III. 主な研究成果の刊行物・別冊

Mishima et al. BMC Research Notes 2011, 4:331
http://www.biomedcentral.com/1756-0500/4/331

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`:

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

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 If `sample.c` is out-of-date, i.e., older than `sample.o`, Rake skips compiling `sample.c` and just links `sample.o` to generate `sample`.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

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

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

Implementation

Rakefiles

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

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

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

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

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

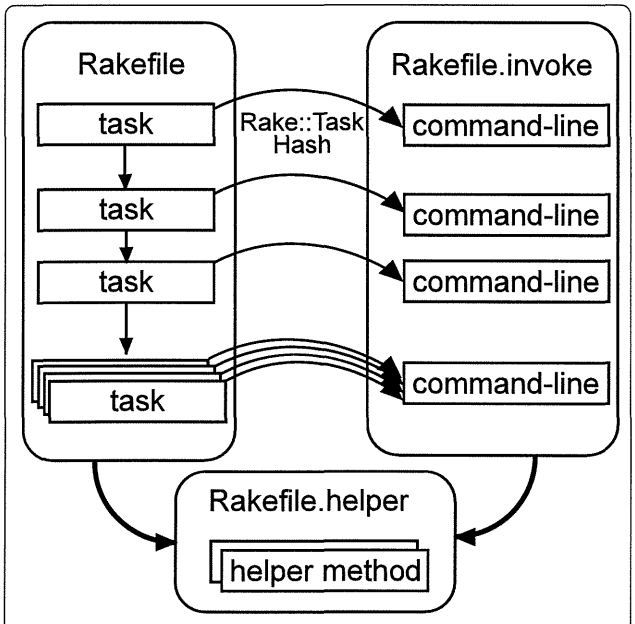

Figure 1 Structure of distinct rakefiles. A `Rakefile` file consists of task dependency descriptions. Tasks may be executed in parallel, if possible automatically. The `rakefile.invoke` file defines a class of the `RakefileInvoke` module. This class defines class methods to invoke command-lines and constants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

Example workflows

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

The GATK workflow

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

The Dindel workflow

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

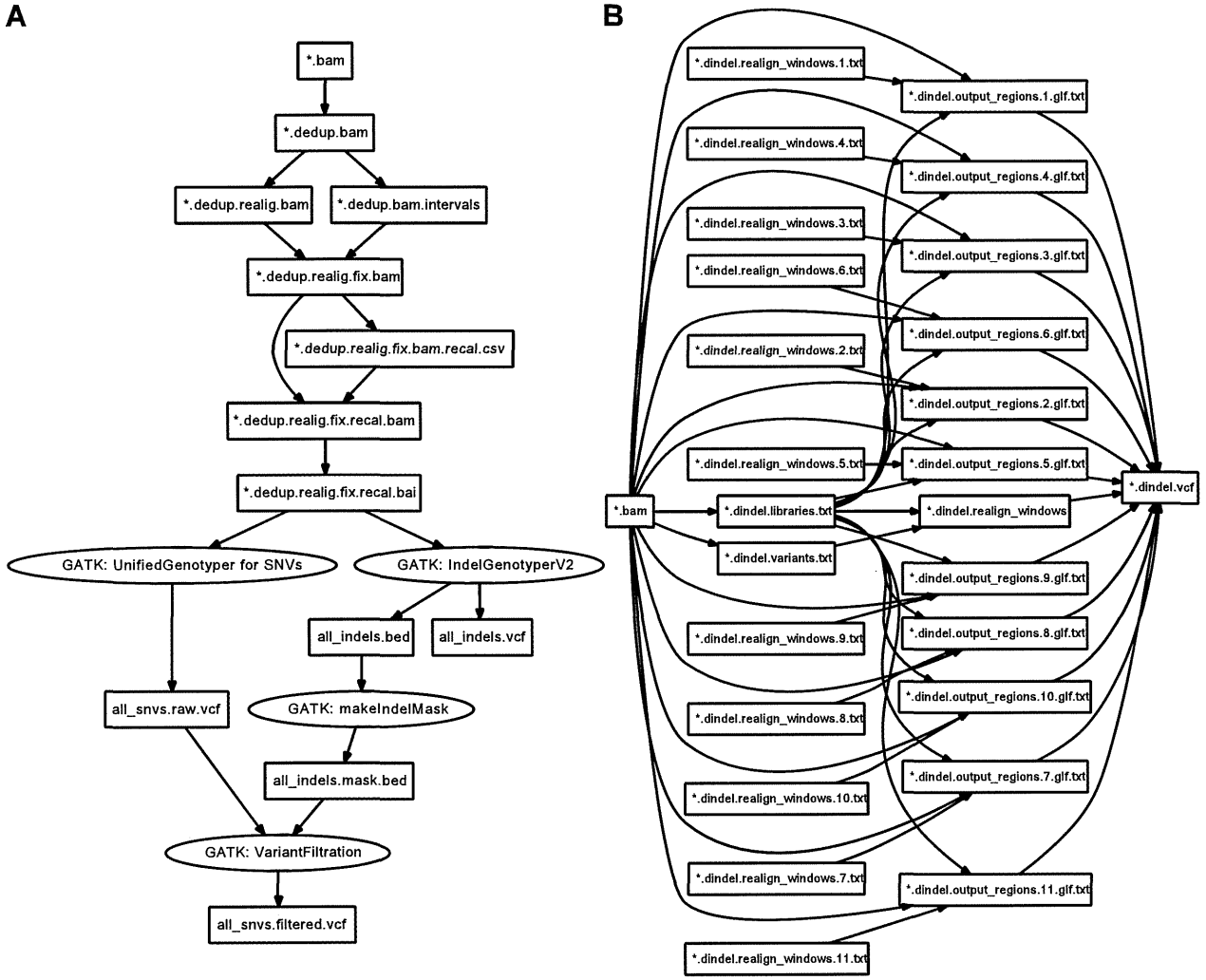

Figure 2 Directed acyclic graphs of GATK and Dindel workflows. Directed acyclic graphs to describe GATK (A) and Dindel (B) workflows are generated by Pwrake, manually adjusted, and visualized using Graphviz. (A) depicts a workflow from a `*.bam` file to an `all_snvs.filtered.vcf`, while (B) depicts a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

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

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

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

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

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

Combination of rakefiles

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

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

Results

Performance

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

Agility in workflow development

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

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

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

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

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

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

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

Procedure to describe new workflows

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

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

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

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

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

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

Discussion

Advantages in workflow execution

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

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

Workflow description flexibility

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

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

Sharing workflows

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

Conclusions

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

Availability and requirements

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

Availability of supporting data

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

List of abbreviations used

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

Acknowledgements

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

Author details

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

Authors' contributions

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

Competing interests

The authors declare that they have no competing interests.

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

References

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

11. Zhao Y, Hategan M, Clifford B, Foster I, Von Laszewski G, Nefedova V, Raicu I, Stef-Praun T, Wilde M: **Swift: Fast, reliable, loosely coupled parallel computation.** *Proceedings - 2007 IEEE Congress on Services, SERVICES 2007* 2007, 199-206.
12. Bartocci E, Corradini F, Merelli E, Scortichini L: **BioWMS: a web-based Workflow Management System for bioinformatics.** *BMC Bioinformatics* 2007, **8**:S2.
13. Fiers M, van der Burgt A, Datema E, de Groot J, van Ham R: **High-throughput bioinformatics with the Cyrille2 pipeline system.** *BMC Bioinformatics* 2008, **9**:96.
14. Berthold MR, Cebron N, Dill F, Gabriel TR, Kotter T, Meinl T, Thiel K, Wiswedel B: **KNIME - The Konstanz Information Miner.** *SIGKDD Explorations* 2009, **11**:26-31.
15. Orvis J, Crabtree J, Galens K, Gussman A, Inman JM, Lee E, Nampally S, Riley D, Sundaram JP, Felix V, Whitty B, Mahurkar A, Wortman J, White O, Angiuoli SV: **Ergatis: a web interface and scalable software system for bioinformatics workflows.** *Bioinformatics* 2010, **26**:1488-1492.
16. Goecks J, Nekrutenko A, Taylor J, Galaxy Team T: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11**:R86.
17. Radetzki U, Leser U, Schulze-Rauschenbach SC, Zimmermann J, Lüssem J, Bode T, Cremers AB: **Adapters, shims, and glue–service interoperability for in silico experiments.** *Bioinformatics* 2006, **22**:1137-1143.
18. Lin C, Lu S, Fei X, Pai D, Hua J: **A Task Abstraction and Mapping Approach to the Shimming Problem in Scientific Workflows.** In *Services Computing, IEEE International Conference on. Volume 0.* Los Alamitos, CA, USA: IEEE Computer Society; 2009:284-291.
19. Kane D, Hohman M, Cerami E, McCormick M, 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.

