ELSEVIER

New technologies for DNA analysis – a review of the READNA Project

Review

Steven McGinn[1], David Bauer[19], Thomas Brefort[17], Liqin Dong[19], Afaf El-Sagheer[6,7,8], Abdou Elsharawy[4,5], Geraint Evans[10], Elin Falk-Sörqvist[13], Michael Forster[4], Simon Fredriksson[27], Peter Freeman[2], Camilla Freitag[24], Joachim Fritzsche[25], Spencer Gibson[2], Mats Gullberg[27], Marta Gut[28,29], Simon Heath[28,29], Isabelle Heath-Brun[28,29], Andrew J. Heron[20], Johannes Hohlbein[10], Rongqin Ke[9,13], Owen Lancaster[2], Ludovic Le Reste[10], Giovanni Maglia[20], Rodolphe Marie[16], Florence Mauger[1], Florian Mertes[3], Marco Mignardi[9,13], Lotte Moens[13], Jelle Oostmeijer[15], Ruud Out[15], Jonas Nyvold Pedersen[16], Fredrik Persson[24], Vincent Picaud[18], Dvir Rotem[20], Nadine Schracke[17], Jennifer Sengenes[1], Peer F. Stähler[17], Björn Stade[4], David Stoddart[20], Xia Teng[15], Colin D. Veal[2], Nathalie Zahra[2], Hagan Bayley[20], Markus Beier[17], Tom Brown[6,7], Cees Dekker[11], Björn Ekström[27], Henrik Flyvbjerg[16], Andre Franke[4], Simone Guenther[21], Achillefs N. Kapanidis[10], Jane Kaye[14], Anders Kristensen[16], Hans Lehrach[3], Jonathan Mangion[21], Sascha Sauer[3], Emile Schyns[26], Jörg Tost[1], Joop M.L.M. van Helvoort[15], Pieter J. van der Zaag[12], Jonas O. Tegenfeldt[23], Anthony J. Brookes[2], Kalim Mir[19], Mats Nilsson[9,13], James P. Willcocks[22] and Ivo G. Gut[28,29]

[1] CEA – Centre National de Génotypage, 2, rue Gaston Cremieux, 91057 Evry Cedex, France
[2] University of Leicester, University Road, Leicester LE1 7RH, UK
[3] Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany
[4] Institute of Clinical Molecular Biology, Christian-Albrechts-University (CAU), Am Botanischen Garten 11, D-24118 Kiel, Germany
[5] Faculty of Sciences, Division of Biochemistry, Chemistry Department, Damietta University, New Damietta City, Egypt
[6] School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, UK
[7] Department of Chemistry, University of Oxford, Chemistry Research Laboratory, 12 Mansfield Rd, Oxford OX1 3TA, UK
[8] Chemistry Branch, Department of Science and Mathematics, Faculty of Petroleum and Mining Engineering, Suez University, Suez 43721, Egypt
[9] Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Box 1031, Se-171 21 Solna, Sweden
[10] Biological Physics Research Group, Clarendon Laboratory, Department of Physics, Parks Road, Oxford OX1 3PU, UK
[11] Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, The Netherlands
[12] Philips Research Laboratories, High Tech Campus 11, 5656 AE Eindhoven, The Netherlands
[13] Department of Immunology, Genetics, and Pathology, Science for Life Laboratory, Uppsala University, Sweden
[14] HeLEX – Centre for Health, Law and Emerging Technologies, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK
[15] FlexGen BV, Galileiweg 8, 2333 BD Leiden, The Netherlands

Corresponding author: Gut, I.G. (ivo.gut@cnag.crg.eu)

http://dx.doi.org/10.1016/j.nbt.2015.10.003
1871-6784/© 2015 Published by Elsevier B.V.

[16] DTU Nanotech, Oerstedsplads Building 345 East, 2800, Kongens Lyngby, Denmark
[17] Comprehensive Biomarker Center GmbH, Im Neuenheimer Feld 583, D-69120 Heidelberg, Germany
[18] CEA-Saclay, Bât DIGITEO 565 – Pt Courrier 192, 91191 Gif-sur-Yvette Cedex, France
[19] The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK
[20] Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Mansfield Road, Oxford OX1 3TA, England, UK
[21] Thermo Fisher Scientific Frankfurter Straße 129B, 64293 Darmstadt, Germany
[22] Oxford Nanopore Technologies, Edmund Cartwright House, 4 Robert Robinson Avenue, Oxford Science Park, Oxford OX4 4GA, UK
[23] Division of Solid State Physics and NanoLund, Lund University, Box 118, 22100 Lund, Sweden
[24] Department of Physics, University of Gothenburg, SE-412 96 Gothenburg, Sweden
[25] Department of Applied Physics, Chalmers University of Technology, Kemivägen 10, 412 96 Göteborg, Sweden
[26] PHOTONIS France S.A.S. Avenue Roger Roncier, 19100 Brive B.P. 520, 19106 BRIVE Cedex, France
[27] Olink AB, Dag Hammarskjölds väg 52A, 752 37 Uppsala, Sweden
[28] Centro Nacional de Análisis Genómico (CNAG-CRG), Center for Genomic Regulation, C/Baldiri Reixac 7, 08028 Barcelona, Spain
[29] Universitat Pompeu Fabra (UPF), Barcelona, Spain

Review

The REvolutionary Approaches and Devices for Nucleic Acid analysis (READNA) project received funding from the European Commission for 4 1/2 years. The objectives of the project revolved around technological developments in nucleic acid analysis. The project partners have discovered, created and developed a huge body of insights into nucleic acid analysis, ranging from improvements and implementation of current technologies to the most promising sequencing technologies that constitute a 3rd and 4th generation of sequencing methods with nanopores and *in situ* sequencing, respectively.

Contents

Introduction

Duration, funding & no. partners

The **RE**volutionary **A**pproaches and **D**evices for **N**ucleic **A**cid analysis (**READNA**) project received 12 million € funding under the European Union Framework Programme 7 from 1st June 2008 to 30th November 2012. The 19 project partners from both academia and industry from in total 7 countries had a project budget of 16 M€ with which they have discovered, created and developed a huge body of insights into nucleic acid analysis. Results have been presented widely in publications and in innumerous public presentations. Results have been moved to spin-offs such as the Olink enrichment kits (now sold by Agilent as Haloplex) and are finding their way to the market, such as the Oxford Nanopore MinIon sequencer that was first released to early-access user sites in the summer of 2014.

Objectives & context

The READNA consortium collaborated in a wide range of developments in nucleic acid analysis methods to accelerate new breakthrough DNA sequencing technologies, to enhance existing analysis methods, and advance nucleic analysis methods to the benefit of patients and society. Participants had diverse and complementary expertise from a wide range of scientific disciplines. The interdisciplinary nature of the consortium allowed the exploration of novel concepts, ease the use and broaden nucleic acid analysis technologies, establish best practices and standards for research and diagnostics, accelerate and simplify methods and create a toolbox for precision medicine. One of the declared goals was to progress technologies enabling sequencing of an entire human genome for 1000€ in less than one day.

The project was organized into five research work packages, and a dissemination and management work package.

The five research work packages were, WP1: Near Term Innovations: Ancillary elements for 2nd generation DNA sequencers; WP2: Near Term Innovations: Improvement and extension of existing methods; WP3: Fluorescence-based Single Molecule Sequencing; WP4: Nanopore Sequencing; and WP5: New Genotyping Challenges.

Project highlights & achievements

Introduction

Progress has been made through all of the fronts that the five work packages have been focusing on towards the overall objectives of READNA. The results have been presented in over 100 scientific publications, many of which are in very high impact factor journals [1–139], including a number of review papers [1–4]. Here we highlight a few selected discoveries and developments.

Discovery of 'TUF' DNA

Methods for genome analysis must contend with the fact that some chromosome regions are extremely difficult to assay, producing weak or no signals and low data accuracy. This behavior is not easily explained by mere differences in C + G content or secondary structures from region to region, as the problem varies in size between DNA samples. We investigated this phenomenon via a series of experiments, and thereby uncovered a novel mechanistic basis for amplification and assay differences between samples and between genome regions [5]. In short, we found that difficult to assay regions contain tiny (<<1 kb) stretches that are extremely resistant to DNA denaturation (a key step in almost all DNA analysis methods), even up to ~130°C. Assay efficiencies are reduced in surrounding zones that can extend several hundreds of kbp from such a site. We call these stretches 'Thermodynamically Ultra-Fastened' (TUF) regions. The 'non-melting' sections within TUF regions are extremely high in C + G content and are made up of elements whose particular arrangements probably enables them to cooperate in holding the two DNA strands together (Fig. 1). Long DNA stretches that flank the non-melting loci are able to denature but remain in close proximity since they are tethered together at the non-melted anchor regions. As such they can very rapidly re-anneal and thereby hinder access by primers or probes, so preventing assays from working over extended domains. The discovery of TUF DNA allows us to explain the inter-sample variability of assay efficiency in these troublesome regions. Specifically, some DNA samples will contain a greater degree of randomly distributed nicking than others, and upon heating these nicks will allow the denatured flanking regions to diffuse away from the non-denatured core TUF domains – such that highly intact DNAs will be most severely affected and assay the least well. A major practical consequence of all of this is that inter-region and inter-sample variability can be largely overcome by employing routine fragmentation methods (e.g. sonication or restriction enzyme digestion) prior to sample interrogation, even though this runs counter to traditional best practice.

