Decision Support and Intelligent Automation of Next-Generation Sequencing Workflows

September 30, 2019

Date: 2019-09-30

Version: 1.0

OTKA K 112915

Contents

1 Introduction

Next-generation sequencing revolutionized medicine, however, its clinical application is still hindered by multiple factors, such as measurement errors and interpretation of the results. The contributions of our project are as follows:

- Novel sequencing error models and error correction methods applicable for multiple single-molecule sequencing platforms, but tailored to Oxford Nanopore Technologies.
- Novel precision sequencing methods using ensemble variant calling.
- Real-time, highly parallelized automation of the sequencing pipeline to support high-throughput computing and adaptive sequencing.
- Automated workflow to support gene level and pathway level analysis.
- A novel adaptive sequencing method, which automatically, in real-time focuses on clinically relevant variants.

The results of the study were applied on real medical data [OTKA1, OTKA2, OTKA3, OTKA4] and were utilized on genetic data in cooperation with the Molecular Genetics Department of the National Oncological Institute and the Genomic Medicine and Rare Diseases Department of the Semmelweis University [OTKA5, OTKA6, OTKA7].

2 Novel sequencing error models, simulation environments and error correction methods

Electronic, nanopore based single molecule real-time DNA sequencing technology offers very long, albeit lower accuracy reads in sharp contrast to existing next-generation sequencing methods, which offer short, high-accuracy reads in abundance. We provided a systematic review of the error characteristics of this new sequencing platform, and demonstrated the most challenging aspects in the field of whole gene sequencing through the human HLA-DQA2 gene using long-range PCR products on multiplexed samples. We consider the limitations of these errors for the applications of this technology, also indicating expected improvements and expected thresholds with respect to these errors [OTKA8].

The initial capital cost of NGS starts above 50k USD, while the MinION device promises capital costs under 1k USD, reagent and consumables costs are similar for all sequencing technologies, but PCR-free library preparation and simplified library preparation protocols allow laboratories with simplified infrastructure – up to the point of using mobile field laboratories.

The ability to perform basecalling in real time on the first couple of hundred bases traversing a pore opens up possibilities for in situ enrichment. Based on the sequences present in the start of a read, the current across the membrane

can be reversed, ejecting the molecule from the pore [1]. A rejected molecule does not have a tether or a motor protein, so it cannot enter the pore again.

Direct RNA sequencing is possible by ligating a poly-T sequencing adapter and annealing the tether molecule to RNA strands. This offers both qualitative and quantitative gene expression analysis and transcriptomic profiling. Real-time species identification, along with PCR free quantitative and qualitative microbiomics are possible, opening up an array of possibilities in the field of metagenomics and pathogen identification, particularly with the possibility of portable sequencing [2,3]. PCR-free library preparation techniques offer quantitative and qualitative cancer profiling, immunome profiling, and mtDNA profiling. Rapid response to antibiotic resistant infections, with a turnaround time under 6 hours is possible with simplified library preparation routines, as well as accurate pathogen identification down to the exact genes responsible for antibiotic resistance.

Long reads allow for the phased genotyping on polyploid samples, eliminating the need for performing paired-end sequencing with large insert sizes to resolve the gametic phase of distant genetic polymorphisms [4, 5]. The extremely long reads produced by the MinION sequencer make it a suitable tool for generating scaffolds for de-novo genome sequencing. In a hybrid approach, NGS reads supplemented by long Oxford Nanopore Technologies (ONT) reads greatly reduce the number and increase the length of contigs a large genome can be assembled into. Shorter genomes or plasmids can be assembled exclusively from long reads [6]. Large scale chromosomal rearrangements and copy number variations can also be resolved, and are only limited by the attainable read lengths.

The error rate of the platform is higher than most mature next-generation sequencing platforms, with many of the deletions accumulating in stretches of identical bases (homopolymers, HPs). However, the mean time each 5-base long subsequence (k-mer) of the molecule spends inside of the pore (dwell time) can also be used to infer the length of the true sequence. We developed a method called NanoTimer [OTKA9], which estimates the homopolymer length from the dwell times. It relies on the redundancy of having multiple reads covering a reference sequence, and the depth of coverage determines its accuracy.

The investigation of the duration and level of signal currents (events) registered by MinKNOW is an active research topic, due to the possibility of improved variant call rate and accuracy. The amplification and barcoding of 2 x 12 5.8kb targets in the human MHC-II region were performed according to ONT protocols, using the Amplicon Sequencing Kit (SQK-006) and the PCR Barcoding Kit. This region was selected for its high degree of variability, as it contains more than 50 polymorphisms per sample, and for the examination of homopolymer indels. Current variant callers have difficulty with such ONT reads, so we have developed an application for visualizing current levels and translocation times on reference-aligned ONT reads. Due to the stochastic nature of single-stranded DNA ratcheting through each sequencing pore by the motor protein, we hypothesized that raw event durations and current levels can be used to improve the inference of the length of homopolymer stretches.

Specifically, in variant calling, the average normalized translocation times allow for the characterization of false positive homopolymer deletions, as well as providing supporting evidence for true positives [OTKA10].

Per base translocation times were calculated by normalizing each event duration by the translocation time of the entire read, to correct for any pore-specific translocation rate bias, as well as correcting for any flow cell level effect. The normalized per base translocation times are independent of the base psition in a read and the kmer context (including GC content). Correspondingly, event durations are not specified in the fast5 model files. The basecalling model does not allow steps between identical homopolymer states (no current changes are detected in homopolymer sequences, thus they do not produce events). However, this stable, independent average translocation rate allows us to infer homopolymer lengths. Additionally, the high variance of per-base translocation times average out when looking at longer homopolymer stretches. The increasing precision of this fundamental parameter for homopolymer length estimations is in sharp contrast with the increasing uncertainty for longer homopolymer lengths present in currently prevailing NGS technologies [7]. The coverage requirements to utilize translocation times depend on the level of accuracy required, higher coverage increases precision.

We examined the dwell time of each subsequence length, by collecting the elapsed time between the events corresponding to the first and last bases of a subsequence (Figure 2.). We found that the relative dwell time (RRTT) distribution of k length sequences can be closely approximated with an Erlang distribution with the shape = 2·k and rate = 0.5 parameters [OTKA9].

The current basecalling performed by ONT is limited to calling HP lengths of a maximum of 5. Considering the uniform distribution of substitution, insertion and non-HP deletions within 1D Nanopore reads, HP deletions present one of the last obstacles for the widespread usage of ONT in genomic DNA sequencing, as they have major (and often deleterious) effects on protein products when found in protein coding regions. Our methods significantly extends the scope of applicability: the accurate resolution of homopolymer stretches is only limited by the depth of coverage available for the target sequence.

The standard method of identifying the individual bases passing through each pore relies on a Hidden Markov Model (HMM), mapping raw current levels to individual bases in a nonlinear, multi-staged fashion. Recent advancements in artificial neural networks (ANN) and related natural language processing (NLP) techniques allow novel neural architectures that may improve the accuracy of the basecalling. We developed and examined a novel deep neural network based method to perform basecalling on raw current level measurements, as well as an efficient method of selecting and curating a training database from a set of real measurements [OTKA11].

We examined the most promising academic approaches, and compared them to the reference solution provided by the platform vendor, using the NA12878 whole genome shotgun sequencing dataset. Multiple types of systematic errors offer challenges to each individual solution, thus we proposed a framework to unify the strengths of each basecaller, and to aggregate their output in order to

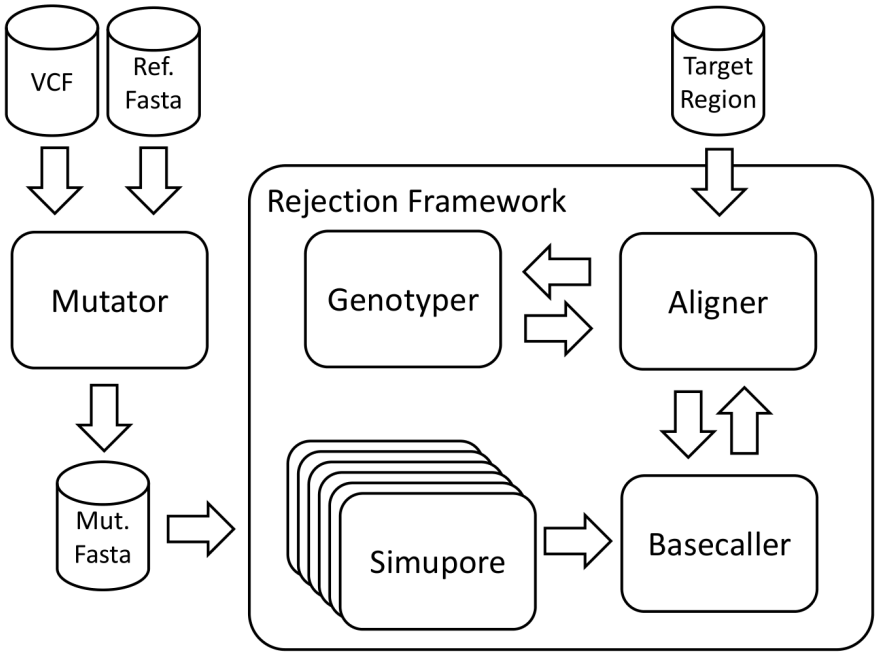

Figure 1: The complete adaptive sequencing and simulation pipeline.

increase their accuracy over any single solution [OTKA12].