Proteasome assembly defect due to a proteasome subunit beta type 8 (PSMB8) mutation causes the autoinflammatory disorder, Nakajo-Nishimura syndrome

Kazuhiko Arima[a,1], Akira Kinoshita[b,1], Hiroyuki Mishima[b,1], Nobuo Kanazawa[c,1], Takeumi Kaneko[d], Tsunehiro Mizushima[e], Kunihiro Ichinose[a], Hideki Nakamura[a], Akira Tsujino[f], Atsushi Kawakami[a], Masahiro Matsunaka[c], Shimpei Kasagi[g], Seiji Kawano[g], Shunichi Kumagai[g], Koichiro Ohmura[h], Tsuneyo Mimori[h], Makito Hirano[i], Satoshi Ueno[i], Keiko Tanaka[j], Masami Tanaka[k], Itaru Toyoshima[l], Hirotoshi Sugino[m], Akio Yamakawa[n], Keiji Tanaka[o], Norio Niikawa[p], Fukumi Furukawa[c], Shigeo Murata[d], Katsumi Eguchi[a], Hiroaki Ida[a,q,2], and Koh-ichiro Yoshiura[b,2]

[a]Unit of Translational Medicine, Department of Immunology and Rheumatology, Graduate School of Biomedical Sciences, Nagasaki University, Nagasaki 852-8501, Japan; [b]Department of Human Genetics, Graduate School of Biomedical Sciences, Nagasaki University, Nagasaki 852-8523, Japan; [c]Department of Dermatology, Wakayama Medical University, Wakayama 641-0012, Japan; [d]Laboratory of Protein Metabolism, Graduate School of Pharmaceutical Sciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan; [e]Department of Life Science, Picobiology Institute, Graduate School of Life Science, University of Hyogo, Kamigori-cho, Ako-gun, Hyogo 678-1297, Japan; [f]Unit of Translational Medicine, Department of Neuroscience and Neurology, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki 852-8501; [g]Department of Clinical Pathology and Immunology, Kobe University Graduate School of Medicine, Kobe 650-0017, Japan; [h]Department of Rheumatology and Clinical Immunology, Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan; [i]Department of Neurology, Nara Medical University, Kashihara, Nara 634-8522, Japan; [j]Department of Neurology, Kanazawa Medical University, Kahoku-gun, Ishikawa 920-0293, Japan; [k]Department of Neurology, Utano National Hospital, Ukyou-ku, Kyoto 616-8255, Japan; [l]Department of Neurology and Medical Education Center, Akita University School of Medicine, Akita 010-8543, Japan; [m]Sugino Pediatric Clinic, Asakita-ku, Hiroshima 731-0231, Japan; [n]Office of Strategic Management, Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo 108-8639, Japan; [o]Laboratory of Protein Metabolism, Tokyo Metropolitan Institute of Medical Science, Setagaya-ku, Tokyo 156-8506, Japan; [p]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Ishikari-Tobetsu, Hokkaido 061-0293, Japan; and [q]Division of Respirology, Neurology, and Rheumatology, Department of Medicine, Kurume University School of Medicine, Kurume, Fukuoka 830-0011, Japan

Edited* by Daniel Kastner, National Institutes of Health, Bethesda, MD, and approved July 21, 2011 (received for review April 14, 2011)

Nakajo-Nishimura syndrome (NNS) is a disorder that segregates in an autosomal recessive fashion. Symptoms include periodic fever, skin rash, partial lipomuscular atrophy, and joint contracture. Here, we report a mutation in the human proteasome subunit beta type 8 gene (PSMB8) that encodes the immunoproteasome subunit β5i in patients with NNS. This G201V mutation disrupts the β-sheet structure, protrudes from the loop that interfaces with the β4 subunit, and is in close proximity to the catalytic threonine residue. The β5i mutant is not efficiently incorporated during immunoproteasome biogenesis, resulting in reduced proteasome activity and accumulation of ubiquitinated and oxidized proteins within cells expressing immunoproteasomes. As a result, the level of interleukin (IL)-6 and IFN-γ inducible protein (IP)-10 in patient sera is markedly increased. Nuclear phosphorylated p38 and the secretion of IL-6 are increased in patient cells both in vitro and in vivo, which may account for the inflammatory response and periodic fever observed in these patients. These results show that a mutation within a proteasome subunit is the direct cause of a human disease and suggest that decreased proteasome activity can cause inflammation.

Nakajo-Nishimura syndrome (NNS) (MIM256040, ORPHA-2615) is a distinct inflammatory and wasting disease. It was first reported by Nakajo in 1939, followed by Nishimura in 1950, and was called "secondary hypertrophic osteoperiostosis with pernio" (1, 2). More than 20 cases of this disease have been reported in various clinical fields, all from Japan (3–8). The disease was soon recognized as a new entity and was called "a syndrome with nodular erythema, elongated and thickened fingers, and emaciation" or "hereditary lipomuscular atrophy with joint contracture, skin eruptions and hyper-γ-globulinemia" on the basis of the common characteristic features (3, 4).

NNS usually begins in early infancy with a pernio-like rash. The patient develops periodic high fever, nodular erythema-like eruptions, and myositis. Lipomuscular atrophy and joint contractures gradually progress, mainly in the upper body, to form the characteristic thin facial appearance and elongated clubbed fingers. Inflammatory changes are marked and include constantly elevated erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP), hyper-γ-globulinemia, hepatosplenomegaly, basal ganglia calcification, and focal mononuclear cell infiltration with vasculopathy on histopathology. Autoantibodies are negative at the onset of NNS; although, in some cases, titers increase as the disease progresses.

Although NNS bears similarities to other autoimmune diseases, particularly dermatomyositis, it is only in recent years that its similarity to autoinflammatory periodic fever syndromes has been pointed out (5, 6). Oral steroids are effective in treating the inflammation, but not the wasting, and most patients die as a result of respiratory or cardiac failure. Despite the predicted segregation in an autosomal recessive fashion, the gene responsible has not been identified. Here, we describe a mutation in the human *PSMB8* that encodes the immunoproteasome subunit β5i in NNS patients.

Proteasomes collaborate with the ubiquitin system, which tags proteins with a polyubiquitin chain and marks them for degradation. The 26S proteasome is a multisubunit protease responsible for regulating proteolysis in eukaryotic cells in collaboration with the ubiquitin system. This ubiquitin–proteasome system is involved in various biological processes, including immune responses, DNA repair, cell cycle progression, transcription and protein quality control. It comprises a single catalytic 20S proteasome with 19S regulatory particles (RPs) attached to the ends (9–11). The 20S proteasome comprises 28 subunits arranged as a cylindrical particle containing four heteroheptameric rings: $\alpha_{1-7}\beta_{1-7}\beta_{1-7}\alpha_{1-7}$. Only three of the β subunits, β1, β2, and β5, are proteolytically active in the standard 20S pro-

Author contributions: N.K., A.Y., N.N., F.F., S.M., K.E., H.I., and K.-i.Y. designed research; K.A., A. Kinoshita, H.M., N.K., T.K., T. Mizushima, K.I., H.N., A.T., A. Kawakami, M.M., S. Kasagi, S. Kawano, S. Kumagai, K.O., T. Mimori, M.H., S.U., Keiko Tanaka, M.T., I.T., H.S., S.M., H.I., and K.-i.Y. performed research; K.A., A. Kinoshita, H.M., N.K., A.Y., S.M., H.I., and K.-i.Y. analyzed data; and N.K., Keiji Tanaka, F.F., S.M., H.I., and K.-i.Y. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]K.A., A. Kinoshita, H.M., and N.K. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: kyoshi@nagasaki-u.ac.jp or ida@med.kurume-u.ac.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1106015108/-/DCSupplemental.

teasome. Each of the three β subunits preferentially cleaves an acidic, basic, or hydrophobic residue, activities often referred to as caspase-like, trypsin-like, or chymotrypsin-like, respectively.

In vertebrates, there are three additional IFN-γ–induced subunits: β1i, β2i, and β5i. These are preferentially incorporated into the 20S proteasome in place of the standard subunits to form the immunoproteasome in immune cells such as macrophages, T and B cells, and dendritic cells, whereas their expression is low in nonlymphoid peripheral tissues. This results in more efficient production of MHC class I epitopes (12). The present study analyzed the activity of proteasomes with a mutated β5i subunit, and the subsequent inflammatory signal transduction pathways in mutant cells. The results suggest that the *PSMB8* mutation evokes an inflammatory response in humans, and that the p38 pathway may play an important role in inflammation in NNS patients.

Recently, a different mutation in the *PSMB8* gene was reported in patients with a disease similar to, but distinct from, NNS: an autosomal recessive syndrome of joint contracture, muscular atrophy, microcytic anemia, and panniculitis-associated lipodystrophy (JMP) (13, 14). The mutation in JMP syndrome, T75M, causes a reduction in chymotrypsin-like activity only, without disrupting the activity of other peptidases (13). In contrast, the G201V mutation identified in NNS patients results in the loss of all peptidase activity because of assembly defects and reduced proteasome levels. Thus, the discovery of *PSMB8* mutations in these related diseases indicates the presence of a distinct class of proteasome-associated autoinflammatory disorders.

