accumulation of oxidized proteins and/or ROS in NNS fibroblasts may be one of the mechanisms responsible for the accumulation of p-p38 (29, 30, 38, 39). Increased p-p38 levels are in agreement with the proposed mechanism for TNFR1-associated periodic syndrome (TRAPS), which is another autoinflammatory syndrome (40).

To date, proteasome inhibitors have been used clinically to treat multiple myeloma and mantle cell lymphoma and are also effective for experimental autoimmune and inflammatory phenotypes, such as arthritis (37) and systemic lupus erythematosus (41). Generally, it is said that proteasome inhibitors induce apoptosis and inhibit immune responses. However, our results indicate that inhibiting the immunoproteasome can induce inflammatory reactions under some circumstances. In this context, the *PSMB8* mutation in NNS can be mimicked by histiocytoid Sweet syndrome (42) and cutaneous vasculitis (43) induced by bortezomib, a nonspecific proteasome inhibitor.

Taken together, the data in the present study suggest that reduction in proteasome activity affects signal transduction and promotes inflammation (Fig. 5). In NNS patients with the *PSMB8* mutation, inflammation causes ubiquitinated proteins to accumulate (compounding the effects on joints, skin, and muscle). These intracellular aggregates may then trigger innate immune responses and increased ROS production (increasing the levels of oxidized proteins), which then, through the activity of p-p38, activate the AP1 transcription factor causing an increase in the secretion of various cytokines such as IL-6.

Materials and Methods

Homozygosity Mapping. The genome-wide ROH overlap pattern was detected using in-house Ruby script (available on request) (44).

Glycerol Density Gradient Separation. Proteins from cell extracts (600 ìg) were separated into 32 fractions by centrifugation (22 h at 100,000 × *g*) in 8 -32 % (vol/vol) linear glycerol gradients.

Additional materials and methods are available in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank the families for their participation. We also thank Prof. M. Nakashima for valuable discussion and Ms. C. Hayashida and M. Ohga for technical assistance. This work was supported, in part, by grants from the Ministry of Health, Labour, and Welfare (to F.F., N.K, and K.-i.Y.), the Japan Society for the Promotion of Science (22591094 to H.I., 21390100 to K.-i.Y., 20590331 to A. Kinoshita, 21791566 to H.M., 23791115 to K.A., and 23591651 to N.K.), the Takeda Scientific Foundation and the Naito Foundation (K.i.-Y.), and the Lydia O'Leary Memorial Foundation (N.K.).

1. Nakajo A (1939) Secondary hypertrophic osteoperiostosis with pernio. *J Dermatol Urol* 45:77–86.
2. Nishimura N, Deki T, Kato S (1950) Secondary hypertrophic osteoperiostosis with pernio-like skin lesions observed in two families. *J Dermatol Venereol* 60:136–141.
3. Kitano Y, Matsunaga E, Morimoto T, Okada N, Sano S (1985) A syndrome with nodular erythema, elongated and thickened fingers, and emaciation. *Arch Dermatol* 121: 1053–1056.
4. Tanaka M, et al. (1993) Hereditary lipo-muscular atrophy with joint contracture, skin eruptions and hyper-gamma-globulinemia: A new syndrome. *Intern Med* 32:42–45.
5. Horikoshi A, Iwabuchi S, Iizuka Y, Hagiwara T, Amaki I (1980) A case of partial lipodystrophy with erythema, dactylic deformities, calcification of the basal ganglia, immunological disorders, and low IQ level (Translated from Japanese). *Rinsho Shinkeigaku* 20:173–180.
6. Kasagi S, et al. (2008) A case of periodic-fever-syndrome-like disorder with lipodystrophy, myositis, and autoimmune abnormalities. *Mod Rheumatol* 18:203–207.
7. Oyanagi K, et al. (1987) An autopsy case of a syndrome with muscular atrophy, decreased subcutaneous fat, skin eruption and hyper gamma-globulinemia: Peculiar vascular changes and muscle fiber degeneration. *Acta Neuropathol* 73:313–319.
8. Muramatsu T, Sakamoto K (1987) Secondary hypertrophic osteoperiostosis with pernio (Nakajo). *Skin Res* 29:727–731.
9. Murata S, Yashiroda H, Tanaka K (2009) Molecular mechanisms of proteasome assembly. *Nat Rev Mol Cell Biol* 10:104–115.
10. Jung T, Catalgol B, Grune T (2009) The proteasomal system. *Mol Aspects Med* 30: 191–296.
11. Tanaka K (2009) The proteasome: Overview of structure and functions. *Proc Jpn Acad Ser B Phys Biol Sci* 85:12–36.
12. Fehling HJ, et al. (1994) MHC class I expression in mice lacking the proteasome subunit LMP-7. *Science* 265:1234–1237.
13. Agarwal AK, et al. (2010) PSMB8 encoding the β5i proteasome subunit is mutated in joint contractures, muscle atrophy, microcytic anemia, and panniculitis-induced lipodystrophy syndrome. *Am J Hum Genet* 87:866–872.
14. Garg A, et al. (2010) An autosomal recessive syndrome of joint contracture, muscular atrophy, microcytic anemia, and panniculitis-associated lipodystrophy. *J Clin Endocrinol Metab* 95:E48–E63.
15. Unno M, et al. (2002) The structure of the mammalian 20S proteasome at 2.75 A resolution. *Structure* 10:609–618.
16. Seemuller E, Lupas A, Baumeister W (1996) Autocatalytic processing of the 20S proteasome. *Nature* 382:468–471.
17. Sijts EJAM, Kloetzel P-M (2011) The role of the proteasome in the generation of MHC class I ligands and immune responses. *Cell Mol Life Sci* 68:1491–1502.
18. Hirano Y, et al. (2008) Dissecting beta-ring assembly pathway of the mammalian 20S proteasome. *EMBO J* 27:2204–2213.
19. Hirano Y, et al. (2005) A heterodimeric complex that promotes the assembly of mammalian 20S proteasomes. *Nature* 437:1381–1385.
20. Seifert U, et al. (2010) Immunoproteasomes preserve protein homeostasis upon interferon-induced oxidative stress. *Cell* 142:613–624.
21. Froment C, et al. (2005) A quantitative proteomic approach using two-dimensional gel electrophoresis and isotope-coded affinity tag labeling for studying human 20S proteasome heterogeneity. *Proteomics* 5:2351–2363.
22. Akira S, Taga T, Kishimoto T (1993) Interleukin-6 in biology and medicine. *Adv Immunol* 54:1–78.
23. Kishimoto T (2005) Interleukin-6: From basic science to medicine—40 years in immunology. *Annu Rev Immunol* 23:1–21.
24. Nishimoto N, Kishimoto T (2006) Interleukin 6: From bench to bedside. *Nat Clin Pract Rheumatol* 2:619–626.
25. Gyrd-Hansen M, Meier P (2010) IAPs: From caspase inhibitors to modulators of NF-kappaB, inflammation and cancer. *Nat Rev Cancer* 10:561–574.
26. Pasparakis M (2009) Regulation of tissue homeostasis by NF-kappaB signalling: Implications for inflammatory diseases. *Nat Rev Immunol* 9:778–788.
27. Thalhamer T, McGrath MA, Harnett MM (2008) MAPKs and their relevance to arthritis and inflammation. *Rheumatology (Oxford)* 47:409–414.
28. Kumar S, Boehm J, Lee JC (2003) p38 MAP kinases: Key signalling molecules as therapeutic targets for inflammatory diseases. *Nat Rev Drug Discov* 2:717–726.
29. Kamata H, et al. (2005) Reactive oxygen species promote TNFalpha-induced death and sustained JNK activation by inhibiting MAP kinase phosphatases. *Cell* 120: 649–661.
30. Park GB, et al. (2010) Endoplasmic reticulum stress-mediated apoptosis of EBV-transformed B cells by cross-linking of CD70 is dependent upon generation of reactive oxygen species and activation of p38 MAPK and JNK pathway. *J Immunol* 185: 7274–7284.
31. Villagomez MT, Bae SJ, Ogawa I, Takenaka M, Katayama I (2004) Tumour necrosis factor-α but not interferon-γ is the main inducer of inducible protein-10 in skin fibroblasts from patients with atopic dermatitis. *Br J Dermatol* 150:910–916.
32. Lee EY, Lee Z-H, Song YW (2009) CXCL10 and autoimmune diseases. *Autoimmun Rev* 8:379–383.
33. Dahlqvist J, et al. (2010) A single-nucleotide deletion in the POMP 5′ UTR causes a transcriptional switch and altered epidermal proteasome distribution in KLICK genodermatosis. *Am J Hum Genet* 86:596–603.
34. Caudill CM, et al. (2006) T cells lacking immunoproteasome subunits MECL-1 and LMP7 hyperproliferate in response to polyclonal mitogens. *J Immunol* 176:4075–4082.
35. Hutchinson S, et al. (2011) A dominant role for the immunoproteasome in CD8+ T cell responses to murine cytomegalovirus. *PLoS ONE* 6:e14646.
36. Basler M, Moebius J, Elenich L, Groettrup M, Monaco JJ (2006) An altered T cell repertoire in MECL-1-deficient mice. *J Immunol* 176:6665–6672.
37. Muchamuel T, et al. (2009) A selective inhibitor of the immunoproteasome subunit LMP7 blocks cytokine production and attenuates progression of experimental arthritis. *Nat Med* 15:781–787.
38. Hou N, Torii S, Saito N, Hosaka M, Takeuchi T (2008) Reactive oxygen species-mediated pancreatic beta-cell death is regulated by interactions between stress-activated protein kinases, p38 and c-Jun N-terminal kinase, and mitogen-activated protein kinase phosphatases. *Endocrinology* 149:1654–1665.
39. McCubrey JA, Lahair MM, Franklin RA (2006) Reactive oxygen species-induced activation of the MAP kinase signaling pathways. *Antioxid Redox Signal* 8:1775–1789.
40. Bulua AC, et al. (2011) Mitochondrial reactive oxygen species promote production of proinflammatory cytokines and are elevated in TNFR1-associated periodic syndrome (TRAPS). *J Exp Med* 208:519–533.
41. Neubert K, et al. (2008) The proteasome inhibitor bortezomib depletes plasma cells and protects mice with lupus-like disease from nephritis. *Nat Med* 14:748–755.
42. Murase JE, et al. (2009) Bortezomib-induced histiocytoid Sweet syndrome. *J Am Acad Dermatol* 60:496–497.
43. Gerecitano J, et al. (2006) Drug-induced cutaneous vasculitis in patients with non-Hodgkin lymphoma treated with the novel proteasome inhibitor bortezomib: A possible surrogate marker of response? *Br J Haematol* 134:391–398.
44. Kurotaki N, et al. (2011) Identification of novel schizophrenia Loci by homozygosity mapping using DNA microarray analysis. *PLoS ONE* 6:e20589.