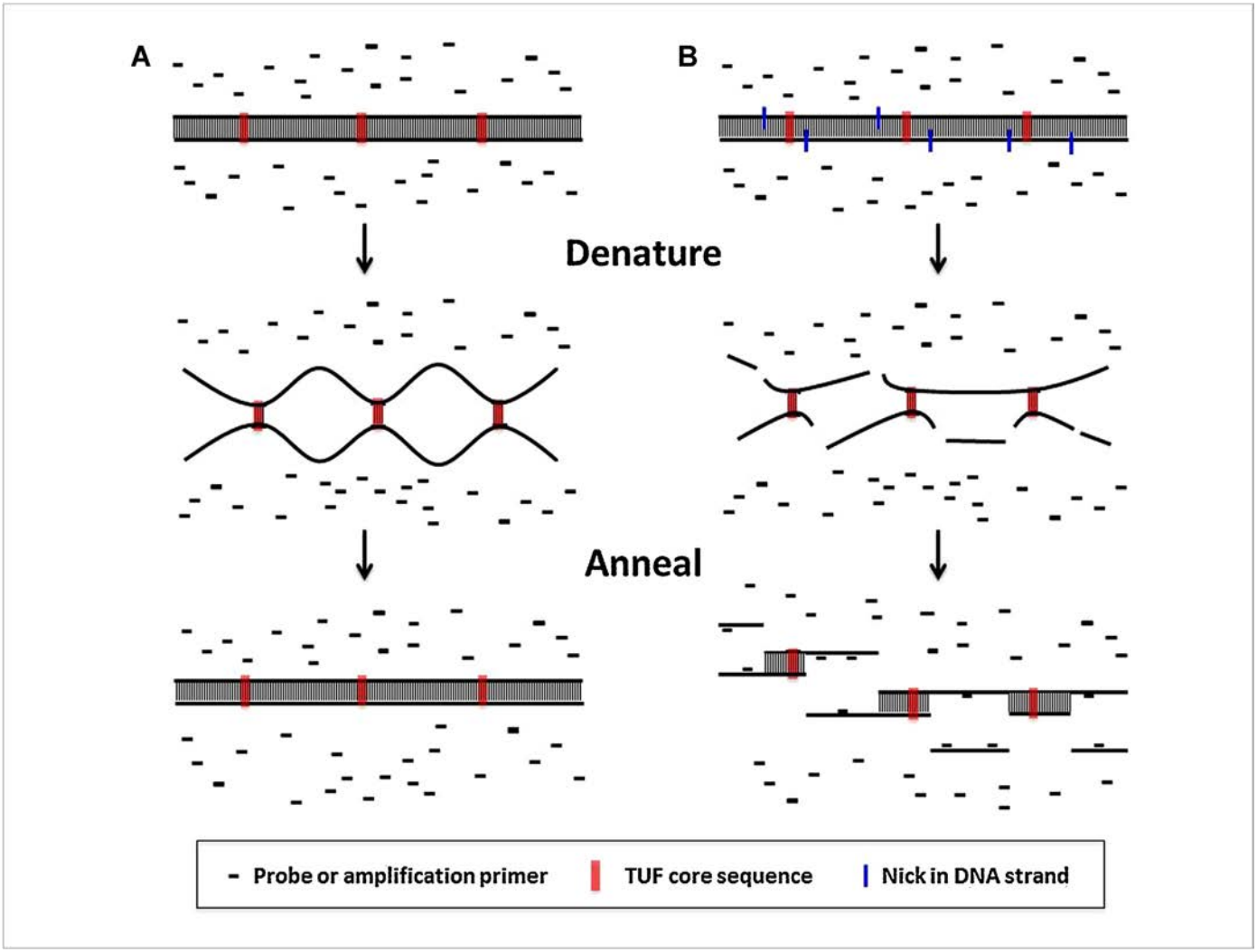

FIGURE 1

A putative TUF region is illustrated for two DNA samples that are either **(a)** high quality (low degree of nicking, so long fragments per strand) or **(b)** low quality (high degree of nicking, so short fragments per strand). Upon denaturing the DNA strands are held together at the core TUF sequences that strongly resist melting. For (a), the entire TUF region and flanks are held together, whereas in (b) sections of single and double stranded material are disconnected from each other and so can diffuse apart. Upon imposing conditions for primer or probe annealing, in (a) the complimentary DNA strands that were held together at core TUF sites rapidly and fully re-anneal, preventing access by primers or probes. Conversely, in (b) many single stranded portions now exist to which the primers and probes can anneal, enabling efficient amplification or detection [though regions very close to the core TUF sequences remain inaccessible]. The overall efficiency of DNA amplification or probe hybridization of target sites within a TUF region will therefore be influenced by the general amount of DNA nicking present and the average per target degree of physical distance from the core TUF sequences. The counter-intuitive result of this is that stronger assay and amplification signals will be produced by lower quality DNAs.

Enrichment methods

Since 2nd generation DNA sequencing emerged its cost has decreased, at first dramatically, then for several years was stable with sequencing of a human genome at 30× coverage costing roughly 5000$ in reagents, and then only recently dropped to ~1000$ with Illumina's HiSeq X system. For high value experiments these costs are acceptable, however, particularly for experiments that involve the analysis of many samples costs become prohibitive. Moreover, for some applications 30× read depth is far from sufficient, for example for application in cancer diagnostics where the tumor cell content may be low, and where sequencing needs to be done to a considerable depth (>1000×) to enable accurate detection of somatic mutations in a 10% cancer cell population. Since very early on, enrichment approaches for focusing only on the regions of interest have been developed. The technical solutions available are very broad, diverse and have different sweet spots. They range from simple PCR for the isolation of a very small fraction of a genome to hybridization-based methods by which several percent of a genome can be enriched. An overview of the different methods for targeted enrichment is depicted in Fig. 2. Within READNA we have developed several enrichment methods (see below). We have also carried out benchmarking of different methods and have developed a concise set of descriptors for an enrichment experiment that precisely captures the essence of an enrichment experiment and method.

Enrichment based on selectors

Prior to the READNA project a novel enrichment technique entitled 'Selectors' was developed to specifically select large sets of DNA sequences for parallel amplification by PCR using

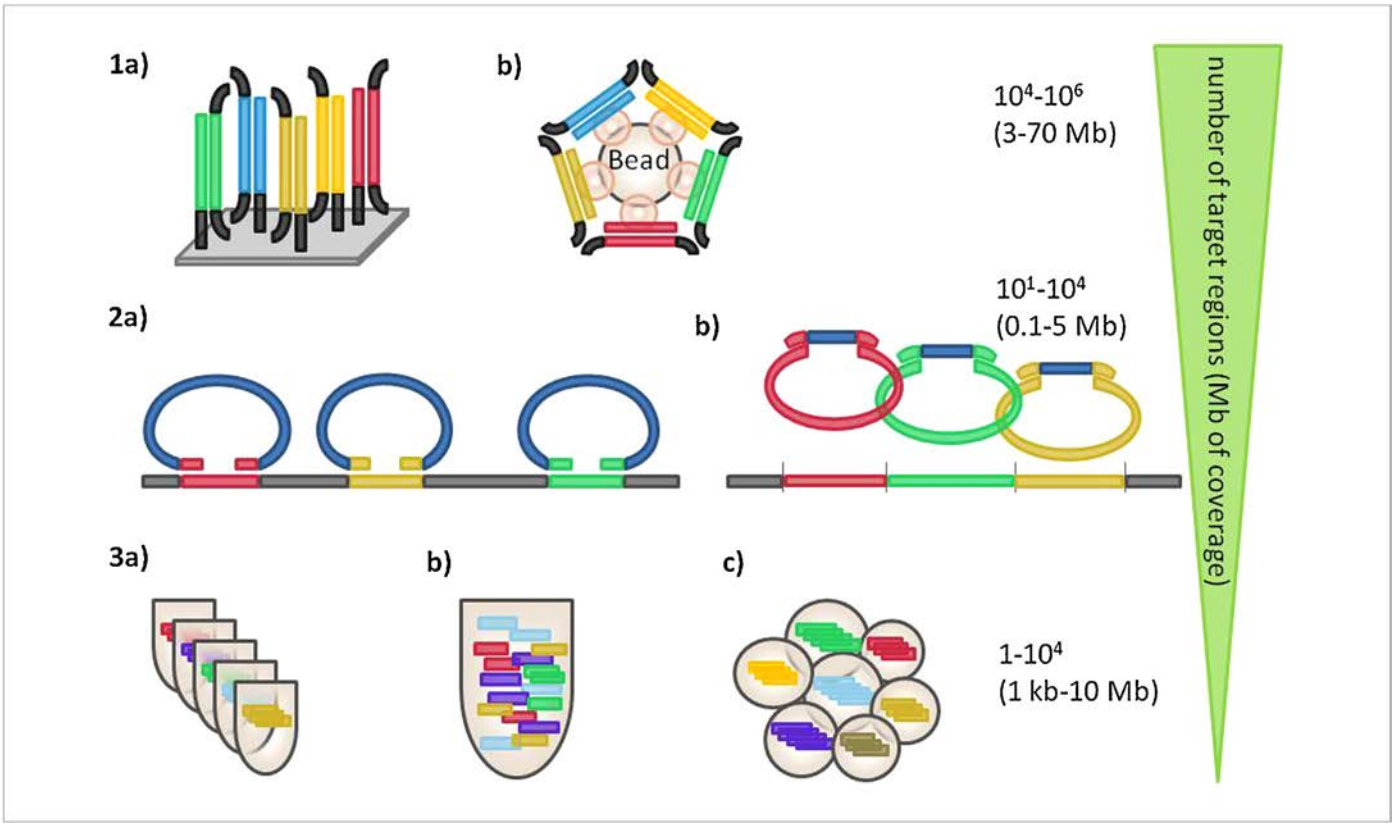

FIGURE 2

Commonly used targeted enrichment techniques. (1) Hybrid capture targeted enrichment either on solid supports such as microarrays (a) or beads (b). A shot-gun fragment library is prepared and hybridized against a library containing the target sequence. After hybridization (and bead coupling) non-target sequences are washed away, the enriched sample can be eluted and further processed for sequencing. (2) Enrichment by Molecular Inversion Probes (MIPs) which are composed of a universal sequence (blue) flanked by target-specific sequences. MIPs are hybridized to the region of interest, followed by a gap filling reaction and ligation to produce closed circles. The classical MIPs are hybridized to mechanically sheared DNA (a), the Selector Probe technique uses a cocktail of restriction enzymes to fragment the DNA and the probes are adapted via the restriction sites (b). (3) Targeted enrichment by PCR-based approaches. Typical PCR with single-tube per fragment assay (a), multiplex PCR assay with up to 50 fragments (b) and RainDance micro droplet PCR with up to 20,000 unique primer pairs (c) utilized for targeted enrichment.

target-specific oligonucleotide constructs called selectors. Selectors are oligonucleotide duplexes with single-stranded target-complementary end-sequences that are linked by a common sequence motif. A pool of selectors is combined with denatured restriction digested DNA and each selector hybridizes to its respective target, forming individual circular complexes. The common sequence that is introduced into the circularized fragments allows the products to be PCR amplified in parallel using a universal primer pair [6]. The technique was rapidly adapted to next generation sequencing (NGS) instruments and in particular the 454/Roche instrument highlighting the potential of the technique to combine highly multiplexed enrichment with NGS [7]. During the READNA project the Selector technology was developed further to adapt it to NGS instruments of higher throughput, such as the Illumina and SOLiD systems. Initially, and due to the short read-length of these systems at the time, this was done by developing a platform independent Multiple Displacement Amplification (MDA) based version, where the sequencing reads were initiated randomly across the enriched DNA. Moreover, selector probe libraries were amplified from programmable DNA microarrays, substantially reducing probe cost. The modified protocol was validated on 28 genes frequently mutated in common solid tumors [8]. Due to the development of longer reads on particularly the Illumina HiSeq and MiSeq instruments, a PCR-based version was introduced that enables integration of the enrichment with the sequencing library construction into a one-step procedure. This version is now commercially available through Agilent Technologies under the name of Haloplex™. In addition the protocol can be used on FFPE samples for cancer diagnostics applications [9,10].

