Published online 7 July 2015

Nucleic Acids Research, 2015, Vol. 43, No. 20 ***e135***
doi: 10.1093/nar/gkv694

A new method to prevent carry-over contaminations in two-step PCR NGS library preparations

Volkhard Seitz[1,2,*], Sigrid Schaper[2], Anja Dröge[2], Dido Lenze[1], Michael Hummel[1,*] and Steffen Hennig[2,*]

[1]Institute of Pathology, Charité—University Medicine Berlin, Campus Benjamin Franklin, Berlin, Germany and [2]HS Diagnomics GmbH, Berlin, Germany

Received March 19, 2015; Revised June 22, 2015; Accepted June 25, 2015

ABSTRACT

Two-step PCR procedures are an efficient and well established way to generate amplicon libraries for NGS sequencing. However, there is a high risk of cross-contamination by carry-over of amplicons from first to second amplification rounds, potentially leading to severe misinterpretation of results. Here we describe a new method able to prevent and/or to identify carry-over contaminations by introducing the K-box, a series of three synergistically acting short sequence elements. Our K-boxes are composed of (i) K1 sequences for suppression of contaminations, (ii) K2 sequences for detection of possible residual contaminations and (iii) S sequences acting as separators to avoid amplification bias. In order to demonstrate the effectiveness of our method we analyzed two-step PCR NGS libraries derived from a multiplex PCR system for detection of T-cell receptor beta gene rearrangements. We used this system since it is of high clinical relevance and may be affected by very low amounts of contaminations. Spike-in contaminations are effectively blocked by the K-box even at high rates as demonstrated by ultra-deep sequencing of the amplicons. Thus, we recommend implementation of the K-box in two-step PCR-based NGS systems for research and diagnostic applications demanding high sensitivity and accuracy.

INTRODUCTION

Two-step polymerase chain reaction (PCR) strategies are a convenient approach to generate amplicon libraries suitable for next generation sequencing (NGS) (Figure 1). For this purpose, in the first amplification reaction the target nucleic acid sequence is amplified using specific primers flanked by a tail sequence (e.g. a M13 or T7 tail). In the second amplification reaction, adaptor sequences required for NGS are introduced utilizing primers complementary at their 3′ end to the tail sequence of the first amplification primers (1). In order to make this method more cost-efficient, multiplexing of several samples for NGS can be performed. To this end so-called barcodes are introduced within or close to the 5′ end of the second amplification primers (1).

However, this highly efficient two-step amplification strategy harbors a considerable risk of cross-contamination during set-up of the second PCR especially by carry-over of amplicons from the first PCR to the second PCR due to the high number of amplicons generated in the first amplification reaction (2,3). The risk of contamination is even higher if resulting PCR products are isolated by gel extraction or PCR purification kits. Furthermore, PCR products of a second amplification may contaminate other second amplification reactions. Other mistakes might occur by pipetting a first-step amplicon inadvertently into the wrong second PCR mix, which can be regarded as a 100% contamination.

There are many applications which rely on two-step PCR strategies for NGS library generation (4–8). Here we focus on the analysis of genomic T-cell receptor beta (TCRβ) gene rearrangements since this application is very suitable to demonstrate the importance of contamination protection due to the inherent risk of detecting extremely low rates of contaminations and due to the clinical importance of detecting very rare events.

T-cells play a central role in cell-mediated immunity with 95% of T-cells expressing T-cell receptors (TCR) consisting of an alpha and beta chain as a heterodimer on their surfaces (9). TCRβ chains are highly diverse and can be regarded as genomic fingerprint of each T-cell (10,11) The high diversity of the human T-cell repertoire is mainly generated by somatic recombination among the Variable (V), Diversity (D) and Joining (J) TCRβ gene segments and the addition of random non-templated bases at recombination junctions (12). The hypervariable complementarity-determining region 3 (CDR3) of the rearranged TCRβ rec-

*To whom correspondence should be addressed. Tel: +49 30 8445 2734; Fax: +49 30 450 536106; Email: volkhard.seitz@charite.de
Correspondence may also be addressed to Prof. Dr. Michael Hummel. Tel: +49 30 8445 2734; Fax: +49 30 450 536106; Email: michael.hummel@charite.de
Correspondence may also be addressed to Dr. Steffen Hennig. Tel: +49 30 79786111/110; Fax: +49 30 79781731; Email: hennig@hsdiagnomics.de