BMC Research Notes

TECHNICAL NOTE Open Access

Agile parallel bioinformatics workflow management using Pwrake

Hiroyuki Mishima[1,2*], Kensaku Sasaki[1,2], Masahiro Tanaka[3,4], Osamu Tatebe[3,4,5] and Koh-ichiro Yoshiura[1]

Abstract

Background: In bioinformatics projects, scientific workflow systems are widely used to manage computational procedures. Full-featured workflow systems have been proposed to fulfil the demand for workflow management. However, such systems tend to be over-weighted for actual bioinformatics practices. We realize that quick deployment of cutting-edge software implementing advanced algorithms and data formats, and continuous adaptation to changes in computational resources and the environment are often prioritized in scientific workflow management. These features have a greater affinity with the agile software development method through iterative development phases after trial and error.

Here, we show the application of a scientific workflow system Pwrake to bioinformatics workflows. Pwrake is a parallel workflow extension of Ruby's standard build tool Rake, the flexibility of which has been demonstrated in the astronomy domain. Therefore, we hypothesize that Pwrake also has advantages in actual bioinformatics workflows.

Findings: We implemented the Pwrake workflows to process next generation sequencing data using the Genomic Analysis Toolkit (GATK) and Dindel. GATK and Dindel workflows are typical examples of sequential and parallel workflows, respectively. We found that in practice, actual scientific workflow development iterates over two phases, the workflow definition phase and the parameter adjustment phase. We introduced separate workflow definitions to help focus on each of the two developmental phases, as well as helper methods to simplify the descriptions. This approach increased iterative development efficiency. Moreover, we implemented combined workflows to demonstrate modularity of the GATK and Dindel workflows.

Conclusions: Pwrake enables agile management of scientific workflows in the bioinformatics domain. The internal domain specific language design built on Ruby gives the flexibility of rakefiles for writing scientific workflows. Furthermore, readability and maintainability of rakefiles may facilitate sharing workflows among the scientific community. Workflows for GATK and Dindel are available at http://github.com/misshie/Workflows.

