

ARTICLE

# Comprehensive massive parallel DNA sequencing strategy for the genetic diagnosis of the neuro-cardio-facio-cutaneous syndromes

Ana Justino<sup>1,2</sup>, Patrícia Dias<sup>3</sup>, Maria João Pina<sup>1</sup>, Sónia Sousa<sup>1</sup>, Luís Cirnes<sup>1</sup>, Ana Berta Sousa<sup>3</sup>, José Carlos Machado<sup>1,4</sup> and José Luis Costa<sup>\*,1,4</sup>

Variants in 11 genes of the RAS/MAPK signaling pathway have been causally linked to the neuro-cardio-facio-cutaneous syndromes group (NCFCS). Recently, *A2ML1* and *RIT1* were also associated with these syndromes. Because of the genetic and clinical heterogeneity of NCFCS, it is challenging to define strategies for their molecular diagnosis. The aim of this study was to develop and validate a massive parallel sequencing (MPS)-based strategy for the molecular diagnosis of NCFCS. A multiplex PCR-based strategy for the enrichment of the 13 genes and a variant prioritization pipeline was established. Two sets of genomic DNA samples were studied using the Ion PGM System: (1) training set ( $n = 15$ ) to optimize the strategy and (2) validation set ( $n = 20$ ) to validate and evaluate the power of the new methodology. Sanger sequencing was performed to confirm all variants and low covered regions. All variants identified by Sanger sequencing were detected with our MPS approach. The methodology resulted in an experimental approach with a specificity of 99.0% and a maximum analytical sensitivity of  $\geq 98.2\%$  with a confidence of 99%. Importantly, two patients (out of 20) harbored described disease-causing variants in genes that are not routinely tested (*RIT1* and *SHOC2*). The addition of less frequently altered genes increased in  $\approx 10\%$  the diagnostic yield of the strategy currently used. The presented workflow provides a comprehensive genetic screening strategy for patients with NCFCS in a fast and cost-efficient manner. This approach demonstrates the potential of a combined MPS–Sanger sequencing-based strategy as an effective diagnostic tool for heterogeneous diseases.

*European Journal of Human Genetics* (2015) 23, 347–353; doi:10.1038/ejhg.2014.97; published online 4 June 2014

## INTRODUCTION

The Noonan (NS, MIM 163950), cardio-facio-cutaneous (CFCS, MIM 115150), Costello (CS, MIM 218040) and LEOPARD (LS, MIM 151100) syndromes are members of the neuro-cardio-facio-cutaneous syndromes group (NCFCS).<sup>1</sup> Although the prevalence of these clinically related developmental disorders has not been determined accurately to date, NS is thought to be relatively common.<sup>2</sup> The overlapping clinical features among these disorders and the wide spectrum of phenotypes within each trait, associated with the absence of consensus on routinely used diagnostic criteria, make diagnosis of NCFCS quite challenging.<sup>1,3</sup> The diagnosis of such disorders is particularly difficult to perform *in utero* and in the first years of life, as most features manifest later during childhood.<sup>4,5</sup> Therefore, molecular analysis has become an important key for the verification of clinical diagnosis and represents a highly informative prognostic tool, with direct impact on the establishment of individual follow-up plans.<sup>6,7</sup>

Heterozygous mutations in 11 genes associated with the RAS/MAPK signaling pathway have been causally linked to these disorders: *PTPN11* (MIM 176876), *SOS1* (MIM 182530), *RAF1* (MIM 164760), *BRAF* (MIM 164757), *MAP2K1* (MIM 176872), *MAP2K2* (MIM601263), *KRAS* (MIM 190070), *HRAS* (MIM 190020), *NRAS* (MIM 164790), *SHOC2* (MIM 602775) and *CBL* (MIM 165360).<sup>8–11</sup>

These genes have been documented to account for  $\sim 70$ – $75\%$  of the NCFCS cases.<sup>1</sup> Recently, two additional genes were also associated with these syndromes: *RIT1* (MIM 609591) that encodes a member of the RAS subfamily,<sup>12</sup> and *A2ML1* (MIM 610627) that encodes a complement factor protease inhibitor capable of activating the MAPK cascade.<sup>13</sup> Independent *A2ML1* and *RIT1* variants were identified in individuals with NS without detectable mutations in known Noonan-related genes. *In vivo* studies of the variants in both genes resulted in developmental defects comparable to defects induced by disease-causing variations in other known NS genes.<sup>12,13</sup>

Sanger sequencing is the current gold standard method for the molecular characterization of patients with NCFCS. However, because of the low throughput of this sequencing strategy, the screening of the 13 genes is costly and time consuming. Current molecular testing strategies focus on multigene panels composed of small sets of genes/exons that are selected based on relative frequencies of mutated genes.<sup>14</sup> Moreover, the phenotypic overlap of the syndromes, the high number of genes and lack of mutation hot spots makes it difficult to define straightforward strategies of gene selection for sequential molecular testing.<sup>10,15</sup> In this study, the benchtop Ion PGM Sequencer (Life Technologies, Carlsbad, CA, USA) was used to develop and validate a massive parallel sequencing (MPS)-based approach, by a targeted resequencing strategy, for the molecular

<sup>1</sup>Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal; <sup>2</sup>Abel Salazar Institute for the Biomedical Sciences of the University of Porto, Porto, Portugal; <sup>3</sup>Department of Genetics, Hospital de Santa Maria, Lisboa, Portugal; <sup>4</sup>Faculty of Medicine, University of Porto, Porto, Portugal  
\*Correspondence: Dr JL Costa, Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Rua Dr Roberto Frias s/n, 4200-465 Porto, Portugal.  
Tel: +351 22 5570 738; Fax: +351 22 5570 799; E-mail: jcosta@ipatimup.pt