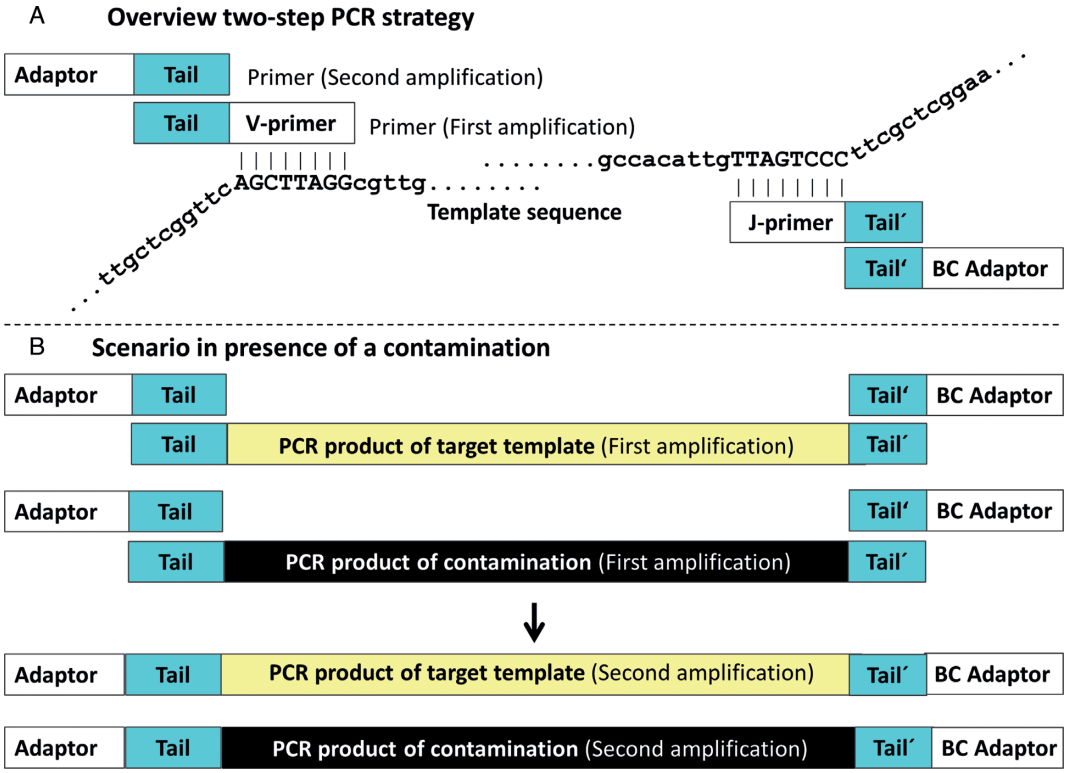

Figure 1. Scheme of a standard two-step PCR approach. In a two-step PCR strategy (**A**) a PCR product is at first generated using specific primers flanked by a tail sequence and is then further amplified in the second reaction with primers that target only the tail sequence (blue color) introduced by the first amplification primers. Therefore, in the second amplification (**B**) contaminations from other first or second PCRs may be co-amplified. The barcodes (BC) allow barcoding for NGS multiplexing but cannot suppress cross-contaminations.

ognizes antigens bound to major histocompatibility complex molecules (10). Therefore, the determination of T-cell repertoires are of interest for a wide range of research and clinical applications, e.g. (i) detection of T-cell clonality, (ii) minimal residual disease analysis in T-cell lymphoma patients, (iii) TCRβ signatures as biomarkers in clinical vaccination studies or (iv) identification of common 'public TCRs' in autoimmune disease as therapeutic targets.

Here we describe a novel contamination protection method for two-step PCRs ideally suited for NGS library preparations. By utilizing a synergistic combination of sequence elements denoted as K-box, this new method operates through both prevention of contaminations and identification of possible residual contaminations. Our method shows effective suppression of even high rates of artificial contaminations. Hence we recommend implementation of the K-box as an additional feature in two-step PCR-based NGS procedures for research and diagnostic applications, where high sensitivity and reliability is mandatory.

MATERIALS AND METHODS

The K-box architecture

We developed a comprehensive method in order to handle inherent contamination problems in two-step PCR-based NGS library preparations. This new method is based on three synergistically acting sequence elements referred to as K-box. As outlined in Figure 2, the first amplification primers comprise all three elements (k1, k2 and s in the forward primer and k1′, k2′ and s′ in the reverse primer), whereas the forward and reverse primers of the second amplification contain only k1 and k1′, respectively. With the three K-box elements one single K-box is designed by bioinformatics methods and optimized not to form hybrids that might lead to mispriming under a multiplex experimental setup.

The rationale behind the introduction of K1 elements is the prevention of carry-over contaminations from previous amplification reactions. As outlined in Figure 2 only matching K1 sequences in the primers of the first and second PCR allow amplification of the respective PCR product. Since K1 sequences are individual for each sample, the amplification of PCR cross-contaminations is suppressed in the second PCR in case of non-concordant K1 elements.

In our TCRβ multiplex analyses K1 sequences comprise exemplary 7 nucleotides (see Table 1). The K1 elements are designed by a minimization procedure which selects the K1 elements by (i) minimal global similarity to any other K1-element, and (ii) a strong dissimilarity rule at the specific 3′ ends of the K1 elements. The resulting K-box sequences have a very low potential for hybrid formation.

A further K-box element is the K2 sequence present only in the first amplification primers. While K1 elements lead to suppression of contaminations, the K2 elements serve to detect possible remaining contaminations from previous amplification reactions. In our TCRβ multiplex analyses we de-

