



SHORT REPORT

DNA-pools targeted-sequencing as a robust cost-effective method to detect rare variants: Application to dilated cardiomyopathy genetic diagnosis

Claire Perret^{1,2} | Carole Proust¹ | Ulrike Esslinger¹ | Flavie Ader^{1,2,3}  |
Jan Haas^{4,5} | Jean-François Prunty⁶ | Richard Isnard^{1,2,7} | Pascale Richard^{1,2,3}  |
David-Alexandre Trégouët¹ | Philippe Charron^{1,2,6,7} | François Cambien¹ |
Eric Villard^{1,2}

¹Sorbonne Université, INSERM, UMR-S1166, Research Unit on Cardiovascular and Metabolic Diseases, Paris, France

²ICAN Institute for Cardiometabolism and Nutrition, Paris, France

³APHP, UF Cardiogénétique et Myogénétique, Service de Biochimie Métabolique, Hôpital Universitaire Pitié-Salpêtrière, Paris, France

⁴Department of Internal Medicine III, University of Heidelberg, Heidelberg, Germany

⁵DZHK (German Centre for Cardiovascular Research), Berlin, Germany

⁶APHP, Centre de Référence Maladies Cardiaques Héritaires, Hôpital Pitié-Salpêtrière, Paris, France

⁷APHP, Cardiology Department, Pitié-Salpêtrière Hospital, Paris, France

Correspondence

Eric Villard, Sorbonne Université, INSERM, UMR-S1166, Research Unit on Cardiovascular and Metabolic Diseases, Paris 75013, France.
Email: eric.villard@sorbonne-universite.fr

Present addresses

Carole Proust and David-Alexandre Trégouët, Univ. Bordeaux, INSERM, BPH, U1219, Bordeaux, 33000, France.

Funding information

The Conny-Maeva Foundation; Aviesan-ITMO Genetique-Genomique-Bioinformatique; ResDiCard AAP 2020; Fondation Leducq “Genomic, epigenomic and systems dissection of mechanisms underlying DCM”

Abstract

Dilated cardiomyopathy (DCM) is a heart disease characterized by left ventricular dilatation and systolic dysfunction. In 30% of cases, pathogenic variants, essentially private to each patient, are identified in at least one of almost 50 reported genes. Thus, while costly, exons capture-based Next Generation Sequencing (NGS) of a targeted gene panel appears as the best strategy to genetically diagnose DCM. Here, we report a NGS strategy applied to pools of 8 DNAs from DCM patients and validate its robustness for rare variants detection at 4-fold reduced cost. Our pipeline uses Freebayes to detect variants with the expected 1/16 allele frequency. From the whole set of detected rare variants in 96 pools we set the variants quality parameters optimizing true positives calling. When compared to simplex DNA sequencing in a shared subset of 50 DNAs, 96% of SNVs/InsDel were accurately identified in pools. Extended to the 384 DNAs included in the study, we detected 100 variants (ACMG class 4 and 5), mostly in well-known morbid gene causing DCM such as TTN, MYH7, FLNC, and TNNT2. To conclude, we report an original pool-sequencing NGS method accurately detecting rare variants. This innovative approach is cost-effective for genetic diagnostic in rare diseases.

KEYWORDS

cardiomyopathy, DNAs-pool, genetic diagnosis, Next Generation Sequencing

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Clinical Genetics* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Dilated cardiomyopathy (DCM) is a heart disease characterized by left ventricular dilatation and systolic dysfunction.¹ It is an important cause of systolic heart failure, and is the first indication for heart transplantation.¹ Pathogenic variants, essentially private to each patient, are identified in about 30% of cases, in more than 50 genes, indicating strong genetic heterogeneity.¹ Then, the most efficient technology to perform genetic diagnosis of DCM is targeted sequencing on a large panel of genes, however it remains expensive while molecular diagnosis prescription increases. In this context, combinatorial pool-DNA sequencing (Pool-Seq) could be a time- and cost-effective approach.² In addition, it has been reported that two-dimensional Pool-Seq allows for the identification of rare variants with determination of carrier DNA, after decoding of the pools.³

In the present report, we performed such a two-dimensional Pool-Seq strategy to sequence 384 DNAs from DCM patients and developed a specific data analysis method for identification of unique variants. We reported the resulting variant atlas in a large panel of 109 genes previously associated with cardiomyopathies and cardiac arrhythmias. In addition, we evaluated the efficiency of the strategy and demonstrated a clear cost and time advantage of Pool-Seq over NGS based simplex DNA capture with minimal loss of sensitivity or specificity for rare variants detection, showing that Pool-Seq can be used for routine genetic diagnosis in rare diseases.