One promising feature of the single-molecule real-time (SMRT) DNA sequencer developed by Oxford Nanopore Technologies is that it enables the termination of the sequencing of a read while traversing the pore (called read-until). We demonstrate that this enables efficient and optimal use of sequencing throughput and flow cell yield in the reinforcement learning framework. To explore the possibilities of such in-silico enrichment methods we have created a complete, end-to-end simulation framework, which spans from target regions to variant calling. To perform a detailed evaluation, we implemented a pipeline [OTKA13] consisting of a (1) raw current level simulator capable of creating signals callable by any nanopore basecaller, (2) a rapid, high-throughput HMM-based basecaller, with customizable early rejection parameters, (3) an in-memory aligner, and (4) a soft real-time incremental variant caller, as shown on Figure 1. This simulation environment is used in development and evaluation of the adaptive sequencing method [OTKA14].

The pipeline automates the tuning of read rejection parameters to achieve uniform error rates over target regions and variants. We performed synthetic evaluation using the NA12878 human sample.

3 Novel precision sequencing methods using ensemble variant calling

The low concordance between different variant calling methods still poses a challenge for the wide-spread application of next-generation sequencing in research and clinical practice. A wide range of variant annotations can be used for filtering call sets in order to improve the precision of the variant calls, but the choice of the appropriate filtering thresholds is not straightforward. Variant quality score recalibration provides an alternative solution to hard filtering, but it requires large-scale, genomic data.

We evaluated germline variant calling pipelines based on BWA and Bowtie 2 aligners in combination with GATK UnifiedGenotyper, GATK HaplotypeCaller, FreeBayes and SAMtools variant callers, using simulated and real benchmark sequencing data (NA12878 with Illumina Platinum Genomes). We argue that these pipelines are not merely discordant, but they extract complementary useful information. [OTKA15] 2

We created VariantMetaCaller to test the hypothesis that the automated fusion of measurement related information allows better performance than the recommended hard-filtering settings or recalibration and the fusion of the individual call sets without using annotations. VariantMetaCaller uses Support Vector Machines to combine multiple information sources generated by variant calling pipelines and estimates probabilities of variants.

This novel method has significantly higher sensitivity and precision than the individual variant callers in all target region sizes, ranging from a few hundred kilobases to whole exomes. We also demonstrated that VariantMetaCaller supports a quantitative, precision based filtering of variants under wider conditions. Specifically, the computed probabilities of the variants can be used to order the variants, and for a given threshold, probabilities can be used to estimate precision. Precision then can be directly translated to the number of true called variants, or equivalently, to the number of false calls, which allows finding problem-specific balance between sensitivity and precision.

The level of uncertainty in next-generation sequencing (NGS) measurements is still higher than what is required for routine clinical use, even for germline variants in targeted gene panels and exome sequencing [8]. The measurement process includes a complex computational variant calling pipeline, which contains many alternative elements with various parameters, heavily influencing the unique characteristics and performance of the whole procedure. Several studies showed that (1) currently there is no single best general individual variant calling method with both superior sensitivity and precision at all circumstances [8, 9], and (2) there are significant discrepancies between commonly used variant calling pipelines, even when applied to the same set of sequence data [8, 10–12] . An ad hoc approach is the fine-tuning of the pipeline for the actual measurement, which requires substantial expertise and time, also hindering standardization and benchmarking.

Generally, variant callers aim to be sensitive, call variants “aggressively” and

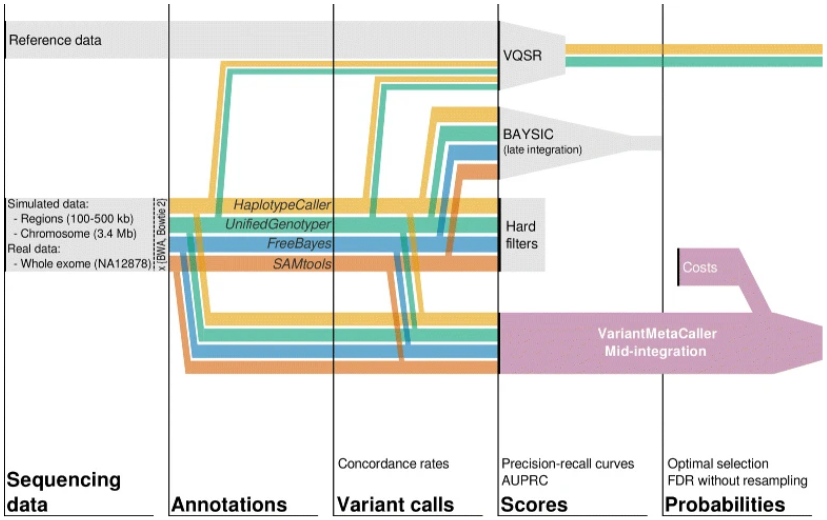

Figure 2: Earlier approaches, current study design including data sets and evaluations, and the conceptual overview of VariantMetaCaller. Study design: Simulated sequences of various target region sizes, and real sequence data covering the whole exome of NA12878 were aligned by BWA and Bowtie 2 to the human genome. Variants were called by GATK HaplotypeCaller, GATK UnifiedGenotyper, FreeBayes and SAMtools. Evaluation: Variant calling pipelines were compared by calculating concordance rates. Precision-recall curves were plotted and the area under the precision-recall curves was calculated for each method. Earlier approaches: Hard filters can be applied to filter variants by specifying annotation cutoffs. VQSR can be applied to recalibrate variant qualities based on gold standard reference data and variant annotations. BAYSIC combines the unfiltered variant calls by late integration. Overview of VariantMetaCaller: VariantMetaCaller (1) combines the unfiltered call sets by SVMs that use variant annotations as features and (2) estimates the probability of each variants being real. The probabilistic output of VQSR and VariantMetaCaller can be used to estimate FDR at each probability cutoff and to optimally select the filtered variants with respect to the cost function of the researchers. AUPRC = Area under the precision-recall curve, FDR = false discovery rate, NGS = Next-generation sequencing, SVM = Support Vector Machine

provide annotations to the user that can help distinguish true variants from false calls originating from sequencing, alignment or data processing artefacts. To further improve the sensitivity of the pipeline, one can use multiple variant calling methods, as it is a well-known fact that different callers produce different results [8, 10–14]. The rationale behind this practice is that the consequence of a false negative variant call (i.e. not discovering a true variant) is usually more serious than the consequence of a false positive (i.e. unreal variant claimed to be real), especially in clinical settings. The union of different call sets (called by different variant callers) could be taken for maximum sensitivity. However, this would result in higher false positive rate, i.e. a decrease in precision. Variants could, in principle, be validated experimentally using complementary measurement methods, but only at the cost of losing the high-throughput efficiency of NGS. Therefore, an application-specific balance between sensitivity and precision is needed.

A possible solution for selecting the appropriate list of variants is the use of hard filters. Variant callers produce a rich set of annotations that provide abundant information about mapping quality and various biases. For example, the evidence for a mutation is usually stronger at higher read depths [12]. A bias in the position of the variant in the read or a bias in the number of reads or base quality scores supporting an alternate allele may denote mapping problems and can be used to identify false variants. However, annotations have complex interrelationships [10,12], they depend on the experimental settings, and in most cases, are difficult to interpret [9]. It is often unclear what an adequate hard filter is; beyond general guidelines each specific study requires experimenting and empirical testing. Besides, most annotation classes depend on the actual read depth, and a filter setting which works for low coverage may not perform equally well for high coverage. The non-uniform coverage often seen in NGS studies [15] makes hard filtering a challenging task. Furthermore, it is also difficult to assess the resulting precision of the hard-filtered variant set.

