PAV-spotter: using signal cross-correlations to identify Presence/Absence Variation in target capture data

Running head: Detecting Presence/Absence with PAV-spotter

Manon de Visser[1,2,3*], Chris van der Ploeg[4,5], Milena Cvijanović[6], Tijana Vučić[1,2,7], Anagnostis Theodoropoulos[1,2], Ben Wielstra[1,2]

[1]Institute of Biology Leiden, Faculty of Science, Leiden University, Leiden, The Netherlands

[2]Naturalis Biodiversity Center, Leiden, The Netherlands

[3]Blijdorp Conservation & Science Center, Royal Rotterdam Zoological & Botanical Gardens, Rotterdam, The Netherlands

[4]Dynamics and Control Group, Department of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

[5]Integrated Vehicle Safety Group, Netherlands Organisation for Applied Scientific Research, Helmond, The Netherlands

[6]Institute for Biological Research "Siniša Stanković", National Institute of the Republic of Serbia, Department of Evolutionary Biology, University of Belgrade, Belgrade, Serbia

[7]Institute of Zoology, Faculty of Biology, University of Belgrade, Belgrade, Serbia

*** Correspondence:**

Manon de Visser

devisser.manon@gmail.com

Keywords: Hyb-Seq, PA variation, target enrichment, copy number variation, pattern recognition, *Triturus*. (Max. 6)

Abstract

High throughput sequencing technologies have become essential in the fields of evolutionary biology and genomics. When dealing with non-model organisms or genomic gigantism, sequencing whole genomes is still relatively costly and therefore reduced-genome representations are frequently obtained, for instance by 'target capture' approaches. While computational tools exist that can handle target capture data and identify small-scale variants such as single nucleotide polymorphisms and micro-indels, options to identify large scale structural variants are limited. To meet this need, we introduce PAV-spotter: a tool that can identify presence/absence variation (PAV) in target capture data. PAV-spotter conducts a signal cross-correlation calculation, in which the distribution of read counts per target between samples of different *a priori* defined classes – e.g. male versus female, or diseased versus healthy – are compared. We apply and test our methodology by studying *Triturus* newts: salamanders with gigantic genomes that currently lack an annotated reference genome. *Triturus* newts suffer from a hereditary disease that kills half their offspring during embryogenesis. We compare the target capture data of two different types of diseased embryos, characterized by unique deletions, with those of healthy embryos. Our findings show that PAV-spotter helps to expose such structural variants, even in the face of medium to low sequencing coverage levels, low sample sizes, and background noise due to mis-mapped reads. PAV-spotter can be used to study the structural variation underlying supergene systems in the absence of whole genome assemblies. The code, including further explanation, is available through the PAV-spotter GitHub repository: https://github.com/Wielstra-Lab/PAVspotter.

Introduction

Next-generation DNA sequencing methods have revolutionized the biological sciences, with an ever-growing amount of sequence data being generated worldwide (Annavi et al, 2019; Morganti et al, 2019). High throughput sequencing techniques have become more affordable and increasingly used, however sequencing entire genomes can still be challenging, for instance when dealing with non-model organisms, genomic gigantism, or a combination of the two (Ekblom & Galindo, 2011; Lemmon &

54 Lemmon, 2013; Wachi et al, 2018). In such cases, well-annotated reference genomes for aligning (re-
55)sequenced reads are generally unavailable, making whole genome sequencing relatively costly in
56 terms of money and (computational) time (e.g. see Etherington et al, 2020; Rovelli et al, 2019; Zaharias
57 et al, 2020). Many biologists therefore still opt for more cheap and efficient 'reduced-representation'
58 high throughput sequencing techniques, which allow for a subset of loci to be sequenced more deeply
59 (Andermann et al, 2019; Da Fonseca et al, 2016; Puritz & Lotterhos, 2018).

60 A technique that has become particularly popular for studying non-model species is target
61 capture, also referred to as hybridization sequencing, exon capture, sequence capture, or exome capture
62 (Bi et al, 2012; Kaur & Gaikwad, 2017; Puritz & Lotterhos, 2018; Yohe et al, 2020). This method
63 facilitates collecting sequence information on hundreds or thousands of pre-selected loci. Due to the
64 consequent rise of large multi-locus DNA sequence datasets, the need for innovative, easy-to-
65 implement bioinformatic applications has surged as well (Gauthier et al, 2019). User-friendly pipelines
66 help to pre-process sequence reads by wrapping and connecting existing software tools, with well-
67 known examples for target capture including HybPiper (Johnson et al, 2016), Assexon (Yuan et al,
68 2019), Sequence Capture Processor (i.e. SECAPR, see Andermann et al, 2018), and CAPTUS (Ortiz
69 et al., 2023). These pipelines and their software dependencies support fast upstream data cleaning and
70 guide the user to phylogenetic applications downstream.

71 Typically, target capture analysis pipelines are utilized to identify small-scale genetic variants
72 such as single nucleotide polymorphisms (SNPs) and relatively small insertions/deletions (microindels,
73 ranging from 1~50bp) to perform, for instance, phylogenetic tree-building (Andermann et al, 2019).
74 However, to identify and analyze larger scale information from target capture data, such as genomic
75 structural variation, few tools are available – especially when focusing on non-model organisms. A
76 particular type of larger-scale variation that is hard to identify in multi-locus DNA sequence datasets
77 is presence/absence variation (hereafter referred to as 'PAV') of relatively big insertions/deletions
78 (macroindels, > 50bp), i.e. above the size of microindels (Wang et al, 2014; Zhang et al, 2014).

79 Structural variants such as PAVs are regularly overlooked because they are harder to identify
80 than SNPs (Mahmoud et al, 2019; Wold et al, 2021). However they are a major source of genetic
81 divergence and diversity (Gerdol et al, 2020; Hu et al, 2022; Marroni et al, 2014; Rosa et al, 2015).
82 PAV in particular poses an extreme example of copy number variation, where fragments in the size of
83 entire exons or (stretches of DNA containing multiple) genes are missing from one genome compared
84 to another (Gabur et al, 2020; Wellenreuther et al, 2019). When comparing such genomes, target

85 capture data would in theory display PAV by showing 'normal data' in the case of target presence,
86 versus a 'data gap' in the case of target absence. This would be an indication of structural variation.

87 Whether there are consistent differences in presence or absence of sequence data can be
88 determined by analyzing the way that reads pile up against a certain reference set of sequences. Some
89 tools can detect copy number variation and PAV patterns by comparing the depth of mapped reads of
90 different samples, such as ExomeCNV (Sathirapongsasuti et al, 2011) and SUPER-CAP (Yuan et al,
91 2019). However, these tools come with strict requirements, such as good quality reference genomes
92 being available, known functional annotation of variants, and/or coverage being consistently high
93 across all samples and targets, with mapped reads ideally following a normal distribution. Yet, most
94 multi-locus datasets, including those resulting from target capture experiments, generally do not meet
95 such requirements (Andermann et al, 2019).

96 We introduce PAV-spotter: a flexible signal cross-correlation method that is able to 'spot'
97 potential PAV in target capture datasets. Our approach borrows the notion of cross-correlation to detect
98 the dissimilarity between datasets obtained through target capture experiments. Cross-correlation
99 methods are generally used in the domain of control engineering. Classically, they are applied on time-
100 series data (Verhaegen & Verdult, 2007), for example in machinery fault detection studies (Gao et al,
101 2015). However, cross-correlation approaches have been proven useful in the field of pattern
102 recognition as well (Jain et al, 2000).

103 Being able to identify structural variation by using pattern recognition would be especially
104 informative when studying supergenes systems, in which individuals can have zero, one or two copies
105 of particular loci (Hall & Wayne, 2013; Thompson & Jiggins, 2014). Supergenes consist of genes that
106 are inherited together as a single locus due to the suppression of recombination (Berdan et al, 2022;
107 Gutierrez-Valencia et al, 2021; Thompson & Jiggins, 2014). As a result, the non-recombining stretches
108 of 'supergene DNA' evolve independently of one another, facilitating the rapid evolution of complex
109 adaptations (Pennisi, 2017; Schwander et al, 2014). These sets of genes are often polymorphic and
110 subjected to balancing selection, as a species generally possesses at least two supergene variants
111 (Llaurens et al, 2017).

112 Sex chromosomes, for instance, are classically considered supergenes (Joron et al, 2006). In
113 diploid organisms the heterogametic sex inherits the sex-determining 'supergene', as well as the
114 alternate sex chromosome, in a hemizygous manner – meaning that they only receive one copy of each

115 (Dufresnes & Crochet, 2022). In the XX-XY sex determination system of mammals, for instance, males
116 generally possess a single copy of the supergene that is the Y chromosome (as well as a single copy of
117 the X-chromosome), whereas in the ZW-WW system it is the females that possess the Z supergene
118 once (next to a single W chromosome). Hence, genes that are hemizygous and thus lie solely on the
119 sex-determining supergene would show PAV in target capture datasets when data of different sexes is
120 compared.