Background

The concept of workflows has traditionally been used in the areas of process modelling and coordination in industries [1]. Now the concept is being applied to the computational process including the scientific domain. Zhao *et al.* found that general scientific workflow systems are employed in and applied to four aspects of scientific computations: 1) describing complex scientific procedures, 2) automating data derivation processes, 3) high-performance computing (HPC) to improve throughput and performance, and 4) provenance management and query [2]. Although naïve methods such as shell scripts or batch files can be used to describe scientific workflows, the necessity of workflow systems arises to satisfy the four aspects mentioned above. Therefore, full-featured scientific workflow systems including Biopipe [3], Pegasus [4], Ptolemy II [5], Taverna [6], Pegasys [7], Kepler [8], Triana [9], Biowep [10], Swift [11], BioWMS [12], Cyrille2 [13], KNIME [14], Ergatis [15], and Galaxy [16] have been applied in the bioinformatics domain. Their features, however, have some disadvantages for actual practices in bioinformatics. It is not always easy to describe actual complex workflows using graphical workflow composition, and some workflow language formats,

* Correspondence: hmishima@nagasaki-u.ac.jp
[1]Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, 1-12-4 Sakamoto, Nagasaki, Nagasaki, Japan
Full list of author information is available at the end of the article

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:

```
CC = "gcc"
rule '.o' = > '.c' do |t|
  sh "#{CC} -c #{t.source}"
end
file "sample" = > ["sample.o"] do |t|
      sh  "#{CC}  -o  #{t.name}  #{t.prerequisites}"
end
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`".

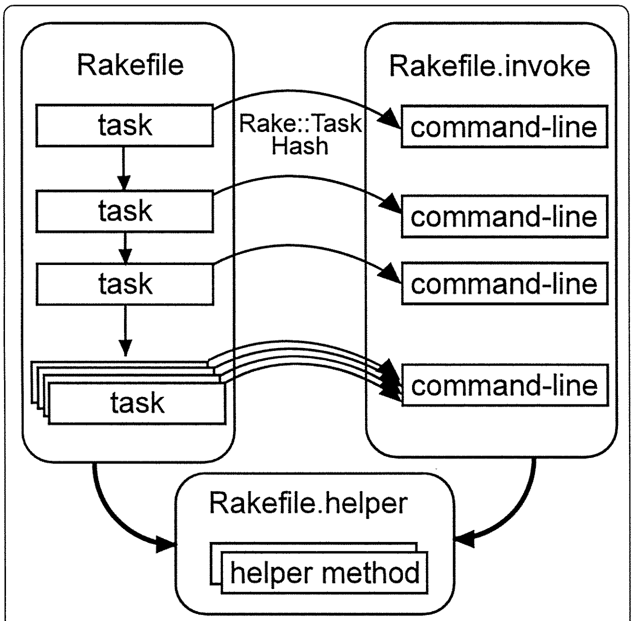

Figure 1 Structure of distinct rakefiles. A `Rakefile` file consists of task dependency descriptions. Tasks may be executed in parallel, if possible automatically. The `rakefile.invoke` file defines a class of the `RakefileInvoke` module. This class defines class methods to invoke command-lines and constants 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.

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.

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

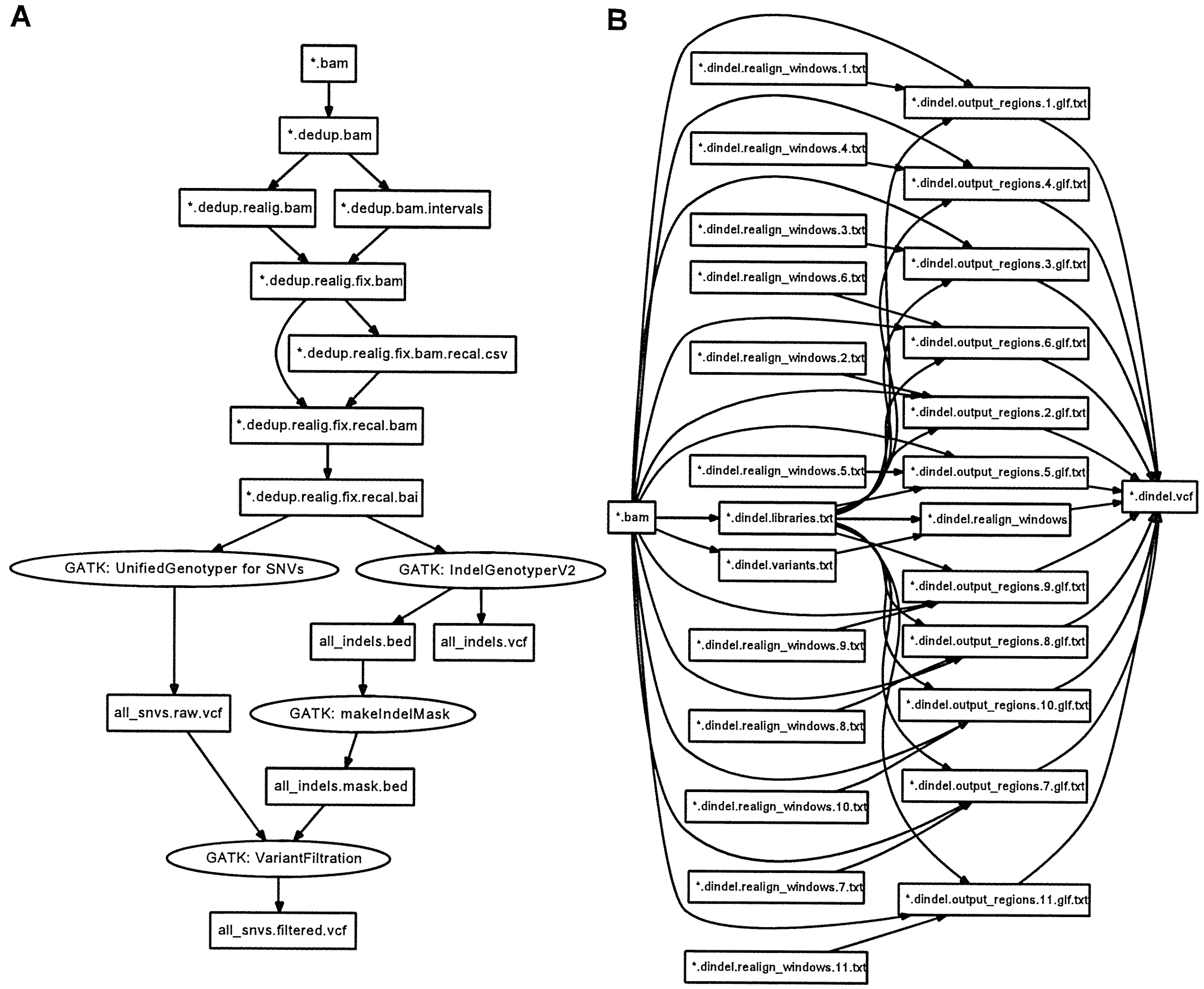
A
*.bam
*.dedup.bam
*.dedup.realig.bam
*.dedup.bam.intervals
*.dedup.realig.fix.bam
*.dedup.realig.fix.bam.recal.csv
*.dedup.realig.fix.recal.bam
*.dedup.realig.fix.recal.bai
GATK: UnifiedGenotyper for SNVs
GATK: IndelGenotyperV2
all_indels.bed
all_indels.vcf
all_snvs.raw.vcf
GATK: makeIndelMask
all_indels.mask.bed
GATK: VariantFiltration
all_snvs.filtered.vcf
B
*.dindel.realign_windows.1.txt
*.dindel.output_regions.1.glf.txt
*.dindel.realign_windows.4.txt
*.dindel.output_regions.4.glf.txt
*.dindel.realign_windows.3.txt
*.dindel.output_regions.3.glf.txt
*.dindel.realign_windows.6.txt
*.dindel.output_regions.6.glf.txt
*.dindel.realign_windows.2.txt
*.dindel.output_regions.2.glf.txt
*.dindel.realign_windows.5.txt
*.dindel.output_regions.5.glf.txt
*.dindel.vcf
*.bam
*.dindel.libraries.txt
*.dindel.realign_windows
*.dindel.variants.txt
*.dindel.output_regions.9.glf.txt
*.dindel.realign_windows.9.txt
*.dindel.output_regions.8.glf.txt
*.dindel.output_regions.10.glf.txt
*.dindel.realign_windows.8.txt
*.dindel.output_regions.7.glf.txt
*.dindel.realign_windows.10.txt
*.dindel.realign_windows.7.txt
*.dindel.output_regions.11.glf.txt
*.dindel.realign_windows.11.txt

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`.

