AmpliSAS: web server for multilocus genotyping using next-generation amplicon sequencing data

[1*]Alvaro Sebastian, [1]Magdalena Herdegen, [1]Magdalena Migalska, [1]Jacek Radwan

[1] Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland (https://sites.google.com/site/evobiolab)

[*] To whom correspondence should be addressed. Email: bioquimicas@yahoo.es

This is the pre-peer reviewed version of the following article:

- Sebastian A, Herdegen M, Migalska M, Radwan J (2015) AmpliSAS: a web server for multilocus genotyping using next-generation amplicon sequencing data. *Molecular ecology resources*

which has been published in final form at doi: 10.1111/1755-0998.12453. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Abstract

Next generation sequencing (NGS) technologies are revolutionizing the fields of biology and medicine as powerful tools for amplicon sequencing (AS). Using combinations of primers and barcodes it is possible to sequence targeted genomic regions with deep coverage for hundreds, even thousands of individuals in a single experiment. This is extremely valuable for genotyping gene families in which locus-specific primers cannot be designed, such as the major histocompatibility complex (MHC). The utility of AS is, however, limited by the high intrinsic sequencing error rates of NGS technologies and other error sources such as polymerase amplification or formation of chimeras. Correcting these errors requires extensive bioinformatics post-processing of NGS data. Amplicon Sequence Assignment tool (AmpliSAS) is a web server analysis tool that performs analysis of AS results in a simple and efficient way, offering customization options for advanced users. AmpliSAS is designed as a three-step pipeline: i) read de-multiplexing, ii) unique sequence clustering, iii) erroneous sequence filtering. Allele sequences and frequencies are retrieved in Excel spreadsheet format, making them easy to interpret. AmpliSAS performance has been successfully benchmarked against previously published genotyped MHC data sets obtained with various NGS technologies.

Availability: AmpliSAS online web server is available at:

https://sites.google.com/site/evobiolab/software/amplisas

Contact: bioquimicas@yahoo.es

Background

Few years after the outbreak of NGS technologies in science, these have reached a stage that makes them available and affordable for most biology laboratories around the world (Glenn 2011; Liu *et al.* 2012; Quail *et al.* 2012; Loman *et al.* 2012). Along with classical NGS approaches, such as whole genome, exome or transcriptome sequencing (Abecasis *et al.* 2010; Ozsolak & Milos 2011; Rabbani *et al.* 2014), there are many adaptations of these techniques that obtain results which would be very expensive and laborious to obtain in other ways. One of these is amplicon sequencing (AS) (Bybee *et al.* 2011), which consists of high-throughput sequencing of amplification products from multiple PCRs. AS is now a widely used technique in metagenomics, ecology, population genetics and evolutionary biology (Sogin *et al.* 2006; Swenson 2012; Di Bella *et al.* 2013; Joly *et al.* 2014).

One of the most useful cases of AS is for typing highly polymorphic, multi-gene families, such as genes of Major Histocompatibility Complex (MHC) or olfactory receptor genes (Babik *et al.* 2009; Bentley *et al.* 2009; Dehara *et al.* 2012). Loci belonging to these families often share conserved parts of sequences in which primers can be located. However, as a consequence, alleles from many loci are co-amplified, and direct or indirect identification of sequences of particular alleles with traditional techniques, such as sequencing, SSCP or RSCA (reviewed in Babik 2010) may become unfeasible in species with high number of loci.

MHC class I and class II gene families, which encode cell surface receptors that present antigens to immune cells, are the most polymorphic genes among vertebrates (reviewed in Sommer 2005; Piertney and Oliver 2006), and have become a paradigm for the study of balancing selection (Garrigan & Hedrick 2003; Spurgin & Richardson 2010). They are also central to the study of the host-parasite coevolution, mate choice and kin recognition (Penn 2002; Milinski 2006).

The number of MHC genes can differ within and among species (Kelley *et al.* 2005), but many species show gene duplications and copy-number variation, which makes application of

traditional methods infeasible. Hence, high-throughput sequencing is becoming a method of choice for the study of multigene MHC family (Babik *et al.* 2009; Radwan *et al.* 2012; Sepil *et al.* 2012; Lighten *et al.* 2014b). A typical experiment consists of amplifying individual samples using barcoded primers, then pooling individual samples together for sequencing. The sequences are then de-multiplexed and genotypes of individuals determined.

However, relatively high error rates associated with AS, stemming both from intrinsic sequencing error rate of high-throughput technologies and PCR errors, such as chimera formation, makes genotyping using NGS challenging. For example, homopolymer regions are a major issue for pyrosequencing and ion semiconductor technologies (454 or Ion Torrent), where erroneous indels are introduced in high rates, whereas technology based on reversible dye-terminators (Illumina) suffers from a high number of not necessarily random substitutions (Table S2) (Gilles *et al.* 2011; Vandenbroucke *et al.* 2011; Liu *et al.* 2012; Loman *et al.* 2012; Bragg *et al.* 2013; Ross *et al.* 2013).

Various approaches to deal with AS errors have been used (Lighten *et al.* 2014a), which rely on the assumption that erroneous sequences (henceforth 'artefacts') are less common than correct ones (henceforth 'true sequences', TS). Artefacts are either sieved out or clustered with TS on the basis of similarity to the more common variants in the amplicon (e.g. Promerová et al. 2013; Kloch et al. 2012), in conjunction with other information such as the presence of a variant in a replicate amplicon and other samples (Sommer *et al.* 2013), relative frequency compared to a dominant variant in a cluster (Stutz & Bolnick 2014), or expected distributions of TS frequencies (Lighten *et al.* 2014b) (See Table S1 for a summary and comparison of available AS genotyping methods).

In a recent review, Lighten *et al.* (2014*a*) advocated a model-based approach that may not be optimal when allele amplification efficiencies are uneven (Sommer *et al.* 2013). The method of choice may thus depend on the particular study system and platform used, and genotyping parameters may need to be optimized on a case-by-case basis (Herdegen *et al.* 2014; Stutz &

85 Bolnick 2014). This is made difficult by the lack of customizable and easy-to-use tools for
86 producing either genotypes or outputs that could be used for further downstream genotyping (Table
87 S1). For example jMHC software (Stuglik *et al.* 2011) can be used to initially de-multiplex reads
88 into amplicons, but it does not perform clustering or any downstream analysis.

