Bioinformatics, 34(16), 2018, 2732–2739
doi: 10.1093/bioinformatics/bty145
Advance Access Publication Date: 10 March 2018
Original Paper

Sequence analysis

Nimbus: a design-driven analyses suite for amplicon-based NGS data

R. W. W. Brouwer,[1] M. C. G. N. van den Hout,[1] C. E. M. Kockx,[1] E. Brosens,[2] B. Eussen,[2] A. de Klein,[2] F. Sleutels[1] and W. F. J. van IJcken[1,]*

[1]Center for Biomics, Department of Cell Biology and [2]Department of Clinical Genetics, Erasmus MC, PO Box 2040, 3000CA Rotterdam, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on September 1, 2017; revised on February 14, 2018; editorial decision on March 6, 2018; accepted on March 9, 2018

Abstract

Motivation: PCR-based DNA enrichment followed by massively parallel sequencing is a straightforward and cost effective method to sequence genes up to high depth. The full potential of amplicon-based sequencing assays is currently not achieved as analysis methods do not take into account the source amplicons of the detected variants. Tracking the source amplicons has the potential to identify systematic biases, enhance variant calling and improve the designs of future assays.
Results: We present Nimbus, a software suite for the analysis of amplicon-based sequencing data. Nimbus includes tools for data pre-processing, alignment, single nucleotide polymorphism (SNP), insertion and deletion calling, quality control and visualization. Nimbus can detect SNPs in its alignment seeds and reduces alignment issues by the usage of decoy amplicons. Tracking the amplicons throughout analysis allows easy and fast design optimization by amplicon performance comparison. It enables detection of probable false positive variants present in a single amplicon from real variants present in multiple amplicons and provides multiple sample visualization. Nimbus was tested using HaloPlex Exome datasets and outperforms other callers for low-frequency variants. The variants called by Nimbus were highly concordant between twin samples and SNP-arrays. The Nimbus suite provides an end-to-end solution for variant calling, design optimization and visualization of amplicon-derived next-generation sequencing datasets.
Availability and implementation: https://github.com/erasmus-center-for-biomics/Nimbus.
Contact: w.vanijcken@erasmusmc.nl
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput identification of deleterious variants in clinical studies has become both technically and economically feasible through the enrichment of specific DNA fragments from the patients genome (Gnirke *et al.*, 2009; Hodges *et al.*, 2007), followed by massively parallel DNA sequencing (Bentley *et al.*, 2008). Enrichment of DNA fragments of interest is either performed by PCR-based amplification or by hybridization capture. PCR-based amplicon assays range from traditional multiplexed PCRs that interrogate a couple of amplicons to more sophisticated technologies that target all coding sequences. Examples of techniques that rely on amplicons include TruSeq Custom Amplicon, HEAT-Seq, GeneRead panels, molecular inversion probes (Hardenbol *et al.*, 2003), pxlence, Multiplicon MASTR and HaloPlex assays. Amplicon-based enrichment assays are popular because they are relatively straightforward and cost effective.

The HaloPlex exome (Agilent technologies, Santa Clara, CA, USA) is one of the most sophisticated examples of amplicon-based enrichment technologies (Dahl *et al.*, 2007; Johansson *et al.*, 2011). Its design consists of over 2 million amplicons targeting most coding exons in the human genome. In the HaloPlex procedure, amplicons are generated by digesting genomic DNA with eight sets of restriction enzymes. Specific restriction fragments are then targeted, enriched through PCR and sequenced. Unlike data from hybridization-based capture, the reads from amplicon-based next-generation sequencing (NGS) experiments are not randomly distributed over the target. Instead the read location is dictated by the start and end positions of the amplicons. In larger designs, such as the HaloPlex exome, a location of interest is covered by multiple overlapping amplicons.

Usually, amplicon-based data are analysed for single nucleotide polymorphism (SNP)/(insertion and deletion (InDel)) detection and copy number variations (Koopmans *et al.*, 2014; Li *et al.*, 2009b; McKenna *et al.*, 2010). A great variety of panels are available that target specific SNPs in disease-associated genes. In traditional variant detection (Li *et al.*, 2009b; Plagnol *et al.*, 2012) the reads are aggregated per region of interest (e.g. an exon). Overlapping amplicons allow for multiple measurements per exon and may thus increase the statistical significance and resolution of variant detection. However, sequencing data acquired from amplicon-based enrichments is currently not capitalized to its full potential as only few analysis tools (Caporaso *et al.*, 2010; Lai *et al.*, 2016) take the amplicon design into account.

Here we report Nimbus, a software suite for the analysis of amplicon-based sequencing data. Nimbus tracks the source amplicons throughout alignment and SNP, insertion and deletion calling. We tested Nimbus on Agilent's HaloPlex exome, the largest amplicon-based design that is currently commercially available.

We demonstrate that exploiting the amplicon design enables (i) filtering of aligned reads that do not originate from the target by using decoy amplicons, (ii) detection of low performance amplicons in the design, (iii) assessment of confidence based on the number of amplicons supporting a variant call and (iv) detection of false-positive variants that are only present in a single amplicon.

2 Materials and methods

2.1 Data sources

To test the tools described in this manuscript, six samples were used (Table 1). All HaloPlex exome data have been deposited with the Sequence Read Archive under BioProject PRJNA 393963, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA393963. These samples consist of two pairs of mono-zygotic twins and two non-twin samples. Written (parental) consent was obtained. Genetic tests were performed according to The Erasmus University Medical Center's local ethics board approved protocol no.193.948/2000/159, addendum Nos. 1 and 2. One sample (NA15510) is obtained from NIGMS Human Genetic Cell Repository at the Coriell institute for medical research (Camden, NJ, USA).

