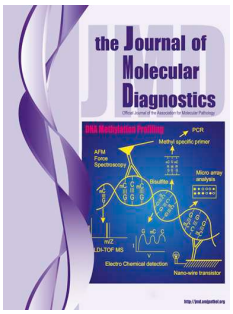
Article:
Watson, CM, Camm, N, Crinnion, LA et al. (6 more authors) (2017) Characterization and Genomic Localization of a SMAD4 Processed Pseudogene. Journal of Molecular Diagnostics, 19 (6). pp. 933-940. ISSN 1525-1578

https://doi.org/10.1016/j.jmoldx.2017.08.002

Accepted Manuscript



Characterization and Genomic Localization of a *SMAD4* Processed Pseudogene

Christopher M. Watson, Nick Camm, Laura A. Crinnion, Agne Antanaviciute, Julian Adlard, Alexander F. Markham, Ian M. Carr, Ruth Charlton, David T. Bonthron

PII: S1525-1578(17)30316-1

DOI: 10.1016/j.jmoldx.2017.08.002

Reference: JMDI 633

To appear in: *The Journal of Molecular Diagnostics*

Accepted Date: 16 August 2017

Please cite this article as: Watson CM, Camm N, Crinnion LA, Antanaviciute A, Adlard J, Markham AF, Carr IM, Charlton R, Bonthron DT, Characterization and Genomic Localization of a *SMAD4* Processed Pseudogene, *The Journal of Molecular Diagnostics* (2017), doi: 10.1016/j.jmoldx.2017.08.002.

Characterisation and genomic localisation of a *SMAD4* processed pseudogene

Characterisation of a SMAD4 pseudogene

Christopher M. Watson,*†‡ Nick Camm,* Laura A. Crinnion,*†‡ Agne Antanaviciute,† Julian Adlard,* Alexander F. Markham,† Ian M. Carr,†‡ Ruth Charlton,* and David T. Bonthron*†‡

From the Yorkshire Regional Genetics Service,* St. James's University Hospital, Leeds; the MRC Medical Bioinformatics Centre,† Leeds Institute for Data Analytics, and the MRC Single Cell Functional Genomics Centre,‡ University of Leeds, St. James's University Hospital, Leeds, United Kingdom

Corresponding author:
Dr. Christopher M. Watson
6.2 Clinical Sciences Building
Yorkshire Regional Genetics Service
St James's University Hospital
Leeds, LS9 7TF
United Kingdom
Email: c.m.watson@leeds.ac.uk

This work was supported by grants MR/M009084/1 and MR/L01629X/1 awarded by the UK Medical Research Council.

ABSTRACT

Like many clinical diagnostic laboratories, we undertake routine investigation of cancer-predisposed individuals by high-throughput sequencing of patient DNA that has been target-enriched for genes associated with hereditary cancer. Accurate diagnosis using such reagents requires alertness against rare non-pathogenic variants that may interfere with variant calling. In a cohort of 2,042 such cases, we identified five that initially appeared to be carriers of a 95-bp deletion of *SMAD4* intron 6. More detailed analysis indicated that these individuals all carried one copy of a *SMAD4* processed gene. Because of its interference with diagnostic analysis, we characterized this processed gene in detail. Whole genome sequencing and confirmatory Sanger sequencing of junction PCR products were used to show that in each of the five cases, the *SMAD4* processed gene was integrated at the same position on chromosome 9, located within the last intron of the *SCAI* gene. This rare polymorphic processed gene therefore reflects the occurrence of a single ancestral retrotransposition event. Compared to the reference *SMAD4* mRNA sequence NM_005359.5 (https://www.ncbi.nlm.nih.gov/nucleotide/), the 5' and 3' UTR regions of the processed gene are both truncated, but its open reading frame is unaltered. Our experience leads us to advocate the use of an RNA-seq aligner, as part of diagnostic assay quality assurance, since this allows their recognition in a comparatively facile automated fashion.

INTRODUCTION

The availability of diagnostic molecular genetic assays has increased significantly in recent years. This has largely been due to the ubiquitous adoption of next generation sequencing (NGS) instruments, which have replaced comparatively low-throughput Sanger sequencing technology, as the standard technique for mutation detection. New laboratory assays combined with ever-increasing automation is resulting in increased patient throughput and more efficient workflows. That several genes can be analysed concurrently has enabled an expansion of testing for heterogeneous genetic disorders which may have previously been considered too rare for a *bona-fide* genetic test to have been established and offered in a routine clinical laboratory. To be able to request a comprehensive analysis of all genes that correspond to a patient's phenotype is transforming diagnostic referral pathways, by eliminating costly 'test and review' processes that are necessary when referrals are made in a consecutive manner.

Operational requirements associated with test portfolios that can accommodate varying combinations of target genes have necessitated a fundamental transformation in assay design. Typically, a far larger range of targets are selected for sequencing than is suggested *a priori* from the patient's presenting phenotype. An *in silico* virtual gene panel is applied to these data thus masking inappropriate results from those requested by the referring clinician. Although this approach generates unnecessary sequence data, laboratories are able to reduce the complexity of wet-laboratory processes thereby streamlining their workflows. As the cost of DNA sequencing continues to fall the number of genes that can be feasibly targeted, while maintaining iteratively comparable test sensitivity, will continue to increase. Indeed, our originally reported 36-gene

reagent has been periodically revised and presently targets the coding exons of 155 cancer-associated genes [1].

