

SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence

Hélène Lopez-Maestre, Lilia Brinza, Camille Marchet, Janice Kielbassa, Sylvère Bastien, Mathilde Boutigny, David Monnin, Adil El Filali, Claudia Marcia Carareto, Cristina Vieira, et al.

▶ **To cite this version:**

Hélène Lopez-Maestre, Lilia Brinza, Camille Marchet, Janice Kielbassa, Sylvère Bastien, et al.. SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. Nucleic Acids Research, 2016, 10.1093/nar/gkw655 . hal-01352586

HAL Id: hal-01352586
https://inria.hal.science/hal-01352586

Submitted on 8 Aug 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nucleic Acids Research Advance Access published July 25, 2016

Nucleic Acids Research, 2016 **1**
doi: 10.1093/nar/gkw655

SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence

Hélène Lopez-Maestre[1,2], Lilia Brinza[3], Camille Marchet[4], Janice Kielbassa[5], Sylvère Bastien[1,2], Mathilde Boutigny[1,2], David Monnin[1], Adil El Filali[1], Claudia Marcia Carareto[6], Cristina Vieira[1,2], Franck Picard[1], Natacha Kremer[1], Fabrice Vavre[1,2], Marie-France Sagot[1,2] and Vincent Lacroix[1,2,*]

[1]Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France, [2]EPI ERABLE - Inria Grenoble, Rhône-Alpes, [3]PT Génomique et Transcriptomique, BIOASTER, Lyon, France, [4]Université de Rennes, F-35000 Rennes; équipe GenScale, IRISA, Rennes, [5]Synergie-Lyon-Cancer, Universite Lyon 1, Centre Leon Berard, Lyon, France and [6]Department of Biology, UNESP - São Paulo State University, São José do Rio Preto, São Paulo, Brazil

Received December 22, 2015; Revised July 05, 2016; Accepted July 11, 2016

ABSTRACT

SNPs (Single Nucleotide Polymorphisms) are genetic markers whose precise identification is a prerequisite for association studies. Methods to identify them are currently well developed for model species, but rely on the availability of a (good) reference genome, and therefore cannot be applied to non-model species. They are also mostly tailored for whole genome (re-)sequencing experiments, whereas in many cases, transcriptome sequencing can be used as a cheaper alternative which already enables to identify SNPs located in transcribed regions. In this paper, we propose a method that identifies, quantifies and annotates SNPs without any reference genome, using RNA-seq data only. Individuals can be pooled prior to sequencing, if not enough material is available from one individual. Using pooled human RNA-seq data, we clarify the precision and recall of our method and discuss them with respect to other methods which use a reference genome or an assembled transcriptome. We then validate experimentally the predictions of our method using RNA-seq data from two non-model species. The method can be used for any species to annotate SNPs and predict their impact on the protein sequence. We further enable to test for the association of the identified SNPs with a phenotype of interest.

INTRODUCTION

Understanding the genetic basis of complex phenotypes remains a central question in biology. A classical approach consists in genotyping a large number of individuals in a population based on a pre-specified catalog of variants, and in associating their genotypes to the studied phenotype. This type of approach can be applied to many loci at once, or even genome wide, through what has been called genome wide association studies (GWAS). These methods have been successfully adopted for human and model species. However, the total cost of GWAS remains very high, and the current framework cannot be applied to non-model species for which genomic resources are sparsely or not available. The recent progress in sequencing technologies together with the recent developments in assembly algorithms are largely changing this view. It can now be envisioned to search for variants associated with a phenotype using NGS data only, without relying on pre-existing genomic resources (that have potential limitations). A possible procedure, applicable to model or non-model species, consists in: (i) sequencing the genome; (ii) assembling it; (iii) identifying the SNPs; (iv) genotyping individuals and (v) associating genotypes with phenotypes. However, such a procedure remains costly and still presents the classical problems of sequential pipelines, namely the potential to accumulate experimental and computational errors at each step.

If the purpose of the study is to identify the variants related to a phenotype, the procedure can be simplified in many ways. First, SNPs can be called *de novo* from the reads, without separating the steps of assembly and SNP calling. Second, cost effective methods like exome or transcriptome sequencing may be adopted as the full genome is not al-

*To whom correspondence should be addressed. Email: vincent.lacroix@univ-lyon1.fr

© The Author(s) 2016. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

ways necessary. Third, pooling individuals may be an attractive option if genotyping is not required. These options have been explored individually and give promising results. *De novo* assembly of SNPs is now computationally possible (1–3). The clear advantage is that it can be applied to non-model species, where no reference genome is available. Even in the case where a reference genome is available, these methods still give good results compared to mapping-based approaches, compensating their lower sensitivity by an ability to call more variants in repeated regions. Transcriptome sequencing is already used in several projects, both in the context of model species (4) and non-model species (5–7). In both cases, it was shown that the SNP calling methods could be tailored to have a good precision, meaning that most of the reported SNPs are true SNPs. However, their recall (i.e. capacity to exhaustively report all SNPs) remains to be clearly determined. Clearly, only SNPs from transcribed regions can be targeted, but they arguably correspond to those with a more direct functional impact. Using RNA-seq technology largely reduces the cost of the experiment, and the obtained data concurrently mirror gene expression, the most basic molecular phenotype. RNA-seq experiments may also provide very high depth at specific loci and therefore allow to discover infrequent alleles in highly expressed genes. Finally, pooling samples is already extensively used in DNA-seq (sometimes termed Pool-seq) (8). The main advantage of this method is that it clearly decreases costs, as library preparation for bar-coding is nowadays approximately the same price as sequencing. The drawback is that genotypes cannot be derived anymore. Instead, we have access to the allele frequency in the population, a result known as the allelotype. In this work, we present a method for the de novo identification, differential analysis and annotation of variants from RNAseq data in non-model species. It takes as input RNA-seq reads from at least two conditions (e.g. the modalities of the phenotype) with at least two replicates each, and outputs variants associated with the condition. The method does not require any reference genome, nor a database of SNPs. It can therefore be applied to any species for a very reasonable cost. We first evaluated our method using RNA-seq data from the human Geuvadis project (9). The great advantage of this dataset is that SNPs are well annotated, since the selected individuals were initially included in the 1000 genomes project (10). This enables to clarify what is the precision and recall of our method, and how it compares to methods which require a reference genome or a reference transcriptome.

We then applied our method in the context of non-model species. First we focused on *Asobara tabida*, an hymenoptera that exhibits contrasted phenotypes of dependence to its symbiont. Using RNA-seq data from two extreme modalities of the phenotype, we were able to establish a catalog of SNPs, stratify them by their impact on the protein sequence, and assess which SNPs had a significant change of allele frequency across modalities. We further selected cases for experimental validation, and were able to confirm that the SNPs were indeed condition specific. We then applied our method on two recently diverged *Drosophila* species, *D. arizonae* and *D. mojavensis*. These species can still produce hybrids that are sterile. In this case, our method identifies differences of 1 nt, which are not SNPs but divergences. On this system also, we were able to validate experimentally that the loci we identify were truly divergent.

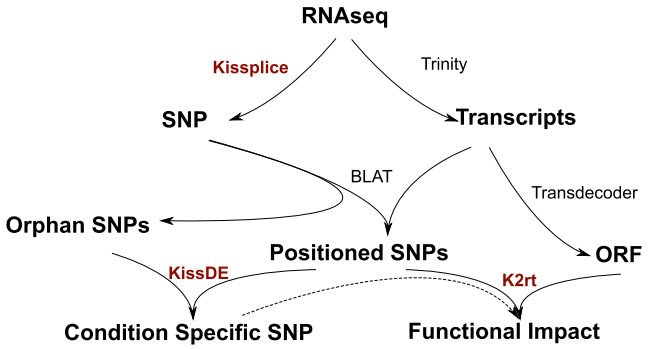


Figure 1. With fasta/fastq input from an RNA-seq experiment, SNPs are found by KISSPLICE without using a reference. As KISSPLICE provides only a local context around the SNPs, a reference can be built with TRINITY, and SNPs can be positioned on whole transcripts. Some SNPs that do not map on the transcripts of TRINITY, called orphan SNPs, are harder to study but can still be of interest. We propose a statistical method, called KISSDE, to find condition-specific SNPs (even if they are not positioned) out of all SNPs found. Finally, we can also predict the amino acid change for the positioned SNPs, and intersect these results with condition-specific SNPs using our package KISSPLICE2REFTRANSCRIPTOME (K2RT).

We outline that, even though the case studies presented in this paper include two replicates, the method can be applied to any number of replicates. Larger cohorts can be helpful to narrow down the list of SNPs likely to be really causal for the phenotype. Our key contribution is that we are able to produce a list of SNPs stratified by their impact on the protein sequence, and ranked by difference of expressed allele frequency across conditions. This list can be further mined for candidates to follow up experimentally.

