyskinesia (PKD [MIM128200]) is a heritable paroxysmal movement disorder characterized by recurrent and brief attacks of involuntary movements.[1,2] Its family histories show an autosomal dominant inheritance pattern. Our previous linkage and haplotype analyses defined the disease locus on 16p11.2-q12.1,[2] Similarly, other linkage studies of PKD assigned the locus to an overlapping region encompassing the centromere of chromosome 16.[3–5] In our previous study, we performed mutation analysis in seven families on 157 genes between D16S3131 and D16S416 (all the genes within this region); however, we failed to identify the causative gene.[1] Based on many linkage studies, we decided to extend the candidate region until more telomeric locus to D16S503 containing 72 RefSeq genes. Because genomic rearrangement could also result in PKD, we also performed copy number analysis for the entire candidate PKD locus.

Here, we describe the results of mutation analysis in 14 PKD families for the 72 genes between D16S416 and D16S503, and the results of copy number analysis in eight PKD families and two sporadic cases.

We collected 14 Japanese families, PKD-1–PKD-14, each of which includes multiple individuals affected by PKD, and two sporadic cases, PKD-S1 and PKD-S2. Among all these families, 64 patients were diagnosed with PKD. Our previous study showed that all affected members in each family have a disease related haplotype on chromosome 16[1,2] except for PKD-1 and PKD-2, which were not analyzed for haplotype because the family members is small.

Direct sequencing of the 72 genes in the segment between D16S416 and D16S503 revealed two substitutions which were not observed among 288 normal controls and not deposited in dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) (Table 1). A substitution, g.25190C>T (p.R282C) in *GPR114* found in the family PKD-12, was considered as rare variant because it was not co-segregated with PKD. The remaining one was g.35905C>T in exon 4 of *NLRC5* (NM_032206) resulting in p.T153T, segregated with PKD in family PKD-3. Even though this mutation in *NLRC5* is "silent," it might be a pathogenic because of splicing disturbance.[6] However, a nucleotide g.35905C in *NLRC5* is not so highly conserved in other species, and g.35905C>T would not affect splicing by prediction of NNSPLICE (http://www.fruitfly.org/seq_tools/splice.html) and GENSCAN (http://genes.mit.edu/GENSCAN.html; data not shown).

Copy number analysis using HumanExon510S-Duo BeadChip (Illumina, San Diego, CA) showed a deletion in 16p11.2 (Fig. 1A), but this has already been reported in the Database of Genomic Variants (DGV) (http://projects.tcag.ca/variation/). In our previous study, two nonsynonymous substitutions, p.P242T in *SCNN1G* and p.K1063R in *ITGAL*, which were segregated with PKD in one family, were still possible pathogenic mutation for PKD.[1] Structural variants including microdeletions/microduplications within three genes, *ITGAL*, *SCNN1G*, and *NLRC5*, were scanned using array comparative genomic hybridization (aCGH: Agilent Technologies, Santa Clara, CA). Two small deletions not registered in DGV were found within *ITGAL* among several patients (Fig. 1B). However, real-time quantitative PCR revealed genomic alterations in only one PKD patient in the ITGAL region1 and region2 (Fig. 1C). No alteration was found in *SCNN1G* and *NLRC5*. Results of copy number analyses showed no causative copy number changes.

Together with our previous study,[1] we have now analyzed almost all the exons and exon-intron boundaries between

***Correspondence to:** Akira Kinoshita, Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan; akino@nagasaki-u.ac.jp

Relevant conflicts of interest/financial disclosures: Nothing to report. This work was supported by Grants-in-Aid for Scientific Research from the Ministry of Health, Labour, and Welfare of Japan.
Full financial disclosures and author roles may be found in the online version of this article.