For many commentators, the long-held aspiration that custom-designed panels will be replaced by exome- and subsequently whole genome-sequencing, is being expedited by large-scale, population based, sequencing projects. Nevertheless, the prevailing approach for performing target enrichment, using probe-based hybridisation, has overcome the need to design and optimise long-range PCR amplicons [2]. This has improved the scalability of targeted loci, as previously only a finite number of long-range PCR primer pairs could be handled by a single laboratory. Despite this advance, hybridisation capture methods have a lower specificity for target enrichment due to the capture of 'off-target' sequences. A comparatively greater number of reads it therefore required to achieve the same depth of coverage (although this is typically off-set by no longer needing to sequence a gene's introns).

Off-target sequences are captured for reasons that may include hybridisation of probes to low-diversity nucleotide sequences, sequence homology between the targeted region and that of a related gene family member or an interfering pseudogene, or reaction kinetics. Although off-target reads are typically ignored, a number of studies have demonstrated their utility for the inadvertent identification of single nucleotide polymorphisms [3] and as a source of low-coverage whole genome sequencing reads for genomewide copy-number analysis [4]. Less useful is the capture and sequencing of DNA fragments that are highly homologous to target loci; it is usually not possible to determine the true genomic origin of these resulting data. As pseudogene sequences

may therefore affect the interpretation of clinical assays their identification and characterisation is of particular importance to the diagnostic community.

A *SMAD4* processed pseudogene was recently detected in a subset of patients referred for diagnostic analysis of hereditary cancer predisposition genes [5]. *SMAD4* is associated with both juvenile polyposis syndrome (OMIM: 174900) and combined juvenile polyposis/hereditary hemorrrhagic telangiectasia syndrome (OMIM: 175050). Here we corroborate this observation and assess the frequency of the *SMAD4* pseudogene in our cohort of 2,042 diagnostically referred hereditary cancer cases. We further define the genomic integration site and report the transcript structure following end-to-end sequencing of the identified *SMAD4* processed pseudogene.

MATERIALS AND METHODS

Patients were referred to the Leeds Genetics Laboratory for diagnostic testing of one or more hereditary cancer predisposition genes using a custom-designed SureSelect hybridisation enrichment assay (Agilent Technologies, Wokingham, UK). The original 36-gene reagent has been iteratively redesigned since the service was launched in 2013 [1] and now targets the exons and immediate flanking sequence of 155 hereditary cancer genes.

DNA was isolated from blood lymphocytes using either a standard salting out method or the Chemagic™ 360 automated extractor (PerkinElmer, Seer Green, UK). For each sample, an Illumina-compatible sequencing library was generated. Initially, 3 µg of genomic DNA was sheared using a Covaris S2 or E220 (Covaris Inc., Woburn, MA, USA) before whole genome library preparation was undertaken using SureSelect XT reagents (Agilent Technologies, Wokingham, UK). This consisted of end-repair, (A)-addition, adaptor ligation and PCR enrichment. A custom RNA probeset was used to perform a targeted capture hybridisation on each of the whole genome libraries, following manufacturer's protocols throughout. Samples were initially prepared manually, but a fully automated solution has since been introduced using a Sciclone G3 liquid handling workstation (PerkinElmer, Seer Green, UK). The quality and concentration of final libraries were confirmed using either an Agilent Bioanalyser or Agilent Tapestation (Agilent Technologies, Wokingham, UK) before, typically, 16 samples were combined into a single batch for sequencing. Each batch was either sequenced on a single lane of an Illumina HiSeq2500 rapid-mode flow cell (2 × 101 bp sequencing reads) or pooled with two additional batches and sequenced on an Illumina NextSeq500 (2 × 151 bp sequencing reads) using a High Output flow cell using version 2 chemistry (Illumina

Inc., San Diego, CA, USA). Raw sequence data was converted to FASTQ.gz format using bcl2fastq v.2.17.1.14.

A common data processing pipeline, running on the Leeds high-performance computer MARC1 (http://arc.leeds.ac.uk/systems/marc1/), was applied to each of the per-sample directories from the SureSelect target enrichment assay. Initially, adaptor sequences and low-quality bases (Q score ≤10) were trimmed from reads using Cutadapt v.1.9.1 (https://github.com/marcelm/cutadapt) [6]. The resulting analysis-ready reads were assessed using FastQC v.0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were next aligned to an indexed human reference genome (hg19) using BWA MEM v.0.7.13 (https://sourceforge.net/projects/bio-bwa/files/) [7] before being sorted by chromosome coordinate and having PCR duplicates marked by Picard v.2.1.1 (http://broadinstitute.github.io/picard/) to create a processed.bam file. These data were realigned using ABRA v.0.97 (https://github.com/mozack/abra) [8] and the Genome Analysis Toolkit (GATK) v.3.6-0 was used to perform variant calling following best practice guidelines. This involved indel realignment, base quality score recalibration and variant calling using the Haplotypecaller to generate a per-sample VCF file [9]. These variant data were annotated using Alamut Batch Standalone v.1.4.4 (database v.2016.03.04) (Interactive Biosoftware, Rouen, France). Coverage metrics were determined using the GATK walkers DepthOfCoverage, CallableLoci and CountReads. Visualisation of aligned sequence reads was performed with the Integrative Genome Viewer v.2.3.80 (http://software.broadinstitute.org/software/igv/) [10]. The analysis-ready reads for five samples with apparent *SMAD4* intron 6 deletions were aligned to an indexed hg19 reference genome annotated using GENCODE Release

26 using the RNA-seq aligner STAR v.2.5.3a with default settings (https://github.com/alexdobin/STAR/) [11].