All the methods presented in this paper are implemented in software that are freely available at http://kissplice.prabi.fr/TWAS. In particular, the statistical procedure that we developed is available through an R package, KISSDE, which is of general interest for researchers who have obtained read counts for pairs of variants in a set of conditions and wish to test if these counts reflect the specificity of the variant in a particular condition.

MATERIALS AND METHODS

Overview

We present here a collection of methods which can be used together to produce, from RNA-seq data alone, a list of condition-specific SNPs, stratified by their predicted impact on the protein. Figure 1 summarises the different steps.

TRINITY, TRANSDECODER and BLAT are third-party software. KISSPLICE was published recently (11), KISSDE and KISSPLICE2REFTRANSCRIPTOME (K2RT) are methods we introduce in this paper.

De novo identification of SNPs

KISSPLICE (11) is a software initially designed to find alternative splicing events (AS) from RNA-seq data, but which

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

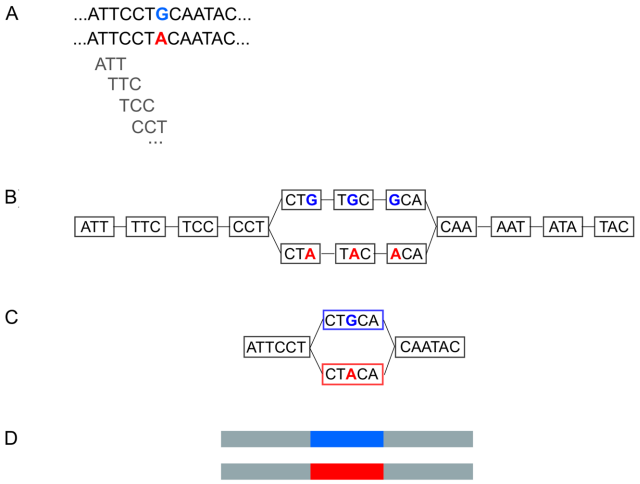


Figure 2. (**A**) A SNP present in two alleles in the data. (**B**) The de Bruijn Graph derived from the data. For the sake of simplicity of exposition, we draw here with $k = 3$. In practice, $k = 41$. (**C**) A compressed de Bruijn graph can be obtained by merging nodes with a single outgoing edge with nodes with a single incoming edge. This compression step is lossless. (**D**) The two paths in the compressed de Bruijn graph correspond to the two alleles of the SNP.

also outputs indels and SNPs. We present here its functionality for SNP detection. The key concept, initially introduced in Peterlongo *et al.* (12) and later used in Iqbal *et al.* and Uricaru *et al.* (1,2) is that a SNP corresponds to a recognisable pattern, called a *bubble*, in a de Bruijn graph (DBG) built from the reads. De Bruijn graphs are widely used data structures in de novo assembly (13–15), as they are well tailored for large amounts of short reads. In our case, DBGs are especially appealing because they model explicitly each nucleotide, a required feature to capture SNPs. The nodes of the graph are words of length k, called k-mers. There is an edge between two nodes if the suffix of length $k-1$ of the first k-mer is identical to the prefix of length $k-1$ of the second k-mer. The DBG that is built from two alleles of a locus will therefore correspond to a pair of vertex-disjoint paths in the graph, which form the bubble. Unlike AS events and indels, bubbles generated by SNPs have two paths of equal length (Figure 2B). Linear paths of the DBG can be further compressed in a single node without loss of information (Figure 2C).

In the special case where there are two SNPs located less than k nt apart on the genome, they will be reported in the same bubble (Supplementary Figure S1). In the case where the two SNPs are perfectly linked, a single bubble is reported. If they are partially linked, each haplotype will correspond to a path, and KISSPLICE will report all pairs of paths. In this case, the number of bubbles does not correspond to the number of SNPs, but to the number of pairs of observed haplotypes. Supplementary Figure S2 illustrates the case of two SNPs and four haplotypes.

KISSPLICE consists in essentially three steps: (i) building the DBG from the RNA-seq reads; (ii) enumerating all bubbles in this graph and (iii) mapping the reads to each path of each bubble to quantify the frequency of each variant. Particular attention was paid to both the memory (16,17) and time (18) requirements of the pipeline. KISSPLICE was able to process 200M reads of 2×75 nt in 20 hours, with less than 16GB of RAM.

Filtering out sequencing errors and inexact repeats

SNPs correspond to bubbles in the de Bruijn graph derived from the reads. However, not all bubbles in the DBG correspond to SNPs. Essentially two types of false positives can be found: sequencing errors and inexact repeats. RNA editing sites may also be mistaken for SNPs but in practice, these correspond to a few cases only, that we discuss in the Results section.

Sequencing errors may generate bubbles in the DBG. A distinctive feature that helps to discriminate them from true variants is that one path of the bubble is expected to be poorly covered. In practice, a common way to filter out sequencing errors when dealing with DNA-seq data is to remove all rare k-mers (seen less than a given number of times) prior to the DBG construction. This simple strategy, implemented for instance in DISCOSNP, is however not sufficient when dealing with RNA-seq data. Since the coverage depends on gene expression, it is therefore very unequal across genes, and the cut-off should be adapted to each gene. To account for this constraint, we introduced a relative cut-off, which enables to remove edges in the DBG that are supported by less than a percentage of all counts outgoing from (or incoming to) the same node. This enables to remove sequencing errors even in highly expressed genes (Figure 3). Clearly, the drawback of these cut-off strategies is that rare variants will be filtered out because they will be mistaken for sequencing errors. Our ability to detect rare variants is therefore limited by this critical parameter. We set the cut-off to 5%. This cut-off corresponds to a good trade-off between precision and recall (Supplementary Figure S3).

Inexact genomic repeats may also generate bubbles in the DBG (Figure 4). This is the case for instance for recently diverged paralogs which still share a lot of sequence similarity and hence may differ locally by one nucleotide flanked by k conserved nucleotides. This is also the case for other types of repeats, including inexact tandem repeats or transposable elements which may be present in the UTRs and introns of genes. In principle, introns are not present in RNA-seq data, but in practice, whatever the protocol used to filter out pre-mRNA, a proportion of at least 5% remains (19).

The question of discriminating SNPs from inexact repeats has already been addressed in the literature in the case of unpooled data. Romiguier *et al.* (5) propose to use the idea that loci corresponding to recently diverged paralogs should present an excess of heterozygous sites. This idea cannot be employed in our case since we want our method to be able to deal with pooled data, where we cannot genotype individuals.

Repeats present in a large number of copies (like transposable elements, or large families of paralog genes) generate a large number of bubbles which are false positives. However, these bubbles have a specific feature that we can use to discriminate them from the others: they are branching (Figure 4). The more (inexact) copies in the repeat family, the higher the number of branches in each bubble. In order to filter them out, we introduced a parameter b, which corresponds to the maximum number of branches allowed.

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

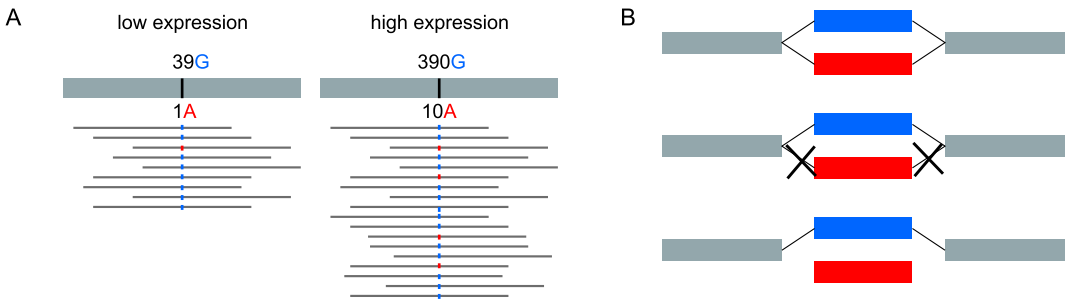


Figure 3. Sequencing errors and rare variants generate bubbles in DBGs with very unbalanced path coverage. (**A**) For ease of exposition of the concept, we represent here the reads mapping to a reference genome. Applying an absolute cutoff would remove the sequencing error for a poorly expressed gene, but not for a highly expressed gene. (**B**) Applying a relative cutoff of 5% in the DBG removes one or two edges from the red path and hence prevents this bubble from being found.