Enrichment based on long DNA fragments

Anticipating improvements in the lengths of fragments that could be sequenced on NGS platforms, we explored a novel solution-phase DNA enrichment approach that used multiple, co-operating, non-overlapping DNA baits per target region, and novel blocking reagents to suppress non-specific hybridization. Additionally, we identified key procedural improvements that: (i) enable bait pools (sets of harvested oligonucleotides from programmable array synthesis) to be massively amplified with high fidelity using solution-phase PCR (rather than emulsion PCR) to give quantities needed for hundreds of enrichment reactions, and (ii) facilitate highly robust long-fragment library preparation. Combining these innovations into a complete procedure yielded combined levels of enrichment power and enrichment evenness

that match or exceed the levels achieved by all alternative solution-phase methods.

MegaPlex PCR

A simple approach to target DNA enrichment would involve direct PCR amplification of the regions of interest. However, this theoretically attractive option is confounded by the fact that high-plexity PCR reactions fail due to excessive inter-primer reactions and off-target priming initiated by the many primers available in the solution. We sought to address this limitation by placing each intended primer pair at a different physical location on a surface within such a reaction, thereby preventing unwanted primer–primer interactions or off-target priming. To make this new 'MegaPlex PCR' [11] reaction format work well, we also established: (i) how to enable surface based PCR priming events to occur efficiently and specifically, and (ii) how to enable all the resulting amplicons to leave the surface carrying common primer sequences, so that they could together be uniformly amplified by a single primer pair in a final solution phase PCR, generating enough product for DNA sequencing.

SIMLY targeted enrichment

Targeted enrichment by hybrid capture has proven to be a highly efficient method (e.g. [12]), particularly if large genomic target regions as well as many samples need to be analyzed. Herein RNA probes are hybridized in solution to specifically bind and enrich the region of interest. The RNA probes are transcribed from a large pool of synthesized DNA template oligonucleotides. As part of the project the goal was to make targeted enrichment by in-solution capture more cost efficient, especially for very large sample numbers. This was achieved by applying emulsion PCR for DNA template library amplification virtually immortalizing a once generated DNA template library and therefore allowing the recurrent synthesis of RNA bait libraries for targeted enrichment experiments. The efficacy of the developed protocol matched the results obtained from commercial kits, such as the SureSelect system commercialized by Agilent Technologies, and furthermore demonstrated that emulsion PCR can be used for bias free amplification of DNA template libraries. The developed 'short immortalized DNA template library' (SIMLY) method for targeted enrichment is a straightforward example for significantly reducing costs for NGS applications which are anticipated to grow rapidly especially in molecular diagnostics within the near future [12].

Benchmarking of enrichment methods

To bring more efficiency to the sequencing process and enable large-scale genomic investigations for complex diseases, we have conducted many studies to evaluate the true capabilities and performance of many upfront targeted and whole-exome approaches using different NGS technology platforms, tested and standardized different scenarios of sample multiplexing to bring down the cost per sample, and examined the applicability of whole-genome amplified (WGA) materials, instead of original gDNA, for NGS to preserve precious patient samples. The results of our studies revealed that pre-enrichment sample multiplexing using hybridization-based methods is technically preferable over PCR-based approaches [14,15]. Indexed NGS libraries, instead of gDNA were successfully enriched using one selected microarray-based capture method [14]. Reproducible coverage profiles (R^2 up to 0.99) for the target regions were achieved across most of the different multiplexed samples, enriching human exonic as well as intronic regions with less than 10% strand bias. In addition, several experimental alternatives were performed in the PCR-based study [15] to investigate the effect of using WGA gDNA material and different pooling strategies on the performance of selective recovery of genomic subregions of interest and on variant detection. The results confirm that WGA can be combined with NGS with no substantial bias in enrichment efficiency. The results of these experiments further support the possibility to process non-pooled and pooled samples (before or after an emulsion PCR (emPCR) step) without substantial bias, and reveal that pooling pre-emPCR can reduce the cost per sample while maintaining good performance [15]. When exome kits became commercially available on different platforms, we performed a multi-platform benchmark of various exome NGS approaches. The results were surprising, showing divergent enrichment efficiencies and low overlap of single nucleotide variant (SNV) calls (Fig. 3a). When we used Sanger sequencing to validate 765 SNV-calls of potential clinical relevance that were concordant between two different kit/sequencer/bioinformatics platforms, only 57-97% concordance was seen in HapMap individual NA12752 (data derived by Sanger sequencing (Fig. 3b)).

The results presented here emphasize that current whole-exome and targeted NGS (T-NGS) methods are not yet optimal, and the long term utility of these methods will depend increasingly on progress towards evenness and enrichment power improvements and also on new and better strategies [16,17]. Taken together, we do expect that even if the cost of routine sequencing of the human exome and whole-exome becomes affordable in the near future, multiplexed targeted NGS will serve alongside as a downstream validation and diagnostic tool to reach sufficient coverage with less sequencing and cost, enabling the investigation of a higher number of samples in parallel [18,19].

Metrics for the assessment of an enrichment experiment

With the wide variety of methods available for targeted enrichment of selected genomic regions for NGS, comparing results among them, and even between experiments using the same enrichment technology, is difficult, as no universal terminology for reporting the quality of target enrichment is in use. In the literature various sets of parameters reporting the results are used, and even if similar parameters seem to be reported, their definition may be different. To deal with this issue, a universal set of Target Enrichment Sequencing Descriptors (TESD) metrics has been developed including clear mathematical formulas to define the key parameters. These TESD metrics enable a complete and unique description of target enrichment experimental results, irrespective of the underling enrichment technology. This set of TES-descriptors uses the following 7 parameters [20]:

1. Size of the Region of Interest, ROI
2. Weight of the input DNA requirement, W
3. Average read depth across the ROI, D_{ROI}
4. Fraction sufficiently covered at a read depth of x, F_x
5. Specificity, S
6. Enrichment factor, EF
7. Evenness, E

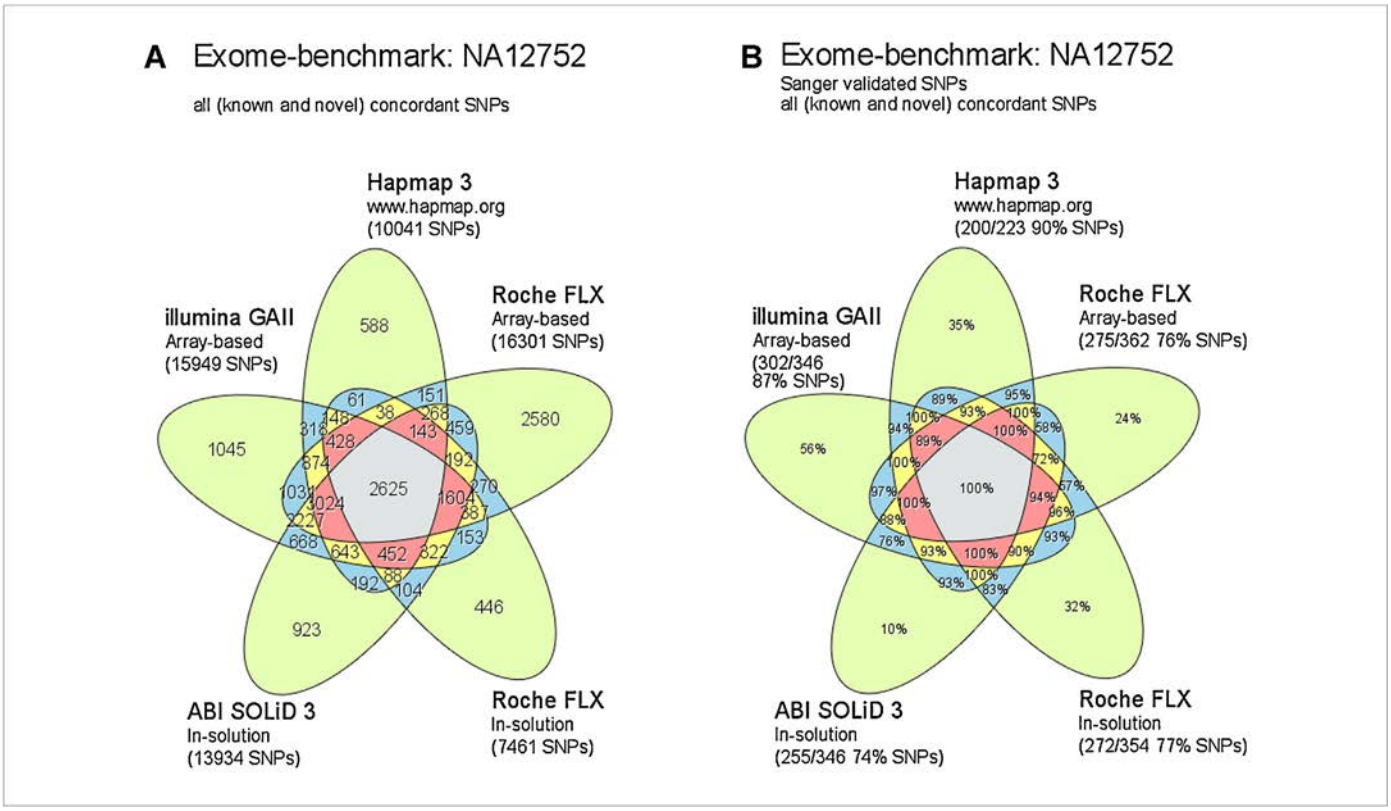

FIGURE 3

Benchmark of different exome solutions based on DNA from HapMap individual NA12752. **(a)** Concordant overlaps of SNVs from different exome solutions with each other and with the high quality HapMap 3 SNPs. Concordant overlaps are defined as shared genomic position and shared genotype. **(b)** Validation rate by Sanger sequencing for 765 representative SNVs selected for potential clinical interest.