Received 14 February 2014; revised 3 April 2014; accepted 8 April 2014; published online 4 June 2014

diagnosis of patients with NCFCS. The development and validation of the MPS-based approach was performed according to the checklist published by the College of American Pathologists,<sup>16</sup> and the guidelines developed by the Next-Generation Sequencing: Standardization of Clinical Testing (Nex-StoCT) workgroup<sup>17</sup> and by the American College of Medical Genetics and Genomics Working Group on Next-Generation Sequencing.<sup>18</sup>

## MATERIALS AND METHODS

### DNA samples selection

Constitutional genomic DNA samples from a total of 35 patients with diagnosis of NCFCS were collected from our archives and used in this study. All samples had been previously screened by Sanger Sequencing for a selected set of exons associated with the clinically suspected syndrome (Supplementary Figure 1). The samples were divided into two groups: (1) training set, composed of 15 selected samples (12 cases of NS, 1 case of CFCS, 1 case of CS and 1 case of LS), of which 10 harbored known disease-causing variants, used to optimize the strategy; and (2) validation set, composed of 20 random samples (19 cases of NS, including 3 cases of prenatal diagnosis, and 1 case of CFCS) studied blindly and used to validate and evaluate the power of the new methodology. In addition, 100 DNA samples from unaffected control subjects were studied to assess the occurrence and frequency of novel and rare variants (minor allele frequency <2%) of unknown clinical significance. The study was performed in accordance with Article 19 ('DNA Banks and other biological products') of Portuguese Law No. 12/2005 of 26 January ('Personal genetic information and health information').

### Targeted enrichment of genomic DNA

For the gene targeted enrichment, a multiplex PCR-based strategy was designed. The sequences of the 13 genes associated with the NCFCS were retrieved from the NCBI Reference Sequence Database ([www.ncbi.nlm.nih.gov/refseq](http://www.ncbi.nlm.nih.gov/refseq); Table 1). Primer pairs were designed to amplify all coding exons, including ~50 bp of noncoding regions flanking the 5' and 3' ends of each exon, using Primer3 web-based tool.<sup>19</sup> Multiplex PCR reactions were performed following the *QIAGEN Multiplex PCR Handbook 10/2010* (QIAGEN, Hilden, Germany) specifications. Each reaction consisted of 2  $\mu$ l of 10  $\times$  Primer mix, 12.5  $\mu$ l of multiplex master mix, 2  $\mu$ l of Q-Solution and 50–100 ng of genomic DNA in a total reaction volume of 25  $\mu$ l. The concentration of each primer in the 10  $\times$  Primer mix was adjusted in order to make uniform the representation of the different amplicons in each multiplex assay (primers concentration varied between 1 and 3  $\mu$ M). Oligonucleotide primer sequences and PCR amplification conditions are detailed in Supplementary Table 1.

**Table 1** NCFCS-associated genes studied

Gene	Locus	Genomic reference	Transcript reference	Coding exons	Bases
					sequenced (bp) <sup>a</sup>
<i>A2ML1</i>	12p13.31	NC_000012.11	NM_144670.4	35	11 626
<i>BRAF</i>	7q34	NG_007873.2	NM_004333.4	18	6870
<i>CBL</i>	11q23.3	NG_016808.1	NM_005188.3	16	6165
<i>HRAS</i>	11p15.5	NG_007666.1	NM_005343.2	4	1358
<i>KRAS</i>	12p12.1	NG_007524.1	NM_004985.3	4	2058
<i>MAP2K1</i>	15q22	NG_008305.1	NM_002755.3	11	4169
<i>MAP2K2</i>	19p13.3	NG_007996.1	NM_030662.3	11	4333
<i>NRAS</i>	1p13.2	NG_007572.1	NM_002524.4	4	2010
<i>PTPN11</i>	12q24.1	NG_007459.1	NM_002834.3	15	5557
<i>RAF1</i>	3p25.2	NG_007467.1	NM_002880.3	16	6472
<i>RIT1</i>	1q21.2	NC_000001.10	NM_006912.5	5	1691
<i>SHOC2</i>	10q25.2	NG_028922.1	NM_007373.3	8	3861
<i>SOS1</i>	2p22.1	NG_007530.1	NM_005633.3	23	9537

<sup>a</sup>Total number of unique bases sequenced for a given gene.

### Ion PGM system library preparation and sequencing

MPS was carried out on the Ion PGM Sequencer according to the manufacturer's instructions. Library preparation was performed according to the protocol Ion Xpress Plus gDNA and Amplicon Library Preparation PN 4471989 Rev. D (Life Technologies). Briefly, for each sample, amplicons generated by the developed PCR multiplex-based strategy were pooled and purified using the AMPure XP reagent (Beckman Coulter, Carnaxide, Portugal). Then, 100 ng of pooled amplicons were sheared by enzymatic digestion for 7 min at 37 °C with the Ion Xpress Plus Fragment Library Kit PN4471269 (Life Technologies). The digested products were purified and adapters or barcode adapters ligated. E-Gel SizeSelect 2% Agarose system (Life Technologies) was used to carry out fragment size selection. The size-selected fragments were nick-translated and PCR-amplified for a total of eight cycles. Following purification, library molarity was determined, and barcoded libraries pooled in equimolar concentrations and used for clonal amplification by emulsion PCR. Template preparation was performed according to the protocol Ion OneTouch System PN4472430 Rev. E (Life Technologies). Quality control of the ISPs was done with the Ion Sphere Quality Control Kit (Life Technologies). The complete ISP sample (for 316 Chip) or half the ISP sample (for 314 Chip) was prepared for sequencing using the protocol Ion PGM 200 Sequencing Kit PN4469714 Rev. C (Life Technologies), loaded on the chip and sequenced on the Ion PGM Sequencer for 130 cycles, using 520-flow runs.