89 Sequence clustering is important when error-distribution is non-random, e.g. when indels
90 occur in some sequences more often than in others (Gilles *et al.* 2011; Bragg *et al.* 2013). Just
91 removing sequences with indels, as is commonly done during MHC typing protocols, may change
92 the frequency estimations of alleles within an amplicon, thus affecting genotyping based on
93 threshold frequencies or expected frequency-distributions. Furthermore, simple clustering based on
94 similarity may overlook TSs which are similar to other TSs within the same amplicon. To help
95 address this, Stutz & Bolnick (2014) proposed a more complex Stepwise Threshold Clustering
96 (STC) algorithm which allows flexible clustering taking into account relative abundance of a
97 variant within a cluster, in addition to sequence similarity.

98 Here we present Amplicon Sequence Assignment tool (AmpliSAS), a publicly available web
99 server that performs all the necessary steps for AS genotyping in a fully automatic way. It extends
100 jMHC functionality by including STC-like clustering algorithm and sequence filtering capabilities,
101 but also offers advanced processing options for customizing genotyping for special genes or
102 samples. AmpliSAS returns results in Excel spreadsheet format, making them easy to interpret.
103 Genotyping can be optimized by setting system-specific clustering and filtering parameters, or
104 clustering results can be easily used for further downstream analysis, such as DOC genotyping
105 algorithm (Lighten *et al.* 2014b). While AmpliSAS has been designed specifically for multilocus
106 genotyping, it can be also used for other AS purposes, such as organism identification in
107 metagenomics, environmental barcoding (barcodes have a different definition in this case, they are
108 individual amplicon sequences that allow species identification), or detecting allelic mutations.

AmpliSAS is accompanied by AmpliCheck module, which allows preliminary exploration of the data to help in setting optimal parameters for AmpliSAS.

We have benchmarked AmpliSAS performance on three datasets. First, to prove the accuracy of genotype assignments, we used class I HLA-A and HLA-B loci in five human cell lines sequenced with Illumina MiSeq paired-end 2×250 cycles, for which allele sequences were assigned based on Sanger sequencing in two independent laboratories (Bai *et al.* 2014). Second, to assess the quality of our clustering algorithm, we compared AmpliSAS results with those generated by STC method in the original dataset of Stutz & Bolnick (2014). This consists of 301 samples from the non-model organism the threespine stickleback (*Gasterosteus aculeatus*), sequenced with 454 GS FLX Titanium technology. Finally, we applied AmpliSAS to 13 guppy (*Poecilia reticulata*) samples for which inter-platform (Ion Torrent PGM 318 chip and Illumina MiSeq) comparison was available (Herdegen *et al.* 2014). This dataset was used to compare directly the results of genotyping that did not use clustering against that utilizing the AmpliSAS clustering algorithm, for both sequencing platforms.

Term	Definition
Sample	A single genetic material to be sequenced (usually from an individual of the study organism).
Barcode / Molecular Identifier Tag (MID)	A unique short DNA sequence that identifies unambiguously a sample. Barcodes are usually ligated after PCR amplification or directly included in one or both primers.
Marker	A DNA region to be amplified.
Read	Each individual sequence (non-unique) retrieved by a sequencing run. A sequence run will retrieve thousands/millions of reads.
Amplicon	A set of reads derived from a single PCR (one marker, one sample).
Amplicon depth	Number of reads per amplicon
Variant/Sequence	Unique sequence retrieved by a sequencing run. Usually multiple reads correspond to a sequence/variant.
Sequence Depth/Coverage	Number of reads per sequence/variant.
Sequence Frequency or Per Amplicon Frequency (PAF)	Number of reads per sequence divided by the total number of reads in a single amplicon.
True Sequence/Allele (TS/TA)	Sequence that matches a real allele or real sequence in the sample genome.
Artefact/Artefactual sequence	Variant resulting from experimental/technical errors: sequencing errors, polymerase errors, non-specific amplifications (paralogues, pseudogenes), contaminants, etc.
Cluster	A set of variants that fulfil the clustering thresholds and are grouped together (similar sequences). Ideally it integrates a real sequence and all its artefacts.
Dominant sequence	Sequence that represents the cluster real sequence. Usually it is a high depth sequence that passes length constrains and is the consensus of the other cluster members.
Subdominant sequence	Sequence with an unusually high frequency with respect to the dominant sequence in a cluster. Such sequences are frequently a TS/TA and should form a new cluster if proved to be true.
Consensus sequence	Sequence created by taking the most frequent nucleotide in each aligned position of the cluster members.
Allele assignment	Identification of a TS/TA in a particular amplicon.
Dropped allele	True allele that is not present in the genotyping results.
Missing allele	True allele that is not present in the amplicon reads.
Chimera	Variant containing partial sequences from two or more true sequences. Chimeras from more than two sequences are very rare.
Singleton	Variant with only 1 read depth.

Table 1. Definitions of commonly used terms in amplicon sequencing and genotyping studies. They can slightly differ from some authors.

Methods

AmpliSAS algorithm

AmpliSAS workflow is divided into three main steps: i) sequence de-multiplexing, ii) clustering, iii) filtering (Figure 1A; a more detailed workflow is shown in Figure S1). Definitions for common technical terms are listed in Table 1.

1. Sequence de-multiplexing

This step is mandatory (Figure 1A), as it classifies reads into amplicons, and searches for matching of primers and barcodes. Other open source tools like jMHC (Stuglik *et al.* 2011) or SESAME (Meglécz *et al.* 2011) and proprietary software like GS Amplicon Variant Analyzer (Roche) perform the same function. In AmpliSAS, it is possible to include multiple pairs of primers in one single analysis, allowing multiple genes to be analysed without having to run the program several times. As in jMHC, previously defined allele names and sequences can be given as input to assign the same names to de-multiplexed sequences. By default, AmpliSAS will name sequences according to the marker name followed by an auto-increment number in descending coverage order (e.g. HLA_A2-00006). A minimum number of reads can be specified to exclude low coverage amplicons from further analysis, which can be adjusted according to the expected number of alleles and other parameters such as amplification efficiency (Sommer *et al.* 2013).

2. Sequence clustering

The important feature of AmpliSAS compared to jMHC is the implementation of a sequence clustering stage between the de-multiplexing and filtering steps (Figure 1A). We followed the STC algorithm principle of Stutz & Bolnick (2014), but simplified it to increase its speed and provide a number of additional options to help the user customize the analysis to their study system and data set. This step is crucial in overcoming the main problems associated with high error rates inherent to high-throughput techniques. These are: i) discarding sequences with wrong length (due to indels), which results in a loss of data and may bias variant frequency estimation if some variants (e.g. homopolymer-rich) are more prone to indel-type error than others; ii) artefacts that have frequencies as high as those of real alleles, due to non-random errors; and iii) two true alleles that are more similar to each other than to their artefacts (see Table 2). AmpliSAS clustering method processes de-multiplexed sequences, amplicon by amplicon (Figure 1B).