DNA was captured with the HaloPlex exome method (Agilent technologies, Santa Clara, CA, USA) according to manufacturer's protocol. The resulting DNA libraries were sequenced on a HiSeq2000 system (Illumina, San Diego, CA, USA) using the TruSeq V3 paired-end 100 base pair sequencing protocol. Sequencing yielded between 41 and 59 million reads per sample (Table 1).

In addition to HaloPlex exome sequencing, the twin samples were also assayed on Illumina HumanExome-12 BeadChip microarray (Illumina, San Diego, CA, USA). The samples were prepared according to the assay protocol prescribed by the manufacturer. The resulting data was processed using the GenomeStudio software (Illumina, San Diego, CA, USA) version V2011.1.

Table 1. Available datasets

Sample	Source	Sex	Twin pair	Reads
NA15510	Cell line	Female	—	41 417 946
Sample 1	Blood	Female	1	54 200 440
Sample 2	Blood	Female	1	51 604 990
Sample 3	Blood	Female	2	58 968 772
Sample 4	Blood	Female	2	55 592 442
Sample 5	Blood	Male	—	51 612 528

For sample NA15510, a truth set of known variants was generated by determining genotypes that were shared between a SureSelect clinical research exome (CRE) exome dataset and a Roche MEDExome dataset (PRJNA393963). Variants are only included in the truth set if they are called by both FreeBayes (Garrison and Marth, 2012) and the GATK HaplotypeCaller (McKenna *et al.*, 2010), are covered by 20 or more reads and have a frequency between 0.4 and 0.6 for heterozygous calls and over 0.99 for homozygous calls.

To simulate somatic variants homozygous SNPs in sample 1 were selected that are absent in NA15510. For each of the 276 selected SNPs, 4 reads were added to NA15510 resulting in expected alternate allele frequencies between 1 and 40% (see Supplementary Material S1). Subsequently, variants were called using Nimbus, FreeBayes, GATK HaplotypeCaller and VarDict (Lai *et al.*, 2016). The expected and actual frequencies of the simulated SNPs were compared to determine the detection limits per variant caller.

Datasets were aligned to the human reference genome (version hg19) (Lander *et al.*, 2001) obtained from the UCSC (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/). The locations of the amplicons targeted by the HaloPlex exome design were retrieved from the manufacturer via the SureDesign web-platform (https://earray.chem.agilent.com/suredesign/). In the HaloPlex exome, over 2 million amplicons are targeted via hybridization/ligation strategy of restriction fragments. This strategy may cause some off-target restriction fragments to be captured. To account for the off-target capture, approximately 739 thousand decoy amplicons were added to the exome design.

2.2 HaloPlex design expansion with decoy amplicons

During the HaloPlex sample preparation, genomic DNA is sheared using mixes of specific restriction enzymes. The resulting DNA fragments are captured and ligated to partial double stranded probes. The specificity for this capture is derived from complementary overhangs in the probes that hybridize to the targets. The subsequent ligation is only efficient for DNA fragments with highly homologous ends to the capture probes. The resulting circular DNA molecule is then amplified and processed further.

Nimbus generates possible decoy amplicons by matching the ends of the original amplicons to the ends of all possible restriction fragments from the genome. A genomic restriction fragment is considered a possible decoy amplicons if its outer 5 bases match a fragment in the design perfectly and the 6 bases further in match with at most 1 mismatch. Candidate decoy amplicons are not considered if they are shorter than 50 bases or larger than 600 bases in length.

A detailed guide on decoy amplicon creation and the scripts used are available at https://github.com/erasmus-center-for-biomics/Nimbus.

2.3 Nimbus tools

The Nimbus tools consist of a set of tools to process amplicon-based sequencing data. A prerequisite for datasets to be analysed with

