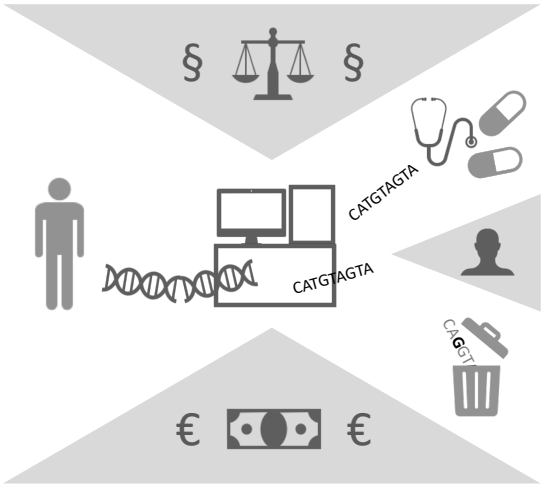
Translating Next-Generation-Sequencing into Precision Medicine



Dissertation for attaining a doctoral degree at the Faculty of Mathematics and Natural Sciences at Kiel University, Christian-Albrechts-Universität zu Kiel, submitted by Dipl.-Ing. Michael Forster

First Referee:	Prof. Dr. rer. nat. Andre Franke
Second Referees:	Prof. Dr. rer. nat. Tal Dagan
	Prof. Dr. rer. nat. Norbert Arnold
Chairman:	Prof. Dr. rer. nat. Stanislav Gorb
Date of the Defense:	27 March 2019
Approved for printing:	27 March 2019

Human genome: UK to become world number 1 in DNA testing

£300 million investment that will transform how diseases are diagnosed and treated announced by the Prime Minister today.

https://www.gov.uk/government/news/human-genome-uk-to-become-world-number-1-in-dna-testing

On January 12, 2016, President Obama tasks Vice President Biden with leading a "Moonshot" to help end cancer as we know it.

https://obamawhitehouse.archives.gov/node/352601

Congress passed the 21st Century Cures Act in December 2016 authorizing $1.8 billion in funding for the Cancer Moonshot over 7 years. An initial $300 million has been appropriated in fiscal year (FY) 2017 to fund Moonshot initiatives.

Cancer Moonshot℠ was originally published by the National Cancer Institute.
https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative

The French government announced that it plans to invest €670 million ($760.8 million) in a genomics and personalized medicine program meant to improve the diagnosis and prevention of disease in the country.

https://www.genomeweb.com/clinical-translational/france-plans-invest-670m-genomics-personalized-medicine

The ballooning costs of healthcare act as a hungry tapeworm on the American economy. Our group does not come to this problem with answers. But we also do not accept it as inevitable. Rather, we share the belief that putting our collective resources behind the country's best talent can, in time, check the rise in health costs while concurrently enhancing patient satisfaction and outcomes.

Warren Buffett
https://www.businesswire.com/news/home/20180130005676/en/Amazon-Berkshire-Hathaway-JPMorgan-Chase-partner-U.S.

Table of Contents

Deutsche Zusammenfassung / English Abstract / Grafische Zusammenfassung

Deutsche Zusammenfassung

Die Next-Generation-Sequenzierung (NGS) wurde zunächst als experimentelles Verfahren für explorative Studien und bei der Suche nach neuen wissenschaftlichen Erkenntnissen eingesetzt. In der Präzisionsmedizin gelten zusätzliche Rahmenbedingungen in Bezug auf konkrete klinische Fragestellungen, Wirtschaftlichkeit, Haftung und Reproduzierbarkeit. Im Rahmen der akkreditierten Diagnostik wird eine Dokumentation der Standardverfahren zur computerbasierten Validierung (*in-silico*-Validierung) von NGS-Ergebnissen in zunehmendem Maße gefordert.

In dieser Dissertation beschreibe ich mögliche *in-silico*-Verfahren anhand meiner Veröffentlichungen zu der `Backmapping`-Methode sowie den Softwareprogrammen `pibase` und `Vy-PER`. Die Verfahren berücksichtigen potentielle Fehlerquellen aus der Biologie, dem Labor und der Bioinformatik.

Meine **Veröffentlichung A** zu `pibase` befasst sich mit der Validierung von individuellen Einzelnukleotidsubstitutionen und einer systematischen Eliminierung von Fehlerquellen.

Meine **Veröffentlichung B** zu `Backmapping` setzt sich mit potentiellen Artefakten auseinander, die auf biologischen Sequenzähnlichkeiten zwischen verschiedenen Bereichen innerhalb des menschlichen Genoms beruhen und daher bioinformatische Mehrdeutigkeiten bei der Ausrichtung der Sequenzen gegen ein Referenzgenom ergeben. Diese potentiellen Artefakte können mit Hilfe von alternativen Genomreferenzen und alternativen Ausrichtungsmethoden validiert werden.

Meine **Veröffentlichung C** zu `Vy-PER` stellt eine verbesserte Methode vor, patientenindividuelle Virusgenompartikelintegrationen in das Patientengenom zu detektieren und Artefakte zu eliminieren.

English Abstract

Next-Generation Sequencing (NGS) was initially used as an experimental method in exploratory research and in the search for novel scientific insights. In precision medicine additional critera apply, for example the testing method must answer specific clinical questions, while being subject to stringent standards of economy, liability and reproducibility. In accredited clinical NGS testing, laboratories must now increasingly document their standard procedures for computational validation (*in silico* validation) of NGS results.

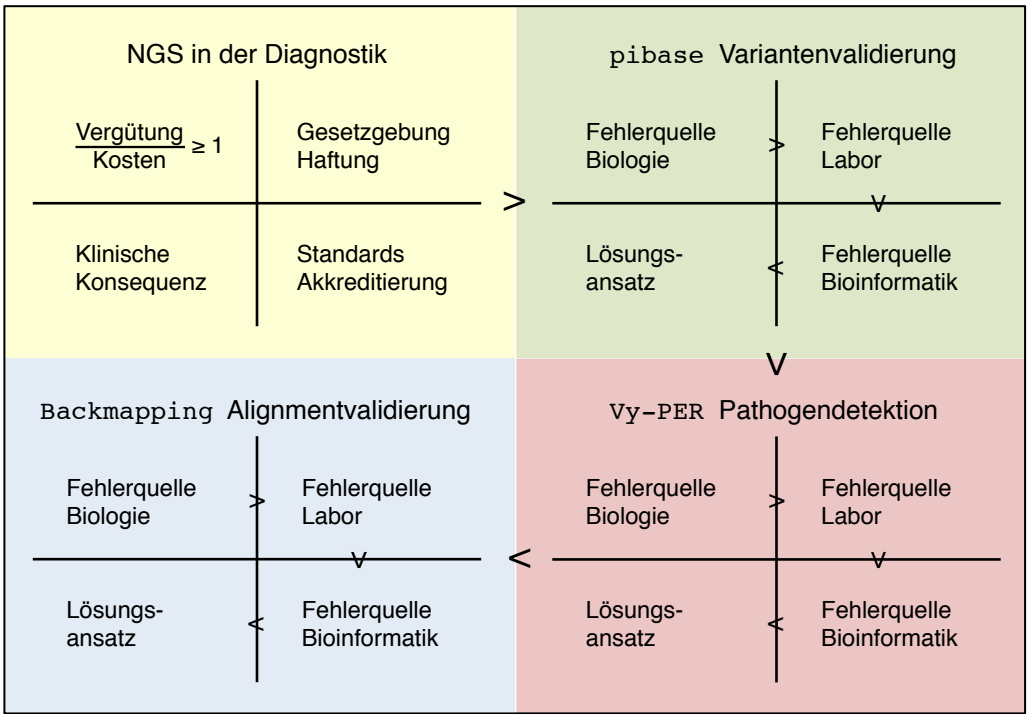
In this dissertation I describe possible *in silico* validation procedures based on my publications on the `Backmapping` method and on the `pibase` and `Vy-PER` software tools. These procedures take into account potential sources of errors from the biology itself, the lab methods, and bioinformatic methods.

My `pibase` **publication A** deals with the validation of individual single nucleotide substitutions and the systematic elimination of error sources.

My `Backmapping` **publication B** looks at potential artifacts that are based on biological sequence similarity between different regions of the human genome. These similarities may or may not lead to bioinformatic ambiguities when sequences are aligned to a reference genome. In consequence, potential artifacts may crop up which may only be exposed with the help of alternative genomic references and alternative alignment methods.

My `Vy-PER` **publication C** presents an improved method to detect patient-individual virus genome particle integrations into the patient's genome and to eliminate artifacts.

Grafische Zusammenfassung



1. Introduction

1.1 Aim of the study

The aim of this study was to formulate reliable methods to discriminate between erroneous DNA variants from Next-Generation-Sequencing (NGS) and error-free DNA variants, so that NGS can be used not only for research but also for precision medicine.

1.2 General Introduction

Over the past decade next-generation sequencing (NGS) has helped to generate an incredible amount of knowledge. Today it is regarded as one of the most important standard methods in molecular biological research and is of fundamental importance for precision medicine. The IKMB has been sequencing human genomes of patients for ten years. Back then, the consumables required for sequencing a human genome had cost around 200,000 euros. The software for data evaluation and the methods of data interpretation had been in their infancy. Today, in 2018, the cost of consumables for genome sequencing has fallen to around 900 euros and data analysis is mature. In addition, the NGS technology and its clear advantages have been recognized by those responsible in the healthcare system. Since the first of July 2016, there exist billing codes in the German public health system (*Abrechnungsziffern im Einheitlichen Bewertungsmaßstab, EBM*), according to which the cost of NGS testing is regulated.

Reimbursement for NGS testing

The Genetic Testing Act (*Gendiagnostikgesetz, GenDG*) was passed in 2009 and came into force on February 1, 2010. At the same time, the Robert-Koch Institut founded the Gene Testing Commission (*Gendiagnostik-Kommission*). Its task is to draw up and regularly revise guidelines for the implementation of genetic testing (https://www.rki.de/DE/Content/Kommissionen/GendiagnostikKommission/Richtlinien/Richtlinien_node.html). These guidelines mainly describe patient counselling and quality assurance for genetic tests. With regard to the analysis of the NGS data, there is however only the briefest hint in the Quality Assurance Guideline (*Qualitätssicherungs-Richtlinie*), with the following 20 words: "generally recognised state of the art in science and technology", which "therefore as a rule defines itself through generally accessible publications in specialised journals and textbooks" *("allgemein anerkannten Stand der Wissenschaft und Technik", der "sich daher i. d. R. durch allgemein zugängliche Publikationen in Fachzeitschriften und Lehrwerken" "definiert"*).

Genetic Testing Act (GenDG) and Guidelines by the Robert-Koch Institut

Currently, in 2019, the Institute of Clinical Molecular Biology (IKMB) is introducing NGS testing for cardiomyopathies, pulmonary hypertension and immunodeficiencies at the University Hospital Schleswig-Holstein (UKSH) in cooperation with the Central Laboratory, the Center of Excellence for Inflammation Medicine and the Clinic for Internal Medicine III.

Against this background, I would like to address my short dissertation not only to my referees but also to those scientists, technicians and clinicians who currently have to familiarize themselves with the "generally recognized state of the art of science and technology" in NGS. In the following, I will summarize the main findings from my publications and the therein presented `Backmapping` method as well as the software `pibase` and `Vy-PER` and I will explain them with regard to NGS data analysis for precision medicine.

Generally recognised state of the art in science and technology

2. Methodological considerations

Genome-specific considerations for NGS analysis

The human genome normally consists of 22 autosomal chromosome pairs and two sex chromosomes. Complete sequencing, i.e. complete coverage of the human genome, is not possible with NGS technology because the short sequences from an NGS instrument cannot be unambiguously mapped to all locations of a human reference genome. In addition, the reference genomes themselves are incomplete. We and others find that some of the human DNA sequences generated by an NGS instrument from a patient sample cannot be mapped using the current version of the human reference genome (hg38), but only using the BLAST database and website of the National Center for Biological Information (NCBI). Furthermore, there are difficult-to-sequence and difficult-to-map regions in the reference genome. These regions include repetitive Short Tandem Repeat (STR) and microsatellite sequences as well as complex sequences that may occur identically or similarly several times in the genome. The latter are referred to as sequence homologies and in certain cases as pseudogenes. In clinical testing, pseudogenes have serious consequences, therefore I will dedicate a section to them in the discussion.

Gaps in the human reference genome

STRs and microsatellites

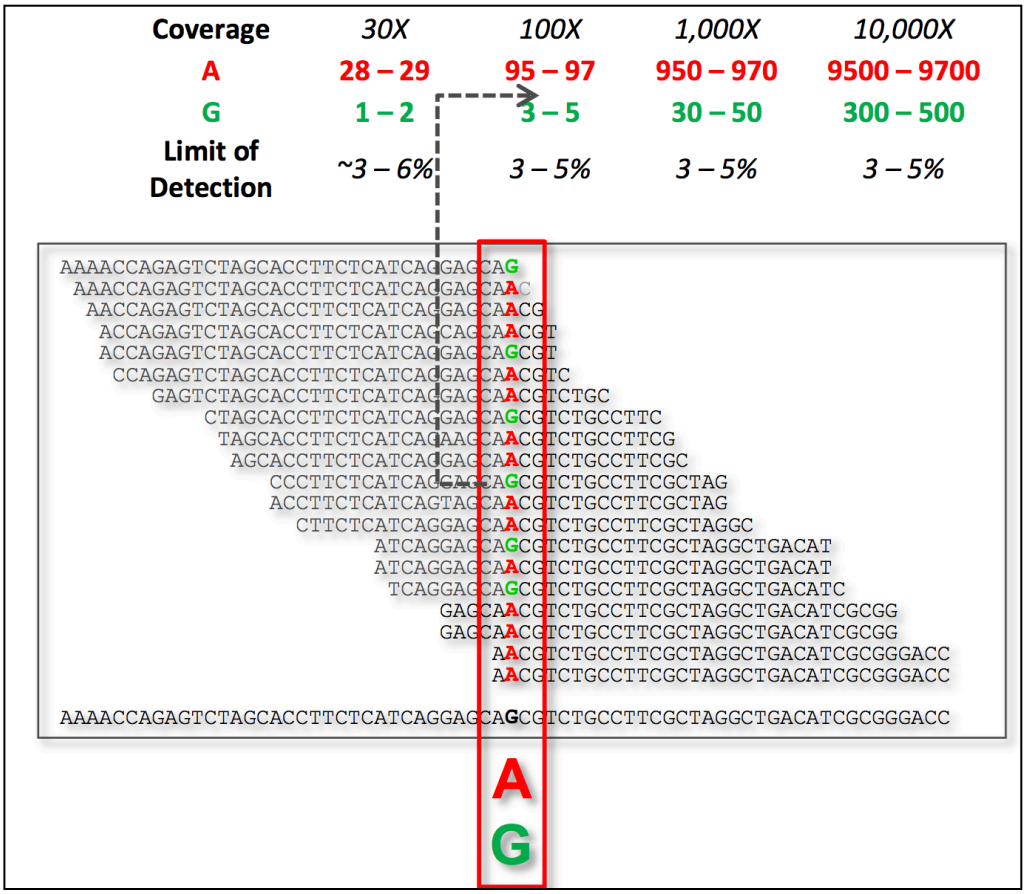
Sequence homologies, pseudogenes

Underlying NGS technology

Today, sequencing-by-synthesis (SBS) technology dominates. It was developed at Cambridge University (Cambridge, UK), brought to market with the help of investors from 1998 to 2006 by Solexa Ltd (London, UK) and taken over in 2007 by Illumina Inc. (San Diego, USA) (https://www.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html).

This NGS method is based on sequencing millions to trillions of short DNA fragments on a sequencing instrument and bioinformatically

analyzing these sequences. In clinical testing, these sequences are usually aligned to human reference genome sequences. Put simply, if many sequences have the same nucleotide mismatch to the reference, then there is a plausible consensus that this mismatch is real (**Fig. 1**).