Results

Clinical Features of NNS Patients. National surveillance in Japan confirms that only around 10 NNS patients are alive today. Therefore, preserved fibroblasts from an autopsy case (patient 1) were provided for genetic analysis, following approval by the local ethical committee. Of the living cases, written informed consent to undertake genetic and molecular analyses was obtained from six patients. The clinical features of all seven cases are summarized in Table S1. Patients 1, 2, and 4 were born to consanguineous parents and their clinical features have been described previously (Fig. S1*A*) (6, 8). The other patients are sporadic cases collected for this study and were born in the limited area between south Osaka and Wakayama. A diagnosis of NNS is not difficult owing to the characteristic features, including the thin facial appearance and long clubbed fingers (Fig. S1*B*). The clinical course throughout childhood was variable: from no medical consultation in the case of patient 7, to administration of oral steroids since infancy in patients 3 and 6. Partial lipomuscular atrophy with long clubbed fingers plus a pernio-like, heliotrope-like, or nodular erythema-like skin rash were observed in all cases, and periodic fever and joint contractures in most but not all. Wheras hyperhidrosis was also observed in some cases, short stature and low IQ were seen only in patients 6 and 1, respectively. Indeed, patient 6 was treated with growth hormone, although growth retardation in this case may have been due, in part, to oral steroids. Chronic inflammation, indicated by elevated ESR and hyper-γ-globulinemia, were observed in all patients, and microcytic anemia, high serum creatine phosphokinase (CPK), hepatosplenomegaly, and basal ganglia calcification were present in most, but not all. Notably, various autoantibodies (with a mildly elevated titer of antinuclear antibodies) were detected in half of the patients. The most striking differences between NNS and JMP are the absence of fever in JMP syndrome and the absence of seizures in NNS (14) (Table S1).

Genetic Mapping and Mutation Searches. We examined genomic DNA samples from five patients (patients 1–5) and three unaffected siblings of patient 4 using an Affymetrix GeneChip Human Mapping 500K array set (Nsp I and Sty I arrays), and the BRLMM genotyping algorithm. Because the runs of homozygosity (ROHs) shared by all patients were expected to be candidate regions containing the gene responsible for the disease, we identified a region spanning 1.1 Mb on chromosome 6p21.31–32 [from 32,798,004–33,903,106; National Center for Biotechnology Information (NCBI) build 36.1] as the sole candidate region responsible for NNS (Fig. 1*A*). We directly sequenced 436 coding exons in the 44 genes within this candidate region, including the splicing sites. A single nonsynonymous variation (not registered in the dbSNP database) was identified in exon 5 of *PSMB8* (NM_148919 in the NCBI database), designated *LMP7* or *RING10*, which encodes the LMP7 protein (β5i subunit) of the immunoproteasome. This mutation was a guanine to thymine transversion at nucleotide position 602 (c.602G > T) (Fig. 1*B*). Haplotype analysis indicated that the G201V mutation was probably introduced into the Japanese population by a single founder, as the haplotype around this mutation was identified in all patients (Fig. S1*C*). Gly201, which is a highly conserved residue in the β5i subunit (Fig. 1*C*) and among mature proteasome subunits in vertebrates (Fig. S1*D*), is substituted by Val (G201V) (Fig. 1*B*).

Impaired Immunoproteasome Assembly and Peptidase Activity. In silico modeling of the mutant β5i ($\beta5i^{G201V}$) subunit was used to infer the conformational impact of this mutation because the assembly of the proteasome is a highly orchestrated and complex process (9, 15). The β5i subunit is cleaved between amino acid residues Gly72 and Thr73 to yield the active form (16), in which the catalytic center is generated by Thr73, Asp89, Arg91, and Lys105. The mutated residue at position 201 was located at the edge of the S8 β-sheet of β5i and was close to its catalytic threonine residue Thr73 (Fig. 2*A*). The G201V substitution caused conformational changes not only in Thr73 but also in Lys105 within the catalytic center (Fig. S2). The mutation resulted in further conformational changes in the S8–H3 loop located at the

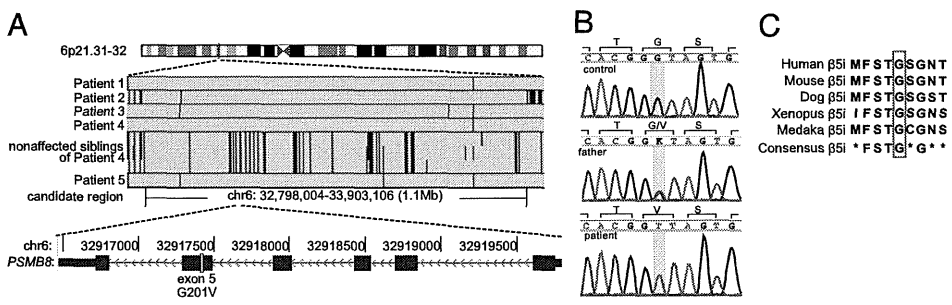

Fig. 1. SNP microarray-based homozygosity mapping and mutation search. (*A*) Homozygosity mapping for NNS patients and nonaffected siblings. ROH regions were detected using a hidden Markov model-based algorithm. The sole candidate region identified within 6p21.31–32 is shown. Green vertical lines indicate heterozygous SNPs and the background gray area indicates a region without heterozygous SNP calls. To be conservative, we did not regard isolated single heterozygous calls as delimiting ROH regions. The physical positions are shown in NCBI build 36.1. Patient numbers correspond to Figs. S1*A* and S1*C* and Table S1. No history of consanguineous marriage was apparent for patients 3 and 5, according to the family history interview. (*B*) Chromatograms for a control, a patient's father, and a patient. A mutation in *PSMB8* exon 5 identified in NNS patients by sequencing is highlighted in yellow. (*C*) Amino acid comparisons with other species. The glycine at the mutation site (red box) is highly conserved among vertebrates.

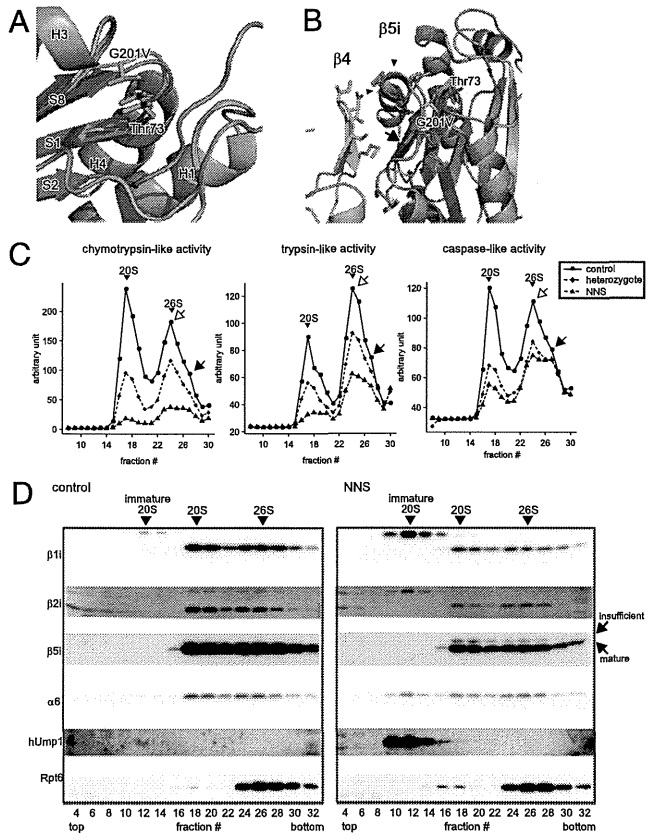

Fig. 2. G201V mutation in β5i reduces proteasome activity in immunoproteasome-expressing cells. (*A*) Close-up view of the mutation site (G201V) within β5i. Structural models of G201V β5i (orange) and wild-type β5i (green) were created from the β5-subunit structure [Protein Data Bank (PDB) ID code 1IRU]. The secondary structure elements for β5i are labeled. Val201 and Thr73 are shown in the stick model. Thr73 is a catalytic residue of β5i. (*B*) A ribbon diagram of the β4–β5i complex. The arrow shows the difference in the β-sheet between β5i (green) and β5i^{G201V} (orange). Arrowheads show the protruding S8–H3 loop of β5i^{G201V}. (*C*) Peptidase activity of LCLs. Extracts were fractionated by glycerol gradient centrifugation (8–32% glycerol from fraction 1–32). Arrowheads indicate the peak positions of the 20S and 26S proteasomes (open arrows, single-capped 26S; closed arrows, double-capped 26S). (*D*) Western blot analysis of fractionated total LCL extracts. Western blot analysis of proteasome subunits from fractions 1–32 fractionated in C. The sedimenting positions of the immature 20S, 20S, and 26S proteasomes are indicated by arrowheads. The mature and incompletely cleaved β5i^{G201V} subunits are indicated by arrows. The mature β5i subunit is cleaved within a C-terminal polypeptide between Gly72 and Thr73. The insufficiently cleaved β5i subunit is probably cleaved at a site toward the N terminus site, yielding a fragment with a higher molecular weight. The same amount of protein was subjected to glycerol gradient ultracentrifugation. The level of proteasome is reduced in NNS patients. Control, LCL extract from healthy control; NNS, LCL extract from patient with NNS.

interface between β4 and β5i, which affected the surface contact of β5i with the adjacent β4 subunit (Fig. 2*B*). These results suggest that the G201V mutation affects both β5i catalytic activity and assembly of the 20S proteasome.