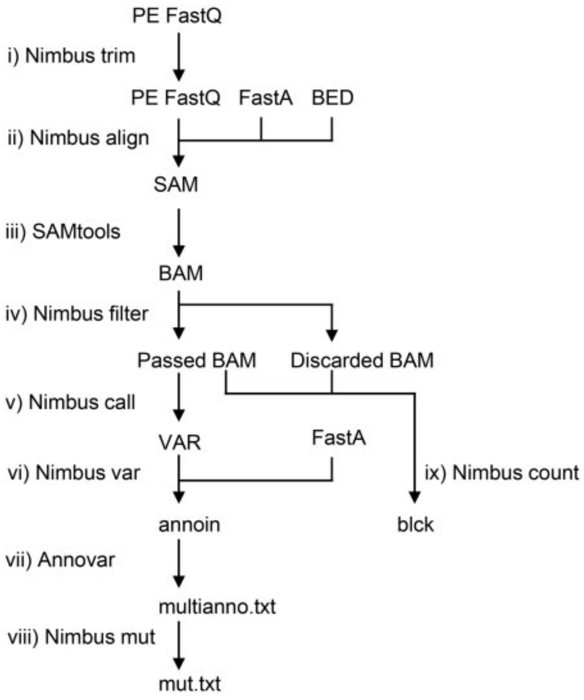


Fig. 1. Nimbus analysis workflow. (i) Reads are first trimmed using Nimbus trim and (ii) subsequently aligned to the reference design using Nimbus align. (iii) The resulting SAM file is sorted and converted to BAM format using SAMtools. (iv) The alignments with four or more differences are copied to the discarded BAM files. Alignments with fewer than four differences are written to the passed BAM files that are used in the downstream analysis. The threshold for filtering can be adjusted. (v) Nimbus call records all differences between the reference sequence and the (passed) alignments in a custom var format. (vi) From these var files, variants are distilled by Nimbus var and reported in the tab-delimited ANNOVAR input format. (vii) The variants files are annotated with ANNOVAR (13) and (viii) converted to the mut.txt format which is readable by the IGV (14). (ix) In parallel with the variant calling, the read depths per amplicon are determined by Nimbus count from both the passed and discarded alignments. This information is recorded in blck files which can be used to determine amplicon performance

Nimbus is known start and end locations of the reads in the reference sequence. Nimbus contains tools to trim, align, count, filter and call variants in amplicon-based datasets (Fig. 1). Nimbus trim detects partial or complete (Illumina adapter) sequences and removes the adapter and subsequent bases. To analyse sequencing data Nimbus makes use of existing open source tools, such as SAMtools, ANNOVAR and integrative genome viewer (IGV) (Li *et al.*, 2009a; Robinson *et al.*, 2011; Wang *et al.*, 2010). The standard workflow to process samples with Nimbus has been implemented in a series of Makefiles (https://www.gnu.org/software/make/).

The tools are implemented in either C++ or Python and are dependent on third-party libraries such as C++ boost (http://www.boost.org), htslib (http://www.htslib.org/) and/or pysam (https://github.com/pysam-developers/pysam).

On our system, the Nimbus workflow has a runtime comparable to the BWA/GATK best practices workflow for exome datasets. Nimbus scales in part with the number of amplicons, so for datasets with the same number of reads and a limited number of amplicons, Nimbus takes far less time. The Nimbus tools, implementation instructions and performance comparisons are available in github repository: https://github.com/erasmus-center-for-biomics/Nimbus.

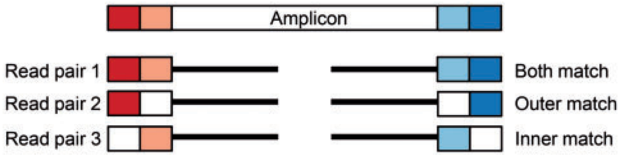


Fig. 2. Reads that match an amplicon during the seeding stage in Nimbus align. Only reads where both ends, either inner or outer, align to the amplicon sequence are selected for alignment to the reference genome and are included in the alignment output file. In all other cases the read pairs, where one or more seeds do not align, are discarded. Matching *k*-mers are indicated with identical colours

The variant calls made by Nimbus are compared with calls made with both the GATK HaplotypeCaller, VarDict and FreeBayes. For these calls, GATK version 3.7, Picard version 2.9.2, VarDict version 1.5.1 and FreeBayes version 1.1.0 (downloaded November 27, 2017) were used. The resulting variant calls were generated and compared as described in Supplementary Material S1.

2.4 Read alignment

Nimbus uses as input a BEDfile with all amplicon, a FASTA file with the genome sequence and two FASTQ files with the sequence reads (Fig. 1). Nimbus performs alignment in three steps. In the first step the beginning of the sequencing reads are perfectly matched to the amplicon ends, which are derived from the genome sequence (Fig. 2). Typically the first seven bases of the reads are matched to the seven bases of the amplicon ends. To overcome match failures due to SNVs in the beginning of the sequence reads, a second match is performed *k* bases inwards in the second step. The amplicons are tracked in memory. In the third step the reads are aligned to the matched amplicons with the Smith-Waterman algorithm (Smith and Waterman, 1981). The alignment with the highest score together with the amplicon is reported in the SAM output file. Alignment parameters can be set via the command-line options.

As Nimbus aligns reads to the amplicon in the design, off-target reads may map to amplicons from which they did not originate. This issue can be minimized by including potential off-target amplicons in the in silico design or by filtering reads that show many mismatches. Many aligners including Nimbus report the mismatch between the alignment and the reference with the Levenshtein or edit distance in the NM tag (Levenshtein, 1966). This score increases with each base in an insertion, deletion or a mismatch. In Nimbus, alignments are not filtered based on the edit distance, but on the number of mismatch events. Each SNP, insertion or deletion increases this score by one. By filtering alignments based this score, alignments with an excessive number of mismatches are discarded without producing an inherent disadvantage to insertions or deletions. The discarded alignments are reported in a separate BAM file. Passed alignments are copied to the passed BAM file. In all our analyses we used a threshold of 4 differences.

2.5 Nimbus call

The Nimbus caller calls variants based on the passed alignments. Variants are called in two stages. In the first stage, differences between the alignments and the reference sequence are reported in a single variant list. This list contains all the alleles (both reference and alternate) between the alignments and reference sequence that meet the calling criteria. Entries in the variant lists represent calls summarized by sequence, sample, strand and the source amplicon. Typically, only non-reference alleles are called that have at least one read, but other criteria concerning the minimum quality, frequency and depth of the alternate alleles can be set. The default analysis does not filter variants for a minimum number of reads, but all variants are called with a frequency of 15% or more and a minimum quality of 200. The minimum quality of 200 corresponds to 6 reads with a PHRED score between 30 and 40 at the position of the variant. Multiple samples can be called simultaneously whereby all the alleles are reported of all samples if the criteria are met in a single sample. The exact sequence content is reported in all samples on that position in the reference. Thus the distinction can be made between alternate alleles being completely absent in a sample or lowly abundant. Nimbus VAR includes the amplicon annotation from Nimbus align. Variant analyses per amplicon and downstream classification are performed in R (scripts provided in Supplementary Material S2).

