DNA Research **19**, 143–152, (2012)
Advance Access Publication on 24 January 2012

doi:10.1093/dnares/dsr047

Detection of Base Substitution-Type Somatic Mosaicism of the *NLRP3* Gene with >99.9% Statistical Confidence by Massively Parallel Sequencing

Kazushi Izawa[1,†], Atsushi Hijikata[2,†], Naoko Tanaka[1], Tomoki Kawai[1], Megumu K Saito[3], Raphaela Goldbach-Mansky[4], Ivona Aksentijevich[5], Takahiro Yasumi[1], Tatsutoshi Nakahata[3], Toshio Heike[1], Ryuta Nishikomori[1,*], and Osamu Ohara[2,6,*]

Department of Pediatrics, Kyoto University Graduate School of Medicine, 54 Shogoin Sakyo, Kyoto 606-8507, Japan[1]; Laboratory for Immunogenomics, RIKEN Research Center for Allergy and Immunology, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho Tarumi-ku, Yokohama, Kanagawa 230-0045, Japan[2]; Clinical Application Department, Center for iPS Cell Research and Application (CiRA), Kyoto University, Kyoto, Japan[3]; Translational Autoinflammatory Disease Section NIH/NIAMS, Bethesda, MD, USA[4]; The National Human Genome Research Institute, Bethesda, MD, USA[5] and Department of Human Genome Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan[6]

*To whom correspondence should be addressed. Tel. +81 75-751-3291 (R.N.); +81 438-52-3913/+81 45-503-9696 (O.O.). Fax. +81 75-752-2361 (R.N.); +81 438-52-3914/+81 45-503-9694 (O.O.). Email: rnishiko@kuhp.kyoto-u.ac.jp (R.N.); ohara@kazusa.or.jp/oosamu@rcai.riken.jp (O.O.)

Edited by Mitsuo Oshimura
(Received 18 November 2011; accepted 26 December 2011)

Abstract

Chronic infantile neurological cutaneous and articular syndrome (CINCA), also known as neonatal-onset multisystem inflammatory disease (NOMID), is a dominantly inherited systemic autoinflammatory disease and is caused by a heterozygous germline gain-of-function mutation in the *NLRP3* gene. We recently found a high incidence of *NLRP3* somatic mosaicism in apparently mutation-negative CINCA/NOMID patients using subcloning and subsequent capillary DNA sequencing. It is important to rapidly diagnose somatic *NLRP3* mosaicism to ensure proper treatment. However, this approach requires large investments of time, cost, and labour that prevent routine genetic diagnosis of low-level somatic *NLRP3* mosaicism. We developed a routine pipeline to detect even a low-level allele of *NLRP3* with statistical significance using massively parallel DNA sequencing. To address the critical concern of discriminating a low-level allele from sequencing errors, we first constructed error rate maps of 14 polymerase chain reaction products covering the entire coding *NLRP3* exons on a Roche 454 GS-FLX sequencer from 50 control samples without mosaicism. Based on these results, we formulated a statistical confidence value for each sequence variation in each strand to discriminate sequencing errors from real genetic variation even in a low-level allele, and thereby detected base substitutions at an allele frequency as low as 1% with 99.9% or higher confidence.

Key words: next generation sequencing; mosaicism; DNA diagnosis; chronic infantile neurological cutaneous and articular syndrome

1. Introduction

Chronic infantile neurological cutaneous and articular syndrome (CINCA; MIM #607115), also known as neonatal-onset multisystem inflammatory disease (NOMID), is a dominantly inherited autoinflammatory disease that is characterized by neonatal onset and a triad of symptoms, including an urticarial-like skin rash, neurological manifestations, and arthritis/arthropathy.[1–3] Patients often experience

[†] These authors contributed equally to this work.

recurrent fever and systemic inflammation. CINCA/NOMID is the most severe clinical phenotype in the spectrum of cryopyrin-associated periodic syndromes (CAPS), which also include two less severe but phenotypically similar syndromes, familial cold autoinflammatory syndrome (FCAS; MIM #120100), and Muckle–Wells syndrome (MWS; MIM #191900). CAPS are caused by mutations in the *NLRP3* gene, which is a member of the Nod-like receptor (NLR) family of the innate immune system.[4–6]

Approximately 60% of CINCA/NOMID patients carry heterozygous germline missense mutations in the *NLRP3* coding region (mutation-positive patients).[7] More than 80 different disease-causing mutations have been reported to date.[8] However, the remaining clinically diagnosed CINCA/NOMID patients (∼40%) show no heterozygous germline *NLRP3* mutation based on conventional DNA sequencing-based genetic analyses (mutation-negative patients). In a previous international collaborative study, we found that there was a high incidence of somatic *NLRP3* mosaicism in mutation-negative CINCA/NOMID patients worldwide.[9] The level of mosaicism ranges from 4.2 to 35.8% (median = 10.2%). Rapidly diagnosing somatic *NLRP3* mosaicism is important to ensure proper treatment. However, the conventional approach used to identify somatic mosaicism of the *NLRP3* gene is time and labour intensive due to the subcloning of the *NLRP3* exon polymerase chain reaction (PCR) products, hereafter designated as amplicons, followed by capillary DNA sequencing of more than 100 subclones for each patient. Thus, this approach is not suitable to routinely diagnose somatic mosaicism of the *NLRP3* gene and additional labour and time will be required to reliably identify somatic mosaicism that occurs at a lower rate. The aim of the present study was to establish a new method that can be used to reliably diagnose somatic mosaicism using the *NLRP3* gene as a model. Massively parallel DNA sequencing (MPS) technology is an obvious method of choice to identify somatic mosaicism, and this approach has been already reported by other groups.[10–12] However, a well-known caveat of MPS is the high rate of sequencing errors, which cannot be disregarded when identifying low-level somatic mosaicism. To our knowledge, there have been no reports of a reliable method to discriminate MPS sequencing errors from somatic mosaicism with statistical confidence.