2 | MATERIALS AND METHODS

See [supplementary file](#) for details.

3 | RESULTS

3.1 | Sequencing metrics

We designed the experiment to over-cover each haplome in the pools with 25 reads, that is, 400 reads (25×16) per target base. The average coverage observed was slightly higher with $496\times$ and despite 25% of the reads being off-target. For 90.4% of captured bases, coverage was at least $128\times$ corresponding to an average haplome coverage of $8\times$ allowing at least one coverage for each haplome ($p = 0.049$) (Table S1).

3.2 | Cost-cutting evaluation

Pool-Seq uses only 96 libraries and capture reactions to process 384 DNAs. Given the additional time and resources required for accurate DNA quantitation and pooling, we estimate that saving reagents and labor time reduces the cost of the

protocol threefold. A similar 4-fold saving is obtained while sequencing since we reached $\sim 500\times$ coverage in average, which is a depth similar to those used in simplex DNAs NGS protocols.

3.3 | Quality filtering of unique variants

Since the known pathogenic variants are essentially private, we selected only variants called in a single DNA concordant pair of pools (unique variant). To maximize true variant identification in pools, we called all variants with at least one mutated read in two concordant pools. Importantly, these variants could be false positives due to sequencing errors, especially in low coverage regions. To isolate true variants from false positives, we take advantage of the Freebayes QUAL score and compared the distribution of the best-of-2-pools QUAL value (mQUAL) in concordant pairs versus in non-concordant pairs, characterizing false positives (Figure 1). The variants identified only in concordant pairs display a $\log(\text{mQUAL}) > 1.5$. This threshold, corresponding to $\text{mQUAL} > 32$, might allow to maximize true positive while decreasing calling false-positive variants.

To refine this mQUAL threshold, we amplified by PCR and re-sequenced a set of candidate variants ($n = 54$) called from unique concordant pairs of pools (Table S2) with mQUAL values in the range [0.1–50]. From Sanger sequencing results, true ($n = 13$) or false-positive ($n = 41$) status was assigned to each variant and each was plotted according to mQUAL and mDPAlt (i.e., the number of reads carrying the alternate allele compare to reference genome) on Figure 2. All but one false positive (1/41; 2.4%) are excluded, and all true-positives are retained (13/13) when applying the following thresholds: (i) $\text{mQUAL} > 12$ and (ii) $\text{mDPAlt} > 1$, suggesting an excellent discriminating potential of the filtering based on mQUAL and mDPAlt.

3.4 | Power-to-detect estimate of the pooling strategy

We compared Pool-Seq unique variant detection rate to a previously published NGS based sequencing of simplex DNA sharing 50 DNAs with our study⁴ (Figure S1). In these 50 DNAs, 319 unique variants were identified in simplex. In the pool sequencing results, after applying $\text{mQUAL} > 12$ and $\text{mDPAlt} > 1$ filters, 305 out of the 319 were identified in the expected concordant pairs. Twelve of the 14 missing SNVs were confirmed as false-negative after Sanger sequencing indicating a false positive rate of 0.63% (2/319) in simplex. Of the 12 false-negative variants in Pool-Seq (12/317; 3.8%), 9 were detected in only 1 of the 2 concordant pools (termed “partial false-negative pool”) and 3 not detected at all (Table S3).

Conversely, of 190 variants found only once in a single pair of concordant pools containing one of the 50 DNAs shared with Haas