An automated approach to improve precision of variant calling, applicable at a larger scale, is the use of variant quality score recalibration (VQSR) [16], which can be used to reclassify variant qualities. However, it requires a large amount of data: it can be used only for whole genomes or for at least 30 whole exomes according to GATK Best Practices. If a smaller region is sequenced, one can rely only on manual hard filters. Besides, VQSR uses gold standard, "error-free" variant sets as reference. In case of organisms for which these resources are unavailable, VQSR cannot be used in a straightforward manner.

In fact, automated recalibration can be also applied using abundant annotations of multiple pipelines instead of large amount of data: in this case the heterogeneous, intermediate annotations from multiple methods can be exploited for automated "recalibration". Indeed, this forms our central hypothesis that popular variant calling pipelines are not merely discordant, but the generated intermediate annotations contain complementary high-dimensional information, which can be combined into a better performing overall model. Our further hypothesis is that fusion of the intermediate annotation information allows the prediction of probabilities of variants in areas not accessible by current

approaches.

Based on these assumptions, we constructed VariantMetaCaller, which combines information from various variant callers using Support Vector Machines (SVM) (for an earlier related method, see [17]). Figure 2 shows the earlier approaches, the current study design including data sets and evaluations, and the conceptual overview of VariantMetaCaller. This novel method predicts the probability that a variant is a true genetic variant and not a sequencing artefact, which provides a principled solution for quantitative support for variant filtering. Specifically, probabilities can be used to order the variants, and for a given threshold, probabilities can be used to estimate precision. Precision then can be directly translated to the number of true called variants, or equivalently to the number of false calls, which allows finding problem-specific balance between sensitivity and precision, i.e. it allows a quantitative, precision-based filtering.

Automated fusion of multiple variant callers has been seen as a promising direction to exploit hidden information with more advanced statistical models. Until now, the arising problem of high-dimensionality and heterogeneity has remained unsolved in earlier fusion approaches, for example BAYSIC [18], used only the predicted calls, implementing late information fusion. To cope with high-dimensionality, a few SVM-based methods have already been introduced, such as the unpublished Ensemble method and the one used for the Exome Sequencing Project [19]. The method of the Exome Sequencing Project was not developed to utilize the combination of multiple variant-callers, and it determines annotation value cutoffs for defining negative training examples and gold standard data sets for defining positive training examples. VariantMetaCaller is conceptually similar to Ensemble, but the latter is limited to single-sample variant sets, and as to our knowledge, does not produce a quantitative score and therefore cannot be used to balance between sensitivity and precision.

Copy number variations (CNV's) are considered a subclass of structural variants in which regions of the genome have varying number of repeats, and the number of these repeats can differ among individuals of a species. The most common CNV's are duplications and deletions of copies of entire coding regions or genes. The heterozygous loss of a copy of a gene is also considered a CNV. Chip-based genome-wide association studies have shown good results [20] in detecting CNV's, where deletions can be inferred from loss of heterozygosity, and additional copies detected from increased heterozigosity of contiguous regions. Whole-exome sequencing (WES) offers new methods of detecting CNV's, where the main approaches involve:

- Identifying areas of outlying coverage along target regions, though some methods of exome capture (e.g. array capture) naturally reduce the spread of individual target coverage.
- Identifying inequal allelic fractions on polymorphisms, where excess homozigosity and imbalanced allelic fractions respectively indicate deletions and copy gains.
- Identification via paired-ends, where discordance among the insert dis-

tances of read pairs aid in the identification of CNV's

- Identification by assembly, via the remapping of soft-clipped reads, which indicate larger breaks from the reference sequence, and are commonly used to identify large structural variations [21].

Detection of copy number variation can also follow different approaches based on the type of dataset used:

- Single sample CNV detection, where the tools call CNVs on a per-sample basis, without the use of data from other samples.
- Multiple sample parallel CNV detection, where tools calls CNVs jointly, using the coverage or allelic fraction data from all samples simultaneously.
- Matched sample CNV detection, most commonly between a normal tissue and tumor sample, where the goal is to identify changes from the normal tissue present in the tumor sample.

A multitude of copy number variation (CNV) callers have been published in the last few of years, yet they show significant disparities in both their sensitivity, and their specificity, as well as their methods and data requirements, as shown on Figure 3.

We tested 4 whole-exome CNV callers, and investigated their performance on 40 whole exome sequencing samples, and created a framework for the unification of their results [OTKA16]. Of the different tools, XHMM [22] found a total of 946 CNVs, CoNIFER [23] indicated 645, while CONTRA [24] noted 32481, and finally CNVnator [25] with 48330 calls. This indicates that XHMM and CoNIFER are highly specific, as their counts are slightly under the expected number of CNVs in 40 samples, while both CONTRA and CNVnator are highly sensitive, with counts far exceeding those that should naturally occur in the samples. The methods show highly discordant calls, despite their being non-empty subset of 73 calls verified by all 4 tools, as show in Figure 4. This discordance is significantly higher than that demonstrated by short variant callers [OTKA15], owing to the enourmous differences between individual call methods.

4 High-throughput automation of NGS sequencing pipelines

In order to support both high-throughput computing and adaptive sequencing, we created an NGS data processing pipeline [OTKA17] which allows the incremental addition of newly sequenced samples while preserving the results of earlier computational steps, while still achieving the ability to jointly call genotypes in a large sample set. Adhering to the best practices of NGS data analysis [16] requires the use of over a dozen different tools sequentially in the data processing pipeline. The computational cost of these tools is highly variable

Name	Release date	Authors	Datatype	Matched?	WES /WGS	Method	State	Distribution
XHMM	2014	Fromer et. al.	Read depths	Normal	WES/WGS	Hidden Markov Model	Maintained	R package
CoNIFER	2012	Krumm et. al.	Read depths	Normal	WES/WGS	Singular Value Decomposition	Maintained	Python package
CNVnator	2011	Abyzov et. al.	Read depths	Normal	WES	Mean-shift	Maintained	C package
ADTEx	2014	Amarasinghe et. al.	B-allele frequencies	Matched normal-tumor	WES	Hidden Markov Model	Maintained	Python/R package
CONTRA	2012	Li et. al.	Read depths	Both	WES	Log-ratios	Maintained	Python package
cn.MOPS	2012	Klambauer et. al.	Read depths	Normal	WES	Mixture of Poissons	Documentation missing	R package
ExomeCNV	2011	Sathirapongsasuti et. al.	Read depths, B-allele frequency	Matched normal-tumor	WES	Log-ratios	Unmaintained	R package
VarScan 2	2012	Koboldt et. al.	Read depths, B allele frequency	Matched normal-tumor	WES	Heuristic and statistical test based classification	Maintained	Java Executable
ExomeDepth	2012	Plagnol et. al.	Read depths	Both	WES	Optimized Classifier	Maintained	R package
EXCAVATOR	2017	Magi et. al.	Read depths	Both	WES/WGS	Read depth log2 ratio	Maintained	Bash, R, Perl scripts

Figure 3: Overview of CNV calling methods applicable to whole exome sequencing datasets (WES)

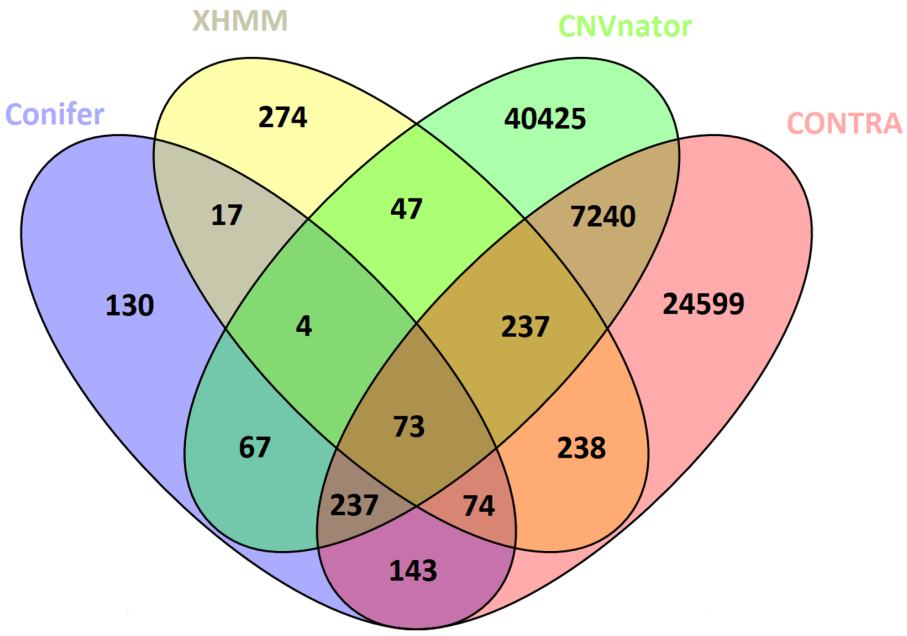

Figure 4: Venn diagram displaying the number of CNVs detected by each tool and the sizes of their overlapping sets