Illumina-compatible whole genome sequencing libraries were subsequently prepared for the same five samples. Approximately 3 μg DNA was sheared using a Covaris S2 prior to end-repair, (A)-addition and adaptor ligation steps being undertaken using NEBNext® Ultra™ reagents, following manufacturer's protocols (New England Biolabs, Ipswich, MA, USA). An ampure size selection ratio for a 300-bp to 400-bp insert and a 6-cycle enrichment PCR was performed. Following an assessment of library quality, the final libraries were pooled in equimolar concentrations and the pooled batch was sequenced using an Illumina NextSeq500 High Output flow cell generating 2 × 151 bp read lengths. For each sample, a processed.bam file was generated using the same bioinformatics pipeline described above. Sequence reads mapping to the *SMAD4* locus (chr18:48550000-48620000) were extracted from the coordinate-sorted duplicate-marked bam file using samtools v.0.1.18 with the options -q 1 and -F 14 [12]. These filters ensured that the mapped read quality score was greater than 0, that neither read in the pair was unmapped and that the pair was not considered to be a "proper pair". Read pairs with one read mapping outside the *SMAD4* locus and whose non-*SMAD4* read clustered within 500 bp of the nearest of non-*SMAD4* read were reviewed and compared between patients.

Three PCR amplicons were generated to amplify across the breakpoints identified by medium coverage whole genome sequencing. The specificity of the amplicons was evaluated for each reaction; one primer was located within *SCAI* intron 18, and the second primer within the *SMAD4* pseudogene.

Two amplicons spanning the 5' end of the *SMAD4* pseudogene were designed and amplified. The first comprised a common *SCAI*-bound forward primer 5'-CTGAGCTTGTGATCTGCCTG-3' and the *SMAD4* exon 2-located reverse primer 5'-TGAAGCCTCCCATCCAATGT-3'. Each PCR reaction consisted of 7.46 µl nuclease-free H_2O, 1.2 µl, 10× Buffer + Mg, 0.12 µl dNTPs, 2.4 µl GC-rich buffer, 0.12 µl Faststart Taq polymerase, 0.1 µl 10 µM forward primer, 0.1 µl 10 µM reverse primer and 0.5 µl of approximately 100 ng/µl DNA (Roche Diagnostics Ltd., Burgess Hill, U.K.). Thermocycling conditions were 96°C for 5 minutes, followed by 35 cycles of 96°C for 30 seconds, 55°C for 30 seconds, 72°C for 2 minutes and a final 72°C extension step for 10 minutes. The second amplicon was amplified using the same common *SCAI*-bound forward primer and a reverse primer specific to *SMAD4* exon 8 5'-TGGAAATGGGAGGCTGGAAT-3'. PCR reagents and volumes were equivalent to the first reaction. Thermocycling conditions were the same, but an additional 5 cycles were performed. PCR products from the second reaction were gel-extracted and purified using a QIAquick column following manufacturer's protocols (Qiagen GmbH, Hilden, Germany). Sanger sequencing was performed on PCR products from both reactions using amplification primers and, for the second reaction, a further two internally sited *SMAD4* exon 2 primers (5'-TTCCTTGCAACGTTAGCTGT-3' and 5'-ACATTGGATGGGAGGCTTCA-3') with an ABI3730 following manufacturer's instructions (Applied Biosystems, Paisley, UK).

The 3' end of the *SMAD4* processed pseudogene was amplified using a forward primer bridging the *SMAD4* exon 5/6 junction 5'-ACAAGTCAGCCTGCCAGTAT-3' and an *SCAI* reverse primer 5'-CAGGAAACAGCTATGACCTGCAATGACTCGATCTCAGC-3'. The reverse

primer contained a universal tag (underlined) for Sanger sequencing using our routine diagnostic workflow. Each reaction consisted of 12.74 µl nuclease-free H_2O, 2 µl SequalPrep™ 10× Reaction Buffer, 0.36 µl SequalPrep™ 5U/µl Long Polymerase, 0.4 µl dimethyl sulfoxide (DMSO), 2 µl SequalPrep™ 10× Enhancer A, 1 µl 10 µM forward primer, 1 µl 10 µM reverse primer and 0.5 µl of approximately 100 ng/µl DNA (Invitrogen, Paisley, UK). Thermocycling conditions comprised a denaturation step of 94°C for 2 minutes, followed by 10 cycles of 94°C for 10 seconds, 60°C for 30 seconds and 68°C for 3 minutes then 25 cycles of 94°C for 10 seconds 60°C for 30 seconds, 68°C for 3 minutes with an additional 20 seconds added per cycle, before a final extension step at 72°C for 5 minutes. PCR products of approximately 2 kb were gel extracted and purified using the QIAquick column following the manufacturer's protocol. Sanger sequencing was performed using the amplification forward primer, universal reverse primer and the following internally sited primers: 5'-AGCCATTGAGAGAGCAAGGT -3' (SMAD4 exon 9/10 forward), 5'-CCTCCAGCTCCTAGACGAAG-3' (SMAD4 exon 12 forward), 5'-CCATGTGGGTGAGTTAATTTTACC-3' (SMAD4 exon 12 forward), 5'-TGGAAATGGGAGGCTGGAAT-3' (SMAD4 exon 8 reverse), 5'-AAAGCAGCGTCACTCTACCT-3' (SMAD4 exon 12 forward) and 5'-TCAGTTTTTGTATCTTGGGGCA-3' (SMAD4 exon 12 forward).

