Using high-resolution variant frequencies to empower clinical genome interpretation

### AUTHORS (to include)

Nicola Whiffin, Roddy Walsh, Elizabeth Edwards, Risha Govind, Mian Ahmad, Xiaolei Zhang, Rachel Buchan, Paul Barton, James Ware and Stuart Cook

## ABSTRACT

Recent guidelines from the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) aim to standardise interpretation of sequence variants. Whilst these guidelines outline the key evidence types that need to be considered, there is a need to define and incorporate disease specific knowledge and thresholds. We have created Cardio Classifier, a powerful tool that aids variant interpretation across a range of inherited cardiac conditions. We use internally derived data along with an extensive collection of publically available datasets and disease/gene-specific knowledge, to pre-populate the criteria from the ACMG/AMP guidelines and output both a predicted classification and the associated evidence. This classification can be interactively refined though addition of clinical data or upload of additional user-defined annotations.

We have tested Cardio Classifier on a panel of 2,081 samples sequenced with the Illumina TruSight Cardio sequencing kit, and show that the annotations and classifications are specific to the disease of interest. Furthermore, using data from ClinVar we demonstrate that the tool successfully identifies true pathogenic variants. Finally, we report that 0.20% and 0.48% of individuals in the ExAC dataset have clinically reportable variants for hypertrophic and dilated cardiomyopathy respectively. We believe that Cardio Classifier represents a unique solution to the challenge of variant interpretation and that it will greatly enhance both research and clinical testing of inherited cardiac conditions. Cardio Classifier is publically available at www.cardioclassifier.org.

## INTRODUCTION

Inherited cardiac conditions (ICCs) represent a major worldwide health burden with a combined prevalence of ~1%. Routine genetic testing for ICCs is widespread and can be critical in confirming a diagnosis and/or dictating clinical management of patients and their relatives.

We recently created a comprehensive targeted sequencing panel of 174 genes associated with ICCs [ref] in a bid to increase the accessibility of clinical grade sequencing for these disorders. The major challenge in genetic testing across all domains, however, is the interpretation of the identified sequence variants and distinguishing the true pathogenic variant(s) from those that are benign bystanders.

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) recently released a set of guidelines that aim to standardise variant interpretation1. These rules are deliberately broad, to allow adoption across the full spectrum of genetic disorders; the challenge to individual disease domains, therefore, is to incorporate both gene and disease specific knowledge, to strengthen variant interpretation. Publication of the guidelines has led to the emergence of interactive tools that aim to facilitate their adoption [refs], however, to date none of these are automated or designed to build in disease specific knowledge.

We describe Cardio Classifier, a novel tool that utilises the framework outlined by the ACMG guidelines, to automatically classify variants for ICCs. This classification incorporates both disease specific and population data and allows for further refinement through user addition of clinical data.

## METHODS

#### Tool implementation

Cardio Classifier is implemented server-side in perl and PHP. Users upload variant data in the form of a single sample VCF file or individual variant details, which are then annotated by the Ensembl variant effect predictor (VEP)[ref] and converted to a table using the tableize\_vcf.py script within LOFTEE (<https://github.com/konradjk/loftee>). Protein altering and splice site variants (coding ±8bps) are analysed for a set of 50 genes associated with inherited cardiac conditions. The full list of definitively associated gene-disease pairs currently analysed can be found in Table 1. We look to expand this list in the future to include the full set of 174 genes available on the Illumina TruSight Cardio sequencing panel [ref].

The classifier automatically assesses each variant for 14 rules across three distinct data categories, as defined by the ACMG/AMP guidelines [ref]. The output is displayed on a PHP webpage that allows the user to interact and add (or remove) additional levels of evidence.

### Rule parameterisation

Full details of how each rule is parameterised can be found in the Supplement to this manuscript.

In addition to the rules outlined in the ACMG/AMP guidelines, we have created three further rules that we believe aid and/or enhance the interpretation of ICC variants. We have coded these as follows:

*PS5 – TTNtv in constitutively expressed exon* Previous studies have shown that truncating mutations in TTN are highly likely to be pathogenic when identified in exons that are constitutively expressed in the heart [refs]. We use previously described RNA-seq data to identify regions with a proportion sliced in (PSI) > 90% [ref] and activate this rule for truncating variants in these regions.