AmpliSAS first orders all sequences in the amplicon by depth, and takes the first sequence

(highest depth). The user can enable an option that checks whether this sequence matches an expected PCR product length or if it complies with a given reading frame (i.e. discrete 3bp deviations from expected length are allowed; see Table 3 for a description of the available clustering parameters). If the sequence complies with the length conditions (or if no conditions are specified), the sequence is labelled as 'dominant sequence' and is then used as the core of a new cluster. Each remaining amplicon sequence (including wrong length ones) is compared with the dominant one, and its sequencing/PCR errors (artefacts) are identified based on user-defined criteria (thresholds for the numbers of substitutions and non-homopolymer indels; Table 3). Note that due to the very frequent homopolymer errors of techniques like Ion Torrent or 454, indels within homopolymer regions are clustered by default; see Table S2 for NGS error rate estimations in different studies. Errors are detected by performing high accuracy pairwise global alignments between the dominant sequence and the others using NEEDLE and NEEDLEALL utilities from EMBOSS package (Rice *et al.* 2000). Instead of sequencing error rates, a more general ‘identity threshold’, can be optionally defined (Table 3). After that, a single cluster is defined as the dominant sequence plus all its artefacts.

The user can define a threshold frequency relative to the dominant sequence (Table 3), the exceeding of which will result in excluding the ‘subdominant sequence’ from the cluster and the formation of a new cluster, even if the sequence is very similar to the dominant (problem case iii). To form a new cluster, the subdominant sequence must be of correct length (± 3bp if such option is selected) and free of frame-shifting indels. Sequences with ‘compensatory indels’ will not form a new cluster when, indels are introduced as a result of a sequencing error, preserving the correct length of a sequence but altering the reading frame. However, potential compensatory indels are ignored by AmpliSAS when they are present at a stretch of 9bp, as, in our experience, such cases are often misalignments of two very similar true alleles rather than sequencing errors.

Finally, all cluster members are merged to create a 'consensus sequence', taking the most frequent nucleotide in each aligned position. If the consensus sequence differs from the dominant one, has not been clustered before, is of correct length, and is not a result of frame shifting indels (see above), then it will replace the dominant sequence. Clustered sequences are removed from further clustering, and their depths are added to the depth of the consensus sequence to increase its coverage (solution of problem i and mitigates ii).

When most of the artefacts have been clustered and only singletons remain to be checked, the clustering process finishes and the non-clustered sequences are discarded. These leftovers are usually contaminants, chimeras or sequences containing many errors that could not be classified into the major clusters.

The full set of clustering parameters is summarized in Table 3, and a graphical schema of the process is shown in Figure 1B. Suggested solutions to problems associated with high error rates of high-throughput sequencing technologies using AmpliSAS clustering algorithm are summarized in Table 2. The AmpliCheck module can be used to explore the sources of possible artefacts and set appropriate clustering parameters.

	Problem description	AmpliSAS solution
i.	Real allele sequence is present at low frequency.	Clustered artefact depths are added to the consensus sequence (putative real allele).
ii.	Artefact sequences are present at high frequencies.	
iii.	Allele sequences are more similar to other alleles than to artefacts.	Adjusting 'dominant frequency' or 'per amplicon frequency' clustering parameters helps to detect these alleles.

Table 2. Genotyping classical problems and suggested solutions with AmpliSAS algorithm.

Clustering parameter	Description
Substitution error rate (%)	Sequences with higher rate of substitutions will be classified into new clusters

Clustering parameter	Description
	(substitutions = error_rate x length).
Indel error rate (%)	Sequences with higher rate of non-homopolymer indels[1] will be classified into new clusters (indels = error_rate x length).
Clustering identity threshold (%)	Sequences with lower sequence identity will be classified into new clusters.
Minimum frequency respect to the dominant (%)	Sequences within a cluster with same or higher frequency respect to the dominant will be classified as subdominants[2] and form a new cluster.
Minimum per amplicon frequency (%)	Sequences with same or higher frequency within the amplicon will be classified as subdominants[2] and form a new cluster.
Cluster only exact length	Only sequences that satisfy theoretical marker lengths can be dominant within a cluster.
Cluster only in-frame	Only sequences in-frame with marker theoretical lengths can be dominant within a cluster.

Table 3. Description of AmpliSAS clustering parameters. [1]Indels in homopolymer regions (3 or more consecutive identical nucleotides) are always clustered. [2]Subdominant sequences must be correct length and free from frame shifting indels.