```
# Rakefile
task :stage3 = > :stage2 do
 Rakefile::Dindel::BAM.each do |bam|
   prefix =
          bam.sub(/\.bam$/, ".dindel.realign_windows")
    FileList["#{prefix}.*.txt"].each do |f|
       target = f.sub(/\.realign_windows\./,
               ".output_regions.").
            sub(/\.txt$/, ".glf.txt")
      prerequisites =
      [f,
     f.sub(/\.dindel\.realign_windows\..*/, ".bam"),
     f.sub(/\.dindel\.realign_windows\..*/,
          ".dindel.libraries.txt"),]
      file target = > prerequisites do |t|
        RakefileInvoke::Dindel.dindel_stage3 t
      end
      file :stage3_invoke = > target
     end
   end
   (task :stage3_invoke).invoke
end
```

```
# Rakefile.invoke
def dindel_stage3(t)
  sh [DINDEL,
    "-analysis indels",
    "-doDiploid",
    "-bamFile #{t.prerequisites[1]}",
    "-ref #{REFERENCE}",
    "-varFile #{t.prerequisites[0]}",
    "-libFile #{t.prerequisites[2]}",
     "-outputFile #{t.name.sub(/\.glf\.txt$/, "")}",
     "1 > #{t.name.sub(/\.glf\.txt$/, "")}.log 2 > &1",
   ].join(" ")
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:

```
load "../GATK/Rakefile"
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:

```
rule '.dedup.bam.intervals' = >
  [ suffix_proc(".bam.intervals" = > ".bam") ] do |t|
    RakefileInvoke::Gatk.gatk_realigner_target_creater t
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:

```
def gatk_realigner_target_creater(t)
 sh [Java,
    "-Xmx#{JavaMemory}",
    "-Djava.io.tmpdir = #{JavaTempFile}",
    "-jar #{GATK_JAR}",
    "-T RealignerTargetCreator",
    "-R #{REFERENCE}",
    "-o #{t.name}",
    "-I #{t.source}",
     "-D #{DBSNP}",
    RakefileInvoke::Gatk::INTERVAL_OPTION,
    " > #{t.name}.log 2 > &1",
    ].join(" ")
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`.

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

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

```
module RakefileInvoke
  def generate_target(t)
    sh "command-line #{t.prerequisite} > #{t.name}"
  end
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, Kuhlmman 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.

THE JOURNAL OF
Obstetrics and Gynaecology Research

The official Journal of Asia and Oceania Federation of Obstetrics and Gynaecology
Japan Society of Obstetrics and Gynecology

doi:10.1111/j.1447-0756.2011.01601.x J. Obstet. Gynaecol. Res. Vol. 37, No. 11: 1666–1670, November 2011

Pre-vaccination epidemiology of human papillomavirus infections in Japanese women with abnormal cytology

Kentaro Yamasaki[1], Kiyonori Miura[1], Takako Shimada[1], Rie Ikemoto[6], Shoko Miura[1], Makoto Murakami[7], Tetsuro Sameshima[2], Akira Fujishita[3], Kouhei Kotera[4], Akira Kinoshita[5], Koh-ichiro Yoshiura[5] and Hideaki Masuzaki[1]

[1]Department of Obstetrics and Gynecology, School of Medicine, Nagasaki University, [2]Department of Obstetrics and Gynecology, The Japanese Red Cross Nagasaki Genbaku Hospital, [3]Department of Obstetrics and Gynecology, Saiseikai Hospital, [4]Department of Obstetrics and Gynecology, Nagasaki Municipal Hospital, [5]Department of Human Genetics, School of Medicine, Nagasaki University, Nagasaki, [6]SRL Corporation, Fukuoka, and [7]Department of Obstetrics and Gynecology, Sasebo Municipal Hospital, Sasebo, Japan

Abstract

Aim: To investigate the pre-vaccination epidemiology of genital human papillomavirus (HPV) infections and genotypes in women with abnormal cytology in Nagasaki, Japan.

Material and Methods: We performed Pap smear tests, biopsies and HPV genotype testing in Nagasaki Prefecture from August 2007 through November 2009.

Results: During the study period, serial samples of uterine cervical specimens were obtained from 539 subjects with abnormal cytology and/or squamous intraepithelial lesions (SIL) confirmed previously, or with clinically suspected invasive cervical cancer. In 119 HPV-positive subjects with low-grade SIL, the three most prevalent high-risk HPV genotypes were HPV52 (21.8%; 26/119), HPV16 (20.2%; 24/119) and HPV56 (17.6%; 21/119). In 199 women, 127 HPV-positive subjects with high-grade SIL and 67 HPV-positive subjects with squamous cell carcinoma (SCC), the three most prevalent high-risk HPV genotypes were HPV16 (44.3%; 86/194), HPV52 (20.6%; 40/194) and HPV58 (16.0%; 31/194).

Conclusion: Compared with the distribution of high-risk HPV genotypes in other countries, HPV52 was a more common genotype in Nagasaki. With disease progression to SCC, the distribution of high-risk HPV56 belonging to the A6 HPV family decreased, while HPV16 and HPV52 belonging to the A9 HPV family persisted. Our data provide an important resource to address the case for vaccination against HPV genotypes other than HPV16 and HPV18 in Japan.