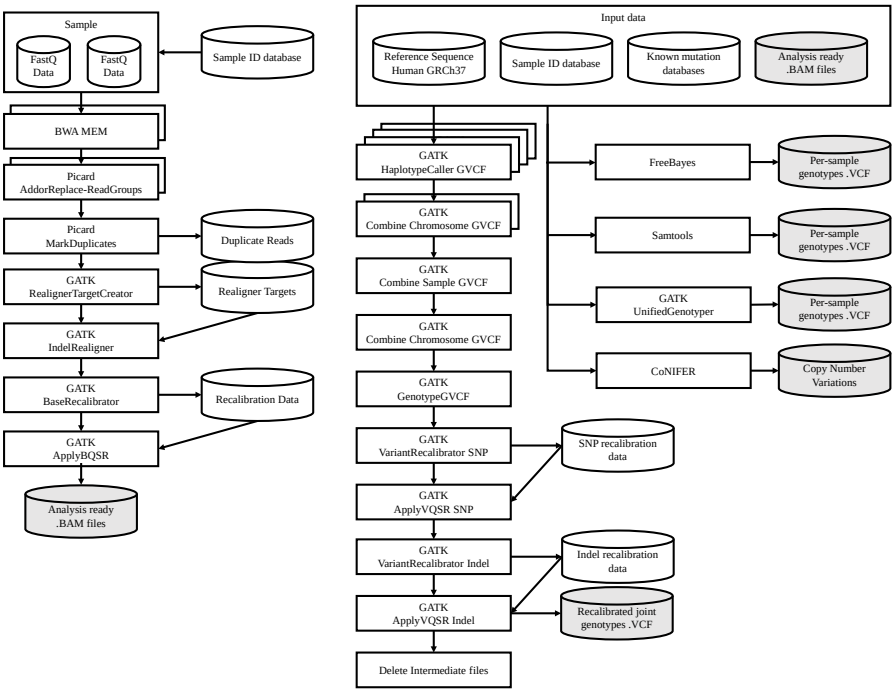

Figure 5: The automated next-generation sequencing pipeline, using optimal resource allocation to allow optimal throughput and incremental genotyping of large whole-genome sequencing datasets. Starting from raw FASTQ data, data preparation is done via PicardTools [26] and GATK [27], with final genotype calls being produced by FreeBayes [28], Samtools [29], UnifiedGenotyper and HaplotypeCaller [27] with variant quality score recalibration (VQSR).

both with respect to CPU time, memory usage, disk usage and parallelizability. We created a system that optimally distributes resources in the computational graph to provide the highest possible processing throughput in a homogenous computing cluster 5.

The system allows for the thorough exploration of the parameter space for each tool, and manages toolchain errors gracefully. We investigated the application of resampling techniques to generate novel, robust quality scores for NGS variant calling [OTKA18].

5 Automated workflow to support gene level and pathway level analysis

The tertiary analysis in the NGS workflow aims the interpretation of the of the variants, which requires large-scale data and knowledge fusion from mulit-

ple levels and domains. Integration of cross-domain information has been targeted at different levels: at the level of data, such as in the joint statistical analysis of cross-domain omic datasets [30], at the level of knowledge, such as in the pharmaceutical integration approaches using semantic web technologies [31–33], and even at the level of computational services, such as in the scientific workflows [34, 35]. However, significant part of scientific knowledge is uncertain, weakly significant, poorly represented and remains inaccessible for cross-domain integration, although the importance of the analysis and interpretation of such weak signs have already been recognized in many standalone high-dimensional omic domains. This is illustrated by data fusion in molecular similarity [36], kernel-based data and knowledge fusion [37], cross-species gene prioritization [38], Bayesian fusion [39] and network boosted analysis of genome-wide polymorphism data [40].

Semantic technologies, relying heavily on the Resource Description Framework (RDF), provide an unprecedented basis for cross-domain data and knowledge fusion, as demonstrated by the emergence of large-scale, unified knowledge space in life sciences (the Life Sciences Linked Open Data Space, LSLODS, see e.g. BIO2RDF [41], CHEM2BIO2RDF [42], Open PHACTS [32], integrated WikiPathways [43], biochem4j [44], DisGeNET-RDF [45, 46]). However, there are serious limitations concerning its computational complexity of inference [47] and practical IT accessibility [48], its inaccessibility for non-technical users [32, 49, 50]. Furthermore, most importantly, its ability to cope with uncertain facts, evidences, and inference is still an open challenge (for representing uncertain scientific knowledge, see e.g. HELO [51]; for combination of uncertain evidences, see e.g. [39, 52–54]).

To tackle these challenges, we developed a methodology utilizing the intermediate, quantitative knowledge level of structured similarities and created a corresponding system to demonstrate its advantages, the Quantitative Semantic Fusion (QSF) system (Fig. 6) [OTKA19]. This approach is related to multiple earlier approaches in fusion, such as (1) Linked Open Data (LOD) cubes to support computationally efficient SPARQL queries [55], (2) knowledge graphs [56], (3) probabilistic logic, Markov logic for semantic web integration inference and approximation of inference in large-scale probabilistic graphical models [57], and (4) relational generalization of kernel-based fusion [37, 58].

The main elements of the proposed framework are as follows.

- *Structure*: Types of entities and their structural dependencies (entities are represented with discrete values, e.g. genes, drugs, diseases).
- *Parameters*: Quantitative pairwise relations, e.g. bioactivities of drug-target interactions, sequence similarity between targets, orthology between genes in different species, genetic variant-disease associations.
- *Inference rules*: Canonical methods for the combination of similarities and relevances, as propagation of evidences.
- *Evidences*: Quantitative, vectorial representation of relevances of entities

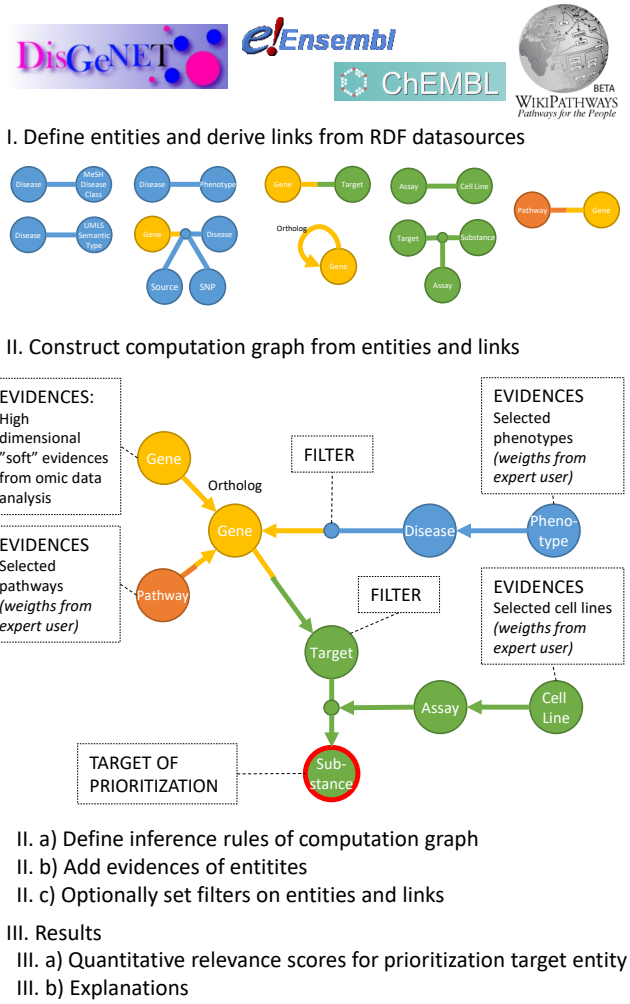

Figure 6: **Quantitative Semantic Fusion (QSF) System** (I.) The QSF System incorporates distinct annotated semantic types (i.e. entities) and their quantitative pairwise relations (i.e. links) by integrating different data sources from the Linked Open Data world. Predefined entities and links from DisGeNET [45], Ensembl [59], ChEMBL [60] and WikiPathways [43] are shown in the top. Together entities and links form the *structure* and *parameters* of the QSF System. (II.) The user can freely construct so-called computation graphs using the available entities and links and can select any entity as the target of the prioritization. An example computation graph is shown in the middle. Then, the user defines the (II.a.) *inference rules*, sets (II.b.) *evidences* of possibly multiple entities and (II.c.) optionally sets filters on specific entities and links. The main results of the prioritization are (III.a) the quantitative relevance scores for the target entity and (III.b.) the most dominant *explanations* of the prioritization results.

in a certain query, such as summary statistics from earlier omic data analysis (i.e. data analytic query) and semantical controls and weights for the inference process.