In this study, we first analysed the patterns of sequencing errors in *NLRP3* coding exons at a single-residue resolution by MPS using a Roche 454 GS-FLX sequencer and then constructed an error rate map for each base position in the *NLRP3* exons. Based on the error rate map, we could formulate a discrimination pipeline of somatic mosaicism from sequencing errors and thereby detect new somatic mosaicism in mutation-negative CINCA/NOMID patients, whose somatic mutations were subsequently confirmed by subcloning and Sanger sequencing. This approach can also be generally used to identify low-level somatic mosaicism in other genes.

2. Patients and methods

2.1. Patients and DNA materials

Patients were clinically diagnosed with CAPS by their referring physicians and the *NLRP3* gene was examined using the conventional Sanger sequencing method. DNA samples were obtained from Japanese *NLRP3* somatic mosaic patients ($n = 5$) who have been previously described,[9,13] CAPS patients ($n = 5$) with heterozygous *NLRP3* mutations, and healthy donors ($n = 50$). Genomic DNA samples from mutation-negative CINCA/NOMID patients ($n = 10$) were obtained from the National Institute of Health, Bethesda, USA. To generate DNA samples with no mosaicism, we constructed a set of subcloned plasmids containing each exon and its flanking intronic regions in the *NLRP3* gene from healthy donor genomic DNA using a Topo TA cloning kit (Invitrogen, San Diego, CA, USA). The cloned plasmids containing each exon and the flanking regions were validated by Sanger sequencing. Written informed consent was obtained from all the patients and their families. The study was approved by the ethical committees of Kyoto University and Kazusa DNA Research Institute and was conducted in accordance with the Helsinki Declaration.

2.2. *MPS of* NLRP3 *gene amplicons*

Genomic DNA samples were extracted from whole blood or peripheral blood mononuclear cells as previously described. We used a two-step PCR assay and pooled sample libraries for MPS. To cover the entire *NLRP3* coding exonic regions and flanking intronic regions, 14 amplicons were designed to be as long as an average read length for a 454 GS-FLX sequencer (up to 450 bases) and then amplified from each genomic DNA sample (Fig. 1A). The sequences of the PCR primers that were used to generate these 14 amplicons are provided in Supplementary Table S1. The upper and lower amplicon-specific primer sequences were flanked by common 15-base adapter sequences (TGTAAAACGACGGCC and GGAAACAGCTATGAC for the upper and lower primers, respectively) at the 5′ end in order to fuse the primer-binding sequence for MPS in the second-step PCR. The first PCR amplifications were performed in 50-μl reactions using 30 ng of genomic DNA, 1× PrimeSTAR GXL buffer, 0.2 mM of each dNTP,