Consensus genotype AG from NGS sequences

***Fig. 1: Sequence stack over reference genome sequence**. Millions of short DNA fragments are sequenced in NGS instruments. In the case that a reference sequence for the organism is known, the seqeunces are then computationally aligned against the corresponding genomic reference sequences. The above example shows the genotype AG that was derived from mismatches in the shown sequences. True genotypes can usually be determined more reliably with increasing sequencing depth (coverage). (Figure: Illumina; permission granted)*

As previously mentioned, the unique alignment of a short sequence against the human reference genome sequences is not always possible due to structural characteristics. Only about 95% of the genome reference can be uniquely mapped using sequences of 75 base pairs (bp) (1000 Genomes Project Consortium *et al.*, 2010). Therefore, it is general practice to sequence both ends of a short DNA fragment (**Fig. 2**). This sequence pair can be uniquely mapped to the reference genome in more cases than a single sequence. Using paired end sequencing, paired sequences of length 2 × 75 bp can be uniquely mapped to approx. 97% of the reference genome, and 2 x 150 bp can be uniquely mapped to approx. 98% (1000 Genomes Project Consortium *et al.*, 2010).

paired end sequencing

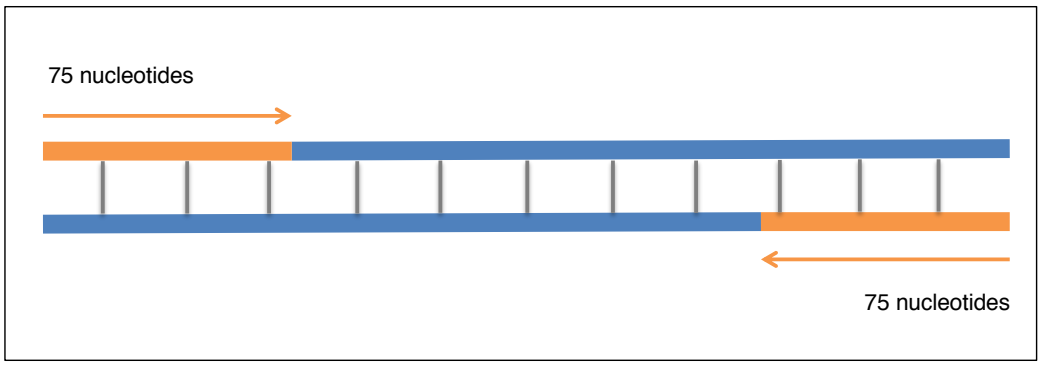


***Fig. 2: Paired end sequencing**. Both ends of a DNA fragment are sequenced (here: 75 nucleotides each) in order to improve the unique bioinformatic alignment of the sequences to the genomic reference sequence.*

In the first step of library construction, large DNA fragments are mechanically or enzymatically fragmented or alternatively transcribed as PCR amplicons from certain loci of the genome. For Illumina Paired end sequencing the required average DNA fragment length is usually about 200-350 bp. Adapters are ligated to the DNA fragments (**Fig. 3**), which are complementary to primer molecules in the Illumina sequencing chip. The single strand molecules are thus bound to the primer molecules (**Fig. 4**).

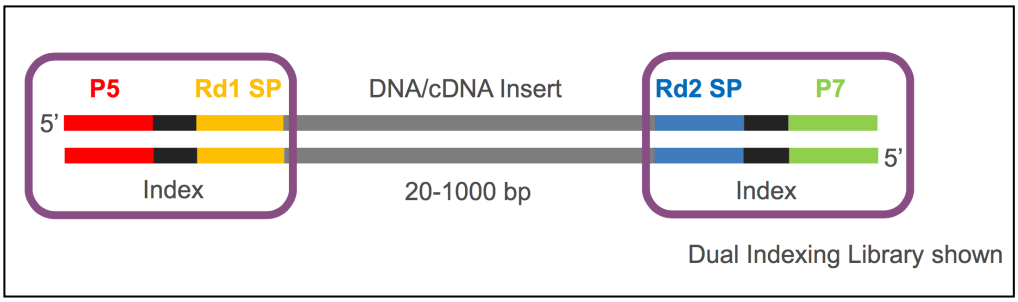


Fig. 3: Illumina sequencing library. *Sequencing adapters with sequencing primer binding sites (Rd1 SP and Rd2 SP, respectively) are attached to each end of the double-stranded DNA fragment. The adapters usually also contain a sample-specific DNA sequence (index), which allows many samples to be pipetted together and sequenced as a pool. The index allows (nearly) each sequence to be bioinformatically reassigned to the correct sample (demultiplexing). At the outer ends of the adapters there are complementary binding sites (P5 and P7) to primers in the Illumina sequencing chip ("flowcell"), to which the binding takes place in the chip. (Figure: Illumina; permission granted)*

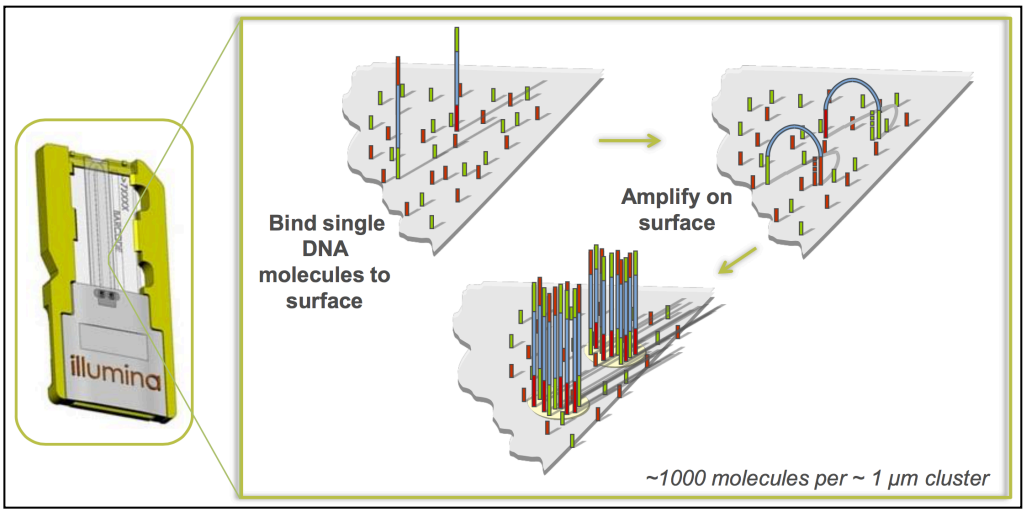


Fig. 4: Illumina sequencing chip showing bridge amplification and clustering. *The sequencing libraries are denatured and bound at their adapters (P5 or P7) to the glass plate in the chip. Each library molecule is amplified with 10 bridge PCR cycles, resulting in a cluster of approximately 1000 library molecules. (Figure: Illumina; permission granted)*

In the sequencing chip, bridge amplification is performed (**Fig. 4**) in order to amplify the optical sequencing signal of each single strand (**Fig. 5**) approximately one thousand times.

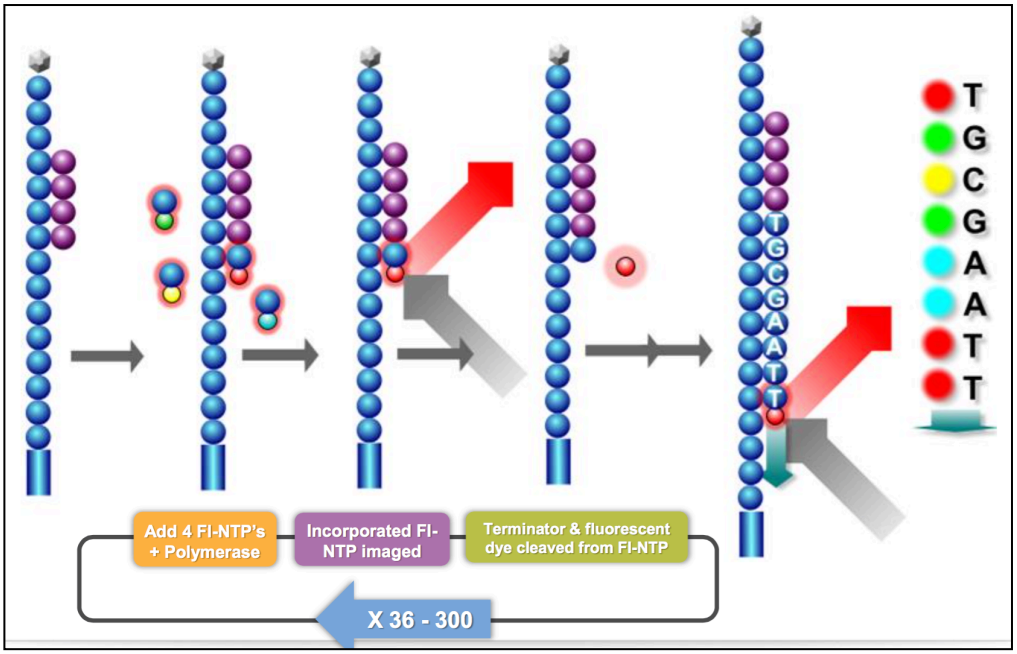


Fig. 5: Illumina's SBS sequencing principle: *Starting with the sequencing primer binding site (violet spheres), individual deoxyribonucleoside triphosphates (dNTPs) with attached fluorophore and reversible terminator are bound to the single strand cycle by cycle. The Fluorophores are laser-excited to emit a light signal, and photographed digitally. Then, the fluorophore and reversible terminator are cleaved off, after which the next cycle can begin. The translation of the light signal of each cluster (approx. 1000 single strands) into a unique nucleotide is performed using "base-calling" software. (Figure: Illumina; permission granted)*

Signal noise due to phasing and prephasing: from adjacent bases

Errors may occur when incorporating individual deoxyribonucleoside triphosphates (dNTPs) during the sequencing process (**Fig. 5**). If a dNTP is not incorporated and the subsequent light signals thus shift, this is known as "phasing" (**Fig. 6**), conversely (if several dNTPs are incorporated instead of one dNTP) as "prephasing" (Kircher, Stenzel and Kelso, 2009; Ledergerber and Dessimoz, 2011).

In the data analysis, phasing or prephasing has the effect that a putative variant is detected which, on closer examination, consists of one of the two neighboring nucleotides in the reference sequence and usually has a much lower allele frequency than a true consensus genotype.

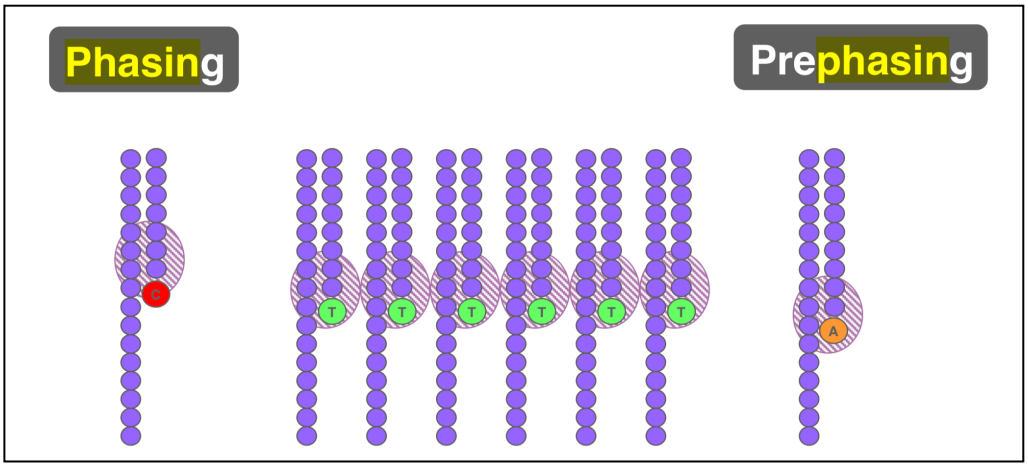


Fig. 6: Phasing and prephasing - dNTP incorporation errors: *Left: Phasing is the failure to incorporate a dNTP, probably because the blocker of the previously incorporated dNTP was not removed correctly. Middle: Single strands without incorporation errors. Right: Prephasing is the incorporation of two dNTPs, probably due to a blocker defect. (Sources: Image courtesy of the Broad Institute, https://www.broadinstitute.org/files/shared/illuminavids/dataSlides.pdf, https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina)*

Standard sequence data format and quality or Phred score

Quality or PHRED

The standard format for NGS sequences is the compressed FASTQ format (FASTQ.gzip), an extended FASTA format: In addition to the base information of the FASTA format, it also contains a quality value for each base, which is based on that of the program phred (Ewing *et al.*, 1998). The PHRED score, also called Base Quality, is defined as follows:

Error probability := 10 $^{-(PHRED/10)}$

e.g. error probability = 10 $^{-(20/10)}$ = 10 $^{-2}$ = 1% for PHRED = 20

The quality of a sequence alignment against a reference sequence is described by a value that is inspired by the PHRED score and referred to as Mapping Quality. Both the Base Quality and the Mapping Quality serve only as indicative estimates and not as absolutely reliable quality values.

STR noise - technical or biological blurring ?

Be aware of STR and homopolymer regions.

As previously mentioned, STRs and homopolymeric sequences (**Fig. 7**) are among the regions of the genome that are difficult to accurately sequence and map to. This has both a biological reason and a technical reason. STRs mutate 100,000 to 1,000,000 times faster than single nucleotide variants (Forster *et al.*, 2015). In some individuals, some STR regions may become genomically unstable in some cells, i.e. they become longer or shorter in the course of life and can initiate diseases such as ataxias and Huntington's disease; coding region trinucleotide STRs become instable when 29-39 repeats are reached (McMurray, 2010). On the other hand, the technical artifact of "PCR stutter" is known from forensic STR analysis. Most bioinformatic methods cannot distinguish PCR stutter from biological length changes and therefore often report false-positive variants.

A validated STR analysis method should therefore be used in clinical testing, e.g. MSI sensor for somatic tumor/normal analysis (Niu *et al.*, 2014). If no validated method for STR analysis is used, the NGS results in STR regions should be ignored. This includes homopolymeric runs, i.e. sequence regions in which the same nucleotide occurs several times in succession. Of clinical importance, certain homopolymeric regions are used in human genetics and pathology as microsatellite markers for Lynch Syndrome and for microsatellite instable tumors (Umar *et al.*, 2004). For mitochondrial DNA it was shown that homopolymers with a length of 8 bp or longer can be highly mutable (Forster *et al.*, 2010). In clinical testing software such as GenSearchNGS (PhenoSystems S.A., Blonay, Switzerland), the filtering of variants in homopolymeric regions is a standard option.

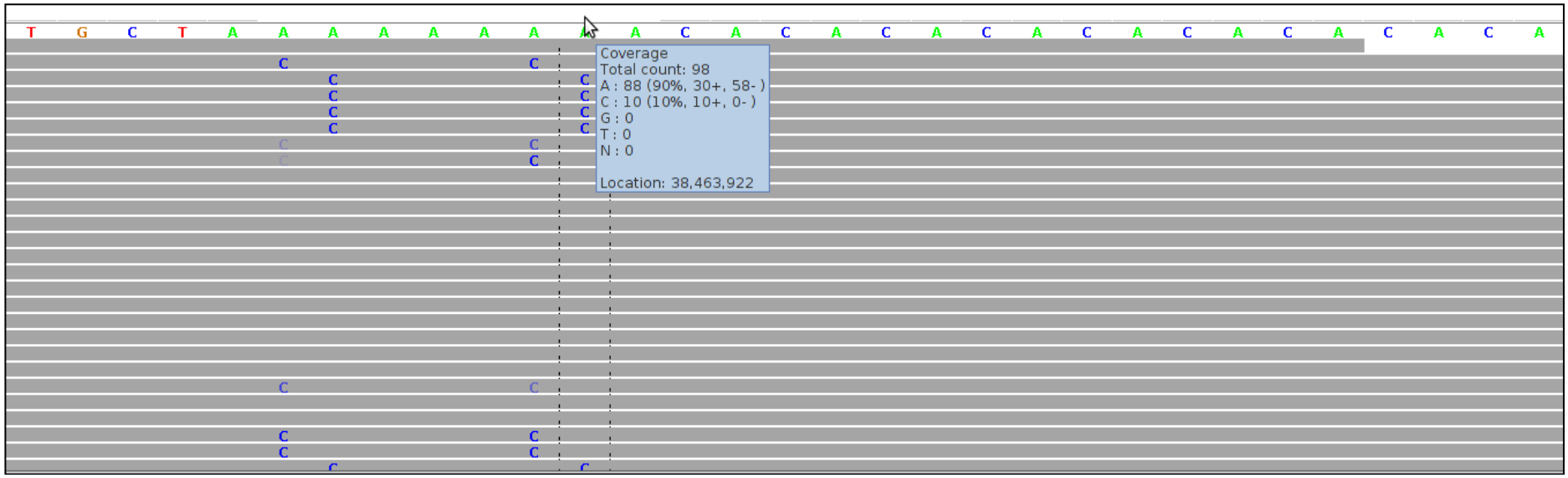


Fig. 7: Short-tandem repeats (STRs) in the `Integrative Alignment Viewer (IGV)`. *This figure shows the alignments of NGS sequences (grey bars) under the reference sequence within an STR region ("CA" repeats) adjoining an "A" homopolymer region. The mismatches ("C") to the reference are shown. It cannot be determined whether these variants are true biological variants or only technical artifacts (e.g. PCR stutter). The free `Integrative Alignment Viewer` was developed at the Broad Institute* (Robinson *et al.*, 2011; Thorvaldsdóttir, Robinson and Mesirov, 2013).