- *Results*: Quantitative relevance scores are inferred, ranked lists (prioritizations) are constructed, e.g. for subsequent enrichment analysis.
- *Explanations*: Dominant chains of reasoning are retrieved and visualized from the inference process in Cytoscape.

We applied this metholodogy and system in multiple domains [61].

As a more constrained approach, we also investigated the use of network propagation methods [62]. We introduced a full-fledged network-based workflow for the analysis of genetic variants, both covering polymorphisms and rare variants using specific gene aggregation tools. We overviewed critical steps, possible solutions, and publicly available resources for this workflow; especially effect of (1) gene definitions and aggregation methods, (2) context-specific molecular networks, and (3) network propagation methods [OTKA20]. The workflow also supports the analysis of multiple traits and diseases, especially from multimorbidity networks [OTKA21, OTKA22, OTKA23, OTKA24]. The developed workflow is shown in Fig. 7.

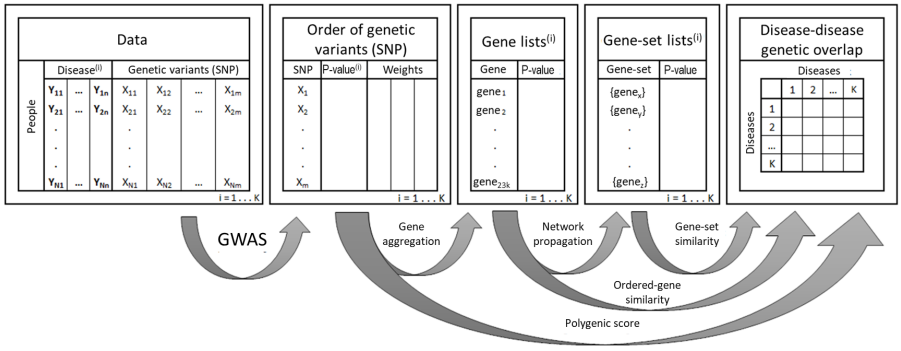

Figure 7: Key steps of the network-based variant post-processing workflow: the variant, gene and gene-set based multimorbidity analysis.

6 A novel adaptive sequencing method

The MinION single-molecule DNA sequencer developed by Oxford Nanopore Technologies (ONT) offers many, diverse novel functionalities compared to mainstream second-generation sequencing technologies [63–68]. Focusing on the scientific aspects beyond its low investment, infrastructural and maintainance requirements, its input/sample preparation is greatly simplified, and can be

PCR-free. Nanopore sequencing produces ultra-long reads enabling novel applications, such as detecting complex structural variants, phasing, and supporting genome assembly [4, 69–73]. Its sequencing technology is based on the measurement of ionic current blockage through the nanopores as single-molecule fragments traverse them. This inference process, called basecalling, is still faces challenges from wide-range of technology specific errors [5, 74–76] [OTKA8]. Besides of the official protocols and methods, this flexible and open sequencing process gave rise to multiple experiments [77, 78], novel computational methods (see e.g., [79–83]), and led to the development of multitudes of simulation environments [84–87]. Despite of its limitations with total sequencing capacity advertised as 4-8 GBs and an accuracy (read identity) only recently surpassing 90% [88], this technology is very promising for many applications/niches, such as surveying repertoire of bacterial communities [89,90], the microbiome [73,91], the immune repertoire [92,93] or mitochondrial landscape [94,95]. Furthermore, the technology is extendable towards mapping DNA methylation [96], RNA [97], and proteins [98].

A notable, but currently underutilized feature of this technology is its real-time nature with the read rejection option, i.e. that it enables the termination of the sequencing of a read while traversing the pore (called read-until, and the measured part of the sequence will be referred to as the *prefix* in this paper) [1]. There are a total of 2048 nanopores, and 512 A/D converters manufactured on each MinION flow cell, with 4 pores multiplexed on to each A/D converter. Since the and the manufacturing of each pore is not perfect, an initial quality control step examines each pore, and selects 4 sets of pores, with the first and each subsequent set selected to maximize the number of usable pores. As sequencing progresses, the pores can become blocked or otherwise unusable. At fixed intervals, defined by default at 12 hours, the multiplexers select the next set of sequencing pores. The A/D converters sample the current values at 4kHz with 10bit ADC resultion. The useful signal range is approximately 8 bits [83]. The system can theoretically provide nearly real-time measurements, but current API limitations introduce an approximately 700 msec delay on the availability of the raw data [1]. This delay corresponds to approximately 300-400 bases for a processing back-end computer, but in principle the sequencing data can be processed in real-time and sequencing can be controlled in nearly real-time (the total delay from the measurement of a base till the rejection of this read is 1 sec (470 bases). This delay is further increased in practice by the processing time of sequencing data, which process may include measurement related calculations, such as basecalling and read alignment, biomedical calculations, such as variant calling and estimation of the functional effects of variants in a given sample, and statistical calculations, such as the expected value of a given read in an adaptive experimental design. Additionally, basecalling and read alignment also pose further requirements on stopping and read length, but in practice read length can be varied from 200 bases up to the capacity of ONT (N50 ¿ 100kb) [72]. Note that the starting position of reads approximately follows a uniform distribution independent of the genome fragmentation technique (for deviations, see e.g., [99]).

Utilizing this feature of the ONT, we developed a novel *adaptive sequencing with rejection* method, assuming that reads arrive randomly according to a distribution over starting position and length, and virtually at any base the sequencing of the read can be stopped, i.e. the rest of the read is rejected.

A central question in this approach is the question of stopping: should we stop at sequencing the read (by ejecting it) or continue? This question can be formalized in multiple theoretical approaches, such as in adaptive study design, sequential decision, multi-armed bandit problems, online learning, active learning, budgeted learning, reinforcement learning. Within this framework, three families of methods can be distinguished based on the supportive methods applied in real-time.

1. **Real-time quality control**: using only standard basecallers on the already read prefix, if quality parameters drop below certain thresholds (e.g. on low complexity regions, or on sequence specific basecalling errors) then the read can be rejected.
2. ***In silico* targeted sequencing with prespecified coverage**: real-time alignment of reads allows the rejection of off-target reads or reads in regions with fulfilled coverage specification.
3. **Precision sequencing with uniform errors**: real-time variant calling allows focusing sequencing on problematic or highly relevant variants/regions.

The application areas of real-time, adaptive sequencing are broad and open-ended, as it offers both general performance improvements and leaves the real-time back-end calculations over raw current measurements, reads or variants open. Better performance may mean more sequencing capacity on regions of interest or less errors, which are critical in many domains, such as in metagenomics, targeted sequencing in cancer research, ultra-deep sequencing for surveying repertoires. Complex loss functions based on the real-time post-processing of the called sequence and discovered variants may arise in personalized medicine, such as in variant effect prediction, haplotyping, phasing or assembling a critical genomial region or genome, and the phylogenic analysis of bacterial communities or tumor cells.

Currently, we are applying this method in targeted sequencing experiments via in silico target enrichment [OTKA13, OTKA14].

7 OTKA References

[OTKA1] Ákos Jobbágy, Judit Schultheisz, Márk Horváth, and Hanna Réfy Vraskóné. Development of an effective therapy and objective assessment for children with birth injuries. *International Journal of Rehabilitation Research*, 39(4):354–360, December 2016.

[OTKA2] Ákos Jobbágy and P. Nagy. The effect of occlusion with the cuff. In Hannu Eskola, Outi Väisänen, Jari Viik, and Jari Hyttinen, ed-

itors, *EMBEC & NBC 2017*, pages 9–12, Singapore, 2018. Springer Singapore.

[OTKA3] Ákos Jobbágy, Miklós Majnár, Lilla K. Tóth, and Péter Nagy. Hrv-based stress level assessment using very short recordings. *Periodica Polytechnica Electrical Engineering and Computer Science*, 61(3):238–245, 2017.

[OTKA4] Ákos Jobbágy, Judit Schultheisz, Márk Horváth, Piroska Bacsó, Péter Csuhaj, and Hanna Réfy Vraskó. Objective assessment of children with birth injuries. In *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, pages 571–575. Springer, 2016.

[OTKA5] Noémi Ágnes Varga, Klára Pentelényi, Péter Balicza, András Gézsi, Viktória Reményi, Vivien Hársfalvi, Renáta Bencsik, Anett Illés, Csilla Prekop, and Mária Judit Molnár. Mitochondrial dysfunction and autism: comprehensive genetic analyses of children with autism and mtdna deletion. *Behavioral and Brain Functions*, 14(1):4, 2018.