121 However, supergene systems are not limited to the biological concept of sex. Other, famous
122 examples of supergenes underlying complex traits are; the Müllerian mimicry complex in *Numata*
123 longwing butterflies (Joron et al, 2006), the striking sexual dimorphism and breeding behaviors of ruffs
124 (Kupper et al, 2016; Lamichhaney et al, 2016) and white-throated sparrows (Tuttle et al, 2016), the
125 social polymorphism observed in several species of ant (Kay et al, 2022), and heterostyly in primrose
126 flowers (Li et al, 2016). Furthermore, hemizygous inheritance of (super)genes also occurs in, for
127 instance, genetic incompatibilities such as with "hybrid necrosis" in plants, which can be linked to
128 PAV in certain genes in for example Asian rice (Li et al, 2023; Li & Lee, 2023); hereditary diseases
129 such as α-thalassaemia, which is caused by large deletions in the alpha globin gene cluster on
130 chromosome 16 in humans (Harteveld & Higgs, 2010); and in balanced lethal systems, in which two
131 distinct chromosome forms exist that are covered by unique lethal mutations (Wielstra, 2020).

132 We demonstrate the application of PAV-spotter using the balanced lethal system in *Triturus*
133 newts as a case study. *Triturus* individuals either are heteromorphic and possess two different versions
134 of their largest autosomal chromosome, characterized by unique deletions, or they are homomorphic
135 and possess two identical versions of this chromosome (De Visser et al, 2024). The two types of
136 homomorphic individuals express a unique disease state and both die during embryogenesis, whereas
137 the heteromorphic individuals are viable – a phenomenon caused by evidential PAV (De Visser et al,
138 2024; France et al, 2024; Macgregor & Horner, 1980; Meilink et al, 2025; Sessions et al, 1988;
139 Wallace, 1987). This 'double hemizygous' system lends itself particularly well for using target capture
140 data to detect PAV, as it allows for a reciprocal test: targets deleted from one chromosome version
141 should be present on the alternate version and the other way around. Based on our findings, we describe
142 the usefulness, as well as the limitations, of our approach.

143

Methods

Sample information & collection

The first chromosome of *Triturus* comes in two forms: 1A and 1B. Homomorphic individuals (1A1A or 1B1B) invariably die during embryogenesis, while heteromorphic individuals (1A1B/1B1A) are viable. We collected *T. macedonicus* x *T. ivanbureschi* F_1 hybrid embryos from an ongoing breeding experiment at the Institute for biological research, „Siniša Stanković", University of Belgrade (with experimental settings, breeding conditions, and other details on the process of raising embryos as described in Vucic et al, 2024; Vučić et al, 2022). Embryo development was followed through observation with a stereomicroscope. Diseased embryos were collected when the process leading up to developmental arrest occurred, which is visible as a 'growth slowdown', during the late tail-bud phase. Diseased embryos were then classified into either the "fat-tailed" (FT) phenotype or the "slim-tailed" (ST) phenotype based on morphological characteristics of the embryo (Sessions et al, 1988; Sims et al, 1984). Healthy/control (HC) embryos that survived this critical phase were subsequently collected. We collected 30 individuals in total (Supplementary Table 1); ten of each class, i.e. ten ST, ten FT and ten HC embryos. Samples were stored in ethanol at -20 °C until further handling.

Laboratory procedures & pre-processing of sequence data

We followed the standard "NewtCap" workflow of salt-based extraction of DNA from embryonic tissue, followed by quantification, library preparation, target capture and Illumina sequencing (as described in De Visser et al, 2025). After obtaining the raw, paired-end sequence reads from Baseclear B.V. (Leiden, the Netherlands), we followed a standard pipeline for checking the quality of, and for cleaning-up and mapping, our sequence data in a Linux environment up to and until the deduplication of the BAM files step (as described in De Visser et al, 2025). These deduplicated BAM files served as input for further data extraction and analyses. Throughout the cleaning and mapping process, we used SAMtools' (Danecek et al, 2021) *stats*, *flagstat* and *coverage* options to calculate basic statistics from the FASTQ and BAM files. The reference FASTA file used for read mapping can be found in Supplementary Material as 'Targets.fasta'. These 7,139 sequences, initially used for probe tiling, were based on *T. dobrogicus* transcripts, and had a maximum length of 450bp (De Visser et al, 2025; Wielstra et al, 2019).

173

174 **Preparing read depth data for PAV-spotter**

175 From the BAM files we extracted information on sequence read depth for all sites per target by using
176 the SAMtools depth option (Danecek et al, 2021). We optimized this extracted information by
177 following several file-manipulation steps in a custom 'prepping' shell script – "Script1_prepping.sh"
178 – to make the input files and folder structure match the requirements of our PAV-spotter tool. The steps
179 in Script1_prepping.sh include automatically merging and sorting of intermediate files where
180 appropriate, changing the tab-delimited format to a CSV format, splitting the overall CSV file into
181 multiple files (one file per separate target/gene), and creating a text file with sample names for later
182 use (details are explained in the script). The exact format of the input folder structure and input files is
183 described in Box 1.

BOX 1:

This is a description of the expected input file format for the main PAV-spotter scripts:

'Species' Directories:

- Each species, or otherwise distinguishable set of data, has its own directory

- PAV-spotter is built in such a way that it will loop over multiple such directories

"individuals.txt" file:

- An automatically generated file, uses the initial sample names for input

- Needs to be located in, and corresponding to the contents of, a particular species directory

- This file contains the individual information, with each sample name on a new line and with information on the classes to be compared included in the name (e.g. 'ST', 'FT', and 'HC')

Gene/Target Data Files:

- Also needs to be located in, and corresponding to the contents of, a particular species directory

- Each file represents a single gene or target

- Each file has columns and is in CSV format (this should be the output from batch script 1 "Script1_prepping.sh"):

- Column 1: Gene/target name

- Column 2: Gene/target position (a number)

- Column 3: Read depth data (a number)

- Column 4: Sample/class name (should match identifiers in "individuals.txt" file names)

184 PAV-spotter assumes background knowledge on phenotype classes that presumably differ in the
185 presence of certain genes (in other words, cross-comparisons are not random: for instance males are
186 compared to females, diseased individuals are compared to control samples, etc.). Here, we work with
187 the a priori classification of three phenotypes: two types of diseased embryos (FT vs. ST) and healthy
188 embryos to serve as a control (HC). Which sample belongs to which class needs to be designated by
189 your input filenames (the BAM files), which will end up in an automatically created text file
190 "individuals.txt" after running Script1_prepping.sh. This is crucial, as PAV-spotter performs the
191 comparisons based on the phenotype information embedded in the names of this text file. In case
192 phenotypic classes as specified by the user cannot be deduced from input file names, the user needs to
193 either alter the input file names manually, or alter the identifiers in Script1_prepping.sh manually – or
194 both – before running Script1_prepping.sh (ideally, the user includes such filename identifiers already
195 at the raw FASTQ file stage for consistency throughout the entire bioinformatics procedure).

196

197 **Applying PAV-spotter**

198 We applied Script1_prepping.sh on our total set of 30 samples (Supplementary Table 2; n=30, ten 'ST',
199 ten 'FT' and ten 'HC' individuals, with these identifier abbreviations occurring in the sample names).
200 Additionally, we applied the script on random subsets of samples (Supplementary Table 3 and
201 Supplementary Table 4); two separate analyses with a sample size of five per class (i.e. two total subsets
202 of n=15, indicated by sample set '5_1' and run '5_2'), and five more separate analyses with a sample
203 size of two per class (i.e. five total subsets of n=6, indicated by sample set '2_1', '2_2', '2_3', '2_4'
204 and '2_5'). This allowed us to assess the performance of PAV-spotter when lower sample sizes are
205 used. We randomized the grouping of samples into subsets by using the 'shuf' command from the
206 standard GNU Core Utilities (http://gnu.org/s/coreutils/).

207 We ran a custom MATLAB script ("Script2b_PAVSpotter.m", hereafter 'PAV-spotter' – also
208 available in Python language as "Script2b_PAVSpotter.py") remotely through SLURM workload
209 manager (example batch scripts attached as "Script2a_scheduler_matlab.sh" and
210 "Script2a_scheduler_python.sh", respectively). Users can compare two, or three phenotypic classes
211 (using the argument "categories", see the SLURM scripts), but if only two are provided, only two are
212 compared. Also, we indicated which class we considered the control group (by implementing
213 "ctrl_category = 'HC'") and we provided a common identifier for all input files/targets (with the

214 argument “common_identifier = 'DN'”). Also, information on the working directory and desired output
215 filename was provided in the MATLAB command depending on the analysis, and the same is the case
216 for the customized #SBATCH lines for running the SLURM job.