3. Method for the technical validation of single nucleotide variants - *pibase*

Publication A (see page 60):

Forster M, Forster P, Elsharawy A, Hemmrich G, Kreck B, Wittig M, Thomsen I, Stade B, Barann M, Ellinghaus D, Petersen BS, May S, Melum E, Schilhabel MB, Keller A, Schreiber S, Rosenstiel P, Franke A.
From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software.
Nucleic Acids Res. 2013 Jan 7;41(1):e16

Most genetic variant detection programs were not developed for clinical testing of individual patients, but for studying population genetics in diploid and healthy individuals. Depending on the software and software settings, a number of real alignment errors and real sequencing errors as well as biologically real genetic variants may be filtered out. An approach to determine real genetic variants more confidently is therefore to call variants using two or more software tools with default settings, and then to perform the intersection of all mutation lists. On the other hand, sensitivity can be improved by forming the union of all mutation lists (Wang *et al.*, 2013).

The functionality of the `pibase` software is based on a similar idea. The software internally uses ten different genotyping methods, which differ in increasingly stringent quality filters. `pibase` takes up the previously investigated idea of mathematical convergence of genotyping stability with sequencing depth (Melum *et al.*, 2010). Thus `pibase` not only provides a "consensus genotype" but also a statement about the variability or stability of this genotype.

Consensus of ten separate, increasingly stringent methods for genotyping

`pibase` uses the following checks to determine whether a genotype is stable:

- Does the genotype change between sequence variants with low base quality and sequence variants with high base quality?
- Does the genotype change between deduplicated sequences (unique start points after (Melum *et al.*, 2010)) and non-duplicated sequences?
- Does the genotype change between short sequences and long sequences?
- Does the genotype change between sequences with one mismatch and sequences with several or many mismatches?
- Does the genotype change between sequences with poor mapping quality and sequence variants with high mapping quality?

Furthermore, `pibase` documents whether the genotype is sequenced on the forward and backward strands (**Fig. 8**, independent supportive observations) or only on one strand. In addition, `pibase` checks and reports whether the genotype lies in a hypervariable (**Fig. 9**) or homologous region (**Fig. 10**) of the genome reference. Finally, `pibase` tests the flanking bases of the reference sequence for homopolymers and dinucleotide STRs.

Double strandedness, hypervariability, homology

As a result, `pibase` provides a comprehensive summary table of the above results (**Fig. 11**) as a valuation aid for real variants or technical artifacts. This table can be filtered automatically to separate stable genotypes from possibly false positive signals. The table is intended for automatic technical validation within a clinical bioinformatics solution as well as for manual technical validation in order to save extensive manual work in the viewer. Input file formats for `pibase` may be a text file of genomic positions or a standard VCF file (variant call format).

Detailed technical report, see **Fig. 11**

To validate potential somatic variants in tumor DNA versus normal DNA, pibase does not compare genotypes but the underlying original alignment files, using an eight-field Fisher's exact test and constraints. This test was also used in the above-mentioned **publication A** for the labour-saving comparison of germline variants of identical twins. The Fisher's exact test is now a generally used standard method for comparing tumor/normal NGS data pairs (Xu, 2018).

Fisher's exact test (8 fields) for comparing alignments

`pibase` is suitable for the detailed examination of given (!) positions in a genome and thus for the decision whether detected variants are real variants or false positive signals. `pibase` is not suitable for the actual detection of variants. To scan for putative variants, it is best to use several variant detection programs with sensitive settings and pool their results. `pibase` can then be used for the technical validation of putative single nucleotide variants. A detailed description of `pibase` can be found on the homepage
http://www.ikmb.uni-kiel.de/pibase .

`pibase` homepage

Key results:

On the basis of publicly available 1000 Genomes alignment data (BAM files) with an average 30-fold sequencing depth, `pibase` achieved a genotyping specificity of 99.97-100.00%.

Specificity 99.97-100.00 at 30X

By analyzing two monozygotic but phenotypically discordant pairs of twins, co-author Britt-Sabina Petersen and I were able to show that `pibase` significantly reduces manual work compared to conventional analyses. Using the p-value of the Fisher's exact test, we were able to reduce the *de novo mutation* candidate list between twins from hundreds of mutation candidates to 5 and 15, respectively, by filtering with $p < 0.01$. Interestingly, only zero to two *de novo mutations* occur in the exome between parent and child (1000 Genomes Project Consortium *et al.*, 2010; Girard *et al.*, 2011). Genetic differences between monozygotic twins are rarely found within the exome regions (Petersen *et al.*, 2014).

10-fold less manual work when comparing identical twins

These convincing results indicate that the `pibase` methods would be useful if integrated into clinical diagnostic solutions.

Fig. 8: Display of forward and reverse sequences in the `Integrative Alignment Viewer (IGV)`. *This figure shows the alignments of Illumina HiSeq 2000 sequences (pointed horizontal gray bars) under the reference sequence (topmost sequence line). The pointed ends of the grey bars indicate whether the sequence was sequenced from the forward or reverse strand of the DNA fragment. Mismatches to the reference sequence are indicated by a FASTA base within the grey bar.*

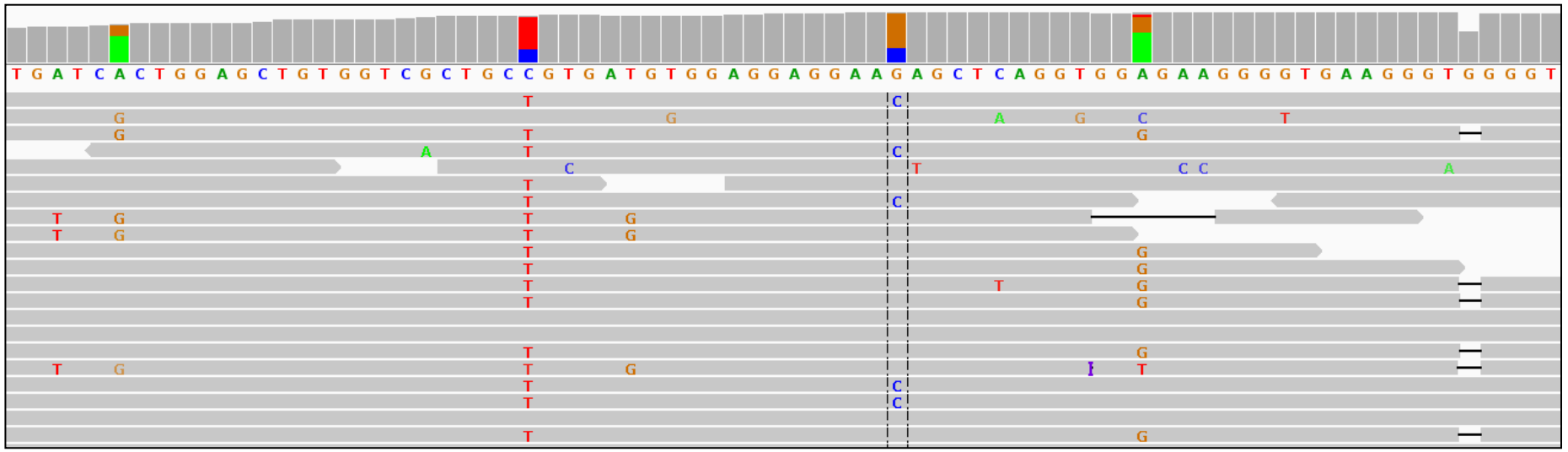


Fig. 9: Display of sequence alignments in a hypervariable region in the `Integrative Alignment Viewer (IGV)`. *This figure shows the alignments of Illumina MiSeq sequences (horizontal gray bars) under the human reference sequence (topmost sequence line). Mismatches to the reference are indicated by a FASTA base (single nucleotide substitution) or by a horizontal line (deletion) within the sequence. The hypervariable region can be recognized by the fact that there are multiple mismatches to the reference sequence in almost every gray sequence bar. On average only 1 single nucleotide mismatch to the reference sequence per 1,000 bp and only one small deletion or insertion per 8,000-9,000 bp is expected* (1000 Genomes Project Consortium *et al.*, 2010).

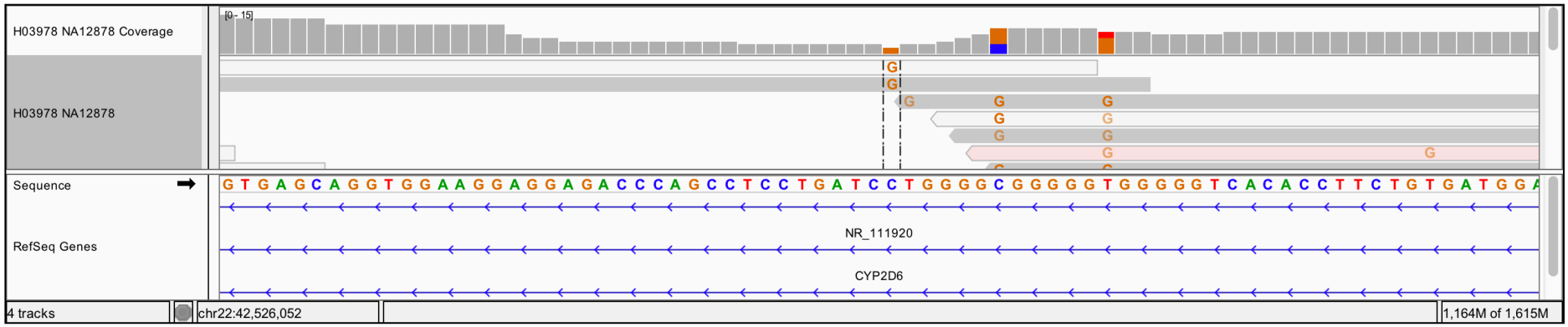


Fig. 10: Display of non-specific sequence alignments in the `Integrative Alignment Viewer (IGV)`. *This figure shows the alignments of Illumina HiSeq sequences (white, grey, and pink bars) under the human reference sequence (lowest sequence line). The white bars indicate `mapping quality 0`, which means that there are homologies to this locus in the reference sequence. The sequence alignments are unspecific and the potential variants found in this locus are often artifacts. To validate the potential variants, a nested PCR (long PCR amplicon, followed by a short PCR amplicon) must be performed and the short product must be sequenced e.g. using the Sanger method.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
27	#							Consensus Genotype														Reads												
28	#							Best				Filt4		Filt3		Filt2		Filt1		Filt0		Filt4		Filt3		Filt2		Filt1		Filt0				
29	# C	Pos[1]	Ref	RefSeqContext	GC[%]	Class	Ign	BestGen	BestQual	A+-	B+-	SP	CV	SP	CV	SP	CV	SP	CV	SP	CV	SP	CV	SP	CV	SP	CV	SP	CV	SP	CV	Homologous	Hypervariable
30	22	17662444	C	TTTTCT[C]ACTCTC	38	5	2	CC	?8	+-	+-			CC	CC	CC	CC	CC	CC	CC	CC	0:0:0:0	0:0:0:0	0:1:0:0	0:1:0:0	0:1:0:0	0:1:0:0	0:4:0:0	0:4:0:0	0:7:0:0	0:7:0:0		
31	22	17662793	A	GAAATC[A]TAGGAC	38	1	1	AA	?7	-	-							AA	AA	AA	AA	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	1:0:0:0	1:0:0:0	3:0:0:0	3:0:0:0		
32	22	17669306	T	TAGTCA[T]GCAAGG	46	1	1	TT	?7	+-	+-	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	0:0:0:3	0:0:0:3	0:0:0:3	0:0:0:3	0:0:0:4	0:0:0:4	0:0:0:6	0:0:0:6	0:0:0:7	0:0:0:7		
33	22	19958811	C	CCTGAC[C]GCGGGC	84	1	1	CC	?7	+	+									CC	CC	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:2:0:0	0:2:0:0		
34	22	19958829	G	GGGCCC[G]GGGGGA	92	9	2	GG	?8	-	-									GG	GG	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:2:0	0:0:2:0		
35	22	19968971	G	GGGCCT[G]GGCCAA	76	1	1	GG	?7	-	-									GG	GG	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:1:0	0:0:1:0		
36	22	19969075	A	ACCCCC[A]CTGCTG	69	8	3	AA	?8	+-	+-									AA	AA	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	3:0:0:0	3:0:0:0		
37	22	19969106	A	GTGGCC[A]CTGGGC	76	1	0	AA	?7	-	-									AA	AA	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	0:0:0:0	1:0:0:0	1:0:0:0		

Fig. 11: Example of a `pibase` results table (excerpt): *This figure shows a `pibase` results file that was generated in tab-separated text format, then imported and formatted in Excel (source: http://www.ikmb.uni-kiel.de/pibase/pibase_consensus_na12878_solid.html). Columns A-B: chromosome and coordinate. C-F: Reference allele and summary information on the reference sequence. G: Number of ignored sequences. H-K: Consensus genotype, quality problem ("?" means quality problem, severity 1-8), detection of the A- or B-allele in +, - or both sequencing directions. L-U: `pibase` uses 5 increasingly stringent filters on the sequences as well as the deduplicated sequences and calculates up to 10 genotypes. V-AE: Number of sequences per base (A:C:G:T) for each of the 10 filters. AF-AG: Marker whether the genomic position was detected as homologous ("H") or hypervariable ("V"). Further information can be generated in `pibase` tables, e.g. sequencing depth.*

4. Method for technical validation of alignments of sequences containing single nucleotide variants - *Backmapping*

Publication B (see page 61):

Elsharawy A*, **Forster M***, Schracke N, Keller A, Thomsen I, Petersen BS, Stade B, Stähler P, Schreiber S, Rosenstiel P, Franke A.
Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing.
***BMC genomics*.** 2012 Aug 22;13:417.
* joint first author

In the early days of the NGS, there were no fast software tools for aligning sequences to a reference sequence. Therefore, researchers studied whether aligning short NGS sequences to a small reference sequence yielded sufficiently accurate results, in analogy to aligning long Sanger sequences to short references. This approach would have accelerated the long computing times. However, in some cases this approach generates artifacts that are reported as variants by the bioinformatics software.

small reference sequence

In **publication B** we presented the `Backmapping` method to detect and undo these artifacts. In the first step we used a very small reference sequence (i.e. the exome reference sequence) instead of the complete human genome reference sequence to identify potential variants. The resulting potential artifacts (**Fig. 12A**) were identified in a postprocessing operation as follows: First, all NGS sequences with the respective mismatch to the reference sequence were extracted from the alignment file. Then these sequences were aligned to the whole human genome

Test "mutated" sequences against whole genome

and finally the genotypes were tested with `pibase` at the genomic coordinates of the previously identified variants. Artifacts were classified as those variants that could no longer be detected after alignment against the complete human genome (**Fig. 12B**). Our automated `backmapping` had thus identified and removed potential artifacts.

automated `backmapping`

In today's interactive sequence viewers such as IGV, Seqpilot (JSI medical systems GmbH) and GenSearchNGS (Phenosystems S.A.), the user can interactively select and copy a sequence (**Fig. 13**). For manual validation of a variant, this sequence can be aligned to the human genome via the `BLAT` website (https://genome.ucsc.edu/cgi-bin/hgBlat) (**Fig. 14**). The sequence can be aligned to known human and non-human sequences using nucleotide `BLAST` (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) (**Fig. 15**)**.** In the special cases of sequences with insertions or deletions or similarities to multiple regions in the reference genome, this validation is good practice (**Fig. 16**). For validating single nucleotide substitutions in sequences without insertions or deletions, `pibase` is usually sufficient.

today: manual `backmapping` with `IGV`, `BLAT` and `BLAST`

Publication B also presents the results of a subproject within the EU project READNA as well as the experiences with faulty and slow bioinformatics tools. A short time later, the variant detection programs `SAMtools` (Li *et al.*, 2009) and `GATK` (McKenna *et al.*, 2010) were revised to enable more time-saving exome and panel analysis. However, the problem of false-positive variant detection is still relevant today.

faster software for exomes

Today, every single variant has to be validated before it can be included in a clinical report. Validation of single nucleotide substitutions can nowadays be assisted by `pibase`. Manual checks remain necessary for insertions and deletions.