Of these parameters the first two (ROI, *W*) give the input requirements, while the remaining five are a measureable report on either the resulting sequencing (D_{ROI}, F_x) or on the enrichment (*S*, EF, *E*). Via these seven parameters a complete description can be given of the performance of any target enrichment method. Adopting this methodology has the following two advantages. Firstly, within a given enrichment method, of say target capture by hybridization arrays, one can directly compare whether a certain adaption of the method is an improvement over the existing method. Consequently results obtained in the literature could be consistently reported and meaningfully compared to each other. Secondly, the strength of different methodologies (e.g. in-solution enrichment methods versus PCR amplification based enrichment) can be compared to determine, given a certain ROI, what would be the optimal and most cost-effective method for enriching a certain ROI. This could be relevant for future clinical application when choosing the optimal method for T-NGS. To this end a comparison and overview has been made between the various, currently used enrichment methods using these seven TESD parameters [20].

Mass spectrometry based DNA analysis methods

Prior to the READNA project, DNA genotyping and mitochondrial DNA re-sequencing MS methods were developed [21,22]. These methods take advantage of a novel class of thermostable DNA polymerases that incorporate ribonucleotides and the high-throughput capacity of MALDI-TOF MS. The basic method involves a PCR reaction, followed by a ribo-extension reaction for the preparation of single-stranded RNA/DNA chimera. The chimera is then cleaved after each ribonucleotide base by sodium hydroxide treatment. Cleaved fragments are then desalted by cation exchange resin charged with H^+ and analyzed by MALDI-TOF MS in negative ion mode. Within READNA, DNA sequencing and allele-specific multiplex genotyping MS methods were developed. They are based on the same concept but, in these newly developed methods a double-stranded RNA/DNA chimera is synthesized to reduce the number of reaction steps. Software was also created to analyze mass spectra for high-throughput analyzes.

Ribo-PCR MS sequencing

A simple ribonucleotide-PCR (ribo-PCR) MS sequencing method has been established [23]. The ribo-PCR method requires only genomic DNA and a PCR mastermix for the preparation of double-stranded RNA/DNA chimera which contains three deoxynucleotides and the fourth nucleotide in its ribonucleotide form. Therefore, the number of steps is reduced as no ribo-extension reaction is required: ribo-PCR, cleavage, desalting, and MS analysis. The mass fingerprint is used to find deviations from the reference sequence and to identify variations (polymorphisms, deletions and insertions) in the sequence being studied. The method takes advantage of the complementary information of both strands of the RNA/DNA chimera in a single reaction. Rare, frequent and unknown polymorphisms can be identified easily with this facile, rapid and accurate method. Ribo-PCR MS sequencing is well suited

for screening DNA sequences of limited regions of interest in a large number of individuals. Software was established to automatically analyze mass spectra, calculate fragment masses, compare mass spectra with those of the reference sequence and identify the sequence of each individual. It contains several steps: loading, file format conversion, peak picking, computational kernel and the control of the candidate sequence by the user.

Ribo-PAP-PCR genotyping

A single-tube allele-specific multiplex genotyping and sequencing method was also developed [24]. It is based on the ribo-PCR method but in this case ribonucleotide analog pyrophosphorolysis-activated polymerization (ribo-PAP PCR) is used for multiplex genotyping. The allele specificity of the ribo-PAP PCR is established with block 3′-PO_4 primers and a new thermostable DNA polymerase. All allele-selective primers also contain a 5′ repetitive motif where each motif has a distinct mass upon reverse copying and alkali fragmentation. Masses of complementary motifs identify primers that were recruited in the ribo-PAP PCR reaction and the genotype of the individual can be determined. This method enables resequencing and multiplex genotyping in a single reaction and in a large number of individuals. It thus constitutes the simplest genotyping protocol with multiplex analysis.

Risk profiling of multifactorial diseases by MALDI mass spectrometry

Mass spectrometry has become an indispensable tool in the life sciences and in diagnostics. Due to the direct measurement of a physical property, the molecular mass to charge ratio, mass spectrometry enables reliable detection of various biomolecules including nucleic acids and proteins. Therefore, mass spectrometry can be applied with great flexibility for integrated diagnostic analyses of complex diseases by making use of the combined information gained from the simultaneous detection of nucleic acids and proteins – instead of detecting only a very limited number of causative or surrogate markers. We could show that the analysis of DNA markers such as SNPs and protein patterns allowed for precise diagnosis to better characterize the disease risk profiles compared to previous state-of-the-art methods [25–27]. This approach was successfully applied for microbial applications [28]. Due to standardized high-throughput workflows the method is now being introduced in many clinical microbiology laboratories replacing conventional biochemical methods [29].

Array based Dynamic Allele-Specific Hybridization (Array-DASH)

Sequence-specific hybridization has been a cornerstone of DNA analysis methods for decades. Most powerfully, it has been applied in single stringency micro-arrays (providing high plexity of target sequence analysis) and in high-resolution dynamic melting procedures (for exquisite target resolution and accuracy). Building on our previous invention of Dynamic Allele-Specific Hybridization (DASH) [30,31] which was conducted on the walls of microtiter plate wells and on plastic membranes, we exploited this core reaction principle in a completely new way that combines all the above assay advantages. Specifically, we established a robust technology for conducting high-resolution dynamic melting of target-probe DNA duplexes on micro-arrays, in a method we call Array-DASH (manuscript in preparation). Array-DASH is simple but powerful. It entails complete hybridization of target DNAs (any of a vast range of sources, complexities, and length ranges) to microarrays comprising up to many tens of thousands of interrogation probes, and then tracking simultaneously the melting of each of the resulting target-probe duplexes in real-time, as the array surface is steadily heated. Convenient, standard run conditions are applied regardless of the assay particulars, and the melt-curves so produced unambiguously resolve and identify all single base and short indel variations in all tested sequence contexts (other than the most extremely C + G rich elements), with quantitative precision that allows allelic alternatives to be detected at ratios as low as 1:99 (cf., Sanger sequencing requires ratios greater than 30:70). Multiple similar alleles are also fully resolved per array feature. We have shown array-DASH to be suitable for genotyping, mutation scanning, genome fingerprinting, and DNA re-sequencing, either within separate experiments or simultaneously on single arrays. Target concentrations work over ~1000 fold range, with fragment sizes from 50b to at least 2 kb. The method is therefore simple to run and yet far more universal, precise and quantitative than traditional static stringency microarrays. Specific applications so far explored with various domain experts include comprehensive profiling of (circulating) miRNAs, detecting emerging drug resistant HIV species, medicinal plant identification and purity testing, and detection of low level mutations in cancer biopsies. However many other uses can also be imagined, such as prenatal diagnostics using fetal DNA in maternal plasma, identification and drug resistance scoring of microorganisms, and mutation scanning or scoring in genetic disease – all on a single device with no requirement for assay specific optimization of run conditions.

Manipulating and visualizing individual DNA molecules

The contiguity of genome sequences spans the length of chromosomes which in humans range in length from 50 to 250 Mbp. Sequencing technologies tend to break genomic sequence into small packets that are approximately a million fold shorter, comprising single reads or paired reads that lack coordinates in the genome. Individual reads can be as short as Complete Genomics 36 bases to Illumina's reads which have increased from 35 to 150 bases. Technologies with longer reads, 454 and Pacific Biosciences have failed to scale into competitive human genome-sequencing technologies. With no upfront information on the location of short reads in the genome, *de novo* genome assembly is challenging due to the non-randomness of genome sequences and particular sequence elements that are enriched in genomes, such as repetitive elements. Instead, sequencing reads must be aligned to a reference genome, thus, missing the long-range structural context of reads in an individual genome [1].

Single molecule mapping, molecule rescue and next generation sequencing

READNA set out to integrate sequence information with its long-range context. We chose to do this through direct, single molecule analysis because this would allow sequence to be in haplotype phase [32–52]. As a chromosome prior to replication is a single molecule, if the length of DNA in a chromosome can be kept intact, long-range phase and structure of sequence information can be retained.