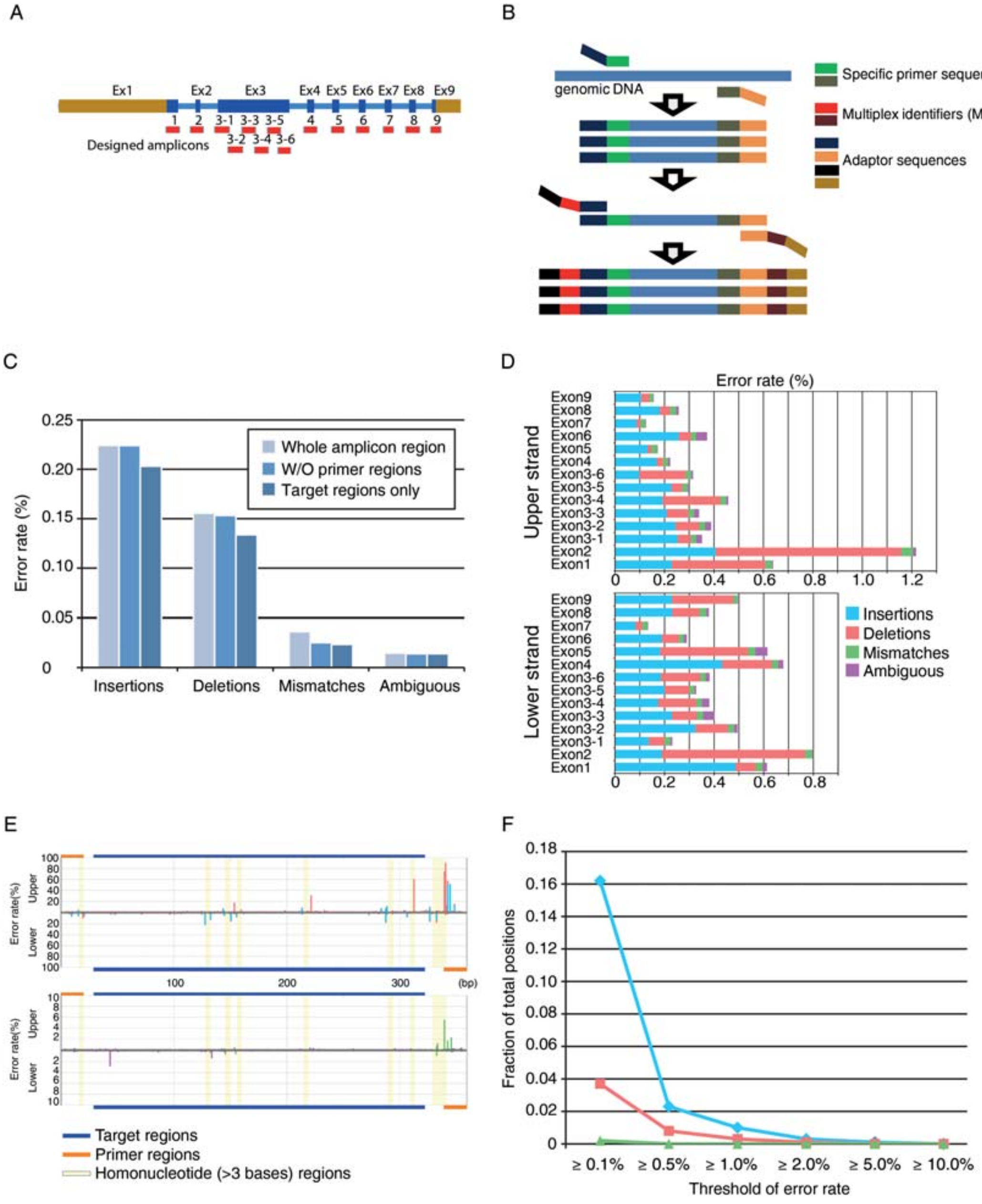

Figure 1. The amplicon analysis for *NLRP3* exons and its error rate. (A) Exon–intron structure of the *NLPR3* gene. Thick and thin rectangles depict exons and introns, respectively. Blue thick rectangles indicate the CDS region. The 14 designed amplicons (red) for nine exons are shown under the exon–intron structure. (B) Amplicon design schema. (C) Error rate for each error category in the region of entire amplicon (pale blue), that without designed primer regions (light blue), and the target regions (CDS + flanking intron; dark blue), respectively. (D) Strand-wise error rate for each amplicon. (E) Error rates along the amplicon sequence of exon 1 in each strand for insertions and deletions in the upper panel and mismatches and ambiguous base calls in the lower panel. The orange and blue lines depict the primer and target regions, respectively. The yellow shaded area depicts the homonucleotide ($n > 3$) region. The colour representation for the bars is the same as (D). (F) Co-occurrence error rate in both strands. The fraction of positions where a certain error occurred with the error rate for insertions, deletions, and mismatches. The colour representation is the same as in (D) and (E).

12.5 pmol of each forward and reverse primer, and 1.25 U of PrimeSTAR GXL DNA polymerase (Takara Bio, Shiga, Japan). The thermal cycling profile consisted of an initial denaturation step at 98°C for 1 min, followed by 28–32 cycles of 10 s denaturation at 98°C, 15 s of annealing at 60°C, and a 30 s extension at 68°C. The lengths of the PCR products ranged from 291 to 421 bp. The second PCR amplifications were performed using primers with adapter sequences at the 3′ end and Multiplex Identifier (MID) sequences at the 5′ end (Fig. 1B), which was used as a tag for each sample. The PCR reactions were performed in 50-μl volumes using 0.5 μl of the first PCR products, 1× PrimeSTAR GXL buffer, 0.2 mM of each dNTP, 12.5 pmol of each forward and reverse primer, and 1.25 U of PrimeSTAR GXL

DNA polymerase to attach the anchor sequences for MPS. The thermal cycling profile consisted of an initial denaturation step at 98°C for 20 s, followed by 20 cycles of 10 s denaturation at 98°C, 15 s of annealing at 60°C, and a 40 s extension at 68°C.

After confirming the amount and integrity of the PCR products by agarose gel electrophoresis, we mixed virtually equal amounts of the respective PCR amplicons that were generated using the same genomic DNA and applied the samples to a 454 Genome Sequencer (GS)-FLX system (Roche Diagnostics Corp., USA). All amplicons were amplified by emPCR and sequenced together in a multiplex fashion. MPS on this platform was performed as instructed by Roche. The sequencing reads from each of the pooled libraries were identified by their MID tags.

2.3. *Sequence data analysis*

The sequence read data were generated using GS RunProcessor ver.2.5.3 with default settings. Reads were sorted according to the MID tag sequences and were mapped to the reference amplicon sequences using the BLAT program[14] with the '-fine' option. In order to identify positions where the bases in a read differed from those in the reference sequence, each read was aligned to its reference sequence with the dpAlign module in the BioPerl package (http://www.bioperl.org/). The 454 pyrosequencing-related errors were categorized as insertions, deletions, mismatches, or ambiguous base calls. When aligning sequences, insertions/deletions are allocated based on the sequence context and strand orientation. To eliminate alignment artefacts due to insertion/deletion positions, the lower strand reads were converted to the reverse complement sequence, i.e. keeping the same strandness as the upper strand reads, when aligned with the reference sequence. A sequence error was defined as discordance in an equivalent position between the reference and control (from the 49 healthy individuals and a cloned plasmid vector). The error rate for a specified category was defined as the number of errors divided by the total number of bases in a read. The error rates of a base position on each strand were calculated from 50 control samples.

2.4. *Confirmation of somatic mosaicism of the* NLRP3 *gene by subcloning and subsequent capillary DNA sequencing*

To confirm the somatic mutational frequency that was identified based on the 454 sequencing data, we subcloned the PCR products and performed capillary DNA sequencing as previously described.[9] A Topo TA cloning kit (Invitrogen, San Diego, CA, USA) was used to subclone each of the 14 amplicons.

2.5. *Functional analysis*

To determine whether the identified *NLRP3* mutants are disease-causing, we assessed both ASC [apoptosis-associated speck-like protein containing a caspase recruitment domain; PYCARD, an approved symbol from the HUGO Gene Nomenclature Committee (HGNC) database]-dependent NF-κB activation in HEK293FT cells and transfection-induced cell death in THP-1 cells, a human monocytic cell line, as previously described.[9,13,15] cDNAs encoding carboxy-terminal green fluorescent protein (GFP)-tagged NLRP3 and its mutants were subcloned into pcDNA5/TO (Invitrogen). Before being introduced into THP-1 cells (10^6) using a Cell Line Nucleofector Kit V (Amaxa Biosystems, Cologne, Germany), phorbol myristate acetate (10 ng/ml) was added to enhance transient expression of *NLRP3* gene with minimizing spontaneous cell death.[15] Four hours after the introduction of plasmids (0.5 μg), cell death of GFP-positive THP-1 cells was measured by flow cytometry.