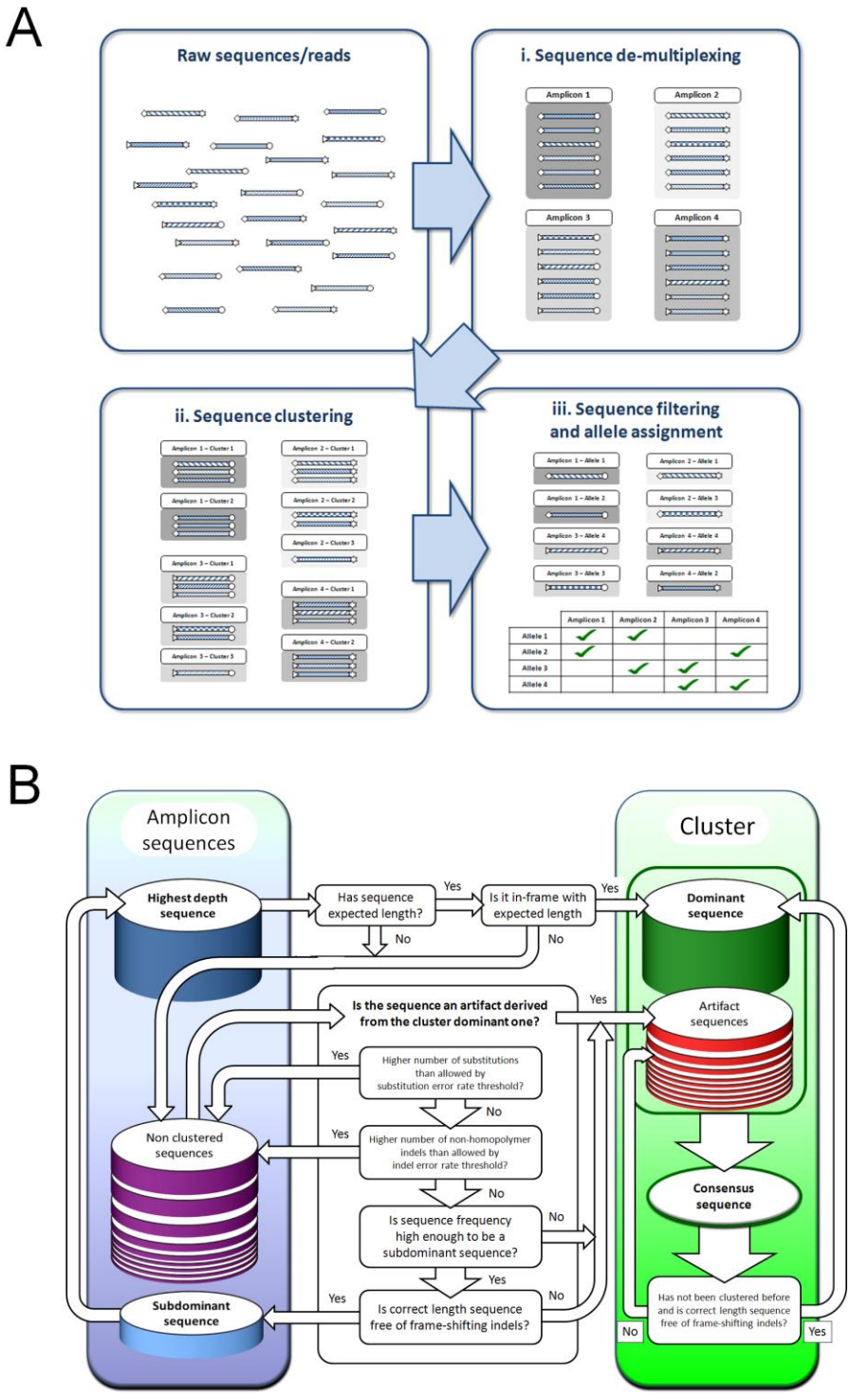


Figure 1. A. AmpliSAS workflow schema: i) sequence de-multiplexing, ii) clustering, iii) filtering and allele assignment. B. Simplified schema of AmpliSAS clustering algorithm decision tree.

3. Sequence filtering

The last step, sequence filtering (Figure 1), implements several user-defined criteria allowing

200 separation of artefacts from putative alleles. Its primary function is to remove PCR chimeras and
201 artefactual non-clustered low depth sequences remaining after clustering.

202 Depending on the genotyping method applied, the settings can be adjusted to yield either an
203 Excel file with final genotypes, or an alternative output for use in downstream analyses. For
204 example, the clustering output containing enriched sequence depths can be readily subjected to
205 DOC analysis (Lighten *et al.* 2014a). AmpliSAS filtering parameters are summarized in Table 4.

206

Filter parameter	Description
*Minimum sequence depth	Sequences with lower amplicon coverage will be discarded.
*Minimum per amplicon frequency (%)	Sequences with lower amplicon frequency will be discarded.
Maximum amplicon length deviation	Sequences longer or shorter than the marker theoretical length±value will be discarded.
Discard chimeras	Sequences that are chimeras from other major sequences will be discarded.
Discard frameshifts	Sequences not in-frame with marker theoretical length will be discarded.
Commonness (number of occurrences and minimum frequency)	Sequences present in an equal or higher number of samples will be kept if they have a minimum frequency set by the user, even if they do not pass other filters.

*Table 4. Description of AmpliSAS filtering parameters. *Depths and frequencies of the unique sequences after clustering will be the sum of depths of all the cluster members.*

207

		Pyrosequencing (455/Ion Torrent)	Illumina
Clustering	[1]Substitution error rate (%)	0.5	1
	[1]Indel error rate (%)	1	0.001
	[2]Minimum frequency respect to dominant (%) or minimum per amplicon frequency (%)	Optional	Optional
	[3]Cluster only exact length/in-frame	YES	Optional
Filtering	[4]Discard chimeras	YES	YES

Table 5. Some suggested AmpliSAS parameters for different techniques. [1]Clustering parameters are

based on technique-specific error profiles (see Table S2). [2]This parameter should be set if the user expects very similar alleles, one of which could be wrongly clustered as an artefact of the other based on the specified error rates. [3]454/Ion Torrent techniques have high sequence position-dependent errors that make this parameter mandatory to avoid wrong length artefactual sequences that are more abundant than true ones. [4]Removal of putative PCR chimeras is highly recommended irrespective of the technique used.

AmpliSAS usage and availability

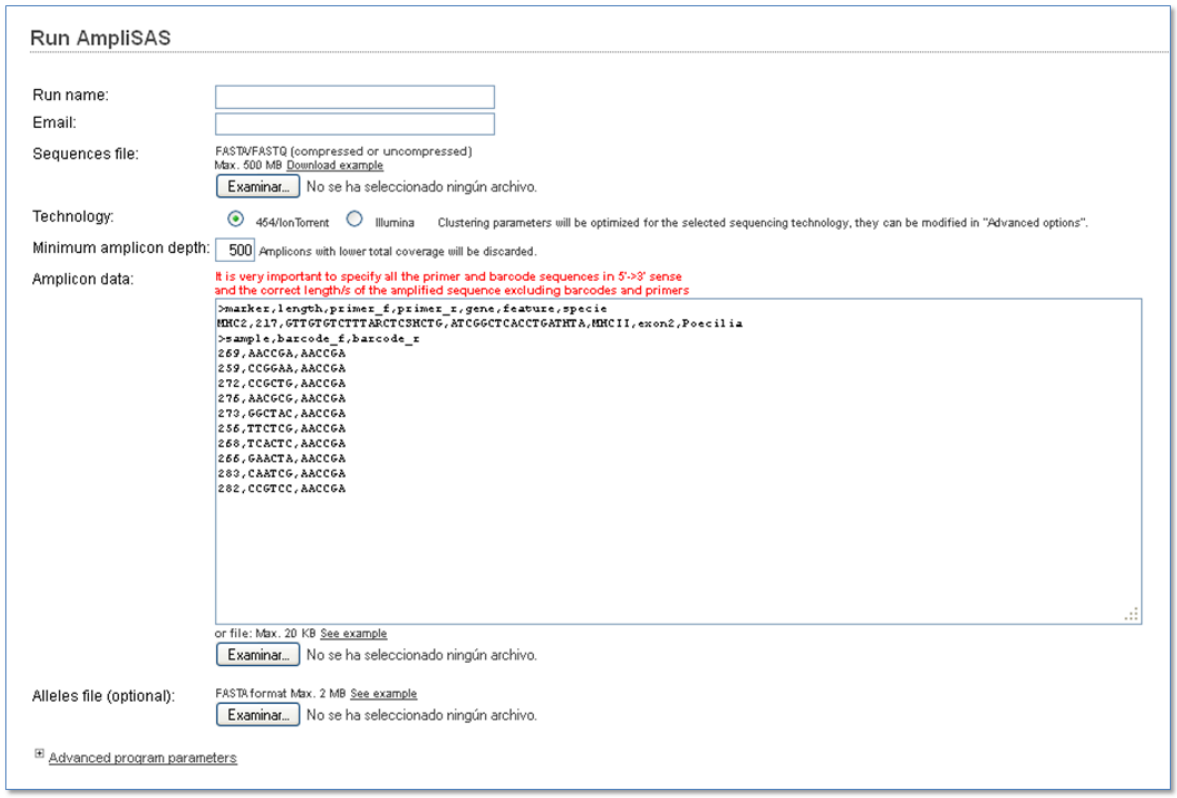
The AmpliSAS main program is written in Perl, with the webserver interface in PHP and JavaScript, running on an Apache server. The online web server is available at: https://sites.google.com/site/evobiolab/software/amplisas.