217 PAV-spotter was run four times on our total set of n=30: once with default settings (no filtering,
218 “reads_threshold” == 0 and “contig_width” == 0), one time with a mild coverage filtering
219 (“reads_threshold” == 5 and “contig_width” == 0), one time with a mild filtering for minimum length
220 of contigs (“reads_threshold” == 0, “contig_width” == 50) and one time with both of the filtering
221 thresholds (“reads_threshold” == 5 and “contig_width” == 50). By setting a soft coverage filter of a
222 minimum of five reads, we filter out reads of any poorly covered target of an individual that does not
223 meet this criterium (thereby assuming a coverage of zero across the target in question). Furthermore,
224 by specifying a minimum contig width, the script will filter out read information of covered regions in
225 between two positions with zero coverage, in case those regions are narrower than, in this case, 50bp
226 (thereby assuming a coverage of zero in that specific target region). For the tests on sample subsets of
227 five individuals per class and two individuals per class, we only used the default filtering settings (no
228 filtering, “reads_threshold” == 0 and “contig_width” == 0). In all analyses, we enabled the script to
229 plot accessory figures (“plot_figures” == TRUE, clarification below). We ran PAV-spotter using
230 MATLAB v.9.13.0 (https://www.mathworks.com) for the main components of our overall study and
231 added the Python version of the tool in order to make it more widely accessible. We used Python v.3.11
232 to run tests and to ensure identical performance and output files.

233

234 **The rationale behind PAV-spotter**

235 After setting up the input files and initiating PAV-spotter successfully, filtering settings are applied as
236 specified. Subsequently, PAV-spotter merges and normalizes the sequence read depth information for
237 all the samples per target per specified sampling class before it starts the actual cross-correlation
238 analyses (but users can turn this setting off in case calculations of all possible cross-comparisons on an
239 individual level are desired). By exploiting the availability of multiple samples per phenotypic class in
240 this way, we ensure that the comparisons will be made on a class level rather than on an individual
241 level. In the latter case, more false outcomes would be expected as a result of the stochastic nature of
242 target capture experiments, something that can be avoided by pooling results. PAV-spotter can loop

243 through a set of input directories if separate datasets need to be analyzed with similar settings
244 consecutively.

245 The cross-correlation analysis in PAV-spotter works as follows (Figure 1). Two target capture
246 results of the same targeted region, but of a different phenotypic class, are defined by D_1 and D_2. These
247 represent two vectors of identical length T, where the vector position represents the target position and
248 the vector values represent the normalized sequence read depth. PAV-spotter then calculates the
249 similarity of the cross-correlation formulated by

250 $$C_{1,2}[n] = \sum_{m=-T}^{T} D_1[m]D_2[m+n]$$

251 for all n between $-T$ and T, with n and m accessing the indices of the vectors $C_{1,2}$, D_1, D_2. This
252 constructs a cross-correlation vector of which the maximum value provides the maximum similarity of
253 D_1 and D_2. For example, in the case that $D_1 = D_2$ (i.e., in the case of autocorrelation), the similarity
254 score C will be 100%.

255 PAV-spotter outputs a CSV file per separate analysis with all the cross-correlation data in the
256 form of percentage similarity, and it outputs a folder with figures. Also, a CSV file with cross-
257 comparisons between data from samples from the control group only (the HCs) is generated. This extra
258 file allows for a calculation of the overall resemblance of the control samples, which should have data
259 present with well-captured targets. This as opposed to the FT and ST samples, which are expected to
260 show 'data gaps', or absence, for some targets.

261 As we are investigating a double hemizygous system, we always have three cross-correlation
262 values to work with. This means we are able to use not only the healthy embryos (HC), but also the
263 other class of diseased embryo (ST or FT), as a control, because genes absent in one class of diseased
264 embryo are expected to be present in both other embryo classes (e.g. to recognize absence in ST
265 embryos, which should have a 1A1A genotype, we can check for presence in the HC embryos which
266 should have the 1A1B genotype, but we can do an additional check for presence in the ST embryos
267 that should have a 1B1B genotype – and the same applies the other way around). Hence, to deduce
268 PAV in the chromosome that is inherited twice in ST embryos, we search for a pattern in which a
269 significant portion of the target was present in both HC and FT embryos (which contain the alternate
270 chromosome form), but absent in ST embryos. Conversely, to deduce PAV in the chromosome

271 inherited twice by FT embryos, we searched for absence in FT, but presence in both HC and ST
272 embryos.

273

274 **Downstream PAV estimation**

275 To automatically deduce PAV patterns, we applied another custom shell script:
276 "Script3_downstream.sh". This script takes the main output file of PAV-spotter, creates an overall
277 matrix of the results, adds columns with information on the cross-correlation data that stand out, and
278 makes lists of the targets that show potential PAV based on a threshold (which can be customized).
279 When, for a certain target, the data of FT versus ST embryos were less than 80% similar, the data of
280 FT versus HC embryos were less than 80% similar, and the data of ST versus HC embryos were more
281 than 80% similar, this target was scored as '1A-linked'. For '1B-linked' genes it was the other way
282 around: FT vs. ST embryos < 80% similar, ST vs. HC embryos < 80% similar, and FT vs ST embryos
283 > 80% similar. This threshold of 20% dissimilarity equals a p-value of between 0.01 (≈ 15%
284 dissimilarity) and 0.001 (≈ 28% dissimilarity). The PAV-spotter output file that shows all similarity
285 scores among the HC embryos only, guided our choice for this threshold (e.g. see Supplementary
286 Figures 1 and 2).

287 Finally, as not much is known about the genetic background of our non-model study species
288 *Triturus*, we performed visual inspections of the read content of all BAM files (n=30) that were used
289 as input for PAV-spotter by checking them in Integrative Genomics Viewer (IGV) software (Robinson
290 et al, 2011) on a Windows environment. This constitutes the last step in our overall workflow (Figure
291 2). We automated obtaining screenshots through IGV by running batch commands for all 7,139 targets
292 (an example batch script is available as "batch_IGV.bat", which is written for Windows, but note that
293 the commands used can similarly be run in Unix-based environments such as Linux and macOS). With
294 a 'checking-by-eye' approach we categorized PAV hits as 'likely true' and 'likely false' in order to
295 assess the quantity of false positive outcomes and we cross-checked the results of all the different runs
296 to identify false negative outcomes (in other words: in case a 'likely true' target with PAV was retrieved
297 in one analysis, but not in another, we counted it as a false negative in the latter analysis).

298 We executed all bioinformatic steps, from pre-processing of reads and read depth information
299 to applying PAV-spotter and extracting information from the output, through the High Performance
300 Computing facility 'ALICE' (Academic Leiden Interdisciplinary Cluster Environment, the

301 Netherlands). The PAV-spotter GitHub repository provides all scripts and further explanation in the
302 "README.md" file: https://github.com/Wielstra-Lab/PAVspotter.

303

304 # Results

305 A mean of 6,483,062 read pairs were generated on average per sample, with a standard deviation of
306 1,450,648 read pairs (Supplementary Table 1). After trimming, this changed to a total of 6,187,680
307 read pairs with a SD of 1,369,254 read pairs. On average, 35.57% of the trimmed reads were
308 successfully mapped against the reference targets after duplicate removal, as an average of 17.09% of
309 all trimmed reads were flagged as duplicates (Supplementary Table 1).

310 Overall, the targets had a mean read depth of 90.09 sequences and a mean coverage of 97.19 %
311 of the sequence bases (Supplementary Table 2, presented per phenotypic class). For the overall set with
312 ten samples per phenotypic class, the average depth of coverage was 84.7 in the FT group, 97.7 in the
313 group of HC embryos, and 87.8 in the ST embryo group (Supplementary Table 2). Moreover, for the
314 batched samples with five individuals per phenotypic class, this average depth of coverage varied
315 between the lowest number of 76.6 in FT batch 5-2 and the highest number of 102.2 in HC batch 5-1
316 (Supplementary Table 3). Lastly, for the batched samples with two individuals per phenotypic class,
317 the averages varied between the lowest number of 37.1 in FT batch 2-1, and the highest number of
318 134.1 in FT batch 2-5 (Supplementary Table 4).