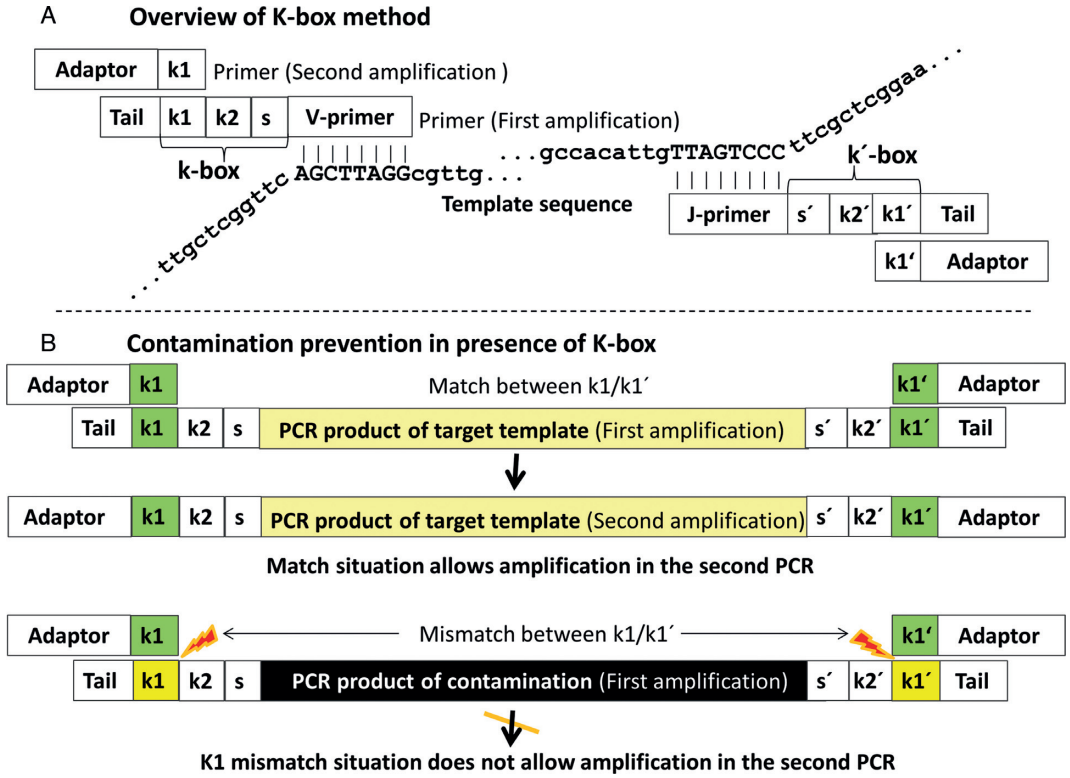

Figure 2. Scheme of the K-box method in two-step PCR. In the K-box method (**A**), three additional sequence elements (K1-, K2- and S-elements) are introduced by the first amplification primers. The combination of k1 and k1′ (K1 elements) differ from sample to sample but are identical (matched) within a specific first and second amplification primer set. (**B**) A second amplification can only occur with matched K1 elements and therefore the amplification of contaminations from other first or second amplifications is suppressed. In Figure 2 mismatched K1 elements are indicated by different colors and marked by two flash symbols. The combination of k2 and k2` (K2 elements) serves as a bioinformatics indicator. Sequences in the PCR mix with unexpected K2 elements are identified as residual contaminations. Furthermore, s and s′ (S elements) are designed as mismatch with the template i.e. the respective genomic TCRβ V- and J-segments. The S elements effectively separate the template matching part of the primers from their 5′- tail thereby reducing the risk of PCR bias.

Table 1. Sequences of k- and k′-box elements for set 1–9 used for the two-step PCR NGS library preparations

Set number	k-box		k′-box	
	k1 sequence	k2 sequence	k1′ sequence	k2′ sequence
Set1	CACCCAA	GAC	GTTGGTT	CGT
Set2	CACCCAA	GAC	GGTCATG	TGG
Set3	CACCCAA	GAC	GCCATTT	TAA
Set4	AGTTTTG	CGG	GTTGGTT	CGT
Set5	AGTTTTG	CGG	GGTCATG	TGG
Set6	AGTTTTG	CGG	GCCATTT	TAA
Set7	CTTTAGA	GTG	GTTGGTT	CGT
Set8	CTTTAGA	GTG	GGTCATG	TGG
Set9	CTTTAGA	GTG	GCCATTT	TAA

The sequences are listed in 5′-3′orientation as present in the forward and reverse primers. The three distinct k-boxes used in the forward primers are marked with green, yellow and red colors and the purple, blue and gray colors mark the three different k′-boxes employed in the reverse primers. By using all possible k/k′ combinations of the 3 k- and 3 k′-boxes 9 sets were generated.

signed K2 elements exemplarily with three sample-specific nucleotides (see Table 1).

For a better understanding of first and second PCR primer matching we introduce here the term ‘set’ which defines a unique k/k′ combination. Within an individual set a specific second primer pair is designed to work only together with a specific first primer pair.

Table 1 gives an example of 9 sets with K1 and K2 sequences used in the TCRβ multiplex analyses. The 9 possible set combinations are formed by 3 distinct k- and k′-boxes in the forward and reverse primers. Basically with 10 distinct k- and k′-boxes 100 unique k/k′ sets can be generated. Therefore, with a linear increase in the effort of primer synthesis an exponential growth in the number of distinct sets is achieved. This makes this approach up-scalable with reasonable efforts.

In principle, K1 and K2 elements can be of any convenient length. We typically use a length of 7 nucleotides for K1 and 3 nucleotides for K2 elements. Since the sequences of K1 and K2 elements are variable in a sample specific manner, certain variations may coincidentally match in their last nucleotides on the 3-′ end with the sequence of the target DNA next to the template. In that case the template hybridizing part of the first amplification primers would be elongated, possibly leading to higher annealing temperatures and PCR bias. To circumvent this problem separator sequences (S elements) are introduced into the first amplification primers (Figure 2). They are designed as mismatch with the template i.e. the respective genomic TCRβ V- and J-segments. The S elements effectively separate the template matching part of the primers from their 5′- tail thereby reducing the risk of PCR bias.

It should be noted that every multiplex PCR assay needs a proper design for S elements, whereas the same K1 sequences and the respective second amplification primers can be used for different multiplex applications. Barcodes are introduced into the second amplification primers to allow multiplex sequencing of NGS libraries. Table 2 summarizes the functions of the different K-box elements and the terminology used throughout the manuscript.

Design and description of target DNA samples

The K-box method was established in two-step PCR-based NGS library preparations with three different types of templates:

(1) Genomic DNA (gDNA) from the T-cell line Peer. PCRs were performed with a primer pair specific for the V- and J-segments of the Peer TCRβ gene rearrangement (Supplementary Figure S1).
(2) Tonsillar gDNA which harbors a physiologically broad range of different TCRβ gene rearrangements. TCRβ multiplex PCRs were performed (Supplementary Methods).
(3) A synthetic template with 616 TCRβ V-J-combinations (Supplementary Figure S3). TCRβ multiplex PCRs were performed (Supplementary Methods).