AmpliSAS functionality

AmpliSAS requires as input two kinds of files/data: i) a file with raw reads in FASTA or FASTQ formats (compressed or not); ii) a file with data on primers, barcodes and amplicons in CSV (comma-separated values) format (example in Figure 2A). After analysis completion, results are downloadable in ZIP compressed format. The compressed file contains three folders ('allseqs', 'clustered' and 'filtered'), an Excel file called 'results.xlsx', and text files with a copy of the input parameters and information about each analysis stage. Final results are saved in an Excel file in a matrix-like format: each predicted allele (TS) is shown in a single row with its sequence, MD5 signature (unique and invariant identifier for each sequence), length, total depth, number of samples in which it is present, mean, maximum and mininum per amplicon frequency (PAF) values, followed by the number of reads corresponding to the sequence found in each sample (samples are represented in columns). An example genotyping results file is shown in Figure 2B. Each worksheet contains results for an individual marker. Output folders store intermediate results after each analysis step ('de-multiplexing', 'clustering' and 'filtering' respectively). FASTA sequence files are generated for individual amplicons, named with the marker followed by the sample name (e.g.

HLA_A3-HEK293.fasta for marker HLA_A3 in sample HEK293). An additional FASTA file is created with all the sequences for a single marker (e.g. HLA_A3.fasta).

A



B

								DEPTH_AMPLICON	1012	1387	1567	1275	894	2439	903	2047	2250	1654
								DEPTH_ALLELES	942	1299	1498	1209	855	2342	855	1946	2161	1605
								COUNT_ALLELES	2	2	2	2	2	1	2	2	1	1
SEQUENCE	**MD5**	**LENGTH**	**DEPTH**	**SAMPLES**	**MEAN_PAF**	**MAX_PAF**	**MIN_PAF**		**269**	**259**	**272**	**276**	**273**	**256**	**268**	**266**	**283**	**282**
AGCTCAAAGAC/	ec35213a	217	8400	7	67,29	97,04	23,84	MHC2-0000001				304	519	2342	457	1012	2161	1605
AGCTCAAAGAC/	e7a698d2	217	1770	2	66,67	70,98	62,36	MHC2-0000005		865		905						
AGCTCAAAGAC/	5141cd1e	217	1709	2	47,54	49,46	45,63	MHC2-0000002			775					934		
AGCTCAAAGAC/	e75233ff7	217	1157	2	38,71	46,14	31,29	MHC2-0000004		434	723							
AGCTGAAGGAC.	8d5bea88	217	734	2	40,83	44,08	37,58	MHC2-0000008					336		398			
AGCTCAAAGAC/	061e1078	217	689	1	68,08	68,08	68,08	MHC2-0000009	689									
AGCTCAAAGAC/	bc6dca7d	217	253	1	25	25	25	MHC2-0000015	253									

Figure 2. A. Example of AmpliSAS web server basic input form. B. Example of Excel file with genotyping results (samples are shown as columns and alleles in rows).

Benchmarking MHC class I and II datasets

We tested the performance of AmpliSAS against three published amplicon sequencing datasets. The first consists of human HLA-A and HLA-B exons 2 and 3 sequenced on Illumina by Bai *et al.* (2014). Here, we applied clustering criteria based on expected error rates typical for this technique

(Table 5) and simple filtering to remove small clusters (note that filtering parameters may vary between species and experiments and should be carefully verified). The purpose of this comparison was to check how well genotypes may be retrieved in the well-characterized human MHC system. The second was the threespined stickleback (*Gasterosteus aculeatus*) class IIβ exon 2, sequenced on 454 and previously genotyped using STC clustering algorithm by Stutz & Bolnick (2014). The purpose of this benchmarking was to see if AmpliSAS one-step clustering gives similar results to those of the recursive clustering algorithm from Stutz & Bolnick (2014). The third was the guppy - (*Poecillia reticulata*) DAβ exon 2, sequenced on both Illumina and PGM and genotyped by Herdegen *et al.* (2014) based on similarity and relative frequency of a variant compared to more common variants within the same amplicon, without clustering and after removal of indels. We replicated the genotyping protocol of Herdegen *et al.* but after AmpliSAS clustering (thus taking into account relative frequency of clusters rather than of unique variants) to see if and how it changed genotyping results.

Human HLA class I genotyping

The data set contains genomic sequences from exon 2 and exon 3 regions from class I HLA-A and HLA-B loci in five human cell lines sequenced with Illumina MiSeq paired-end 2×250 cycles (EBI accession number PRJEB4744) (Bai *et al.* 2014). Real allele sequences were assigned by Sanger sequencing in 2 independent laboratories. To make data compatible with AmpliSAS input format, barcode sequences were incorporated at primer ends for each sample file, and all samples have been merged into a single FASTA file. AmpliSAS was run with parameters adjusted for Illumina data for clustering (substitution error rate: 1%, indel error rate: 0.001%, Table 5). For filtering, we set min. per amplicon frequency as 10 %, and ‘discard chimeras’ as ‘yes’. The threshold of 10% was chosen for this exploratory analysis because most sequences above this threshold should be true variants

based on frequency distribution (Galan et al. 2010) of non-duplicated loci (human MHC-A and B heterozygous cells will have maximum two alleles).