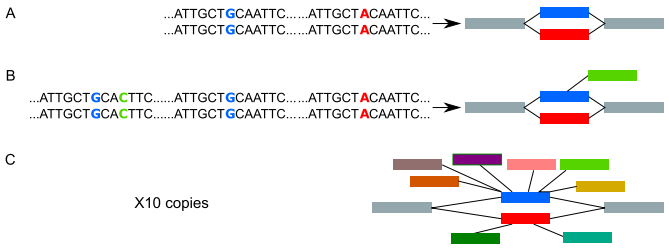


Figure 4. Two inexact repeats give rise to a pattern in the DBG that resembles a SNP (**A**). Very often, repeats are present in more than two copies (**B**) and therefore generate branching bubbles. Bubbles with more than five branches (**C**) are filtered out.

If one path of the bubble has more than b branches, then the bubble is filtered out. In practice, we set this parameter to 5, which appeared to be a good trade-off between recall and precision as shown in Supplementary Figure S3.

Repeats present in a small number of copies are not filtered out by this criterion. Some can be filtered by focusing on bubbles whose path length is strictly $2k + 1$, not larger. We found that this simple strategy was efficient and we used it in this work. It can however be modified in KISSPLICEwith the s parameter, which we recommend if the purpose is to find multiple SNPs. In any case, most inexact repeats are actually filtered out at the next step of the pipeline, when we test for the enrichment of one variant in one condition (as described in the Statistical analysis section). Indeed, most repeats do not have expression levels that are condition-specific. The ones that are *not* filtered out at this step correspond to paralogous genes, where one copy is more expressed in the first condition and the second copy is more expressed in the other condition. Although these are not SNPs, we can argue that they are still relevant candidates for an association study aiming at proposing causes for the difference of phenotype.

Predicting the impact of SNPs on the protein sequence

KISSPLICE predicts SNPs, but outputs only a very local context around the SNP. In order to predict the amino acid change it causes, if any, we need to place the SNP in a larger genomic context. For this, we relied on a widely used global transcriptome assembler: TRINITY (15), which takes as input RNA-seq reads and outputs contigs that correspond to either full-length transcripts (if the expression level of the transcript is sufficient) or to fragments of transcripts. The results of KISSPLICE were aligned onto the transcripts predicted by TRINITY using BLAT (20). Concurrently, we Fdsearched for coding potential in the transcripts using TRANSDECODER. Once we had the location of the SNP within the transcript and the location of the open reading frame (ORF), we could assess if the SNP was located within the CDS or not, and if so, if it was a synonymous or non synonymous SNP. In the case where no ORF was predicted for the transcript, we concluded that the SNP was within a non coding region. In practice, this can correspond to a non coding RNA, a UTR or an intron. Prediction of the amino acid change of a SNP was included in a Python package, called KISSPLICE2REFTRANSCRIPTOME (K2RT), which takes as input a set of predicted ORFs (bed format), the output of KISSPLICE (fasta format), and a mapping of the results of KISSPLICE to the transcripts (psl format). Importantly, TRINITY, TRANSDECODER and BLAT are third party software which can be replaced by others, provided the exchange formats are respected (bed and psl).

In the case where a SNP mapped to several TRINITY transcripts, we reported the amino acid change of the SNP in each transcript. This happened in particular when a SNP was located in a constitutive exon of a gene that gave rise to multiple alternative transcripts through alternative splicing. We further show in the Results Section that our ability to call SNPs both in constitutive exons and alternative exons is a strong advantage of our method against others that first map the reads to the assembled transcriptome and then call SNPs using a genotyper.

In the case where a SNP mapped to no transcript, then it could not be treated by K2RT and it was filtered out. Those SNPs were called orphan SNPs. They were mostly located in poorly expressed genes and/or highly repeated regions. Indeed, repeated regions are notoriously difficult to assemble. When repeated regions are located within genes, they may either generate chimeric transcripts in the assembly if the assembler is too permissive, or a series of truncated short contigs if the assembler is too conservative. By default, TRINITY does not output contigs shorter than 200 nucleotides. Because these contigs are highly enriched in repeats and poorly expressed genes, it explains the origin of the majority of our orphan SNPs.

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

As mentioned in the model section, the number of bubbles does not always correspond to the number of SNPs. In the case of SNPs located less than k nucleotides apart, the number of bubbles corresponds to the number of pairs of haplotypes out of the total number of haplotypes. The same SNP may therefore be present in multiple bubbles. When mapping the bubbles to a reference transcriptome, it is possible to remove this redundancy and count the true number of SNPs. Indeed, if two bubbles map to the same transcript at the same location, then it means that they refer to the same SNP, and we count it only once.

The software versions that we used were: TRINITY r20140717, TRANSDECODER v2.0.1, BLATSUITE36, KISSPLICE v2.4, KISSPLICE2REFTRANSCRIPTOME v1.0.

All were used with default parameters. We set the minimum query coverage to 90% in K2RT. Changing this from 70% to 90% only marginally affected our results.

A critical parameter in de novo assembly is the k-mer size. In TRINITY, this value is set to 25 and cannot be modified. In KISSPLICE the default value is 41 as we found it is a good compromise between recall and precision. We also tested 25 and this resulted in an increase of 10% in recall but a decrease of 10% in precision (Supplementary Figure S3). For advanced users interested in obtaining a more exhaustive list of candidates (hence optimising recall), we recommend to decrease the value of k in KISSPLICE.

Statistical analysis

Testing the association between a variant and a condition. Given the number of SNPs (n) and the number of replicates (m), our data set is a count matrix of size $2n \times m$, with two lines corresponding to one SNP (upper and lower path representing the two different alleles with one nucleotide differing between both paths). For each individual, we aimed to compare read counts per allele and per condition. As we worked with biological replicates, several sources of variance were added and the variance parameter of the Poisson distribution was in general not flexible enough to describe the data (21,22). Hence, our statistical analysis adopted the framework of count regression with Negative Binomial distribution.

We considered a two-way design with interaction, with *alleles* and *experimental conditions* as main effects. Following the Generalized Linear Model framework, the expected intensity of the signal was denoted by λ_{ijk} and was decomposed as:

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

where μ is the local mean expression of the transcript that contains the SNP, α_i the effect of allele i on the expression, β_j the contribution of condition j to the total expression, and $(\alpha\beta)_{ij}$ the interaction term. In order to properly model the variability of the data that are characterised by overdispersion (as in any RNASeq data (21,22)), we considered the Negative Binomial distribution. In this setting, Y_{ijk} denotes the counts of a sample k with allele i in condition j. We assume that:

$$Y_{ijk} \sim NB(\lambda_{ijk}, v_{ijk}),$$

with λ_{ijk} defined as above. With this model, the variance of the observations becomes:

$$v_{ijk} = \lambda_{ijk} + \phi \times \lambda_{ijk}^2,$$

with ϕ the over-dispersion, which is the excess of variance seen in the data in comparison to a Poisson distribution.

Due to numerical instabilities associated with the estimation of Negative Binomial parameters, we adopted a model selection approach to determine which model was best suited to handle the over-dispersion parameter ϕ. Our strategy was first to estimate a model without over-dispersion using the GLMNET package (model $\mathcal{M}(\phi = 0)$). We then considered two different estimation methods for the parameter ϕ, namely a global estimation approach using the package AOD (model $\mathcal{M}(\phi = \phi_{\text{global}})$), and a SNP-specific parameter using the DSS package (model $\mathcal{M}(\phi = \phi^i_{\text{DSS}})$. We used a BIC to choose the best model out of the three. Before comparing the allele read counts from different libraries, the count data were normalised by library sizes as proposed in the DESEQ package (23). This software has been shown to be the most efficient according to a recent normalisation comparison study (24). Pseudo-counts (*i.e.*, systematic random allocation of ones) were considered for SNPs showing many zeros to avoid singular hessian matrices while fitting the generalised linear model. Some events were then filtered out based on their counts: if global counts (for all replicates and all conditions) for both variants were too low (less than 10 counts), we considered that we did not have enough power to conclude on this event and we did not test it.

We then performed the core test on the association between variant and condition. The target hypothesis was H_0: $\{(\alpha\beta)_{ij} = 0\}$, i.e. no interaction between the allele and the condition. If this interaction term is not null, a differential usage of an allele across conditions occurred. The test was performed using a Likelihood Ratio Test with one degree of freedom, which corresponds to the supplementary interaction parameter that is included in the second model and not in the first (25). To account for multiple testing, p-values were adjusted with a 5% false discovery rate (FDR) following a Benjamini–Hochberg procedure (26).

Quantifying the magnitude of the effect. When a variant is found to be differentially represented in two populations, one remaining difficulty is to quantify the magnitude of this effect. Indeed, significant ($P < 0.05$) but weak effects are often detected, especially in RNA-seq data in which some genes are very highly expressed (and hence have very high read counts).

A natural measure for quantifying the magnitude of the effect would be the difference of allele frequencies between the two conditions. In practice, the true difference of allele frequencies is not known, and we estimated it using the RNA-seq counts. The precision of this estimation is discussed in the Results Section.