Sequence chromatograms for all Sanger sequencing reactions were analysed using 4Peaks v.1.8 (http://nucleobytes.com/4peaks/index.html).

RESULTS

Since 2013, we have used a custom-hybridisation enrichment assay and NGS pipeline for the diagnostic analysis of hereditary cancer genes [1]. In the present study, we

retrospectively examined 2,042 patient libraries that had been sequenced in 131 batches. We noticed five cases in which our standard variant-calling pipeline identified an apparent 95-bp deletion, corresponding to the entire *SMAD4* intron 6 nucleotide sequence (c.787+1_788-1del, NM_005359.5, https://www.ncbi.nlm.nih.gov/nucleotide/). Assay performance metrics for each of these five libraries are displayed in Supplemental Table S1.

Visualisation of *SMAD4* read coverage charts for the five cases with an apparent intron 6 deletion revealed plots with prominent 'cliff-edge' shaped profiles, the discontinuities in which aligned with the *SMAD4* exon-intron boundaries (Supplemental Figure S1). This was particularly conspicuous for *SMAD4* exon 8. Close inspection of these data established that reads at the exon-intron boundaries had been "soft-clipped". To further investigate whether these soft-clipped reads spanned *SMAD4* exon-to-exon splice junctions, sequence reads were mapped to a transcript-annotated human genome, using the RNA-seq aligner STAR. Resulting Sashimi plots displaying splice junction read counts were consistent with the presence of a spliced *SMAD4* sequence whose exon structure matched that of the reference mRNA sequence NM_005359.5 (https://www.ncbi.nlm.nih.gov/nucleotide/) (Figure 1). These data thus suggested the presence of a processed (intron-lacking) *SMAD4* pseudogene in these five individuals.

The existence of such a pseudogene was indeed recently reported [5], although its structure was not characterized in detail. The frequency (5/2042 = 0.24%) of cases we observed carrying the *SMAD4* pseudogene was in keeping with that reported by Millson et al. (12/4672 = 0.26%). The likely interference of the pseudogene with diagnostic

testing prompted us to define its exact structure, and address the question of whether its sequence and location are identical among carriers.

To assess the relative number of copies of the *SMAD4* pseudogene, we determined the ratio of gapped (pseudogene-derived) to non-gapped (non-pseudogene) read alignments spanning intron 6. (Although this region was not specifically targeted when designing the capture enrichment probes, its small size and proximity to *SMAD4* exons 5 and 6 ensured that the intron was fortuitously sequenced.) The ratio of gapped:non-gapped reads was approximately 1:2, suggesting that only a single copy of the pseudogene was present in each case (Table 1). Further, by comparing the normalised read-depths of these cases to controls from the same sequencing batches, we determined relative dosage values for each *SMAD4* exon. These results indicated the presence of three copies of most of the *SMAD4* exons, again indicative of a single copy of the *SMAD4* pseudogene (Supplemental Table S2). Although data for exons 4 and 8 deviate from this interpretation, this is probably due to the small genomic intervals represented by these exons (30 bp and 51 bp, respectively). Additionally, the greater variability displayed by sample 1 is probably attributable to the reduced number of available intra-batch controls (9 samples, *vs.* 15 samples for the other 4 cases).

Retrospective variant calling was undertaken using VarScan2 [13], to assess the allelic ratios of coding and non-coding variants. No non-reference coding variants were identified. However, for sample 4, two variants c.905-52A>G (rs948589) and c.955+58C>T (rs948588) were present, in introns 7 and 8 respectively. The non-reference read frequencies were 47% for c.905-52A>G (681 of 1455 reads) and 46% for c.955+58C>T (722 of 1562 reads). This diploid allelic ratio further supports the

inference that the *SMAD4* pseudogene is processed, allelic ratios of intronic SNPs being unaffected by the presence of the pseudogene.

To determine whether a common *SMAD4* pseudogene integration site was shared between the five cases, medium-coverage whole genome sequencing (approximately 9× per sample) was performed. Evidence for the integration site being located on chromosome 9, within intron 18 of the *SCAI* gene, was provided by the 16 read pairs detailed in Table 2. These data characterise DNA fragments whose opposite ends were each mapped to (a) *SCAI* intron 18 and (b) either the 5' (14 read pairs) or 3' (2 read pairs) end of *SMAD4*. Soft-clipped reads spanning the precise integration site indicated that this was identical, at least among samples 2-5. (For sample 1, no supporting read-pairs were identified, despite there being no obvious difference between the assay performance metrics, as displayed in Supplemental Table S3.) *SMAD4* mapped reads indicated that the pseudogene sequences for exons 1 and 12 were shorter than those reported in transcript record NM_005359.5 (https://www.ncbi.nlm.nih.gov/nucleotide/). However, the precise terminal nucleotide of the 3'-UTR could not be determined from this dataset. This was probably due to the presence of the poly-A tail, hindering DNA sequencing and mapping, and resulting in an underrepresentation of exon 12 mapped read pairs. Interestingly, the library insert for the sample 5 read pair 4:23601:11116:11521 was sufficiently large that the SMAD4-mapped read spanned the exon 1-2 splice junction.