In the second stage, the variant list is converted to a tab-delimited table in ANNOVAR input format. These files are annotated with ANNOVAR (Wang *et al.*, 2010). For visualization purposes, the output of ANNOVAR is converted to the mutation file format (*.mut.txt) with Nimbus mut. The mutation files can be visualized by the IGV (Robinson *et al.*, 2011). The IGV assigns colours to the variants based on the labels in the mutation type column. Nimbus mut assigns labels to the mutations based on ‘exonic function’ (ExonicFunc) and ‘function in reference gene’ (Func.refGene) columns from the ANNOVAR annotation. Splicing mutations are obtained specifically from the ‘function in reference gene’ column.

In the analyses below, genotypes are imputed from the ANNOVAR input files based on the frequency of the alternate allele. A frequency of 0.05 or lower is called as reference, 0.3–0.7 as heterozygous, and 0.95 and greater as homozygous, which are mainly appropriate for germline variant detection. Other frequencies are called as atypical and not included in downstream analyses.

3 Results

3.1 Alignment

We constructed the Nimbus tools, which are specifically designed to align reads and call variants for amplicon-based data. We performed HaloPlex exome capture and sequencing on six samples and analysed them with Nimbus to demonstrate its capabilities and advantages for amplicon-based data. Nimbus aligns reads to the target amplicons and not to the whole genome. As the search space for amplicons is substantially smaller, more mismatches are permitted in the Nimbus alignment. In the case of the HaloPlex exome analysis, the search space is reduced to 2.8 million positions instead of the 3.2 billion bases in the human genome. The drawback of alignment to the target amplicons is that reads either fail to align due to mismatches in the alignment seed or are forced to align to the amplicons while they actually originate from off-target locations. Nimbus employs three strategies to improve its alignment results (i) double seed-based alignment, (ii) filtering of reads with many mismatches and (iii) filtering by decoy amplicons. The effect of those three strategies is visualized for our test samples (Fig. 3). About 98% of the sequencing data is aligned to the reference genome. The double seed-based alignment rescues ~6% of the reads.

Of the 98% aligned reads, ~3% was filtered as they contained 4 or more mismatches to the reference genome indicative for alignment errors (Fig. 3). The goal of this filter is to remove aligned reads that required many unrelated permutations to fit on the reference sequence. This filter considers insertions and deletions as a single mismatch event irrespective of their length. Finally, about 1% of the reads is mapped to decoy amplicons preventing inadvertently enriched off-target reads to align to the amplicon design and can cause false-positive variants. All in all, ~94% of the reads are mapped to amplicons with a high confidence. These form the basis for downstream analyses such as variant calling.

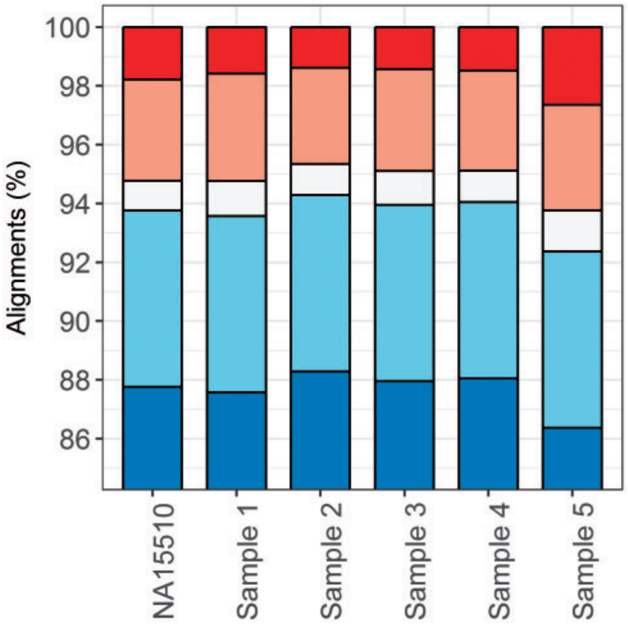


Fig. 3. Impact of the alignment improvement strategies. The alignment percentages are shown for five distinct categories from bottom to top: aligned reads (blue), reads rescued by the double seeding strategy (light blue), reads filtered by decoy amplicons (white), reads filtered due to high number of mismatches (light red), not aligned reads (red) (Color version of this figure is available at *Bioinformatics* online.)

3.2 Short nucleotide variants

The primary goal for most amplicon studies is to detect short nucleotide variants such as SNPs and InDels. To facilitate variant calling in amplicon-derived sequencing datasets, the Nimbus variant caller has been developed. This tool retains the amplicon annotation throughout the variant calling process. Thus, it allows to trace calls back to the original amplicon even if multiple amplicons overlap the same genomic position. Variants can be called in multiple samples at once to allow for direct comparisons. Nimbus call therefore allows to determine which variants are present solely in the index samples and completely absent in unaffected individuals or lowly abundant. This functionality is useful for example in studies where the index patient and its parents are sequenced and the variant is expected to be *de novo*. All deviations from the reference sequence are reported. Filters are applied to remove low-quality variants. Genotypes are imputed by applying filters on the alternate allele frequencies as described in the materials and methods.