Clinical testing: Validate each variant

In summary, **publication B** contains three significant messages that have a practical value for the clinical application of NGS to this day:

(1) We recommended the `backmapping method` for validating potential false positive variants. In interactive form, I recommend `Backmapping` with today's alignment viewer programs: extract an individual sequence containing the detected variant and check its alignment with `BLAST` and `BLAT`, as well as checking for possible contamination with `BLAST`.

`Backmapping` as QC

(2) In READNA, we demonstrated for the first time that targeted enrichment of genomic target regions using hybridization probes works not only for DNA fragments - as had been standard protocol - but also for NGS libraries with sequencing adapters as well as for pooled ready-to-use libraries with sequencing adapters and sample-specific molecular barcodes/indexes (**Fig. 3**). The latter two methods are standard protocol today.

Targeted enrichment of NGS libraries

(3) In READNA, we recommended technical replicates. Today, many laboratory methods are more mature than in the past. For DNA from fresh EDTA blood or fresh frozen tissue, technical replicates are usually no longer necessary. In contrast, I recommend technical replicates for potentially modified DNA or RNA - e.g. DNA from FFPE tissue (Do and Dobrovic, 2012) - even today.

Technical replicates

Fig. 12: Display of sequence alignments for two versions of the reference sequence in the `Integrative Alignment Viewer (IGV)`. *(A) The first reference sequence consisted of only the protein-coding regions of the human genome reference. The figure shows a heterozygous C mismatch to this reference sequence. (B) The second reference consisted of the whole human genome reference sequence. The alignments show no trace of the C variant. Thus, `Backmapping` eliminates sequences that have been force-mapped into a small reference sequence "into the wrong locus".*

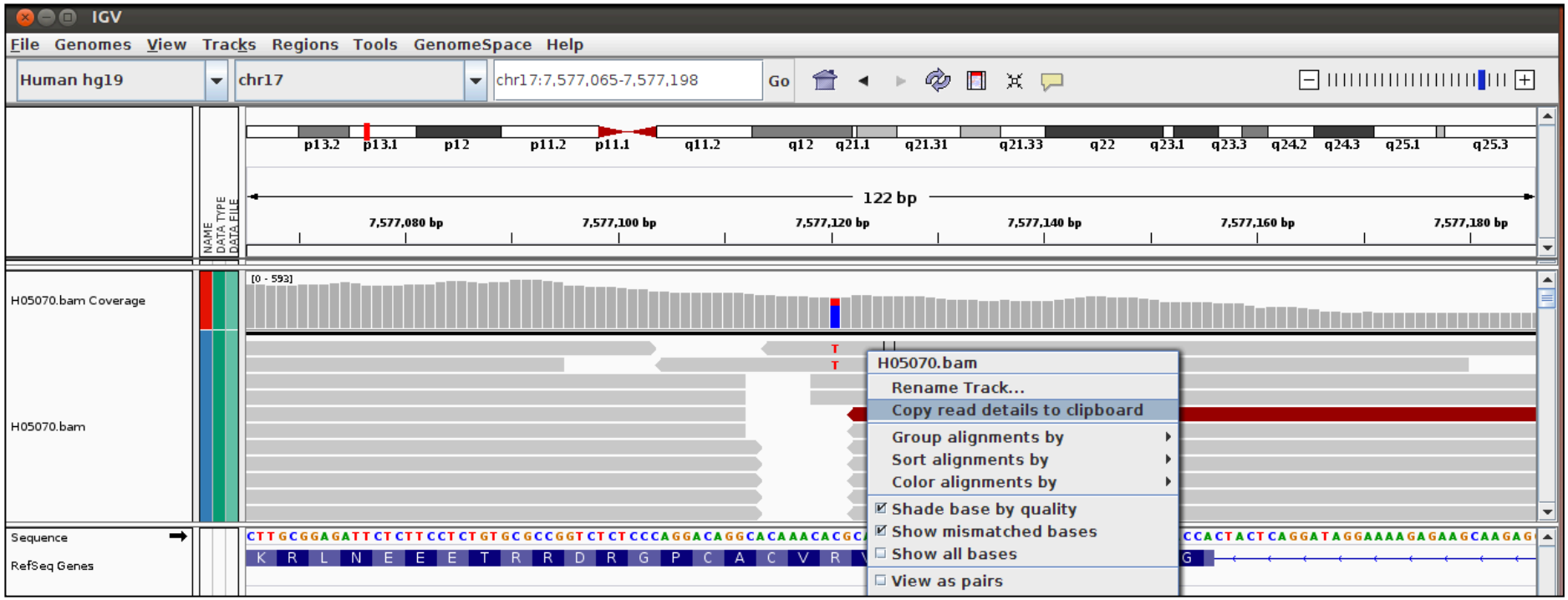
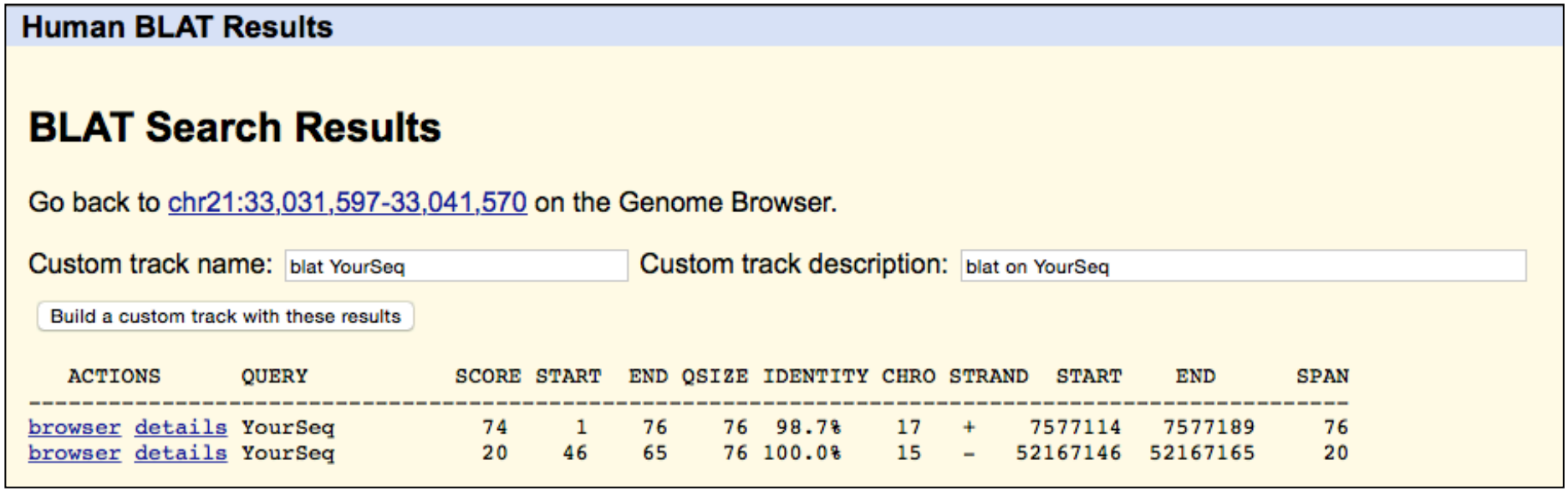


Fig. 13: Selecting and copying a sequence in the `Integrative Alignment Viewer (IGV)`. *In today's interactive sequence viewers, the user can interactively select and copy a sequence, here in `IGV` by right-clicking on the sequence of interest (grey bar). The sequence with the mutation (here: "T") can be selected and UCSC `BLAT` or NCBI `BLAST` can be used to check for alignment artifacts, reference artifacts or contamination:*
CAAACATGCACCTCAAAAGCTGTTCCGTCCCAGTAGATTACCACTACTCAGGATAGGAAAAGAGAAGCAAGAGGCAG.



Human BLAT Results

BLAT Search Results

Go back to chr21:33,031,597-33,041,570 on the Genome Browser.

Custom track name: blat YourSeq Custom track description: blat on YourSeq

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	74	1	76	76	98.7%	17	+	7577114	7577189	76
browser details	YourSeq	20	46	65	76	100.0%	15	-	52167146	52167165	20

Fig. 14: UCSC `BLAT` sequence search validates `BWA` alignment. *This figure shows two results for the query sequence CAAACATGCACCTCAAAGCTCTGTTCCGTCCCAGTAGATTACCACTACTCAGGATAGGAAGGAAGAGAAGCAAGGCAG (copied from `IGV`). The upper result applies to the entire sequence (bases 1-76) with a hit in the human genome reference hg19 on chromosome 17, which also spans 76 bases (SPAN). The second result shows a hit for a subsequence (bases 46-65). Thus, the upper result is clearly the correct result. It is good practice to test the query sequence against the newer human genome reference hg38. (Note: Gene panels from Illumina and other manufacturers are defined in hg19 coordinates.)*

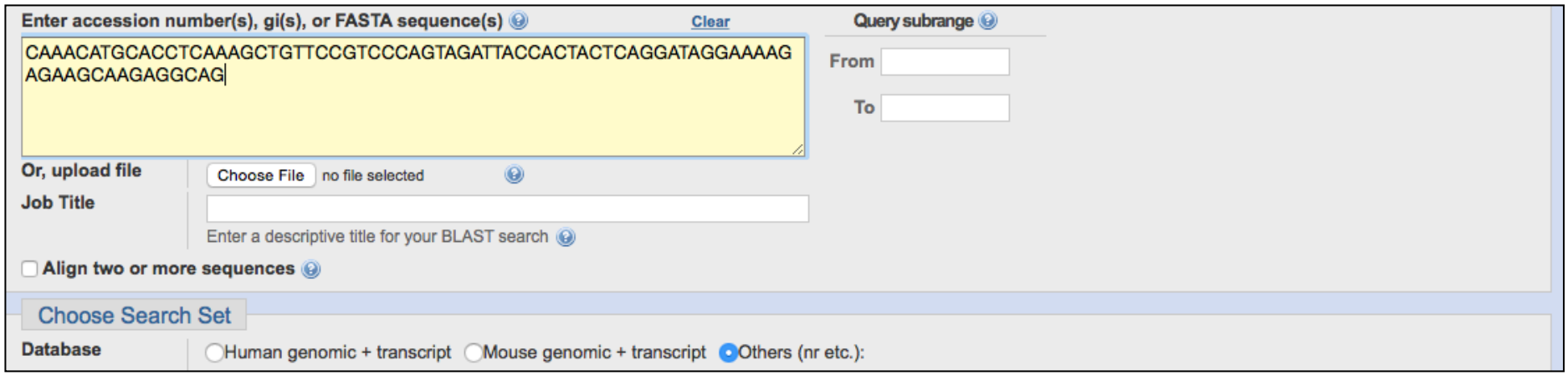
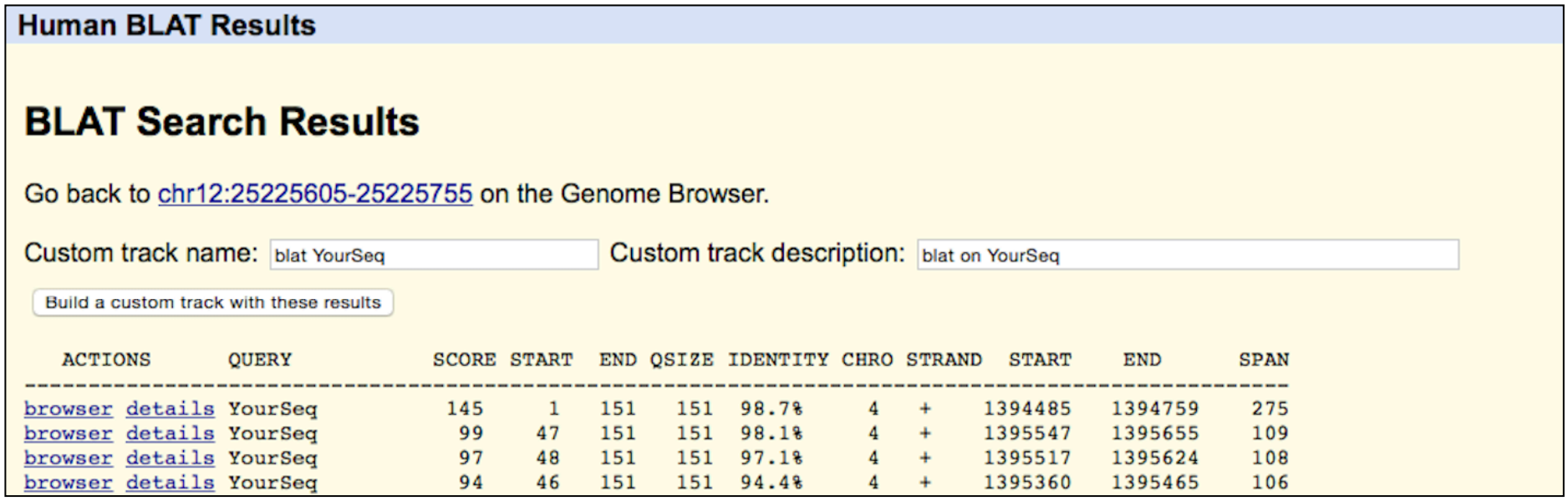


Fig. 15: Search mask of the NCBI nucleotide BLAST database: *This figure shows the search input window with the query sequence CAAACATGCACCTCAAAGCTGTTCCGTCCCAGTAGATTACCACTACTCAGGATAGGAAGGAAGAGAAGAAGCAAGAGGCAG (copied from IGV), using the option to search for hits in the alternative databases ("Others"). This search serves to exclude potential contamination with non-human DNA or hits in human sequences that have not yet been included in the human genome reference. When not using the new unique-dual indexes we regularly see sequencer-related cross-contamination artifacts between samples on the same sequencing chip ("index hopping"* (Costello *et al.*, 2018)*), e.g. wheat DNA in human samples, and highly diluted DNA traces from one patient in the samples of other patients.*



Human BLAT Results

BLAT Search Results

Go back to chr12:25225605-25225755 on the Genome Browser.

Custom track name: blat YourSeq Custom track description: blat on YourSeq

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	145	1	151	151	98.7%	4	+	1394485	1394759	275
browser details	YourSeq	99	47	151	151	98.1%	4	+	1395547	1395655	109
browser details	YourSeq	97	48	151	151	97.1%	4	+	1395517	1395624	108
browser details	YourSeq	94	46	151	151	94.4%	4	+	1395360	1395465	106

Fig. 16: UCSC BLAT sequence search suggests BWA alignment problem. *This figure shows four results for the query sequence (copied via IGV): TTGCAGTTCTGGCACACTTTGCACTCAGGACATTGCCAACCTGCACGTTTTAATGGAGTAACCGCTATATCCAGGCACATTCCATGATAGTGCTGACCACAA GTAGTACAAAAGAACTGATCTAAGAGGTCTCCTGGGCTGTTGCACACTG. The CIGAR code 151M had been computed by the BWA software, i.e. BWA had not found any Indel. In contrast, the BLAT results show one deletion (SPAN 275) and two mismatches (IDENTITY 98.7%) in the top row. Additionally, BLAT found hits for partial sequences in other loci of chromosome 4. From this, we conclude that (a) the BWA alignment was questionable, and (b) homologies are present that may cause artifacts. Due to these problems, this locus must be excluded from the NGS test.*

5. Method for the detection of virus integrations into the human genome - *Vy-PER*

Publication C (see page 62):

Forster M, Szymczak S, Ellinghaus D, Hemmrich G, Rühlemann M, Kraemer L, Mucha S, Wienbrandt L, Stanulla M; UFO Sequencing Consortium within I-BFM Study Group, Franke A.
Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data.
Sci Rep. 2015 Jul 13;5:11534.