We denote by f_e the estimation of the allele frequency based on RNA-seq counts:

$$f_e = \frac{\#counts_variant_1}{\#counts_variant_1 + \#counts_variant_2}.$$

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

The value of f_e was computed for each replicate of each condition. We then took the mean of these values for all replicates within each condition. Finally, we calculated the difference across conditions and obtained the magnitude of the effect: $Df_e = f_{e_{cond1}} - f_{e_{cond2}}$. In the special case where the two variants had low counts (less than 10) within one replicate, then f_e was not calculated. Finally, if at least half of the replicates of one condition had low counts, Df_e was not computed either. Overall, this prevented from over-interpreting large magnitudes obtained from low counts.

Our method is embedded and distributed in an R package, called KISSDE, which can take as input either the output file of KISSPLICE or any count matrix with two lines representing an event.

Methodology for testing and validating our approach

We first evaluated our method in human, because it is a species for which a reference genome is available and SNPs are well annotated. We then used our method on a non-model species: *Asobara tabida*, an hymenoptera that exhibits contrasted phenotypes and for which no reference genome is available. Finally, we applied our method on a different evolutionary timescale, working on two recently diverged *Drosophila* species, *D. mojavensis* and *D. arizonae*, for which a draft reference genome is available only for *D. mojavensis*.

The Geuvadis dataset. Our method enables to find SNPs from RNA-seq data. In order to assess if the SNPs we find are correct, and if the list we output is exhaustive, we chose to test our method on RNA-seq data from the Geuvadis project. Indeed, the individuals whose transcriptome was sequenced in this project were already included in the 1000 genome project. Hence, their SNPs have already been well annotated. We downloaded fastq files from SRA (see Data access) and selected 10 Toscans and 10 Central Europeans. We sampled 10M reads for each individual and concatenated the fastq files in pools of five individuals.

Definition of the set of true SNPs and their genotypes. We downloaded the vcf file from the 1000G webpage. For each SNP called in the 1000 Genomes project, we had at our disposal the genotype of each individual. We focused on the genotypes of the 20 individuals selected for our analysis. Whenever only one allele was represented in the 20 individuals, we filtered out this SNP, as it simply cannot be discovered based on these 20 individuals only.

Whenever one SNP was covered by less than 5 reads out of the total number of reads in the 20 individuals, we considered that the SNP was located in a too poorly expressed region and could not be discovered by RNA-seq. Other levels of poorly/medium/highly expressed regions are discussed in the Results section. The read coverage was computed using SAMTOOLS depth, on the .sam file obtained after mapping the reads with STAR (v2.3.0) (27).

Calling SNPs from reads mapped to a reference genome: GATK-GENOME. In order to clarify if the performances of our method were on par with other methods, we chose to benchmark against GATK, which is the most widely used method for variant calling in eukaryote samples when a reference genome is available.

We employed the GATK Best Practices workflow for SNP and indel calling on RNA-seq data (https://www.broadinstitute.orggatkguidearticle?id=3891 posted on 6 March 2014, last updated on 31 October 2014) which considers the following steps: (i) mapping to the reference genome with the STAR aligner, 2-pass method (28) with the suggested parameters allowing to obtain the best sensitivity for the variant call task, where during the second pass of STAR a new reference index is created from the splice junction information determined during the first step alignment and a new alignment step is done with the new index reference; (ii) adding read group information, sorting, marking duplicates and indexing, using Picard's tools; (iii) splitting reads into exon segments (removing Ns but maintaining grouping information) and hardclipping sequences overhanging into the intronic regions, using the SplitNCigarReads GATK tool; (iv) realigning indels and recalibrating Base quality; (v) calling variant with HAPLOTYPECALLER, and finally filterimg the variants with VARIANTFILTRATION.

Calling SNPs from reads mapped to a reference transcriptome MPILEUP-TRANSCRIPTOME. The reference transcriptome was assembled using TRINITY (as described previously) and reads were mapped to this reference using BOWTIE2 (29). We then used MPILEUP and BCFTOOLS (30) to call SNPs from the mapped reads. TRINITY, BOWTIE2 and MPILEUP were used with default parameters. BCFTOOLS was used with the options –multiallelic-caller and –variants-only.

As outlined in the Results section, this pipeline performs poorly in the context of alternative splicing, as it misses most of the SNPs located in exons shared by several transcripts.

A way to deal with this issue is to filter the redundancy caused by alternative splicing. The first approach we considered was described in Pante *et al.* (31) and consists in applying CD-HIT (7), a widely used greedy clustering method, to the transcriptome assembled by TRINITY. The second approach we considered was described in Van Belleghem *et al.*, 2012 (6) and consists in keeping only the longest isoform for each gene assembled by TRINITY.

In both cases, we obtained a filtered transcriptome, with reduced redundancy, and we then used BOWTIE2, MPILEUP and BCFTOOLS to call SNPs.

Comparison of genome-based and transcriptome-based approaches. In order to compare the SNPs predicted by KISSPLICE with our set of true SNPs, we needed to obtain a genomic position for each of our predictions. To this purpose, we aligned each variant of each bubble to the reference genome using STAR (v2.3.0). In the case where a variant mapped to several locations, we used the default behaviour of STAR, which is to assign the variant to the location with the fewer number of mismatches. In case of ties, we kept all equally good locations, and if at least one of the possible locations corresponded to an annotated SNP, we considered that the prediction of KISSPLICE was correct.

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

For MPILEUP, we aligned the transcripts assembled by TRINITY on the reference genome with BLAT.

Asobara tabida *lines, RNA sequencing and SNP verification.* *Asobara tabida* (Hymenoptera: Braconidae) is a parasitoid species which develops on *Drosophila* hosts. *A. tabida* is naturally infected by three strains of *Wolbachia*, among which one (*w*Atab3) is necessary for oogenesis completion (32,33). However, when *Wolbachia* are removed by antibiotic treatment, the degree of oogenetic defect exhibits genetic variation within populations (34). We thus founded two lineages of *A. tabida* from a natural population (Sainte Foyles-Lyon, France) based on their extreme phenotype after elimination of *Wolbachia*: the SFR2 lineage whose females do not produce any eggs and the SFR3 lineage whose females produce half the normal content of eggs. In both cases, dependence is complete as the eggs produced are sterile. These two lineages were founded by three females and were kept for 15 generations (three founders at each generation) before RNA extraction.

The experimental design for RNA-seq sequencing aimed at describing the transcriptomic changes associated with the presence / absence of *Wolbachia*, and the variations observed in the two *A. tabida* lineages exhibiting an extreme phenotype. To this purpose, cDNA libraries were constructed from infected and non-infected ovaries in these two lineages. Because these RNA-seq data were issued from two distinct lineages from a non-model species, we exploited this dataset to validate the method developed here and to discover biologically relevant SNPs, using libraries obtained from infected ovaries. The samples used for RNA extraction were young female (0–1 day old) ovaries dissected in a drop of A-buffer (two replicates of 30 ovaries per lineage). RNA was extracted as described in Kremer *et al.* (35). These RNA extracts were used to generate corresponding cDNA libraries, following the recommendations given by the manufacturer of the SMARTer PCR cDNA synthesis and BD Advantage two PCR kits (Clontech). These cDNA libraries were then purified with the Qiaquick kit (Qiagen) and their quality checked. Sequencing of cDNA was performed by the Genoscope (Evry), on an Illumina GA-IIx instrument, to obtain 1x75bp reads. These data were trimmed using the ShortRead package with default parameters and then used as input of the pipeline defined in Figure 1.

Based on these results, 34 SNPS were chosen for verification. For each SNP, primers were designed on the corresponding transcript to amplify the surrounding genomic region. PCRs were performed from an aliquot of the purified cDNA libraries. The reaction was performed in a total volume of 25 μl, and the mixture consisted in 2.5 μl of 5× green DreamTaq mastermix, 200 nM of dNTP, forward and reverse primers (see Supplementary Table S1 for primer sequences), and 5U of DreamTaq DNA polymerase (ThermoFisher). PCR amplification was performed on a Tetrad thermocycler (Biorad) as follows: 2 min at 94°C, 35 times (30 s at 94°C, 30 s at 58°C, 30s at 72°C), and 10 min at 72°C. The PCR products were sequenced using the Sanger method from forward and reverse primers by the Biofidal company. The sequences were aligned and their respective chromatograms analysed by the CLC Main workbench.