Determination of parameters for effective suppression of carry-over contamination

In order to optimize the sets of K-boxes for practical use in a laboratory, systematic variations of different parameters were tested with respect to their effects on contamination suppression.

In standardized experiments 100 ng genomic DNA of the T-cell line Peer was used as template in the first amplification reaction employing only primers specific for the Peer TCRβ rearrangement. The product of this first PCR round was used as template for the second amplification. The PCR conditions are given as Supplementary Methods. With this two-step PCR workflow both matching and mismatching K1 elements were tested. Supplementary Table S1 illustrates the sequences of the 4 primer sets with matching K1 elements (sets A–D). Each set is characterized by a unique k/k′ combination. Supplementary Tables S4 and S6 explain the experimental set-up using primers with matching and mismatching K1 elements. Mismatching alterations of one and two base pairs (bp) were carried out and their gradual effects on contamination suppression were determined.

Proofreading polymerases exhibit 3′- 5′exonuclease activity and may remove K1 elements at the 3′- end of the second amplification primer. We examined therefore the protective effect of phosphorothioate (PT) bonds introduced at (i) the first, (ii) the first and second and at (iii) the first, second and third position from the 3′ end of the second amplification primers (Supplementary Table S4). For comparison we performed the second amplification with a non-proofreading polymerase and second amplification primers without PT-bonds (Supplementary Table S6).

PCR products were analyzed on 6% acrylamide gels and Tif images were generated with the Biorad Geldoc 2000 (München, Germany) using default settings. PCR bands were quantified with the FusionCapt Advance software (Vilber Lourmat, Eberhardzell, Germany). The mean and standard deviation of replicated experiments were determined to obtain statistically reliable results.

The degree of contamination suppression was determined by quantifying the visible product of the second PCR. Complete contamination suppression was assumed when the second amplification yielded no visible PCR product.

Testing of contamination protection by the K-box in multiplex two-step PCR analyses

To test the effectiveness of the K-box method in a multiplex TCRβ analysis we first used two different experimental set-ups: (i) runs with matching K1 elements and (ii) runs with mismatching K1 elements simulating e.g. pipetting errors. We used 400 ng tonsillar DNA as template to allow the amplification of high numbers of different TCRβ gene rearrangements. The two-step PCR protocol employed is given as Supplementary Methods. PCR results were analyzed by gel electrophoresis. In the experiments with matching K1 elements the 9 primer sets depicted in Table 1 were used. In line with the K-box principle these 9 sets were characterized by unique k/k′ combinations between the sets and concordant k/k′ combinations within a set. The same two-step PCR workflow was used also in a total of 15 experiments with mismatching K1 elements simulating pipetting mistakes (Supplementary Table S2). To this end the first amplification product of a given set (e.g. set 1) was used as template for amplification with second amplification primers from another (not matching) set (e.g. set 2). Such a simulation of sample confusion can be regarded as 100% contamination in the second PCR mix.

Next, we applied NGS to evaluate the K-box method. We used again the 9 TCRβ multiplex primer sets depicted in Table 1 and a synthetic template. Generation of the synthetic template is illustrated in the Supplementary Methods in Figure S3. This synthetic template covered all TCRβ V-J combinations ($N = 616$) relevant for binding of the 44 V- and 14 J-segment specific TCRβ primers. Using the aforementioned 9 primer sets (Table 1) 18 PCR assays were performed, 9 of which without and 9 assays with spike-in contaminations (see Supplementary Table S3 and Table 3). Spike-in contaminations are here defined as first amplification products with not matching K-boxes to the second amplification primers employed. For example, a mix of eight first amplification products amplified with set 1–8 primers are spike-in contaminations if the second amplification is performed with set 9 primers. If the amplification of the spike-in contaminations is not completely suppressed by the K1 mismatches, residual contaminations can be detected by the unique k2/k2′combinations. The 9 assays with spike-in contaminations were assembled as follows: 60% represented the PCR product of the first amplification with matching K-box combination whereas the remaining 40% were derived from the PCR products of the other 8 first amplifications (5% each) with mismatching K-box combinations (Supplementary Table S3). This first amplification mixture was diluted 1:100 and the second amplification assay was

Table 2. Description of the K-box sequence elements

Abbreviation	Description
K-box	Comprises the sequence elements k1/k1′, k2/k2′, s/s′.
k-box	K-box of the forward first and second amplification primers.
k′-box	K-box of the reverse first and second amplification primers.
k1/k1′ (K1 elements)	K-box elements of the first and second amplification primers for suppression of contaminations.
k2/k2′ (K2 elements)	K-box elements of the first amplification primers. Their main purpose is the detection of contaminations in subsequent bioinformatics steps (NGS data analysis)
s/s′ (S elements)	S-elements are designed as mismatch with the template i.e. the respective genomic TCRβ V- and J-segments. The S elements effectively separate the template matching part of the primers from their 5′- tail thereby reducing the risk of PCR bias.

The typical lengths of K1, K2 and S-elements as used in our TCRβ multiplex analyses are 7, 3 and 2 nucleotides, respectively.

processed as described in the Supplementary Methods. Spin column-purified PCR products were sequenced with MiSeq (Illumina, San Diego, USA) in paired-end mode (2 × 150 bp). For multiplexing standard TruSeq barcodes and MiSeq sequencing default conditions were employed allowing 1 mismatch in the barcode.

Bioinformatics analysis

Resulting reads were clustered and classified with respect to the K-box elements by employing a tailored algorithm: in step 1, reads delivered by the Illumina sequencer in paired-end mode, were joined by an in-house developed optimal assembly procedure, which uses the phred-like quality values per base to deliver a proper consensus sequence per each read pair. The resulting joined read is (i) of higher quality than each of the single reads alone and (ii) is able to deliver effective read-length of > 500 bp, if a current 2×300 bp MiSeq run is used (Supplementary Methods). These joined reads were used in step 2 to find the respective K-box elements by sequence comparison.