After de-multiplexing 123876 reads, 41302 were assigned to HLA-A exon 2, 54257 to HLA-A exon 3, 22903 to HLA-B exon 2 and 5318 to HLA-B exon 3. However, for HLA-B exon 3 the most abundant unique sequence consisted of only 14 reads (compared to 3925, 7441 and 1244 reads, respectively, for the other markers), likely because of the presence of many non-specific sequences within an amplicon. We therefore excluded this marker from further analysis.

AmpliSAS HLA-A (exons 2 and 3) and HLA-B (exon 2) allele predictions fully matched real allele sequences obtained by Sanger sequencing. For exon 2 and 3 regions of HLA-A, the 5 real alleles were predicted with 100% accuracy without any false positive (Table 6). HLA-B exon 2 region predictions also cover all alleles confirmed with Sanger sequencing, but AmpliSAS retrieves one additional sequence (Table 6). This sequence matches the HLA-E locus, which suggests that HLA-B exon 2 primers simultaneously amplified a gene of the same family and that our algorithm was accurate enough to retrieve its sequence. When we relaxed the filtering parameters (e.g. min. per amplicon frequency: 3%), we discovered more sequences from HLA-E, HLA-G, HLA-Cw1 and HLA-K alleles (data not shown), which are likely to be non-specific PCR products present among Illumina reads. Full genotyping results are shown in Appendix S1.

Stickleback MHC class IIβ genotyping

The second data set is from Stutz & Bolnick (2014), and consists of genomic sequences of MHC class IIβ loci, exon 2 region, from 301 samples of the non-model organism the threespine stickleback (*Gasterosteus aculeatus*), sequenced with 454 GS FLX Titanium technology. This data had previously been analysed with the Stepwise Threshold Clustering (STC) genotyping algorithm (Stutz & Bolnick 2014), and the original raw SFF file is available from NCBI (accession number

287 SRR1177032). The STC algorithm is accurate but slow, as it performs multiple clustering rounds
288 with increasing similarity thresholds and repeats clustering 100 times in each round reordering
289 sequences. Our aim was thus to assess whether the reduced computational intensity of AmpliSAS
290 could produce clusters of comparable accuracy.

291 Reads from the original STC article were given as input for AmpliSAS. For clustering, we
292 used the following parameters: substitution error rate = 0.5%; indel error rate = 1%; minimum
293 frequency respect to dominant = 22%; cluster only exact length = 'yes'. For the filtering step, we set
294 min. per amplicon frequency = 4.5%, discard chimeras = 'yes', and min. amplicon depth = 500.
295 'Minimum frequency respect to dominant' and 'min. per amplicon frequency' parameters are
296 equivalent to 'dominance threshold' and 'size threshold' parameters used by Stutz & Bolnick
297 (2014). Following the original article, we used the commonness thresholds in AmliSAS to retain
298 sequences with that had low frequencies after clustering (small clusters) but which were present in
299 at least three other samples. However, we note that such inclusion of very low frequency sequences
300 as TS is highly controversial, because they could derive from contaminants or from tag-swapping
301 (Schnell *et al.* 2015). A total of 92 samples which passed the criterion of 500 sequences per
302 amplicon were retained. The same dataset was analysed with the original STC software
303 implemented in R (Stutz & Bolnick 2014).

304 STC produced 530 clusters above the size threshold of 4.5%, while AmpliSAS formed 586
305 clusters. Average per amplicon frequencies of clusters were 12.2% with STC and 14.0% with
306 AmpliSAS. Of the 530 clusters identified by STC, 495 (93%) were also identified by AmpliSAS,
307 sharing the same dominant sequences. Among the 35 clusters found only by STC, 14 were present
308 among AmpliSAS small clusters (freq. < 4.5%) and the remaining 21 had a sequence with wrong
309 length as dominant. These clusters are removed later by STC, but AmpliSAS retains them because a
310 correct-length dominant sequence is present among cluster members. Ion Torrent and 454

311 technologies produce a high number of position specific errors (particularly in homopolymer
312 regions), and sometimes some artefacts have higher depths than the true sequences (Gilles *et al.*
313 2011). These cases would be incorrectly discarded by STC when removing clusters with wrong
314 length dominant sequences, but retained by AmpliSAS. Among clusters found by AmpliSAS, but
315 not by SCT, 54 were found among STC small clusters. The remaining 37 had dominant sequences
316 of correct length and an average frequency of 11.9%, which suggests they were correctly assigned.
317 Apart from clustering strategy, AmpliSAS differs from STC in its strategy of aligning
318 amplicon sequences, which may account for some of the inconsistencies between STC and
319 AmpliSAS clusterings. STC performs a multiple global alignment of all amplicon sequences using
320 CLUSTALW to produce a matrix of distances, whereas AmpliSAS performs pairwise global
321 alignments with the DNA version of the Needleman-Wunsch algorithm (Needleman & Wunsch
322 1970; Larkin *et al.* 2007). Pairwise global alignments are more time-consuming but much more
323 accurate. In the early design stages of AmpliSAS, we trialled the use of multiple alignment of the
324 amplicon, but found that it returned too many alignment errors. The presence within an amplicon of
325 divergent allele sequences accompanied by multiple insertions and deletions resulting from
326 sequencing errors makes the multiple alignment error-prone, especially in large datasets.
327 Both STC and AmpliSAS retrieved 163 putative alleles, 159 of which (98%) were identical.
328 STC performed 667 allele assignments (total number of alleles assigned in all individuals; see
329 definition of assignment in Table 1), and AmpliSAS 655, having 620 (93%) in common with SCT
330 (Table 6). Analysing the differences in more detail, we found that allele assignments made by STC
331 and not by AmpliSAS corresponded with allele sequences with very low depth, which are filtered
332 by AmpliSAS because their clusters are too small (<1% frequency after clustering; Figure S3).
333 Meanwhile, the few allele assignments made by AmpliSAS and not by STC correspond to clear true
334 alleles. For example in sample 317, three clear alleles were dropped by STC (alleles 83, 124 and

882). These three alleles are present in other samples, have correct length, high frequencies, and are not chimeras (Figures S3 y S4A). Further examination showed that these three alleles, all of length 213bp, are members of clusters where an artefactual 212bp sequence is the major one, with the length difference arising from a homopolymer indel (Figure S5). STC initially recognizes these 212bp sequences as true alleles but later removes them because of their incorrect length. This is a clear case where a particular artefact is more abundant than the real sequence from which it derives. In contrast, AmpliSAS recognizes the correct length allele sequences as a 'dominant sequence' at the clustering stage and retains them in the final results (the clustering parameter 'cluster only exact length/in-frame' is crucial in this case; Figure S5). Full genotyping results are shown in Appendix S1.