Drosophila *strains, RNA sequencing and SNP verification.* *D. mojavensis* and *D. arizonae* are two Drosophila species that are endemic of the arid southwestern United States and Mexico. These species diverged recently (less than 1 MYA) (36,37). In the laboratory, hybridisation of these two species is possible while in nature it does not occur (or is very rare). The ovarian transcriptome of these two species (and their reciprocal crosses) was sequenced to investigate the first step of hybrid incompatibility and look for deregulated genes in hybrids. In this paper, we did not study the transcriptomes of the hybrids, we only used the transcriptomes of the parents to test for the validity of our pipeline at a different evolutionary scale. The sequenced strains were *Drosophila mojavensis* from the Anza Borrego Desert, CA (stock number: 15081–1352.01) and *Drosophila arizonae*, from Metztitlan – Hidalgo, Mexico (stock number: 15081–1271.17), both obtained in the US San Diego Drosophila Stock Center. Virgin female flies were collected after hatching and isolated until they reached ten days. The RNA was extracted from a pool of 30 ovaries of 10-days-old flies for each line. The extractions were performed using the RNeasy kit (Qiagen) and samples were then treated with DNase (DNA-free Kit, Ambion) and stored at −80°C. The samples were quantified by fluorescence in the Bioanalyser 2100 (Agilent), according to pre-established criteria by the sequencing platform. For each line, the extracted RNA was divided into two parts in order to generate two cDNA libraries (two replicates per condition). RNA was sequenced by Illumina Technology, in the IlluminaHiseq 2000. We sequenced 2 × 51 bp paired-end reads and the medium size of the inserts was 300 bp. We used URQT (38) with the default parameters to remove the low quality bases and the polyA tail from the dataset before running the pipeline described in Figure 1. The protocol for SNP verification is identical to the one used for *Asobara tabida* (see Supplementary Table S2 for primer sequences).

Data access

The human data used in this study can be found through the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) under the accession number E-GEOD-29342 and we used the individuals named NA20808, NA20809, NA20810, NA20811, NA20812, NA20813, NA20814, NA20815, NA20819, NA20826, NA06984, NA11840, NA06986, NA06989, NA06994, NA07346, NA07357, NA10851, NA11829 and NA11832.

The RNAseq libraries from *D. mojavensis* and *D. arizonae* are available through the NCBI Sequence Read Archive (SRA : http://www.ncbi.nlm.nih.gov/sra) under the accession no. SRX1272419 and SRX1277353.

The *A. tabida* dataset is available through the NCBI Sequence Read Archive (SRA : http://www.ncbi.nlm.nih.gov/sra) under the accession no. SRX1701817, SRX1701824, SRX1701826 and SRX1701855.

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

RESULTS

Validation of the SNP calling method using available data from a model species

Identification of variants. In order to evaluate the performance of our method, we needed to test it in the case where we knew which SNPs should be found. We thus focused on a dataset from human in which SNPs were already annotated. We selected two populations (Toscans and Central Europeans) from the Geuvadis project (39), and downloaded the RNA-seq data of 10 individuals in each population. We sampled 10M reads from each individual and pooled individuals 5 × 5, to obtain two replicates of five pooled individuals per population. We ran KISSPLICE and TRINITY on these four read sets and we aligned the variants of KISSPLICE to the TRINITY transcripts using BLAT (with at least 90% query coverage and 90% identity). Out of the 64824 bubbles initially found by KISSPLICE, 53494 (82%) mapped to TRINITY-assembled transcripts, 8024 partially aligned, and 3306 did not align. As explained in the Methods Section, SNPs located near other SNPs may be enumerated more than once, but with different contexts (see Supplementary Figure S2). After removing this redundancy, we ended up with 51,235 bubbles.

To assess whether these bubbles were true SNPs, we first aligned the sequences of the variants (i.e. each path of the bubble) to the human reference genome and compared their genomic positions to a set of SNPs downloaded from the 1000 genome project webpage. We also benchmarked our method against two software: GATK, a widely used method to call SNPs in the presence of a reference genome and MPILEUP, part of the SAMTOOLS/BCFTOOLS, used here to call SNPs on the transcriptome assembled by TRINITY using the same RNAseq data.

GATK was run with parameters recommended from the GATK web page for RNA-seq data. MPILEUP was run on top of BOWTIE2, both on the transcriptome assembled by TRINITY (MP-TRANSCRIPTOME), and on the reduced transcriptome. In the latter case, we either kept the longest isoform for each gene (MP-LONG-TRANS) as described in Van Belleghem *et al.* (6), or we applied CD-HIT to cluster similar isoforms (MP-CD-HIT) as described in Pante *et al.* (31).

For each method, we calculated the Precision, i.e. the number of true SNPs out of the total number of predicted SNPs, and the Recall, i.e. the number of predicted SNPs out of the total number of true SNPs.

As outlined in Figure 5, the recall of all methods is extremely low if no filter is applied to the set of true SNPs (True SNPs minimum coverage set to 0). This is an expected result, because true SNPs were identified using DNA-seq data and recovering them using RNA-seq data requires that they are located in sufficiently expressed regions. The higher the expression, the higher the recall of all methods. For SNPs located in regions covered by at least 100 reads, the best recall is reached for GATK-GENOME (42%), which is better than KISSPLICE (35%) and MP-TRANSCRIPTOME (28%). The low recall of MP-TRANSCRIPTOME is essentially due to its poor ability to find SNPs in constitutive exons, a limitation which can be adressed using MP-LONG-TRANS (but not MP-CD-HIT). The recall of KISSPLICE can also be improved by modifying its relative threshold parameter from 5% to 2%. Interestingly, it even slightly outperforms GATK-GENOME. The reason is that KISSPLICE finds more SNPs located in repeated regions of the genome, while GATK filters them out based on their low mapping quality. Finally, we show that a large number of SNPs are still not found by any method. The majority of those are rare alleles (Supplementary Figure S4) and the remaining are SNPs located in repeated regions or very polymorphic genes, like immune genes.

As outlined in Figure 5, with the exception of KISSPLICE, the precision of all methods was very poor if no filter was applied on the number of reads supporting each prediction. This is an interesting advantage of KISSPLICE. Its predictions can be taken as is, and the precision will already be 80%. If we now focus on predicted SNPs supported by at least 100 reads, then GATK-GENOME was the best and reaches a precision of almost 90%, while MP-TRANSCRIPTOME was the worst with a precision of 70%.

The false positives of all methods can essentially be divided into two categories: sequencing errors, and inexact repeats. The impact of RNA editing was minor (less than 5% of cases were annotated in RADAR v2 (40)).

Filtering out SNPs supported by few reads effectively deals with the issue of sequencing errors, but this consequently affects the ability to find true SNPs in poorly expressed regions.

The issue of inexact repeats affects mostly transcriptome-based methods, not genome-based methods. While KISSPLICE partially deals with this issue with the branching parameter and the filtering of long bubbles, MP-TRANSCRIPTOME does not address this problem.

Overall, we conclude that, although we do not use a reference genome, the recall and precision of our method are comparable to those which use one, such as GATK. Furthermore, we show that our method has a better ability to call SNPs in the context of alternative splicing and a more efficient way to filter out inexact repeats than methods which call SNPs after mapping reads to an assembled transcriptome.

Quantification of variants and statistical differential analysis. The quantification we obtain for variants called from pooled RNA-seq data reflects both the allele frequency of the variant in the pool and the expression level of the gene. An 'expressed' allele frequency can be derived from these counts, by simply taking the ratio, but the obtained frequency is expected to be distorted compared to the allele frequency estimated from DNA-seq data. Several causes may be listed. First, within a heterozygous individual, one allele may be more expressed than the other, a process known as Allele Specific Expression (ASE). Second, RNA expression from different individuals (hence possibly different genotypes) can be variable within a pool, thus distorting the allelotype. In order to evaluate the magnitude of this distortion, we computed within each pool the correlations between the true allelic frequencies, and the estimated allele frequencies. To obtain the true allelic frequency within a pool, we took advantage of the availability of the genotypes of each individual from the Geuvadis dataset, and we simply summed up the number of alternative alleles over the

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

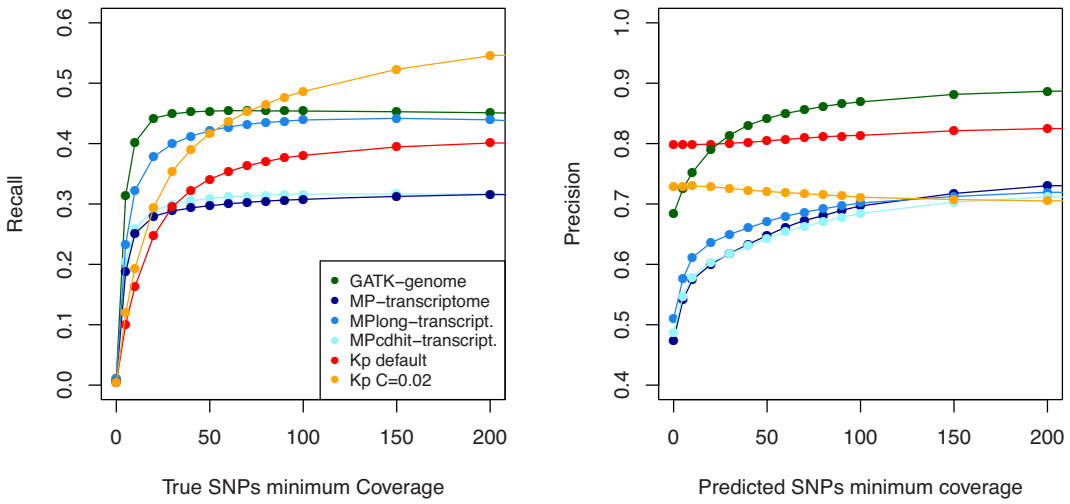


Figure 5. Precision and recall of KISSPLICE, GATK-GENOME, MP-transcriptome and MP-LONG-TRANS as a function of the expression level of the locus. For the recall, all predictions are taken into account, but the set of true SNPs is restricted to those covered by at least a given number of reads. For the precision, only SNPs supported by at least a given number of reads are taken into account.

total number of alleles within the pool. The expressed allele frequencies were obtained from KISSPLICE calls, summing the alternative allele counts of each individual over all allele counts of the pool.