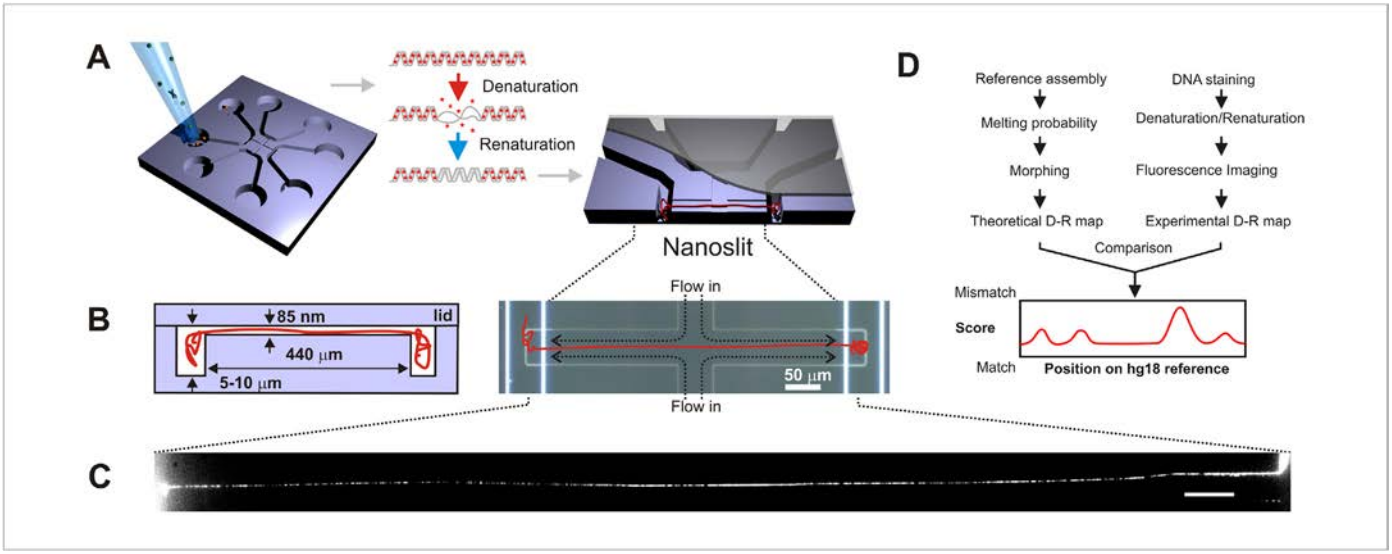

FIGURE 4

Microfluidic chip design, denaturation-renaturation mapping, and detection of structural variation in a single DNA molecule. **(a)** The chip is loaded with cell extract enriched in metaphase chromosomes. Stained DNA is partially denatured and renatured, creating a fluorescence pattern (D-R map). **(b)** The inlet ports of the chip connect to 5–10-μm-deep microchannels for DNA handling, which feed into an 85-nm shallow nanoslit. This nanoslit effectively confines DNA molecules to 2D and stretches them by opposing fluid flows from a second, perpendicular nanoslit. A megabase pair-long DNA fragment is (b) flow-stretched and **(c)** imaged, and **(d)** its D-R map is compared locally to a reference genome; chromosomal origin and structural variations are assessed as good versus poor matches.

We found individual metaphase chromosomes – which are compact and isolatable packets of DNA to be effective vehicles for delivering whole, by definition in-phase, DNA, into a nano/microfluidic device. The chromosomal DNA could then be separated from chromosomal proteins before being stretched [32–34]. At this stage the naked chromosomal DNA retains the canonical X-shape, even though the proteins were digested [32,33]. Once naked DNA was obtained, we found effective means to manipulate and stretch megabases of its length in a microfluidic device. One approach to stretching megabase DNA provided close to 100% stretching, required no nanolithography and is thus amenable to inexpensive mass production [35,36]. A second approach was able to display the entire length of a *S. pombe* chromosome within the field of view of a single CCD array (Fig. 4) with the potential to scale to whole human chromosomes in the future [37].

Extracting, handling & imaging ultra-long DNA and obtaining sequence information in a long-range context

We next developed three approaches for obtaining sequence information in its long-range context. In the first approach, we chose to provide long-range context to current generation sequence and demonstrated that this could be done effectively by denaturation-renaturation (DR) mapping of the DNA [32,36,38–41]. We found that a single molecule viewed in a chip could be rescued from the chip, amplified and then sequenced on an Illumina Genome Analyzer as well as used to probe a metaphase chromosome spread. Consequently, this allowed long-range structure to be reconciled with short range sequence [36]. In order to view the barcode the DNA must be stretched. We investigated local thermophoretic stretching as an easy alternative to flow stretching [39–41]. The decision on which molecule to stretch may be based on their length and topology. Sorting molecules according to length and topology was investigated [42]. In the second approach we used a competitive binding assay to barcode the genomic DNA [43]. In the third approach we captured sequence reads directly on long lengths of DNA stretched in nanofluidic channels. Passivation with lipids proved to be an effective means to enable a plethora of enzymatic reactions to be conducted in the chips [44]. With such passivation we showed that template-directed extension, incorporating fluorescent nucleotides could be carried out within nanochannels (Dong *et al.*; manuscript in press), opening the way to applying fluorescent sequencing-by-synthesis approaches in nanofluidic channels. Finally, in order to ultimately obtain a high-density of sequence information along stretched DNA, we investigated super resolution microscopy and algorithms [45–47].

Sequencing chemistry, sensors

This part of the project focused on developing technologies for fluorescence-based single-molecule sequencing in various contexts [52–55]. One approach involved the use of labeled DNA or RNA polymerases as sequencing engines, with different values of fluorescence intensities and fluorescence ratios providing the information needed to assign bases in a real-time sequencing platform [53]. To enable such an assay, we selected a high-fidelity DNA polymerase (specifically, the Klenow fragment of DNA polymerase I) and used it to develop strategies to monitor distances and conformational changes within various fluorescently labeled polymerase derivatives and inextensible DNA substrates [56–59], this work has recently been extended to extensible DNA substrates, allowing real-time incorporation of individual bases. Furthermore, we explored the use of dark quenchers (non-fluorescent chromophores) as acceptors for single-molecule FRET [60], this ability would be useful for several DNA-sensing assays that require high concentration of an interactant but would be desirable to implement without sophisticated nanofabrication (such as zero-mode waveguides). To increase the spatial and temporal resolution of FRET-based sequencing assays, we provided means to predict and

evaluate single-molecule FRET efficiencies within biomolecules while immobilized on solid supports [61], to monitor several distances within a single molecule or molecular complex [62,63], to detect single-molecule FRET in gel matrices [64] and to detect several sequence-dependent protein-based operations on DNA substrates [65–68], even in cases where single molecules of the relevant proteins (DNA polymerase and DNA ligase) were performing their functions in single living bacterial cells [69]; the latter possibility may allow the development of *in vivo* DNA sequencing assays in the future. We also provided tools that enable several biosensing assays that detect DNA-binding proteins, RNA and DNA. Specifically, we offered novel routes for probing chromosomes with high-resolution by introducing a new algorithm for crowded-field single-molecule super-resolution imaging [70]; we also introduced new ways for detecting and manipulating DNA-binding proteins at the single-molecule level [45,71,72].

Ligation-based sequencing

We developed a ligation-based sequencing-by-synthesis approach using DNA/RNA chimeric oligonucleotides in which only the base immediately following the ligation junction was coded and following detection the non-coded nucleotides/label were removed by RNAses before each next cycle. Compared to a concurrently developed ligation sequencing approach (ABI SOLiD™), our method is less complex not requiring color-space decoding, and as it does not require proprietary oligonucleotide chemistry, it is easily implemented by others using off-the shelf reagents [73].

Click chemistry – enzyme-free ligation based sequencing

In sequencing by synthesis or ligation, the complementary strand of DNA could in principle be assembled by a template mediated chemical reaction instead of a polymerase or ligase enzyme. However, for this approach to be viable the chemistry must be very efficient, compatible with aqueous buffers, and should proceed with similar sequence-fidelity to that of current enzymatic methods. Assuming this demanding set of criteria can be met, chemical sequencing methods that utilize fluorescence or other reporter methodologies [74] might have a number of advantages, including low cost. With this in mind we have investigated several chemical ligation reactions including the Diels-Alder reaction [75], the CuAAC reaction [76] and the SPAAC reaction [77]. Extensive work on the Sharpless-Meldal Cu(I)-catalysed alkyne-azide cycloaddition reaction (CuAAC reaction) in the nucleic acid context [78] led us to design an inter-nucleoside triazole linkage that can be read through by DNA and RNA polymerase enzymes, albeit with the omission of a single nucleotide at the site of chemical ligation [76,79]. Nevertheless this was a highly promising advance and further optimization yielded a DNA triazole linkage that can be read through accurately by DNA and RNA polymerase enzymes [80,81]. Surprisingly DNA containing this linkage is functional in bacterial and mammalian cells [81] and the structural basis of this has been recently established [82]. The triazole linkage is a phosphodiester mimic, and when present in a DNA template, its ring nitrogen atoms can form hydrogen bonds with polymerase enzymes [82]. This linkage was also able to support the activity of a hammerhead ribozyme when positioned close to the cleavage site of the substrate, and has been used for the chemical synthesis of large RNA constructs [83]. The CuAAC reaction has one potential drawback; the use of Cu(I) leads to DNA damage unless the reaction conditions are carefully controlled (exclusion of oxygen). The procedure might not prove to be robust in the hands of inexperienced operators, and might also be difficult to automate. Although the biocompatibility of the triazole linkage is fascinating from a theoretical point of view, it is not essential in the context of DNA sequencing by chemical methods. To circumvent the potential problems caused by the presence of Cu(I) we then focused on the Strain-Promoted Alkyne Azide Cycloaddition reaction (SPAAC reaction) which does not require metal ion catalysis. This reaction has proved to be highly efficient in a DNA templated context [77,84] and we are encouraged by the finding that chemical ligation by the SPAAC reaction is inhibited by the presence of mismatched base pairs [77]. Our work on chemical ligation of DNA and RNA has recently been reviewed [85]. Chemical methods of DNA sequencing are likely to be more flexible than the current enzyme-based methods, and could accommodate a wider range of substrates. This would be an advantage when sequencing small RNAs and other modified nucleic acids.

Nanopore sequencing

The initial concept of using the measured ionic conductance of nanoscale pores for sequencing DNA was conceived over twenty years ago. Nevertheless, prior to the start of the READNA project, significant challenges remained: one being the discrimination of single nucleotides within a nanopore, including modifications to the standard bases, the other being a method of controlling the translocation speed as DNA is drawn electrophoretically through a nanopore. A third challenge, that of providing a system for interrogating many nanopores in parallel in clinically useful devices, whilst primarily drawn from commercial considerations, also presented significant technical hurdles. Three projects within READNA were therefore designed to investigate methods and techniques to overcome these challenges. The first project aimed to explore the utility of direct sensing by nanopores not only to enable the detection and discrimination of single nucleotides, but also to enable the detection of modifications to DNA bases [86–101]. Modifications to nucleotides, such as cytosine methylation play an important role in biological function and malfunction, so such analysis may be vital for a true description of DNA. The second project aimed to develop a novel chemistry for sequencing, wherein an exonuclease enzyme would be coupled to a nanopore and digestion of single stranded DNA by the enzyme would produce a series of individual monophosphate nucleosides in a controlled manner that would pass sequentially through a nanopore detector appropriately modified for recognition of nucleosides [102–108]. The third project focused on critical steps towards the development of high density arrays of individually addressable nanopores to enable scale up of sequencing power, as well as reducing the hardware requirement to a minimum size for portability [109,110]. Overall, the collaborative effort between Oxford Nanopore Technologies, University of Oxford and Delft University of Technology enabled many key steps to be taken in achieving the goals, with Nanopore Sequencing now a commercial reality.