*PP6 – Missense change at equivalent amino residue of a paralogous gene is reported as pathogenic* *PM7 – Equivalent amino change in a paralogous gene is reported as pathogenic* For genes with paralogues that are known to be involved in disease, mutations at equivalent residues in these paralogues can identify residues that are likely intolerant to variation and hence may be more likely to harbour disease causing variants [ref]. We use previously published data [ref] to activate two rules relating to mutations at equivalent residues and the same amino acid change at these residues. Only mutations with high-confidence mapping to a paralogous gene are used (M-coffee mapping score >3 and the same reference amino acid). As this data relies on mutations ‘reported’ as pathogenic that have not been manually curated, the user can select whether to include these additional rules or not.

#### Test data

In order to fully test the classifier we utilised data from the following sources:

1. ExAC – all variants passing filters in the 0.3.1 (March 2016) release of the Exome Aggregation Consortium (ExAC) dataset
2. ClinVar – all variants identified as ‘Pathogenic’ or ‘Likely Pathogenic’ for Long QT syndrome (LQTS) with no conflicting data (i.e. reports of ‘Benign’ or ‘Likely Benign’) extracted from the X.X.X release of ClinVar using the MacArthur lab scripts on Github (<https://github.com/macarthur-lab/clinvar>)
3. An Internal dataset of 880 DCM, 327 HCM and 874 healthy volunteer samples recruited to the NIHR Royal Brompton cardiovascular BRU and all confirmed by cardiac MRI. All samples were sequenced using the IlluminaTruSight Cardio Sequencing Kit [ref] on the Illumina NextSeq platform. This study had ethical approval (REC: 09/H0504/104+5) and informed consent was obtained for all subjects
4. A set of 57 protein-altering MYH7 variants from the ClinGen pilot study [ref]