[OTKA6] Péter Balicza, Noémi Ágnes Varga, Bence Bolgár, Klára Pentelényi, Renáta Bencsik, Anikó Gál, András Gézsi, Csilla Prekop, Viktor Molnár, and Mária Judit Molnár. Comprehensive analysis of rare variants of 101 autism-linked genes in a hungarian cohort of autism spectrum disorder patients. *Frontiers in genetics*, 10:434, 2019.

[OTKA7] Péter Balicza, Zoltán Grosz, Viktor Molnár, Anett Illés, Dora Csabán, Andras Gézsi, Lívia Dézsi, Dénes Zádori, László Vécsei, and Mária Judit Molnár. Nkx2-1 new mutation associated with myoclonus, dystonia, pituitary dysfunction and empty sella. *Frontiers in genetics*, 9:335, 2018.

[OTKA8] Péter Sárközy, Viktor Molnár, Dóra Fogl, Csaba Szalai, and Péter Antal. Beyond homopolymer errors: a systematic investigation of nanopore-based dna sequencing characteristics using hla-dqa2. *Periodica Polytechnica Electrical Engineering and Computer Science*, 61(3):231–237, 2017.

[OTKA9] Peter Sarkozy, Ákos Jobbágy, and Peter Antal. Calling homopolymer stretches from raw nanopore reads by analyzing k-mer dwell times. In Hannu Eskola, Outi Väisänen, Jari Viik, and Jari Hyttinen, editors, *EMBEC & NBC 2017*, pages 241–244, Singapore, 2018. Springer Singapore.

[OTKA10] Peter Sarkozy, Viktor Molnár, Dóra Fogl, and Péter Antal. Time and current domain exploration of homopolymer lengths in ont reads. 5 2016. London Calling: Oxford Nanopore Technologies Annual Meeting ; Conference date: 26-05-2016 Through 27-05-2016.

[OTKA11] Máte Borkó, Bence Bolgár, and Peter Sarkozy. Basecalling raw nanopore dna sequencing reads using neural networks. *Proceedings of the 25th Minisymposium of BME MINISY@DMIS2018*, 25(1), 2018.

[OTKA12] Erik Jagyugya and Peter Sarkozy. Comparison of nanopore dna sequencing basecallers on whole human data. *Proceedings of the 25th Minisymposium of BME MINISY@DMIS2018*, 25(1), 2018.

[OTKA13] Peter Sarkozy, András Antos, and Péter Antal. Online variant calling using read rejection: evaluation in a comprehensive raw current based simulation framework. 5 2019. London Calling: Oxford Nanopore Technologies Annual Meeting ; Conference date: 22-05-2019 Through 24-05-2019.

[OTKA14] Peter Sarkozy, András Antos, Zsolt Bihary, and Peter Antal. adaseq: theoretical bounds, methods, and simulation environment for adaptive sequencing using read rejection in nanopore. *In Preparation*, 2019.

[OTKA15] András Gézsi, Bence Bolgár, Péter Marx, Peter Sarkozy, Csaba Szalai, and Péter Antal. Variantmetacaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. *BMC Genomics*, 16(1):875, 2015.

[OTKA16] Peter Sarkozy, Ákos Jobbágy, and Peter Antal. Comparison of nanopore dna sequencing basecallers on whole human data. *In Preparation*, 2019.

[OTKA17] Peter Sarkozy, Ákos Jobbágy, and Peter Antal. Újgenerációs dns szekvenálási adatok automatizált feldolgozása homogén számítási rendszerben. *In Preparation*, 2019.

[OTKA18] P. Sarkozy, Á Jobbágy, and P. Antal. Bootstrap-based quality scores for ngs variant calling. In Ákos Jobbágy, editor, *First European Biomedical Engineering Conference for Young Investigators*, pages 44–47, Singapore, 2015. Springer Singapore.

[OTKA19] Andras Gezsi, Bence Bruncsics, Gabor Guta, and Peter Antal. Constructing a quantitative fusion layer over the semantic level for scalable inference. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 41–53. Springer, 2018.

[OTKA20] Bence Bruncsics and Peter Antal. A multi-trait evaluation of network propagation for gwas results. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–6. IEEE, 2019.

[OTKA21] Péter Antal, József Reiter, and Péter Mátyus. Biomarkerek hálózatától a klinikai döntéstámogatásig. *Orvosi Hetilap*, 156(51):2077–2081, December 2015.

[OTKA22] Péter Antal and Csaba Szalai. Hálózatok vizsgálata betegségekben (networks analysis in diseases). *Természet Világa*, 146(1):68–72, December 2015.

[OTKA23] P. Marx and P. Antal. Decomposition of shared latent factors using bayesian multi-morbidity dependency maps. In Ákos Jobbágy, editor, *First European Biomedical Engineering Conference for Young Investigators*, pages 40–43, Singapore, 2015. Springer Singapore.

[OTKA24] Peter Marx, Peter Antal, Bence Bolgar, Gyorgy Bagdy, Bill Deakin, and Gabriella Juhasz. Comorbidities in the diseasome are more apparent than real: What bayesian filtering reveals about the comorbidities of depression. *PLoS computational biology*, 13(6):e1005487, 2017.

8 References

[1] Matthew Loose, Sunir Malla, and Michael Stout. Real-time selective sequencing using nanopore technology. *Nature methods*, 13(9):751, 2016.

[2] Sissel Juul, Fernando Izquierdo, Adam Hurst, Xiaoguang Dai, Amber Wright, Eugene Kulesha, Roger Pettett, and Daniel J Turner. Whats in my pot, real-time species identification on the minion. November 2015.

[3] D. Nichols, N. Cahoon, E. M. Trakhtenberg, L. Pham, A. Mehta, A. Belanger, T. Kanigan, K. Lewis, and S. S. Epstein. Use of ichip for high-throughput in situ cultivation of uncultivable microbial species. *Applied and Environmental Microbiology*, 76(8):2445–2450, 2010.

[4] Ron Ammar, Tara A Paton, Dax Torti, Adam Shlien, and Gary D Bader. Long read nanopore sequencing for detection of hla and cyp2d6 variants and haplotypes. *F1000Research*, 4, 2015.

[5] Thomas W Laver, Richard C Caswell, Karen A Moore, Jeremie Poschmann, Matthew B Johnson, Martina M Owens, Sian Ellard, Konrad H Paszkiewicz, and Michael N Weedon. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Scientific reports*, 6:21746, 2016.

[6] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, June 2015.

[7] Peter Sarkozy, Márton Enyedi, and Péter Antal. Flow index based characterization of next generation sequencing errors. 03 2014.

[8] Jason O'Rawe, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson, W. Evan Johnson, Zhi

Wei, Kai Wang, and Gholson J. Lyon. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine*, 5(3):28–28, Mar 2013. 23537139[pmid].

[9] Mi-Hyun Park, Hwanseok Rhee, Jung Hoon Park, Hae-Mi Woo, Byung-Ok Choi, Bo-Young Kim, Ki Wha Chung, Yoo-Bok Cho, Hyung Jin Kim, Ji-Won Jung, et al. Comprehensive analysis to improve the validation rate for single nucleotide variants detected by next-generation sequencing. *PloS one*, 9(1):e86664, 2014.

[10] Xiaoqing Yu and Shuying Sun. Comparing a few snp calling algorithms using low-coverage sequencing data. *BMC bioinformatics*, 14(1):274, 2013.

[11] Xiangtao Liu, Shizhong Han, Zuoheng Wang, Joel Gelernter, and Bao-Zhu Yang. Variant callers for next-generation sequencing data: a comparison study. *PloS one*, 8(9):e75619, 2013.

[12] Mehdi Pirooznia, Melissa Kramer, Jennifer Parla, Fernando S Goes, James B Potash, W Richard McCombie, and Peter P Zandi. Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, 8(1):14, 2014.

[13] Joseph A Neuman, Ofer Isakov, and Noam Shomron. Analysis of insertion–deletion from deep-sequencing data: software evaluation for optimal detection. *Briefings in Bioinformatics*, 14(1):46–55, 2012.

[14] Anthony Youzhi Cheng, Yik-Ying Teo, and Rick Twee-Hee Ong. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*, 30(12):1707–1713, 2014.

[15] Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC genomics*, 13(1):341, 2012.

[16] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Chris Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1110):11.10.1–11.10.33, 2013. 25431634[pmid].

[17] Brendan D O'Fallon, Whitney Wooderchak-Donahue, and David K Crockett. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics*, 29(11):1361–1366, 2013.

[18] Brandi L Cantarel, Daniel Weaver, Nathan McNeill, Jianhua Zhang, Aaron J Mackey, and Justin Reese. Baysic: a bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC bioinformatics*, 15(1):104, 2014.