We found that the distortion highly depends on the expression levels (Supplementary Figure S5). While the correlation was weak (0.65) for poorly expressed loci (less than 3 reads), it increased steadily with the expression level up to a plateau of 0.98. When we restricted to loci with at least 10 reads, the correlation reached 0.95.

We therefore conclude that, whenever a locus was sufficiently expressed (at least 10 reads), the expressed allele frequency was a good predictor of the true allele frequency.

If we now compute the difference of allele frequencies across conditions (denoted by *df*), and compare it to the difference of expressed allele frequencies across conditions (denoted by *dfe*), the correlations remain high, but are weaker, reaching a plateau of 80% for highly expressed loci. The reason is that most SNPs do not have a significant difference of allele frequencies across our two populations, hence these correlations are contaminated by SNPs with (almost) equal allele frequencies. In this case, the difference of allele frequencies is just a random fluctuation. When considering all SNPs, the correlation between *df* and *dfe* is significant but weak (Figure 6-A)

If we restrict to SNPs that are found as condition specific by KISSDE, then the correlation is much stronger (Figure 6B). Finally, if we restrict to SNPs covered by a total of at least 100 reads (an average of 25 reads per sample), then the correlation is again higher (Figure 6C). The more a gene is expressed, the higher the fit between *df* and *dfe*. A few SNPs ($n = 22$), however, exhibited a large difference between *df* and *dfe* (>0.3). A detailed analysis of these cases reveals that they are located in immune genes ($n = 5$), in genes showing a very variable expression across individuals ($n = 9$), or in genes exhibiting an allele specific expression ($n = 8$).

Overall, we conclude that, provided we restrict to condition specific SNPs, the metric we output with KISSDE for the difference of expressed allele frequencies, that is *dfe*, can largely be interpreted as a measure of the true difference of allele frequencies.

Prediction of the amino acid change. When no reference genome is available, it is not possible to obtain a genomic location for each SNP and therefore to apply SNPEFF (41), or POLYPHEN (42), which are widely used software for assessing the impact of a SNP on the protein sequence. In the absence of any reference genome, a reference transcriptome can nevertheless be obtained, using a full-length transcriptome assembler like TRINITY (15). Based on this transcriptome, it is possible to assess the coding potential of each transcript using TRANSDECODER, to position the predicted SNPs onto the assembled transcripts using BLAT (20), and finally to assess the impact of each SNP on its transcript(s). In the end, each positioned SNP is classified as coding or non coding. In the case where the SNP is located in the coding region, it is then classified as synonymous or non-synonymous (See Methods).

Out of 47,243 positioned SNPs (those which aligned to TRINITY transcripts), 14,804 cases (31%) fell in CDSs and the other 32,439% fell in non-coding regions (including UTRs). Among the ones falling in CDSs, we found that 53% (7788) were synonymous, while the other half (7016) were non synonymous.

To validate our predictions, we then intersected the genomic positions of our predicted SNPs with the genomic positions of SNPs in dbSNP, for which the impact on the protein sequence is known. Out of the 47,243 SNPs we predict, 39313 could be assigned a genomic position which matched a SNP annotated in dbSNP. Out of those 39313 cases, 2725 have no functional annotation in dbSNP, 35,141 had a correct prediction and 1447 cases wrongly predicted. A thorough examination of the 1447 cases wrongly predicted reveals that in most cases, the transcript predicted by TRINITY was very partial and was overlapping an intron (this happens when pre-mRNA is sampled together with mRNA at the RNA extraction step, despite selection

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

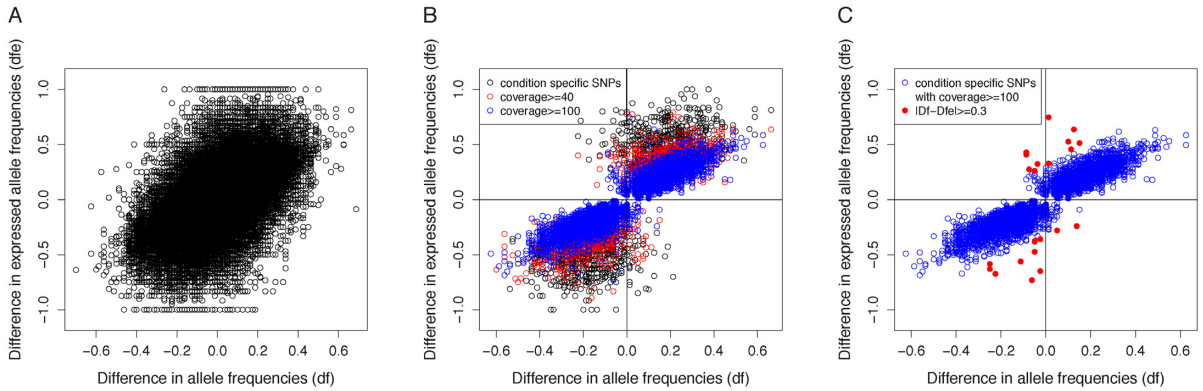


Figure 6. Difference of allele frequencies (df) Vs Difference of expressed allele frequencies (dfe). (**A**) All SNPs. (**B**) Condition-specific SNPs. (**C**) Conditions-specific SNPs covered by at least 100 reads.

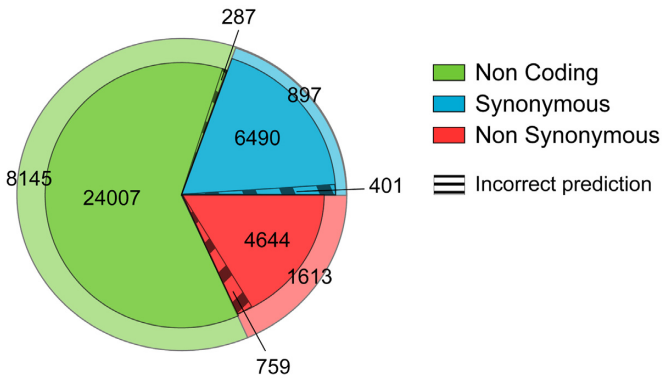


Figure 7. Results of KisSplice2RefTranscriptome The green, red and blue areas correspond respectively to non-coding, synonymous and non-synonymous SNPs. The dashed area corresponds to errors of our predictions of the impact on the protein sequence. The outer area corresponds to SNPs that are not in dbSNP or for which the prediction cannot be evaluated due to a lack of annotation in dbSNP.

of polyA+RNAs). In this case, the ORF predictor can over-predict coding regions, and our pipeline therefore tends to over-predict non synonymous cases. Figure 7 summarises our results for the prediction of the impact on the protein sequence. Overall, when SNPs can be evaluated, the precision of K2RT is 96% (35,141 out of 36,588).

Performance of the full pipeline. In the previous section, we evaluated our capacity to predict the impact on the protein independently of the remaining of our pipeline. We now turn to its evaluation within the full pipeline. Two situations can be discussed here. First, if only one experimental condition is considered, then no differential analysis is carried out. SNPs are identified and their impact on the protein is predicted. In this case, the prediction inherits from the errors made at the identification step. Out of 47,243 predicted SNPs, 39313 were in dbSNP and 35,141 had a correctly predicted impact. In the worst-case scenario, if we consider that the 7930 SNPs for which there was no dbSNP entry and the 2715 SNPs for which the dbSNP entry is incomplete were false positives, the precision of the pipeline was 74%. In practice, dbSNP is not exhaustive, and the true precision is between 74% and 96%. Second, if two conditions were considered (which is the original purpose of this study), then many of the false positives of the identification step were filtered out. Out of the 47,243 predicted SNPs, 5518 were condition-specific, and 5309 had a correct prediction of the impact on the protein sequence. Hence the precision, in the worst-case scenario, for condition-specific SNPs was 96% (5309 out of 5518).

