Addressing False Positive Variants Arising from Pseudogenes

Risha Govind[1,2], **Sam Wilkinson**[1,3], **Nicola Whiffin**[1,2], **Shibu John**[1,2], **Rachel J. Buchan**[1,2], **Elizabeth Edwards**[1,2], **Deborah J. Morris-Rosendahl**[1,3], **James S. Ware**[1,2], **P.J. Barton**[1,2], **Stuart A. Cook**[1,2]

[1]National Heart and Lung Institute, Imperial College, London, UK, [2]NIHR Cardiovascular BRU, Royal Brompton and Harefield NHS Trust, London, UK, [3]Royal Brompton and Harefield NHS Trust, London, UK.

INTRODUCTION

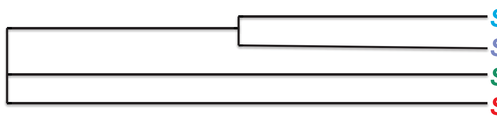
Clinical genetic testing has been transformed in recent years by the introduction of Next-Generation Sequencing (NGS). Targeted gene panels have made it possible to simultaneously analyse hundreds of genes with high confidence. However, since current aligners lack the sensitivity to distinguish reads that come from homologous parts of the genome, it is a challenge to work with genes with paralogues or pseudogenes. Pseudogenes arise from duplication of protein-coding genes, and have been considered to be degraded paralogues, due to loss of functionality.

SDHA is Homologous with 3 Pseudogenes

During the validation of our bioinformatics pipeline, we compared variants called by the two variant callers from GATK v3.2, HaplotypeCaller and UnifiedGenotyper. SNVs in *SDHA* which were only called by GATK HaplotypeCaller did not validate by Sanger Sequencing.

SDHA is a mitochondrial protein associated with Dilated Cardiomyopathy and is part of our 174 gene targeted panel for Inherited Cardiac Conditions.

Pair-wise alignment showed that 40-80% of the SDHA mRNA has 95% similarity with the non-coding RNAs of three pseudogenes (*SDHAP1, SDHAP2, SDHAP3*). There were less homology in the intronic regions.



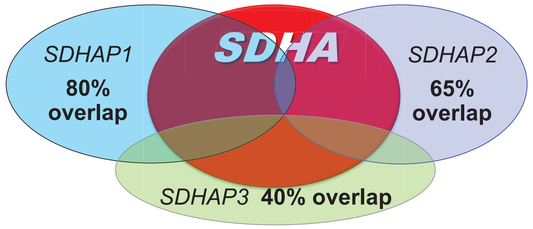


Figure 1: Phylogenetic tree showing evolutionary relationships of *SDHA* and the relative pseudogenes. Venn Diagram showing the Percentage of *SDHA* overlapping with pseudogene with 95% Similarity

False Positive Variants Arising from Pseudogenes

Due to high sequence similarity, sequenced reads which arise from a pseudogene may be wrongly aligned to the functional gene, and result in false positive variant calls, even in targeted panels. Furthermore, variants in regions of homology to a pseudogene are difficult to validate with Sanger sequencing.

We performed Multiple Sequence Alignment (MSA) to identify the regions of *SDHA* which overlap with pseudogenes. The variants which did not validate by Sanger sequencing were in regions which were homologous to all three of pseudogenes.

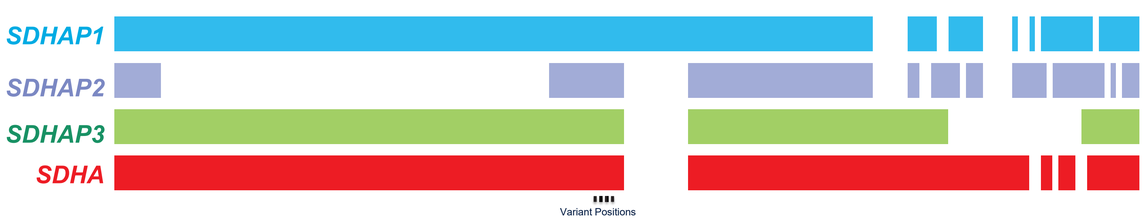


Figure 2: MSA of *SDHA* and pseudogenes. Positions of the variants which did not validate by Sanger Sequencing are marked.

Bioinformatics Approach

We found four ways to identify these variant computationally.

Odd Allelic Balance

Allelic Depth (AD) information from VCF files can be used to calculate Allelic Balance, the ratio of AD between reference and alternate; a ratio not close to 50 or 100 is a sign of a false positive variant.

```
chr5   236628   rs201139275   C   T   831.77   PASS   ...   GT:AD:GQ:PL   0/1:229,36:99:860,0,16855
chr5   236678   rs111387770   G   A   930.77   PASS   ...   GT:AD:GQ:PL   0/1:241,45:99:959,0,5886
```

Figure 3: Allelic balance is odd when the ratio of Allelic Depth (AD) is not ~100% or ~50%. VCF files contains AD.

Corresponding position in Pseudogenes

Perform Multiple Sequence Alignment (MSA) to compare the alternate allele with the base at corresponding position in pseudogenes.

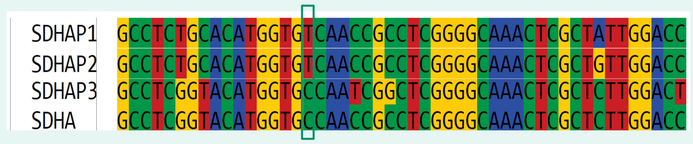


Figure 4: MSA showing the variant and the corresponding position in pseudogenes

Variant Caller with Local Realignment

Genome Analysis Toolkit (GATK) provides two variant calling tools, UnifiedGenotyper (UG) and HaplotypeCaller (HC). HC being the advanced variant caller performs an extra step of local realignment before making variant calls. The purpose of this step is to resolve any reads misaligned by the alignment tool, however, when reads arising from pseudogenes and wrongly align to functional genes, this step could lead to false variants.

Presence of Alternate Alignment

XA tags in SAM files provide alternative positions where reads align. SAM flags also provide information on primary and secondary alignments. Variants called from reads which align to multiple locations tend to be false positives.

Sanger Confirmation using Sequence Specific Primers

Multiple Sequence Alignment (MSA) on *SDHA* and its pseudogenes showed positions in *SDHA* which were homologous and discordant with the three pseudogenes. We used this information to design primers from the regions of discordance which distinguished *SDHA* from the pseudogenes and performed Sanger sequencing.

Figure 5: MSA showing discordant region between *SDHA* and pseudogenes are 576 bases upsteam and 197 bases downstream of the variant

CONCLUSION

Our results confirmed that the “variants” were indeed false positive calls. Variant that cannot be identified computationally, for example, private variants, we recommend to use sequence specific Sanger to identify the true source of the variants. This approach is not unique to *SDHA* and is promising for all pseudogenes.

This poster presents independent research commissioned by the Health Innovation Challenge Fund (HICF-R6-373), a parallel funding partnership between the Department of Health and Wellcome Trust. The views expressed in this work are those of the authors and not necessarily those of the Department of Health or Wellcome Trust