Key words: epidemiology, genotype, human papillomavirus, infection, 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] However, HPV infections in most cases disappear naturally in a relatively short period, and induce little risk of developing disease.[2–4] We do not fully know the pathological mechanism that results in HPV infection developing into invasive cervical cancer (CC). During persistent infection, different viral characteristics along with HPV genotype may be important, such as the distribution of each type in the population and the ability to evade the host's immune system. Another important factor in persistent infection could be related

Received: November 2 2010.
Accepted: January 27 2011.
Reprint request to: Dr Kiyonori Miura, Department of Obstetrics and Gynecology, School of Medicine, Nagasaki University, 1-7-1 Sakamoto, Nagasaki 852-8501, Japan. Email: kiyonori@nagasaki-u.ac.jp

to the host, such as the host immune reaction against a specific HPV genotype, and sexual behavior.

The distribution of infectious high-risk HPV genotypes and the prevalence of CC in women varies worldwide. Clarification of the relation between clinical characteristics of CC and specific HPV genotypes in a local region may lead not only to implementation of a preventive strategy in that region, but also to an elucidation of the natural history of HPV infections compared with other regions in the world. In Japan, data on the distribution of HPV genotypes remains inadequate. To evaluate the possible effect of an HPV vaccine, we require knowledge of the pre-vaccination epidemiology of genital HPV infections. Thus, to determine the distribution and natural history of HPV infections in Nagasaki, Japan, we performed HPV genotype testing, cervical cytology and colposcopic biopsies.

Methods

Study population

The study included 625 subjects with abnormal cytology and/or histologically confirmed squamous intraepithelial lesions (SIL), or with clinically suspected invasive CC who required examination by colposcopy and directed biopsy. Cytology and HPV DNA test samples were collected in five hospitals in Nagasaki Prefecture from August 2007 through November 2009. Exclusion criteria were patients who had received therapeutic excisions previously or who had non-squamous neoplasms confirmed histologically. Thus, 86 subjects were excluded from the study.

The study protocol was approved by the Ethical Review Board of Nagasaki University and the other hospitals involved. All women were informed of the purpose of the study and gave their consent.

Sample collection and pathologic diagnoses

Specimens were collected using a Cervex Brush (Rovers Medical Devices, the Netherlands) and suspended in 10 mL of SurePath preservative fluid (Becton, Dickinson & Co., Franklin Lakes, NJ, USA). We used the samples from the same vial for cytology with the Bethesda III system (2001) and for HPV genotype testing. Cervical specimens for cytology and HPV genotyping were obtained at each visit from participants who received regular follow-up examinations. The cytologic diagnoses of the specimens were performed by experienced cytoscreeners in a commercial laboratory (SRL, Inc., Tokyo, Japan), and they were blinded from the HPV genotyping test. The histopathological review was performed by experienced pathologists of the Division of Pathology at Nagasaki University Hospital.

HPV genotyping test

Genotyping of HPV DNA in the SurePath preservative fluid after preparing glass slides was carried out using the Linear Array HPV Genotyping Test kit (Roche Molecular Systems, Indianapolis, IN, USA). The kit uses the PGMY09/PGMY11 primers[5] to amplify the L1 conserved region. Following polymerase chain reaction amplification, hybridization of the HPV amplicon was performed using an array of oligonucleotide probes that allowed independent identification of individual HPV genotypes. This kit can detect 37 HPV genotypes (6, 11, 16, 18, 26, 31, 33, 35, 39, 40, 42, 45, 51, 52, 53, 54, 55, 56, 58, 59, 61, 62, 64, 66, 67, 68, 69, 70, 71, 72, 73 (MM9), 81, 82 (MM4), 83 (MM7), 84 (MM8), IS39 and CP6108 (89)). For consistency with previous studies, we considered 16 HPV genotypes (16, 18, 31, 33, 35, 45, 51, 52, 53, 56, 58, 59, 66, 68, 73 and 82) as high-risk genotypes, which are related to CC based on previous reports.[6–8]

Results

Uterine cervical specimens for cytology and HPV DNA testing were collected from 539 women, with a mean age of 42 years in the age range of 19–94, with abnormal Pap smears and/or previously confirmed squamous intraepithelial lesions (SIL) or with clinically suspected invasive CC. In 154 subjects who were negative for intraepithelial lesion or malignancy (NILM), HPV was positive in 67 women (43.5%), with a mean age of 41 years at their initial HPV DNA test. The three most prevalent high-risk HPV genotypes in the NILM group were HPV 52 (23.9%; 16/67), HPV 16 (23.9%; 16/67) and HPV 71 (11.9%; 8/67) (Fig. 1a). In 125 subjects with cytologically low-grade SIL (LSIL), HPV was positive in 119 women (95.2%), with a mean age of 40 years at their initial test. The three most prevalent high-risk HPV genotypes were HPV 52 (21.8%; 26/119), HPV 16 (20.2%; 24/119) and HPV 56 (17.6%; 21/119) (Fig. 1b). In 128 participants diagnosed with cytologically high-grade SIL (HSIL), HPV infection was present in 127 women (99.2%), with a mean age of 41 years. In 71 women diagnosed cytologically with squamous cell carcinoma (SCC), HPV infection was positive in 67 women (94.4%), with a mean age of 56 years. In these latter two groups, the three most prevalent high-risk HPV genotypes were HPV 16