According to Sijts and Kloetzel (17), the β1 subunit has a caspase-like function, the β2 subunit has trypsin-like activity, and the β5 subunit has chymotrypsin-like activity. Although it has not been clearly confirmed which of the immunoproteasome subunits possess which peptidase activity, it is generally thought that β5i has chymotrypsin-like activity. We next examined the influence of the β5i mutation on proteasome peptidase activity. Extracts from immortalized lymphoblastoid cell lines (LCLs) that constitutively expressed the immunoproteasome, rather than the standard proteasome, were obtained from an NNS patient, his heterozygous parent, and a healthy control, and were separated by glycerol gradient centrifugation. The fractions were then assayed for chymotrypsin-like, trypsin-like, and caspase-like activity mediated by the 20S/26S proteasomes. The results showed that not only was chymotrypsin-like activity markedly decreased in NNS cells, but the other two enzyme-like activities were also decreased (Fig. 2*C*).

Reduced Proteasome Levels. To gain further insight into the molecular mechanisms affecting peptidase activity in the mutant cells, the glycerol density gradient fractions were subjected to Western blot analysis (Fig. 2*D*). Assembly of the mammalian 20S proteasome begins with the formation of the α-ring in conjunction with a dedicated assembly chaperone, PAC1-4. The β-ring is then formed on the α-ring with the aid of another chaperone, hUmp1, resulting in the formation of half-sized immature proteasomes. The immature proteasomes then dimerize to form the 20S proteasome accompanied by cleavage of β-subunit propeptides and the degradation of hUmp1 (9). Our most noteworthy finding was the accumulation of immature 20S proteasome precursors in NNS cells before incorporation of β5i and dimerization, as indicated by the presence of the proforms of β1i and β2i, α6 and hUmp1, and the absence of β5i (Fig. 2*D*, fractions 10–14) (18). Computer modeling suggests that this assembly defect could be due to the fact that β5i, β4, and β6 line up next to each other and that the interaction between mutant β5i^{G201V} and β4 may be disturbed (Fig. 2*B*). The reduction in peptidase activity was unlikely due to differences in the ability of 20S to associate with 19S RP, because single-capped and double-capped 26S proteasomes were detected in the glycerol fractions from an NNS patient and control LCLs (Fig. 2*C*). The assembly defect caused a reduction in the number of 20S and 26S proteasomes in NNS cells (Fig. 2*D*), which accounts for the observed decrease in activity of all three peptidases. Another intriguing observation was that a portion of the β5i^{G201V} subunit incorporated into the mature proteasome appeared as a slower migrating band, suggesting the presence of an insufficiently cleaved form of β5i^{G201V} (Fig. 2*D*) (16). This may have contributed to the markedly reduced chymotrypsin-like activity seen in NNS cells compared with the other two peptidase activities.

Decreased Proteolytic Activity and Accumulation of Ubiquitinated and Oxidized Proteins. To examine proteolytic activity in vitro, the ornithine decarboxylase (ODC) degradation assay was performed (19). Proteolytic activity was significantly decreased in mutant proteasomes (Fig. 3*A*). As a consequence of the altered proteasome levels and incomplete cleavage of the subunits, proteolytic activity decreased and ubiquitinated proteins accumulated in LCLs (Fig. 3*B*) and fibroblasts from NNS patients (Fig. 3*C*). In particular, there was an obvious accumulation of K48 polyubiquitinated proteins in fibroblasts (Fig. 3*C*).

Because the immunoproteasome is important for degrading oxidized proteins and defective ribosomal products (20), we examined whether such proteins accumulated in NNS cells. We found that the level of oxidized proteins increased in cultured NNS fibroblasts and after stimulation with IFN-γ (Fig. 3*D*). Taken together, these results show that the G201V substitution within β5i severely impairs assembly of the immunoproteasome, leading to decreased proteasome levels and activity in β5i-expressing cells.

We then examined whether the defect in proteasome activity was apparent in situ in NNS patients. We stained skin biopsy sections obtained from an NNS patient and used sections from a monocytic fasciitis patient as a control. CD68 is a marker for monocyte/macrophages, a cell type known to predominantly express the immunoproteasome rather than the standard proteasome (21). Inflammatory responses characterized by the infiltration of numerous CD68$^+$ cells into the skin were observed in both NNS and fasciitis samples. However, the CD68$^+$ cells in the NNS sec-

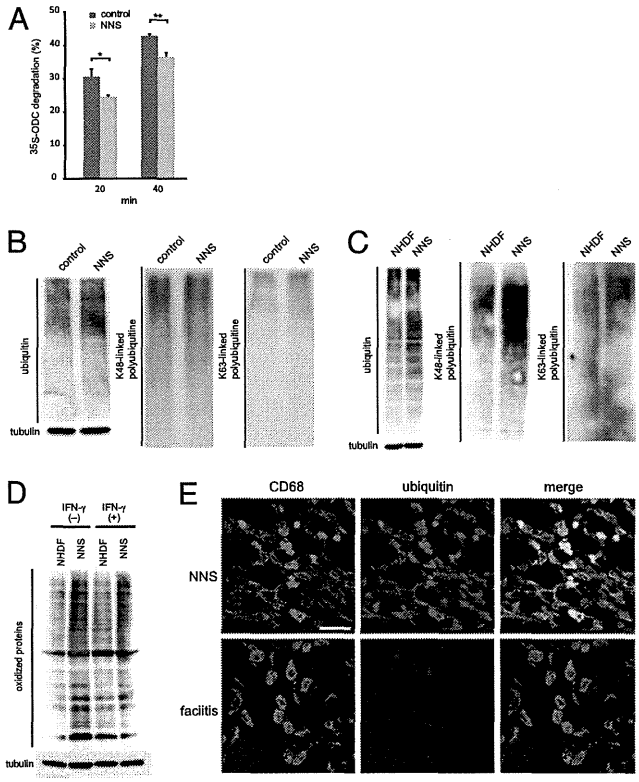

Fig. 3. Decrease of proteolytic activity and accumulation of polyubiquitinated and oxidized proteins in NNS cells. (*A*) In vitro proteolytic activity of the mutant proteasome. Degradation of recombinant ^{35}S-labeled ODC was expressed as % total ODC as described previously (11). Error bars indicated the SD of the mean (n = 3). *$P < 0.05$, **$P < 0.01$. (*B* and *C*) Accumulation of ubiquitinated proteins in LCLs (*B*) and fibroblasts (*C*). Western blot analysis of ubiquitinated proteins using an antiubiquitin antibody (*Left*), an anti-K48 polyubiquitinated protein antibody (*Middle*), and an anti-K63 polyubiquitinated protein antibody (*Right*). Tubulin was used as a loading control (*Lower*). NHDF, adult normal human dermal fibroblasts. (*D*) Levels of oxidized proteins determined by Oxyblot. NHDF and NNS fibroblasts were stimulated with or without 100 units of IFN-γ for 24 h. Tubulin was used as a loading control. (*E*) Immunofluorescence staining of CD68 and ubiquitinated proteins. Staining for CD68 (green) and ubiquitinated proteins (red) in skin sections from an NNS patient and a fasciitis patient. NNS ubiquitin signals showed a 4.7-fold increase with ImageJ (http://rsb.info.nih.gov/ij/) compared with fasciitis signals. (Scale bar, 10 μm.)