Expression plasmids for NLRP3 and ASC in the pEF-BOS vector background have been previously described.[13] HEK293FT cells (10^5) were transfected using TransIT-293 Transfection Reagent (Milus Bio, Madison, WI, USA) with an NF-κB reporter construct (pNF-κB-luc; 20 ng; BD Biosciences Clontech, Palo Alto, CA, USA), an internal control construct (pRL-TK; 5 ng; Toyo Ink, Tokyo, Japan), and wild-type or mutant NLRP3 expression plasmid (20 ng) in the presence or absence of ASC expression plasmid (20 ng). The amounts of total plasmid DNA used for transfection experiments were kept constant by adding pEF-BOS vector DNA. Twenty-four hours later, the transfected cells were harvested and subjected to dual luciferase assay by which the ability of each construct to induce NF-κB activation was assessed as previously described.[9]

3. Results

3.1. *Construction of base- and strand-specific error rate maps of* NLRP3 *exons from the MPS data of 50 control samples*

Errors in sequence reads generated by a Roche 454 GS-FLX sequencer are not randomly distributed along the sequence and depend on various factors.[16] Although this is a well-known characteristic of 454 sequencing, the occurrence pattern of these errors has not been explored in detail simply because these sequencing errors are considered noise that can be filtered out in most cases. However, it is highly critical to understand the occurrence pattern of sequencing errors on the MPS platform because low-level somatic mosaicism might appear at a rate close to that of sequencing errors. To address this, we collected

~1 million sequence reads using the 454 GS-FLX sequencer for 14 amplicons of *NLRP3* exons from 50 control samples that were thought to be free from somatic mosaicism, and ~94% of those reads were mapped to one of the reference *NLRP3* exon sequences. The number of sequencing depths for each amplicon of each sample on each strand was between 65 and 2139 (mean = 565.3, Supplementary Table S2). We found that the average error rate for each mutation category (insertion, deletion, mismatch, and ambiguous base calls) at each base position on each strand of the amplicons in the control samples was 0.22, 0.16, 0.036, and 0.014%, respectively (Fig. 1C). These values were consistent with those reported in a recent study on the error rates with 454 sequencing data.[16] The sequencing error in the 454 GS-FLX system tends to occur at the beginning and end of the reads,[11,16] and we confirmed this trend in our amplicon sequencing data (Supplementary Fig. S1). Moreover, after removing the end regions of the read sequences, we found that the error rates of the target regions for each category were 0.20, 0.134, 0.023, and 0.014%, respectively (Fig. 1C and Supplementary Table S3). When generating the amplicon sequences for the *NLRP3* exons, the target sequence (CDS region and flanking intron in 10-bp length) was designed to be 300−400 bp and not adjacent to primer sequences in order to obtain relatively low sequencing error rates (Fig. 1C). However, when the base- and strand-specific error rates of the respective amplicons were compared, we noticed that there were large variations in the error rate among amplicons in a strand-specific manner (Fig. 1D). We further examined the occurrence pattern of sequencing errors, as shown in Fig. 1E; the average sequencing error rates at each base in the 50 control amplicons are shown in a bar graph, where the bars in the upper or lower direction show the sequence error rates at the base position on the upper or lower strand of the amplicons, respectively. As evident in Fig. 1E, the error rates at most residues were low (<1%) with some hotspots for each type of error. Most of the insertion/deletion errors preferentially occurred at a homonucleotide region (yellow regions in Fig. 1E) as previously described,[17] but it was not always the case for all of homonucleotide regions. We could not find any tight relationship between other sequence patterns and the error rate. In addition, there was almost no position where sequencing errors occurred at a similar rate on both strands. This is more clearly shown in Fig. 1F, which indicates the numbers of positions with sequence variations in both strands that were higher than the threshold along the horizontal axis. These results indicate that the sequence errors can be discriminated from real genetic alterations when the sequence is read in both directions. However, it is important to keep in mind that PCR errors are not distinct from real genetic alterations. We did not observe any base substitution at a rate higher than 1% in our experiments (Fig. 1F), and the overall PCR error rate under MPS conditions was lower than 1% as long as a high-fidelity DNA polymerase was used to generate the amplicons.

Because Gilles *et al.*[16] recently reported that the occurrence of sequencing errors using the Roche 454 GS-FLX DNA sequencer depends on various factors, we first examined variations in the sequencing error rates of *NLRP3* exons among samples in the same run. For each mutation category, we found a similar trend in the error distribution rate in the amplicon sequences among the control samples (Supplementary Figs S2−S4). We confirmed that, for almost all residues, the error rate distributions among the 50 control samples fitted a Poisson distribution (data not shown). We next examined the run-to-run variation of the sequencing error rate for *NLRP3* exons. For this purpose, we performed an additional MPS run with seven amplicons (exons 3, 4, and 6) that were newly prepared and compared the number and positions of the sequencing errors between two independent sequencing runs. Out of 1993 base positions in the target regions, there was a low occurrence rate of mismatch errors in both runs and this seemed to fit a Poisson distribution. However, insertion/deletion errors (>1% error rate) were observed at ~100 base positions (<5% in the target regions) in each run, and only a half of these errors were shared between both runs (Table 1).

Table 1. Run-to-run variations in the error occurrence (>1% frequency)

Error category	Upper strand			Lower strand			All[a]
	First run	Second run	Overlap	First run	Second run	Overlap	
Insertions	63	73	42	76	96	52	10
Deletions	36	44	24	29	65	20	2
Mismatches	0	0	0	3	0	0	0
Ambiguous base calls	6	8	6	12	10	10	0

[a]The number of positions where the error rates in each category were commonly >1% for both strands in two independent runs.

This indicated that the occurrence of insertion/deletion errors was considerably affected by the run conditions (probably due to variations in the absolute signal strengths of pyrosequencing). Thus, as previously reported, the detection of insertion/deletion mutations by MPS on the 454 GS-FLX system was quite error-prone at least at a limited number of residues. However, the results also implied that false-positive mosaic mutations could be avoided by considering the sequencing data for both strands because these run-dependent insertion/deletion errors occur only in a single strand. Taken together, we conclude that the obtained sequence error map is stable and sufficiently robust to discriminate substitution sequencing errors from low-level mosaicism.