Published online 10 February 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.23475

Table 1. Novel SNPs identified in this study

Gene name	Nucleotide change	Location	AA change	Family	Notes
SLC12A3	g.14369C>T	Exon 11	A464A	PKD-13	Synonymous substitution Observed among 96 controls
NUDT21	g.21841T>C	3′UTR		PKD-11, PKD-14	Observed among 96 controls
PLLP	g.4C>T	5′UTR		PKD-13	Observed among 96 controls
KATNB1	g.8648G>A	Exon 3	S58N	PKD-3, PKD-10, PKD-11, PKD-13, PKD-14	Nonsynonymous substitution Observed among 96 controls
	g.15561G>A	IVS5+6		PKD-14	Near the splice site Observed among 96 controls
	g.16858C>T	IVS8+4		PKD-1, PKD-10, PKD-14	Near the splice site Observed among 96 controls
	g.18427C>T	Exon 12	D409D	PKD-1, PKD-13	Synonymous substitution Observed among 96 controls
	g.19491C>T	Exon 14	P472P	PKD-S2	Synonymous substitution Observed among 96 controls
SNORA46[a]	g.122G>A			PKD-8	Observed among 96 controls
	g.128G>A			PKD-4	Observed among 96 controls
NLRC5	g.35905C>T	Exon 4	T153T	PKD-3	Synonymous substitution Not observed among 288 controls
	g.36018C>T	Exon 4	P191L	PKD-11, PKD-12	Nonsynonymous substitution Observed among 96 controls
SLC38A7	g.4673A>G	Exon 3	T10C	PKD-13	Nonsynonymous substitution Observed among 96 controls
	g.13614G>A	Exon 10	Q378Q	PKD-3	Synonymous substitution Observed among 96 controls
GPR114	g.25190C>T	Exon 9	R282C	PKD-12	Nonsynonymous substitution Not observed among 288 controls

[a]*SNORA46* is a noncoding RNA.IVS, intervening sequence; AA, amino acid.