Base distinction including 5meC

With the unique ability to directly sense chemical structure or composition, nanopore detection has the potential to provide

DNA modification analysis. Key mutations to the alpha hemolysin pore ultimately showed that individual bases on intact strands of DNA could be resolved and discriminated when held stationary inside the pore. Through careful optimization, it was demonstrated that not only could the 'standard' four bases be determined, but also 5-methylcytosine (5-mC) and 5-hydroxymethylcytosine (5-hmC) without the need for conversion or labeling. Using this ability during a nanopore sequencing process could radically improve the workflow and accuracy for measurement of DNA modifications in biological systems [102]. These developments to detect DNA were then subsequently shown to be transferable to RNA, where it was demonstrated that there too, nanopores have the unique ability to resolve 'normal' and 'modified' RNA bases on a static strand [103]. In a different approach to base discrimination, it was also shown that a cyclic adaptor attached in the interior of a protein nanopore could detect monophosphate nucleosides in solution. This was achieved after a thorough investigation of the method and site of attachment of a cyclodextrin adaptor at a single point within the pore. In order to achieve recognition of the bases passing through the protein pore α-hemolysin, a particular mutant was shown to be superior [104]. Refinement of nanopore-based measurement techniques were also used to show that not only was this mutant pore capable of accurate recognition and discrimination of the 'standard' four bases of DNA, but it was also capable of distinguishing 5-mC and 5-hmC.

The exonuclease nanopore

The concept of using an exonuclease enzyme to provide individual nucleotides via digestion of a single strand of DNA that are subsequently detected in series by a suitable detector is not new, and was originally conceived in the mid-1980s. However, the original designs suffered from the requirement to label every base as well as to position the enzyme a significant distance from the site of detection, leading to errors from mislabeling and diffusion-driven jumbling of the nucleotide series. The exonuclease-nanopore approach by comparison would remove the need for labels given the ability of the nanopore to identify small molecules directly, and would enable nanometer-scale localization of the enzyme next to the detector through the coupling of the two proteins components together. However, ionic flux measurements in nanopores typically use non-physiological conditions and are incompatible with most enzyme activity. Therefore, it became important to discover suitable enzymes active under nanopore measurement conditions, and once this was achieved, to attach the enzyme to the pore whilst retaining the activities of both enzyme and pore. Many enzymes and mutants were screened before one group was selected for attachment. Novel chemical coupling techniques were developed to enable the efficient placement of the preferred enzymes such that the active site was ~1 nm from the entrance to the pore protein. It was discovered that although enzyme component activity remained after attachment and in the presence of the buffers used in the nanopore measurements, the application of the applied potential across the nanopore, vital for both capture and measurement of the individual bases of DNA, destabilized the enzyme after a relatively short period of time (seconds to minutes) resulting in deactivation. Furthermore, even in the cases where enzyme lifetime was long enough for digestion of 100–200-mer fragments of DNA under an applied potential, no DNA bases were confirmed as detected by the nanopore. Although it was difficult to be certain as to how many bases translocated through the nanopore after digestion, it was likely that the binding and/or detection by the cyclodextrin-modified pore was too inefficient. Further work is required in a number of areas if this approach is to become scientifically and commercially viable [104].

Arrays of nanopores: protein and solid state

One approach taken towards building robust arrays of nanopores was to integrate protein nanopores into apertures drilled into solid state materials. To achieve this, mutant forms of α-hemolysin were conjugated to 3 kbp of double stranded DNA via a 12-base oligonucleotide linker. This complex was driven into a solid state aperture of 2–4 nm diameter fabricated in silicon nitride using an applied potential to create an electrophoretic force. Once localized, the electrical properties of the bionanopore hybrid were analyzed and shown to have similar behavior to hemolysin in lipid bilayer systems. The functionality of the hemolysin was probed by adding ssDNA to the cis side of the bionanopore hybrid and demonstrating the translocation of these strands through the hybrid pore, giving similar characteristics of blockade of current flow and time duration of the block previously observed with hemolysin suspended in lipid bilayers [109]. Another approach for achieving robust, scalable arrays of nanopores was pursued that made use of the novel finding that droplets of water in oil/lipid mixtures can form bilayers at the interface between two adjacent droplets. It was shown that these droplet-interface bilayers are capable of very robustly supporting a nanopore protein and when connected to electrodes, high quality measurements can be made of analytes interacting with the nanopore sensors, even in the presence of complex enzymatic or subcellular activity, such as amplification or transcription. Further work explored the connection of multiple droplets via interfaces with imbedded nanopore channels, and demonstrated that these multisomes are capable of complex connectivity that may provide a path for bottom-up synthesis of cellular systems [110].

Strand sequencing

Critical for the development of devices capable of using these novel nanopore array systems, Oxford Nanopore developed the first highly parallel integrated electronic read-out circuit in silicon chip format. Combining these chips with proprietary nanopore arrays as well as technology to control and measure intact strands of DNA during translocation led ultimately to the presentation at the Advances in Genome Biology and Technology conference in February 2012 of the first nanopore-based sequenced genomes, PhiX and Lambda phages, as well as the unveiling of the company's first product, the small, portable, USB-connectivity enabled device (MinION TM). The core nanopore array architecture is designed to scale and at The American Society of Human Genetics in October 2014 the company displayed its more powerful desktop version (PromethION TM) with nearly 150,000 nanopore arrays. The MinION nanopore sequencers are currently being used in many laboratories around the world ((Fig. 5) a screenshot of a MinION run).

In situ *DNA analysis methods*

Most high-throughput molecular analysis techniques use tissue homogenates as substrates. By doing so, an average molecular state

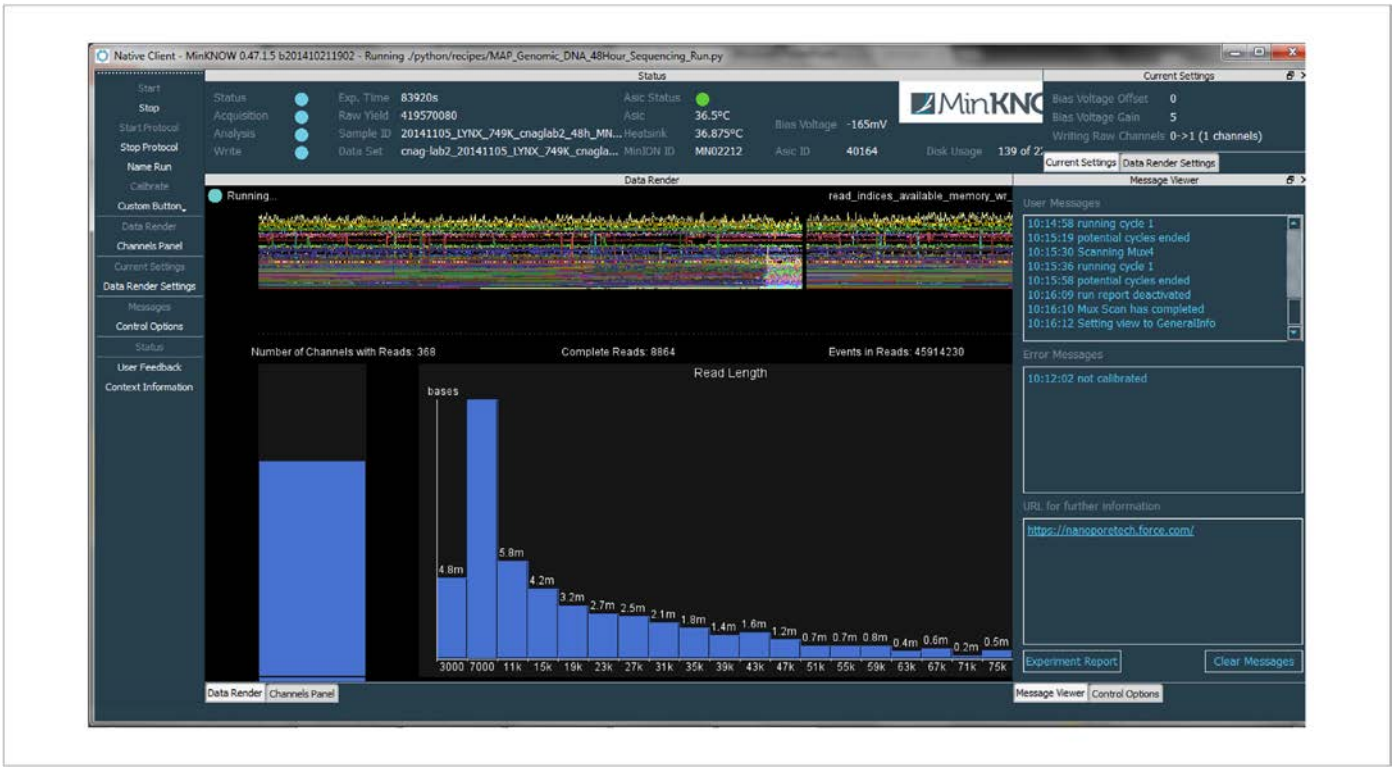

FIGURE 5

Screenshot of an Oxford Nanopore MinIon during a sequencing run. At the moment of this screenshot 75% of the nanopores were sequencing. Readlengths of several 10 kbs were achieved. Squiggles of all nanopores are displayed. Squiggles are then converted into sequences by submission to a cloud server running the Metrichor software.

is measured providing information about the molecular composition of the average cell in the tissue. The problem with this approach is that it does not make sense in tissues composed of different cell types, which is the case for almost all tissues in a mammalian organism, except on the DNA level. Some of these tissue cell-types will have common molecular components, while others will have unique ones. And regardless of these qualitative differences or similarities, there will be quantitative differences between cells of different developmental origin, and in similar cells differences may occur due to differences in the local environment. The only way to decompose this mix of different molecular states is to do cellular resolved analysis, such as single-cell analysis, and histology based *in situ* analyses, also linking measurements of abundance of molecules with spatial information. In READNA, we have developed *in situ* techniques that enable NGS *in situ*, a contextual sequencing approach that may be referred to as 4th generation sequencing [111–123].