319 Through the different runs with a sample size of ten per class, we discovered large-scale PAV
320 for in total 72 targets. Genes without PAV had sequence reads with similar distributions for all three
321 classes (Figure 3A), whereas of those 72 aberrant genes that we discovered, 32 showed an absence in
322 FT, but a presence in HC and ST embryos (and are thus "1A-linked", Table 5 and Figure 3B), while
323 the remaining 40 were absent in ST, but not in HC and FT embryos (and are thus "1B-linked", Table
324 6 and Figure 3C). We confirmed our findings using the visual IGV inspection (for examples, see
325 Supplementary Figures 3-5).

326 After running PAV-spotter without any filtering options on the full dataset, we correctly
327 identified all 32 1A targets and generated one false positive in the 1A list, which we identified by the
328 visual IGV inspection (see Supplementary Figure 6). By re-running PAV-spotter with the previously
329 described filtering options, this false positive was removed from the list in some cases, however these

330 additional analyses also generated more false positives and false negatives: the frequency of them
331 depending on the combination of filtering settings used (Supplementary Table 5). For the 40 1B targets,
332 we discovered 37 true positives and three false positives in the unfiltered run. Depending on the
333 combination of filtering settings applied, two of these three false positives again disappeared from the
334 list. However, these extra runs with filtering settings also highlighted three additional 1B targets that
335 were overlooked (as false negatives) in the initial analysis (Supplementary Table 6).

336 The false positive outcomes consistently had either the lowest - or in a single case, the highest
337 - mean depth values, reflected by the 'MAXpeak' output of the overall results matrix generated by
338 PAV-spotter. This value is the peak number of reads in one position of a certain target observed across
339 all the individual samples included in the analyses. The false negative outcomes that came to light as
340 true positive results after additional filtering was applied, all had similarity scores for the resembling
341 classes (FT and HC) above 99% in the unfiltered analysis. But the lower similarity scores of these
342 targets between the non-resembling classes (both between FT and ST and between ST and HC) were
343 lower than 90%, but not lower than the 80% threshold (which is why they were initially overlooked).

344 The results for the n=5 per class runs (5_1 and 5_2) resemble our earlier findings. We re-
345 discovered an average of 68.5 (95.1% success) out of the total of 72 PAV targets previously discovered,
346 but with half the sample size. For run 5_1, this number was 70 out of 72 PAV targets (97.2% success)
347 and for run 5_2 this number was 67 out of 72 (93.1% success). Overall, for 1A, the results of both the
348 analyses with a sample size of five individuals per class were complementary, as all 32 previously
349 identified 1A targets were re-discovered at least once (Supplementary Table 7). The same goes for the
350 40 previously identified 1B targets (Supplementary Table 8). Between the two runs, the overall mean
351 depth of coverage was the lowest in 5_2, the analyses that also showed less successful out of the two.

352 The analysis of the 5_1 subset (n=15) resulted in the discovery of 31 true 1A targets plus one
353 false negative and one two positives, and in 39 true 1B targets with one false negatives and one false
354 positive. The analysis of the 5_2 subset (n=15) again resulted in 31 true 1A targets (with one difference)
355 plus one false negative and two false positives. These false outcomes were not the same as with the
356 5_1 subset analysis. For 1B, the 5_2 analysis yielded 36 true 1B targets with four false negatives and
357 seven false positives.

358 The results of the n=2 per class runs (2_1 through 2_5, each with n=6 in total) again highlighted
359 the same PAV exhibiting targets. We re-discovered an average of 70 (97.2% success) out of the total

360 of 72 PAV targets previously discovered, but with a fifth of the sample size (Supplementary Tables 9
361 and 10). For each of the 2_1, 2_2 and 2_4 runs, these numbers were indeed 70 out of 72 (97.2%
362 success), run 2_3 retrieved 71 out of 72 PAV targets (98.6% success) and run 2_5 69 out of 72 (95.8%).

363 Overall, for both the 32 previously identified 1A targets and the 40 previously identified 1B
364 targets, these five analyses with a sample size of only two individuals per class appeared
365 complementary, as all true positive targets were re-discovered at least once. For 1A, one of the false
366 negative outcomes came forward as a false negative in two out of the five analyses (and as a true
367 positive in the three other analyses). The other three false negative outcomes in the 1A list were
368 incidental. For 1B, each of the false negative outcomes occurred in only one of the five analyses (and
369 formed a true positive result in the four alternative analyses). The number of false positive outcomes
370 was slightly higher with these low sample size tests (Supplementary Tables 11 and 12), however this
371 was especially noticeable for the 1A results of the first batch (2_1). The mean depth of coverage was
372 also the lowest for the FT (1B1B) samples in this batch (Supplementary Table 4). We therefore tested
373 for a correlation and show that the mean depth of the samples exhibiting absence (i.e. the mean depth
374 of FT samples with determining 1A absence, and the mean depth of the ST with determining 1B
375 absence) appeared to be negatively and significantly correlated to the number of false positives brought
376 forward (Spearmann's rank correlation, n=10, p=0.018).

377

378 # Discussion

379 We employ a signal cross-correlation approach to discern PAV patterns in target capture data of
380 *Triturus* newt DNA. By comparing the read depth in sequence data of embryos of different phenotypic
381 classes, and by manually checking the results of the read alignments, we are able to identify over
382 seventy targets that appear to be either present in, or absent from, the genome, depending on the
383 phenotype.

384 The three example targets displayed in Figure 3 have been independently tested using multiplex
385 (mx) PCR techniques, including mxKASP (Meilink et al, 2025). Control marker CDK is present in all
386 three embryo classes. This corresponds with our results from PAV-spotter, where we discover high
387 cross-correlations values between the distribution of mapped reads of all three phenotypic classes for
388 this target marker (>99% similarity, way above our 80% cutoff threshold). On the other hand,

389 PLEKHM1 is observed in HC and ST embryos, but not in FT embryos, and NAGLU is observed in
390 HC and FT embryos, but not in ST embryos (Meilink et al, 2025). Again, this matches our PAV-spotter
391 findings – even with sample sizes as low as two individuals per phenotypic class.

392 Evidently, PAV-spotter relies on the correct, a priori classification of samples. When multiple
393 samples are provided, read-depth data are merged per phenotypic class, as the core aim of PAV-spotter
394 is to detect PAV *between* predefined groups rather than *within* them. If desired, the merging and
395 normalization procedure can be disabled to compute cross-correlations among individual samples – for
396 example to explore potential polymorphisms within phenotypic classes. However, this is not the default
397 or recommended use – thus future users should carefully consider what they want to compare, and
398 why. For instance, we underline that cross-contamination of DNA is a main concern here when using
399 target capture data of ancient DNA (Zavala et al, 2022), as this could potentially distort the similarity
400 values calculated by PAV-spotter – something to be wary of.

401 The fact that PAV-spotter is able to, on average, re-discover 97,2% of true positive PAV target
402 outcomes in our trials is especially convenient for studies where scientists must rely on a limited
403 amount of available DNA, as is often the case with herbarium specimens (Hart et al, 2016; Kates et al,
404 2021). However, the highest yield of false positive outcomes is observed in our 1A results of batch
405 2_1, but not the accessory 1B results, which firstly shows that there is a likely trade-off between sample
406 size and sequence coverage. Preferably, the quality of DNA is as high as possible when working with
407 small sample sizes. Although also preferred in case of a larger sample size per phenotypic class, the
408 chances are then higher that any poor coverage sample(s) will be compensated for by sample(s) with
409 better coverage. In general, spreading sequencing efforts across at least a couple samples, with slightly
410 lower – but still informative – depths per sample, is considered a safer option than working with
411 extremely low sample sizes (Lou et al, 2021; Pezzini et al, 2023).

412 Regardless of sample size, identifying and characterizing structural variation from target
413 capture data is widely recognized to be difficult (Jones & Good, 2016) and it is especially challenging
414 with non-model organisms. This is because capture-rates may vary considerably depending on bait
415 design, sample quality, species relatedness, batch effects, and stochastic factors (Andermann et al,
416 2019; Feng et al, 2015). In most cases where targets showed a significant absence of mapped reads, we
417 observe a small amount of reads being (mis)mapped against reference targets when none are expected,
418 for instance (visible in the PAV-spotter output figures and the IGV screenshots). Occasionally, these
419 consist of (clipped/partially matching) reads, something that can be caused by sequencing errors,

420 chimeric reads, errors in the reference sequence, tandem duplications, or genomic rearrangements and
421 structural variants (Schroder et al, 2014; Suzuki et al, 2011). Due to the case-specificity of such
422 potential errors, benchmarking across a broader range of target capture datasets – including those
423 analyzed in independent follow-up studies using PAV-spotter – will be relevant to further assess the
424 accuracy of the PAV-spotter method and, where applicable, to keep building upon its performance over
425 time.