To confirm the identified integration site, and establish the terminal nucleotide of the *SMAD4* 3'-UTR, three overlapping PCR amplicons, each anchored at one end by a primer bound to *SCAI* intron 18, were amplified and sequenced (Figure 2). All five cases were

confirmed to have the same genomic integration site, at which the inserted pseudogene is flanked by a 4-nt microduplication (TTTC). The exon-exon arrangement was identical to transcript record NM_005359.5 (https://www.ncbi.nlm.nih.gov/nucleotide/), and no nucleotide sequence variants were identified in any of the pseudogene exons. Compared to the mRNA reference sequence, 41 nt are missing from the beginning of the *SMAD4* 5'-UTR and 5,265 nucleotides are absent from the end of the 3'-UTR. A schematic representation of the integration site and scale drawing of the gene structure are displayed in Figure 3.

DISCUSSION

In recent years, the significantly increased number of genes that are attributable to clinically recognisable phenotypes have resulted in far greater scope for genetic testing. Laboratories typically create target enrichment panels that sequence more loci than are requested by the referring clinician and the unwanted variant data is masked by creating virtual gene panels *in silico*. While this approach facilitates the creation of efficient wet-laboratory processes, it also generates sequence data that is not routinely analysed. For the purposes of this study we harnessed these data to determine the frequency of a reported *SMAD4* processed pseudogene in our cohort of patients that had been referred for hereditary cancer testing. We determined the pseudogene to be present at a frequency of 1 in 408, which is consistent with the previously reported frequency of 1 in 389 [5]. That the integration site was common to all five patients suggests that this reflects a single ancestral founder event. Given that the majority of our laboratory's referrals are of northern European ancestry it will be interesting to determine whether this variant is also detected in more diverse ethnic populations. Unsurprisingly, many other polymorphic processed genes have been found to be

344 restricted to certain ethnic groups [14]. Polymorphic processed genes of the present
345 type have been revealed by large-scale sequencing surveys to be a frequent feature of
346 the human (and mouse) genome. Although the insertion site of the *SMAD4* processed
347 gene was not determined in the large-scale studies of Ewing et al., (2013) and Shrider et
348 al., (2013) [14, 15].
349
350 Most processed genes in the reference human genome are known to be non-functional
351 (*i.e.* they are processed pseudogenes), either because they lack promoter sequences
352 ("dead on arrival") or have acquired inactivating mutations subsequent to
353 retrotransposition. However, processed genes whose existence is polymorphic within
354 the normal population are likely to have been recently transposed, and therefore (as in
355 the present case) not to have acquired many inactivating mutations. There is
356 population-level evidence that new processed genes are frequently subject to positive
357 or negative evolutionary selection [15] as well as anecdotal examples of individual
358 functional effects of processed genes (discussed in Richardson et al., (2014)) [16].
359
360 Since the coding region of the *SMAD4* processed gene is unaltered in comparison to its
361 parent gene, we cannot be completely certain that it is non-functional (*i.e.* that it really
362 is a processed pseudogene). We have been unable to address this question, since RNA is
363 not available from any of the five carrier individuals, to permit analysis of whether the
364 processed gene is transcribed. For the same reason, we cannot address any possible
365 effect of the retrotransposed gene on the splicing of the *SCAI* gene, within which it is
366 integrated. A newly transposed processed gene can be disease-causing as a result of
367 disruption of splicing of its target gene [17].
368

SCAI itself is a nuclear protein that was first characterized for its suppressive effects upon tumour cell invasiveness, through regulation of beta1-integrin expression [18]. It has also been shown to be a TP53BP1 interaction partner with an important role in double-strand break repair [19]. It has been reported that the *SCAI* 3'-UTR contains a binding site for miR-1228. When bound, this microRNA is capable of down-regulating endogenous SCAI protein [20]. Furthermore, SCAI levels have been observed to be down-regulated in human tumours leading to reports of its tumour suppressor characteristics. RNA interference experiments of *SCAI* have shown an upregulation of β1-integrin gene expression and a resulting increase in invasive cell migration. Despite these observations, we were unable to obtain relevant tissue specimens from our patients to determine whether *SCAI* expression is perturbed by the presence of the *SMAD4* pseudogene.

Pseudogenes commonly interfere with the diagnostic analysis of clinically important genes. In extreme cases, unambiguous analysis may be impossible without resort to highly specialized methodologies; such is the case for mutations in *PMS2*, which in the heterozygous or biallelic state cause low-penetrance colorectal cancer predisposition (Lynch syndrome; OMIM: 614337), and a young-onset mismatch repair cancer syndrome (OMIM: 276300), respectively [21, 22]. Typically, however, because pseudogenes are not polymorphic, assay designs can be tailored to avoid interference and allow robust and reliable clinical diagnosis.

The ad-hoc discovery of polymorphic processed pseudogenes is likely to become more frequent as an increasingly genomic approach is applied to molecular diagnostic investigations. It is perhaps therefore surprising that given the clinical importance of

SMAD4 [23], no comprehensive analysis of the *SMAD4* pseudogene integration site had hitherto been undertaken.

While the initial identification of the *SMAD4* pseudogene stemmed from aberrant MLPA result, the clinical adoption of NGS-based hybridisation enrichment panels is outpacing the production of gene-specific MLPA kits. Consequently, the per-exon cost of performing MLPA to detect novel pseudogenes, on a large-scale, would likely be cost-prohibitive. Our study demonstrates a convenient approach of using an RNA-seq aligner to detect processed pseudogenes from hybridisation capture data. We also report how comparative read depth methods can effectively determine the allelic copy number of novel pseudogene sequences. Increased demand for genetic testing has meant laboratories are becoming ever-more reliant on automated variant calling pipelines that do not involve visualisation of the directly sequenced reads, and clinical scientists are required to interpret sequence variants for unfamiliar genes. To maintain quality assurance of these tests, we advocate the inclusion of an RNA-seq aligner into laboratory pipelines as a means of detecting as-yet unreported polymorphic processed pseudogenes which, if they remain undetected, could interfere with the interpretation of clinical results.