### Bioinformatic analysis

Data from Ion PGM System sequencing runs were automatically transferred to the dedicated Torrent Server and processed using platform-specific pipeline software Torrent Suite v3 (Life Technologies) to generate sequencing reads and base-call quality scores, filter low-quality reads, trim adapter sequences and split the reads according to the barcode. Processed reads were stored in the FASTQ file format. The sequence files were aligned to reference sequences previously annotated for the covered amplicons, using SeqMan NGen v4 (DNASTar, Madison, WI, USA). Coverage analysis, variant discovery and base-calling statistics were performed using SeqMan Pro v10 (DNASTar).

A dedicated variant prioritization pipeline (VPP) was developed for data analysis in order to detect all the true positive calls and minimize the number of false positive results.<sup>20</sup> The VPP was fine-tuned using all known variants present in the training set as surrogate markers. Genetic alterations observed in exons that were not included in our original strategy (Supplementary Figure 1) were validated by Sanger sequencing. Different filter thresholds related to sequencing calling required for variant identification were weighted: percentage of the single most prevalent nonreference base (SNP%  $\geq$  25), phred-like quality score of the called genotype (Qcall  $\geq$  40) and depth of coverage (Depth  $\geq$  30). An additional interpretative exclusion setting related to the Ion PGM System sequencing error profile was used to minimize the number of false positive calls. Ion PGM System is prone to falsely call single-base indels mostly because of nonlinear pH variation relative to homopolymer length during the DNA polymerization.<sup>21</sup> Such sequencing miscalls are nonrandomly distributed, systematic (appear in more than one case and in different runs) and have a strand-specific bias pattern (occur preferentially on one DNA strand compared with the other).<sup>20</sup> Calls that fulfill these characteristics were excluded from the analysis. The established VPP was then applied to the validation set. Sanger sequencing was performed in genes/exons that were not included in the previously used multigene panel to confirm all variants in replicate PCR amplicons and to fill-in regions with insufficient coverage. The framework for the validation of clinical molecular genetic tests set by the EuroGentest Validation Group was used to calculate the performance of the new approach.<sup>22</sup> For these calculations, the regions sequenced by both the gold standard Sanger and Ion PGM Sequencer were used. These regions correspond to the exons detailed in Supplementary Figure 1 and the amplicons covering variants confirmed by Sanger sequencing (all true positive and false positive calls Ion PGM System calls).

### Sanger sequencing

The PCR reactions were enzymatically purified using FastAP Thermosensitive Alkaline Phosphatase and Exonuclease I (Fermentas, Pittsburgh, PA, USA), sequenced with BigDye Terminator v3.1 Cycle Sequencing Kit

(Applied Biosystems, Carlsbad, CA, USA) and run on the ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems) according to the instructions of the manufacturers. Variant analysis was carried out with the software Mutation Surveyor 3.24 (Softgenetics, State College, PA, USA).

### Sequence variant interpretation

Validated variants were compared with mutations reported in literature as causative of NCFCS and with the data from National Center for Biotechnology Information ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)), 1000 Genomes project ([www.1000genomes.org](http://www.1000genomes.org)), Human Gene Mutation Database ([www.hgmd.cf.ac.uk](http://www.hgmd.cf.ac.uk)) and NHLBL Exome Sequencing Project ([evs.gs.washington.edu/EVS/](http://evs.gs.washington.edu/EVS/)). In addition, for novel and rare variants (minor allele frequency <2%) of unknown clinical significance, a set of 200 control alleles (100 unaffected subjects) were sequenced by MPS on Ion PGM Sequencer according to the method described above. The functional effects of novel and rare variants that were absent from the control population were predicted using different *in silico*-based methods: PolyPhen-2,<sup>23</sup> MutationTaster,<sup>24</sup> SNPs&Go<sup>25</sup> and Human Splicing Finder.<sup>26</sup> These variants were submitted to a public database ([www.ncbi.nlm.nih.gov/SNP/](http://www.ncbi.nlm.nih.gov/SNP/)).

### Sequence variant nomenclature

The sequence variant nomenclature used in this work is in accordance with the Human Genome Variation Society's guidelines.<sup>27</sup> DNA alteration numbering is based on the cDNA sequence with position +1 corresponding to the A of the ATG translation initiation codon in the reference sequence. Sequence variant descriptions on the protein level consider the initiator methionine as codon 1 and have been checked using the Mutalyzer program.<sup>28</sup>

## RESULTS

### Multiplex PCR strategy and sequencing

For the enrichment of all coding regions of the 13 genes a multiplex PCR-based strategy was designed (Table 1). A total of 65.7 kb of genomic DNA was amplified using 158 primer pairs distributed in 12 multiplex and 8 simplex PCR reactions. Amplicon sizes ranged from 217 to 943 bp, with an average length of 413 bp (Supplementary Table 1). Taking into consideration the 65.72 kb of genomic sequence for each sample and the minimum desired coverage per base (30 reads), it was estimated a maximum of 1 sample per 314 chip and up to 6 samples to be barcoded and pooled per 316 chip. The Ion PGM System sequencing output was dependent on the chip-loading efficiencies. For the 314 chip, an average loading density of 61% was obtained that resulted in >350 000 reads. For the 316 chip, an average loading density of 70% was obtained that resulted in >2 000 000 reads. The average read length was 184 bases.

### Training set

The training set was primarily used to develop and optimize the analytical wet bench process and the subsequent data analysis pipeline. This set was composed of 15 samples from our archives with known synonymous polymorphisms ( $n=14$ ) and different disease-causing alterations ( $n=10$ ) that were used as positive controls. The coverage of all amplicons was ascertained before the development of the bioinformatic pipeline. To avoid the introduction of bias in calculations, the noncovered amplicons were annotated and excluded from the analysis. The post-assembly Ion PGM System data analysis was performed without the application of variant calling filters. Within the list of 2133 calls produced by the Ion PGM System, all the known 24 variants (14 polymorphisms + 10 disease-causing variants) were correctly identified. Furthermore, a total of 164 genetic alterations in the genes/exons that were not included in the previously used multigene panel were also identified. To validate these new variants, Sanger sequencing was carried out in replicate PCR amplicons.