426 A solution to remove any unwanted (mis-)mapped reads would be to filter more strictly
427 upstream. However, in case of samples or targets that show poor or limited coverage – an issue that is
428 not uncommon with target capture procedures (Bragg et al, 2016) – strict filtering may not be desired.
429 This means that, due to this potential stochasticity, merely using existing tools to check whether there
430 are any mapped reads at all in a sample/target (i.e. checking for the presence of zero reads versus >0
431 reads), or building *de novo* assembled contigs per target on a sample-per-sample basis, will not be
432 sufficient to identify PAV accurately. Our method offers an alternative solution, as PAV-spotter
433 appears robust enough to detect PAV, even in the face of low coverage and (partially) mis-mapped and
434 clipped reads. However, the trade-off in false positive and false negative outcomes will largely depend
435 on the similarity thresholds and other criteria set by users, as well as on any manual, double-checks
436 performed. General risks associated with target capture experiments, such as inadvertently sequencing
437 paralogues, pseudogenes, or other off-target or multi-copy regions due to probe non-specificity, should
438 therefore always be taken into account: in principle, PAV-spotter is applied under the assumption that
439 targets represent orthologous loci.

440 In conclusion, we show that considering the genomic position as a variable for signal
441 displacement instead of time – which is generally the case in more classic cross-correlation applications
442 (Verhaegen & Verdult, 2007) – makes it possible to identify markers of PAV/structural variation in
443 target capture data, without needing any prior knowledge on large-scale, genomic context. And while
444 future extensions of the tool may include the incorporation of whole-genome sequence data to further
445 guide interpretation, the current implementation enables robust PAV detection under the constraint of
446 lacking a reference genome. Our study therefore shows that a multidisciplinary bioengineering and
447 biotechnological approach can help bioinformatics, and thus the fields of evolutionary and molecular
448 research, forward when dealing with challenging research questions, datasets, and study organisms.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 802759). Sample collection was supported by the Serbian Ministry of Science, Technological Development and Innovation (grants nos. 451-03-66/2024-01/200007, 451-03-65/2024-03/200178, 451-03-66/2024-03/200178). Dr. Ana Ivanović supported breeding and collecting *Triturus* samples. Individuals of *T. ivanbureschi* and *T. macedonicus* were originally collected from natural populations: *T. ivanbureschi* from Zli Dol, Serbia with permission obtained from the Serbian Ministry of Energy, Development and Environmental Protection (permit No. 353-01-75/2014-08) and *T. macedonicus* in Ceklin, Montenegro with permission obtained from the Agency for Environmental Protection, Montenegro (permit No. UPI-328/4).The experimental procedures were approved by the Ethics Committee of the Institute for Biological Research „Siniša Stanković", University of Belgrade (decisions no. 03-03/16 and 01-1949). All experiments were performed in accordance with the Directive 2010/63/EU.

References

Andermann, T., Cano, A., Zizka, A., Bacon, C. & Antonelli, A. (2018) SECAPR-a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. *PeerJ*, 6, e5175.

Andermann, T., Torres Jimenez, M. F., Matos-Maravi, P., Batista, R., Blanco-Pastor, J. L., Gustafsson, A. L. S., Kistler, L., Liberal, I. M., Oxelman, B., Bacon, C. D. & Antonelli, A. (2019) A guide to carrying out a phylogenomic target sequence capture project. *Frontiers in Genetics*, 10, 1407.

Annavi, G., Uli, J. E. & Arumugam, R. (2019) A review of the application of Next Generation Sequencing (NGS) in wild terrestrial vertebrate research. *Annual Research & Review in Biology*, 1-9.

Berdan, E. L., Flatt, T., Kozak, G. M., Lotterhos, K. E. & Wielstra, B. (2022) Genomic architecture of supergenes: connecting form and function. *Philosophical Transactions of the Royal Society B*, 377(1856), 20210192.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C. & Good, J. M. (2012) Transcriptome-based exon capture enables highly cost-effective comparative genomic data colection at moderate evolutionary scales. *BMC Genomics*, 13(403), 1471-2164.

Bragg, J. G., Potter, S., Bi, K. & Moritz, C. (2016) Exon capture phylogenomics: efficacy across scales of divergence. *Molecular Ecology Resources*, 16(5), 1059-68.

Da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., Zepeda-Mendoza, M. L., Campos, P. F., Heller, R. & Pereira, R. J. (2016) Next-generation

483 biology: sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, 30,
484 3-13.

485 Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane,
486 T., McCarthy, S. A., Davies, R. M. & Li, H. (2021) Twelve years of SAMtools and BCFtools.
487 *Gigascience*, 10(2).

488 De Visser, M. C., France, J., McCartney-Melstad, E., Bucciarelli, G. M., Theodoropoulos, A.,
489 Shaffer, H. B. & Wielstra, B. (2025) NewtCap: an efficient target capture approach to boost genomic
490 studies in Salamandridae (True Salamanders and Newts). *Ecology and Evolution* 15(8), e71835..

491 De Visser, M. C., France, J., Paulouskaya, O., Brown, T., Fahrbach, M., Van der Ploeg, C. &
492 Wielstra, B. (2024) Conserved gene content and unique phylogenetic history characterize the
493 'bloopergene' underlying *Triturus*' balanced lethal system *bioRxiv*, 2024.10.25.620277.

494 Dufresnes, C. & Crochet, P. A. (2022) Sex chromosomes as supergenes of speciation: why
495 amphibians defy the rules? *Philosophical Transactions of the Royal Society B*, 377(1856), 20210202.

496 Ekblom, R. & Galindo, J. (2011) Applications of next generation sequencing in molecular ecology of
497 non-model organisms. *Heredity*, 107(1), 1-15.

498 Etherington, G. J., Heavens, D., Baker, D., Lister, A., McNelly, R., Garcia, G., Clavijo, B.,
499 Macaulay, I., Haerty, W. & Di Palma, F. (2020) Sequencing smart: de novo sequencing and assembly
500 approaches for a non-model mammal. *Gigascience*, 9(5).

501 Feng, Y., Chen, D., Wang, G. L., Zhang, V. W. & Wong, L. J. (2015) Improved molecular diagnosis
502 by the detection of exonic deletions with target gene capture and deep sequencing. *Genetics in*
503 *Medicine*, 17(2), 99-107.

504 France, J., De Visser, M., Arntzen, J. W., Babik, W., Cvijanović, M., Ivanović, A., Smith, J. &
505 Vučić, T. (2024) Genomic evidence suggests the balanced lethal system in *Triturus* newts originated
506 in an instantaneous speciation event. *bioRxiv*, 2024.10.29.620207.

507 Gabur, I., Chawla, H. S., Lopisso, D. T., von Tiedemann, A., Snowdon, R. J. & Obermeier, C. (2020)
508 Gene presence-absence variation associates with quantitative *Verticillium longisporum* disease
509 resistance in *Brassica napus*. *Scientific Reports*, 10(1), 4131.

510 Gao, Z., Cecati, C. & Ding, S. X. (2015) A survey of fault diagnosis and fault-tolerant techniques -
511 Part I: fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on*
512 *Industrial Electronics*, 62(6), 3757-3767.

513 Gauthier, J., Vincent, A. T., Charette, S. J. & Derome, N. (2019) A brief history of bioinformatics.
514 *Briefings in Bioinformatics*, 20(6), 1981-1996.

515 Gerdol, M., Moreira, R., Cruz, F., Gomez-Garrido, J., Vlasova, A., Rosani, U., Venier, P., Naranjo-
516 Ortiz, M. A., Murgarella, M., Greco, S., Balseiro, P., Corvelo, A., Frias, L., Gut, M., Gabaldon, T.,
517 Pallavicini, A., Canchaya, C., Novoa, B., Alioto, T. S., Posada, D. & Figueras, A. (2020) Massive
518 gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome*
519 *Biology*, 21(1), 275.

520 Gutierrez-Valencia, J., Hughes, P. W., Berdan, E. L. & Slotte, T. (2021) The Genomic Architecture
521 and Evolutionary Fates of Supergenes. *Genome Biology and Evolution*, 13(5).

522 Hall, D. W. & Wayne, M. L. (2013) Ohno's "peril of hemizygosity" revisited: gene loss, dosage
523 compensation, and mutation. *Genome Biology and Evolution*, 5(1), 1-15.

Hart, M. L., Forrest, L. L., Nicholls, J. A. & Kidner, C. A. (2016) Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon*, 65(5), 1081-1092.

Harteveld, C. L. & Higgs, D. R. (2010) α-thalassaemia. *Orphanet Journal of Rare Diseases*, 5(13).

Hu, H., Scheben, A., Verpaalen, B., Tirnaz, S., Bayer, P. E., Hodel, R. G. J., Batley, J., Soltis, D. E., Soltis, P. S. & Edwards, D. (2022) *Amborella* gene presence/absence variation is associated with abiotic stress responses that may contribute to environmental adaptation. *New Phytologist*, 233(4), 1548-1555.