In the following we used sequence analysis to quantify the rate of contamination suppression by determining the amount of contamination carried over between different sets. The principles to analyze the experiment outlined in Supplementary Table S3 are given below:

- The k2/k2′ elements are identified in each read pair as part of the respective K-box (V and J side).
- Since the correct k2/k2′ elements are known, as well as the potentially contaminating k2/k2′ elements generated by spike-ins (see above), we are able to precisely determine the respective rates.
- In each experiment one set is given as correct, i.e. there is one defined 'correct' k2 element on V and one correct k2′ on J side. Other sets which are studied in parallel and which are also used as spike-in contaminations can be uniquely identified by their different K2 elements. The respective matches are counted and flagged as 'contaminated'.
- However, in addition to the methodologically introduced K2 elements there are so-called random K2 elements, which can be caused e.g. by sequencing artifacts. These random K2 elements can be simply all possible triplets (61 on the V and 61 on the J side) except the 2×3 K2 elements on the V or J side used in the 9 sets (see Table 1). The respective matches are counted and flagged as 'random'.
- The random K2 elements are complicating the analysis. To obtain a clear-cut read out for each of the experiments described in Supplementary Table S3, we measured the events with simultaneous occurrence of the 2 contaminating k2 elements on the V side and 2 contaminating k2′ elements on the J side. This robust approach detects in each experiment 6 of the 8 spike-in contaminations.
- Furthermore, we united all reads containing random K2 elements on the V and/or the J side and compared the rates in the experiments with and without spike-in contaminations by a standard two-sided *t*-test. The aim was to determine if the random error is equally distributed in both groups.

Comparison of the respective rates will allow a precise measurement of how efficient the K-boxes are in suppressing carry-over contaminations.

RESULTS

Determination of parameters for effective suppression of carry-over contamination

In order to determine the factors needed for effective suppression of carry-over contamination in two-step PCR analyses we analyzed (i) the impact of the number of base pairs of K1 mismatches and (ii) the effect of PT-bonds at the 3′-end of second amplification primers using a proofreading polymerase. For this purpose we used DNA of the T-cell line Peer as a template and amplified in a non-multiplex setting specifically the Peer TCRβ gene rearrangement (Supplementary Figure S1). As visualized in Figure 3 the absence of K1 mismatches led—as expected—to strong amplification products after second amplification (Lane 1) whereas a K1 mismatch of 1 bp was able to strongly reduce the PCR product intensity (Lanes 2 and 3). K1 mismatches of 2 bp (Lane 4) were sufficient to suppress the contaminating second amplification products almost completely. As another important factor an increasing number of PT-bonds at the 3′-end of second amplification primers resulted in improved contamination suppression (Figure 4). Details are given as Supplementary Table S4).

Consequently, conditions for effective carry-over contamination suppression should comprise k1/k1′ mismatches of more than 2 nucleotides and the use of at least 2 PT-bonds at the 3′-end of second amplification primers.

For comparison to the experiments using a proofreading polymerase and primers with PT-bonds in the sec-

Table 3. 18 experiments were performed with or without spike-in contaminations as outlined in Supplementary Table S3

Experiment number	Set-ID	Spike-in contamination	Number of read pairs	Correct reads (%)	Contamination reads (%)	Random reads (%)
1	set1	No	467378	97,5	0	2,5
2	set2	No	514960	98,4	0	1,6
3	set3	No	415902	98,0	0	2,0
4	set4	No	447288	98,0	0	1,9
5	set5	No	454124	98,1	0	1,9
6	set6	No	415391	98,0	0	1,9
7	set7	No	491452	97,9	0	2,1
8	set8	No	540915	98,2	0	1,8
9	set9	No	455199	98,0	0	2,0
10	set1	40%	481099	97,9	0	2,1
11	set2	40%	486962	98,4	0	1,5
12	set3	40%	446743	98,0	0	2,0
13	set4	40%	460226	97,9	0	2,0
14	set5	40%	480041	98,1	0	1,9
15	set6	40%	464568	97,9	0	2,0
16	set7	40%	527910	97,7	0	2,2
17	set8	40%	434073	98,0	0	2,0
18	set9	40%	431367	97,9	0	2,1

The percentages of correct reads and of contaminations were determined according to the respective K2 sequences employed. A contamination rate of 0% indicates that no contaminating read was detectable. Furthermore, the percentage of random reads with unexpected K2-elements occurring stochastically on V/J-side is given.

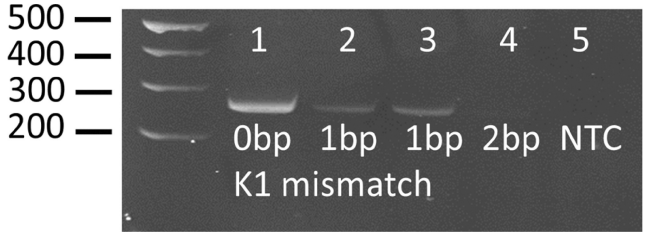

Figure 3. Effect of base pair length of K1 mismatches on contamination suppression. Gel electrophoresis results (Ethidium bromide staining) of two-step PCR demonstrate a positive correlation between length of K1 mismatches and contamination suppression. Whilst under K1 matching conditions (0 bp) the expected prominent readout is achieved (Lane 1), K1 mismatches of only 1 bp have already pronounced negative effects on the second amplification (Lanes 2 and 3). A K1 mismatch of 2 bp prevents almost completely the second amplification (Lane 4). In this experiment 3 PT-bonds were used at the 3′end of the second amplification primers. PCR products of the first amplification were defined as 100% contamination and were used as template for the second amplification. The size of the gel ladder is given in base pairs.

ond amplification we performed experiments with a non-proofreading polymerase in the second amplification and primers without PT-bonds. In both settings the k-box and k′-box contamination suppression increases with an increasing number of K1 mismatches. However, contamination suppression is less effective if non-proofreading polymerases are used (Supplementary Tables S5 and S7). A further advantage of proofreading polymerases is their less error-prone DNA-synthesis.