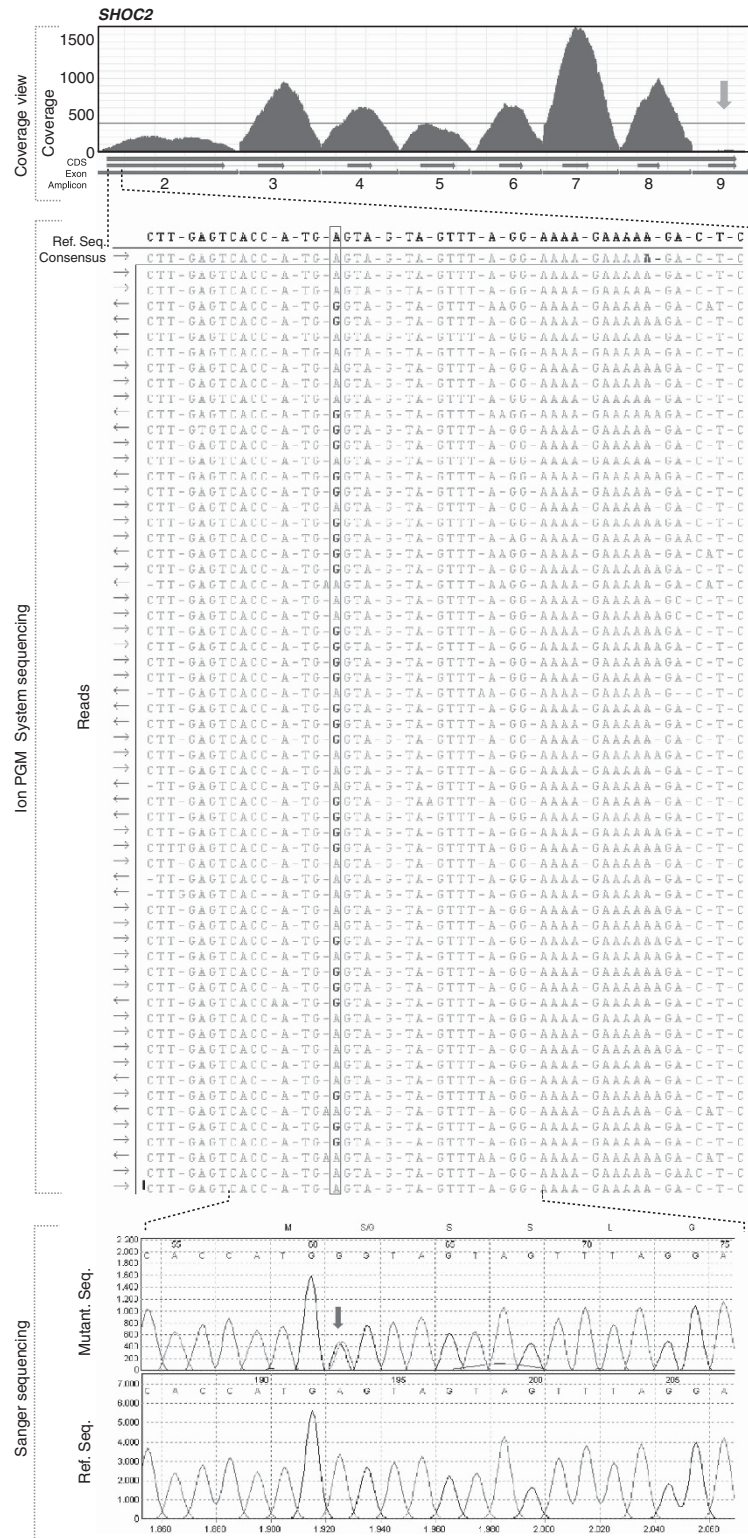
The validation of these genetic alterations resulted in 188 true variants (24 previously known variants + 164 new variants) that were used as surrogate markers to develop the custom bioinformatic parameters for variant prioritization (Supplementary Table 2). Different parameters related to sequencing calling were set in order to determine the thresholds required to filter-in the real variants. To exclude the possibility of false negative results and to minimize the number of false positives, the filter thresholds were set at SNP%  $\geq 25$ , Q call  $\geq 40$  and Depth  $\geq 30$ . To further attenuate the false positive rate, known limitations of the Ion PGM System sequencing chemistry were used as an additional interpretative exclusion criterion. The application of this VPP to our training set resulted in a list of 211 variants, containing all 188 known variants and 23 additional false positives.