If we compare the genotypes called by Nimbus from HaloPlex for sample NA15510 with a ‘truth’ set of genotypes generated from CRE and MEDexome datasets for the sample, the true positive rate of Nimbus 0.97 and a false negative rate of 0.03 (see Supplementary Material S3). The false positive and True negative rates were not determined as the ‘truth’ set does not contain reference calls. Relaxing the call criteria for heterozygous variants from 0.3–0.7 to 0.2–0.8 increases the true positive rate to 0.991 and decreases the false negative rate to 0.009.

Table 2. HaloPlex exome to HumanExome SNP-array concordance

Label	Positions	Same genotype	Same genotype (%)
Sample 1	13515	13087	96.83
Sample 2	13508	13148	97.33
Sample 3	13523	13214	97.72
Sample 4	13528	13204	97.60

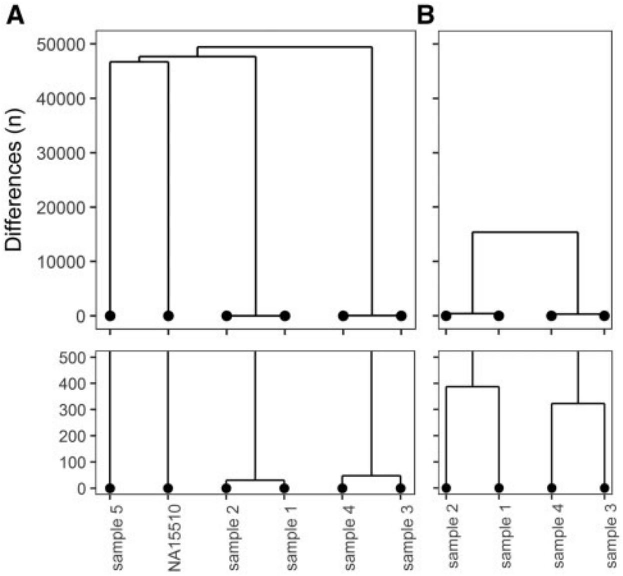


Fig. 4. Concordance between samples. Comparative trees showing **(A)** The number of different genotype calls between the samples based on the sequencing data. (**B**) the number of different genotype calls based on Illumina HumanExome v12 SNP-arrays. In the bottom panels, the trees are zoomed in to 0–500 genotype differences

Genotypes from HaloPlex exome called by Nimbus were compared with genotypes derived from HumanExome-12 BeadChip microarray data to determine to the concordance. Of the 242 901 genotypes queried on the arrays, ~13 500 overlapped with Nimbus computed/called genotypes (Table 2). Over 96% of the shared calls were concordant between the SNP-arrays and HaloPlex exome data. From the verification with DNA microarray data, we conclude that variant calling performed by Nimbus is both accurate and reproducible.

The variants in all six exome samples were called together to compare them directly. As expected, the monozygotic twin samples had much fewer differences when compared with the non-related samples (Fig. 4A). Of the considered genotypes (quality over 600 and a valid genotype), only 30 genotype calls differed between Samples 1 and 2. From Samples 3 and 4, a total of 48 differing genotypes were observed. These differences are three orders of magnitude smaller than when unrelated samples are considered. For example, between NA15510 and sample 1 have 47 680 discordant calls. The number of differences between the twin members is smaller compared with the DNA microarray data (Fig. 4B). In the HumanExome SNP-array data, 323 and 387 different genotype calls were made in the twins. Whereas unrelated samples differed in approximately 15 000 calls (~2.3 * 10^{-2} errors per call). The variant calling performed by Nimbus outperformed a widely used commercially available SNP-array.

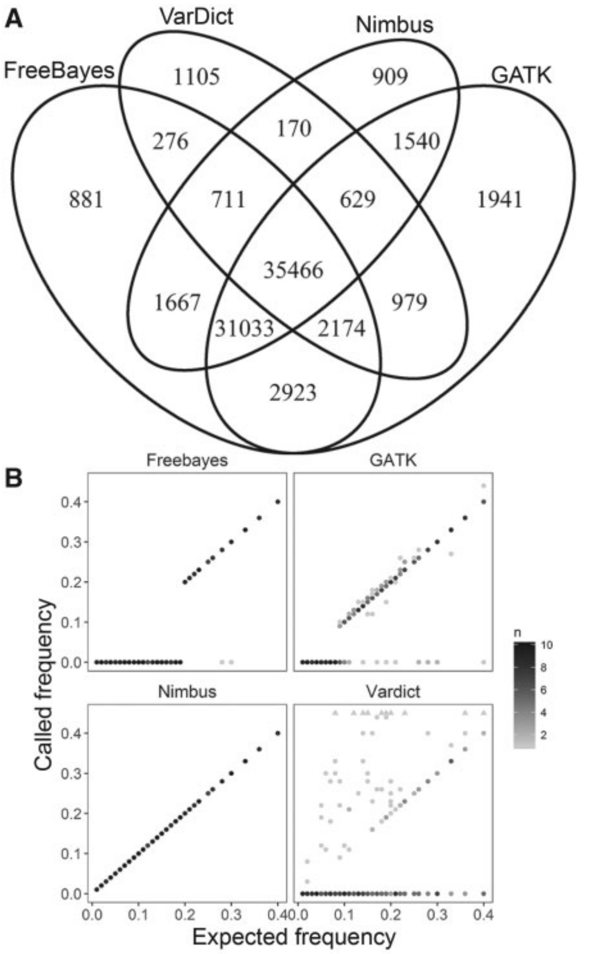


Fig. 5. (**A**) SNP concordance. SNPs were called with Nimbus, VarDict, GATK and FreeBayes and compared with each other. InDels and SNPs with fewer than five reads for the alternate allele were omitted from this analysis. (**B**) Low-frequency variant detection Variants with frequencies ranging from 0 to 40% were added to NA15510 and called with Nimbus, GATK HaplotypeCaller, FreeBayes and VarDict. Expected and observed frequencies are shown