Testing of contamination protection by the K-box in multiplex two-step PCR analyses

To test the impact of the K-box method for contamination protection in a multiplex TCRβ analysis we utilized in a first experiment 9 primer sets in 9 match and 15 mismatch settings. In these experiments tonsillar DNA was used as a template. The length of K1 mismatches in the mismatch settings was 4-5 base pairs (Supplementary Table S2). As shown in Figure 5 and Supplementary Figure S4 the matched settings led to the production of the expected PCR products. On the other hand the K1 mismatch settings between first and second amplification primers completely suppressed the amplification during the second PCR.

In order to demonstrate the effectiveness of contamination protection by a highly sensitive NGS we investigated the 9 primer sets in 9 match settings and in 9 assays with matching sets plus spike-in contaminations (Supplementary Table S3). For these experiments an artificial template was used (Supplementary Figure S3).

In total, 18 NGS libraries were prepared, which were tagged by standard Illumina barcodes. The libraries were sequenced on Illumina MiSeq and demultiplexed by Illumina software (default parameters). For 8.52 % of the reads the barcode could not be identified unambiguously. These reads were excluded from further analysis. Another quality filtering step was imposed by our read-joining algorithm which requires that both reads of each read pair exhibit significant overlaps. 94.8% ± 1.81% of all read pairs could by joined successfully to one consensus sequence. Finally, the joined reads were classified and filtered again for occurrence of possible k1/k1′-boxes, which was the case for 90.6% ± 2.4%. In the end, compared to the original raw read numbers, on average 83.5% ± 2.2% of reads were used for the data analysis, which was equivalent to around 8.5 million read pairs.

Our observations are summarized in the following and in Table 3:

- Most importantly, evaluation of NGS results revealed complete suppression of spike-in contaminations.
- We compared the percentages of random reads in the runs without spike-ins to the ones with 40% effective contamination by a standard two-sided *t*-test. The resulting

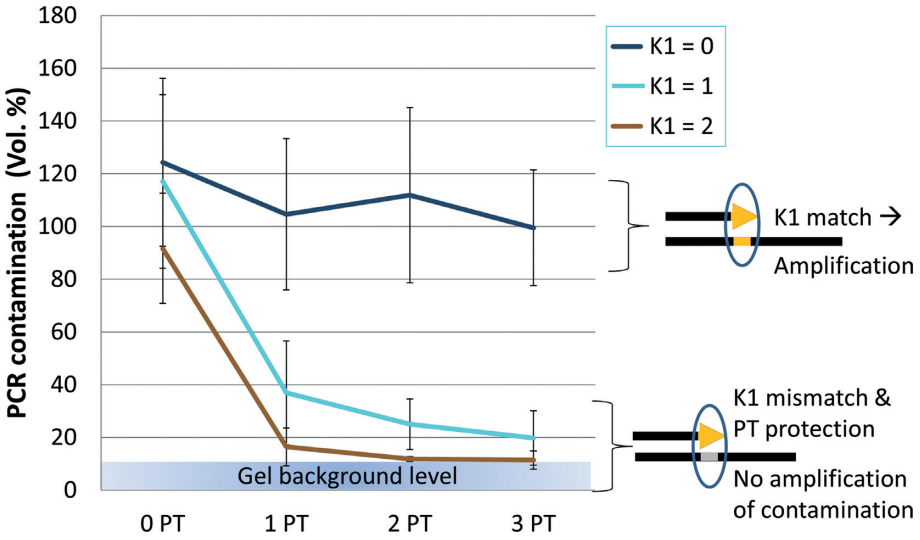

Figure 4. Impact of K1 mismatch length and number of PT-bonds in the second amplification primer on contamination suppression. The base pair length of K1 mismatch sequences and the number of PT-bonds was varied in two-step PCR assays. The quantity of PCR products from the second PCR are shown as volume percent (Vol. %) which were determined as outlined in Supplementary Figure S2 and the results were statistically analyzed as shown in Supplementary Table S5. The base pair length of K1 mismatches and the number of PT-bonds positively correlate with suppression of contaminations. The mean for the NTCs ($N = 16$) and background ($N = 16$) measurements was 9.9 ± 1.9 Vol. % and 9.4 ± 2.3 Vol. % respectively.

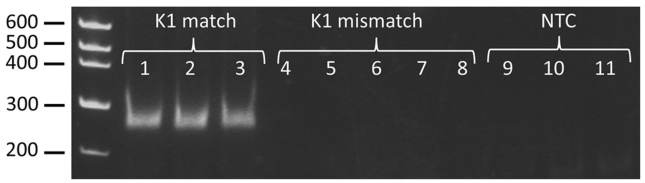

Figure 5. Testing of contamination protection by the K-box in two-step PCR TCRβ analyses. Results analyzed by gel electrophoresis (Ethidium bromide staining) for 3 match settings, 5 mismatch settings and 3 NTCs are shown. In the match setting the expected TCRβ amplification product was observed whereas in the mismatch settings amplification was completely suppressed. The experimental setup for the PCR reactions 1–11 is given in Supplementary Table S2. The size of the gel ladder is given in base pairs.

P-value is not significant ($P = 0.8$) which means that the random errors occurred at equal rates in both groups.

- The design of our primer sets with integrated K-boxes leads to a highly efficient reduction of contamination in the secondary PCR step, and, if necessary, allows detection and of course computational removal of possibly persisting minor contaminations by bioinformatics sequence analysis.