### Validation set

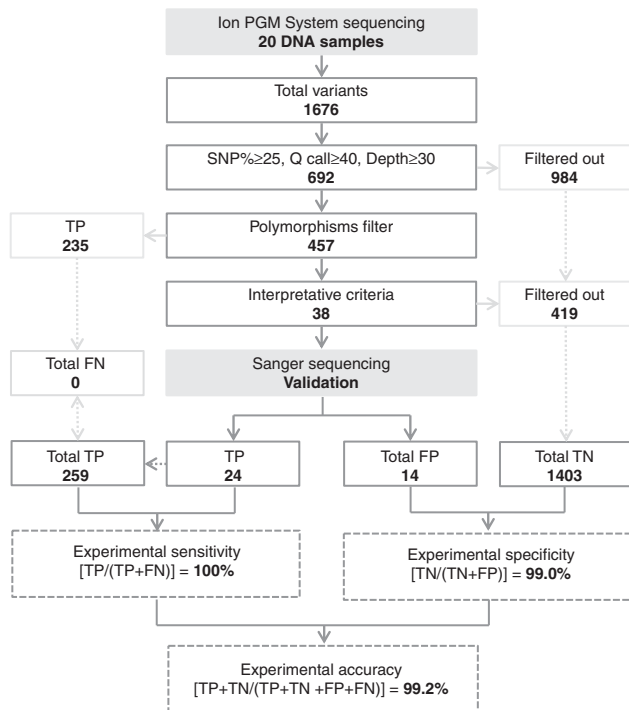
The validation set, composed of 20 random samples, was used to validate and independently evaluate the performance of the workflow. Upon Ion PGM System sequencing, the coverage of all amplicons was confirmed in order to exclude the possibility of false negative results. The rare amplicons that did not achieve the established 30 reads of coverage were annotated and re-sequenced by Sanger method (average of 3 amplicons per sample; Figure 1). The Ion PGM System sequencing data were then analyzed in a blind manner through the application of the developed VPP to the 2855 generated calls. Briefly, the sequencing calling analysis step reduced the number of putative variants to 1106. All calls that were known polymorphisms (235) were identified and confirmed by Sanger sequencing. The exclusion of these polymorphisms, followed by the application of the interpretative exclusion setting, resulted in the identification of 38 variants. Validation of these variants through Sanger sequencing resulted in the detection of 24 single-base substitutions. Validation of one of these variants is comprehensively illustrated in Figure 1. The remaining 14 discordant variants were falsely called indels identified by the Ion PGM System, 9 of them in homopolymer regions longer than 4 nucleotides (Supplementary Table 3). To calculate the sensitivity and specificity of the assay, the regions sequenced by both technologies were considered. This resulted in a total of 1676 Ion PGM System calls. The application of the developed VPP to these data correctly filtered out 1403 calls (true negatives), reducing the number of putative variants from 1676 to 273. No true variant went undetected and, of the 273 identified variants, 259 were confirmed by Sanger sequencing and 14 were false positive calls. According to these results, the sensitivity and specificity of this assay were 100% and 99.0%, respectively. The experimental outcome of the validation set is comprehensively illustrated in Figure 2.

### Genetic variants identified in the validation set

Concerning the 259 Sanger-confirmed variants, 236 were in coding regions whereas 23 were in intronic regions (Supplementary Table 4). Only the intronic change c.3347-1G>A in *SOS1* had not been reported previously. Of the 236 variants detected in coding regions, 126 were synonymous (63 heterozygous and 63 homozygous), 106 nonsynonymous (46 heterozygous and 60 homozygous) and 4 were heterozygous deletion c.3676\_3677delGC (p.(Ala1226Glnfs\*34)) within the coding region of *A2ML1*. Of the 106 missense variants, 83 were common polymorphisms located in *A2ML1* gene. Within the remaining 23 missense alterations, 13 were variants that had already been established as causative of NCFCS. The remaining 10 missense variants were novel and rare variants of unknown clinical significance located in the *A2ML1*, *CBL* and *SOS1* genes.



**Figure 1** Illustrative example of the detection and validation of a variant within the *SHOC2* gene. The coverage view (top panel) depicts the depth of coverage profile for the coding region of the *SHOC2* gene. The orange arrow points to an amplicon (exon 9) with insufficient coverage (<30 reads) that was studied through Sanger Sequencing. The beginning of exon 2 is expanded to show the underlying DNA sequence. The Ion PGM System sequencing (middle panel) shows the sequence alignment of Ion PGM System reads against the reference sequence of exon 2 of the *SHOC2* gene. The green and red arrows indicate the forward and reverse reads, respectively (for readability, not all aligned reads are shown). The red box highlights the detected heterozygous missense variant c.4A>G. Sanger sequencing (bottom panel) confirms the genetic alteration. The red arrow indicates the variant position in the electropherogram. The full colour version of this figure is available at [European Journal of Human Genetics](http://www.ejhg.com) online.



**Figure 2** Variant prioritization pipeline applied to the validation set and experimental outcome of the assay. The analysis of the genomic regions covered by both Sanger and Ion PGM System sequencing resulted in a total of 1676 Ion PGM System calls. Sequencing quality filters (SNP%  $\geq$  25, Q call  $\geq$  40 and Depth  $\geq$  30) filtered out 984 calls, reducing the number of putative variants from 1676 to 692. Of these 692 calls, 235 were known polymorphisms that were confirmed by Sanger sequencing. Filtering out the 235 validated calls reduced the number of putative variants to 457. Interpretative exclusion criteria filtered out 419 calls. A total of 38 putative variants were directly obtained from the VPP output. Of these 38 variants, 24 were validated through Sanger sequencing as real and 14 were falsely called indels. Experimental outcome of this validation assay resulted in 259 Sanger-confirmed calls or true positives (TP); 14 miscalls or false positives (FP); 1403 correctly filtered out calls or true negatives (TN); 0 incorrectly filtered out calls or false negatives (FN).

A known disease-causing variant for NCFCS was detected in 13 patients (12 NS and 1 CFCS). In two of these cases (patients 12 and 13), one rare variant of unknown significance in *A2ML1* gene was found to co-occur (Table 2). For the remaining seven patients with no described NCFCS genetic defects, five had new or rare variants of uncertain disease causality. In two of these patients (patients 14 and 18), genetic alterations of uncertain pathogenicity were found to co-occur (Table 2). All novel and rare variants of unknown clinical significance were absent from the 100 control samples. The experimental outcome of *in silico* prediction tests for these variants is presented in Supplementary Table 5.