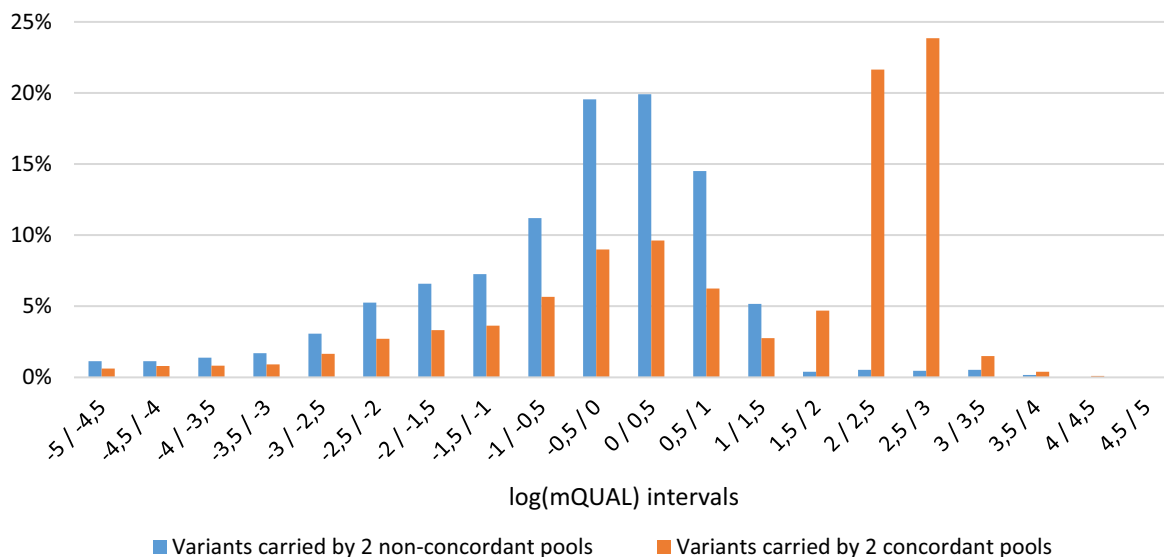


FIGURE 1 Distribution of the Freebayes mQUAL logarithmic value for all variants identified once in two concordant pools, expected to be true variants (orange bars) or two non-concordant pools, assumed to be false positives (blue bars). [Colour figure can be viewed at wileyonlinelibrary.com]

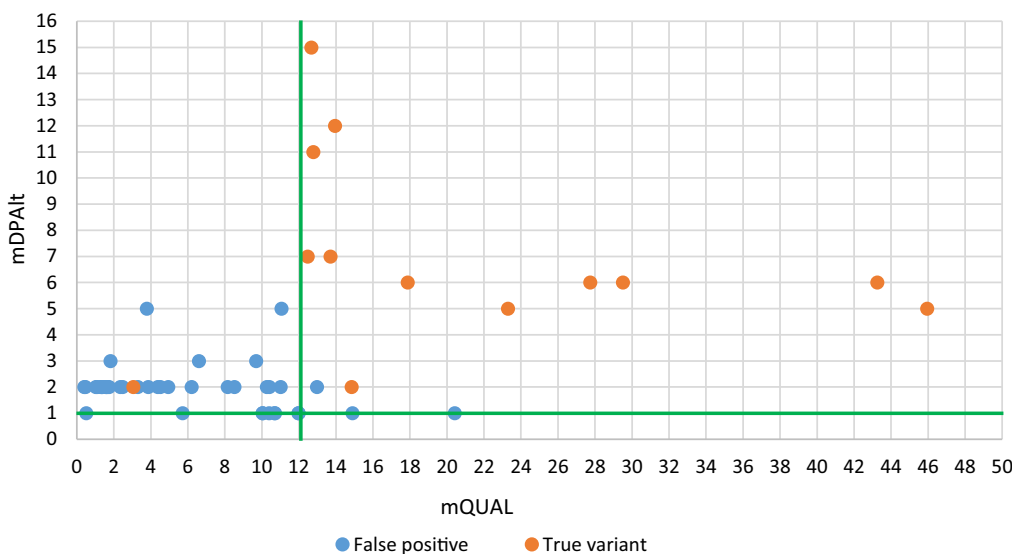


FIGURE 2 Random set of true or false variant replication status after Sanger sequencing. Variants detected in two concordant pools ($n = 56$) were randomly selected from those with a low mQUAL value (range [0.1–50]) and plotted for the best of two pools of QUAL value (mQUAL) on the abscissa and the best of two pools mutant allele coverage (mDPAIt) in the ordinate. Orange dots: True variants, validated by Sanger method. Blue dots: False positive variants, not validated by Sanger method. [Colour figure can be viewed at wileyonlinelibrary.com]

et al. study, 187 were also present after standard simplex capture NGS (98.9%), in the same DNA. Two were false negatives in simplex as they were confirmed after Sanger sequencing (false-negative rate = 1.1%). The other was a concordant pools false positive (0.5%) (Table S4).