3.2. Discrimination formula for detection of somatic mosaicism with statistical confidence

We next examined known SNPs, known heterozygous mutations and somatic mosaic mutations of CAPS patients using MPS. All of these variations appeared on both strands at the expected allele frequencies as shown in Fig. 2, again indicating that filtering the strand-specific sequence variations is unlikely to eliminate real genetic variations.

Based on the experimentally observed sequencing errors with the 454 GS-FLX system described above, we established a discrimination formula to detect low-level somatic mosaicism as follows. In previous studies, the number of reads with the sequence error of a certain category in a sequence position was modelled based on the Poisson distribution with two parameters λ and k where the expected number of reads containing an error and the observed number of reads containing a sequence alteration, respectively, are as shown below[18]:

$$\mathrm{Pois}(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \qquad (1)$$

This model assumes that the error rate is constant across the different sequence regions but our data described above pointed out that the sequence error rate varies with the sequence content.[19] Thus, we introduced a position- and strand-specific error rate $q_{i,j,d}$ for a certain error category j in amplicon position i with strand d based on the sequencing data from 50 control samples. With the error rate $q_{i,j,d,}$ the upper probability (P) that the number of reads (R) with a certain sequence alteration of category j in position i is equal or greater than the number of observed reads r out of N reads with a sequenced direction d for an unknown sample was defined as:

$$P(R \geq r_{i,j,d}|\lambda_{i,j,d}) = 1 - \sum_{k=0}^{r-1} \frac{\lambda_{i,j,d}^{k} \mathrm{e}^{-\lambda_{i,j,d}}}{k!}, \qquad (2)$$

where, $\lambda_{\mathrm{i,j,d}} = N_{i,d} \times q_{i,j,d}$.

For the mismatch error rate, we did not consider the type of base substituted in an amplicon position in this study. We took $(1 - P)$ as a measure of the statistical confidence of the data and conventionally set a threshold of the statistical confidence to be 99.9%. In other words, if P-value was <0.001, the sequence alteration was considered to be a real sequence variation, not an error. For the final identification of real genetic variation with low-level somatic mosaicism, we determined that both of the P-values for the ith residue in the upper and lower strands must be smaller than the threshold.

To evaluate the lower detection limit for the allele frequencies of somatic mosaicism based on the statistical formulation shown above, we generated a series of known allele frequencies by diluting DNA from CAPS patients carrying heterozygous *NLRP3* mutations (c.1043C>T, c.1316C>T, and c.1985T>C) with DNA from normal donors carrying the wild-type *NLRP3* gene. In the dilution series, the mutant allele frequencies were adjusted to be 10, 5, 3, 2, 1, and 0.5% (Table 2). The data indicated that somatic mosaicism at these sites and at an allele frequency $\geq$1% could be convincingly detected with statistical significance ($P < 0.001$) if more than 350 reads for each strand were obtained for an amplicon. We also applied this statistical method to detect somatic mosaicism in patients with known low-level mosaic mutations described above and confirmed that all of

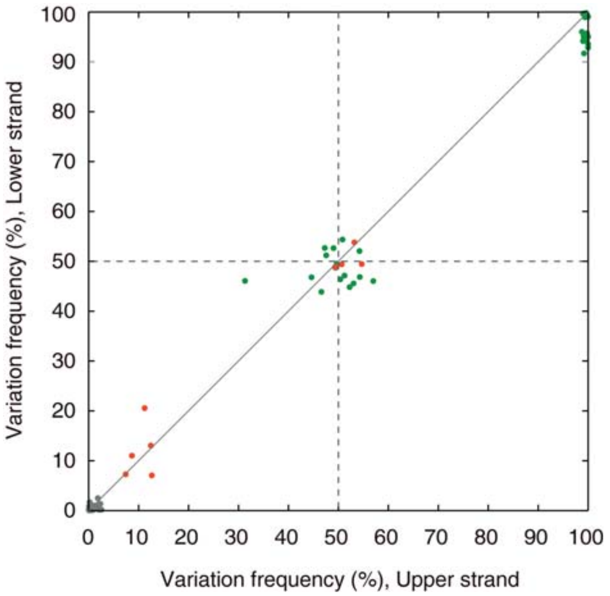

Figure 2. Scatter plot of the observed frequency variation in both strands. The colours depict known SNPs (green), heterozygous and mosaic mutations (orange) and errors (grey).

Table 2. Evaluation of the lower detection limit for mosaicism with three sets of dilution series

Mutation	Dilution (%)	Upper strand				Lower strand			
		Total reads	Mutant reads	%Mutant	*P*-value	Total reads	Mutant reads	%Mutant	*P*-value
c.1043C>T; p.Thr348Met	10.0	724	61	8.43	8.62E−130	520	57	10.96	1.73E−117
	5.0	453	24	5.30	2.86E−47	372	15	4.03	1.26E−25
	3.0	876	27	3.08	1.16E−46	757	21	2.77	6.83E−32
	2.0	737	10	1.36	1.05E−14	645	7	1.09	8.68E−09
	1.0	715	9	1.26	4.73E−13	624	4	0.64	1.11E−04
	0.5	1025	7	0.68	1.15E−14	756	3	0.40	3.22E−03[a]
c.1431C>A; p.Asn477Lys	10.0	542	65	11.99	1.22E−113	346	24	6.94	6.84E−49
	5.0	491	30	6.11	1.13E−44	356	17	4.78	2.42E−32
	3.0	487	21	4.31	1.26E−28	374	19	5.08	1.78E−36
	2.0	577	18	3.12	2.78E−22	495	9	1.82	4.57E−14
	1.0	491	4	0.82	9.17E−04	354	5	1.41	7.34E−08
	0.5	483	0	0	NA	424	3	0.71	NA
c.1985T>C; p.Met662Thr	10.0	658	79	12.01	1.13E−179	643	74	11.51	4.64E−167
	5.0	643	31	4.82	2.56E−59	608	33	5.43	9.96E−65
	3.0	777	27	3.48	4.65E−48	704	29	4.12	1.26E−53
	2.0	929	21	2.26	7.59E−34	835	15	1.80	3.92E−23
	1.0	735	17	1.09	2.74E−11	709	9	1.27	4.06E−13
	0.5	702	2	0.29	3.90E−03[a]	590	1	0.17	1.37E−01[a]

[a]Not significant.