## DISCUSSION

The aberrant activation of the RAS/MAPK cascade underlies a number of clinically and genetically related developmental disorders that are grouped within the NCFCS family. Four of these disorders – NS, CFCS, CS and LS – share numerous similarities of phenotypic features, making the differential diagnosis challenging, particularly prenatally and in newborns.<sup>1,5</sup> Until now, heterozygous variants in 13 genes of this pathway (*PTPN11*, *SOS1*, *RAF1*, *BRAF*, *KRAS*, *NRAS*, *HRAS*, *MAP2K1*, *MAP2K2*, *SHOC2*, *CBL*, *A2ML1* and *RIT1*) have been documented to be causally associated with NCFCS.<sup>3,11–13</sup>

Currently, because of genetic heterogeneity, only the most prevalent disease genes are assessed in routine molecular diagnosis of the NCFCS. Moreover, the low throughput of current gene-diagnostic tools makes it both challenging and time consuming. Although MPS has the potential to offer a solution for this problem, its implementation in a diagnostic setting is complex as it requires several steps of validation.<sup>20</sup>

This study presents the validation of a MPS-based approach for the molecular diagnosis of patients with NCFCS and a strategy for its implementation in a clinical laboratory environment. For that, a multiplex PCR-based strategy for the enrichment of the entire coding region of the 13 NCFCS disease genes was designed. A PCR-based strategy was chosen to directly adapt some of the primers and conditions previously used in conventional Sanger sequencing. Although multiplex PCR has been shown to be a robust enrichment technique, it becomes less practical as the target size increases. Capture arrays and the recently available commercial alternatives will facilitate scaling up to larger targets.

The training set was used to optimize the PCR conditions in order to ensure that all exons were sequenced and adjust the representation of the different amplicons in the library pool. Despite efforts for equimolar amplicon pooling, the read-coverage patterns were not uniform along the targeted region. The issue of coverage variability seems to be linked to unequal enrichment efficiencies, library preparation aspects, particularities of local sequence context, alignment issues<sup>29</sup> and, in the specific case of Ion PGM System, chip-loading efficiencies. Variability in coverage requires samples to be oversequenced to achieve the minimum desired coverage per base required for variant identification (30 reads).<sup>20</sup> The adjustments made in the PCR and sample pooling conditions largely improve amplification uniformity and reduce the number of amplicons below the coverage threshold. Nevertheless, because of unpredictability of the missing amplicons, we observed that the initial step of amplicon coverage analysis was of utmost importance to avoid false negative results. In the rare cases where amplicons could not be properly analyzed, mostly because of PCR skipping or high GC content, they were Sanger sequenced in parallel.

Although MPS provides several advantages over the Sanger method, the amount of data generated by these sequencing techniques poses informatic challenges for data analysis.<sup>30</sup> Currently, there is no standard data analysis pipeline and different criteria for variant identification have been used in MPS studies.<sup>29,31</sup> In this study, known true positive variants of the training set were used as surrogate markers to customize the data analysis process and fine-tune the VPP. The basic principle of the VPP was to filter out random sequence errors without the introduction of false negatives. The validation set was then used to assess the power of the new methodology and report its diagnostic accuracy.<sup>22</sup> Thus, at a 99% confidence, the maximum analytical sensitivity of the study is  $\geq$  98.2% and this means that the probability of a future false negative result is 1.8%. The level of specificity obtained (99.0%) is higher than the levels reported in previous studies concerning the Ion Torrent technology.<sup>21,31</sup> As expected, the false positive calls were mostly single-base indels located in short homopolymer regions with a strand-specific bias. Furthermore, a high number of these errors were observed within motifs of two nucleotide homopolymers, principally CC and GG. As the sequencing chemistry and bioinformatic algorithms continually improve, a reduction in the Ion PGM System error rate in the near future is expected.

In this study, the molecular characterization of the validation set samples revealed 13 known disease-causing alterations and 11 novel and rare variants of unknown significance (10 missense and one in a canonical splice site). Six of these variants were found to co-occur

**Table 2 Genetic profile of the 20 validation set samples**

Patient (syndrome)	Gene	Variant		Classification	MAF <sup>a</sup>
		Nucleotide	Amino acid		
1 (CFCS)	<i>BRAF</i>	c.1502A>G	p.(Glu501Gly)	Disease causing <sup>6</sup>	—
2 (NS)	<i>PTPN11</i>	c.218C>T	p.(Thr73Ile)	Disease causing <sup>35</sup>	—
3 (NS)	<i>PTPN11</i>	c.854T>C	p.(Phe285Ser)	Disease causing <sup>35</sup>	—
4 (NS)	<i>PTPN11</i>	c.922A>G	p.(Asn308Asp)	Disease causing <sup>36</sup>	—
5 (NS)	<i>PTPN11</i>	c.922A>G	p.(Asn308Asp)	Disease causing <sup>36</sup>	—
6 (NS)	<i>PTPN11</i>	c.923A>G	p.(Asn308Ser)	Disease causing <sup>35</sup>	—
7 (NS)	<i>PTPN11</i>	c.1403C>T	p.(Thr468Met)	Disease causing <sup>37</sup>	—
8 (NS)	<i>PTPN11</i>	c.1403C>T	p.(Thr468Met)	Disease causing <sup>37</sup>	—
9 (NS)	<i>RAF1</i>	c.770C>T	p.(Ser257Leu)	Disease causing <sup>38</sup>	—
10 (NS)	<i>RAF1</i>	c.782C>T	p.(Pro261Leu)	Disease causing <sup>38</sup>	—
11 (NS)	<i>SHOC2</i>	c.4A>G	p.(Ser2Gly)	Disease causing <sup>39</sup>	—
12 (NS)	<i>RIT1</i>	c.270G>A	p.(Met90Ile)	Disease causing <sup>12</sup>	—
13 (NS)	<i>A2ML1</i>	c.289C>G	p.(Arg97Gly)	VUS	0.001
	<i>PTPN11</i>	c.785T>G	p.(Leu262Arg)	Disease causing <sup>40</sup>	—
14 (NS)	<i>A2ML1</i>	c.861C>A	p.(Asp287Glu)	VUS	0.003
	<i>A2ML1</i>	c.619G>C	p.(Gly207Arg)	VUS	0.020
	<i>A2ML1</i>	c.1109T>C	p.(Phe370Ser)	VUS	0.019
	<i>CBL</i>	c.2345C>T	p.(Pro782Leu)	VUS	0.000
	<i>SOS1</i>	c.3347-1G>A	—	VUS	0.000
15 (NS)	<i>A2ML1</i>	c.1918G>A	p.(Asp640Asn)	VUS <sup>b</sup>	0.000
16 (NS)	<i>A2ML1</i>	c.3287C>T	p.(Ser1096Phe)	VUS <sup>b</sup>	0.000
17 (NS)	<i>CBL</i>	c.1463C>T	p.(Ala488Val)	VUS	0.012
18 (NS)	<i>CBL</i>	c.1711G>A	p.(Asp571Asn)	VUS <sup>b</sup>	0.000
	<i>SOS1</i>	c.2138G>A	p.(Arg713Gln)	VUS <sup>b</sup>	0.000
19 (NS)	—	—	—	—	—
20 (NS)	—	—	—	—	—

Abbreviations: MAF, minor allele frequency; VUS, variant of unknown significance.