[19] Jacob A Tennessen, Abigail W Bigham, Timothy D O'Connor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *science*, 337(6090):64–69, 2012.

[20] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F. A. Grant, Hakon Hakonarson, and Maja Bucan. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome research*, 17(11):1665–1674, Nov 2007. 17921354[pmid].

[21] Ramesh Rajaby and Wing-Kin Sung. Transurveyor: an improved database-free algorithm for finding non-reference transpositions in high-throughput sequencing data. *Nucleic acids research*, 46(20):e122–e122, Nov 2018. 30137425[pmid].

[22] Menachem Fromer and Shaun M. Purcell. Using xhmm software to detect copy number variation in whole-exome sequencing data. *Current protocols in human genetics*, 81:7.23.1–7.23.21, Apr 2014. 24763994[pmid].

[23] Niklas Krumm, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, NHLBI Exome Sequencing Project, Aaron R. Quinlan, Deborah A. Nickerson, and Evan E. Eichler. Copy number variation detection and genotyping from exome sequence data. *Genome research*, 22(8):1525–1532, Aug 2012. 22585873[pmid].

[24] Jason Li, Richard Lupat, Kaushalya C. Amarasinghe, Ella R. Thompson, Maria A. Doyle, Georgina L. Ryland, Richard W. Tothill, Saman K. Halgamuge, Ian G. Campbell, and Kylie L. Gorringe. Contra: copy number analysis for targeted resequencing. *Bioinformatics (Oxford, England)*, 28(10):1307–1313, May 2012. 22474122[pmid].

[25] Alexej Abyzov, Alexander E. Urban, Michael Snyder, and Mark Gerstein. Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome research*, 21(6):974–984, Jun 2011. 21324876[pmid].

[26] Broad Institute. Picard tools. http://broadinstitute.github.io/picard/, 2018. Accessed: 2018-06-02; version 2.18.9,.

[27] Poplin R Garimella K Maguire J Hartl C Philippakis A del Angel G Rivas MA Hanna M McKenna A Fennell T Kernytsky A Sivachenko A Cibulskis K Gabriel S Altshuler D Daly M. DePristo M, Banks E. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature Genetics*, 43:491–498, 2011.

[28] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing, 2012.

[29] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, Aug 2009. 19505943[pmid].

[30] Zhihong Zhu et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics*, 48(5):481–487, 2016.

[31] Huajun Chen, Li Ding, Zhaohui Wu, Tong Yu, Lavanya Dhanapalan, and Jake Y Chen. Semantic web for integrated network analysis in biomedicine. *Briefings in bioinformatics*, 10(2):177–192, 2009.

[32] Antony J. Williams, Lee Harland, Paul Groth, Stephen Pettifer, Christine Chichester, Egon L. Willighagen, Chris T. Evelo, Niklas Blomberg, Gerhard Ecker, Carole Goble, and Barend Mons. Open PHACTS: Semantic interoperability for drug discovery. *Drug Discovery Today*, 17(21-22):1188–1198, 2012.

[33] Bin Chen, Huijun Wang, Ying Ding, and David Wild. Semantic breakthrough in drug discovery. *Synthesis Lectures on the Semantic Web*, 4(2):1–142, 2014.

[34] Robert Stevens, Patricia Baker, Sean Bechhofer, Gary Ng, Alex Jacoby, Norman W Paton, Carole A Goble, and Andy Brass. Tambis: transparent access to multiple bioinformatics information sources. *Bioinformatics*, 16(2):184–186, 2000.

[35] Md Rezaul Karim, Audrey Michel, Achille Zappa, Pavel Baranov, Ratnesh Sahay, and Dietrich Rebholz-Schuhmann. Improving data workflow systems with cloud services and use of open data for bioinformatics research. *Briefings in Bioinformatics*, page bbx039, 2017.

[36] Claire MR Ginn, Peter Willett, and John Bradshaw. Combination of molecular similarity measures using data fusion. In *Virtual Screening: An Alternative or Complement to High Throughput Screening?*, pages 1–16. Springer, 2000.

[37] Gert RG Lanckriet, Tijl De Bie, Nello Cristianini, Michael I Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[38] Léon-Charles Tranchevent, Amin Ardeshirdavani, Sarah ElShal, Daniel Alcaide, Jan Aerts, Didier Auboeuf, and Yves Moreau. Candidate gene prioritization with endeavour. *Nucleic acids research*, 44(W1):W117–W121, 2016.

[39] Michael A Province and Ingrid B Borecki. Gathering the gold dust: methods for assessing the aggregate impact of small effect genes in genomic scans. In *Pacific Symposium on Biocomputing*, volume 13, pages 190–200, 2008.

[40] Priyanka Nakka, Benjamin J Raphael, and Sohini Ramachandran. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics*, 204(2):783–798, 2016.

[41] Alison Callahan, José Cruz-Toledo, Peter Ansell, and Michel Dumontier. Bio2rdf release 2: improved coverage, interoperability and provenance of life science linked data. In *Extended Semantic Web Conference*, pages 200–212. Springer, 2013.

[42] Bin Chen, Xiao Dong, Dazhi Jiao, Huijun Wang, Qian Zhu, Ying Ding, and David J Wild. Chem2bio2rdf: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC bioinformatics*, 11(1):255, 2010.

[43] Andra Waagmeester, Martina Kutmon, Anders Riutta, Ryan Miller, Egon L Willighagen, Chris T Evelo, and Alexander R Pico. Using the semantic web for rapid integration of wikipathways with other biological online data resources. *PLoS computational biology*, 12(6):e1004989, 2016.

[44] Neil Swainston, Riza Batista-Navarro, Pablo Carbonell, Paul D Dobson, Mark Dunstan, Adrian J Jervis, Maria Vinaixa, Alan R Williams, Sophia Ananiadou, Jean-Loup Faulon, et al. biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PloS one*, 12(7):e0179130, 2017.

[45] Núria Queralt-Rosinach, Janet Piñero, Àlex Bravo, Ferran Sanz, and Laura I Furlong. Disgenet-rdf: harnessing the innovative power of the semantic web to explore the genetic basis of diseases. *Bioinformatics*, 32(14):2236–2238, 2016.

[46] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, 45(D1):D833–D839, 2017.

[47] Alasdair JG Gray, Paul Groth, Antonis Loizou, Sune Askjaer, Christian Brenninkmeijer, Kees Burger, Christine Chichester, Chris T Evelo, Carole Goble, Lee Harland, et al. Applying linked data approaches to pharmacology: Architectural decisions and implementation. *Semantic Web*, 5(2):101–113, 2014.

[48] Wouter Beek, Laurens Rietveld, Stefan Schlobach, and Frank van Harmelen. Lod laundromat: Why the semantic web needs centralization (even if we don't like it). *IEEE Internet Computing*, 20(2):78–81, 2016.

[49] Xiao Dong, Y Ding, H Wang, B Chen, and D Wild. Chem2bio2rdf dashboard: Ranking semantic associations in systems chemical biology space. *Future of the Web in Collaboratice Science (FWCS), WWW*, 2010.

[50] Maulik R Kamdar and Mark A Musen. Phlegra: Graph analytics in pharmacology over the web of life sciences linked open data. In *Proceedings of the 26th International Conference on World Wide Web*, pages 321–329. International World Wide Web Conferences Steering Committee, 2017.

[51] Larisa N Soldatova, Andrey Rzhetsky, Kurt De Grave, and Ross D King. Representation of probabilistic scientific knowledge. *Journal of biomedical semantics*, 4 Suppl 1(Suppl 1):S7, 2013.

[52] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1):496, 2011.

[53] Alison Callahan, Juan José Cifuentes, and Michel Dumontier. An evidence-based approach to identify aging-related genes in caenorhabditis elegans. *BMC bioinformatics*, 16(1):40, 2015.

[54] Gang Fu, Ying Ding, Abhik Seal, Bin Chen, Yizhou Sun, and Evan Bolton. Predicting drug target interactions using meta-path-based semantic network analysis. *BMC bioinformatics*, 17(1):160, 2016.

[55] Alberto et al. Abelló. Fusion cubes: towards self-service business intelligence. 2013.

[56] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508, 2017.

[57] Pedro Domingos, Daniel Lowd, Stanley Kok, Hoifung Poon, Matthew Richardson, and Parag Singla. Just add weights: Markov logic for the semantic web. *Uncertainty Reasoning for the Semantic Web I*, pages 1–25, 2008.

[58] Tijl De Bie, Léon-Charles Tranchevent, Liesbeth MM Van Oeffelen, and Yves Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–i132, 2007.