Table 3. Potential mosaic mutations detected in patients with unknown mutations

Patient ID	Amplicon #	Variation		% Variation frequency		*P*-value		dbSNP	State
				Forward	Reverse	Forward	Reverse		
P1	Exon3_2	c.907G>C	p.Asp303His	7.12	11.56	3.0E−44	1.7E−84	rs121908153	Known
P2	Exon3_5	c.1699G>A	p.Glu567Lys	5.94	5.79	2.0E−69	8.9E−47	—	Known
P3	Exon3_5	c.1699G>A	p.Glu567Lys	18.28	15.33	0.0E+00	1.0E−312	—	Known
P4	Exon3_2	c.906C>A	p.Phe302Leu	9.78	9.70	1.7E−86	2.2E−122	—	Novel

the mutations could be detected with statistical significance without any false positives (data not shown).

3.3. *Detection and characterization of* NLRP3 *somatic mosaicism using the MPS platform*

To demonstrate the power of this approach in practice, we applied our new pipeline for 10 CINCA/NOMID patients in whom we failed to detect mutations in the *NLRP3* gene using a conventional direct DNA sequencing approach. The mutations detected by the analysis formulated using the MPS platform in this study are listed in Table 3. We successfully identified four out of the 10 patients with *NLRP3* somatic mosaicism, which was confirmed by subcloning and Sanger sequencing. The nucleotide substitutions were as follows (parentheses indicate the corresponding amino acid change): c.907G>C (p.Asp303His), c.1699G>A (p.Glu567Lys) in two patients, and c.906C>A (p.Phe302Leu). The frequencies of mosaicism identified in these patients by the MPS approach were consistent with those that were identified by the subcloning and subsequent capillary DNA sequencing method (data not shown). Both c.907G>C and c.1699G>A variants were reported as CINCA/NOMID-associated mutations in Infevers database (http://fmf.igh.cnrs.fr/ISSAID/infevers/) and in the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/).[8]

Because the NLRP3 p.Phe302Leu mutation was novel and not detected in the 50 healthy controls, we performed an *in vitro* functional analysis to see the effect of p.Phe302Leu on the protein function. We used two different *in vitro* transfection experiments,

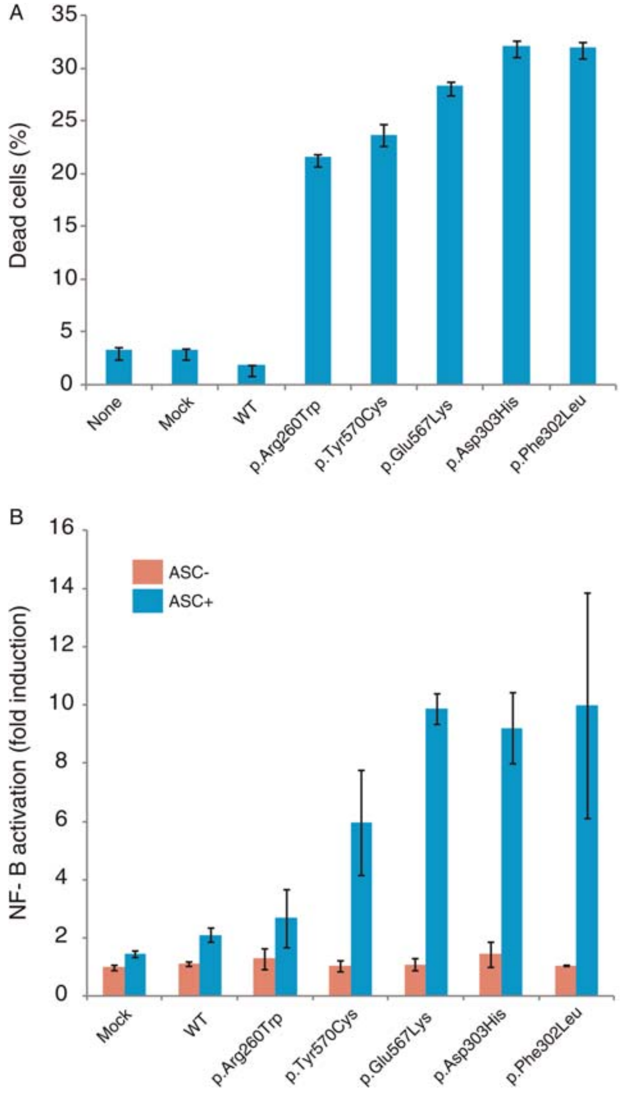

Figure 3. *In vitro* functional analysis of the identified *NLRP3* mosaic mutations. (A) Rapid cell death in transfected THP-1 cells. A GFP-fused wild-type or mutant *NLRP3* was transfected into THP-1 cells and incubated with PMA (10 ng/ml) for 4 h. The percentage of dead cells (7-amino-Actinomycin D [7-AAD]-positive) among the GFP-positive cells is shown. Data represent the means ± SD of triplicate experiments and are representative of two independent experiments. The data for previously reported mutations as well as the mutations found in this study are shown. (B) ACS-dependent NF-κB activation in transfected HEK293FT cells. HEK293FT cells were co-transfected with wild-type or mutant *NLRP3* in the presence or absence of ASC. NF-κB induction is shown as the fold-change compared with cells that were transfected with a control vector without ASC (set equal to one). Values are the means ± SD of triplicate experiments, and the data are representative of three independent experiments. The data for previously reported mutations (p.Arg260Trp and p.Tyr570Cys) and the mutations found in this study are shown. For each mutation, the data obtained in the presence and absence of ASC are shown. These findings identified p.Phe302Leu as a novel disease-causing mutation.