The genotypes determined with Nimbus call were also compared with genotypes called by 3 other popular variant callers, namely FreeBayes, VarDict and the GATK HaplotypeCaller. Most of the SNPs called by FreeBayes and the GATK were also called by Nimbus (Fig. 5A). VarDict fails to call 47% of the SNPs compared with the other variant callers. Between FreeBayes and GATK, 5097 variants are called that were absent from the Nimbus genotypes. Of these variants 2751 had a variant frequency between 0.7 and 0.95 and were not called as heterozygous. Nimbus variant calls are thus predominantly concordant to those of the GATK HaplotypeCaller and FreeBayes. The performance of Nimbus to detect somatic variants with low frequencies was compared with other popular callers (Fig. 5B). Nimbus outperforms the other callers and is able to detect variants at lower frequencies.

3.3 The added value of Nimbus

3.3.1 Design optimization by amplicon performance

As Nimbus annotates the aligned reads with the source amplicons, the performance of individual amplicons can be assessed. In individual samples 7.3–11% of the autosomal amplicons were not detected (Fig. 6). Approximately 2.3% of the amplicons did not yield reads in any of the six samples. The design can be optimized by either removing or improving consistently missing amplicons.

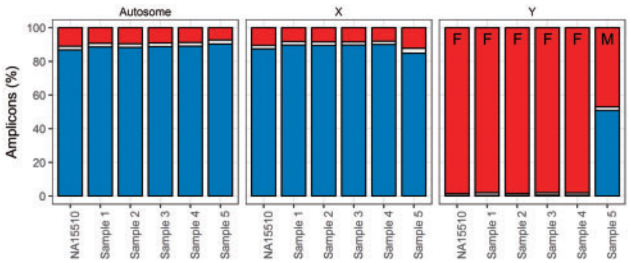


Fig. 6. Amplicon fate per sample. For each sample the percentages of not detected (top/red), error-prone (middle/white) and good amplicons (bottom/blue) per sample in the HaloPlex exome design is shown. Off-target regions were not included in this graph. Male (M) and female (F) samples are indicated for chromosome Y (Color version of this figure is available at *Bioinformatics* online.)

In addition to missing amplicons, other amplicons consistently yield alignments with many mismatches. Nimbus discards these erroneous alignments in filtering step iv (Fig. 1). In each sample, 45–52 thousand amplicons yield over 20% discarded alignments (4 or more mismatch events; Supplementary Table S1). Per sample, 11–20% of these amplicons are sequenced by 30 reads or more. The relatively few amplicons that cause issues in the HaloPlex exome design are thus easily identified using Nimbus allowing them to be corrected in future iterations of the design.

3.3.2 Design optimization by decoy performance

Approximately 739 000 potential off-target sites (decoys) were added to the HaloPlex exome design prior to alignment. Most of these decoy amplicons (~80%) were not observed in any of the six datasets (Supplementary Table S1). Between 74 and 154 decoy amplicons generated in excess of a 1000 reads. Those highly sequenced decoy amplicons point to candidates for design removal or optimization, as these are apparently responsible for a significant portion of the off-target reads (Supplementary Table S1).

3.3.3 Sex determination

A clear difference between the male and the female samples is observed for amplicons on sex chromosome Y. The amplicons on the X chromosome behave similarly compared with the autosomes. In female samples, over 95% of the amplicons on chromosome Y are absent (Fig. 6). In the male sample, 47% of these chromosome Y amplicons are detected. Therefore, the sex of a sample can easily be determined by the detection of amplicons on chromosome Y.

3.3.4 Rescue of variants in the alignment seeds

Nimbus align uses a double seed strategy to match the read-pair to the target amplicons (Fig. 2). When mismatches occur in the first seed pair, a second seed pair is used to rescue the alignment and include the amplicon in the list. To demonstrate the effectiveness of the double seed strategy, the variants were matched with the amplicons with which they overlapped. Across all six samples 533 347 matches were detected. In 30 694 of those matches the variant was present in the outer seed pair of an amplicon. Thus, the double seed approach is able to rescue 6% of variants in an amplicon.

3.3.5 Selection of high confidence SNPs

Tracking the amplicons throughout variant calling adds a new quality parameter to classify individual variant calls, which we can exploit to improve variant calling. We assume that variants observed in multiple overlapping amplicons are more likely to be correct than a variant found with the same number of reads in only a single amplicon. To investigate whether this assumption holds true, the frequency of variants in Sample 1 is considered. In the ideal situation, a heterozygous variant would be represented by an equal number of reads from allele A and B resulting in a frequency of 0.5. If the variants called as heterozygous are better centered around the expected value of 0.5, they are better calls than non-centered values.