In **Chapter 4** I noted that the ubiquitously used `BWA` software from Richard Durbin and Heng Li can occasionally produce erroneous results (**Fig. 16**). These errors can also lead to false-positive conclusions about pathogens, such as viruses, bacteria and protozoa. At the 2012 annual conference of the *Arbeitsgemeinschaft für GenDiagnostik e.V.* Zemin Zhang reported that his group was able to detect hepatitis B virus integrations into the human genome in some cases of liver cancer, using NGS.

Building on Zemin Zhang's method (i.e. mapping sequences to the human genome and using `BLAST` to align the unmapped sequences to virus databases), I introduced step-by-step improvements with the aim of eliminating the large number of false-positive virus sequences (**Fig. 17**).

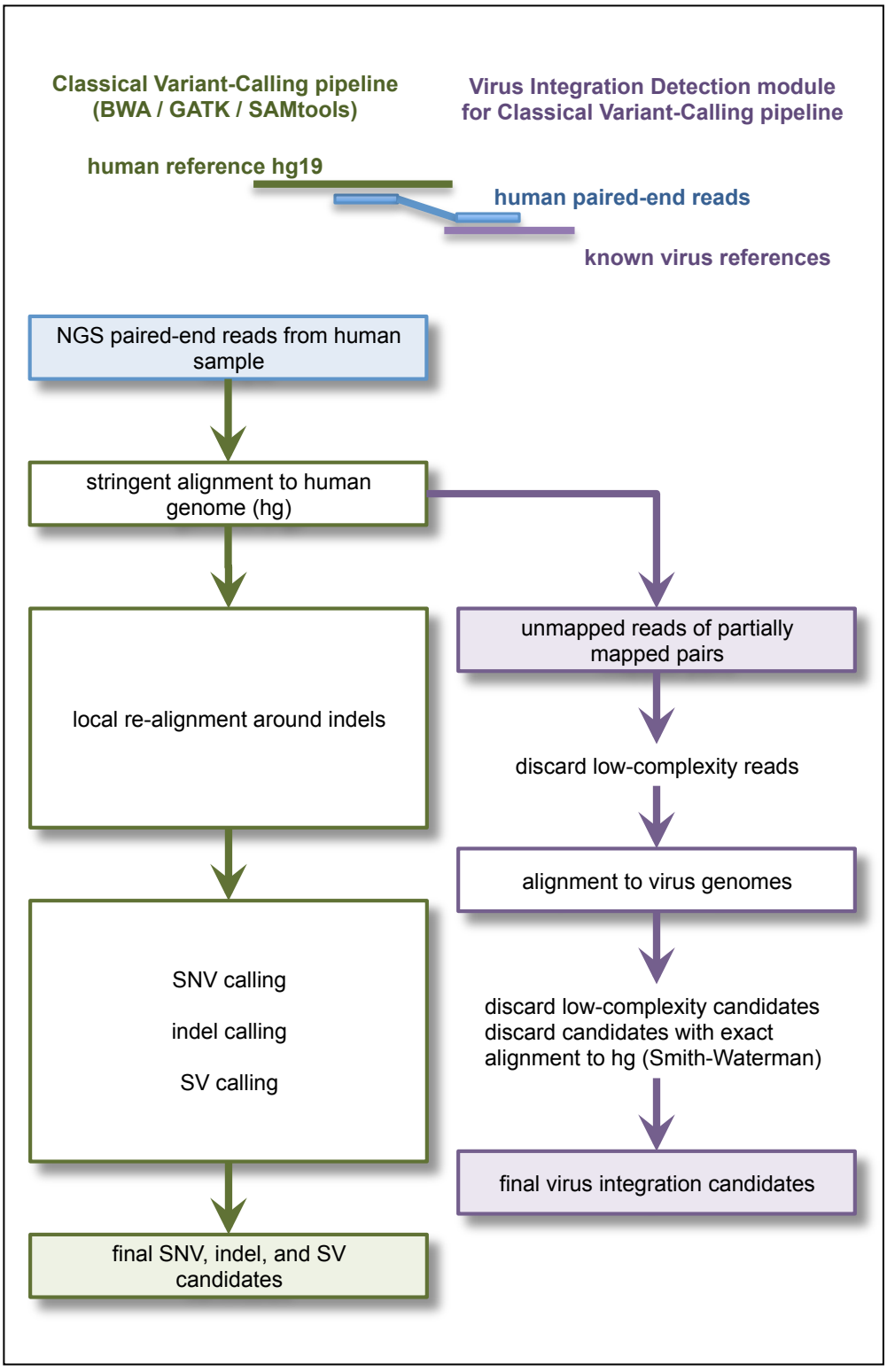


Fig. 17: Paired-end sequencing enables the detection of patient-individual virus integrations into the human genome. *The image is reproduced from my publication and summarizes the algorithm.*

Vy-PER searches for human virus chimeras in paired end sequence pairs. The user may choose whether Vy-PER should consider single chimeric pairs (highest sensitivity) or a specific minimal number of chimeric pairs in the same locus (higher specificity). Vy-PER filters non-specific sequences that consist only of STR sequences or homopolymer sequences in the "virus" subsequence. Vy-PER tests the more complex virus sequences for a possible human origin by mapping them to the human genome using the exact Smith-Waterman algorithm. Using Vy-PER on more than a hundred patients' sequences has shown that even complex-looking "virus" sequences can be identical to sequences in the human genome.

Overcoming my initial reservations concerning the results from Zhang's group, Vy-PER reproduced that HBV integrations can be detected in public liver cancer genome sequence data (Sung *et al.*, 2012) and transcriptome sequence data (Chen *et al.*, 2013): **Figure 18** shows the integration loci of hepatitis B virus fragments into a patient genome determined by Vy-PER. The patient's sequence data originate from transcriptome sequencing by Chen and colleagues. **Figure 19** shows an analysis of the same data set with the highest sensitivity: Allowing individual human-virus chimeras, Vy-PER also detected the PhiX spike-in, which is necessary for the Illumina sequencing method. These PhiX control libraries are normally spiked into the sequencing libraries at a ratio of 1:99 in order to better eliminate the phasing errors (see chapter 1 "Introduction"). We interpret the chimeric PhiX sequences as being PhiX libraries ligated end-to-end with human libraries. Similarly, I occasionally find chimeric sequences where human libraries appear to be ligated end-to-end with libraries from non-integrating pathogens. I speculate that the known PhiX spike-in may potentially allow the estimation of the non-integrated pathogen load.

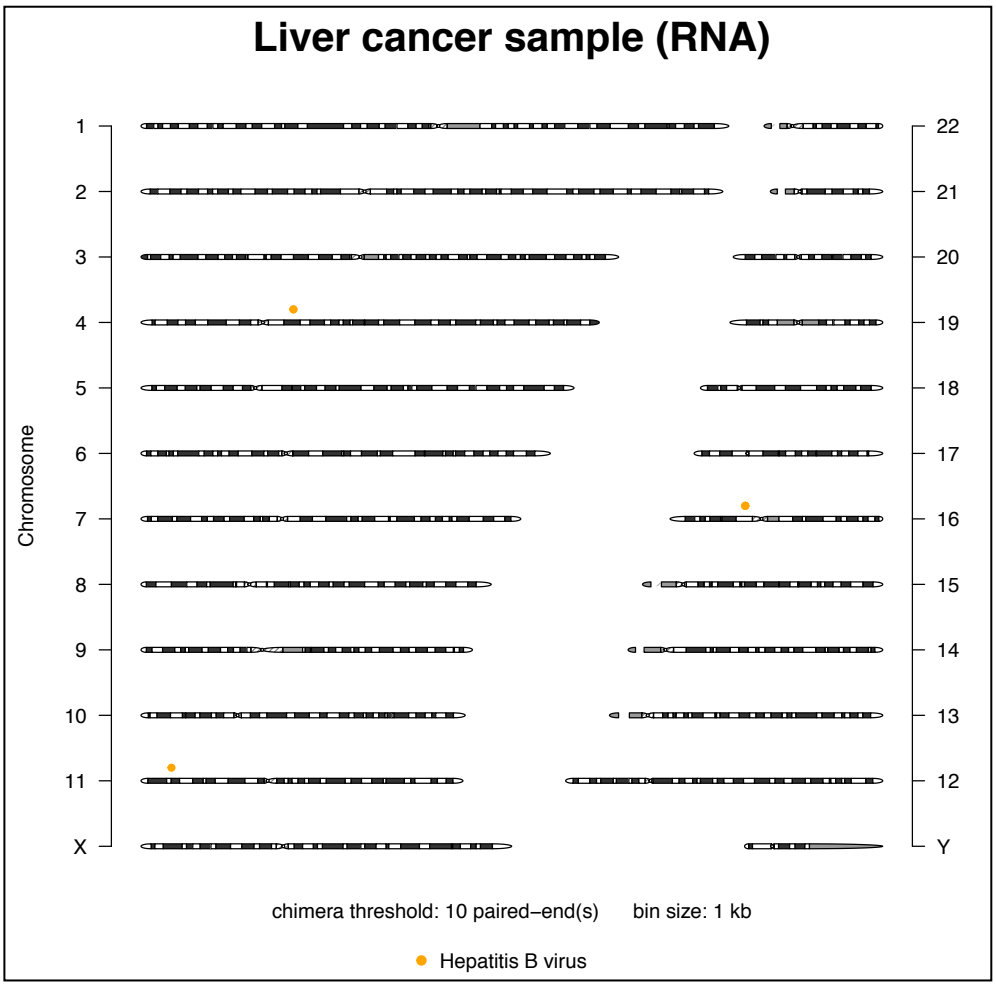


Fig. 18: Vy-PER high specificity results ideogram. *Plot from my publication with HBV integration loci on chromosomes 4, 11 and 16. Only loci with 10 or more chimeric sequence pairs within a genomic seuqence window of 1000 bp were considered. These loci were concordant with the loci published by Chen et al.*

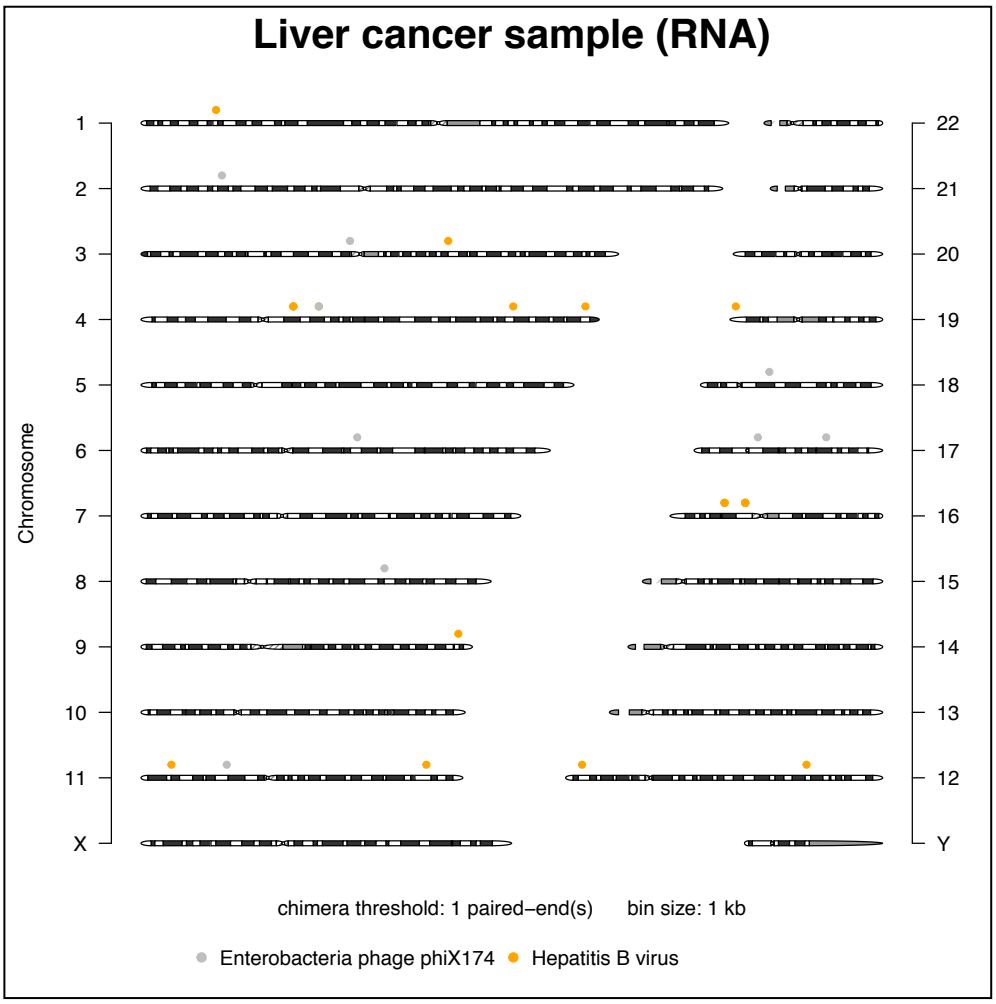


Fig. 19: Vy-PER high sensitivity results ideogram. *Plot from my publication. The HBV integration loci on chromosomes 4, 11 and 16 are concordant to the loci published by Chen et al. In addition, I detected further potential integration loci when chimeric sequence pair singletons were considered. Compared to Fig. 16, these further potential HBV integration loci into the patient genome as well as PhiX chimeric sequence pairs (which are not integrated into the patient genome) can be seen.*

6. Discussion

Unvalidated raw NGS results are currently not suitable for clinical use. This applies to results generated by popular research bioinformatics software as well as data generated by commercial clinical testing software. Therefore, standard practice is to validate the raw data using independent methods, mostly by visually inspecting the sequence alignments in visualization software and, in the case of individual uncertainties, often also by Sanger sequencing. These validation processes are increasingly being formalized with the aim of standardized, reproducible processes. During the accreditation inspection at IKMB in January 2018 by DAkkS auditor Prof. Dieter Lohmann, our clinical testing lab was for the first time requested that our standard procedure for identifying artifacts be recorded in writing. This request is now answered with my publicly available dissertation, which I would like to encourage other clinical testing labs to re-use.

The `pibase` scripts enable the automated validation of single nucleotide variants, except for the case of sequences with indels. I use `pibase` regularly in research to filter technical artifacts and unclear variants from raw variant lists, e.g. in (Zeissig et al. 2014; Fischer et al. 2015). According to `Google Scholar` (`scholar.google.com`) `pibase` is currently cited by 22 publications.

I have used `Vy-PER` for virus analysis in the published leukaemia cases and also for exploratory analysis of other patient sequences. Occasional feedback shows that the topic of false-positive virus detection continues to be one of the research interests of bioinformaticians on the American, Asian, Australian and European continents in 2018 and 2019. According to `Google Scholar`, `Vy-PER` is currently cited by 16 publications.

The Backmapping method from our BMC genomics publication is not used in our current standard bioinformatics pipelines. Nowadays, computing time can be saved by specifying the target regions of interest using a BED file. However, we nowadays use Backmapping "manually" in routine clinical testing when we use the independent - and often more accurate, but slow - mapping software BLAT and BLAST for the technical validation of suspicious sequence alignments to the human genome reference. **Figure 20** shows an example of such an artifact. The sequences probably originate from the patient's pseudogene locus *TYRO3P,* but were assigned to the locus of the similar gene *TYRO3* (**Figure 21**). This shows how important it is to perform Backmapping validation for suspicious variant-calls, especially when exome sequencing or unfamiliar gene sets are used.
According to Google Scholar, the Backmapping paper is currently cited by 10 publications.