Guppy MHC class II genotyping

To assess how clustering affects allele assignment based on Ion Torrent and Illumina sequencing, we used a dataset on the guppy alleles of MHC class II (exon 2) obtained by sequencing 13 individuals on both platforms (Herdegen *et al.* 2014). Herdegen *et al.* (2014) assigned alleles without clustering, using the empirical threshold method (Radwan *et al.* 2012; Promerová *et al.* 2013). Using a representative sample of sequences, they determined that the lower threshold, below which vast majority of variants could be explained as 1-2 bp substitution artefacts, was 3%, and the upper threshold, above which such artefacts are not found, was 12%. During genotyping, after removing sequences with indels, variants with frequencies less than the threshold of 3% were removed. The remaining variants were screened for chimeras, as well as 1-2 bp substitutions of more common variants on a case-by-case basis; such variants were removed, except when they constituted >12% of the reads within an amplicon (see Herdegen *et al.* 2014 for details).

In our analysis, we used similar parameters for AmpliSAS as used in the original study

359 (<3% for removal, >12% for variants with 1-2 bp substitutions to form a separate cluster), but
360 sequences less frequent than 12% which contained 1-2 bp substitutions compared to a more
361 common variant within the same amplicon were clustered together with this variant, rather than
362 removed. Likewise, variants with indels (1-2bp) were retained for clustering.

363 For Illumina data, all 46 assignments made by Herdegen *et al.* (2014) were also called by
364 AmpliSAS clustering, but one additional allele was called by AmpliSAS. For Ion Torrent, 43 of the
365 44 assignments of Herdegen *et al.* (2014) were also called by AmpliSAS clustering, with AmpliSAS
366 identifying three additional variants. The few detected differences in allele assignments were all due
367 to changes in per amplicon frequencies of the reads forming a cluster compared to per amplicon
368 frequencies of unclustered variants. These relatively minor changes (<6 %) caused some variants to
369 shift over or under the thresholds that determined whether they were called as artefacts or TAs.

370 The greater effect of AmpliSAS clustering on results from Ion Torrent allele assignment
371 relative to Illumina was to be expected, as the former is prone to sequence-specific generation of
372 indels, the removal of which may bias estimates of per-amplicon variant frequencies. While this had
373 a very minor effect on genotyping results from the guppy dataset, the effect is likely to vary
374 between systems according to the properties of the sequence sets analysed.
375

Marker	NGS technology	Sample number	Method	Allele number	Common alleles	Total allele assignments	Common assignments
Human HLA-A exon 2	Illumina MiSeq	5	Sanger	5	5	8	8
			AmpliSAS	6		8	
Human HLA-A exon 3	Illumina MiSeq	5	Sanger	5	5	8	8
			AmpliSAS	5		8	
Human HLA-B exon 2	Illumina MiSeq	5	Sanger	5	5	6	6
			AmpliSAS	6		7	
Stickleback MHCII- exon 2	454 GS FLX Titanium	92	STC	163	159	667	620

			AmpliSAS	163		655	
Guppy MHCII exon 2	Illumina MiSeq	13	MPAF	19	18	46	46
			AmpliSAS	18		47	
Guppy MHCII exon 2	Ion Torrent PGM	13	MPAF	22	21	44	43
			AmpliSAS	21		46	

Table 6: Statistics of AmpliSAS allele predictions and assignments compared to human HLA typing by Bai et al. (2014), stickleback MHC class IIb typing by Stutz & Bolnick (2014) and guppy MHC class II typing by Herdegen et al. (2014)

Conclusion

The utility of AS as a ground-breaking tool for characterisation of sequences of multi-gene families is hampered by high frequency of errors introduced by next generation sequencing, which requires complex bioinformatic post-processing of the data. This can now be facilitated by the AmpliSAS web server described here. It builds on the genotyping strategy introduced by the STC algorithm of Stutz & Bolnick (2014), and, like STC, allows clustering artefacts with the real sequences from which they come from. Artefact recognition is not always straightforward, and can be particularly problematic when using pyrosequencing (454) or ion semiconductor technologies (Ion Torrent) that produce high rates of non-random sequencing errors in homopolymer regions. In benchmarking against three published data sets that had utilised a range of NGS technologies and genotyping approaches, we have shown that the pairwise global sequence alignment clustering approach of AmpliSAS is an efficient and accurate tool for error annotation and artefact recognition, and after setting experiment-dependent parameters by the user, it is a useful tool for genotyping. By clustering artefacts with true variants, it increases the depth of allele sequences, making it easier to distinguish alleles from the remaining low frequency artefacts at later filtering stages.

AmpliSAS clustering outputs can be adjusted by frequency, depth or other desired parameters to yield both putative genotypes and files for downstream analyses, such as DOC method (Lighten *et al.* 2014b). While different genotyping approaches should produce similar

results even in species with highly polygenic MHC, given sufficiently deep coverage and careful primer design (Biedrzycka *et al.* unpublished), comparison of protocols and optimising genotyping parameters is recommended for each study, based on replicated genotyping of a subset of individuals. For example, while in guppies sequences with per amplicon frequency < 2% appeared to be mostly artefacts (Herdegen *et al.* 2014; Lighten *et al.* 2014b), in sedge warbler (*Acrocephalus schoenbaenus*), characterised by much higher number of co-amplifying alleles (up to 51) and sequenced at much higher depth, all sequences >1% could be classified as TA (Biedrzycka *et al.* unpublished).

Our benchmarking has shown that AmpliSAS reliably replicates clustering and genotyping results obtained in earlier studies across different NGS platforms. Due to its accuracy, versatility and user-friendly interface, AmpliSAS, in conjunction with AmpliCHECK, would facilitate optimisation of genotyping parameters and the choice of optimal genotyping method. We believe it will prove to be a useful tool for many applications involving amplicon sequencing.

Data Accessibility

Supporting information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Excel file with AmpliSAS genotyping assignments for the benchmarking datasets (human, stickleback and guppie). Original results are also included for comparison.

Table S1. Summary of up to date multilocus genotyping methods for amplicon targeted sequencing.

Table S2. Error rate comparison among several NGS technologies and sources.