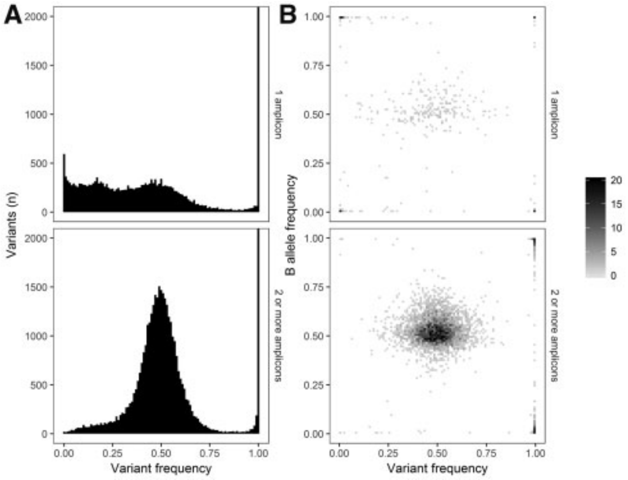


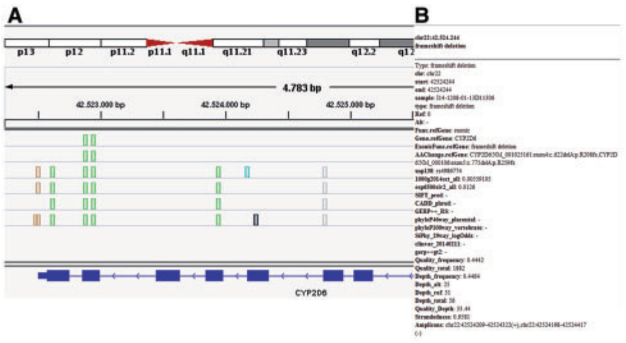
Fig. 7. Variant frequencies (**A**) The variant frequency is set out against the number of amplicons in which the variant is found. Variants found in one amplicon are shown in the top row. Variants found in multiple amplicons are in the bottom row. Variants of which the locations were called with quality below 600 were omitted from the figure. (**B**) Comparison of variant frequencies from SNP array with those of Nimbus split between one (top) or multiple amplicons (bottom)

In Figure 7A, the frequency of the alternate allele in Sample 1 is depicted for variants called across all six samples. The frequencies of variants detected in only a single amplicon range from nearly 0 to 1. A large number of variants are found with a frequency between 0.05 and 0.3. The genotypes suggested by these frequencies would lie in between reference and heterozygous calls which is unlikely for a diploid organism. In total 7679 of the variants (quality > 600) in Sample 1 do not have a clear heterozygous or reference call versus 9158 clear heterozygous calls. In contrast, variants called in two or more amplicons have seven times less unclear calls (3428 unclear versus 29 716 clear heterozygous calls). So, tracking amplicons during alignment and calling enables the selection of high confidence variants.

Another way to show the added value of the tracking amplicons during calling is to compare with genotypes derived by SNP-array. Variants called in multiple amplicons (4031) have an accuracy of 0.960 whereas the variants present in only a single amplicon have an accuracy of only 0.889 (Fig. 7B). Even though most heterozygous variants calls are supported by multiple amplicons, the difference in accuracy is striking. In conclusion, the number of unclear calls is greatly reduced and accuracy is increased when variants are detected in multiple amplicons.

3.3.6 Easy visualization of variants and annotation

In order to visualize variants easily across multiple samples in a single overview, Nimbus creates mutation files that can be loaded in the IGV (Robinson *et al.*, 2011). The mutation files contain all detected variants with rich annotation from ANNOVAR (Wang *et al.*, 2010). The mutation files are loaded in the IGV and variants are coloured based on functional impact (Fig. 8). Mouse-over on the individual variants displays detailed annotations from ANNOVAR.



chr22:42,524,244
frameshift deletion

Type: frameshift deletion
chr: chr22
start: 42524244
end: 42524244
sample: [illegible]
type: frameshift deletion
Ref: G
Alt: -
Func.refGene: exonic
Gene.refGene: CYP2D6
ExonicFunc.refGene: frameshift deletion
AAChange.refGene: CYP2D6:NM_001025161:exon4:c.622delA:p.R208fs,CYP2D6:NM_000106:exon5:c.775delA:p.R259fs
snp138: rs4986774
1000g2014oct_all: 0.00559105
esp6500siv2_all: 0.0126
SIFT_pred: -
CADD_phred: -
GERP++_RS: -
phyloP46way_placental: -
phyloP100way_vertebrate: -
SiPhy_29way_logOdds: -
clinvar_20140211: -
gerp++gt2: -
Quality_frequency: 0.4442
Quality_total: 1882
Depth_frequency: 0.4464
Depth_alt: 25
Depth_ref: 31
Depth_total: 56
Quality_Depth: 33.44
Strandedness: 0.8581
Amplicons: chr22:42524209-42524322(+),chr22:42524188-42524417
0-1

Fig. 8. Variant visualization in IGV. (**A**) Variants visualized in the IGV are colour coded based on their functional impact. Synonymous and intronic variants are shown in grey, non-synonymous variants are green, non-frameshift deletions are blue, downstream variants are brown and frameshift deletions are depicted in black. (**B**) Via an on-click pop-up, more information is obtained for the variant including the amplicons in which this variant was found (Color version of this figure is available at *Bioinformatics* online.)