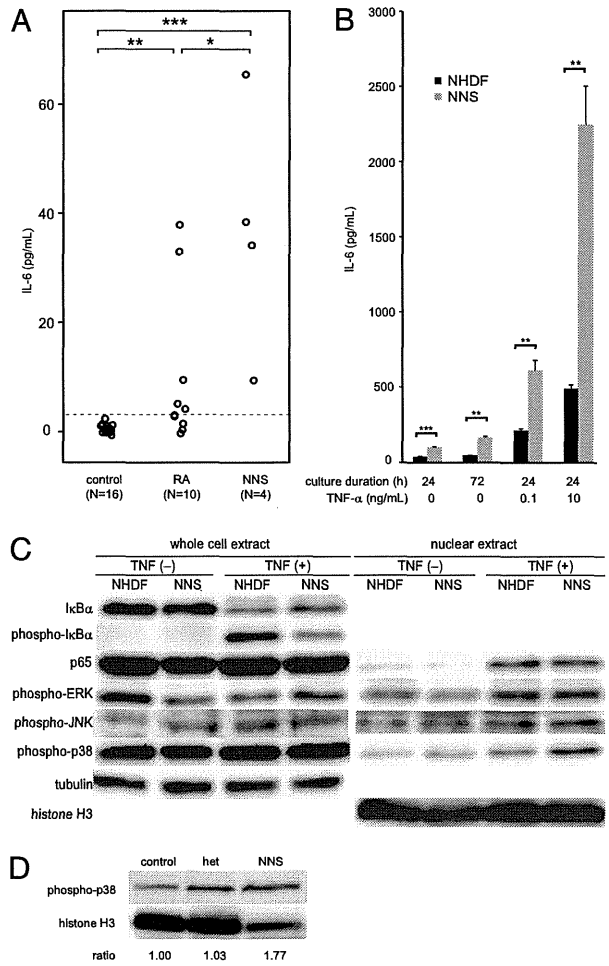

Fig. 4. Analyses of the level of IL-6 in NNS and the signal transduction system related to cytokine production. (*A*) IL-6 concentrations in sera from healthy controls, patients with NNS, and patients with rheumatoid arthritis. IL-6 levels in sera were determined by ELISA. (*B*) IL-6 production by cultured fibroblasts. The concentrations of IL-6 in conditioned media were determined by ELISA (in triplicate). (*C*) Western blot analysis for NF-κB and MAPK. Whole cell extracts and nuclear extracts were immunoblotted using antibodies against IκBα, p-IκBα, p65, p-ERK, p-JNK, and p-p38. (*D*) Western blot analysis of p-p38 in peripheral blood lymphocytes. Nuclear extracts from the peripheral blood lymphocytes of a healthy control, a heterozygous family member, and a NNS patient were blotted and visualized with anti–p-p38. Error bars indicate SD of the mean. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$ [Mann-Whitney u test (*A*) and two-tailed Welch's t test (*B*)]. Signal intensities were quantified using ImageJ and expressed as fold changes relative to a healthy control normalized to histone H3 (*D*).

tions were strongly positive for ubiquitin, whereas ubiquitin was only faintly detectable in the fasciitis sections (Fig. 3*E*).

Increased IL-6 and IP-10 Levels in NNS Patient Sera and Signal Transduction in NNS Fibroblasts. We next screened NNS patient sera for inflammatory cytokines using a multiplex bead-based ELISA on a suspension array. The results showed a significant increase in the levels of interleukin (IL)-6, IFN-γ–inducible protein (IP)-10, granulocyte colony stimulating factor, and monocyte chemoattractant protein-1 (Fig. S3*A*). IL-6 was of particular interest because it is a pleiotropic cytokine with a wide range of biological activities, and it plays a key role of immune regulation, hematopoiesis, oncogenesis, and inflammation (22–24). Increased IL-6 levels in NNS sera were confirmed using a standard ELISA (Fig. 4*A*). IL-6 production was significantly higher in NNS patient fibroblasts than in healthy control fibroblasts both in the presence and absence of TNF-α (Fig. 4*B*). The serum concentration of IP-10 was also higher than that in healthy controls (Fig. S3 *A* and *B*). We measured the level of IP-10 in conditional media from cultured fibroblasts using an ELISA, but found no significant difference under the conventional culture condition, although NNS cells tended to overproduce IP-10 after stimulation with 10 ng/mL TNF-α (Fig. S3*C*).

We next investigated the various signal transduction pathways that could be responsible for IL-6 overproduction by NNS fibroblasts. Nuclear factor (NF)-κB and AP-1 are the two major transcription factors that induce proinflammatory cytokines, including IL-6 (25, 26). We used an EMSA to detect activated NF-κB in cells treated with TNF-α; however, no differences in the amount of the p65/p50 heterodimer were observed in nuclear extracts from NNS fibroblasts and healthy control fibroblasts (Fig. S4 *A* and *B*). Consistent with this result, IκBα degradation and nuclear translocation of NF-κB were not enhanced in NNS fibroblasts (Fig. 4*C*). Although activation of NF-κB is largely dependent on the ubiquitin–proteasome system, these results

suggest that decreased proteasome activity does not have much influence on the regulation of NF-κB signaling in NNS cells.

We next measured the molecules that activate AP-1, including JNK1/2/3, ERK1/2, and p38, by Western blot analysis (27, 28). The amount of phosphorylated p38 (p-p38) in the nuclear extracts from NNS fibroblasts was increased (Fig. 4*C*), irrespective of TNF-α stimulation; however, there was no obvious difference in the levels of JNK1/2/3 and ERK1/2 (Fig. 4*C*). We also observed increased levels of p-p38 in the nuclear extracts from NNS peripheral blood lymphocytes (Fig. 4*D*). The build-up of oxidized proteins and/or reactive oxygen species (ROS) within NNS fibroblasts may be one of the mechanisms responsible for the accumulation of p-p38 (29, 30).

Discussion

We have identified a point mutation in the gene encoding the immunoproteasome subunit β5i as the cause of NNS. This mutation interferes with the assembly of the 20S proteasome in cells expressing immunoproteasomes. The mutation is described as c.602G > T, and results in a Gly201 to Val (G201V) (NM_148919) substitution in the immunoproteasome β5i subunit. Although a heterozygous carrier showed reduced proteasome peptidase activity, carriers had no clinical symptoms. Thus, the NNS phenotype may be due to a reduction in total proteasome enzymatic activity below the threshold necessary for maintaining cellular homeostasis in homozygous individuals.

The *PSMB8* mutation, c.224C > T (Thr75Met), occurs in patients with JMP syndrome (13). Mutant β5i in JMP patients results in a clear reduction in chymotrypsin-like activity only, with no disruption of other peptidase activities. However, the G201V mutation we identified in NNS patients causes losses of all peptidase activity owing to assembly defects and reduced proteasome levels. The T75M mutation is probably rapidly incorporated to the proteasome complex during biogenesis and is specific for chymotrypsin-like activity. The differences between the JMP syndrome and NNS phenotypes, including cytokine production by various cells during inflammatory or noninflammatory states, need to be clarified because these differences could result from a reduction in chymotrypsin-like activity in JMP syndrome or from reductions in chymotrypsin-, trypsin-, and caspase-like activity in NNS. One of the main differences between NNS and JMP syndrome is the level of IFN-γ. IFN-γ levels are increased in JMP patients, but are within the normal range in NNS patients (Fig. S3*A*). The basis for this difference is unclear. It is possible that IFN-γ levels may not increase when all three peptidase activities are inhibited.

We also found increased IP-10 levels in patient sera using ELISA on suspension arrays. There were no significant differences in IP-10 levels between nonstimulated NNS fibroblasts and control cells, although NNS fibroblasts tended to overproduce IP-10 after stimulation with TNF-α (Fig. S3 *B* and *C*). This may reflect the proinflammatory state in NNS cells, or an increased sensitivity to cytokines (31). Because IP-10 is categorized as an inflammatory chemokine produced by various types of cells, it may play an important role in leukocyte homing to inflamed tissues and in perpetuating inflammation in various autoimmune diseases such as rheumatoid arthritis, systemic lupus erythematosus, systemic sclerosis, and multiple sclerosis (32). Thus, IP-10 may enhance inflammation in NNS patients and be associated with the autoantibody production that is occasionally observed.

A single base deletion in the 5′-UTR of hUmp1 causes keratosis linearis with ichthyosis congenita and sclerosing keratoderma (KLICK) syndrome, which is characterized by palmoplantar keratoderma (33) related to proteasome activity. This mutation results in changes in hUmp1 levels and alterations in the epidermal distribution of hUmp1 and proteasomal subunits. It is unclear how the proteasome functions in KLICK syndrome, although it is clear that disturbances in proteasome function cause clinical phenotypes in humans.

Studies in animal models indicate that cells deficient in various immunoproteasome subunits show poor CD8 responses when challenged with epitopes (34, 35) and may display alterations in the T-cell receptor (TCR) repertoire (36). In particular, β5i-deficient mice show increased susceptibility to pathogens, most likely due to the reduced efficiency of antigen presentation by β5i-deficient cells (12). Actually, in NNS patients, unresponsiveness to an intradermally applied purified protein derivative of *Mycobacterium tuberculosis* has been reported; however, there are no documented changes in susceptibility to pathogens, and no abnormalities in the number of any particular T-cell subset have been observed, apart from reduced NK activity (4). Conversely, there are no reports that β5i-deficient mice show the type of systemic inflammation observed in NNS patients.

In general, gene-deficient mice are very useful tools for analyzing the functions of target genes; however, the $\beta5i^{G201V}$ mutation shows a type of "enzymatic dominant-negative interference," which abrogates not only chymotrypsin-like activity (due to the mutation) but also the activity of the entire proteasome (due to defective assembly). Thus, it is not surprising that the phenotype seen in NNS is different from that seen in *Psmb8* knockout mice (12, 20) or after treatment with PR-957 inhibitors (37). Thus, analysis of patients with NNS and JMP syndrome and mice knocked in with these mutations would provide new insights into the function of the immunoproteasome in vivo.