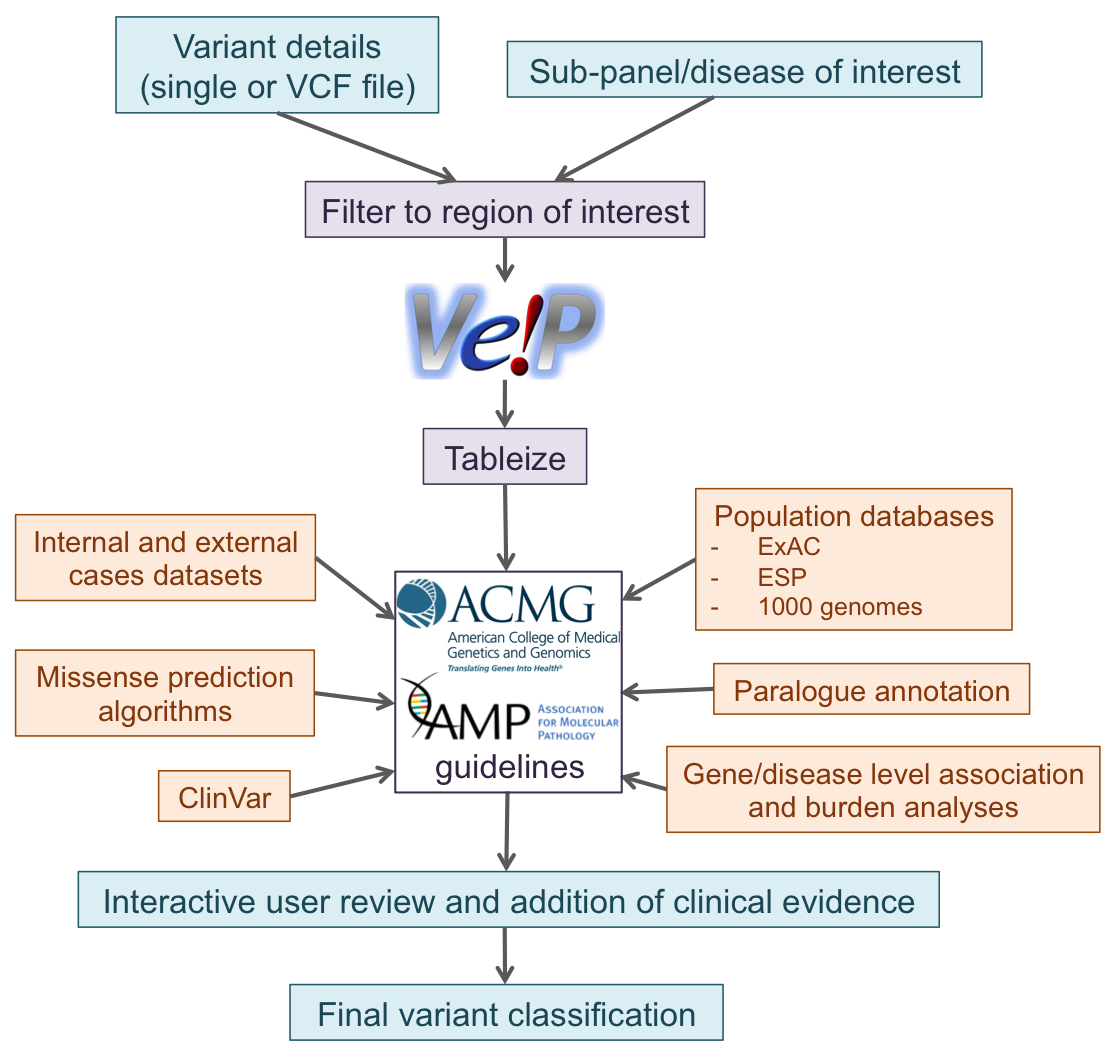
## RESULTS

#### Key features of Cardio Classifier aid clinical interpretation of ICC variants

Cardio Classifier takes as input either a single sample VCF file (variant call file) from the end of a sequencing analysis pipeline or individual variant details. Interpretation is performed over a set of 50 genes definitively associated with inherited cardiac conditions (ICCs; Table 1). We pre-populate a grid representing the ACMG/AMP framework with computationally predicted and defined data (see Methods). This data is disease, gene and variant-type specific. It is important to note that it is only possible to computationally predict a sub-portion of the ACMG/AMP rules and that manual addition of clinical data is crucial in variant interpretation. The user is therefore able to interactively add and remove clinical data to refine the classification of each variant. It is important to keep this in mind when interpreting the results presented in this manuscript, where we often expect the classifier to output a classification one level lower than those that are manually curated.

### Figure 1

Flow diagram



### Table 1

Details of gene-disease relationships

Table continues below

|  |  |
| --- | --- |
| Disease | DiseaseClass |
| DCM | Cardiomyopathy |
| HCM | Cardiomyopathy |
| ARVD/C | Cardiomyopathy |
| RCM | Cardiomyopathy |
| ncCM | Cardiomyopathy |
| Noonan syndrome | Cardiomyopathy |
| LQTS | Arrhythmia |
| Brugada syndrome | Arrhythmia |
| CPVT | Arrhythmia |
| Marfans syndrome | Aortopathy |
| FH | - |
| Heterotaxy | - |
| Alagille syndrome | - |
| Carney Complex | - |
| Holt-Oram syndrome | - |
| Duane radial ray syndrome | - |
| Ulnar-Mammary syndrome | - |

Table continues below

|  |
| --- |
| Genes |
| LMNA,TNNT2,SCN5A,TTN,TCAP,MYH7,VCL,TPM1,TNNC1,RBM20,DSP,BAG3 |
| MYH7,TNNT2,TPM1,MYBPC3,PRKAG2,TNNI3,MYL3,MYL2,ACTC1,CSRP3,PLN,TNNC1,GLA,FHL1,LAMP2,GAA,TTR |
| DSP,PKP2,DSG2,DSC2,JUP |
| TNNI3 |
| MYBPC3,MYH7 |
| RAF1,SOS1,PTPN11,KRAS |
| KCNQ1,KCNH2,SCN5A,KCNE1,KCNE2 |
| SCN5A |
| RYR2 |
| FBN1 |
| LDLR |
| ZIC3,CRELD1,NKX2-5,NODAL |
| JAG1 |
| PRKAR1A |
| TBX5 |
| SALL4 |
| TBX3 |

|  |
| --- |
| TotalGenes |
| 12 |
| 17 |
| 5 |
| 1 |
| 2 |
| 4 |
| 5 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |

We have added additional levels of interactivity to the tool including links out to external webpages commonly used in variant interpretation for ICCs. These include the ExAC browser [ref], ClinVar [ref], the Beacon Network [ref] and the Atlas of Cardiac Genetic Variation [ref].

As our full list of 50 genes represents a range of ICCs, when a diagnosis has already been confirmed for a patient it may be more appropriate to restrict analysis to a subset of genes relating to the condition of interest. We have therefore defined a set of 20 sub-panels for common and rare ICCs (Table ?) that can be selected for analysis.

#### Paralogue annotation enhances the interpretation in specific arrhythmia variants

In addition to the standard ACMG/AMP guidelines, we have developed three additional rules that we believe enhance the interpretation of specific ICC genes. The first is specific to the Titin (TTN) gene, which has a role in up to 20% of dilated cardiomyopathy (DCM) cases [ref]. We have shown in previous work that DCM causing TTN truncating variants (TTNtv) are exclusively found in exons that are constitutively expressed in the heart (proportion spliced in (PSI) > 0.9) [ref]. We have therefore added an additional strong piece of evidence for TTNtvs that are found in these highly expressed exons (rule PS5).

The other two new lines of evidence utilise known disease-causing mutations in related genes/proteins (paralogues), which likely identify residues intolerant to variation. This approach has been described previously [ref]. We use high confidence mutations as evidence if they either affect the equivalent residue in a paralogue (rule PP6), or represent the equivalent amino acid change (i.e. the same alternate amino acid; rule PM7). This analysis is currently restricted to the arrhythmia genes described in this previous work [ref].

To demonstrate the utility of our paralogue annotation rules we used variants reported as ‘Pathogenic’ or ‘Likely Pathogenic’ for Long QT syndrome (LQTS) in the ClinVar database. Of 144 identified variants, 51 were output as ‘Likely Pathogenic’ and 8 as ‘Pathogenic’ using the ACMG/AMP rules alone. On inclusion of the paralogue annotation rules, 15 variants were reclassified from ‘Uncertain Significance’ to ‘Likely Pathogenic’ and 3 from ‘Likely Pathogenic’ to ‘Pathogenic’ (Table 2). This represents a 25.4% increase in variants that would be considered clinically actionable. Additionally, PP6 and PM7 were never activated for a set of 19 variants annotated as ‘Benign’ for LQTS in ClinVar.

#### Results show high concordance with manually curated data

As a gold-standard manually-curated dataset with which to compare our tool we used 57 protein-altering variants from the recent ClinGen pilot for MYH7 in hypertrophic cardiomyopathy. Overall we observed very good concordance across rules where Cardio Classifier makes a computational prediction, with the output of these rules matching between Cardio Classifier and the ClinGen pilot on X/X occasions (Table X/Figure X). Furthermore, after manually curating these variants using the ClinVar and Google links and references found in the classifier output, we reached an identical conclusion for XX% of variants.

#### Classifications are disease specific

To assess the performance of our tool on large cohorts of cases and controls we used internal data from a set of 2,081 samples comprising 880 DCM, 327 HCM and 874 healthy volunteers (HVOLs). We ran all variants found in these samples for both a HCM and DCM test. For both analyses, ‘Pathogenic’ and ‘Likely Pathogenic’ variants for the test disease were almost exclusively found in the samples corresponding to that disease (Figure X). In addition, the VUS category is enriched in the test disease samples over the other cohorts. With manually added clinical data, a proportion of these variants would become clinically actionable.