We achieved an almost identical rare variant detection rate with both methods (96% vs. 99%) at similar coverage (Figure S2) indicating that the pooling strategy could reduce the sequencing depth and thus the sequencing cost by 4-fold.

3.5 | Variants calling in cardiomyopathy patient's DNAs

3.5.1 | Variants identification

From the pooled DNAs of 384 patients with standard criteria for DCM (see supporting information), we selected unique variants present in a single pair of pools after applying pre-defined QUAL and coverage thresholds (mQUAL < 12, mDPAIt > 1).

We have therefore identified a total of 1596 unique variants with expected 1/16 average allele frequency in exons or intron-exon boundaries in 107 genes for 383 samples (Table S5). After annotation, 102 variants of interest (ACMG class 4 or 5) were identified in 100 DCM patients and 36 genes. Among them, 22 (21.6%) were missense and 80 (78.4%) were truncating variants (34 frameshift, 11 splice-site, and 35 Stop codon) (Table S6). If only the 19 genes with strong evidence for association with DCM were considered, according to ClinGen consortium,⁵ 95 of this 102 variants are retained (Figure S3). Regarding the genes spectrum, class 4 and 5 variants are mainly TTN truncation, but also MYH7 missense, FLNC truncation and TNNT2 missense for the most frequent. Regarding the variants spectrum, 49 are already reported and 46 are new ones, absent in gnomAD or ClinVar database.

4 | DISCUSSION

In the present targeted-NGS strategy, we describe a simple and robust combinatorial 8-DNA pooling strategy to detect unique variants with high sensitivity and reduced cost.

Our objective was to select, in DNA-pools, all unique variants that could cause the disease, without increasing sequencing depth. To be able to detect them in low coverage region, we reduced the minimal mutated base coverage to 1×, with the cost of increasing sequencing error calling. Our original combinatorial pool strategy drastically limits the false-positives rate. Moreover, filtering optimization based on mQUAL and mDPAlt parameters allowed to limit false-positive rate to ~0.5% only.

We also calculated a very good true-positive detection rate of 96.2% in pool, with a mean coverage similar to which generally applied to genetic diagnosis (400×), constituting a strong improvement compare to previously published pool-sequencing.³ The 12 false-negative (3.8%), may be mainly explained by local low coverage. Indeed, 10 are under-covered compared to an established 15× necessary for simplex sequencing calling,⁶ indicating they would have been recovered with deeper sequencing, reaching 99.4% sensitivity identical to what calculated from standard simplex capture results reported in the present study. A very accurate quantitation of DNAs before and after pooling is mandatory to avoid allelic drop-out that could increase false negative rate.

We also demonstrated the combinatorial pooling method capacity to identify the DNA carrier in pools for unique variants. Nevertheless, this method can be used exclusively to detect unique variants because the carrier DNA determination can be ambiguous with more than 2 carrier pools. So, it is particularly suitable for genetic diagnosis in disease associated with private variants. Identification of hotspots, or familial variants, occurring more than once in the DNAs included in the pools, or variants validation would need optimization on pools constitution, filtering strategy and/or extra orthogonal validation using sequencing on an independent sample, generating extra costs.

Interestingly, our Pool-seq strategy allows to reduce significantly the cost of genetic diagnosis since the cost of library preparation and

target sequence capture were reduced by 3 and the sequencing cost by 4.

Considering only the 19 most confirmed DCM genes,⁵ they harbor 93% of ACMG class 4 and 5 variants in 25% of the included DNAs. Our variant detection rate is in the lower part of previously reported rates (20% and 35%),⁵ probably due to high number of sporadic cases in our study (75%) as previously reported.⁷ We observed 57% of variants being TTN truncating variant (TTNtv), nearly twice more than reported.⁸ This TTNtv enrichment could be explained by high rate of carriers with severe phenotype since 56% of them had heart transplantation.⁹

The majority of detected variations was in known, high-confident, DCM genes such as TTN, MYH7, FLNC, and TNNT2. However, we also identified class 4 and 5 variants in genes mainly associated with other cardiomyopathies such as HCM with MYBPC3tv ($n = 2$) or ARVC with PKP2tv ($n = 1$). This might be related to (i) uncertainty when defining the genes responsible for DCM, or (ii) in difficulties in phenotypic classification of some patients since end-stage phase of some cardiomyopathies may mimic DCM, especially when early conventional phenotypic phase was not recognized.¹⁰

To progress towards personalized medicine and optimal treatment of cardiomyopathies and other genetic diseases, genetic diagnosis is increasingly requested by clinicians. Here we demonstrated that Pool-Sequencing is a cost- and time-effective NGS strategy for the diagnostic identification of rare variants. In addition, our study increases knowledge of the atlas of DCM gene variants.