Finally, we observed increased levels of p-p38 in nuclear extracts from NNS peripheral blood lymphocytes (Fig. 4*D*), although it remains unknown precisely how attenuation of proteasome activity causes accumulation of p-p38 in the nucleus. The

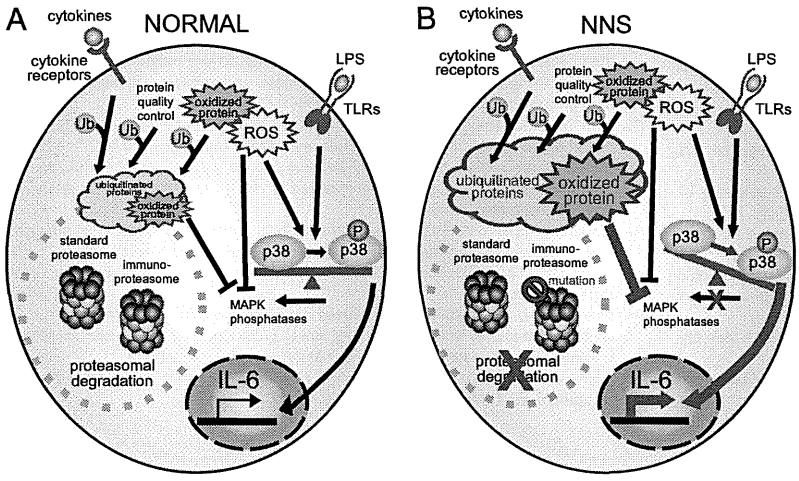

Fig. 5. Schematic model showing induction of inflammation in NNS patients with the *PSMB8* mutation. Our data are based on the scheme proposed by Bulua et al. (40). (*A*) In a normal cell, ubiquitinated or oxidized proteins generated by various stressors, including cytokines, are cleared by proteasomes. (*B*) The ubiquitinated and oxidized proteins accumulate in a cell with the *PSMB8* mutation (NNS cell). ROS and/or oxidized proteins may cause phosphorylation of p-38 to predominate over the nonphosphorylated form by inhibiting MAPK phosphatase or by activating MAPK.

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.

RESEARCH ARTICLE

AMERICAN JOURNAL OF medical genetics PART A

Spectrum of *MLL2* (*ALR*) Mutations in 110 Cases of Kabuki Syndrome

Mark C. Hannibal,[1,2] Kati J. Buckingham,[1] Sarah B. Ng,[3] Jeffrey E. Ming,[4] Anita E. Beck,[1,2] Margaret J. McMillin,[2] Heidi I. Gildersleeve,[1] Abigail W. Bigham,[1] Holly K. Tabor,[1,2] Heather C. Mefford,[1,2] Joseph Cook,[1] Koh-ichiro Yoshiura,[5] Tadashi Matsumoto,[5] Naomichi Matsumoto,[6] Noriko Miyake,[6] Hidefumi Tonoki,[7] Kenji Naritomi,[8] Tadashi Kaname,[8] Toshiro Nagai,[9] Hirofumi Ohashi,[10] Kenji Kurosawa,[11] Jia-Woei Hou,[12] Tohru Ohta,[13] Deshung Liang,[14] Akira Sudo,[15] Colleen A. Morris,[16] Siddharth Banka,[17] Graeme C. Black,[17] Jill Clayton-Smith,[17] Deborah A. Nickerson,[3] Elaine H. Zackai,[4] Tamim H. Shaikh,[18] Dian Donnai,[17] Norio Niikawa,[13] Jay Shendure,[3] and Michael J. Bamshad[1,2,3]*

[1]Department of Pediatrics, University of Washington, Seattle, Washington

[2]Seattle Children's Hospital, Seattle, Washington

[3]Department of Genome Sciences, University of Washington, Seattle, Washington

[4]Department of Pediatrics, The Children's Hospital of Philadelphia, The University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

[5]Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan

[6]Department of Human Genetics, Yokohama City University Graduate School of Medicine, Yokohama, Japan

[7]Department of Pediatrics, Tenshi Hospital, Sapporo, Japan

[8]Department of Medical Genetics, University of the Ryukyus, Okinawa, Japan

[9]Department of Pediatrics, Dokkyo Medical University, Koshigaya Hospital, Saitama, Japan

[10]Division of Medical Genetics, Saitama Children's Medical Center, Saitama, Japan

[11]Division of Clinical Genetics, Kanagawa Children's Medical Center, Yokohama, Japan

[12]Department of Pediatrics, Chang Gung Children's Hospital, Taoyuan, Taiwan, Republic of China

[13]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Hokkaido, Japan

[14]National Laboratory of Medical Genetics, Xiangya Hospital, Central South University, Republic of China

[15]Department of Pediatrics, Sapporo City General Hospital, Sapporo, Japan

[16]University of Nevada School of Medicine, Las Vegas, Nevada

[17]Department of Genetic Medicine, Manchester Academic Health Sciences Centre, University of Manchester, England

[18]Department of Pediatrics, University of Colorado, Denver, Colorado

Received 25 February 2011; Accepted 30 March 2011

Additional supporting information may be found in the online version of this article.

Grant sponsor: National Institutes of Health/National Heart Lung and Blood Institute; Grant number: 5R01HL094976; Grant sponsor: National Institutes of Health/National Human Genome Research Institute; Grant numbers: 5R21HG004749, 1RC2HG005608, 5RO1HG004316, T32HG00035; Grant sponsor: National Institute of Heath/National Institute of Environmental Health Sciences; Grant number: HHSN273200800010C; Grant sponsor: National Institute of Neurological Disorders and Stroke; Grant number: RO1NS35102; Grant sponsor: NIHR Manchester Biomedical Research Centre; Grant sponsor: Ministry of Health, Labour and Welfare of Japan; Grant sponsor: Japan Science and Technology Agency; Grant sponsor: Society for the Promotion of Science; Grant sponsor: Life Sciences Discovery Fund; Grant numbers: 2065508, 0905001; Grant sponsor: Washington Research Foundation; Grant sponsor: National Institutes of Health/National Institute of Child Health and Human Development; Grant numbers: 1R01HD048895, 5K23HD057331.

Mark C. Hannibal, Kati J. Buckingham, and Sarah B. Ng contributed equally to this work.

*Correspondence to:
Michael J. Bamshad, M.D., Department of Pediatrics, University of Washington School of Medicine, Box 356320, 1959 NE Pacific Street, Seattle, WA 98195. E-mail: mbamshad@u.washington.edu

Published online 10 June 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/ajmg.a.34074

Kabuki syndrome is a rare, multiple malformation disorder characterized by a distinctive facial appearance, cardiac anomalies, skeletal abnormalities, and mild to moderate intellectual disability. Simplex cases make up the vast majority of the reported cases with Kabuki syndrome, but parent-to-child transmission in more than a half-dozen instances indicates that it is an autosomal dominant disorder. We recently reported that Kabuki syndrome is caused by mutations in *MLL2*, a gene that encodes a Trithorax-group histone methyltransferase, a protein important in the epigenetic control of active chromatin states. Here, we report on the screening of 110 families with Kabuki syndrome. *MLL2* mutations were found in 81/110 (74%) of families. In simplex cases for which DNA was available from both parents, 25 mutations were confirmed to be de novo, while a transmitted *MLL2* mutation was found in two of three familial cases. The majority of variants found to cause Kabuki syndrome were novel nonsense or frameshift mutations that are predicted to result in haploinsufficiency. The clinical characteristics of *MLL2* mutation-positive cases did not differ significantly from *MLL2* mutation-negative cases with the exception that renal anomalies were more common in *MLL2* mutation-positive cases. These results are important for understanding the phenotypic consequences of *MLL2* mutations for individuals and their families as well as for providing a basis for the identification of additional genes for Kabuki syndrome.

Key words: Kabuki syndrome; *MLL2*; *ALR*; Trithorax group histone methyltransferase

How to Cite this Article:
Hannibal MC, Buckingham KJ, Ng SB, Ming JE, Beck AE, McMillin MJ, Gildersleeve HI, Bigham AW, Tabor HK, Mefford HC, Cook J, Yoshiura K-i, Matsumoto T, Matsumoto N, Miyake N, Tonoki H, Naritomi K, Kaname T, Nagai T, Ohashi H, Kurosawa K, Hou J-W, Ohta T, Liang D, Sudo A, Morris CA, Banka S, Black GC, Clayton-Smith J, Nickerson DA, Zackai EH, Shaikh TH, Donnai D, Niikawa N, Shendure J, Bamshad MJ. 2011. Spectrum of *MLL2* (*ALR*) mutations in 110 cases of Kabuki syndrome.
Am J Med Genet Part A 155:1511–1516.