#### Pathogenic mutations in the general population do not exceed expected rates

In order to investigate the burden of pathogenic mutations in the general population, and ensure we are not overcalling variants as pathogenic, we used the ExAC dataset of 60,706 individuals. We ran all variants identified in this cohort through the classifier for both a HCM and DCM test.

For DCM, we identified 293 Likely Pathogenic variants in 374/60,706 (0.48%) individuals (assuming all individuals have only one DCM variant). Although this proportion is slightly above the reported prevalence of DCM of 1 in 250 (0.40%) [ref], the vast majority of these, 264/374, are truncating mutations in TTN, which are known to be present at an appreciable frequency in the general population. Overall we find ‘Likely Pathogenic’ mutations in TTN in 0.43% of ExAC individuals.

For HCM, we identify 4 Pathogenic and 71 Likely Pathogenic variants in 6 and 116 individuals respectively. In total, this corresponds to 0.20% (122/60,706) of ExAC individuals with a clinically actionable variant for HCM, a frequency that exactly matches the reported penetrance of the condition [ref]. Although we might expect this proportion to be lower, as the classifier does not include clinical data, this calculation does not take into account the reduced penetrance of many mutations involved in HCM.

#### Manual curation of known variants

We have manually curated a set of X variants to include clinical and functional data in Cardio Classifier. These comprise the most commonly observed variants in published cohorts for HCM, DCM and ARVC (i.e. all those that occur more than X times in the combined dataset) [refs], the 57 protein altering mutations from the ClinGen pilot MYH7 analysis described above and the 144 variants identified as ‘Pathogenic’ in ClinVar for LQTS. Overall, this dataset comprises X ‘Pathogenic’, X ‘Likely Pathogenic’ and X ‘VUS’ variants after curation (Table X).

## DISCUSSION

We have described Cardio Classifier, an automated and interactive tool to aid clinical variant interpretation and classification across a wide range of inherited cardiac conditions. To the best of our knowledge our tool represents a unique solution that incorporates data from the literature with in-house knowledge and analysis to pre-calculate variant classifications in line with the ACMG/AMP guidelines. The tool is transparent, with all the information incorporated into interpreting each variant displayed along with the final classification. It is also flexible, and designed to be fully interactive, with the user able to add and remove evidence specific to the patient/family of interest.

The true strength of our tool is in its disease specificity. We have used our wide knowledge of ICCs to define criteria and thresholds for each ACMG/AMP rule that are specific to the disorder of interest. We believe this adds huge power to variant interpretation and are able to demonstrate the effectiveness of this approach using both internal and external datasets. Incorporation of disease specific knowledge is, however, limited by our current understanding and the availability of disease datasets on which to perform analyses. Consequently, the power of our tool will continue to increase over time as these data emerge across the full compliment of ICCs.

We believe the main limitation to the effectiveness of our tool is the inability to computationally incorporate clinical or patient specific data into the initial classification, leading to us under-call variants as pathogenic. We hope, however, that the combination of pre-populated computational data with the ability to interactively add additional levels of evidence will help to overcome this hurdle.

Cardio Classifier is designed to work seamlessly, but not exclusively, with the Illumina TruSight Cardio sequencing kit [ref], and takes as input the final output VCF from a standard bioinformatics analysis pipeline. We believe the combination of these two powerful tools is a crucial step in broadening the availability of genetic testing for ICCs, and standardising variant interpretation in this field.

## TOOL AVAILABILITY

Cardio Classifier is freely available for non-commercial use at www.cardioclassifier.org.

## ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust (107469/Z/15/Z), the Medical Research Council (UK), the NIHR Biomedical Research Unit in Cardiovascular Disease at Royal Brompton & Harefield NHS Foundation Trust and Imperial College London, the Fondation Leducq (11 CVD-01) and a Health Innovation Challenge Fund award from the Wellcome Trust and Department of Health, UK (HICF-R6-373).

This publication includes independent research commissioned by the Health Innovation Challenge Fund (HICF), 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.

## SUPPLEMENTARY INFORMATION

## REFERENCES

1. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine* **17,** 405–423 (2015).