[59] Andrew Yates, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, et al. Ensembl 2016. *Nucleic acids research*, 44(D1):D710–D716, 2015.

[60] Simon Jupp, James Malone, Jerven Bolleman, Marco Brandizi, Mark Davies, Leyla Garcia, Anna Gaulton, Sebastien Gehant, Camille Laibe, Nicole Redaschi, Sarala M Wimalaratne, Maria Martin, Nicolas Le Novère, Helen Parkinson, Ewan Birney, and Andrew M Jenkinson. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–9, may 2014.

[61] András Gézsi, Árpád Kovács, Tamás Visnovitz, and Edit I Buzás. Systems biology approaches to investigating the roles of extracellular vesicles in human diseases. *Experimental & molecular medicine*, 51(3):33, 2019.

[62] Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551, 2017.

[63] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature biotechnology*, 34(5):518, 2016.

[64] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17(1):239, 2016.

[65] Erwin L van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. The third revolution in sequencing technology. *Trends in Genetics*, 34(9):666–681, 2018.

[66] Fritz J Sedlazeck, Hayan Lee, Charlotte A Darby, and Michael C Schatz. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19(6):329, 2018.

[67] Tuomo Mantere, Simone Kersten, and Alexander Hoischen. Long-read sequencing emerging in medical genetics. *Frontiers in genetics*, 10:426, 2019.

[68] Alberto Magi, Roberto Semeraro, Alessandra Mingrino, Betti Giusti, and Romina D'aurizio. Nanopore sequencing data analysis: state of the art, applications and challenges. *Briefings in bioinformatics*, 19(6):1256–1272, 2017.

[69] Mohammed-Amin Madoui, Stefan Engelen, Corinne Cruaud, Caroline Belser, Laurie Bertrand, Adriana Alberti, Arnaud Lemainque, Patrick Wincker, and Jean-Marc Aury. Genome assembly using nanopore-guided long and error-free dna reads. *BMC genomics*, 16(1):327, 2015.

[70] Alexis L Norris, Rachael E Workman, Yunfan Fan, James R Eshleman, and Winston Timp. Nanopore sequencing detects structural variants in cancer. *Cancer biology & therapy*, 17(3):246–253, 2016.

[71] Mircea Cretu Stancu, Markus J Van Roosmalen, Ivo Renkens, Marleen M Nieboer, Sjors Middelkamp, Joep De Ligt, Giulia Pregno, Daniela Giachino, Giorgia Mandrile, Jose Espejo Valle-Inclan, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature communications*, 8(1):1326, 2017.

[72] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338, 2018.

[73] Anton Bankevich and Pavel A Pevzner. Joint analysis of long and short reads enables accurate estimates of microbiome complexity. *Cell systems*, 7(2):192–200, 2018.

[74] Christopher R O'Donnell, Hongyun Wang, and William B Dunbar. Error analysis of idealized nanopore sequencing. *Electrophoresis*, 34(15):2137–2144, 2013.

[75] Thomas Laver, J Harrison, PA O'neill, Karen Moore, Audrey Farbos, Konrad Paszkiewicz, and David J Studholme. Assessing the performance of the oxford nanopore technologies minion. *Biomolecular detection and quantification*, 3:1–8, 2015.

[76] Raga Krishnakumar, Anupama Sinha, Sara W Bird, Harikrishnan Jayamohan, Harrison S Edwards, Joseph S Schoeniger, Kamlesh D Patel, Steven S Branda, and Michael S Bartsch. Systematic and stochastic influences on the performance of the minion nanopore sequencer across a range of nucleotide bias. *Scientific reports*, 8(1):3159, 2018.

[77] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the minion nanopore sequencer. *Nature methods*, 12(4):351, 2015.

[78] Matthew T Noakes, Henry Brinkerhoff, Andrew H Laszlo, Ian M Derrington, Kyle W Langford, Jonathan W Mount, Jasmine L Bowman, Katherine S Baker, Kenji M Doering, Benjamin I Tickman, et al. Increasing the accuracy of nanopore dna sequencing using a time-varying cross membrane voltage. *Nature biotechnology*, 37(6):651, 2019.

[79] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. Deepnano: deep recurrent neural networks for base calling in minion nanopore reads. *PloS one*, 12(6):e0178751, 2017.

[80] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, May 2018.

[81] Sheng Wang, Zhen Li, Yizhou Yu, and Xin Gao. Wavenano: a signal-level nanopore base-caller via simultaneous prediction of nucleotide labels and move labels through bi-directional wavenets. *Quantitative Biology*, 6(4):359–368, 2018.

[82] Robert Lanfear, Miriam Schalamun, David Kainer, W Wang, and Benjamin Schwessinger. Minionqc: fast and simple quality control for minion sequencing data. *Bioinformatics*, 35(3):523–525, 2018.

[83] Franka J Rang, Wigard P Kloosterman, and Jeroen de Ridder. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology*, 19(1):90, 2018.

[84] Chen Yang, Justin Chu, René L Warren, and Inanç Birol. Nanosim: nanopore sequence read simulator based on statistical characterization. *GigaScience*, 6(4):gix010, 2017.

[85] Yu Li, Renmin Han, Chongwei Bi, Mo Li, Sheng Wang, and Xin Gao. Deepsimulator: a deep simulator for nanopore sequencing. *Bioinformatics*, 34(17):2899–2908, 2018.

[86] Christian Rohrandt, Nadine Kraft, Pay Gießelmann, Björn Brändl, Bernhard M Schuldt, Ulrich Jetzek, and Franz-Josef Müller. Nanopore simulation–a raw data simulator for nanopore sequencing. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1–8. IEEE, 2018.

[87] Héctor Rodríguez-Pérez, Tamara Hernández-Beeftink, José M Lorenzo-Salazar, José L Roda-García, Carlos J Pérez-González, Marcos Colebrook, and Carlos Flores. Nanodj: a dockerized jupyter notebook for interactive oxford nanopore minion sequence manipulation and genome assembly. *BMC bioinformatics*, 20(1):234, 2019.

[88] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome Biology*, 20(1):129, 2019.

[89] Satomi Mitsuhashi, Kirill Kryukov, So Nakagawa, Junko S Takeuchi, Yoshiki Shiraishi, Koichiro Asano, and Tadashi Imanishi. A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Scientific reports*, 7(1):5657, 2017.

[90] Lee J Kerkhof, Kevin P Dillon, Max M Häggblom, and Lora R McGuinness. Profiling bacterial communities by minion sequencing of ribosomal operons. *Microbiome*, 5(1):116, 2017.

[91] Jongoh Shin, Sooin Lee, Min-Jeong Go, Sang Yup Lee, Sun Chang Kim, Chul-Ho Lee, and Byung-Kwan Cho. Analysis of the mouse gut microbiome using full-length 16s rrna amplicon sequencing. *Scientific reports*, 6:29681, 2016.

[92] Mandeep Singh, Ghamdan Al-Eryani, Shaun Carswell, James M Ferguson, James Blackburn, Kirston Barton, Daniel Roden, Fabio Luciani, Tri Giang Phan, Simon Junankar, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nature communications*, 10(1):3120, 2019.

[93] James M Heather, Mazlina Ismail, Theres Oakes, and Benny Chain. High-throughput sequencing of the t-cell receptor repertoire: pitfalls and opportunities. *Briefings in bioinformatics*, 19(4):554–565, 2017.

[94] Roxanne R Zascavage, Kelcie Thorson, and John V Planz. Nanopore sequencing: An enrichment-free alternative to mitochondrial dna sequencing. *Electrophoresis*, 40(2):272–280, 2019.

[95] Juvid Aryaman, Iain G Johnston, and Nick S Jones. Mitochondrial heterogeneity. *Frontiers in genetics*, 9, 2018.

[96] Arthur C Rand, Miten Jain, Jordan M Eizenga, Audrey Musselman-Brown, Hugh E Olsen, Mark Akeson, and Benedict Paten. Mapping dna methylation with high-throughput nanopore sequencing. *Nature methods*, 14(4):411, 2017.

[97] Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, et al. Highly parallel direct rna sequencing on an array of nanopores. *Nature methods*, 15(3):201, 2018.

[98] Laura Restrepo-Pérez, Chirlmin Joo, and Cees Dekker. Paving the way to single-molecule protein sequencing. *Nature nanotechnology*, 13(9):786, 2018.

[99] Maria S Poptsova, Irina A Il'Icheva, Dmitry Yu Nechipurenko, Larisa A Panchenko, Mingian V Khodikov, Nina Y Oparina, Robert V Polozov, Yury D Nechipurenko, and Sergei L Grokhovsky. Non-random dna fragmentation in next-generation sequencing. *Scientific reports*, 4:4532, 2014.