INTRODUCTION

Kabuki syndrome (OMIM#147920) is a rare, multiple malformation disorder characterized by a distinctive facial appearance, cardiac anomalies, skeletal abnormalities, and mild to moderate intellectual disability. It was originally described by Niikawa et al. [1981] and Kuroki et al. [1981] in 1981, and to date, about 400 cases have been reported worldwide [Niikawa et al., 1988; White et al., 2004; Adam and Hudgins, 2005]. The spectrum of abnormalities found in individuals with Kabuki syndrome is diverse, yet virtually all affected persons are reported to have similar facial features consisting of elongated palpebral fissures, eversion of the lateral third of the lower eyelids, and broad, arched eyebrows with lateral sparseness. Additionally, affected individuals commonly have severe feeding problems, failure to thrive in infancy, and height around or below the 3rd centile for age in about half of cases.

We recently reported that a majority of cases of Kabuki syndrome are caused by mutations in *mixed lineage leukemia 2* (*MLL2*; OMIM#602113), also known as either *MLL4* or *ALR* [Ng et al., 2010]. *MLL2* encodes a SET-domain-containing histone methyltransferase important in the epigenetic control of active chromatin states [FitzGerald and Diaz, 1999]. Exome sequencing revealed that 9 of 10 individuals had novel variants in *MLL2* that were predicted to be deleterious. A single individual had no mutation in the protein-coding exons of *MLL2*, though in retrospect, his phenotypic features are somewhat atypical of Kabuki syndrome. In a larger validation cohort screened by Sanger sequencing, we found *MLL2* mutations in approximately two-thirds of 43 Kabuki cases, suggesting that Kabuki syndrome is genetically heterogeneous.

Herein we report on the results of screening *MLL2* for mutations in 110 families with one or more individuals affected with Kabuki syndrome in order to: (1) characterize the spectrum of *MLL2* mutations that cause Kabuki syndrome; (2) determine whether *MLL2* genotype is predictive of phenotype; (3) assess whether the clinical characteristics of *MLL2* mutation-positive cases differ from *MLL2* mutation-negative cases; and (4) delineate the subset of Kabuki cases that are *MLL2* mutation-negative for further gene discovery studies.

MATERIALS AND METHODS

Subjects

Referral for inclusion into the study required a diagnosis of Kabuki syndrome made by a clinical geneticist. From these cases, phenotypic data were collected by review of medical records, phone interviews, and photographs. These data were collected from five different clinical genetics centers in three different countries and over a protracted period of time and forwarded for review to two of the authors (M.B. and M.H.). Data on certain phenotypic characteristics including stature, feeding difficulties, and failure to thrive was not uniformly collected or standardized. Therefore, we decided to be conservative in our analysis and use only phenotypic traits that could be represented by discrete variables (i.e., presence or absence) and for which data were available from at least 70% of cases. In addition, these clinical summaries were de-identified and therefore facial photographs were unavailable from most cases studied. Written consent was obtained for all participants who provided identifiable samples. The Institutional Review Boards of Seattle Children's Hospital and the University of Washington approved all studies. A summary of the clinical characteristics of 53 of these individuals diagnosed with Kabuki syndrome has been reported previously [Ng et al., 2010].

Mutation Analysis

Genomic DNA was extracted using standard protocols. Each of the 54 exons of *MLL2* was amplified using Taq DNA polymerase (Invitrogen, Carlsbad, CA) following manufacturer's recommendations and using primers previously reported [Ng et al., 2010]. PCR products were purified by treatment with exonuclease I (New England Biolabs, Inc., Beverly, MA) and shrimp alkaline phosphatase (USB Corp., Cleveland, OH), and products were sequenced using the dideoxy terminator method on an automated sequencer (ABI 3130xl). The electropherograms of both forward and reverse strands were manually reviewed using CodonCode Aligner (Dedham, MA). Primer sequences and conditions are listed in Supplementary Table I.

For *MLL2* mutation-negative samples, DNA was hybridized to commercially available whole-genome tiling arrays consisting of one million oligonucleotide probes with an average spacing of 2.6 kb throughout the genome (SurePrint G3 Human CGH Microarray 1 × 1 M, Agilent Technologies, Santa Clara, CA). Twenty-one probes on this array covered *MLL2* specifically. Data were analyzed using Genomics Workbench software according to manufacturer's instructions.

RESULTS

All 54 protein-coding exons and intron–exon boundaries of *MLL2* were screened by Sanger sequencing in a cohort of 110 kindreds with Kabuki syndrome. This cohort included 107 simplex cases (including a pair of monozygotic twins) and 3 familial (i.e., parent-offspring) cases putatively diagnosed with Kabuki syndrome. Seventy novel *MLL2* variants that were inferred to be disease-causing were identified in 81/110 (74%) kindreds (Fig. 1 and Supplementary Table II online). These 81 mutations included 37 nonsense mutations (32 different sites and five sites with recurrent mutations), 3 in-frame deletions or duplications (2 different sites and 1 site with a recurrent mutation), 22 frameshifts (22 different sites), 16 missense mutations (11 different sites and 4 sites with recurrent mutations), and 3 splice consensus site (or intron–exon boundary) mutations. None of these variants were found in dbSNP (build 132), the 1000 Genomes Project pilot data, or 190 chromosomes from individuals matched for geographical ancestry. In total, pathogenic variants were found at 70 sites. Additionally, there were 10 sites at which recurrent mutations were observed.

For 25 simplex cases in which we identified *MLL2* mutations, DNA was available from both unaffected parents, and in each case the mutation was confirmed to have arisen de novo (Supplementary Table II online). These included 14 nonsense, 5 frameshift, 3 missense, 2 splice site mutations, and 1 deletion. De novo events were confirmed at 6 of the 10 sites where recurrent mutations were noted. In addition to the 81 kindreds in which we identified causal *MLL2* mutations, we found two *MLL2* variants in each of three simplex cases. In each case, neither *MLL2* mutation could unambiguously

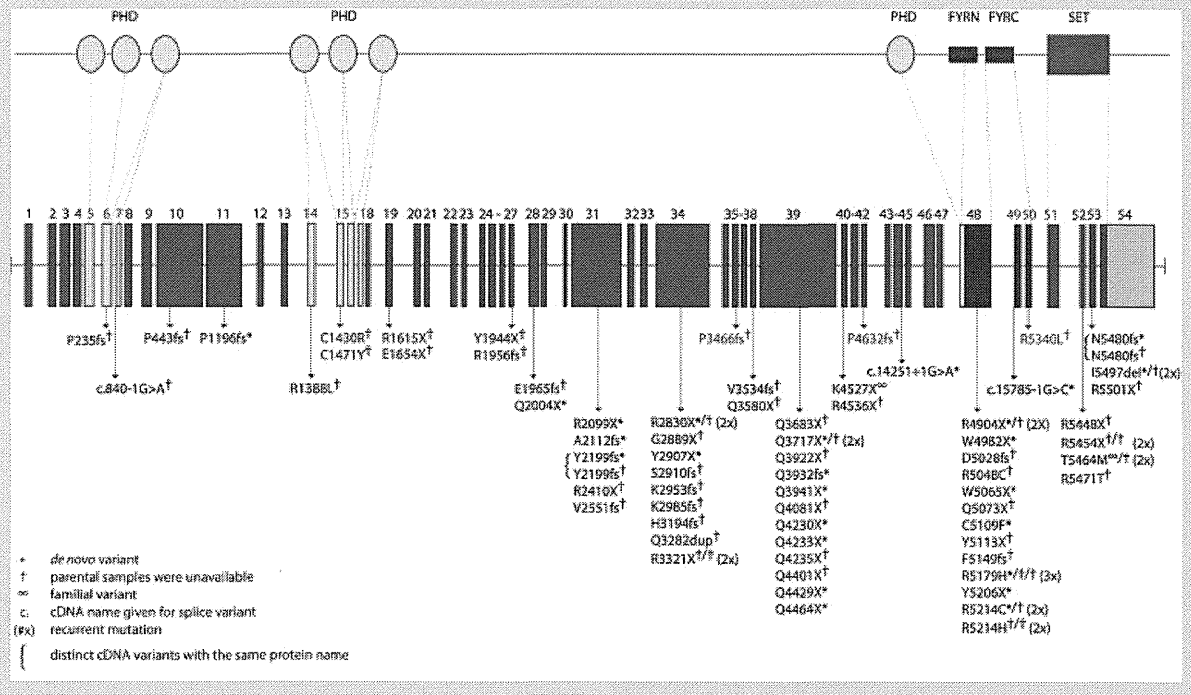

FIG. 1. Genomic structure and allelic spectrum of *MLL2* mutations that cause Kabuki syndrome. *MLL2* is composed of 54 exons that include untranslated regions (orange) and protein coding sequence (blue) including 7 PHD fingers (yellow), FYRN (green), FYRC (green), and a SET domain (red). Arrows indicate the locations of 81 mutations affecting 70 sites found in 110 families with Kabuki syndrome including: 37 nonsense, 22 frameshifts, 16 missense, 3 in-frame deletions/duplications, and 3 splice-site mutations. Asterisks indicate mutations that were confirmed to be de novo and crosses indicate cases for which parental DNA was unavailable. Figure adapted from Ng et al. [2010].

be defined as disease-causing (Supplementary Table II online). In one case, we found both a 21 bp in-frame insertion in exon 39 and a 1 bp insertion in exon 46 predicted to cause a frameshift. However, the unaffected mother also carried the 21 bp insertion suggesting that this is a rare polymorphism, and that the 1 bp deletion is the pathogenic mutation responsible for Kabuki syndrome.