Jain, A. K., Duin, R. P. W. & Mao, J. (2000) Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1), 4-37.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega, N. J. & Wickett, N. J. (2016) HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, 4(7).

Jones, M. R. & Good, J. M. (2016) Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, 25(1), 185-202.

Joron, M., Papa, R., Beltran, M., Chamberlain, N., Mavarez, J., Baxter, S., Abanto, M., Bermingham, E., Humphray, S. J., Rogers, J., Beasley, H., Barlow, K., ffrench-Constant, R. H., Mallet, J., McMillan, W. O. & Jiggins, C. D. (2006) A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biology*, 4(10), e303.

Kates, H. R., Doby, J. R., Siniscalchi, C. M., LaFrance, R., Soltis, D. E., Soltis, P. S., Guralnick, R. P. & Folk, R. A. (2021) The effects of herbarium specimen characteristics on short-read NGS sequencing success in nearly 8000 specimens: old, degraded samples have lower DNA yields but consistent sequencing success. *Frontiers in Plant Science*, 12, 669064.

Kaur, P. & Gaikwad, K. (2017) From genomes to gene-omes: exome sequencing concept and applications in crop improvement. *Frontiers in Plant Science*, 8, 1-7.

Kay, T., Helleu, Q. & Keller, L. (2022) Iterative evolution of supergene-based social polymorphism in ants. *Philosophical Transactions of the Royal Society B*, 377(1856), 20210196.

Kupper, C., Stocks, M., Risse, J. E., Dos Remedios, N., Farrell, L. L., McRae, S. B., Morgan, T. C., Karlionova, N., Pinchuk, P., Verkuil, Y. I., Kitaysky, A. S., Wingfield, J. C., Piersma, T., Zeng, K., Slate, J., Blaxter, M., Lank, D. B. & Burke, T. (2016) A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics*, 48(1), 79-83.

Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., Kerje, S., Gustafson, U., Shi, C., Zhang, H., Chen, W., Liang, X., Huang, L., Wang, J., Liang, E., Wu, Q., Lee, S. M., Xu, X., Hoglund, J., Liu, X. & Andersson, L. (2016) Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nature Genetics*, 48(1), 84-8.

Lemmon, E. M. & Lemmon, A. R. (2013) High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44(1), 99-121.

Li, C., Binaghi, M., Pichon, V., Cannarozzi, G., Brandao de Freitas, L., Hanemian, M. & Kuhlemeier, C. (2023) Tight genetic linkage of genes causing hybrid necrosis and pollinator isolation between young species. *Nature Plants*, 9(3), 420-432.

Li, J., Cocker, J. M., Wright, J., Webster, M. A., McMullan, M., Dyer, S., Swarbreck, D., Caccamo, M., Oosterhout, C. V. & Gilmartin, P. M. (2016) Genetic architecture and evolution of the S locus supergene in *Primula vulgaris*. *Nature Plants*, 2(12), 16188.

566 Li, J. & Lee, C. R. (2023) The role of gene presence-absence variations on genetic incompatibility in
567 Asian rice. *New Phytologist*, 239(2), 778-791.

568 Llaurens, V., Whibley, A. & Joron, M. (2017) Genetic architecture and balancing selection: the life
569 and death of differentiated variants. *Molecular Ecology*, 26, 2430-2448.

570 Lou, R. N., Jacobs, A., Wilder, A. P. & Therkildsen, N. O. (2021) A beginner's guide to low-
571 coverage whole genome sequencing for population genomics. *Molecular Ecology*, 30(23), 5966-
572 5993.

573 Macgregor, H. C. & Horner, H. (1980) Heteromorphism for chromosome 1, a requirement for normal
574 development in crested newts. *Chromosoma*, 76, 111-122.

575 Mahmoud, M., Gobet, N., Cruz-Davalos, D. I., Mounier, N., Dessimoz, C. & Sedlazeck, F. J. (2019)
576 Structural variant calling: the long and the short of it. *Genome Biology*, 20(1), 246.

577 Marroni, F., Pinosio, S. & Morgante, M. (2014) Structural variation and genome complexity: is
578 dispensable really dispensable? *Current Opinion in Plant Biology*, 18, 31-6.

579 Meilink, W. R. M., De Visser, M., Theodoropoulos, A., Fahrbach, M. & Wielstra, B. (2025)
580 Determining zygosity with multiplex Kompetitive Allele-Specific PCR (mxKASP) genotyping.
581 *Ecology and Evolution* 15(6), e71642.

582 Morganti, S., Tarantino, P., Ferraro, E., D'Amico, P., Duso, B. A. & Curigliano, G. (2019) Next
583 generation sequencing (NGS): a revolutionary technology in pharmacogenomics and personalized
584 medicine in cancer, in Ruiz-Garcia, E. A.-d. l. V., H. (ed), *Translational Research and Onco-Omics*
585 *Applications in the Era of Cancer Personal Genomics*. Advances in Experimental Medicine and
586 Biology: Springer Nature Switzerland, 9-30.

587 Ortiz, E. M., Höwener, A., Shigita, G., Raza, M., Maurin, O., Zuntini, A., Forest, F., Baker, W. J.,
588 Schaefer, H. (2023). A novel phylogenomics pipeline reveals complex pattern of reticulate evolution
589 in Cucurbitales. *bioRxiv*, 2023.10.27.564367.

590 Pennisi, E. (2017) 'Supergenes' drive evolution. *Science*, 357, 1083.

591 Pezzini, F. F., Ferrari, G., Forrest, L. L., Hart, M. L., Nishii, K. & Kidner, C. A. (2023) Target
592 capture and genome skimming for plant diversity studies. *Applications in Plant Sciences*, 11(4),
593 e11537.

594 Puritz, J. B. & Lotterhos, K. E. (2018) Expressed exome capture sequencing: a method for cost-
595 effective exome sequencing for all organisms. *Molecular Ecology Resources*, 18, 1209-1222.

596 Rosa, R. D., Alonso, P., Santini, A., Vergnes, A. & Bachere, E. (2015) High polymorphism in big
597 defensin gene expression reveals presence-absence gene variability (PAV) in the oyster *Crassostrea*
598 *gigas*. *Developmental and Comparative Immunology*, 49(2), 231-8.

599 Rovelli, V., Ruiz-González, A., Vignoli, L., Macale, D., Buono, V., Davoli, F., Vieites, D. R.,
600 Pezaro, N. & Randi, E. (2019) Genotyping-by-Sequencing (GBS) of large amphibian genomes: a
601 comparative study of two non-model species endemic to Italy. *Animal Biology*, 69(3), 307-326.

602 Schroder, J., Hsu, A., Boyle, S. E., Macintyre, G., Cmero, M., Tothill, R. W., Johnstone, R. W.,
603 Shackleton, M. & Papenfuss, A. T. (2014) Socrates: identification of genomic rearrangements in
604 tumour genomes by re-aligning soft clipped reads. *Bioinformatics*, 30(8), 1064-1072.

605 Schwander, T., Libbrecht, R. & Keller, L. (2014) Supergenes and complex phenotypes. *Current*
606 *Biology*, 24, R288-R294.

607 Sessions, S. K., Macgregor, H. C., Schmid, M. & Haaf, T. (1988) Cytology, embryology, and
608 evolution of the developmental arrest syndrome in newts of the genus *Triturus* (Caudata:
609 Salamandridae). *The Journal of Experimental Zoology*, 248, 321-334.

610 Sims, S. H., Macgregor, H. C., Pellatt, P. S. & Horner, H. A. (1984) Chromosome 1 in crested and
611 marbled newts (*Triturus*) - An extraordinary case of heteromorphism and independent chromosome
612 evolution. *Chromosoma*, 89, 169-185.

613 Suzuki, S., Yasuda, T., Shiraishi, Y., Miyano, S. & Nagasaki, M. (2011) ClipCrop: a tool for
614 detecting structural variations with single-base resolution using soft-clipping information. *BMC*
615 *Bioinformatics*, 12, S7.

616 Thompson, M. J. & Jiggins, C. D. (2014) Supergenes and their role in evolution. *Heredity*, 113, 1-8.

617 Tuttle, E. M., Bergland, A. O., Korody, M. L., Brewer, M. S., Newhouse, D. J., Minx, P., Stager, M.,
618 Betuel, A., Cheviron, Z. A., Warren, W. C., Gonser, R. A. & Balakrishnan, C. N. (2016) Divergence
619 and functional degradation of a sex chromosome-like supergene. *Current Biology*, 26, 344-350.