the rapid cell death in transfected THP-1 cells and the ASC-dependent NF-κB activation in transfected HEK293FT cells (Fig. 3A and B, respectively). Both assays clearly showed that p.Phe302Leu was a disease-causing mutation similar to known CINCA/NOMID-associated pathogenic mutations (p.Asp303His and p.Glu567Lys).[9]

4. Discussion

Although the somatic mutation rate at the nucleotide level *in vivo* was difficult to quantitatively measure due to the complexity of the genome and laborious molecular detection processes, recent advances in MPS technologies have allowed us to directly quantitate somatic mutations in human genome.[20–22] The current estimate for the somatic (*de novo*) mutation rate is $1-2 \times 10^{-8}$ residues/generation/haploid, and this estimate is sufficiently low that we would expect to never observe somatic mosaicism in the *NLRP3* gene by chance; although the error rate of the high-fidelity DNA polymerase used to produce the amplicons is two orders of magnitude larger than the somatic mutation rate,[23,24] we could not detect PCR-generated mosaicism higher than 1% in the 454 sequencing error maps. Based on the literature, the single base substitutions are the most frequent type of somatic mutations (~500 times more frequent than short insertions/deletions)[25] and protein-coding sequences are less mutagenic than sequences in non-coding regions, assuming that the somatic mutation spectrum in malignant cells is the same as in normal cells. Somatic mosaicism is thought to result from *de novo* gain-of-function-type mutations that are introduced at a very early and limited stage of development, and it is reasonable to focus our efforts on detecting base substitutions for somatic mosaicism in the *NLRP3* gene.

It is challenging but highly important in many areas of research, such as cancer, to detect low-level somatic mutations, which we designated as somatic mosaicism in this study, from apparently mutation-negative samples by conventional sequencing. Subcloning followed by the capillary DNA sequencing has been a *de facto* standard to identify somatic mosaicism, but this is not the method of choice for routine diagnostics because it is laborious, time consuming, and costly. Thus, it is reasonable for us to explore MPS as a new tool for this purpose. Although previous studies have used MPS technology to detect somatic mosaicism, it was unclear how sensitive this method is to detect a low-level somatic mosaicism using the MPS platform because this platform is generally error-prone. To address this challenge, we developed a new pipeline to detect low-level somatic mosaicism with statistical confidence using base position- and strand-specific error rate maps for the *NLRP3* amplicons to be studied. Whereas the detection limit of somatic mosaicism depends on the base position and the read depth of the amplicons, the limit of detection could be as low as 1% allele frequency with no false positives for substitutions (the precision is higher than 99.9%). Our error map shows that 98.1% of base positions (3343 out of 3407 target positions) in the *NLRP3* exonic amplicons can be detected with ~1% mosaicism when more than ~350 reads were accumulated for each strand. Although the remaining region (64 base positions out of 3407 target positions) was too error-prone (the error rate ranged from 0.1 to 1.7% in either the upper or lower strand) to detect low-level mosaicism by MPS, medium-level mosaicism (5% or high) could be identified in all base positions in the target region with the same significance level. Based on this pipeline, we successfully identified four cases of somatic mosaicism among 10 apparently mutation-negative CINCA/NOMID patients. These results were subsequently confirmed by functional analysis and subcloning followed by capillary DNA sequencing method.

As described above, we revealed that a read depth of ~350 for each strand of each amplicon would be sufficient to detect somatic mosaicism as low as 1% with statistical confidence. This means that an analysis of somatic mosaicism (detection limit of 1% allele frequency) of the *NLRP3* gene for one sample requires $350 \times 2 \times 14 = 9800$ reads with the 454 GS-FLX sequencer, which has a capacity to obtain 1 000 000 reads per run. Thus, we could analyse ~100 patient samples with a single run (~10 h) using this MPS platform. For this purpose, a miniaturized 454 sequencer might be more convenient because it could analyse 10 patient samples at once with a reasonably reduced running cost.

The approach used to detect somatic mosaicism is very similar to that for low-frequency alleles in pooled DNA samples, for which MPS applications have been reported by many groups.[18,26,27] However, the main aim of these previous studies was to screen for a rare allele in a population. Thus, the discovery phase on the MPS platform must be followed by an evaluation phase using conventional methods. Therefore, when diagnosing somatic mosaicism of the *NLRP3* gene based solely on the MPS platform, we could not use the same approach to detect rare alleles in a population due to its low accuracy. The sequencing error rate on the Roche MPS platform was sufficiently stable and low enough as shown in this study. Using our pipeline, we were able to detect 1% somatic mosaicism in the *NLRP3* gene with 99.9% confidence. Although another research group recently used a similar approach with a short-read MPS,[28] the Roche long-read MPS is more suitable as a diagnostic tool mainly because of the short run time. If we could diagnose somatic mosaicism of the *NLRP3* gene within a reasonable time with low labour and costs as shown in this study, the success rate of CINCA/NOMID genetic diagnosis will increase from 60 to 80% or higher,[9] which will greatly advance the health and care of these patients and prevent irreversible bone and neurological complications of disease.

This pipeline would also be efficient to detect somatic mosaicism in mutation-negative patients with other diseases, including cancer. The error rate map for a given gene should be constructed from authentic plasmids, and used to detect somatic mosaicism of other genes as well as rare alleles in various populations.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This study was supported by the Japanese Ministry of Education, Science, Sports, and Culture, and the Japanese Ministry of Health, Labor, and Welfare.