Apparent disease-causing variants were discovered in nearly half (i.e., 22/54) of all protein-coding exons of *MLL2* and in virtually every region known to encode a functional domain (Fig. 1). However, the distribution of variants appeared non-random as 13 and 12 novel variants were identified in exons 48 and 39, respectively. These sites accounted for 25, or more than one-third, of all the novel *MLL2* variants and 31/81 mutations that cause Kabuki syndrome in our cohort. Eleven of the 12 pathogenic variants in exon 39 were nonsense mutations and occurred in regions that encode long polyglutamine tracts.

Four of the families studied herein had two individuals affected with Kabuki syndrome. A pair of monozygous twins with a c.15195G>A nonsense mutation were concordant for mild developmental delay, congenital heart disease, preauricular pits, and palatal abnormalities, but discordant for hearing loss, and a central nervous system malformation. Concordance for mild developmental delay between an affected parent and child was observed in two families with *MLL2* mutations, one with a nonsense mutation, c.13579A>T, p.K4527X, and the other with a missense mutation, c.16391C>T, p.T5464M that was also found in a simplex case. No *MLL2* mutation was found in the remaining affected parent and child pair (Fig. 2).

To examine the relationship between genotype and phenotype, we first compared the frequency of developmental delay, congenital heart disease, cleft lip and/or palate, and structural renal defects between *MLL2* mutation-positive versus *MLL2* mutation-negative cases. No significant difference was observed between groups for three of these four phenotypes (Table Ia). However, renal anomalies were observed in 47% (31/66 cases) of *MLL2* mutation-positive cases compared to 14% (2/14 cases) of *MLL2* mutation-negative cases and this difference was statistically significant ($\chi^2 = 5.1$, df = 1, $P = 0.024$). In 35 cases in two clinical cohorts for whom more complete phenotypic data were available, short stature was observed in 54% (14/26) of *MLL2* mutation-positive cases compared to 33% (3/19 cases) of *MLL2* mutation-negative cases. We also divided the *MLL2* mutation-positive cases into those with nonsense and frameshift mutations and those with missense mutations and compared the frequency of developmental delay, congenital heart disease, cleft lip and/or palate, and structural renal defects between groups. No significant differences were observed between groups (Table Ib).

In 26 independent cases of Kabuki syndrome, including one parent-offspring pair, no *MLL2* mutation was identified. Both persons in the mother–child pair had facial characteristics consistent with Kabuki syndrome (Fig. 2), mild developmental delay, and no major malformations. The mother is of Cambodian ancestry and her daughter is of Cambodian and European American ancestry. In general, most of the *MLL2* mutation-negative Kabuki cases had facial characteristics (Fig. 3) similar to those of the *MLL2* mutation-positive Kabuki cases, and a similar pattern of major malformations (Table I) with the exception of fewer renal abnormalities.

TABLE I. Phenotypic Traits Grouped by *MLL2* Mutation Status (a) and Type (b)

Trait	*MLL2* +	*MLL2* −
Intellectual disability	74/74 (100%)	19/20 (95%)
Mild	51/74 (69%)	10/20 (50%)
Moderate	18/74 (24%)	4/20 (20%)
Severe	4/74 (5%)	3/20 (15%)
Cleft palate, CL/CP	29/72 (40%)	8/18 (44%)
Congenital heart defect	36/71 (51%)	8/19 (42%)
Renal abnormality	31/66 (47%)	2/14 (14%)

Trait	Truncating (N = 59)	Missense (N = 16)
Intellectual disability	54/54 (100%)	15/15 (100%)
Mild	36/54 (67%)	11/15 (73%)
Moderate	13/54 (24%)	4/15 (27%)
Severe	5/54 (9%)	0/15
Cleft palate, CL/CP	23/54 (43%)	3/14 (21%)
Congenital heart defect	30/54 (55%)	4/13 (30%)
Renal anomaly	9/44 (20%)	2/12 (17%)

We screened the *MLL2* mutation-negative cases by aCGH for large deletions or duplications that encompassed *MLL2*. Abnormalities were found in four cases. In one case, a 1.87 kb deletion of chromosome 5 (hg18, chr5:175,493,803–177,361,744) that included *NSD1* and had breakpoints in flanking segmental duplications identical to the microdeletion commonly found in Sotos syndrome, was found. This suggests that this individual has Sotos syndrome, not Kabuki syndrome [Kurotaki et al., 2002]. A second case had a novel 977-kb deletion of chromosome 19q13 (hg18, chr19:61,365,420–62,342,064) encompassing 20 genes. The majority of genes within the deleted region are zinc finger genes, some of which are known to be imprinted in both human and mouse. A third case had a complex translocation t(8;18)(q22;q21). Finally, a fourth case was found to have extra material for the entire chromosome 12. Average log 2 ratio across chromosome 12 was 0.49, most likely representing mosaic aneuploidy of chromosome 12. No aCGH abnormalities were observed in 21 cases and aCGH failed for one case.

DISCUSSION

We have expanded the spectrum of mutations in *MLL2* that cause Kabuki syndrome and explored the relationship between *MLL2* genotype and some of the major, objective phenotypic characteristics of Kabuki syndrome. The majority of variants found to cause Kabuki syndrome are either novel nonsense or frameshift mutations, and appear to arise de novo. While mutations that cause Kabuki syndrome are found throughout the *MLL2* gene, there appear to be at least two exons (39 and 48) in which mutations are identified with a considerably higher frequency. Mutations in these two exons account for nearly half of all mutations found in *MLL2*, while the length of these exons represents ~24% of the *MLL2* open reading frame (ORF). Furthermore, exon 48, the exon in which mutations are most common, comprises only ~7% of the

FIG. 2. Facial photographs of mother and daughter with Kabuki syndrome in whom no causative mutation in *MLL2* was identified. Both have mild developmental delay and no known major malformations.

FIG. 3. Facial photographs of four children diagnosed with Kabuki syndrome in whom no causative mutation in *MLL2* was found. The photograph in the upper left was reprinted from Ng et al. [2010].

MLL2 ORF. Exon 39 contains several regions that encode long polyglutamine tracts suggesting the presence of a mutational hotspot, although no such explanation is obvious for exon 48. A stepwise approach in which these regions are the first screened might be a reasonable approach to diagnostic testing. However, capture of all introns, exons, and nearby *MLL2* regulatory regions followed by next-generation sequencing would be more comprehensive and likely to be less costly over the long term.

Comparison of four of the objective clinical characteristics of *MLL2* mutation-negative versus *MLL2* mutation-positive cases allowed us to explore both the relationship between *MLL2* genotype and Kabuki phenotype and the phenotype of *MLL2* mutation-negative cases. Overall, the clinical characteristics of *MLL2* mutation-positive cases did not differ significantly from *MLL2* mutation-negative cases with the exception that renal anomalies were more common in *MLL2* mutation-positive cases. Similarly, we observed no significant phenotypic—including the severity of developmental delay—differences between individuals grouped by mutation type. However, the phenotypic data available to us for analysis was limited and, for many cases, we lacked specific information about each malformation present. Furthermore, the most typical phenotypic characteristic, the distinctive facial appearance, was not compared in detail between cases although it would be of interest to study facial images "blinded" to mutation status to investigate its power to predict genotype. Analysis of genotype–phenotype relationships using both a larger set of Kabuki cases, and with access to more comprehensive phenotypic information would be valuable.

No *MLL2* mutation could be identified in 26 of the cases referred to us with a diagnosis of Kabuki syndrome. In three of these cases, aCGH identified structural variants that could be of clinical significance although additional investigation is required. A fourth case had the classical deletion observed in individuals with Sotos syndrome, and in retrospect it appears that this case was included in the cohort erroneously. The 22 remaining cases, including 1 parent-offspring pair, represent individuals with fairly classic phenotypic features of Kabuki syndrome without a *MLL2* mutation. This observation suggests that Kabuki syndrome is genetically heterogeneous. To this end, in these 22 cases, we sequenced the protein-coding exons of *UTX*, a gene that encodes a protein that directly interacts with *MLL2* but no pathogenic changes were found (data not shown). Exome sequencing of a subset of these *MLL2* mutation-negative cases to identify other candidate genes for Kabuki syndrome is underway.

Whether Kabuki syndrome is the most appropriate diagnosis for the *MLL2* mutation-negative cases is unclear. Some of the *MLL2* mutation-negative cases appear to have a facial phenotype that differs somewhat from that of the *MLL2* mutation-positive cases. Whether these *MLL2* mutation-negative cases diagnosed by expert clinicians should be considered Kabuki syndrome, a variant thereof, or a separate disorder remains to be determined. Our opinion is that