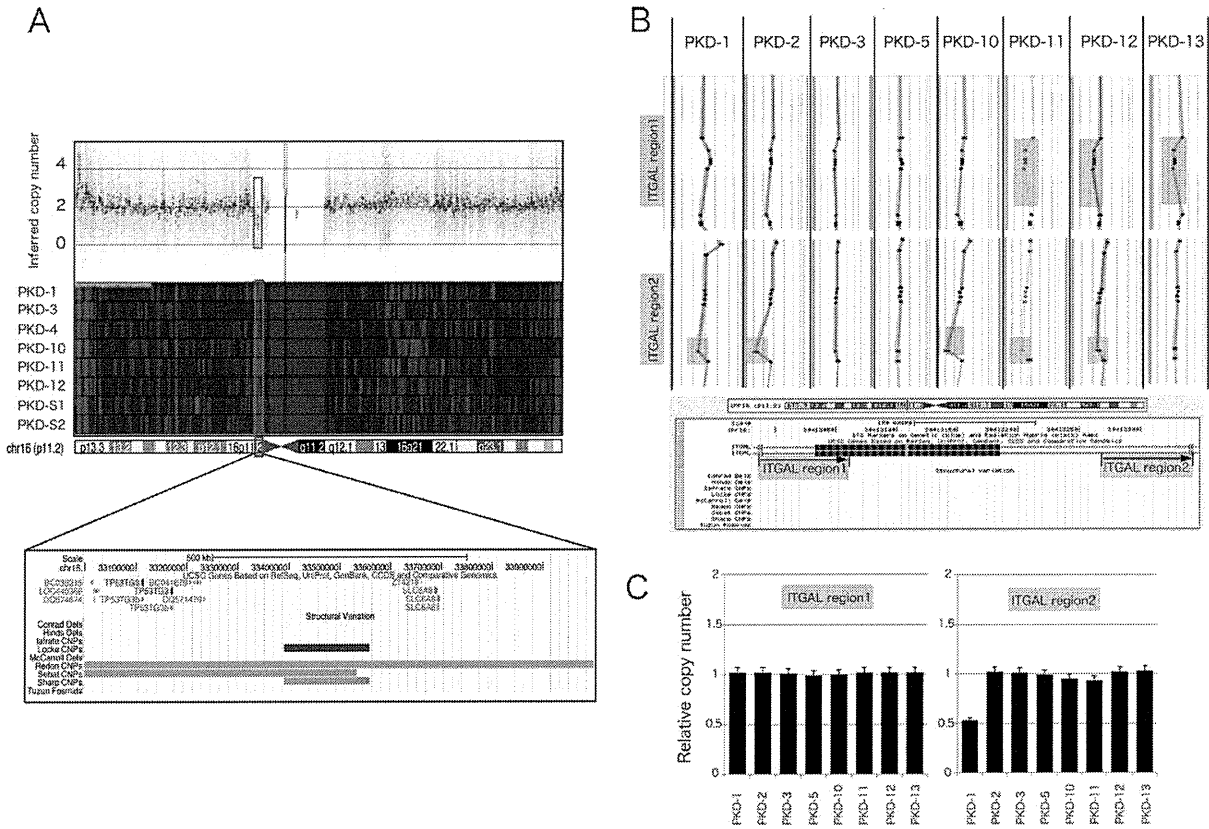


FIG. 1. Copy number analysis in patients with PKD. **A:** Eight patients (PKD-1, PKD-3, PKD-4, PKD-10, PKD-11, PKD-12, PKD-S1, and PKD-S2) are screened for copy number change using HumanExon510S-Duo BeadChip. Representative microarray data for chromosome 16. The region surrounded with a rectangle indicates the deleted region in affected individuals. Three horizontal lines indicate the inferred copy number. The deletion is registered in DGV. **B:** Eight patients (PKD-1, PKD-2, PKD-3, PKD-5, PKD-10, PKD-11, PKD-12, and PKD-13) were screened for copy number change using aCGH. Two microdeleted regions within *ITGAL* detected by array CGH are indicated by gray shaded areas (*ITGAL*, intron 11, Region 1) and light red shaded areas (*ITGAL*, intron 11, Region 2). Array CGH in PKD-11, PKD-12, and PKD-13 showed a ~100 bp loss of copy number in Region 1, and in PKD-1, PKD-2, PKD-10, PKD-11, and PKD-12 showed a ~100 bp loss of copy number in Region 2. **C:** The left and right bar graphs represent the results of quantitative PCR in *ITGAL* Region 1 and Region 2, respectively. No copy number changes were detected in Region 1, and the deletion in Region 2 was detected only in PKD-1, among eight patients with PKD.

D16S3131 and D16S503, but failed to identify the causative gene for PKD. Why have we failed? It is possible that PKD is caused by a recurrent structural aberration beyond the detection level of our experimental approaches. Because the 90th percentile largest gap on the HumanExon510S-Duo BeadChip is 14 kb, it is highly unlikely that our analysis could detect <15 kb structural variants accurately.

The peri-centromeric region (16p11.2–12.1) represents the largest zone of interchromosomal duplications and is composed of ~54 intrachromosomal duplications.[7] It is difficult to find pathogenic copy number changes and base changes in such a complex region. Even though our sequence analysis was comprehensive, some genes could be incomplete. For example, gene conversion in multiple copy number genes would be overlooked in our screening strategy. Furthermore, it is possible that the recurrent or founder intronic mutation cause PKD. den Hollander et al.[8] showed that a recurrent intronic mutation affected the splicing of *CEP290* gene, which resulted in Leber congenital amaurosis. The limitation of our strategy is that most of intronic regions in candidate locus could not be analyzed. We could not have an evidence of founder mutation in Japanese; it is still possible that a founder mutation in an intron is shared in patients.

We conclude that the causative mutation for PKD has not yet been identified. New technologies may be required to identify the PKD mutation in this complex genomic region.

Acknowledgments: We are grateful to the family members for their participation in this study. We thank Ms. Hayashida, Mrs. Oga, and Mrs. Iida for their technical assistance.

Shinji Ono, MD,[1,2] Koh-ichiro Yoshiura, MD, PhD,[1] Naohiro Kurotaki, MD, PhD,[2] Taeko Kikuchi, MD, PhD,[2] Norio Niikawa, MD,[3] and Akira Kinoshita, PhD[1]*

[1]Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan; [2]Department of Psychiatry, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan; [3]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Hokkaido, Japan