Acknowledgements: We thank all patients who participated in the study. We are grateful to Ms. Yuki Takaoka at the Department of Pediatrics, Kyoto University Graduate School of Medicine and Mr. Takashi Watanabe at the Department of Human Genome Research, Kazusa DNA Research Institute for their technical assistance.

References

1. Prieur, A.M. and Griscelli, C. 1981, Arthropathy with rash, chronic meningitis, eye lesions, and mental retardation, *J. Pediatr.*, **99**, 79–83.
2. Hassink, S.G. and Goldsmith, D.P. 1983, Neonatal onset multisystem inflammatory disease, *Arthritis Rheum.*, **26**, 668–73.
3. Torbiak, R.P., Dent, P.B. and Cockshott, W.P. 1989, NOMID–a neonatal syndrome of multisystem inflammation, *Skeletal Radiol.*, **18**, 359–64.
4. Feldmann, J., Prieur, A.M., Quartier, P., et al. 2002, Chronic infantile neurological cutaneous and articular syndrome is caused by mutations in CIAS1, a gene highly expressed in polymorphonuclear cells and chondrocytes, *Am. J. Hum. Genet.* United States, 198–203.
5. Aksentijevich, I., Nowak, M., Mallah, M., et al. 2002, De novo CIAS1 mutations, cytokine activation, and evidence for genetic heterogeneity in patients with neonatal-onset multisystem inflammatory disease (NOMID)—a new member of the expanding family of pyrin-associated autoinflammatory diseases, *Arthritis Rheum.*, **46**, 3340–8.

6. Hoffman, H.M., Mueller, J.L., Broide, D.H., Wanderer, A.A. and Kolodner, R.D. 2001, Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome, *Nat. Genet.*, **29**, 301–5.
7. Goldbach-Mansky, R. 2011, Current status of understanding the pathogenesis and management of patients with NOMID/CINCA, *Curr. Rheumatol. Rep.*, **13**, 123–31.
8. Milhavet, F., Cuisset, L., Hoffman, H.M., et al. 2008, The infevers autoinflammatory mutation online registry: Update with new genes and functions, *Hum. Mutat.*, **29**, 803–8.
9. Tanaka, N., Izawa, K., Saito, M.K., et al. 2011, High incidence of NLRP3 somatic mosaicism in patients with chronic infantile neurologic, cutaneous, articular syndrome: results of an International Multicenter Collaborative Study, *Arthritis Rheum.*, **63**, 3625–32.
10. Qin, W., Kozlowski, P., Taillon, B.E., et al. 2010, Ultra deep sequencing detects a low rate of mosaic mutations in tuberous sclerosis complex, *Hum. Genet.*, **127**, 573–82.
11. Campbell, P.J., Pleasance, E.D., Stephens, P.J., et al. 2008, Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing, *Proc. Natl. Acad. Sci. USA*, **105**, 13081–6.
12. Rohlin, A., Wernersson, J., Engwall, Y., Wiklund, L., Bjoerk, J. and Nordling, M. 2009, Parallel sequencing used in detection of mosaic mutations: Comparison with four diagnostic DNA screening techniques, *Hum. Mutat.*, **30**, 1012–20.
13. Saito, M., Nishikomori, R., Kambe, N., et al. 2008, Disease-associated CIAS1 mutations induce monocyte death, revealing low-level mosaicism in mutation-negative cryopyrin-associated periodic syndrome patients, *Blood*, **111**, 2132–41.
14. Kent, W.J. 2002, BLAT—the BLAST-like alignment tool, *Genome Res.*, **12**, 656–64.
15. Fujisawa, A., Kambe, N., Saito, M., et al. 2007, Disease-associated mutations in CIAS1 induce cathepsin B-dependent rapid cell death of human THP-1 monocytic cells, *Blood*, **109**, 2903–11.
16. Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T. and Martin, J.F. 2011, Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing, *BMC Genomics*, **12**, 245.
17. Margulies, M., Egholm, M., Altman, W.E., et al. 2005, Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, **437**, 376–80.
18. Altmann, A., Weber, P., Quast, C., Rex-Haffner, M., Binder, E.B. and Muller-Myhsok, B. 2011, vipR: variant identification in pooled DNA using R, *Bioinformatics*, **27**, i77–84.
19. Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. 2008, Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, *Nucleic Acids Res.*, **36**, e105.
20. Lee, W., Jiang, Z., Liu, J., et al. 2010, The mutation spectrum revealed by paired genome sequences from a lung cancer patient, *Nature*, **465**, 473–7.
21. Awadalla, P., Gauthier, J., Myers, R.A., et al. 2010, Direct Measure of the de novo mutation rate in Autism and Schizophrenia Cohorts, *Am. J. Hum. Genet.*, **87**, 316–24.
22. Conrad, D.F., Keebler, J.E.M., DePristo, M.A., et al. 2011, Variation in genome-wide mutation rates within and between human families, *Nat. Genet.*, **43**, 712–4.
23. Cha, R.S. and Thilly, W.G. 1993, Specificity, efficiency, and fidelity of PCR, *PCR Methods Appl.*, **3**, S18–29.
24. Vandenbroucke, I., Marck, H.V., Verhasselt, P., et al. 2011, Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications, *BioTechniques*, **51**, 167–77.
25. Pleasance, E.D., Cheetham, R.K., Stephens, P.J., et al. 2010, A comprehensive catalogue of somatic mutations from a human cancer genome, *Nature*, **463**, 191–6.
26. Fakhrai-Rad, H., Zheng, J.B., Willis, T.D., et al. 2004, SNP discovery in pooled samples with mismatch repair detection, *Genome Res.*, **14**, 1404–12.
27. Bansal, V. 2010, A statistical method for the detection of variants from next-generation resequencing of DNA pools, *Bioinformatics*, **26**, i318–24.
28. Flaherty, P., Natsoulis, G., Muralidharan, O., et al. 2012, Ultrasensitive detection of rare mutations using next-generation targeted resequencing, *Nucleic Acids Res.*, **40**, e2.