In summary, we report a common genomic integration site for the polymorphic *SMAD4* processed pseudogene. We demonstrate how alignment of these data using an RNA-seq aligner can confirm the presence of splice-junction containing reads. And advocate that as the number of genes analysed by clinical laboratories continues to expand this would provide a worthwhile quality assurance approach for target enrichment experiments.

REFERENCES

1. Watson CM, Crinnion LA, Morgan JE, Harrison SM, Diggle CP, Adlard J, Lindsay HA, Camm N, Charlton R, Sheridan E, Bonthron DT, Taylor GR, Carr IM: Robust diagnostic genetic testing using solution capture enrichment and a novel variant-filtering interface. Hum Mutat 2014, 35:434-441

2. Morgan JE, Carr IM, Sheridan E, Chu CE, Hayward B, Camm N, Lindsay HA, Mattocks CJ, Markham AF, Bonthron DT, Taylor GR: Genetic diagnosis of familial breast cancer using clonal sequencing. Hum Mutat 2010, 31:484-491

3. Guo Y, Long J, He J, Li CI, Cai Q, Shu XO, Zheng W, Li C: Exome sequencing generates high quality data in non-target regions. BMC Genomics 2012, 13:194

4. Bellos E, Coin LJ: cnvOffSeq: detecting intergenic copy number variation using off-target exome sequencing data. Bioinformatics 2014, 30:i639-645

5. Millson A, Lewis T, Pesaran T, Salvador D, Gillespie K, Gau CL, Pont-Kingdon G, Lyon E, Bayrak-Toydemir P: Processed Pseudogene Confounding Deletion/Duplication Assays for SMAD4. J Mol Diagn 2015, 17:576-582

6. Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 2011, 17:10-12

7. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754-1760

8. Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS: ABRA: improved coding indel detection via assembly-based realignment. Bioinformatics 2014, 30:2813-2815

9. 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, 43:491-498

10. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013, 14:178-192

11. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29:15-21.

12. 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

13. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012, 22:568-576

14. Ewing AD, Ballinger TJ, Earl D; Broad Institute Genome Sequencing and Analysis Program and Platform, Harris CC, Ding L, Wilson RK, Haussler D: Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol 2013, 14:R22

15. Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ: Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet 2013, 9:e1003242

16. Richardson SR, Salvador-Palomeque C, Faulkner GJ: Diversity through duplication: whole-genome sequencing reveals novel gene retrocopies in the human population. Bioessays 2014, 36:475-481

17. de Boer M, van Leeuwen K, Geissler J, Weemaes CM, van den Berg TK, Kuijpers TW, Warris A, Roos D: Primary immunodeficiency caused by an exonized retroposed gene copy inserted in the CYBB gene. Hum Mutat 2014, 35:486-496

18. Brandt DT, Baarlink C, Kitzing TM, Kremmer E, Ivaska J, Nollau P, Grosse R: SCAI acts as a suppressor of cancer cell invasion through the transcriptional control of beta1-integrin. Nat Cell Biol 2009, 11:557-568

19. Hansen RK, Mund A, Poulsen SL, Sandoval M, Klement K, Tsouroula K, Tollenaere MA, Räschle M, Soria R, Offermanns S, Worzfeld T, Grosse R, Brandt DT, Rozell B, Mann M, Cole F, Soutoglou E, Goodarzi AA, Daniel JA, Mailand N, Bekker-Jensen S: SCAI promotes DNA double-strand break repair in distinct chromosomal contexts. Nat Cell Biol 2016, 18:1357-1366

20. Lin L, Liu D, Liang H, Xue L, Su C, Liu M: MiR-1228 promotes breast cancer cell growth and metastasis through targeting SCAI protein. Int J Clin Exp Pathol 2015, 8:6646-6655

21. De Vos M, Hayward BE, Picton S, Sheridan E, Bonthron DT: Novel PMS2 pseudogenes can conceal recessive mutations causing a distinctive childhood cancer syndrome. Am J Hum Genet 2004, 74:954-964

22. Goodenberger ML, Thomas BC, Riegert-Johnson D, Boland CR, Plon SE, Clendenning M, Win AK, Senter L, Lipkin SM, Stadler ZK, Macrae FA, Lynch HT, Weitzel JN, de la Chapelle A, Syngal S, Lynch P, Parry S, Jenkins MA, Gallinger S, Holter S, Aronson M, Newcomb PA, Burnett T, Le Marchand L, Pichurin P, Hampel H, Terdiman JP, Lu KH, Thibodeau S, Lindor NM: PMS2 monoallelic mutation carriers: the known unknown. Genet Med 2016, 18:13-19

23. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, Herman GE, Hufnagel SB, Klein TE, Korf BR, McKelvey KD, Ormond KE, Richards CS, Vlangos CN, Watson M, Martin CL, Miller DT: Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0):

a policy statement of the American College of Medical Genetics and Genomics. Genet Med 2017, 19:249-255

FIGURE LEGENDS

Figure 1: *SMAD4* Sashimi plots generated following alignment of targeted capture data using the RNA-seq aligner STAR. Each arc's corresponding value records the number of reads crossing the reported splice junction. Alignment coverage data is displayed with *y*-axis values ranging from 0-20,000 for Sample 1 and 0-6,000 for all other samples.