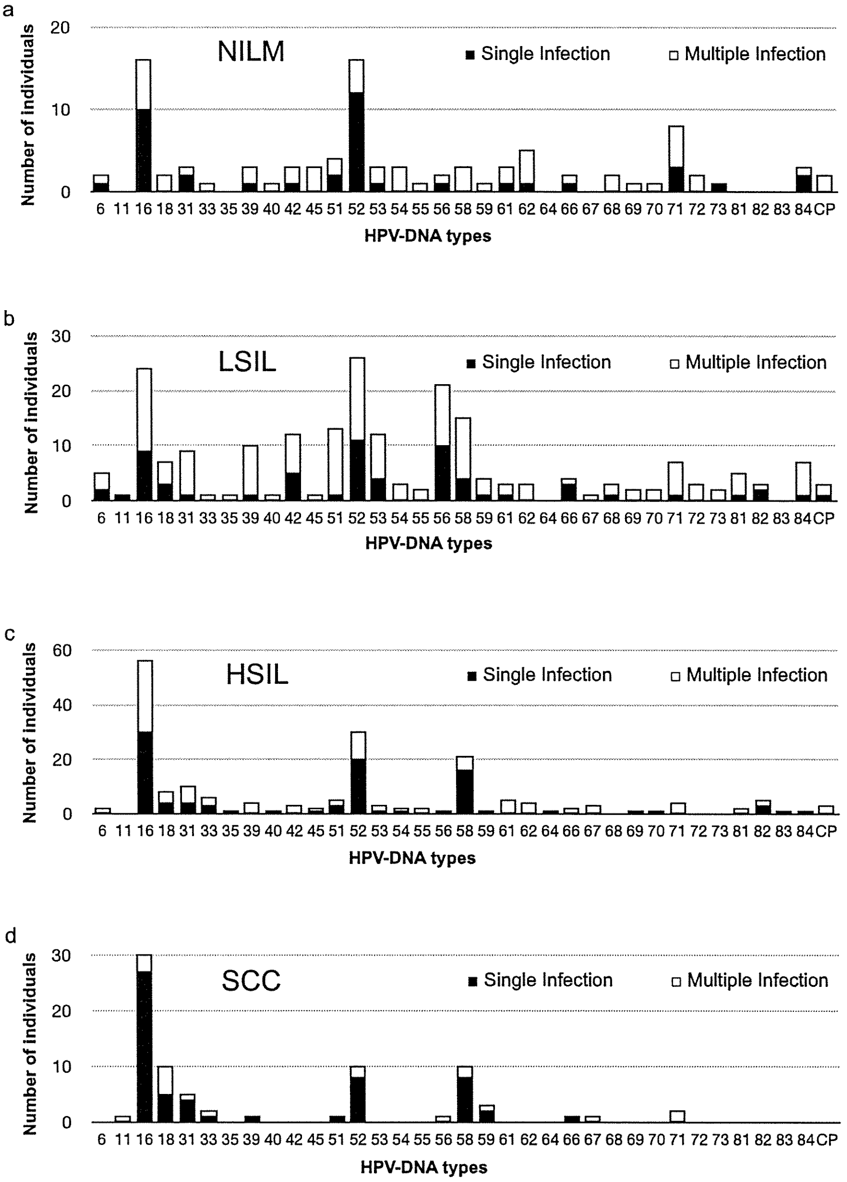

Figure 1 Prevalence of human papillomavirus (HPV) genotype among participants who were diagnosed with NILM (a), participants who were diagnosed with LSIL (b), participants who were diagnosed with HSIL (c) and with SCC (d). Closed boxes show infection with a single type of HPV DNA, and open boxes show multiple infections with two or more HPV-DNA types. NILM, negative for intraepithelial lesion or malignancy; LSIL, low-grade squamous intraepithelial lesion; HSIL, high-grade squamous intraepithelial lesion; SCC, squamous cell carcinoma; CP: CP6108 (HPV 89).

(44.3%; 86/194), HPV 52 (20.6%; 40/194) and HPV 58 (16.0%; 31/194) (Fig. 1c,d). The results of the other 61 subjects who were diagnosed with atypical squamous cells of undetermined significance (ASC-US) and atypical cells-cannot exclude HSIL (ASC-H) were not considered in this report. However, the HPV genotype distribution of the ASC-US group was similar to that of the LSIL group. The number of participants in the ASC-H group was too small to determine the distribution of HPV genotypes.

The Table 1 shows the multiple HPV infection status by cytological diagnosis. The percentage of single HPV infection was significantly higher in the SCC group (86.6%; 58/67) than in the LSIL group (53.8%; 64/119) ($P < 0.01$, Fisher's exact test).

Figure 2 shows a histogram of patient age distribution and the number of single infections of HPV 16, HPV 18 and HPV 52 in patients diagnosed histologically with cervical intraepithelial lesions grade 3 (CIN3) and invasive CC. The mean age of patients infected with HPV 52 tended to be higher than that of patients with HPV 16 infection but the difference was not statistically significant ($P = 0.07$, Student's *t*-test).

Table 1 Multiple human papillomavirus (HPV) infection status by cytologic diagnosis

Cytology	*n*	Single type	Two types	Three types	More than four types
LSIL	119	53.8% (64)	24.4% (29)	10.1% (12)	11.8% (14)
HSIL	127	67.7% (86)	22.8% (29)	5.5% (7)	3.9% (5)
SCC	67	86.6% (58)	10.4% (7)	3.0% (2)	0

The table shows the percentage and the number of participants who were infected with a single HPV genotype, and two, three or more than four HPV genotypes.

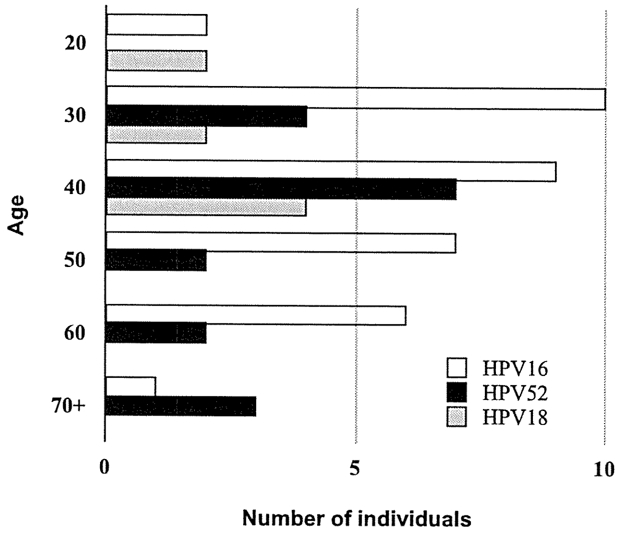

Figure 2 Distribution of ages in CIN3 and SCC patients who were infected with one HPV type, HPV 16, HPV 18 or HPV 52. Open boxes show the number of patients infected with only HPV 16, closed gray boxes show the number of patients infected with only HPV 18 and closed black boxes show the number of patients infected with only HPV 52. 20: 20–29 years; 30: 30–39 years; 40: 40–49 years; 50: 50–59 years; 60: 60–69 years; 70+: ≥70 years; CIN3: cervical intraepithelial lesions grade 3.

Discussion

The distribution of HPV genotypes in the LSIL group suggests that HPV 52 is the most frequently observed genotype among subjects with persistent HPV infections in Nagasaki. Other investigators also have reported that HPV 52 was dominant among women with normal cytology or cervical neoplastic lesions in Japan.[8–10] In the general population, the prevalence of HPV genotypes exhibit geographic differences in different countries, though HPV 16 is found to be most prevalent worldwide.[7,11] In pre-neoplastic and cancer cases, the geographic differences in prevalence of HPV genotype are diminished and HPV 16 tends to be the most dominant all over the world. HPV 18 and HPV 31 infections have also been reported to show higher prevalence in CIN and CC patients, but in the current study, there was a low prevalence of HPV 18 in the HSIL group. In the SCC group, the prevalence of HPV 18 was similar to that of HPV 52 and 58, although there was a lower number of single infection cases. Because four cases of multiple infections included HPV 18 as well as HPV 16 and 52 infections, the contribution of HPV 18 infection in the SCC group was difficult to evaluate.