The mutation files created by Nimbus enables efficient and easy data visualization and analyses in a multiple sample context that is essential to both research and diagnostics.

4 Discussion

Here we report Nimbus, dedicated tools for the analysis of amplicon-based NGS data. Nimbus includes tools for data preprocessing, alignment, variant calling, quality control and visualization. Throughout the analysis, Nimbus keeps track of the amplicons from which reads originate and variants are called. We show that by tracking the amplicons, false-positive variants can be distinguished and separated true positive variants. Furthermore, Nimbus guides the evaluation and re-design process of successive amplicon enrichments for the same target regions by quantifying the individual amplicon performance in terms of (relative) read-depth, number of filtered alignments and number of off-target reads. Amplicon designs can be optimized more effectively by considering these metrics, than on empirical observations. Finally, Nimbus provides an easy way to visualize the analysed samples with relevant annotation.

The qualitative performance of Nimbus is demonstrated using six HaloPlex exome example samples which included two pairs of monozygotic twins. Variants called by Nimbus were highly concordant between the members of the twin samples and to genotypes detected with SNP-arrays. Between the members of the twins ∼40 discordant variants were found which are more than usually reported in twin studies (Chaiyasap *et al.*, 2014). The goal of the filters applied here is to preserve most variants present in the sample and not to over filter. Therefore, we did not filter the variants based on the number of amplicons in which they were called. Comparison with GATK and FreeBayes showed that the variant calls are predominantly concordant with the Nimbus calls. VarDict seems to be unsuitable for calling SNPs in HaloPlex datasets. The observed lower SNP concordance with VarDict is possibly due to the exclusion of variants by VarDict if they are not present in all amplicons. Nimbus outperforms other callers for low-frequency variants. This makes Nimbus broadly applicable in cancer studies where it is often important to detect new emerging variants at low frequencies.

Variants called in multiple amplicons are more accurate than those present in only a single amplicon. In terms of frequency, heterozygous variants called in multiple amplicons are closer to the expected frequency of 0.5 than heterozygous variants present in only a single amplicon. Furthermore, multi-amplicon variants show fewer differences to SNP-arrays. The number of conflicting variants between the monozygotic twin siblings can be significantly reduced through direct comparisons and amplicon filtering. Based on these observations we recommend to use multiple overlapping amplicons to increase variant confidence.

Designs for specific targets are optimized quickly with Nimbus as Nimbus carries over the amplicons throughout the analysis. Thus, amplicons yielding too few reads and/or too many 'bad' alignments are easily identified. By aggregating this information over multiple samples, candidate amplicons for redesign are identified. Through the amplicon performance metrics, Nimbus guides (clinical) researchers in obtaining an optimal design for target regions.

The sex of samples can be easily inferred based on the amplicons on chromosome Y for the used HaloPlex exome design. The sensitivity of sex inferral for custom designs is dependent on both the number of amplicons and the coverage per amplicon.

Methods to detect SNPs and short insertions and deletions are included in Nimbus. Methods to find larger copy-number variants (CNVs) such as duplicated or deleted exons are currently not included. However, the per-amplicon counts can serve as input for dedicated exome CNV detection algorithms such as ExomeDepth (Plagnol *et al.*, 2012) and XHMM (Fromer *et al.*, 2012) as demonstrated in the guides at the Nimbus GitHub repository. With the read depths per amplicon provided by Nimbus, these methods can base their analyses on multiple measurements per exon which can potentially allow for the detection of small CNVs.

The Nimbus suite provides an end-to-end solution for variant calling and design optimization of amplicon-derived NGS datasets. By providing specific input formats for the IGV (Robinson *et al.*, 2011), Nimbus provides easy and accurate visualization of multi-sample variants with added contextual information.

Acknowledgements

We thank the members of the Erasmus Center for Biomics lab for their valuable contribution to discussion and experiments.

Conflict of Interest: none declared.

References

Bentley,D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Chaiyasap,P. *et al.* (2014) Whole genome and exome sequencing of monozygotic twins with trisomy 21, discordant for a congenital heart defect and epilepsy. *PLoS One*, **9**, e100191.

Dahl,F. *et al.* (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. USA*, **104**, 9387–9392.

Fromer,M. *et al.* (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.*, **91**, 597–607.

Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*.

Gnirke,A. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.

Hardenbol,P. *et al.* (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.*, **21**, 673–678.

Hodges,E. *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nat. Genet*, **39**, 1522–1527.

Johansson,H. *et al.* (2011) Targeted resequencing of candidate genes using selector probes. *Nucleic Acids Res.*, **39**, e8.

Koopmans,A.E. *et al*. (2014) Clinical significance of immunohistochemistry for detection of BAP1 mutations in uveal melanoma. *Mod. Pathol*., **27**, 1321–1330.

Lai,Z. *et al*. (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*., **44**, e108–e108.

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Levenshtein,V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Dokl. Phys*., **10**, 707–710.

Li,H. *et al*. (2009a) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,R. *et al*. (2009b) SNP detection for massively parallel whole-genome resequencing. *Genome Res*., **19**, 1124–1132.

McKenna,A. *et al*. (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*., **20**, 1297–1303.

Plagnol,V. *et al*. (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, **28**, 2747–2754.

Robinson,J.T. *et al*. (2011) Integrative genomics viewer. *Nat. Biotechnol*., **29**, 24–26.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.

Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*., **38**, e164.