It should always be borne in mind that NGS databases and analysis software seldom conform to good craftsmanship. Outstanding exceptions are IGV (Robinson *et al.*, 2011; Thorvaldsdóttir, Robinson and Mesirov, 2013) and HLAssign (Wittig *et al.*, 2015). Three important problems should be remembered:

1. As a rule, bioinformatics software has too many "dependencies" on other software or databases, which means that it is practically unusable from the start or becomes unusable later when an updated version is released. The term for this problem is "dependency hell". When deciding on a bioinformatics solution, particular attention should be paid to this problem.

2. Annotations can be misleading, even in some of the commercial clincial testing software versions that I have tested so far, as well as in the leading open clinical variant database ClinVar.

Annotations must therefore be validated in routine clinical testing. The database issues are known (Lek *et al.*, 2016) and are currently being addressed (keyword search "theatlantic Heidi Rehm PTPN11 Noonan" or https://bit.ly/2mFgxeJ). Today, it is standard practice in clinical testing that variants with population allele frequencies in cohorts of healthy control subjects greater than 1% must be classified as benign. The most important publicly accessible healthy controls are `GnomAD` (http://gnomad.broadinstitute.org/about) and two cohorts of healthy old control persons: the `Welderly` cohort (Erikson *et al.*, 2016) and the cancer-free `FLOSSIES` - Fabulous Ladies Over Seventy (https://whi.color.com).

3. Quality assurance of databases or software is passed on to the users when "agile software development" is performed, even in the case of some software from commercial providers (**Figure 22**). Therefore, in accredited clinical testing labs a careful validation of each new software version becomes necessary, if solutions from agile software developers are used.

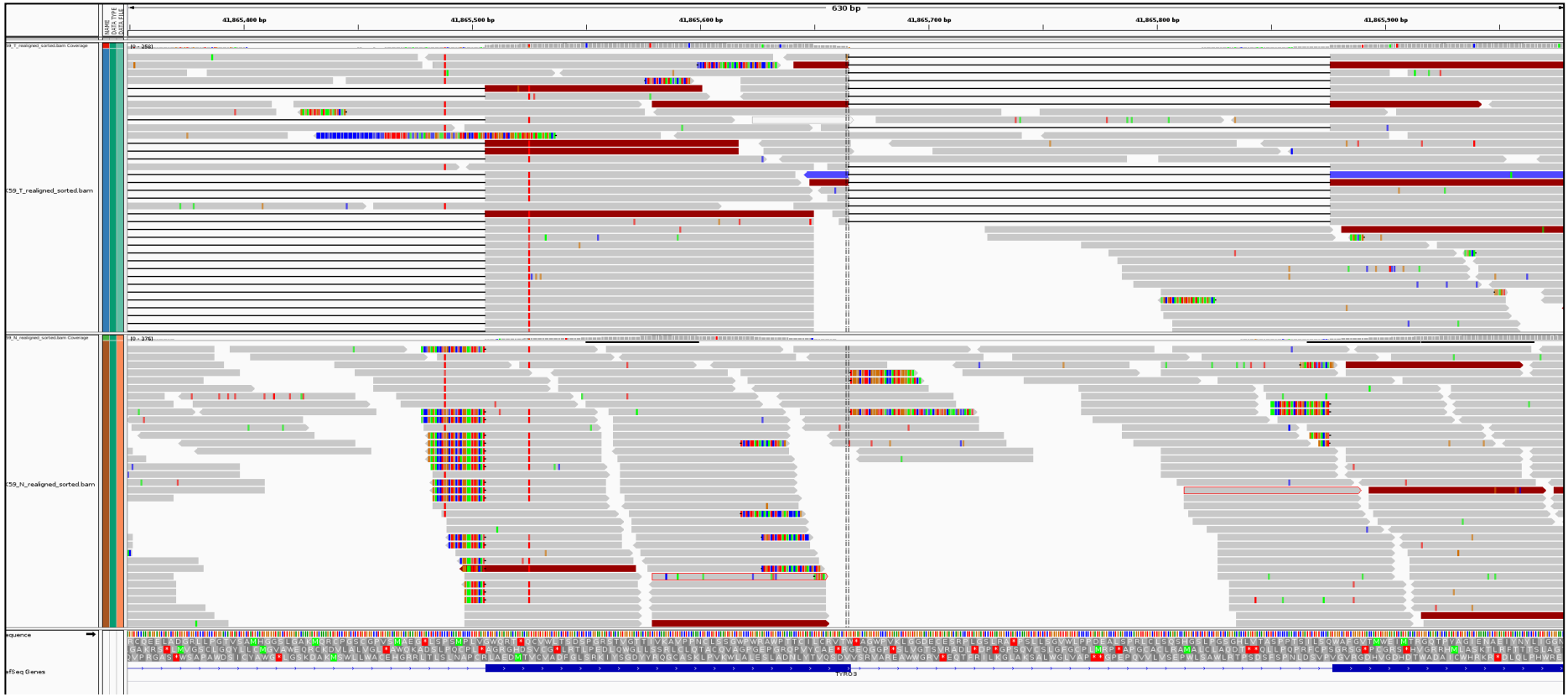


Fig. 20: Mismapping sequences in the TYRO3 gene region, displayed in IGV. *Top: Sequences of length 150 bp with deletions of intron length. Below: Sequences of length 75 bp, whose ends were cut off by BWA ("softclipping"), represented by the colored ends. The sequences come from the pseudogene TYRO3P. (Sequences from FK59, from an ongoing sequencing project with Prof. Dirk Bauerschlag and Prof. Norbert Arnold.)*

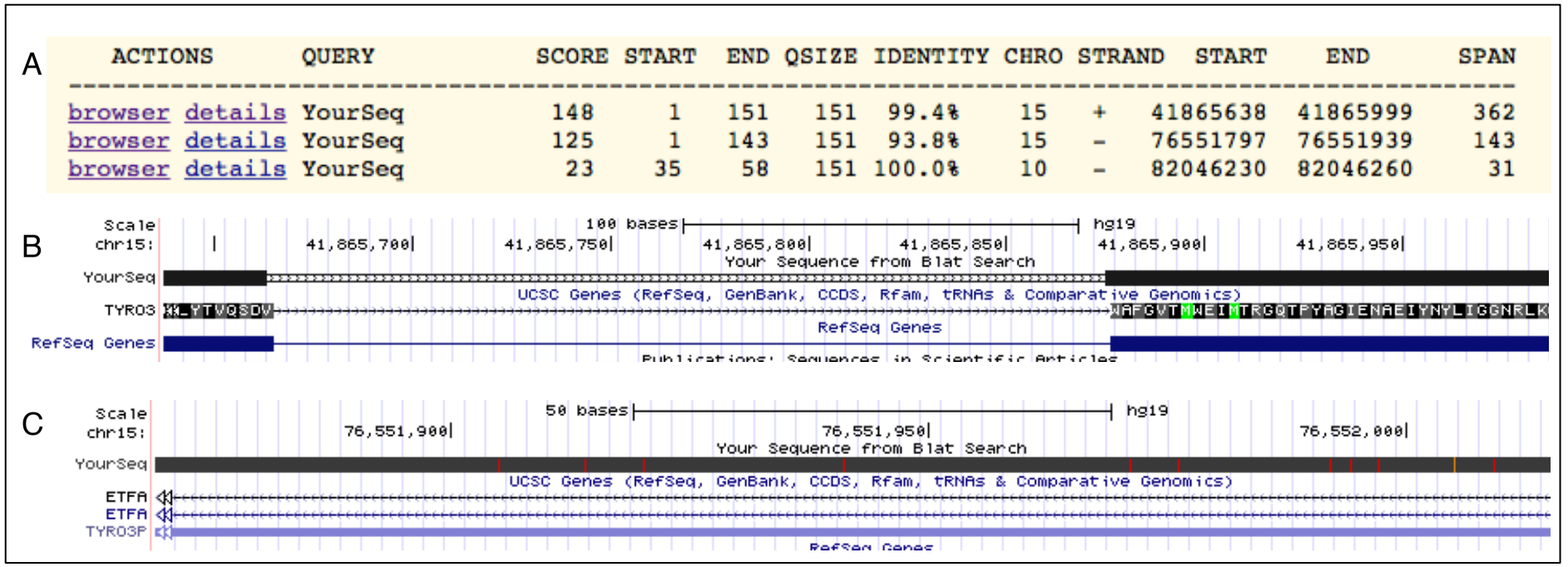


Fig. 21: Inspection with UCSC BLAT and the UCSC Genome Browser of a sequence mismapping onto TYRO3. *(A) BLAT results. (B) First BLAT result in the Genome Browser in the TYRO3 locus (with deletion). (C) Second BLAT result in the TYRO3P locus (without deletion, but with 11 single nucleotide mismatches). Interpretation: The sequence originates from the pseudogene TYRO3P; presumably this is a highly polymorphic locus.*

README_NISTv3.3.2.txt

```
Best Practices for Using High-confidence Calls:
Benchmarking variant calls is a complex process, and best practices are still being developed by the Global Alliance for Genomics and
Health (GA4GH) Benchmarking Team (https://github.com/ga4gh/benchmarking-tools/).  Several things are important to consider when
benchmarking variant call accuracy:
1. Complex variants (e.g., nearby SNPs and indels or block substitutions) can often be represented correctly in multiple ways in the
vcf format.  Therefore, we recommend using sophisticated benchmarking tools like those developed by members of the GA4GH benchmarking
team.  The latest version of hap.py (https://github.com/Illumina/hap.py) now allows the user to choose between hap.py's xcmp and RTG's
vcfeval comparison tools, both of which perform sophisticated variant comparisons.  Preliminary tests indicate they perform very
similarly, but vcfeval matches some additional variants where only part of a complex variant is called.
```

```
##FORMAT=<ID=BD,Number=1,Type=String,Description="Decision for call (TP/FP/FN/N)">
##FORMAT=<ID=BK,Number=1,Type=String,Description="Sub-type for decision (match/mismatch type)">
##FORMAT=<ID=BI,Number=1,Type=String,Description="Additional comparison information">
##FORMAT=<ID=QQ,Number=1,Type=Float,Description="Variant quality for ROC creation.">
##FORMAT=<ID=BVT,Number=1,Type=String,Description="High-level variant type (SNP|INDEL).">
##FORMAT=<ID=BLT,Number=1,Type=String,Description="High-level location type (het|homref|hetalt|homalt|nocall).">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  TRUTH   QUERY
1       201334382       .       C       A       50      .       BS=201334382    GT:BD:BK:BI:BVT:BLT:QQ  0|1:FN:.:tv:SNP:het:.   ./.:.:.:.:NOCALL:nocall:0
1       201334382       .       C       A       100     .       BS=201334382    GT:BD:BK:BI:BVT:BLT:QQ  ./.:.:.:.:NOCALL:nocall:.       0/1:FP:.:tv:SNP:het:100
14      23892888        .       T       G       50      .       BS=23892888     GT:BD:BK:BI:BVT:BLT:QQ  0/1:FN:.:tv:SNP:het:.   ./.:.:.:.:NOCALL:nocall:0
14      23892888        .       T       G       100     .       BS=23892888     GT:BD:BK:BI:BVT:BLT:QQ  ./.:.:.:.:NOCALL:nocall:.       0/1:FP:.:tv:SNP:het:100
14      23899027        .       G       T       50      .       BS=23899027     GT:BD:BK:BI:BVT:BLT:QQ  0/1:FN:.:tv:SNP:het:.   ./.:.:.:.:NOCALL:nocall:0
14      23899027        .       G       T       100     .       BS=23899027     GT:BD:BK:BI:BVT:BLT:QQ  ./.:.:.:.:NOCALL:nocall:.       0/1:FP:.:tv:SNP:het:100
14      23902753        .       G       A       100     .       BS=23902753     GT:BD:BK:BI:BVT:BLT:QQ  0|1:TP:gm:ti:SNP:het:100        0/1:TP:gm:ti:SNP:het:100
```

Fig. 22: Error in the comparison software `hap.py`. *Top: Extract from the README file downloaded along with the Genome-in-a-Bottle reference data (standard recommended by National Institute of Standards and Technology, NIST). Bottom: VCF file with multiple comparison errors by `hap.py` between the reference dataset and the test dataset: The first three of the four variants shown here were compared incorrectly. The `hap.py` summary files reported only 25% concordance between the Genome-in-a-Bottle reference variants and our panel sequencing variants using the Genome-in-a-Bottle reference DNA (from the Illumina MiSeq instrument!). Note : Various `hap.py` settings were tried, including "Genotype" setting instead of "Haplotype" setting.*

Significant new findings

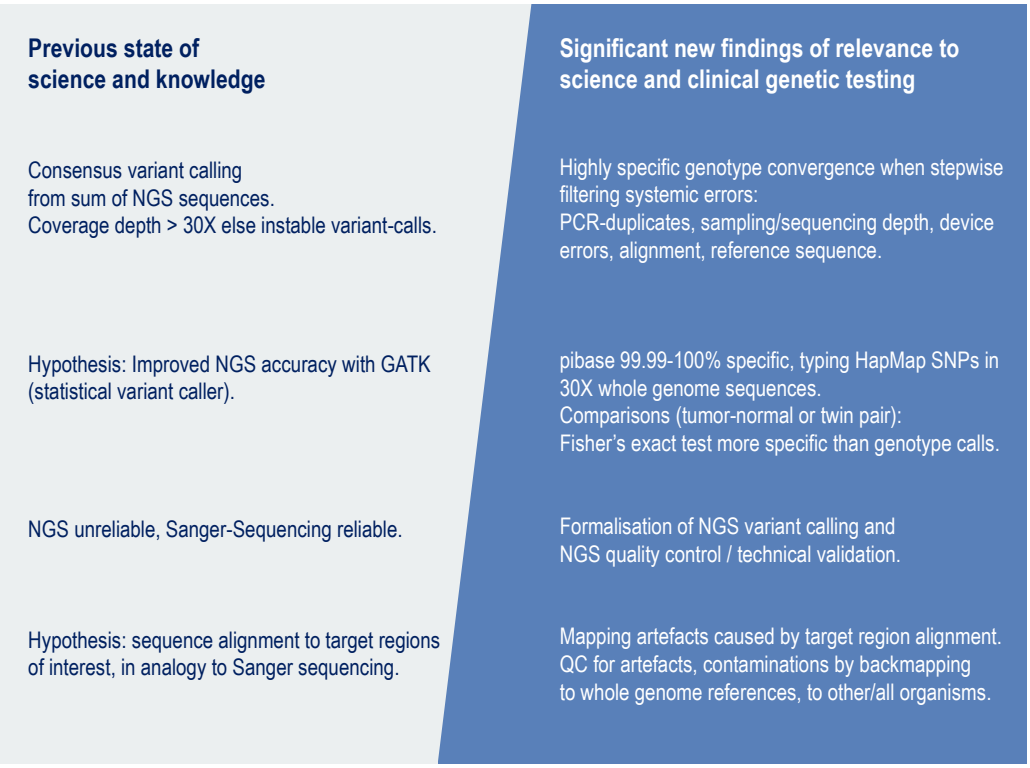


Fig. 23: Comparison of the previous state of science and knowledge with significant new findings from my works

Figure 23 summarizes significant new findings that have been gained through my contributions presented here. Some of my findings build on previous work that others carried out at IKMB (Melum *et al.*, 2010). Melum and colleagues used the concept of *unique start points* to eliminate potential sequence duplicates; by randomly subsampling sequences they observed that NGS variant-calling became increasingly uncertain at sequencing depths of less than 30X.

I built on their work, exploring additional criteria of my own choice (comparison between duplicated and deduplicated sequences, comparison between short and long sequences, comparison between sequences with many or few mismatches to the reference genome, comparison between sequences with high or low alignment quality, comparison between variants with high or low `Phred` score, use of a

strand bias criterion as well as the reference sequence context). When I submitted my first `pibase` manuscript version, these criteria were still quite novel considerations and they generated resistance. Today, it has become common practice to use some of the criteria described in the `pibase` publication.

The use of Fisher's exact test to improve tumor/normal comparisons and intra-twin comparisons was described in the `pibase` manuscript and proposed by co-author Peter Forster. We were the first to publish the use of the Fisher's exact test for NGS and it is now used in many analysis programs for tumor NGS data. The idea of the `Backmapping` method is derived from co-author Andreas Keller's variant quality control procedure: He used `BLAST` for mapping to the human reference his test sequences that consisted of each putative variant plus 50 bases of upstream and downstream padding from the reference sequence.