In situ genotyping and sequencing

A technique that enables *in situ* genotyping and sequencing of transcripts in fixed cells and tissues has been developed. The approach is based on target-primed Rolling Circle Amplification (RCA) of padlock probes and can be used to read barcodes in the probes that correspond to a certain mRNA sequence or sequence variants, or read short stretches of native mRNA sequence. The RCA is used to generate clonally amplified sequencing substrates of specifically circularized padlock probes, and the cDNA strand is used as primer for this amplification to preserve the localization of the amplification product. Thus, it is important that there is a free 3′end close to the padlock probe binding site. A strategy to cut the target strand in a site specific manner using a combination of MutY endonuclease and AP-lyase activities has been developed [112]. The basic general method to detect mRNA sequences and single nucleotide variants thereof was presented by Larsson *et al.* [113]. We have also developed assays to detect the seven most common mutations in codons 12 and 13 of the KRAS transcript, as well as mutations in the EGFR and TP53 genes [114]. This approach to detect somatic mutations in tumor tissue could be very attractive to clinical diagnostics use since it can be performed directly in the tissue section used for pathology examination and scoring (Fig. 6).

We have further developed the *in situ* genotyping approach to allow *in situ* sequencing. To achieve sequence information from the *in situ* synthesized cDNA, we replaced the regular padlock probe with a gap-fill padlock probe that introduces cDNA sequence into the RCA products (Fig. 7). The RCA products were then sequenced by ligation, essentially according to the protocol published by Shendure *et al.* [115]. We have successfully sequenced four bases in the human and mouse β-actin transcript (ACTB) in cell lines and also HER2 and ACTB transcripts in a HER2 positive breast cancer tissue [116]. This is the first study that succeeds with sequencing RNA directly in intact tissue sections, linking sequences to the location of the molecules with micrometer resolution. We also demonstrated by sequencing codon 12 of the KRAS transcript, that five KRAS mutant cells could be detected in the background of 5000 KRAS wild-type cells.

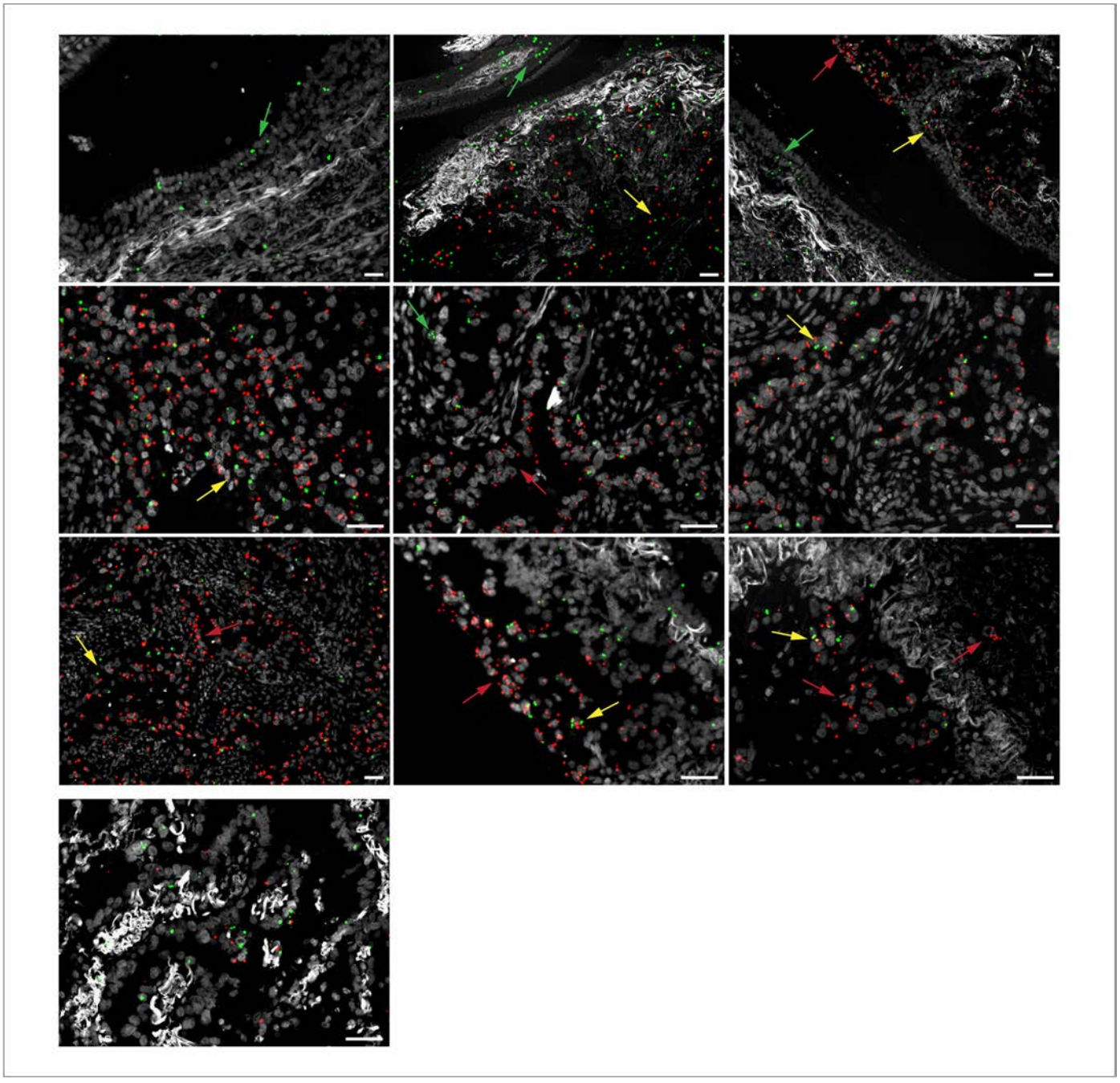

FIGURE 6

In situ detection of EGFR L858R point mutation in FFPE lung cancer tissue, visualizing intra-tumoral genetic heterogeneity. Fluorescence microscopy images from nine fields of views (20× magnification) are shown displaying the results of a padlock probe and RCA based *in situ* scoring of a point mutation in the EGFR gene. Probes for the wild-type transcript generate green fluorescence signals, while probes for the mutant transcript generate red signals. Nuclear staining and auto fluorescence are displayed in grey. The top left image shows normal lung tissue, while the other images contain more or less cancer tissue. Green arrows highlights areas with predominantly wild-type transcripts, yellow arrows mixed staining, and red arrows predominantly mutant transcripts.

In situ expression profiling

The *in situ* sequencing approach described was also applied to readout multiplex padlock probe reactions to achieve expression profiles *in situ*. We designed padlock probes targeting 39 transcripts to study their expression and localization in tissue sections with microscopic resolution. The transcripts were selected to be expressed in breast cancer tissue and included 21 transcripts that are used in a breast cancer prognostic expression panel (OncoType DX). We applied all probes in one reaction and determined the expression pattern of each transcript by sequencing their unique four-base long barcodes *in situ* (Fig. 8). Gene expression profiling was performed on three HER2 positive fresh frozen breast cancer tissue sections that were fixed on microscope slides. After filtering for base-calling quality, we were able to extract reliable expression data from 24 transcripts, requiring a frequency of detection that was higher than the most abundant unexpected barcode read to