Figure 2: DNA sequence at the common *SMAD4* processed gene integration site, located within *SCAI* intron 18 (using reference transcript NM_173690.4, https://www.ncbi.nlm.nih.gov/nucleotide/). Genomic coordinates refer to human reference genome build hg19. **(A)** The dashed red line marks the breakpoint 5' to the processed gene. **(B)** The last nucleotide matching the *SMAD4* 3' untranslated region is identified, immediately to the left of the vertical dashed line. To the right of this line is a poly(A) sequence. **(C)** The *SCAI* intron 18 integration site beyond the poly(A) tail. This sequence was generated using a reverse strand primer. The four nucleotides located between the dashed red lines are duplicated from the proximal breakpoint.

Figure 3: A schematic representation of the *SCAI* locus, displaying the exon arrangement of the *SMAD4* processed pseudogene, which is consistent with that reported for NM_005359.5 (https://www.ncbi.nlm.nih.gov/nucleotide/). Exons (green boxes) are drawn to scale using GeneDrawer (www.insilicase.com/Desktop/GeneDrawer.aspx, last accessed August 18, 2017).

Table 1: The ratio of gapped to non-gapped sequence alignments in cases with an apparent *SMAD4* intron 6 deletion.

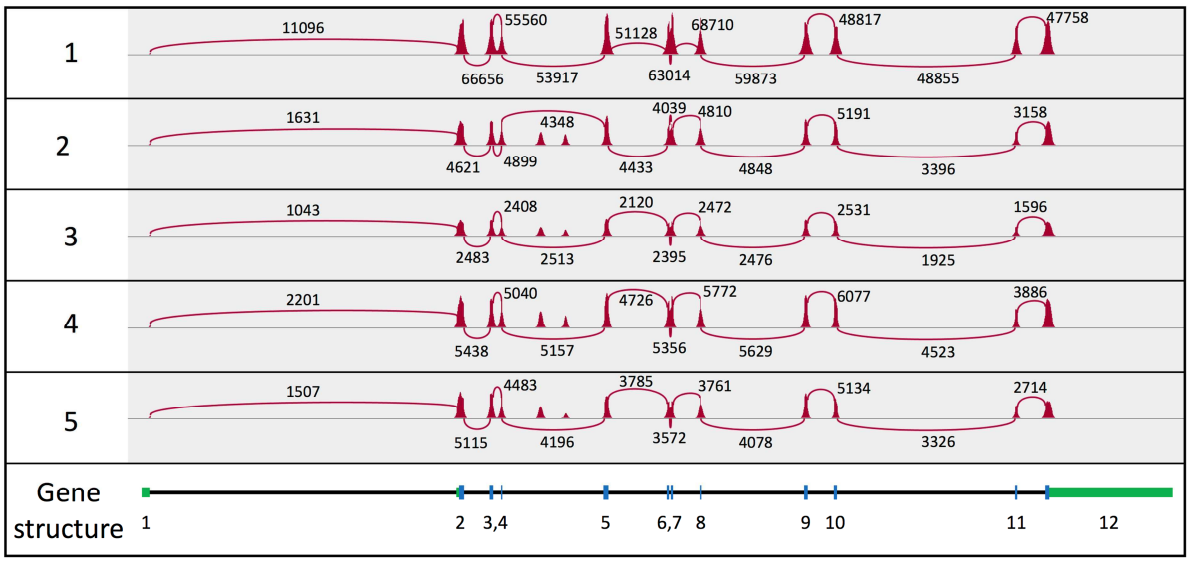
Sample	Gapped read alignments spanning intron 6	Mean per-base read depth for intron 6 nucleotides	Ratio of gapped to non-gapped reads
1	1,620	3,069	1:1.89
2	739	1,750	1:2.37
3	571	1,183	1:2.07
4	1,198	2,250	1:1.88
5	770	1,323	1:1.72