Application of the method using biological data from species without any reference genome

From our study on the human dataset, we conclude that our method has a precision and recall similar to methods which require a reference genome. We now turn to the application of our method to non-model species.

Application to intraspecific polymorphism: the case of Asobara tabida. We first applied our method to *Asobara tabida*, for which RNA-seq data from two lineages (SFR2 and SFR3) were available. These lineages come from the same population, but they differ by their phenotype of dependence to their symbiont *Wolbachia*. In the absence of *Wolbachia*, SFR2 individuals produce no eggs, while SFR3 produce some. Consequently, we suspect a low but significant genetic differentiation between lineages that could be associated with the phenotypes, or to genetic drift associated with maintenance in the laboratory. While the experimental design, with a single lineage for each phenotype, does not enable us to separate between these two effects, we think that this dataset is still well tailored for a validation of our method because: (a) no reference genome is available for this species; (b) individuals were pooled for RNA extraction and (c) replicates are available for each lineage.

The transcriptomes of two replicates of pools of 30 individuals were sequenced through RNA-seq for each lineage, leading to 15M reads for each replicate. We ran our pipeline and found a total of 18609 positioned SNPs out of which 17,031 are condition-specific. The large proportion of condition-specific SNPs is largely due to the fact that most of them are fixed in at least one lineage. Indeed, 21% of them are fixed in both lineages, 63% are fixed in one lineage and polymorphic in the other, and 7% are polymorphic in both lineages (Supplementary Figure S6B).

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

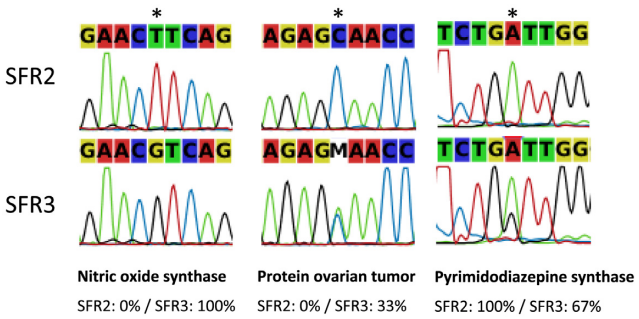


Figure 8. Three examples of SNPs validated by Sanger sequencing. The first is fixed in both the SFR2 and SFR3 lineages. The second and third are polymorphic in SFR3 but fixed in SFR2. In the third case, the base caller does not reflect the polymorphism but it can be seen from the chromatogram

Out of the 17,031 condition-specific variants, we found that 5608 (32%) were non coding, 6137 (36%) were synonymous and 3876 (22%) were non-synonymous.

Based on these results, we selected 27 cases for experimental validation: 10 were cases where the two lineages were fixed for a different nucleotide, 15 were cases where one lineage was fixed and the other polymorphic, 2 were cases where the two lineages were polymorphic. For all the 10 first cases, we were able to validate that the SNP was real and that the two lineages were indeed fixed for a different nucleotide (Supplementary Table S1, Figure 8). Out of the 17 remaining cases, we were able to validate that the SNP was real in all cases, but only in 9 cases were we able to validate that the site was polymorphic in one lineage (Supplementary Table S1, Figure 8). The rate of validation of the polymorphic status of the site within a lineage largely depended on the frequency of the minor allele (Supplementary Figure S5). Rare variants were harder to validate in terms of polymorphism detection. These rare variants could be false positives of our method, but they may also very well be true variants, not detectable experimentally using a direct sequencing technique without cloning. Importantly, although we could not always validate the fact that one site is polymorphic within a lineage, we systematically confirmed that the SNP was real, and that each lineage had a specific major allele. Therefore, we validated the condition-specificity of all SNPs.

As discussed earlier, our method outputs SNPs that are found by no other method. In order to test if these SNPs were true, we further tested specifically 7 such cases, and were able to validate all seven SNPs (Supplementary Table S1).

Because our RNA-seq data were initially obtained to compare the transcriptome of these two lineages, the design was not optimized for QTL analysis. In particular, each phenotype is represented by a single inbred genotype, making it difficult to separate the SNPs linked to the phenotype from those linked to drift. Despite this issue, we further characterised the impact on the protein sequence of the condition-specific SNPs. Among all these genes, some called our attention regarding their possible implication in the dependence phenotype. For instance, some genes, such as *Dorsal* and *Hypoxia up-regulated protein 1*, presented SNPs in their UTRs and were differentially expressed between lineages. These genes are involved in immunity and oxidative stress homeostasis, two functions that have been shown as particularly important in this biological system. Another example concerns genes involved in oogenesis, like *OTU-domain containing protein* or *Female sterile*, that exhibit non-synonymous SNPs in their CDS regions. These few examples show how the suite we propose in this paper rapidly allows to link the SNPs detected to their impact on the protein sequence, thus permitting to pinpoint candidate genes involved in phenotypic variation. Validation of these genes could involve either genetic studies (*e.g.*, knockdown experiments) and/or other linkage analyses targeted to these candidates.

Application to Interspecific Divergence: the case of Drosophila mojavensis *and* Drosophila arizonae. Similarly to the *Asobara* dataset, the drosophila dataset corresponds to non-model species, where individuals had to be pooled prior to RNA sequencing. In this case however, the two modalities of the phenotypes are not two populations of the same species, but two recently diverged species. This therefore enabled us to assess if our method also applies to a very different evolutionary scale, where differences of one nucleotide are no longer SNPs, but divergences. Additionally, the availability of the reference genome for *D. mojavensis* (and not *D. arizonae*) enabled us to study in depth the case of condition-specific inexact repeats.

D. mojavensis and *D. arizonae* are two closely related species that diverged 1MYA. We sequenced through RNA-seq the ovarian transcriptomes of two replicates of pools of 30 individuals for each species. We obtained 55M paired-end reads per replicate. We ran our pipeline on the data and obtained 51,730 positioned SNPs, and most of them (51,135) were condition-specific.

The condition-specific SNPs were mostly in coding regions (60%, i.e. 40,674 SNPs). We could classify 34,382 of them as synonymous, and the other 6292 SNPs as non-synonymous.

We selected 11 cases for experimental validation, six of which were divergent sites, and five were cases where the site was polymorphic in one species and fixed in the other. We were able to validate that the variation was condition-specific for all the divergent sites, and for four cases out of five for the polymorphic cases. Additionally, for two cases out of these four, we were able to amplify the two alleles in the species where the site was predicted to be polymorphic (Supplementary Table S2).

In most cases, an observed variation in the transcriptome is caused by the presence of two alleles at one locus. However, it is also possible that two mono-allelic loci, if they exhibit the same sequence except for one nt, generate a variation that resembles a SNP. In order to quantify this phenomenon, we explicitely selected in the results of KisSplice, the variations for which one path was mapping to one locus and the other path was mapping to another locus. This was only possible because we had at our disposal a draft genome of *D. mojavensis*. We selected explicitely cases where we knew that the variation we detected was potentially caused by two loci. There were only 224 cases like this,

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

which is very few compared to the total number of variations detected. We however tested three of them experimentally, and we were able to validate all of them. These cases are not true SNPs, but they correspond to recent paralog genes where one copy is more expressed in *D. arizonae*, and the other copy is more expressed in *D. mojavensis*.

CONCLUSION AND PERSPECTIVES

We present a method that can discover condition-specific SNPs from raw RNA-seq data. The individuals may be pooled, which decreases the costs of library preparation, while still enabling to allelotype and to find variants specific to one condition. As no reference genome is required, the range of applications of the method is very large. We first evaluated our method in human, where a reference genome is available and SNPs are extensively annotated. We show that our method has similar performances in terms of precision and recall, compared to GATK, a widely used mapping-based approach. We then evaluated our method on two non-model species.

In both cases, we were able to call variants, to classify them, and to discuss their impact. We selected a fraction of them for experimental validation through RT-PCR + Sanger sequencing. In all cases, we were able to validate that the variant was condition-specific. However, when the locus was predicted to be polymorphic in one condition, we were able to validate the presence of the two alleles only in cases where the minor allele frequency was at least 15%.

This work is a first approach toward transcriptome-wide association studies in non-model species. The method can readily be applied to RNA-seq data from any species, whenever two phenotypes are clearly identified and the goal is to find candidates for their genetic bases. In the case of continuous phenotypes, like height, the statistical framework can be generalised to quantitative trait loci (QTL).

This work focuses on SNP identification and analysis and does not address the question of the experimental design of a transcriptome-wide association study. A systematic evaluation of the optimal design is beyond the scope of this paper, but we would like to provide here briefly some basic advice.