In the LSIL group and the HSIL-SCC group, the distribution of HPV genotypes was different; the most marked differences between the HSIL-SCC group and LSIL group were a more than doubling of the HPV 16 genotype and the disappearance of HPV 56 infection in the former. HPV 16, HPV 52 and HPV 58 belong to the same alpha-papillomavirus species no. 9 family (A9 HPV family), which also includes HPV 31, 33, 35 and 67. However, HPV 56 belongs to the A6 HPV family, which also includes HPV 53 and 66.[12] The results indicated that the prevalence of the A6 HPV family was not small, especially in the Nagasaki LSIL group, but this HPV family was less likely to continue into persistent infection, and the observed prevalence of HPV 56 infection was found to be reduced in the HSIL and SCC groups.

Interestingly, the LSIL group had the lowest single infection rate of HPV (53.8%) (Table 1) and the rate of single infection was higher (67.7%) in the HSIL group. The SCC group showed the highest rate of single infection (86.6%). This finding has also been reported by other investigators and was suggested to support a monoclonal origin for cancer.[10]

We analyzed samples from patients with CIN3 and invasive SCC histologically and counted the number of patients in each age group, 20–29, 30–39, 40–49, 50–59, 60–69 and 70-plus. The most prevalent and dangerous HPV genotypes appeared to be HPV 16 and HPV 18, but the degree of risk of HPV 16/18 remained to be quantified. The histogram in Figure 2 shows the number of individuals who had a single infection with HPV 16, HPV 18 or HPV 52. The mean age and distribution of ages among HPV 16 and 52 types was different (paired *t*-test, $P = 0.07$), but HPV 52 appeared to be associated with slower progression of carcinogenesis, and HPV 16 and 18 with faster progression. The differences in HPV genotypes may be related not only to development of persistent infection, but also to the speed of progression to SCC.

In Japan, one commercial CC vaccine became available in December 2009. Although this study has some limitations because we included data only from SIL/CC women, our pre-vaccination data on the distribution of genital HPV infections in a region where HPV 52 and 58 are prevalent is valuable to determine the potential usefulness of a bivalent HPV vaccine. Paavonen *et al.* reported an estimated cross-reactivity against CIN2+ lesions with non-vaccine oncogenic HPV types of 37–54%.[13] Further study of the distribution of HPV genotypes in a SIL/CC population and the transition of pathological changes in patients according to HPV genotype is warranted.

Acknowledgments

This study was supported by Nagasaki University President's Fund Grant. K. Yamasaki was supported by a Grant-in-Aid for Young Scientists (B) (No. 20791150) from the Ministry of Education, Sports, Culture, Science and Technology of Japan. K. Miura was supported by a Grant-in-Aid for Young Scientists (B) (No. 21791567) from the Ministry of Education, Sports, Culture, Science and Technology of Japan, Grants from a Grant for Child Health and Development (No. 20C-1) from the Ministry of Health, Labour and Welfare, Japan. K. Yoshiura was supported by Grants-in-Aid for Scientific Research from the Ministry of Health, Labour and Welfare, Japan.

References

1. Shimada T, Miyashita M, Miura S *et al.* Genital human papilloma virus infection in mentally-institutionalized virgins. *Gynecol Oncol* 2007; **106**: 488–489.
2. Moscicki AB, Palefsky J, Smith G, Siboshski S, Schoolnik G. Variability of human papillomavirus DNA testing in a longitudinal cohort of young women. *Obstet Gynecol* 1993; **82** (Pt 1): 578–585.
3. Woodman CB, Collins S, Winter H *et al.* Natural history of cervical human papillomavirus infection in young women: A longitudinal cohort study. *Lancet* 2001; **357**: 1831–1836.
4. Ho GY, Bierman R, Beardsley L, Chang CJ, Burk RD. Natural history of cervicovaginal papillomavirus infection in young women. *N Engl J Med* 1998; **338**: 423–428.
5. Gravitt PE, Peyton CL, Alessi TQ *et al.* Improved amplification of genital human papillomaviruses. *J Clin Microbiol* 2000; **38**: 357–361.
6. Walboomers JM, Jacobs MV, Manos MM *et al.* Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol* 1999; **189**: 12–19.
7. Munoz N, Bosch FX, de Sanjose S *et al.* Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med* 2003; **348**: 518–527.
8. Asato T, Maehama T, Nagai Y, Kanazawa K, Uezato H, Kariya K-I. A large case-control study of cervical cancer risk associated with human papillomavirus infection in Japan, by nucleotide sequencing-based genotyping. *J Infect Dis* 2004; **189**: 1829–1832.
9. Miura S, Matsumoto K, Oki A *et al.* Do we need a different strategy for HPV screening and vaccination in East Asia? *Int J Cancer* 2006; **119**: 2713–2715.
10. Inoue M, Sakaguchi J, Sasagawa T, Tango M. The evaluation of human papillomavirus DNA testing in primary screening for cervical lesions in a large Japanese population. *Int J Gynecol Cancer* 2006; **16**: 1007–1013.
11. de Sanjose S, Diaz M, Castellsague X *et al.* Worldwide prevalence and genotype distribution of cervical human papillomavirus DNA in women with normal cytology: A meta-analysis. *Lancet Infect Dis* 2007; **7**: 453–459.
12. de Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H. Classification of papillomaviruses. *Virology* 2004; **324**: 17–27.
13. Paavonen J, Naud P, Salmeron J *et al.* Efficacy of human papillomavirus (HPV)-16/18 AS04-adjuvanted vaccine against cervical infection and precancer caused by oncogenic HPV types (PATRICIA): Final analysis of a double-blind, randomised study in young women. *Lancet* 2009; **374**: 301–314.

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

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. Hayashi[ƒ], 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

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

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,
ƒDept 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

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$^{\mathrm{G100S}}$).

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.