620 Verhaegen, M. & Verdult, V. (2007) 4.3 Random Signals, *Filtering and system identification: a least*
621 *squares approach.* Cambridge university press, 100-103.

622 Vucic, T., Drobnjakovic, M., Ajdukovic, M., Bugarcic, M., Wielstra, B., Ivanovic, A. & Cvijanovic,
623 M. (2024) A staging table of Balkan crested newt embryonic development to serve as a baseline in
624 evolutionary developmental studies. *Journal of Experimental Zoology Part B: Molecular and*
625 *Developmental Evolution*.

626 Vučić, T., Ivanović, A., Ajduković, M., Bajler, N. & Cvijanović, M. (2022) The reproductive success
627 of *Triturus ivanbureschi* x *T. macedonicus* F_1 hybrid females (Amphibia: Salamandridae). *Animals*,
628 12(4).

629 Wachi, N., Matsubayashi, K. W. & Maeto, K. (2018) Application of next-generation sequencing to
630 the study of non-model insects. *Entomological Science*, 21(1), 3-11.

631 Wallace, H. (1987) Abortive development in the crested newt *Triturus cristatus*. *Development*, 100,
632 65-72.

633 Wang, Y., Lu, J., Chen, S., Shu, L., Palmer, R. G., Xing, G., Li, Y., Yang, S., Yu, D., Zhao, T. &
634 Gai, J. (2014) Exploration of presence/absence variation and corresponding polymorphic markers in
635 soybean genome. *Journal of Integrative Plant Biology*, 56(10), 1009-19.

636 Wellenreuther, M., Merot, C., Berdan, E. & Bernatchez, L. (2019) Going beyond SNPs: the role of
637 structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*,
638 28(6), 1203-1209.

639 Wielstra, B. (2020) Balanced lethal systems. *Current Biology*, 30, R742-R743.

640 Wielstra, B., McCartney-Melstad, E., Arntzen, J. W., Butlin, R. K. & Shaffer, H. B. (2019)
641 Phylogenomics of the adaptive radiation of *Triturus* newts supports gradual ecological niche
642 expansion towards an incrementally aquatic lifestyle. *Molecular Phylogenetics and Evolution*, 133,
643 120-127.

644 Wold, J., Koepfli, K. P., Galla, S. J., Eccles, D., Hogg, C. J., Le Lec, M. F., Guhlin, J., Santure, A.
645 W. & Steeves, T. E. (2021) Expanding the conservation genomics toolbox: incorporating structural
646 variants to enhance genomic studies for species of conservation concern. *Molecular Ecology*, 30(23),
647 5949-5965.

Yohe, L. R., Davies, K. T. J., Simmons, N. B., Sears, K. E., Dumont, E. R., Rossiter, S. J. & Dávalos, L. M. (2020) Evaluating the performance of targeted sequence capture, RNA-Seq, and degenerate-primer PCR cloning for sequencing the largest mammalian multigene family. *Molecular Ecology Resources*, 20, 140-153.

Yuan, H., Atta, C., Tornabene, L. & Li, C. (2019) Assexon: assembling exon using gene capture data. *Evolutionary Bioinformatics*, 15, 1176934319874792.

Zaharias, P., Pante, E., Gey, D., Fedosov, A. E. & Puillandre, N. (2020) Data, time and money: evaluating the best compromise for inferring molecular phylogenies of non-model animal taxa. *Molecular Phylogenetics and Evolution*, 142, 106660.

Zavala, E. I., Aximu-Petri, A., Richter, J., Nickel, B., Vernot, B. & Meyer, M. (2022) Quantifying and reducing cross-contamination in single- and multiplex hybridization capture of ancient DNA. *Molecular Ecology Resources*, 22(6), 2196-2207.

Zhang, L. M., Luo, H., Liu, Z. Q., Zhao, Y., Luo, J. C., Hao, D. Y. & Jing, H. C. (2014) Genome-wide patterns of large-size presence/absence variants in sorghum. *Journal of Integrative Plant Biology*, 56(1), 24-37.

Data Accessibility and Benefit-Sharing

The raw, Illumina sequencing reads used in this study have been submitted to the NCBI Sequence Read Archive (SRA) and are publicly available through BioProject PRJNA1111729 (https://www.ncbi.nlm.nih.gov/sra/PRJNA1111729). The PAV-spotter tool, including further explanation on how to use and customize the underlying code, is available through the PAV-spotter GitHub repository: https://github.com/Wielstra-Lab/PAVspotter. Furthermore, Supplementary Materials are provided through a Zenodo repository: https://zenodo.org/records/13991751. Figures created by PAV-spotter for each separate analysis that are not shown in the paper can be provided upon request. The same goes for all IGV screenshot images generated by the batch script.

Author contributions

MdV and CvdP designed the tool. BW, MdV and CvdP designed the experiments. MC, TV and AI collected the samples and conducted the phenotypic classification. MdV and AT performed the molecular laboratory work. CvdP wrote the PAV-spotter scripts (Scripts 2a and 2b, both the MATLAB and the Python versions), which comprises of the signal cross-correlation code. Other scripts (Script 1, Script 3, and the IGV batch script) were written by MdV. MdV conducted the main bioinformatics, data acquisition, and data interpretation. MdV drafted the work and CvdP added mathematical details

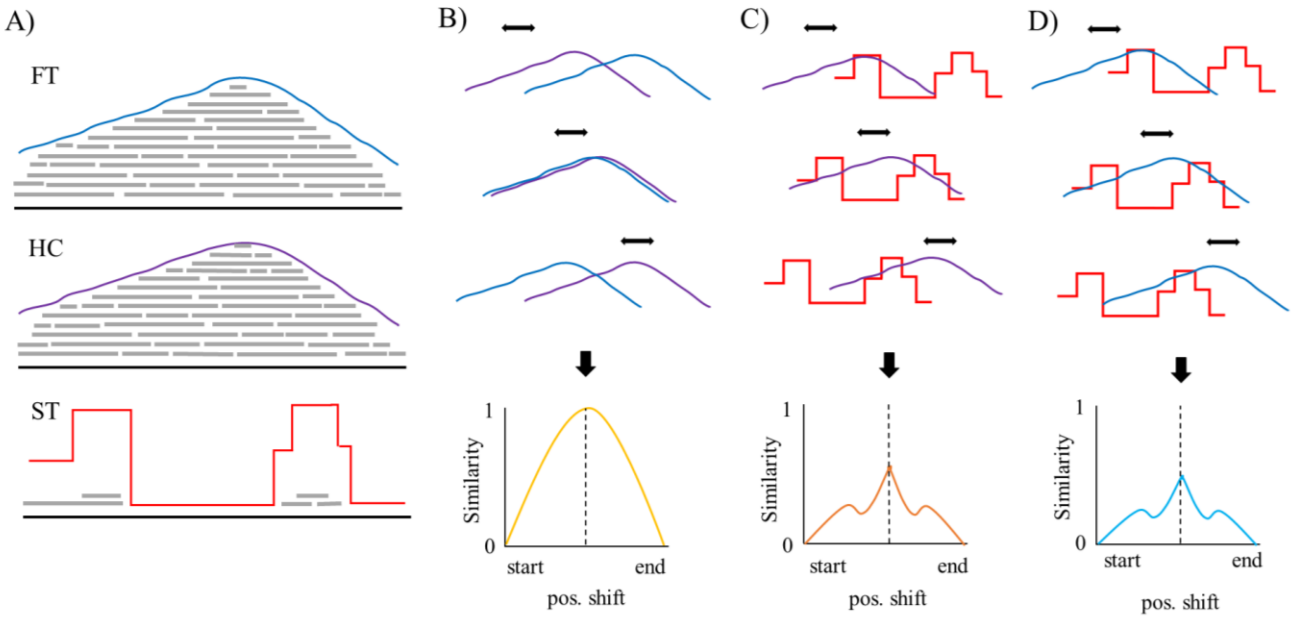
to the text and conducted p-value/threshold estimations. All authors revised the manuscript and approved of the submitted version.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

705 # Figures

706



707

708 ***Figure 1:*** *A simplified visualization of the cross-correlation methodology in PAV-spotter, showing a*
709 *hypothetical gene/target that is missing in only ST embryos as an example.* ***(A)*** *A number of sequence*
710 *reads are mapped (in grey) against a gene/target. This information is taken from depth files and*
711 *merged per phenotypic class (in case of multiple samples per class), then the absolute distributions of*
712 *the read depths are normalized, here represented by the colored lines; dark blue = FT, purple = HC,*
713 *red = ST.* ***(B)*** *The distribution data of the FT and the HC classes are compared. A measure of similarity*
714 *is determined as a function of the displacement of one read depth distribution relative to the other, as*
715 *if they were to 'slide over' each other (indicated by black arrows). The similarity appears close to 1*
716 *(=100%) and the graph produced by PAV-spotter also follows a smooth line.* ***(C)*** *The distribution data*
717 *of the HC and ST classes are compared: the similarity is not close to 100% and the similarity graph*
718 *produced by PAV-spotter does not follow a smooth line.* ***(D)*** *The distribution data of the HC and ST*
719 *classes are compared: the similarity is not close to 100% and the similarity graph produced by PAV-*
720 *spotter does not follow a smooth line.*