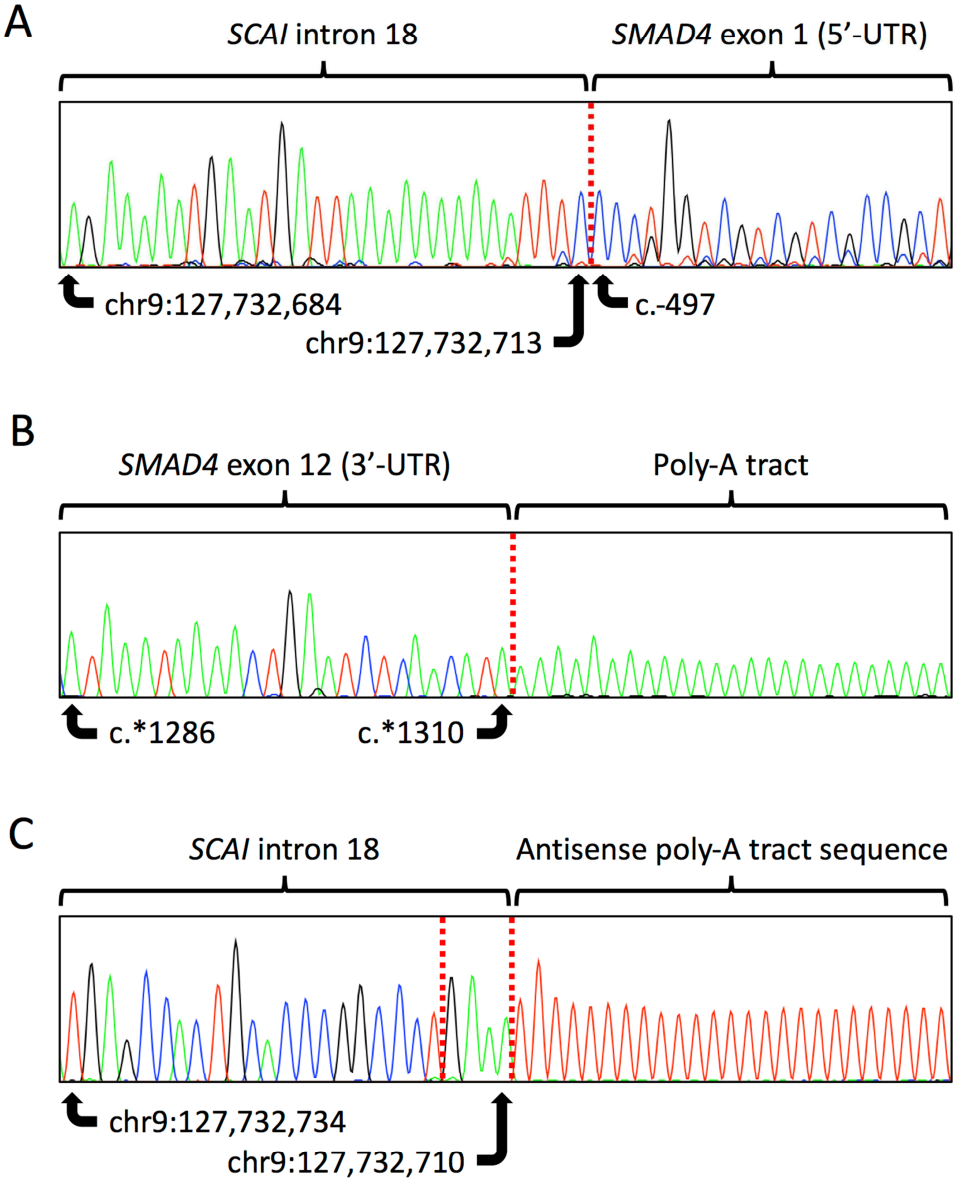
Intron numbering determined according to NM_005359.5 (https://www.ncbi.nlm.nih.gov/nucleotide/).

Table 2: Characteristics of whole genome sequencing reads supporting the intragenic *SCAI* integration site.

Sample	Read pair ID	Read 1					Read 2				
		Locus	Chr.	Start	Str.	CIGAR	Locus	Chr.	Start	Str.	CIGAR
2	4:13608:15564:14605	5'-SMAD4	18	48,556,641	-	151M	SCAI	9	127,732,358	+	150M
2	2:13210:1908:2419	SCAI	9	127,732,501	+	151M	5'-SMAD4	18	48,556,701	-	6S143M
2	4:22402:24489:4305	5'-SMAD4	18	48,556,700	-	151M	SCAI	9	127,732,506	+	150M
2	1:11204:8894:18752	5'-SMAD4	18	48,556,624	-	60S91M	SCAI	9	127,732,633	+	81M70S
2	4:21606:19277:14299	SCAI	9	127,732,700	-	24S127M	3'-SMAD4	18	48,605,924	+	101M49S
3	3:22511:22720:13254	5'-SMAD4	18	48,556,622	-	151M	SCAI	9	127,732,422	+	151M
3	2:11311:9152:15694	5'-SMAD4	18	48,556,624	-	60S91M	SCAI	9	127,732,437	+	151M
3	2:21212:20833:4797	SCAI	9	127,732,556	+	150M	5'-SMAD4	18	48,556,624	-	64S87M
3	1:21211:11117:11719	3'-SMAD4	18	48,605,995	+	150M	SCAI	9	127,732,700	-	87S62M
4	3:13407:5904:6688	5'-SMAD4	18	48,556,711	-	151M	SCAI	9	127,732,459	+	150M
4	1:11210:11536:3015	5'-SMAD4	18	48,556,624	-	58S93M	SCAI	9	127,732,487	+	151M
4	2:21206:5607:15745	5'-SMAD4	18	48,556,624	-	48S103M	SCAI	9	127,732,604	+	110M41S
5	4:13501:18475:17089	5'-SMAD4	18	48,556,624	-	45S106M	SCAI	9	127,732,552	+	150M
5	4:11605:22916:12006	SCAI	9	127,732,569	+	145M6S	5'-SMAD4	18	48,556,635	-	151M
5	4:12410:4727:7249	5'-SMAD4	18	48,556,800	-	151M	SCAI	9	127,732,582	+	132M19S
5	4:23601:11116:11521	SCAI	9	127,732,607	+	107M44S	5'-SMAD4	18	48,556,882	-	2S114M35S

Str.: Strand. CIGAR: The mapping defined by the BWA alignment. All coordinates are provided according to human genome build hg19. Locus represents the read mapping to one of three possible loci, either the 5' end of the *SMAD4* pseudogene (5'-SMAD4), the 3' end of the *SMAD4* pseudogene (3'-SMAD4), or the *SCAI* integration site (SCAI).




A
SCAI intron 18
SMAD4 exon 1 (5'-UTR)
chr9:127,732,684
chr9:127,732,713
c.-497
B
SMAD4 exon 12 (3'-UTR)
Poly-A tract
c.*1286
c.*1310
C
SCAI intron 18
Antisense poly-A tract sequence
chr9:127,732,734
chr9:127,732,710