DISCUSSION

PCR-based systems are known to be vulnerable to false positive results caused by contaminations with products from previous amplifications (2,3). Sources of contaminating DNA include assay reagents, equipment, operator handling and aerosols (2,3). The rapid development of NGS technologies permits a much higher sensitivity than most of the previously established methods. However, this sensitivity may lead to the detection of previously unrecognized cross-over contaminations from other samples. This is especially critical in the case of assays which require the reliable detection of just a few amplicons (e.g. minimal residual disease (MRD)). MRD is an important prognostic factor and has clinical impact (13–15).

Two-step PCR NGS library preparations are used for various applications such as targeted re-sequencing of cancer genes, 16S microbiome analyses or the analysis of B- and T-cell receptor genes (4–8). As described by Faham and Willis amplification bias may be decreased by reducing the number of amplification cycles in a first step using primers with tails non-complementary to the target sequences (16). These tails include primer binding sites that are added to the ends of the sequences of the primary amplicon and are specifically targeted in the second amplification step. By using only a single forward and a single reverse primer, the primary cause of amplification bias is confined (16).

Here we introduce a novel method to effectively prevent contaminations in two-step PCR systems. This new approach sharply contrasts with previously described methods which are merely able to detect contaminations or to prevent first round PCR contaminations (17–21). Our approach is additionally capable to prevent contaminations in the second PCR (Table 4). Especially the most commonly used UTP/UNG system where dTTP is replaced by dUTP in the PCR mixture leading to degradation of unintentionally transferred DNAs into the samples is not able to detect or suppress contaminations present in the second PCR (17–19). The reason is that the products from the first reaction would be cleaved by an UNG treatment in the second reaction. This is especially true with bisulfite amplification / sequencing, where conversion to U precludes using UDG in the first place.

An alternative method to detect amplicon cross-contaminations is the use of NGS adaptor-specific primers (20,21). These primers anneal to NGS adaptor parts being present in the first and second PCR products (e.g. a M13 tail). By this method, cross-contaminations from the first

Table 4. Cross-contaminations and their possible prevention in two-step PCR settings for NGS library generation

	Possible contamination site	
Source of contamination	First amplification PCR mix	Second amplification PCR mix
Amplicon from the first PCR	(A) Prevention: UTP/UNG system Detection: PCR with first amplification tail-specific primers	(C) Prevention and detection: K-box protection (this manuscript)
Amplicon from the second PCR	(B) Prevention: UTP/UNG system Detection: PCR with first or second amplification tail- or adaptor-specific primers	(D) Prevention and detection: K-box protection (this manuscript)

A & B: Contaminations of the first amplification by PCR products derived from another first or second amplification can be prevented by the UTP/UNG system and detected by tail- or adaptor-specific primers (19–21). C & D: Contaminations of the second amplification by PCR products derived from another first or second amplification can be prevented and detected by the K-box protection method described in this manuscript.

or second PCR amplicons, either in the template or the PCR reagents, can be detected but not prevented.

Our K-box method as described in this manuscript is a synergistic combination of suppression of carry-over contaminations and detection of possible residual contaminations in the two-step PCR workflow. The design of the sequence elements of this K-box is sample-specific. Specific amplification is only possible in case of a perfect match of the K1 elements used for first and second amplification. K2 sequences serve for detection of possible residual contaminations.

We found that our method is most effective with a proofreading polymerase in the second amplification providing the additional advantage of amplification with lesser errors. In order to fully prevent cross-contaminations in the second PCR, mismatches between K1 sequence elements of more than 2 nucleotides should be used in addition to at least 2 PT-bonds at the 3′-end of second amplification primers (Figures 3 and 4).

Applying these K-box parameters we showed that in two step NGS library preparations a high contamination grade of up to 100% was completely suppressed (Figure 5, Table 3 and Supplementary Figure S4). Thus, our method is not only able to prevent low level contaminations but also very high rates of contaminations which can be caused by confusing samples.

The suppression of contaminations is a much more elegant approach as compared to mere detection of contaminations. The latter approach necessitates time-consuming repetition of the entire procedure whereas the K-box method is perfectly integrated in the two-step PCR workflow without additional time requirements. Especially for diagnostic amplicon-based NGS the avoidance of repetition represents an important advantage due to the time pressure associated with this type of analysis.

Examples for distinct practical applications of the K-box method are (i) the analysis of consecutive samples from one patient in MRD diagnostics, (ii) targeted re-sequencing of cancer genes or (iii) 16S microbiome analyses. The K-box method can be adapted to high throughput formats, e.g. a two-step PCR performed in 96 well plates. Thereby, K2 elements are essential to detect possible remaining contaminations from previous amplification reactions and thus certify the efficiency of contamination suppression by K1 elements. If the effective contamination suppression by K1 elements is firmly established for the given primer sets, a reduction to a sole K1-S combination (without K2 elements) may be considered.

In summary, we describe a new method capable of suppressing contaminations and/or detecting residual contaminations in NGS library preparations employing a two-step PCR protocol. This new method is based on three synergistically acting sequence elements referred to as K-box. The K-box method provides a significant improvement for NGS diagnostics in terms of accuracy, time requirement and costs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank Anke Seegebarth for her excellent technical assistance, Julia Ritter for help with primer design and Rudolf Hammer for fruitful discussions in preparation of the manuscript.

FUNDING

Investitionsbank Berlin and the European Regional Development Fund [10155447 and 10155355 to the Charité and HS Diagnomics, respectively].

Conflict of interest statement. S.H. is CEO as well as cofounder of HS Diagnomics, and S.S., A.D. and V.S. are shareholders of HS Diagnomics. V.S., M.H., S.H. are inventors of a patent (EP 2746405 A1) which describes the K-box system. D.L. has no conflict of interest to disclose.