721

722

723

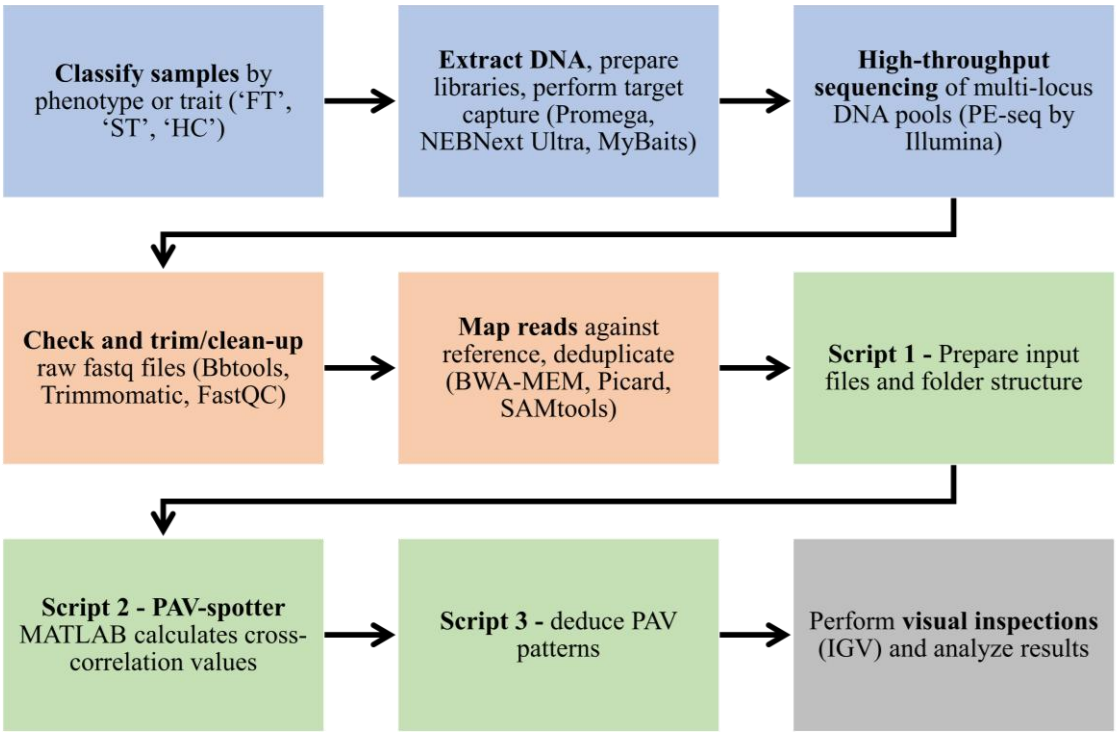


Figure 2: *A summary of the consecutive steps of our methods in which the order of the steps is indicated by black arrows. Blue boxes show the laboratory process, orange boxes represent bioinformatic pre-processing steps, green boxes stand for the application of the main PAV-spotter scripts, and the grey box covers the conclusive steps of inspecting and interpreting the results.*

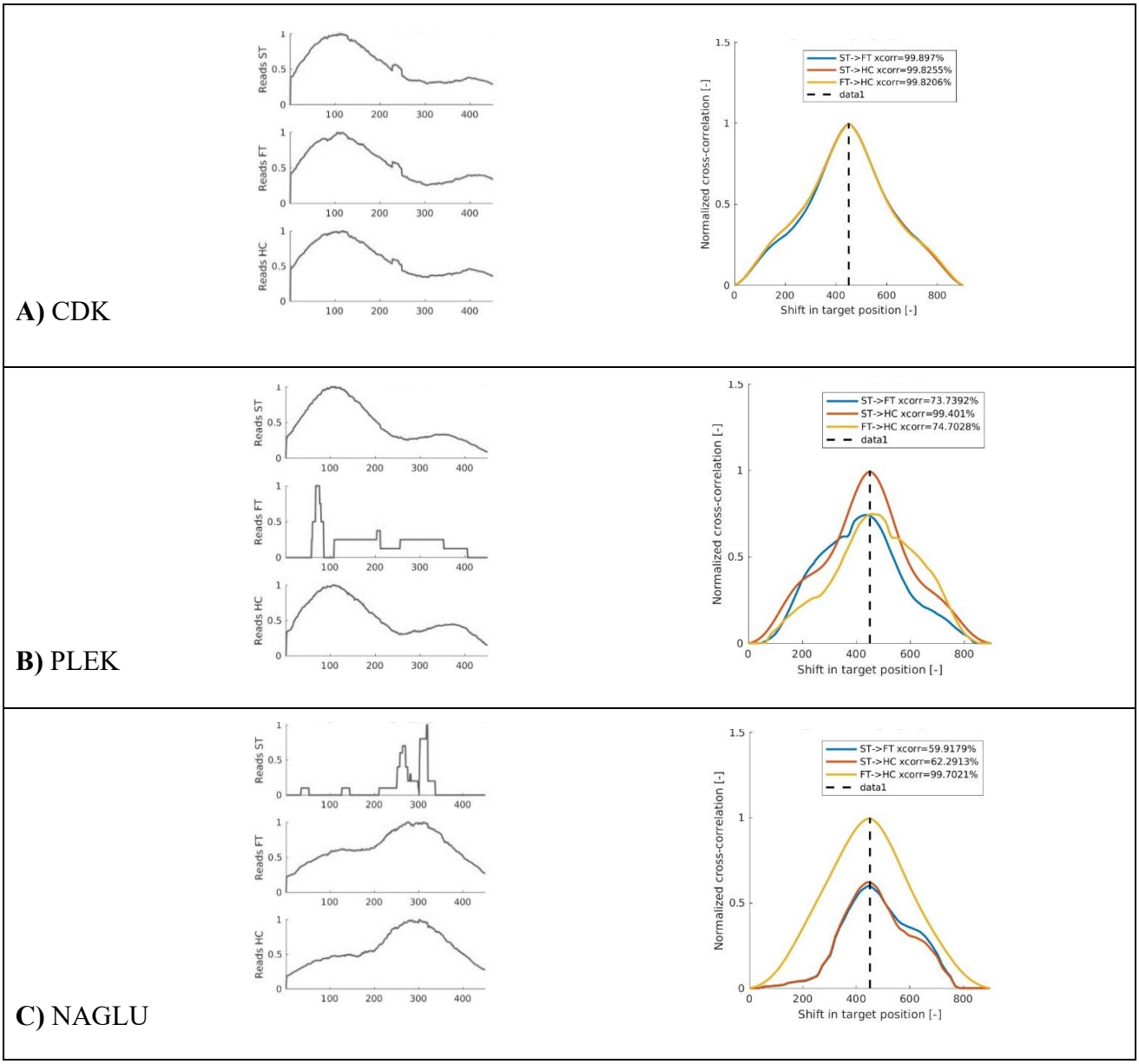


Figure 3: *Examples of target-specific plots produced by PAV-spotter in the overall run with no filtering (n=10 per class). The stacked plots on the left side within the panels A-C show the merged and normalized distributions of sequence read depths per phenotypic class, and the colored plots on the right side within the panels A-C show associated measures of similarity of those distributions as a function of the displacement of one relative to the other (including a legenda explaining the colors and cross-correlation values).* ***A)*** *An example of a 'normal' gene/target, showing the cross-correlation analyses of control marker 'CDK' (see Discussion), with a similar shape of the sequence read depth distributions and high correlation values (above the 80% similarity threshold) between all three classes.* ***B)*** *An example of a 1A-linked gene/target, showing the cross-correlation analyses of 'PLEKHM1' (see Discussion), with a deviant read depth distribution for the FT class and a high*

751 *correlation value (>80%) for ST vs. HC samples, but a lower cross-correlation value (<80%) for the*
752 *ST vs. FT and FT vs. HC sample comparisons.* ***C)*** *An example of a 1B-linked gene/target, showing the*
753 *cross-correlation analyses of 'NAGLU' (see Discussion), with a deviant read depth distribution for the*
754 *ST class and a high correlation value (>80%) for FT vs. HC samples, but lower cross-correlation*
755 *values (<80%) for the FT vs. ST and the ST vs. HC samples.*