4.1 | Limitations

These results are not accounting for diversity of human genetic backgrounds. Accordingly, refinement of the thresholding strategy with more ancestry-diverse and larger cohorts will be required to validate the selected parameters. CNVs were not searched for since it is a rarely reported cause for cardiomyopathy and they are still not optimally detected by capture-based NGS strategy.

AUTHOR CONTRIBUTIONS

Conceptualization: C. Perret; F. Cambien; E. Villard. **Funding acquisition:** R. Isnard; P. Charron; F. Cambien; E. Villard. **Data production:** C. Perret; C. Proust; U. Esslinger; J. Haas; J. F. Prunty; D. A. Trégouët; F. Cambien. **Data analysis:** C. Perret; U. Esslinger; F. Ader; J. Haas; R. Isnard; P. Richard; D. A. Trégouët; P. Charron; F. Cambien; E. Villard. **Writing—original draft:** C. Perret; E. Villard. **Review & editing:** All Authors.

ACKNOWLEDGMENTS

Thanks to Nadjim Chelgoum and Florian Thibord for contribution to bioinformatics.

CONFLICT OF INTEREST STATEMENT

Nothing to declare.

PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/cge.14427>.

DATA AVAILABILITY STATEMENT

The new pathogenic variants identified are available on ClinVar (SUB13444431). The sequencing data (FASTQ Files) are accessible on demand.

ORCID

Flavie Ader  <https://orcid.org/0000-0001-7891-3385>

Pascale Richard  <https://orcid.org/0000-0002-2390-3005>

REFERENCES

1. Pinto YM, Elliott PM, Arbustini E, et al. Proposal for a revised definition of dilated cardiomyopathy, hypokinetic non-dilated cardiomyopathy, and its implications for clinical practice: a position statement of the ESC working group on myocardial and pericardial diseases. *Eur Heart J*. 2016;37:1850-1858.
2. Cao C, Sun X. Combinatorial pooled sequencing: experiment design and decoding. *Quant Biol*. 2016;4:36-46.
3. Zuzarte PC, Denroche RE, Fehringer G, Katzov-Eckert H, Hung RJ, McPherson JD. A two-dimensional pooling strategy for rare variant detection on next-generation sequencing platforms. *PLoS One*. 2014; 9:e93455.
4. Haas J, Frese KS, Peil B, et al. Atlas of the clinical genetics of human dilated cardiomyopathy. *Eur Heart J*. 2015;36:1123-1135.

5. Jordan E, Peterson L, Ai T, et al. Evidence-based assessment of genes in dilated cardiomyopathy. *Circulation*. 2021;144:7-19.
6. Hartman P, Beckman K, Silverstein K, et al. Next generation sequencing for clinical diagnostics: five year experience of an academic laboratory. *Mol Genet Metab Rep*. 2019;19:100464.
7. Hershberger RE, Givertz MM, Ho CY, et al. Genetic evaluation of cardiomyopathy: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2018;20:899-909.
8. Herman DS, Lam L, Taylor MRG, et al. Truncations of titin causing dilated cardiomyopathy. *N Engl J Med*. 2012;366:619-628.
9. Roberts AM, Ware JS, Herman DS, et al. Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Sci Transl Med*. 2015;7:270ra6.
10. Stolfo D, Collini V, Sinagra G. Advanced heart failure in special population: cardiomyopathies and myocarditis. *Heart Fail Clin*. 2021;17(4): 661-672.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Perret C, Proust C, Esslinger U, et al. DNA-pools targeted-sequencing as a robust cost-effective method to detect rare variants: Application to dilated cardiomyopathy genetic diagnosis. *Clinical Genetics*. 2024; 105(2):185-189. doi:[10.1111/cge.14427](https://doi.org/10.1111/cge.14427)