<sup>a</sup>MAF from 1000 Genomes project.

<sup>b</sup>No rs-identifier is known for this variant, nor has it been observed in 1000 Genomes project and/or ESP.

with alterations described as causative of NCFCS or with other uncommon variants. Co-occurring variants in different RAS/MAPK pathway-related genes have already been described in NCFCS patients.<sup>32,33</sup> However, it is not known whether these co-occurring variants lead to more severe or atypical phenotypes. In fact, to predict the disease causality of the novel variants, either alone or in co-occurrence, remains challenging. For a deeper understanding of the effects of novel sequence variants, it would be necessary to assess segregation patterns of the variants, analyze mRNA transcripts and perform functional assays in genetically manipulated cell cultures and animal models. Bioinformatic prediction tools can also be used, but should be interpreted with caution.<sup>34</sup>

The presented workflow provides a comprehensive genetic screening tool for patients suspected of NCFCS, allowing the detection of sequence variants in previously reported genes regardless of clinical diagnosis. Importantly, the analysis of the 13 genes allowed the identification of 13 NCFCS patients with a previously described disease-causing variant out of the 20 cases of the validation set. If the molecular characterization of these cases had been restricted to the previously used multigene panel (Supplementary Figure 1), only 11 of the cases would have been reported as positive, as the panel did not encompass the genetic screening of *RIT1* and *SHOC2* genes. As a result, by screening less frequently altered genes, the new methodology increased in ~10% the diagnostic yield of the strategy currently used for the molecular diagnosis of NCFCS. Additionally, with the new methodology the costs and turn-around-time were reduced. The time reduction was of particular importance for prenatal genetic diagnosis.<sup>4</sup> For instance, the results of the screening of the 20

validation set cases were obtained within 5 days: 2 days for target enrichment and library preparation, 1 day for template preparation and sequencing and 2 days for data analysis and interpretation. In contrast, to screen the same 20 cases using Sanger sequencing, 90 days would have been needed: 2 days for target enrichment and enzymatic purification, 60 days to perform 6320 sequencing reactions and electrophoretic runs (158 amplicons per case in forward and reverse sequencing directions) and 30 days for data analysis and interpretation.

In the current validation study, we designed a combined targeted MPS–Sanger sequencing-based strategy for the molecular diagnosis of NCFCS using the Ion PGM System. This methodology is faster, more comprehensive and more cost effective than the routinely used Sanger sequencing multigene panel-based approach. The study provides practical evidences that a combined targeted MPS–Sanger sequencing-based strategy is suitable for daily clinical practice of genetic testing and can be used as an effective molecular diagnostic tool for multigenic diseases and for prenatal genetic testing.

#### CONFLICT OF INTEREST

The authors declare no conflict of interest.

#### ACKNOWLEDGEMENTS

This work was funded by Research Grant from Fundação para a Ciência e Tecnologia (ERA-PTG/0001/2010 to AJ; PTDC/BIM-MEC/0650/2012 to JLC) and PPS5 Consórcio DoIT (ADI–Agência de Inovação to JLC). IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Education and Science and is partially supported by FCT, the Portuguese Foundation for Science and Technology.