First, in all the case studies presented here, we considered only two replicates, which is the minimum required by our method. We clearly advise that for a pre-determined cost, it is wiser to have a low coverage for each replicate, but to increase the number of replicates. Second, the type of replicates to choose is probably a more central issue. In the case of *Asobara*, we sequenced two biological replicates, but both replicates were derived from the same lineage. Having replicates when extracting RNA is useful, but not as useful as replicates at the line-establishment step. Only this type of replicate can allow to discriminate between SNPs in the original population and genetic drift in the lab. Finally, if pooling is envisioned, the number of individuals per pool should be as large as possible, especially for very polymorphic species. The larger the pool, the more representative of the population it is.

From the point of view of our method itself, there is of course also room for improvement. In particular, we found that, while easy SNPs are identified by all methods, a large amount of difficult SNPs are currently being overseen. This is the case of SNPs located in repeated regions of the genome, and that are notoriously difficult to annotate. SNPs located very close to each other are also challenging to annotate. Without a reference genome, we found that they are particularly difficult to tell apart from inexact repeats. Finally, SNPs located within very polymorphic regions of the genome, like immune genes, are also very challenging, even for mapping-based approaches. The use of a single reference genome is clearly limiting. De novo assembly methods are a promising direction for these, but still need to be optimised.

For future work, we see two lines of research, which could ultimately be combined. First, we could take advantage of the availability of long reads coming from third generation sequencing platforms (Pacbio, Minion). In principle, long reads have the potential to solve most of the issues we mentioned, but currently, the error rates are too high (10–15%) and the sequencing depth is not sufficient to apply to RNA-seq. In the meantime, it seems still relevant to keep on working in the context of short reads, but we think that the best resolution we can achieve for the prediction of difficult SNPs is not well captured by sequences. Graphs could instead well represent close SNPs and a partial quantification of their phasing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was performed using the computing facilities of the CC LBBE/PRABI.
The authors would like to thank Marie-Christine Carpentier and Delphine Charif for help on the analysis of the *Asobara* dataset; Gustavo Sacomoto for help on the analysis of the Geuvadis dataset; Sebastien Deraison for help on experimental validation of the candidates for the Drosophila dataset, and Vincent Miele and Alice Julien-Laferrière for help on developing KISSPLICE and KISSDE.

FUNDING

Agence Nationale de la Recherche [ANR-12-BS02-0008, ANR-11-BINF-0001-06, ANR-2010-BLAN-170101]; São Paulo Research Foundation – FAPESP/Brazil [2010/10731-4 to C.M..C.]; European Research Council under the European Community's Seventh Framework Programme [FP7 /2007–2013)/ERC Grant Agreement No. [247073]10]. Funding for open access charge: INRIA.
Conflict of interest statement. None declared.

REFERENCES

1. Iqbal,Z., Caccamo,M., Turner,I., Flicek,P. and McVean,G. (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.*, **44**, 226–232.
2. Uricaru,R., Rizk,G., Lacroix,V., Quillery,E., Plantard,O., Chikhi,R., Lemaitre,C. and Peterlongo,P. (2015) Reference-free detection of isolated SNPs. *Nucleic Acids Res.*, **43**, e11.
3. Leggett,R.M., Ramirez-Gonzalez,R.H., Verweij,W., Kawashima,C.G., Iqbal,Z., Jones,J.D.G., Caccamo,M. and Maclean,D. (2013) Identifying and classifying trait linked

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016

polymorphisms in non-reference species by walking coloured de bruijn graphs. *PLoS One*, **8**, e60058.
4. Piskol,R., Ramaswami,G. and Li,J.B. (2013) Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.*, **93**, 641–651.
5. Romiguier,J., Gayral,P., Ballenghien,M., Bernard,A., Cahais,V., Chenuil,A., Chiari,Y., Dernat,R., Duret,L., Faivre,N. *et al.* (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, **515**, 261–263.
6. Van Belleghem,S.M., Roelofs,D., Van Houdt,J. and Hendrickx,F. (2012) De novo transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle Pogonus chalceus (Coleoptera, Carabidae). *PLoS One*, **7**, e42605.
7. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
8. Schlötterer,C., Tobler,R., Kofler,R. and Nolte,V. (2014) Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.*, **15**, 749–763.
9. Lappalainen,T., Sammeth,M., Friedländer,M.R., 't Hoen,P.A.C., Monlong,J., Rivas,M.A., Gonzàlez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
10. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
11. Sacomoto,G.A.T., Kielbassa,J., Chikhi,R., Uricaru,R., Antoniou,P., Sagot,M.-F., Peterlongo,P. and Lacroix,V. (2012) KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, **13**(Suppl 6), S5.
12. Peterlongo,P., Schnel,N., Pisanti,N., Sagot,M.F. and Lacroix,V. (2010) Identifying SNPs without a reference genome by comparing raw reads. https://hal.inria.fr/inria-00514887/document.
13. Pevzner,P.A., Tang,H. and Waterman,M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 9748–9753.
14. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
15. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.
16. Chikhi,R. and Rizk,G. (2013) Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorith. Mol. Biol: AMB*, **8**, 22.
17. Salikhov,K., Sacomoto,G. and Kucherov,G. (2014) Using cascading Bloom filters to improve the memory usage for de Brujin graphs. *Algorit. Mol. Biol.*, **9**, 2.
18. Sacomoto,G., Sinaimeri,B., Marchet,C., Miele,V., Sagot,M.-F. and Lacroix,V. (2014) Navigating in a Sea of Repeats in RNA-seq without Drowning. *Lect. Notes Bioinformatics*, **8701**, 82–96.
19. Tilgner,H., Knowles,D.G., Johnson,R., Davis,C.A., Chakrabortty,S., Djebali,S., Curado,J., Snyder,M., Gingeras,T.R. and Guigó,R. (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.*, **22**, 1616–1625.
20. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
21. Lu,J., Tomfohr,J.K. and Kepler,T.B. (2005) Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, **6**, 1.
22. Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
23. Anders,S. (2010) Analysing RNA-Seq data with the DESeq package. *Mol. Biol.*, **43**, 1–17.
24. Dillies,M.-A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinformatics*, **14**, 671–683.
25. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.
26. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)*, 289–300.
27. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
28. Engström,P.G., Steijger,T., Sipos,B., Grant,G.R., Kahles,A., Alioto,T., Behr,J., Bertone,P., Bohnert,R., Campagna,D. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185–1191.
29. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
30. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
31. Pante,E., Rohfritsch,A., Becquet,V., Belkhir,K., Bierne,N. and Garcia,P. (2012) SNP detection from de novo transcriptome sequencing in the bivalve Macoma balthica: marker development for evolutionary studies. *PLoS One*, **7**, e52302.
32. Dedeine,F., Vavre,F., Fleury,F., Loppin,B., Hochberg,M.E. and Bouletreau,M. (2001) Removing symbiotic Wolbachia bacteria specifically inhibits oogenesis in a parasitic wasp. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 6247–6252.
33. Dedeine,F., Vavre,F., Shoemaker,D.D. and Boulétreau,M. (2004) Intra-individual coexistence of a Wolbachia strain required for host oogenesis with two strains inducing cytoplasmic incompatibility in the wasp Asobara tabida. *Evolution*, **58**, 2167–2174.
34. Kremer,N., Dedeine,F., Charif,D., Finet,C., Allemand,R. and Vavre,F. (2010) Do variable compensatory mechanisms explain the polymorphism of the dependence phenotype in the Asobara tabida-wolbachia association? *Evolution*, **64**, 2969–2979.
35. Kremer,N., Voronin,D., Charif,D., Mavingui,P., Mollereau,B. and Vavre,F. (2009) Wolbachia interferes with ferritin expression and iron metabolism in insects. *PLoS Pathog.*, **5**, e1000630.
36. Matzkin,L.M. (2004) Population genetics and geographic variation of alcohol dehydrogenase (Adh) paralogs and glucose-6-phosphate dehydrogenase (G6pd) in Drosophila mojavensis. *Mol. Biol. Evol.*, **21**, 276–285.
37. Reed,L., Nyboer,M. and Markow,T. (2007) Evolutionary relationships of Drosophila mojavensis geographic host races and their sister species Drosophila arizonae. *Mol. Ecol.*, **16**, 1007–1022.
38. Modolo,L. and Lerat,E. (2015) UrQt: an efficient software for the Unsupervised Quality trimming of NGS data. *BMC Bioinformatics*, **16**, 137.
39. Lappalainen,T., Sammeth,M., Friedländer,M.R., 't Hoen,P.A.C., Monlong,J., Rivas,M.A., Gonzàlez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
40. Ramaswami,G. and Li,J.B. (2014) RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.*, **42**, D109–D113.
41. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.
42. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Downloaded from http://nar.oxfordjournals.org/ by guest on August 8, 2016