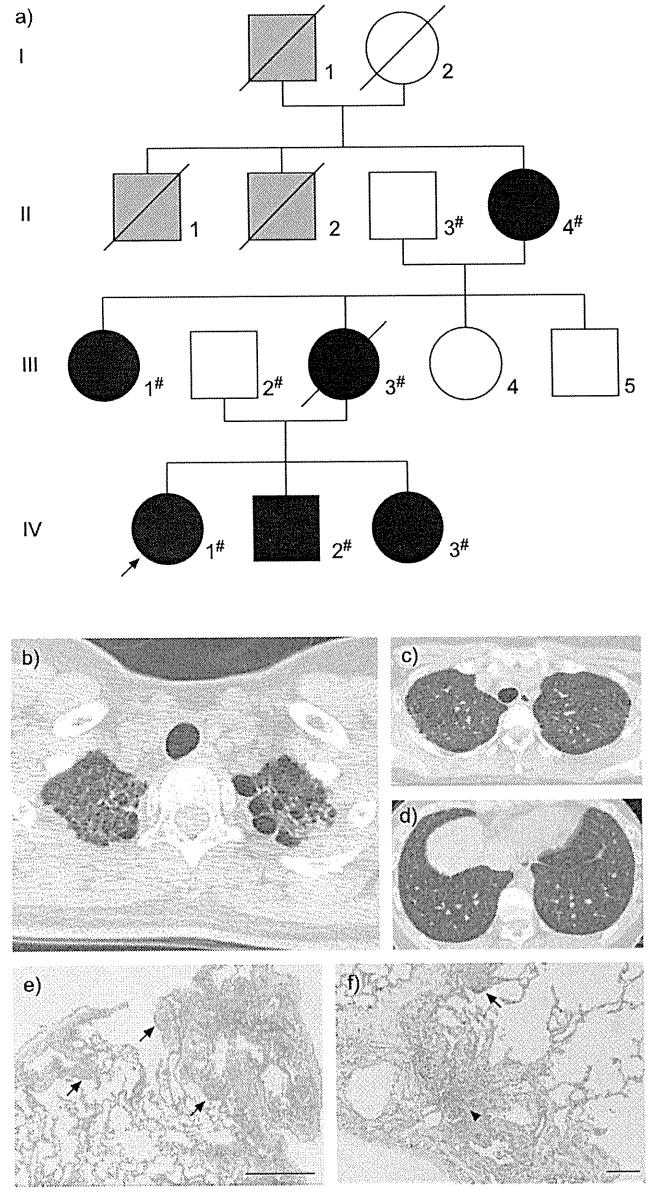

FIGURE 1. Pedigree of a family with familial pulmonary fibrosis (FPF), and radiological and histopathological findings of the proband. a) Pedigree of the family with FPF. Squares: males; circles: females; black: individuals diagnosed with pulmonary fibrosis; grey: individuals who had died due to respiratory 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′-ATCGGCTCCACT$\underline{A}$GCCTCGTGGTGT-3′ (forward) and 5′-ACACCACGAGGC$\underline{T}$AGTGGAGCCGAT-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) over-expressing 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 μg·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·mL^{-1} (normal range <500)	1560	386	NA	245	309	332
LDH IU·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
$P_{(A-a),O_2}$ Torr (normal range <10)	24.2¶	16.2	6.1	-8.5	-14	-10.8
BAL						
Cell count 10^5·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; $P_{(A-a),O_2}$: 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

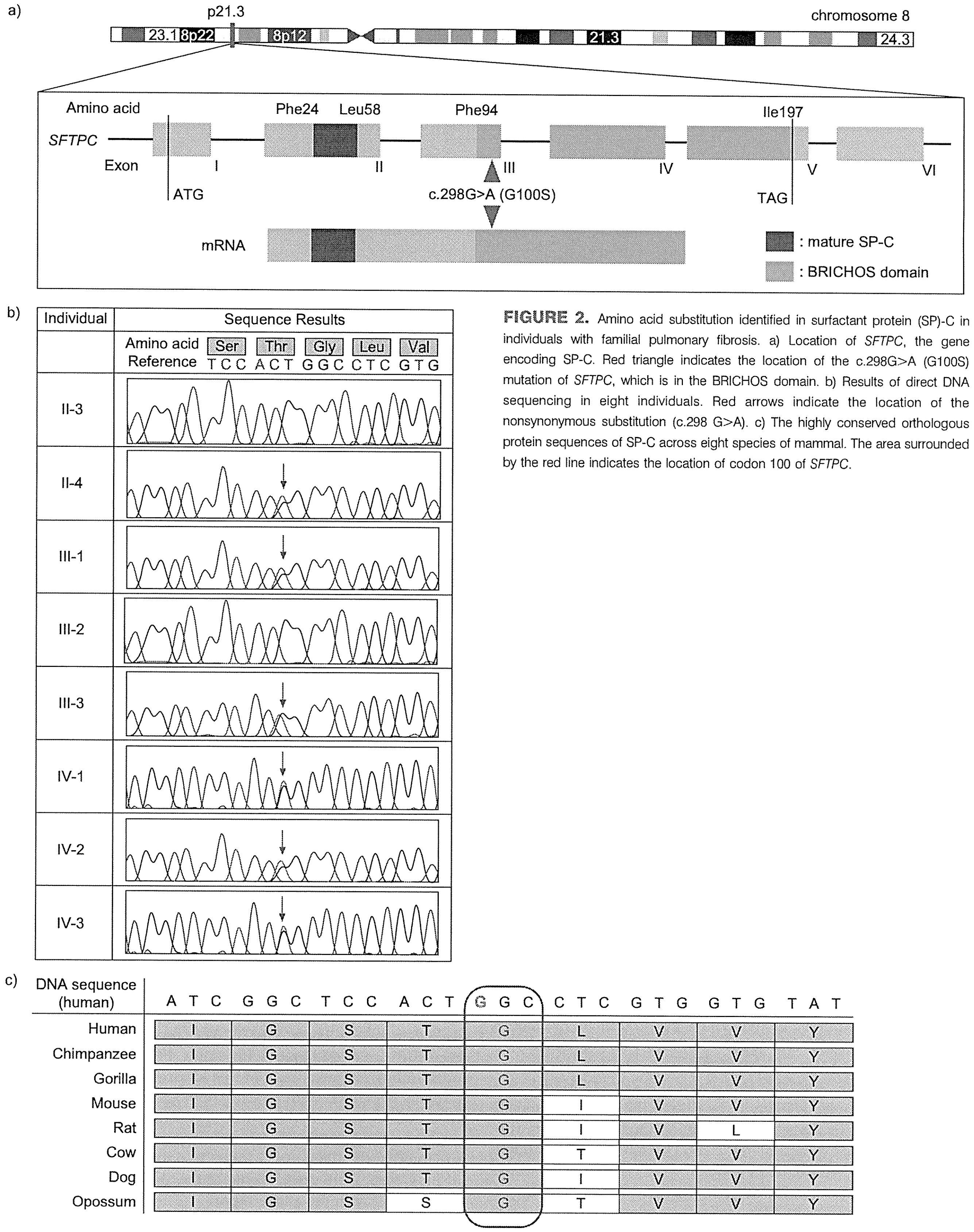

FIGURE 2. Amino acid substitution identified in surfactant protein (SP)-C in individuals with familial pulmonary fibrosis. a) Location of *SFTPC*, the gene encoding SP-C. Red triangle indicates the location of the c.298G>A (G100S) mutation of *SFTPC*, which is in the BRICHOS domain. b) Results of direct 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*.