- 1 Tartaglia M, Gelb BD: Disorders of dysregulated signal traffic through the RAS-MAPK pathway: phenotypic spectrum and molecular mechanisms. *Ann N Y Acad Sci* 2010; **1214**: 99–121.
- 2 Nora JJ, Nora AH, Sinha AK, Spangler RD, Lubs HA: The Ullrich-Noonan syndrome (Turner phenotype). *Am J Dis Child* 1974; **127**: 48–55.
- 3 Tidyman WE, Rauen KA: Noonan, Costello and cardio-facio-cutaneous syndromes: dysregulation of the Ras-MAPK pathway. *Expert Rev Mol Med* 2008; **10**: e37.
- 4 Baldassarre G, Mussa A, Dotta A *et al*: Prenatal features of Noonan syndrome: prevalence and prognostic value. *Prenat Diagn* 2011; **31**: 949–954.
- 5 Digilio MC, Lepri F, Baban A *et al*: RASopathies: clinical diagnosis in the first year of life. *Mol Syndromol* 2011; **1**: 282–289.
- 6 Rodriguez-Viciana P, Tetsu O, Tidyman WE *et al*: Germline mutations in genes within the MAPK pathway cause cardio-facio-cutaneous syndrome. *Science* 2006; **311**: 1287–1290.
- 7 Romano AA, Allanson JE, Dahlgren J *et al*: Noonan syndrome: clinical features, diagnosis, and management guidelines. *Pediatrics* 2010; **126**: 746–759.
- 8 Gripp KW, Lin AE: Costello syndrome: a Ras/mitogen activated protein kinase pathway syndrome (rasopathy) resulting from HRAS germline mutations. *Genet Med* 2012; **14**: 285–292.
- 9 Martínez-Quintana E, Rodríguez-González F: LEOPARD syndrome: clinical features and gene mutations. *Mol Syndromol* 2012; **3**: 145–157.
- 10 Nava C, Hanna N, Michot C *et al*: Cardio-facio-cutaneous and Noonan syndromes due to mutations in the RAS/MAPK signalling pathway: genotype-phenotype relationships and overlap with Costello syndrome. *J Med Genet* 2007; **44**: 763–771.
- 11 Roberts AE, Allanson JE, Tartaglia M, Gelb BD: Noonan syndrome. *Lancet* 2013; **381**: 333–342.
- 12 Aoki Y, Niihori T, Banjo T *et al*: Gain-of-function mutations in RIT1 cause Noonan syndrome, a RAS/MAPK pathway syndrome. *Am J Hum Genet* 2013; **93**: 173–180.
- 13 Yntema H, Nillesen W, Overman J *et al*: The identification of a novel gene identified by exome sequencing reveals the upstream components of the RAS/MAPK signaling pathway involved in Noonan syndrome (Abstract/Program #215). Presented at 62nd Annual Meeting of the American Society of Human Genetics, 6–10 November 2012 in San Francisco, California.
- 14 Schulz AL, Albrecht B, Arici C *et al*: Mutation and phenotypic spectrum in patients with cardio-facio-cutaneous and Costello syndrome. *Clin Genet* 2008; **73**: 62–70.
- 15 Nyström AM, Ekvall S, Berglund E *et al*: Noonan and cardio-facio-cutaneous syndromes: two clinically and genetically overlapping disorders. *J Med Genet* 2008; **45**: 500–506.
- 16 Program CLA: Molecular Pathology Checklist. [http://www.cap.org/apps/docs/laboratory\\_accreditation/checklists/new/molecular\\_pathology\\_checklist.pdf](http://www.cap.org/apps/docs/laboratory_accreditation/checklists/new/molecular_pathology_checklist.pdf)
- 17 Gargis AS, Kalman L, Berry MW *et al*: Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol* 2012; **30**: 1033–1036.
- 18 Rehm HL, Bale SJ, Bayrak-Toydemir P *et al*: ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* 2013; **15**: 733–747.
- 19 Rozen S, Skaletsky H: Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 2000; **132**: 365–386.
- 20 Costa JL, Sousa S, Justino A *et al*: Nonoptical massive parallel DNA sequencing of BRCA1 and BRCA2 genes in a diagnostic setting. *Hum Mutat* 2013; **34**: 629–635.
- 21 Quail MA, Smith M, Coupland P *et al*: A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012; **13**: 341.
- 22 Mattocks CJ, Morris MA, Matthijs G *et al*: A standardized framework for the validation and verification of clinical molecular genetic tests. *Eur J Hum Genet* 2010; **18**: 1276–1288.
- 23 Adzhubei IA, Schmidt S, Peshkin L *et al*: A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
- 24 Schwarz JM, Rödelsperger C, Schuelke M, Seelow D: MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; **7**: 575–576.
- 25 Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R: Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* 2009; **30**: 1237–1244.
- 26 Desmet FO, Hamroun D, Lalande M, Collod-Bérout G, Claustres M, Bérout C: Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 2009; **37**: e67.
- 27 den Dunnen JT, Antonarakis SE: Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 2000; **15**: 7–12.
- 28 Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE: Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat* 2008; **29**: 6–13.
- 29 Dames S, Durtschi J, Geiersbach K, Stephens J, Voelkerding KV: Comparison of the Illumina Genome Analyzer and Roche 454 GS FLX for resequencing of hypertrophic cardiomyopathy-associated genes. *J Biomol Tech* 2010; **21**: 73–80.
- 30 Jia P, Li F, Xia J *et al*: Consensus rules in variant detection from next-generation sequencing data. *PLoS One* 2012; **7**: e38470.
- 31 Chan M, Ji SM, Yeo ZX *et al*: Development of a next-generation sequencing method for BRCA mutation screening: a comparison between a high-throughput and a benchtop platform. *J Mol Diagn* 2012; **14**: 602–612.
- 32 Brasil AS, Malaquias AC, Wanderley LT *et al*: Co-occurring PTPN11 and SOS1 gene mutations in Noonan syndrome: does this predict a more severe phenotype? *Arq Bras Endocrinol Metabol* 2010; **54**: 717–722.
- 33 Ekvall S, Hagenas L, Allanson J, Anneren G, Bondeson ML: Co-occurring SHOC2 and PTPN11 mutations in a patient with severe/complex Noonan syndrome-like phenotype. *Am J Med Genet A* 2011; **155A**: 1217–1224.
- 34 Richards CS, Bale S, Bellissimo DB *et al*: ACMG recommendations for standards for interpretation and reporting of sequence variations: revisions 2007. *Genet Med* 2008; **10**: 294–300.
- 35 Tartaglia M, Kalidas K, Shaw A *et al*: PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity. *Am J Hum Genet* 2002; **70**: 1555–1563.
- 36 Ferrario VF, Sforza C, Ciusa V, Dellavia C, Tartaglia GM: The effect of sex and age on facial asymmetry in healthy subjects: a cross-sectional study from adolescence to mid-adulthood. *J Oral Maxillofac Surg* 2001; **59**: 382–388.
- 37 Digilio MC, Conti E, Sarkozy A *et al*: Grouping of multiple-lentiginos/LEOPARD and Noonan syndromes on the PTPN11 gene. *Am J Hum Genet* 2002; **71**: 389–394.
- 38 Pandit B, Sarkozy A, Pennacchio LA *et al*: Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nat Genet* 2007; **39**: 1007–1012.
- 39 Cordeddu V, Di Schiavi E, Pennacchio LA *et al*: Mutation of SHOC2 promotes aberrant protein N-myristoylation and causes Noonan-like syndrome with loose anagen hair. *Nat Genet* 2009; **41**: 1022–1026.
- 40 Ferrero GB, Picco G, Baldassarre G *et al*: Transcriptional hallmarks of Noonan syndrome and Noonan-like syndrome with loose anagen hair. *Hum Mutat* 2012; **33**: 703–709.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (<http://www.nature.com/ejhg>)