Figure S1. AmpliSAS extended workflow schema.

Figure S2. BLASTN alignments of a HLA real allele and a PCR sub-product to human genome.

Figure S3. Examples of genotyping discrepancies between AmpliSAS and STC methods in stickleback MHC class IIβ.

Figure S4. Alignment examples of AmpliSAS predicted allele sequences for stickleback MHC class IIβ.

Figure S5. AmpliSAS clusters for alleles 83, 124 and 882 (213bp) in stickleback sample 317.

Acknowledgements

We thank William Stutz for his kind support in running STC method and benchmarking, Michal Stuglik for his help with chimera detection code and Karl Phillips for his elaborated suggestions and corrections. This work was supported by MAESTRO grant UMO-2013/08/A/NZ8/00153 from National Science Centre to JR.

References

Abecasis GR, Altshuler D, Auton A *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–73.

Babik W (2010) Methods for MHC genotyping in non-model vertebrates. *Molecular ecology resources*, **10**, 237–51.

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular ecology resources*, **9**, 713–9.

Bai Y, Ni M, Cooper B, Wei Y, Fury W (2014) Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC genomics*, **15**, 325.

Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G (2013) High throughput sequencing methods and analysis for microbiome research. *Journal of microbiological methods*, **95**, 401–14.

Bentley G, Higuchi R, Hoglund B *et al.* (2009) High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue antigens*, **74**, 393–403.

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS computational biology*, **9**, e1003031.

Bybee SM, Bracken-Grissom H, Haynes BD *et al.* (2011) Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome biology and evolution*, **3**, 1312–23.

Dehara Y, Hashiguchi Y, Matsubara K *et al.* (2012) Characterization of squamate olfactory receptor genes and their transcripts by the high-throughput sequencing approach. *Genome biology and evolution*, **4**, 602–16.

Garrigan D, Hedrick PW (2003) Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution; international journal of organic evolution*, **57**, 1707–22.

Gilles A, Meglécz E, Pech N *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics*, **12**, 245.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular ecology resources*, **11**, 759–69.

Herdegen M, Babik W, Radwan J (2014) Selective pressures on MHC class II genes in the guppy (Poecilia reticulata) as inferred by hierarchical analysis of population structure. *Journal of Evolutionary Biology*, **27**, 2347–2359.

Joly S, Davies TJ, Archambault A *et al.* (2014) Ecology in the age of DNA barcoding: the resource, the promise and the challenges ahead. *Molecular ecology resources*, **14**, 221–32.

Kelley J, Walter L, Trowsdale J (2005) Comparative genomics of major histocompatibility complexes. *Immunogenetics*, **56**, 683–95.

Kloch A, Baran K, Buczek M, Konarzewski M, Radwan J (2012) MHC influences infection with parasites and winter survival in the root vole Microtus oeconomus. *Evolutionary Ecology*, **27**, 635–653.

Larkin MA, Blackshields G, Brown NP *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, **23**, 2947–8.

Lighten J, van Oosterhout C, Bentzen P (2014a) Critical review of NGS analyses for de novo genotyping multigene families. *Molecular ecology*, **23**, 3957–72.

Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014b) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (Poecilia reticulata). *Molecular ecology resources*, 1–15.

Liu L, Li Y, Li S *et al.* (2012) Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, **2012**, 251364.

Loman NJ, Misra R V, Dallman TJ *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, **30**, 434–9.

Meglécz E, Piry S, Desmarais E *et al.* (2011) SESAME (SEquence Sorter & AMplicon Explorer): genotyping based on high-throughput multiplex amplicon sequencing. *Bioinformatics (Oxford, England)*, **27**, 277–8.

Milinski M (2006) Fitness consequences of selfing and outcrossing in the cestode Schistocephalus solidus. *Integrative and comparative biology*, **46**, 373–80.

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**, 443–53.

Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics*, **12**, 87–98.

Penn DJ (2002) Major Histocompatibility. *Enciclopedia of Life Sciences*.

Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity*, **96**, 7–21.

Promerová M, Králová T, Bryjová A, Albrecht T, Bryja J (2013) MHC class IIB exon 2 polymorphism in the Grey partridge (Perdix perdix) is shaped by selection, recombination and gene conversion. *PloS one*, **8**, e69135.

Quail M a, Smith M, Coupland P *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, **13**, 341.

Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, **59**, 5–15.

Radwan J, Zagalska-Neubauer M, Cichoń M *et al.* (2012) MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Molecular Ecology*, **21**, 2469–2479.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, **16**, 276–7.

Ross MG, Russ C, Costello M *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome biology*, **14**, R51.

Schnell IB, Bohmann K, Gilbert MTP (2015) Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular ecology resources*.

Sepil I, Moghadam HK, Huchard E, Sheldon BC (2012) Characterization and 454 pyrosequencing of major histocompatibility complex class I genes in the great tit reveal complexity in a passerine system. *BMC evolutionary biology*, **12**, 68.

514 Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the
515 underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the*
516 *United States of America*, **103**, 12115–20.

517 Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and
518 conservation. *Frontiers in zoology*, **2**, 16.

519 Sommer S, Courtiol A, Mazzoni CJ (2013) MHC genotyping of non-model organisms using next-
520 generation sequencing: a new methodology to deal with artefacts and allelic dropout. *BMC*
521 *genomics*, **14**, 542.

522 Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and
523 misunderstandings. *Proceedings. Biological sciences / The Royal Society*, **277**, 979–88.

524 Stuglik MT, Radwan J, Babik W (2011) jMHC: software assistant for multilocus genotyping of
525 gene families using next-generation amplicon sequencing. *Molecular ecology resources*, **11**,
526 739–42.

527 Stutz WE, Bolnick DI (2014) Stepwise Threshold Clustering: A New Method for Genotyping MHC
528 Loci Using Next-Generation Sequencing Technology. *PloS one*, **9**, e100587.

529 Swenson NG (2012) Phylogenetic analyses of ecological communities using DNA barcode data.
530 *Methods in molecular biology (Clifton, N.J.)*, **858**, 409–19.

531 Vandenbroucke I, Van Marck H, Verhasselt P *et al.* (2011) Minor variant detection in amplicons
532 using 454 massive parallel pyrosequencing: experiences and considerations for successful
533 applications. *BioTechniques*, **51**, 167–77.

534 Westerdahl H, Wittzell H, von Schantz T, Bensch S (2004) MHC class I typing in a songbird with
535 numerous loci and high polymorphism using motif-specific PCR and DGGE. *Heredity*, **92**,
536 534–42.

537