My dissertation thus clarifies the specificity of NGS variant-calling and provides tools or procedures how artifacts can be detected and removed. This expert guidance is solidly founded on nearly ten years of intensive study of NGS data.

7. Outlook

The primary aim of my future work is to develop precision medicine approaches and translate these into general patient care through interdisciplinary cooperations with natural scientists and clinicians.

Several ongoing pilot projects have already generated human sample collections and noteworthy results that merit potential scientific publication, initiation of clinical trials, and grant applications for third-party funding to finance the required human and material resources.

The currently ongoing projects are based at several clinics that collect tissue and/or blood of cancer patients. Their goals are to understand and ultimately use patient-specific molecular biomarkers in longitudinal follow-ups to help guide therapeutic interventions. These promising strategies are already being implemented by the large US cancer centers, especially for rich or privately insured patients. However, my wish is that normal patients with just the basic German public health insurance may also benefit from these precision medicine strategies at economically acceptable costs.

References

1000 Genomes Project Consortium *et al.* (2010) 'A map of human genome variation from population-scale sequencing.', ***Nature***, 467(7319), pp. 1061–73.

Chen, Y. *et al.* (2013) 'VirusSeq: Software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue', ***Bioinformatics***, 29(2), pp. 266–267.

Costello, M. *et al.* (2018) 'Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms.', ***BMC Genomics***, 19(1):332.

Do, H. and Dobrovic, A. (2012) 'Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase.', ***Oncotarget***, 3(5), pp. 546–58.

Erikson, G. A. *et al.* (2016) 'Whole-Genome Sequencing of a Healthy Aging Cohort.', ***Cell***, 165(4), pp. 1002–11.

Ewing, B. *et al.* (1998) 'Base-calling of automated sequencer traces using phred. I. Accuracy assessment.', ***Genome Res.***, 8(3), pp. 175–85.

Fischer, U. *et al.* (2015) 'Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options.', ***Nat. Genet.***, 47(9), pp. 1020–9.

Forster, L. *et al.* (2010) 'Evaluating length heteroplasmy in the human mitochondrial DNA control region.', ***Int. Journal Legal Med.***, 124(2), pp. 133–42.

Forster, P. *et al.* (2015) 'Elevated germline mutation rate in teenage fathers.', ***Proc. R. Soc. B.***, 282(1803):20142898.

Girard, S. L. *et al.* (2011) 'Increased exonic de novo mutation rate in individuals with schizophrenia.', ***Nat. Genet.***, 43(9), pp. 860–863.

Kircher, M., Stenzel, U. and Kelso, J. (2009) 'Improved base calling for the Illumina Genome Analyzer using machine learning strategies.', ***Genome Biol.***, 10(8):R83.

Ledergerber, C. and Dessimoz, C. (2011) 'Base-calling for next-generation sequencing platforms.', ***Brief. Bioinformatics***, 12(5), pp. 489–97.

Lek, M. *et al.* (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', ***Nature***, 536(7616), pp. 285–291.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools.', ***Bioinformatics***, 25(16), pp. 2078–9.

McKenna, A. *et al.* (2010) 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.', ***Genome Res.***, 20(9), pp. 1297–303.

McMurray, C. T. (2010) ‘Mechanisms of trinucleotide repeat instability during human development.’, ***Nature Rev. Genetics***, 11(11), pp. 786–99.

Melum, E. *et al.* (2010) ‘SNP discovery performance of two second-generation sequencing platforms in the NOD2 gene region.’, ***Human Mutat.***, 31(7), pp. 875–85.

Niu, B. *et al.* (2014) ‘MSIsensor: Microsatellite instability detection using paired tumor-normal sequence data’, ***Bioinformatics***, 30(7), pp. 1015–1016.

Petersen, B.-S. *et al.* (2014) ‘Whole genome and exome sequencing of monozygotic twins discordant for Crohn’s disease.’, ***BMC Genomics***, 15(1):564.

Robinson, J. T. *et al.* (2011) ‘Integrative genomics viewer.’, ***Nat. Biotechnology***, 29(1), pp. 24–6.

Sung, W.-K. *et al.* (2012) ‘Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma.’, ***Nature Genet.***, 44(7), pp. 765–9.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) ‘Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration’, ***Brief. Bioinformatics***, 14(2), pp. 178–192.

Umar, A. *et al.* (2004) ‘Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability.’, ***J. Natl. Cancer Inst.***, 96(4), pp. 261–8.

Wang, Q. *et al.* (2013) ‘Detecting somatic point mutations in cancer genome sequencing data: A comparison of mutation callers’, ***Genome Med.***, 5(10), pp. 1–8.

Wittig, M. *et al.* (2015) ‘Development of a high-resolution NGS-based HLA-typing and analysis pipeline.’, ***Nucleic Acids Res.***, 43(11):e70.

Xu, C. (2018) ‘A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data’, ***Comput. Struct. Biotechnol. J.***, 16, pp. 15–24.

Zeissig, S. *et al.* (2014) ‘Early-onset Crohn’s disease and autoimmunity associated with a variant in CTLA-4.’, ***Gut***, 64(12), pp. 1889–97.

Appendices

I. Declaration

II. Acknowledgments

III. Curriculum Vitae

IV. Peer-Reviewed Publications 2012-2018

V. Publication A for doctoral degree

VI. Publication B for doctoral degree

VII. Publication C for doctoral degree

VIII. Patent for pibase

IX. Vy-PER presented as poster and spotlight talk at ISMB Annual Conference Satellite Symposium HiTSeq 2014

X. HiTSeq 2014 Poster Award for Vy-PER

I. Declaration

I declare that I wrote this dissertation entirely by myself, both in its German draft version and also in its submitted English version, in form and content. Apart from the advice of my academic supervisors, all relevant scientific sources or permissions are listed in the references or in the figure legends.

I declare that this dissertation has not been submitted elsewhere within the procedure of an examination for an academic degree, neither in its entirety nor in part nor in a derivative work thereof.

I declare that this dissertation has not been published in its entirety or in part except for the three underlying peer reviewed publications.

I declare that this dissertation has been produced in accordance with the rules of good scientific practice of the Deutsche Forschungsgemeinschaft.

I declare that no previously awarded academic degree of mine has ever been revoked.

Three peer-reviewed scientific publications, in *Nucleic Acids Research*, *BMC Genomics* and *Scientific Reports* respectively, form the basis of my submission for a doctoral degree. These publications are co-authored with others, as is standard practice in this field of science.

I declare that my own work contributions are as detailed in the following:

Publication A in *Nucleic Acids Research*: I conceived and wrote the manuscript draft, managed co-author contributions, edited the revisions, and co-wrote the point-by-point responses to the reviewers. I developed

and tested the underlying `pibase` tools and wrote up the website. I performed the bioinformatic analyses (except that Britt-Sabina Petersen performed most of the monozygotic twin analyses) and data interpretations for the initial manuscript submission and all revisions.

Publication B in *BMC Genomics*: I performed bioinformatic analyses and collaboratively interpreted the data with co-authors. I formulated the `Backmapping` method (which Ingo Thomsen then programmed). I co-wrote the manuscript and all revisions.

Publication C in *Scientific Reports*: I conceived and wrote the manuscript and all revisions. I conceived, developed and tested the `Vy-PER` python and bash scripts and the website. (Silke Szymczak wrote the R graphics script). I performed the bioinformatic analyses using `Vy-PER.` (Malte Rühlemann and Georg Hemmrich-Stanisak performed the comparative bioinformatics using `VirusFinder`, `VirusSeq`, and `SURPI`).

I used DeepL Pro (https://www.deepl.com/translator) extensively to help translate into English my own German texts. I then corrected the DeepL Pro translations sentence by sentence.

II. Acknowledgments

I am sincerely grateful to my wife, family, and friends,

the board of directors at the Institute of Clinical Molecular Biology,
especially Professor Andre Franke,

my colleagues at the Institute of Clinical Molecular Biology,
at Kiel University, also known as Christian-Albrechts-Universität zu Kiel,
and at the University Hospital Schleswig Holstein,

and my cooperation partners

for their friendship and loyalty.

I am deeply grateful to Professor Andre Franke, Professor Philip Rosenstiel, Professor Stefan Schreiber, Professor Norbert Arnold and my colleagues for supporting my strategic goals.

My work was funded by grants from the EU Seventh Framework Programme FP7/2007-2013, 201418 (READNA) and 262055 (ESGI), the German Ministry of Education and Research (BMBF) within the e:Med project 01ZX1306A (sysINFLAME), the Deutsche Forschungsgemeinschaft (DFG) Cluster of Excellence ‘Inflammation at Interfaces’, and the German Federal Office for Radiation Protection (BfS, St.Sch. 3611S70014). I thank the funding bodies and the successful applicants Professor Andre Franke, Professor Philip Rosenstiel, Professor Stefan Schreiber, Professor David Ellinghaus and Professor Martin Stanulla.

III. Curriculum Vitae

Personal Data

1964	born in Kassel
since 2002	married

Graduation

1983	Abitur, Max-Planck-Schule Kiel
1989	Dipl.-Ing., Technische Universität Braunschweig

Work stations

1989-1999	Freelance software developer
since 1999	Director of Fluxus Technology Ltd (commercial consulting and sales, free phylogenetic software Network > 8000 citations, academic software)
since 2009	Scientist at Institut für Klinische Molekularbiologie, Christian-Albrechts-Universität zu Kiel
since 2016	Head of oncological research in the Franke Group, Institut für Klinische Molekularbiologie

EU grant as coordinator - for Fluxus Technology Ltd:

2000-2001	ECITTT, High-level Scientific Conference HPCF-CT-2000-00118, EU FP5, €35000

Participation in consortia - for Institut für Klinische Molekularbiologie:

since 2009	EU-Konsortia: READNA, ESGI, EASI-Genomics, Changing Cancer Care
since 2016	Deutsches Konsortium Familiärer Brust- und Eierstockkrebs
since 2017	Kieler Onkologienetzwerk (KON)
since 2017	Nordic Alliance for Sequencing and Precision Medicine

Further cooperation partners in the field of translational precision medicine research:

since 2014 Hannover Medical School

since 2016 Avera Molecular and Experimental Medicine, Sioux Falls

since 2017 Department of Gynaecology and Obstetrics, UKSH Kiel

since 2017 Institut für Klinische Chemie UKSH Kiel

since 2017 Department of Dermatology, UKSH Kiel

since 2017 Department of General and Thoracic Surgery, UKSH Kiel

since 2017 Lithuanian University of Health Science, Kaunas

IV. Peer-Reviewed Publications 2012-2018

Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, **Forster M**, Stade B, Franke A, Mayer J, Spangler J, McLaughlin S, Shah M, Lee C, Harkins TT, Sartori A, Moreno-Estrada A, Henn B, Sikora M, Semino O, Chiaroni J, Rootsi S, Myres NM, Cabrera VM, Underhill PA, Bustamante CD, Vigl EE, Samadelli M, Cipollini G, Haas J, Katus H, O'Connor BD, Carlson MR, Meder B, Blin N, Meese E, Pusch CM, Zink A.
New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing.
Nat Commun. 2012 Feb 28;3:698.

ElSharawy A, Warner J, Olson J, **Forster M**, Schilhabel MB, Link DR, Rose-John S, Schreiber S, Rosenstiel P, Brayer J, Franke A.
Accurate variant detection across non-amplified and whole genome amplified DNA using targeted next generation sequencing.
BMC Genomics. 2012 Sep 20;13:500.

Elsharawy A*, **Forster M***, Schracke N, Keller A, Thomsen I, Petersen BS, Stade B, Stähler P, Schreiber S, Rosenstiel P, Franke A.
Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing.
BMC Genomics. 2012 Aug 22;13:417.
* joint first author

Publication for doctoral degree

Filtering false positives due to alignment

Forster M*, Forster P*, Elsharawy A*, Hemmrich G, Kreck B, Wittig M, Thomsen I, Stade B, Barann M, Ellinghaus D, Petersen BS, May S, Melum E, Schilhabel MB, Keller A, Schreiber S, Rosenstiel P, Franke A.
From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software.
Nucleic Acids Res. 2013 Jan 7;41(1):e16

Publication for doctoral degree

Filtering false positive single nucleotide substitutions

Ellinghaus D, Zhang H, Zeissig S, Lipinski S, Till A, Jiang T, Stade B, Bromberg Y, Ellinghaus E, Keller A, Rivas MA, Skieceviciene J, Doncheva NT, Liu X, Liu Q, Jiang F, **Forster M**, Mayr G, Albrecht M, Häsler R, Boehm BO, Goodall J, Berzuini CR, Lee J, Andersen V, Vogel U, Kupcinskas L, Kayser M, Krawczak M, Nikolaus S, Weersma RK, Ponsioen CY, Sans M, Wijmenga C, Strachan DP, McArdle WL, Vermeire S, Rutgeerts P, Sanderson JD, Mathew CG, Vatn MH, Wang J, Nöthen MM, Duerr RH, Büning C, Brand S, Glas J, Winkelmann J, Illig T, Latiano A, Annese V, Halfvarson J, D'Amato M, Daly MJ, Nothnagel M, Karlsen TH, Subramani S, Rosenstiel P, Schreiber S, Parkes M, Franke A.
Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies.
Gastroenterology. 2013 Aug;145(2):339-47.

Wittig M, Anmarkrud JA, Kässens JC, Koch S, **Forster M**, Ellinghaus E, Hov JR, Sauer S, Schimmler M, Ziemann M, Görg S, Jacob F, Karlsen TH, Franke A.
Development of a high-resolution NGS-based HLA-typing and analysis pipeline.
Nucleic Acids Res. 2015 Jun 23;43(11):e70.

Publication for doctoral degree

Filtering false positive virus sequences in human data

Forster M*, Szymczak S*, Ellinghaus D, Hemmrich G, Rühlemann M, Kraemer L, Mucha S, Wienbrandt L, Stanulla M; UFO Sequencing Consortium within I-BFM Study Group, Franke A.
Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data.
Sci Rep. 2015 Jul 13;5:11534.

Fischer U*, **Forster M***, Rinaldi A,* Risch T*, Sungalee S*, Warnatz HJ*, Bornhauser B, Gombert M, Kratsch C, Stütz AM, Sultan M, Tchinda J, Worth CL, Amstislavskiy V, Badarinarayan N, Baruchel A, Bartram T, Basso G, Canpolat C, Cario G, Cavé H, Dakaj D, Delorenzi M, Dobay MP, Eckert C, Ellinghaus E, Eugster S, Frismantas V, Ginzel S, Haas OA, Heidenreich O, Hemmrich-Stanisak G, Hezaveh K, Höll JI, Hornhardt S, Husemann P, Kachroo P, Kratz CP, Te Kronnie G, Marovca B, Niggli F, McHardy AC, Moorman AV, Panzer-Grümayer R, Petersen BS, Raeder B, Ralser M, Rosenstiel P, Schäfer D, Schrappe M, Schreiber S, Schütte M, Stade B, Thiele R, von der Weid N, Vora A, Zaliova M, Zhang L, Zichner T, Zimmermann M, Lehrach H, Borkhardt A, Bourquin JP, Franke A, Korbel JO, Stanulla M, Yaspo ML.
Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options.
Nat Genet. 2015 Sep;47(9):1020-1029.
* joint first author

Zeissig S, Petersen BS, Tomczak M, Melum E, Huc-Claustre E, Dougan SK, Laerdahl JK, Stade B, **Forster M**, Schreiber S, Weir D, Leichtner AM, Franke A, Blumberg RS.
Early-onset Crohn's disease and autoimmunity associated with a variant in CTLA-4.
Gut. 2015 Dec;64(12):1889-97.