REFERENCES

1. Baetens,M., Van Laer,L., De Leeneer,K., Hellemans,J., De Schrijver,J., Van De Voorde,H., Renard,M., Dietz,H., Lacro,R.V., Menten,B. *et al.* (2011) Applying massive parallel sequencing to molecular diagnosis of Marfan and Loeys-Dietz syndromes. *Hum. Mutat.*, **32**, 1053–1062.
2. Asbury,T., Carlton,V., Faham,M., Macevicz,S., Moorhead,M., Wills,T. and Zheng,J. (2013) Detection and Quantification of Sample Contamination in Immune Repertoire Analysis. Patent WO2013155119 A1.
3. Urban,C., Gruber,F., Kundi,M., Falkner,F.G., Dorner,F. and Hammerle,T. (2000) A systematic and quantitative analysis of PCR template contamination. *J. Forensic Sci.*, **45**, 1307–1311.

4. Hullein,J., Jethwa,A., Stolz,T., Blume,C., Sellner,L., Sill,M., Langer,C., Jauch,A., Paruzynski,A., von Kalle,C. *et al.* (2013) Next-generation sequencing of cancer consensus genes in lymphoma. *Leuk. Lymphoma*, **54**, 1831–1835.
5. Logan,A.C., Zhang,B., Narasimhan,B., Carlton,V., Zheng,J., Moorhead,M., Krampf,M.R., Jones,C.D., Waqar,A.N., Faham,M. *et al.* (2013) Minimal residual disease quantification using consensus primers and high-throughput IGH sequencing predicts post-transplant relapse in chronic lymphocytic leukemia. *Leukemia*, **27**, 1659–1665.
6. Hadd,A.G., Houghton,J., Choudhary,A., Sah,S., Chen,L., Marko,A.C., Sanford,T., Buddavarapu,K., Krosting,J., Garmire,L. *et al.* (2013) Targeted, high-depth, next-generation sequencing of cancer genes in formalin-fixed, paraffin-embedded and fine-needle aspiration tumor specimens. *J. Mol. Diagn.*, **15**, 234–247.
7. Liang,S., Gliniewicz,K., Mendes-Soares,H., Settles,M.L., Forney,L.J., Coats,E.R. and McDonald,A.G. (2014) Comparative analysis of microbial community of novel lactic acid fermentation inoculated with different undefined mixed cultures. *Bioresour. Technol.*, **179**, 268–274.
8. Camarinha-Silva,A., Jauregui,R., Chaves-Moreno,D., Oxley,A.P., Schaumburg,F., Becker,K., Wos-Oxley,M.L. and Pieper,D.H. (2014) Comparing the anterior nare bacterial community of two discrete human populations using Illumina amplicon sequencing. *Environ. Microbiol.*, **16**, 2939–2952.
9. Zinkernagel,R.M. (1997) The Nobel Lectures in Immunology. The Nobel Prize for Physiology or Medicine, 1996. Cellular immune recognition and the biological role of major transplantation antigens. *Scand. J. Immunol.*, **46**, 421–436.
10. Murphy,K., Travers,P., Walport,M. and Janeway,C. (2012) *Janeway's immunobiology*. Garland Science, NY.
11. Robins,H.S., Campregher,P.V., Srivastava,S.K., Wacher,A., Turtle,C.J., Kahsai,O., Riddell,S.R., Warren,E.H. and Carlson,C.S. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*, **114**, 4099–4107.
12. Strominger,J.L. (1989) Developmental biology of T cell receptors. *Science*, **244**, 943–950.
13. Campana,D. (2009) Role of minimal residual disease monitoring in adult and pediatric acute lymphoblastic leukemia. *Hematol. Oncol. Clin. North Am.*, **23**, 1083–1098.
14. Buccisano,F., Maurillo,L., Del Principe,M.I., Del Poeta,G., Sconocchia,G., Lo-Coco,F., Arcese,W., Amadori,S. and Venditti,A. (2012) Prognostic and therapeutic implications of minimal residual disease detection in acute myeloid leukemia. *Blood*, **119**, 332–341.
15. Szczepanski,T., Flohr,T., van der Velden,V.H., Bartram,C.R. and van Dongen,J.J. (2002) Molecular monitoring of residual disease using antigen receptor genes in childhood acute lymphoblastic leukaemia. *Best Pract. Res. Clin. Haematol.*, **15**, 37–57.
16. Faham,M. and Willis,T. (2011) Monitoring Health and Disease Status using Clonotype Profiles. Patent US20110207134 A1.
17. Borst,A., Box,A.T. and Fluit,A.C. (2004) False-positive results and contamination in nucleic acid amplification assays: suggestions for a prevent and destroy strategy. *Eur. J. Clin. Microbiol. Infect. Dis.*, **23**, 289–299.
18. Niederhauser,C., Hofelein,C., Wegmuller,B., Luthy,J. and Candrian,U. (1994) Reliability of PCR decontamination systems. *PCR Methods Appl.*, **4**, 117–123.
19. Kox,L.F., Rhienthong,D., Miranda,A.M., Udomsantisuk,N., Ellis,K., van Leeuwen,J., van Heusden,S., Kuijper,S. and Kolk,A.H. (1994) A more reliable PCR for detection of Mycobacterium tuberculosis in clinical samples. *J. Clin. Microbiol.*, **32**, 672–678.
20. Shuber,A.P. (1999) Methods for Detecting Contamination in molecular Diagnostics using PCR. Patent WO/1999/020798.
21. Abbott,L.Z., Spicer,T., Bryz-Gornia,V., Kwok,S., Sninsky,J. and Poiesz,B. (1994) Design and use of signature primers to detect carry-over of amplified material. *J. Virol. Methods*, **46**, 51–59.