References

1. Kikuchi T, Nomura M, Tomita H, et al. Paroxysmal kinesigenic choreoathetosis (PKC): confirmation of linkage to 16p11-q21, but unsuccessful detection of mutations among 157 genes at the PKC-critical region in seven PKC families. J Hum Genet 2007;52: 334–341.
2. Tomita H, Nagamitsu S, Wakui K, et al. Paroxysmal kinesigenic choreoathetosis locus maps to chromosome 16p11.2-q12.1. Am J Hum Genet 1999;65:1688–1697.
3. Valente EM, Spacey SD, Wali GM, et al. A second paroxysmal kinesigenic choreoathetosis locus (EKD2) mapping on 16q13-q22.1 indicates a family of genes which give rise to paroxysmal disorders on human chromosome 16. Brain 2000;123:2040–2045.
4. Bennett LB, Roach S, Bowcock AM. A locus for paroxysmal kinesigenic dyskinesia map to human chromosome 16. Neurology 2000;54:125–130.
5. Cuenca-Leon E, Cormand B, Thomson T, et al. Paroxysmal kinesigenic dyskinesia and generalized seizures: clinical and genetic analysis in a Spanish pedigree. Neuropediatrics 2002;33:288–293.
6. Liu W, Qian C, Francke U. Silent mutation induces exon skipping of fibrillin-1 gene in Marfan syndrome. Nat Genet 1997;16: 328–329.
7. Martin J, Han C, Gordon LA, et al. The sequence and analysis of duplication-rich human chromosome 16. Nature 2004;432: 9889–9894.
8. den Hollander AI, Koenekoop RK, Yzer S, et al. Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. Am J Hum Genet 2006;79:556–561.

Imaging Evidence of Nigral Damage in Dystonia Secondary to Disulfiram Intoxication

There are a few reports on patients who developed dystonia or parkinsonism after disulfiram (DSF) intoxication. Such patients had lesions in the pallidostriatal nucleus on brain magnetic resonance imaging (MRI) studies.[1] However, the integrity of the nigrostriatal dopaminergic neurons has not been studied.

We describe a 25-year-old man who developed generalized dystonia following DSF intoxication. [^{18}F]-FP-CIT brain positron emission tomography (PET) studies showed reduced uptake in the bilateral posterior putamen.

A 25-year-old man presented with generalized dystonia following DSF intoxication. At the age of 15, he committed suicide by taking a 15 g of DSF. When he was transported to the emergency room, he was drowsy but had no focal neurological deficits. One month later, he developed action-induced dystonia in both feet. He also developed sudden onset of transient generalized dystonia. The attacks occurred about 10 times a day and lasted up to 5 min. On neurological examinations, he could understand simple questions and his speech and hand writing were unintelligible. He showed initiation delay in the vertical and horizontal saccadic eye movements. When he stretched out his arms, there was no dystonia. His finger tapping was very slow, but there was no fatigue or akinetic block. Once he made fists, he could hardly release them. While tapping his feet, there was dystonia in the toes. His foot tapping was slow bilaterally, and the amplitude was small. There was no motor weakness or sensory deficit. He needed help to stand and walk due to severe dystonic plantar flexion of the feet. There was a mild spasticity in the legs, but no rigidity in the arms and legs. The knee and ankle jerks increased mildly, but plantar reflexes were flexor bilaterally and there was no ankle clonus. Occasionally, he developed sudden onset of transient generalized dystonia, consisting of retrocollis, upward deviation of the eyes, grasp of the hands, and flexion and

***Correspondence to:** Myung Sik Lee, Department of Neurology, Gangnam Severance Hospital, Brain Korea 21 Project for Medical Science, Yonsei University College of Medicine, Seoul, South Korea; mslee@yuhs.ac

Relevant conflicts of interest/financial disclosures: Nothing to report. This work was supported by a faculty research grant of Yonsei University College of Medicine (grant number 6-2009-0157). Full financial disclosures and author roles may be found in the online version of this article.

Published online 10 February 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.23496

Journal of Human Genetics (2011) 56, 313–315

www.nature.com/jhg

ORIGINAL ARTICLE

Epidemiology of human papillomavirus genotypes in pregnant Japanese women

Kentaro Yamasaki[1], Kiyonori Miura[1], Takako Shimada[1], Shoko Miura[1], Shuhei Abe[1], Makoto Murakami[2], Tetsuro Sameshima[3], Akira Fujishita[4], Kouhei Kotera[5], Akira Kinoshita[6], Koh-ichiro Yoshiura[6] and Hideaki Masuzaki[1]