McGinn S, Bauer D, Brefort T, Dong L, El-Sagheer A, Elsharawy A, Evans G, Falk-Sörqvist E, **Forster M**, Fredriksson S, Freeman P, Freitag C, Fritzsche J, Gibson S, Gullberg M, Gut M, Heath S, Heath-Brun I, Heron AJ, Hohlbein J, Ke R, Lancaster O, Le Reste L, Maglia G, Marie R, Mauger F, Mertes F, Mignardi M, Moens L, Oostmeijer J, Out R, Pedersen JN, Persson F, Picaud V, Rotem D, Schracke N, Sengenes J, Stähler PF, Stade B, Stoddart D, Teng X, Veal CD, Zahra N, Bayley H, Beier M, Brown T, Dekker C, Ekström B, Flyvbjerg H, Franke A, Guenther S, Kapanidis AN, Kaye J, Kristensen A, Lehrach H, Mangion J, Sauer S, Schyns E, Tost J, van Helvoort JM, van der Zaag PJ, Tegenfeldt JO, Brookes AJ, Mir K, Nilsson M, Willcocks JP, Gut IG.
New technologies for DNA analysis--a review of the READNA Project.
N Biotechnol. 2016 May 25;33(3):311-30.

Meißner T, Mark A, Williams C, Berdel WE, Wiebe S, Kerkhoff A, Wardelmann E, Gaiser T, Müller-Tidow C, Rosenstiel P, Arnold N, Leyland-Jones B, Franke A, Stanulla M, **Forster M**.
Metastatic triple-negative breast cancer patient with TP53 tumor mutation experienced 11 months progression-free survival on bortezomib monotherapy without adverse events after ending standard treatments with grade 3 adverse events.
Cold Spring Harb Mol Case Stud. 2017 Jul 5;3(4). pii: a001677.

Flachsbart F, Dose J, Gentschew L, Geismann C, Caliebe A, Knecht C, Nygaard M, Badarinarayan N, ElSharawy A, May S, Luzius A, Torres GG, Jentzsch M, **Forster M**, Häsler R, Pallauf K, Lieb W, Derbois C, Galan P, Drichel D, Arlt A, Till A, Krause-Kyora B, Rimbach G, Blanché H, Deleuze JF, Christiansen L, Christensen K, Nothnagel M, Rosenstiel P, Schreiber S, Franke A, Sebens S, Nebel A.
Identification and characterization of two functional variants in the human longevity gene FOXO3.
Nat Commun. 2017 Dec 12;8(1):2063. Erratum in: ***Nat Commun***. 2018 Jan 17;9(1):320.

Kachroo P, Szymczak S, Heinsen FA, **Forster M**, Bethune J, Hemmrich-Stanisak G, Baker L, Schrappe M, Stanulla M, Franke A.
NGS-based methylation profiling differentiates TCF3-HLF and TCF3-PBX1 positive B-cell acute lymphoblastic leukemia.
Epigenomics. 2018 Feb;10(2):133-147.

Streleckiene G, Reid HM, Arbold N, Bauerschlag DO, **Forster M**.
Quantifying cell free DNA in Urine: The Impact of Gender, Inter-Individual Variation and Isolation Method.
Biotechniques. 2018 May;64(5):225-230.

Forster M, Mark A, Egberts F, Rosati E, Rodriguez E, Stanulla M, Bauerschlag D, Schem C, Maass N, Amallraja A, Murphy KK, Prouse BR, Sulaiman RA, Young BM, Mathiak M, Hemmrich-Stanisak G, Ellinghaus D, Weidinger S, Rosenstiel P, Arnold A, Leyland-Jones B, Williams CB, Franke A*, Meißner T*.
RNA based individualized drug selection in breast cancer patients without patient-matched normal tissue.
Oncotarget. 2018 Aug 17;9(64):32362-32372.

von Frieling J, Fink C, Hamm J, Klischies K, **Forster M**, Bosch TCG, Roeder T, Rosenstiel P, Sommer F.
Grow With the Challenge - Microbial Effects on Epithelial Proliferation, Carcinogenesis, and Cancer Therapy.
Front Microbiol. 2018 Sep 20;9:2020.

V. Publication A for doctoral degree: From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software.

Forster M*, Forster P*, Elsharawy A*, Hemmrich G, Kreck B, Wittig M, Thomsen I, Stade B, Barann M, Ellinghaus D, Petersen BS, May S, Melum E, Schilhabel MB, Keller A, Schreiber S, Rosenstiel P, Franke A.

From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software.

Nucleic Acids Res. 2013 Jan 7;41(1):e16

VI. Publication B for doctoral degree: Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing.

Elsharawy A*, **Forster M***, Schracke N, Keller A, Thomsen I, Petersen BS, Stade B, Stähler P, Schreiber S, Rosenstiel P, Franke A.

Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing.

***BMC Genomics*.** 2012 Aug 22;13:417.

* joint first authors

VII. Publication C for doctoral degree: Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data.

Forster M*, Szymczak S*, Ellinghaus D, Hemmrich G, Rühlemann M, Kraemer L, Mucha S, Wienbrandt L, Stanulla M; UFO Sequencing Consortium within I-BFM Study Group, Franke A.

Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data.

Sci Rep. 2015 Jul 13;5:11534.

VIII. Patent for pibase

US Patent Application for ACCURATE COMPARISON AND VALIDATION OF SINGLE NUCLEOTIDE VARIANTS Patent Application (Application #20130245958)

Publication number: 20130245958

Type: Application
Filed: Mar 12, 2013
Publication Date: Sep 19, 2013
Applicant: Siemens Aktiengesellschaft (Munich)

Inventors: **Michael FORSTER** (Kiel), Andre FRANKE (Kronshagen), Andreas KELLER (Puettlingen)

Application Number: 13/795,492

IX. ISMB HiTSeq 2014 Poster and Spotlight Talk on Vy-PER

ISMB Annual Conference Satellite Symposium HiTSeq 2014

Title of Poster and Spotlight Talk:

Herpes beware: eliminating false positive virus detections in NGS data resulting from alignment biases

Herpes beware: eliminating false positive virus detections in NGS data resulting from alignment biases

Michael Forster*, **Silke Szymczak***, **David Ellinghaus**, **Georg Hemmrich**, **Lars Kraemer**, **Sören Mucha**, **Lars Wienbrandt**, **Martin Stanulla for the UFO Sequencing Consortium within the ALL-BFM Study Group and Andre Franke**

* these two authors share first authorship, manuscript in revision

Motivations: therapeutical safety and etiological research

1. Safety of therapeutical virus integrations[1,2]

REVIEW ARTICLE

Retrovirus Vectors: Toward the Plentivirus?

Christopher Baum,[1,2,*] Axel Schambach,[1] Jens Bohne,[1] and Melanie Galla[1]

IJC International Journal of Cancer

Fowlpox-based survivin vaccination for malignant mesothelioma therapy

Pietro Bertino[1], Maddalena Panigada[2], Elisa Soprana[2], Valentina Bianchi[2], Sabrina Bertilaccio[2], Francesca Sanvito[2], Aaron H. Rose[1], Haining Yang[3,4], Giovanni Gaudino[3], Peter R. Hoffmann[3], Antonio Siccardi[2] and Michele Carbone[3,4]

The Atlantic

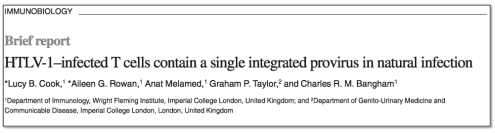
How to Cure a Bubble Boy

2. Virus integrations and disease associations[3,4,5]

Original Article

Incidence of Potentially Human Papillomavirus-Related Neoplasms in the United States, 1978 to 2007

George Kurdgelashvili, MD[1,2]; Graça M. Dores, MD, MPH[1,3]; Samer A. Srour, MD[1,2]; Anil K. Chaturvedi, PhD[3]; Mark M. Huycke, MD[1,2]; and Susan S. Devesa, PhD[3]



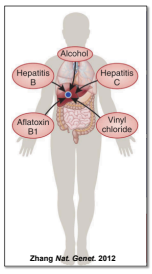
IMMUNOBIOLOGY

Brief report

HTLV-1–infected T cells contain a single integrated provirus in natural infection

*Lucy B. Cook,[1] *Aileen G. Rowan,[1] Anat Melamed,[1] Graham P. Taylor,[2] and Charles R. M. Bangham[1]

[1]Department of Immunology, Wright Fleming Institute, Imperial College London, United Kingdom; and [2]Department of Genito-Urinary Medicine and Communicable Disease, Imperial College London, London, United Kingdom



Zhang *Nat. Genet.* 2012

Methods: classical and improved approaches (Vy-PER**)

Classical

Map paired-ends to human genome.
Map unmapped reads to virus database (NCBI, etc.).
Filter low-complexity reads.

Problem: >1500 false positive chimeras per HiSeq lane

Virus candidate sequence	Probable virus candidate	More likely explanation
TTAGGGTTAGGGCTAGGGCTAGGGCTAGGGCTAGGGCTAGGGCTAGGGCT	Cyprinid herpesvirus 3	STR (human telomere)
CTCTCTCTCTCTCTCACACACACACACACACACACACACACACACACACAC	Ictalurid herpesvirus 1	STRs
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTA	Cuvid herpesvirus 2	mainly homopolymer
TATATATATATATATATATATATATATTTTTTTTTTTTTTTTTTTTTTTT	Cotesia congregata bracovirus	STR and homopolymer
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA	Human herpesvirus 6A	STR (human telomere)
TAACCCTAACCCTAACCCTAACCCTAGCCCTAACCCTAACCCTAACCCTA	Human herpesvirus 7	STR (human telomere)

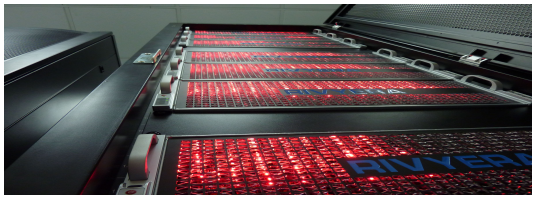
Improved (VirusSeq[6])

Problem fix:
Cluster virus integration sites

New problem:
Loss of sensitivity if
1) only fraction of cells affected
2) integration locus heterogeneity

Further improved (Vy-PER)

**acronym for Virus integration detection by Paired-End Reads



Multi-step elimination of false positives without impairing sensitivity.
Use of exact alignments (Smith-Waterman).
Speed-up of Smith-Waterman: FPGA (3000× single core speed)

Results

- False positive human/virus chimeras reduced more than 1000-fold
- Singleton human/virus chimeras easily detectable

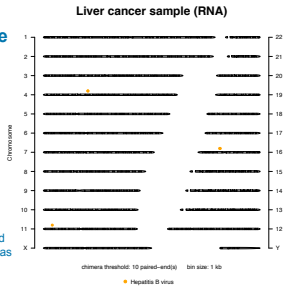
Vy-PER virus-positive example:

HBV virus integration

RNA-Seq

FASTQ files from VirusSeq web page[6]

This plot only shows integrations supported by 10 or more chimeras



Vy-PER virus-negative example:

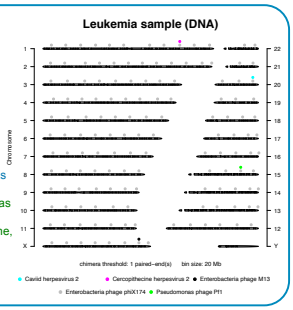
No virus integration

Whole genome 80× coverage

This plot shows integration sites down to single chimeras

12,000 false positive chimeras were reduced to 3: Caviid herpes, Cercopithecine, Pseudomonas

PhiX + M13: Illumina control library spike-in



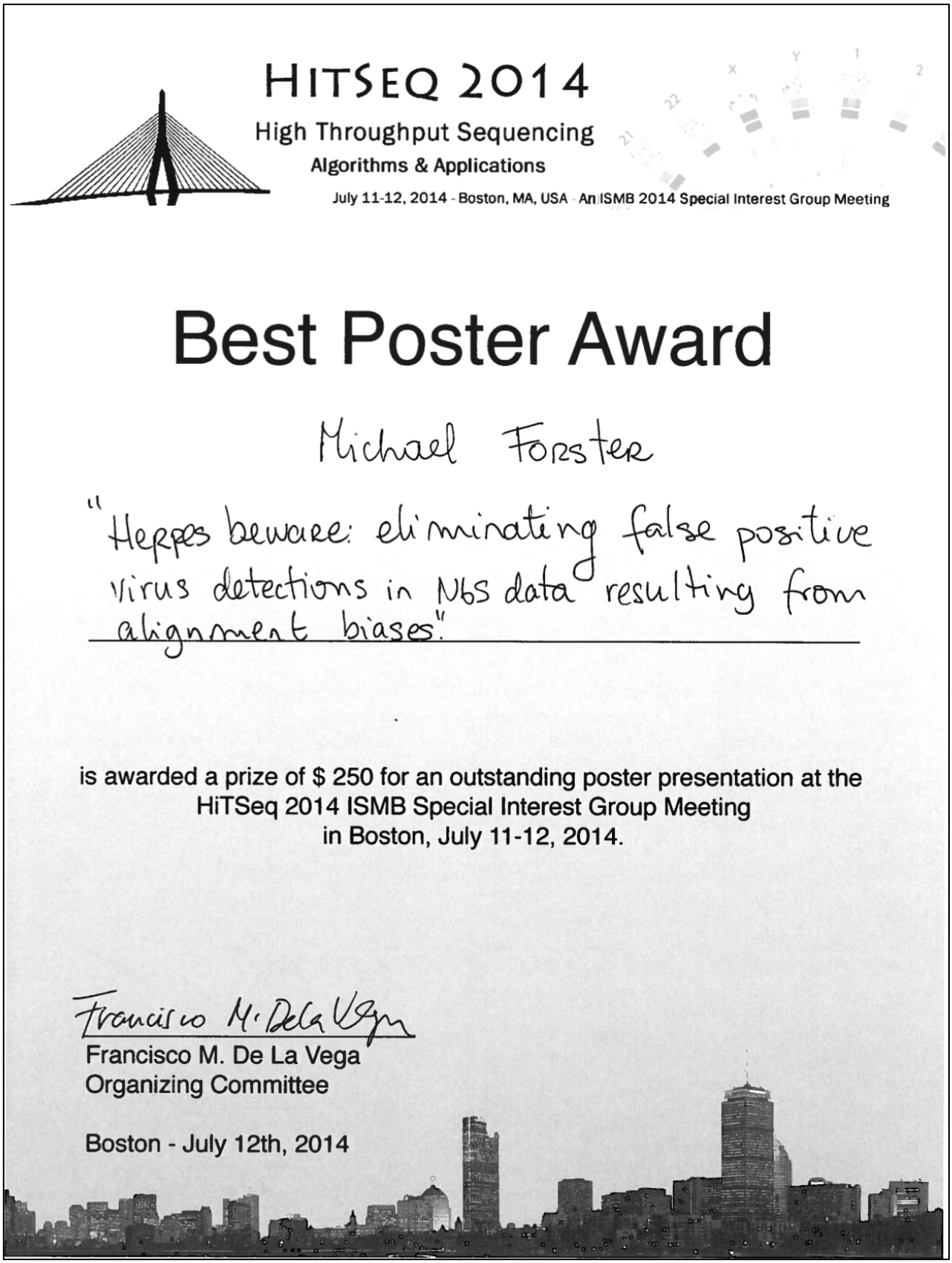
References

1) Baum *et al.* ***Mol. Ther.*** 2006. PMID: 16632409
2) Bertino *et al.* ***Int. J. Cancer*** 2013. PMID: 23335100
3) Kurdgelashvili *et al.* ***Cancer*** 2013. PMID: 23580435
4) Cook *et al.* ***Blood*** 2012. PMID: 22955925
5) Zhang. ***Nat. Genet.*** 2012. PMID: 2301122
6) Chen *et al.* ***Bioinformatics.*** 2013. PMID: 23162058

SPONSORED BY THE
Federal Ministry of Education and Research
Bundesamt für Strahlenschutz

Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Germany
m.forster@ikmb.uni-kiel.de, http://www.ikmb.uni-kiel.de

X. HiTSeq 2014 Poster Award for Vy-PER



HITSEQ 2014

High Throughput Sequencing

Algorithms & Applications

July 11-12, 2014 - Boston, MA, USA - An ISMB 2014 Special Interest Group Meeting

Best Poster Award

Michael Forster

"Herpes beware: eliminating false positive virus detections in NGS data resulting from alignment biases"

is awarded a prize of $ 250 for an outstanding poster presentation at the HiTSeq 2014 ISMB Special Interest Group Meeting in Boston, July 11-12, 2014.

Francisco M. De La Vega

Francisco M. De La Vega
Organizing Committee

Boston - July 12th, 2014