To investigate the pre-vaccination epidemiology of genital human papillomavirus (HPV) infections and genotypes in pregnant Japanese women, we performed Pap smear tests and HPV genotype testing in patients attending Nagasaki University Hospital and collaborating hospitals from August 2007 to July 2010. Serial uterine cervical specimens were obtained from 151 pregnant women. The HPV test was positive on the first visit in 54 women (35.8%; 54/151, average age 30). A total of 49 women (32.5%; 49/151) were infected by at least one high-risk HPV and 5 women were infected by only low-risk HPV. The three most prevalent high-risk HPV genotypes were HPV 52 (31.5%; 17/54), HPV 16 (29.6%; 16/51) and HPV 31 (13.0%; 7/51). The HPV infection pattern (negative, single infection and multiple infection) differed significantly according to the pregnancy trimester (χ^2-test; $P<0.01$(Pearson)). Among HPV-infected pregnant Japanese women, HPV52 was the most common genotype. The second most common genotype was HPV16, and these two genotypes accounted for ~60% of HPV-positive pregnant women. Infection with multiple HPV genotypes was observed more frequently in the first trimester of pregnancy and the pattern of infection changed significantly depending on pregnancy stage.
Journal of Human Genetics (2011) **56,** 313–315; doi:10.1038/jhg.2011.11; published online 10 February 2011

Keywords: epidemiology; genotype; HPV; infection; pregnancy; uterine cervical neoplasia

INTRODUCTION

Persistent infections with human papillomavirus (HPV) are recognized as a major cause of cervical cancer. Genital infections with HPV are very common, and these infections are transmitted by sexual contact.[1] In most cases, HPV infections disappear naturally in a relatively short time period and carry only a small risk of disease development;[2–4] however, some investigators believe that pregnancy affects the host immune system; for example, pregnancy is believed to reduce seroreactivity against HPV infection.[5] Estrogen and progesterone have been shown to activate the upstream regulatory region of HPV18 (see ref. 6). Thereby altering HPV clearance rates compared with non-pregnant women. During persistent infection, important factors of disease susceptibility include HPV genotypes and viral characteristics, such as the distribution of each type in the population and the evasive ability against the host's immune system. It is not clear how persistent infection relates to the host, such as how the host immune system reacts against a specific HPV genotype and how sexual behavior in pregnant women affects infection rates. In particular, there are very few data detailing HPV genotype prevalence in pregnant Japanese women. To determine the distribution of HPV infections in pregnant Japanese women, we performed HPV genotype testing.

MATERIALS AND METHODS

Study population and sample collection

Cytology and HPV DNA test samples were collected by six hospitals that collaborated with Nagasaki University Hospital from August 2007 to July 2010. All pregnant women whose first visit during pregnancy was to one of our collaborating hospitals were invited to join the study. An obstetrics doctor explained the study aim, procedures and complications. We did not inform participants of the HPV genotyping results. Exclusion criteria were patients who had previously received therapeutic surgery or who had histologically confirmed non-squamous neoplasms. We registered 151 pregnant women having a routine Pap test screen for the study. Specimens were taken with a Cervex brush (Rovers Medical Devices, Oss, Netherlands) and suspended in 10 ml of SurePath preservative fluid (Becton Dickinson, Franklin Lakes, NJ, USA). We used sample from the same vial for cytology and for HPV genotype testing. Cervical specimens for HPV genotyping

[1]Department of Obstetrics and Gynecology, School of Medicine, Nagasaki University, Nagasaki, Japan; [2]Department of Obstetrics and Gynecology, Sasebo Municipal Hospital, Sasebo, Japan; [3]Department of Obstetrics and Gynecology, the Japanese Red Cross Nagasaki Genbaku Hospital, Nagasaki, Japan; [4]Department of Obstetrics and Gynecology, Saiseikai Hospital, Nagasaki, Japan; [5]Department of Obstetrics and Gynecology, Nagasaki Municipal Hospital, Nagasaki, Japan and [6]Department of Human Genetics, School of Medicine, Nagasaki University, Nagasaki, Japan
Correspondence: Dr K Miura, Department of Obstetrics and Gynecology, School of Medicine, Nagasaki University, 1-7-1 Sakamoto, Nagasaki 852-8501, Japan.
E-mail: kiyonori@nagasaki-u.ac.jp
Received 30 November 2010; revised 26 December 2010; accepted 12 January 2011; published online 10 February 2011