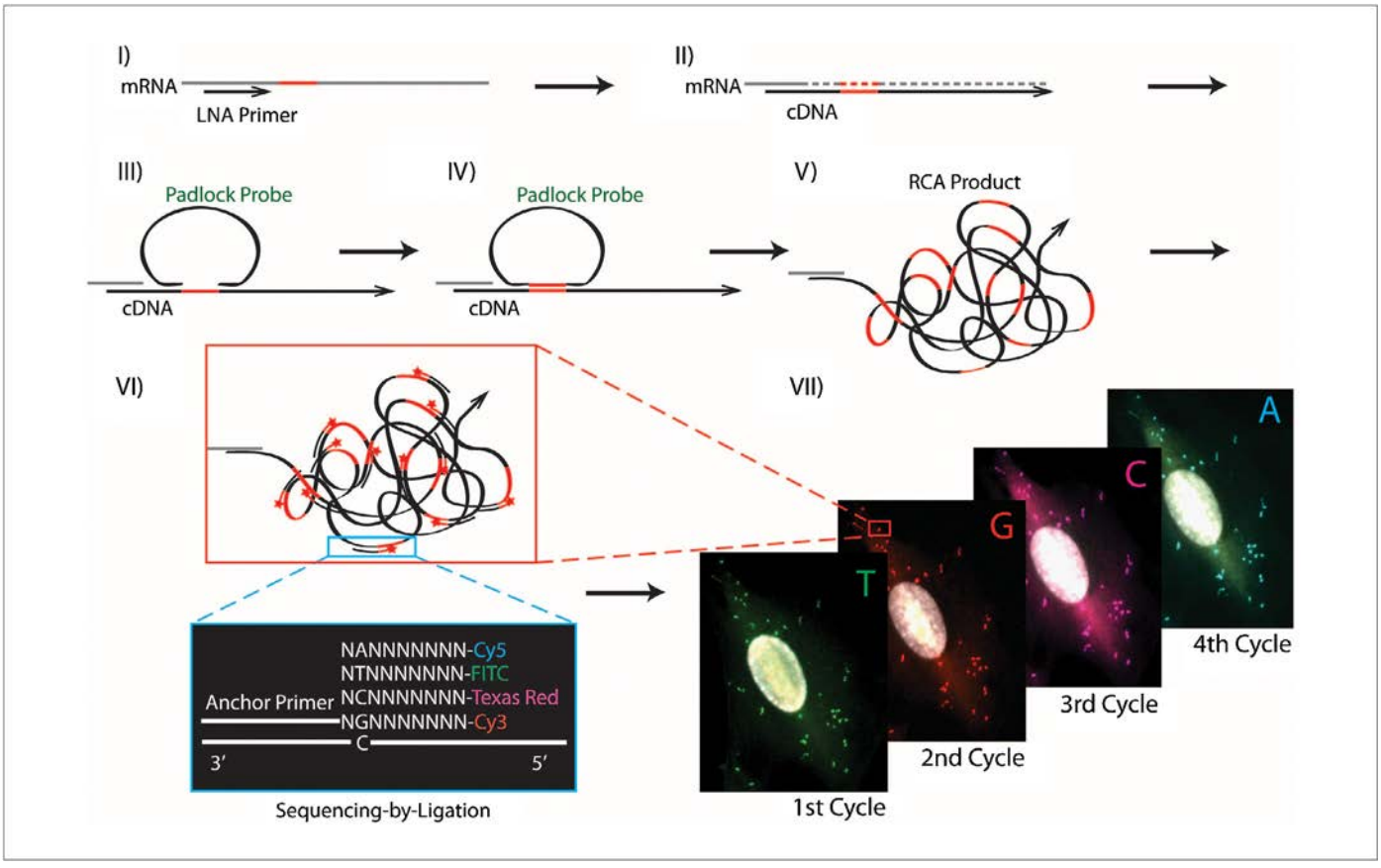

FIGURE 7

Schematic illustration of *in situ* sequencing of RNA. (I) cDNA is synthesized using locked nucleic acid (LNA)-modified primers or random decamers. (II) The RNA strand is degraded by RNase H, (III) followed by hybridization of a padlock probe, which is designed such that a gap between the two ends is formed after the hybridization. (IV) The gap, which is the target for sequencing, is then filled by DNA polymerization. The 3′ end and 5′ end are then joined by DNA ligase to form a completed DNA circle. (V) Target primer RCA is performed to clonally amplify the DNA circle and generate RCA products, which are then subjected for sequencing by ligation. (VI) The anchor primer is hybridized right next to the target on the 3′ end, followed by ligation of four interrogation probes, consisting of eight random positions and one specific. The interrogation probes carry four different fluorophores and the best matching probe will be ligated to the anchor probe, thus incorporating fluorophores to the RCA products in a base-specific manner. (VII) Finally, the cells and RCA products are imaged and base calling is performed based on which fluorophore is recorded for each RCA product in that cycle. Steps VI and VII are repeated using interrogation probes with specific bases at a different position as many times as needed to achieve a certain read-length. Sequences are generated by aligning the base-calling data from the different sequencing cycles.

consider it reliable. The location of the sequencing reads correlated well with the expected expression patterns of the corresponding genes, and overall the number of *in situ* sequencing reads correlated exceptionally well with the number of RNA sequencing reads from breast-cancer derived cell lines for the HER2 positive cells and with normal breast tissue for the vimentin positive cells, defining the cells surrounding the cancer cells [116].

Improvement of protocols

Many different protocols have been proposed since 2nd generation sequencers were introduced. However, many of the presented protocols have shortcomings. Within READNA we re-worked and optimized many protocols and set them up for potential commercialization.

DNA-seq

We noticed that the commercially offered sample preparation procedures for Illumina sequencing had strong GC-bias, with good representation of sequences in the medium GC-composition range, while representation of high and low GC-content reads was weak. This was so marked that we found entire genes that were not sufficiently covered in human whole genome sequencing for reliable variant or somatic mutation calling, for example, NOTCH1 gene (Fig. 9). Upon investigation of this, we developed a PCR-free sample preparation that was combined with a bead-based affinity purification procedure [137].

RNA-Seq

Another development within READNA has been the modification of the RNA-Seq protocol to determine the polarity of transcripts, important for correct annotation of novel genes and providing essential information about the potential function of the gene. The simple modification involves incorporation of deoxy-UTP during the second strand cDNA synthesis and subsequent destruction of the uridine-containing strand in the sequencing library for correct identification of the orientation of transcripts [119].

Optimization of whole genome methylation pipelines

5-Methyl-cytosine analysis is achieved by converting non-methylated cytosines to uracile with bisulphite followed by sequencing.

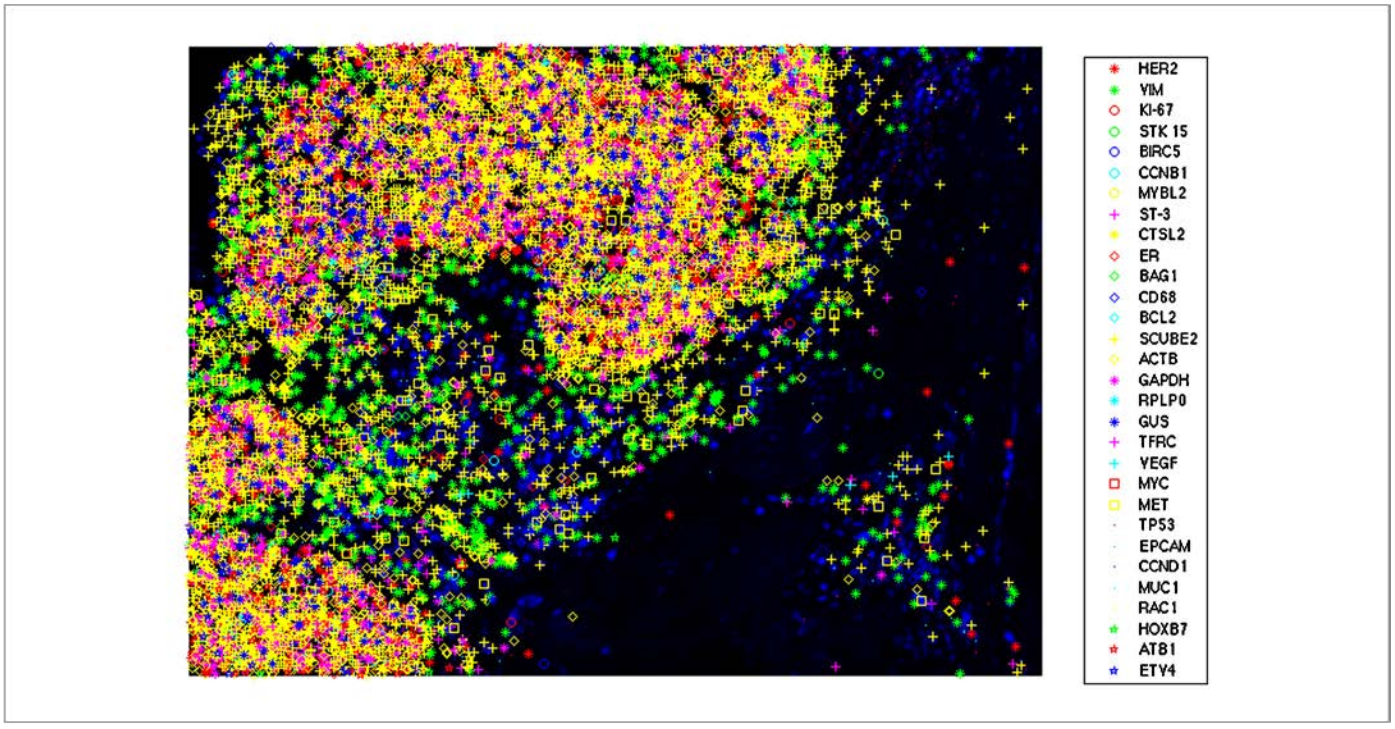

FIGURE 8

In situ gene expression profiling in a breast cancer tumor section. A 0.6 mm^2 sized area of a breast cancer tumor tissue section was analyzed using a pool of padlock probes targeting 39 different transcripts. The picture shows the distribution of the sequencing reads from the 30 most abundant transcripts, listed to the right of the picture.

We developed this into a process for whole genome 5-methylcytosine analysis with 2nd generation sequencing [138,139]. The process is based on preparing a sequencing library, followed by the inclusion of a conversion control spike-in, bisulphite conversion, library quality control system, a sequencing procedure that allows balancing the uneven response of the four bases after bisulphite conversion in Illumina sequencing instruments. We integrated the pipeline with a data analysis pipeline in which all spike-ins are tracked automatically. This pipeline is applied in the EU-funded project of the International Human Epigenome Project BLUEPRINT.

Other applications of nucleic acid analysis – ProteinSeq

READNA has developed a method entitled ProteinSeq that measures in parallel candidate protein biomarkers in many samples. A multiplex proximity ligation assay (PLA) is performed and the readout is done using realtime PCR or DNA sequencing (ProteinSeq). We demonstrate improved sensitivity over conventional sandwich assays for simultaneous analysis of sets of 35 proteins in 5 ml of blood plasma. Importantly the method can be used with multiplexing as background signal remains low. The level of multiplexing that is possible will be investigated further. ProteinSeq was used to analyze proteins in plasma samples from cardiovascular disease (CVD) patient cohorts and matched controls. Three proteins, namely P-selectin, Cystatin-B and Kallikrein-6, were identified as putative diagnostic biomarkers for CVD. The latter two have not been previously reported in the literature and their potential roles in CVD will be validated in larger patient cohorts. ProteinSeq has a potential for screening large numbers of proteins and samples while the technology can provide a much-needed platform for validation of diagnostic markers in biobank samples and in clinical use [121,122]. In a recent publication [123] Proximity Ligation Assay (PLA) has been proven to be a robust protein detection method. The technique is characterized by high sensitivity and specificity, but the assay precision is probably limited by the PCR readout. To investigate this potential limitation and to improve precision, we developed a digital PLA for protein measurement in fluids based on amplified single molecule detection (ASMD). The assay showed significant improvements in precision, and thereby also detection sensitivity, over the conventional real-time PCR readout. The assay is complementary to the ProteinSeq method also developed within READNA.

Software development

The sequence output of 2nd generation sequencers puts huge strain on computation. The main reasons are that the sequencing reads are short and that it is easy to produce huge numbers of short reads with 2nd generation sequencers. Within READNA we have developed software to respond to several challenges for the analysis of NGS data.

Alignment of 2nd generation sequencing reads

Sequence alignment is one of the first steps that needs to be executed in a data analysis pipeline of NGS data. The millions of reads generated in a sequencing run need to be positioned on a reference sequence. Matching sequences that are identical to the reference and are unique in the genome does not pose particular challenges. However, the objective is to capture small differences to the reference sequence. Another problem of many genomes is that in genomes, Ts, Cs, Gs, and As are not randomly distributed,

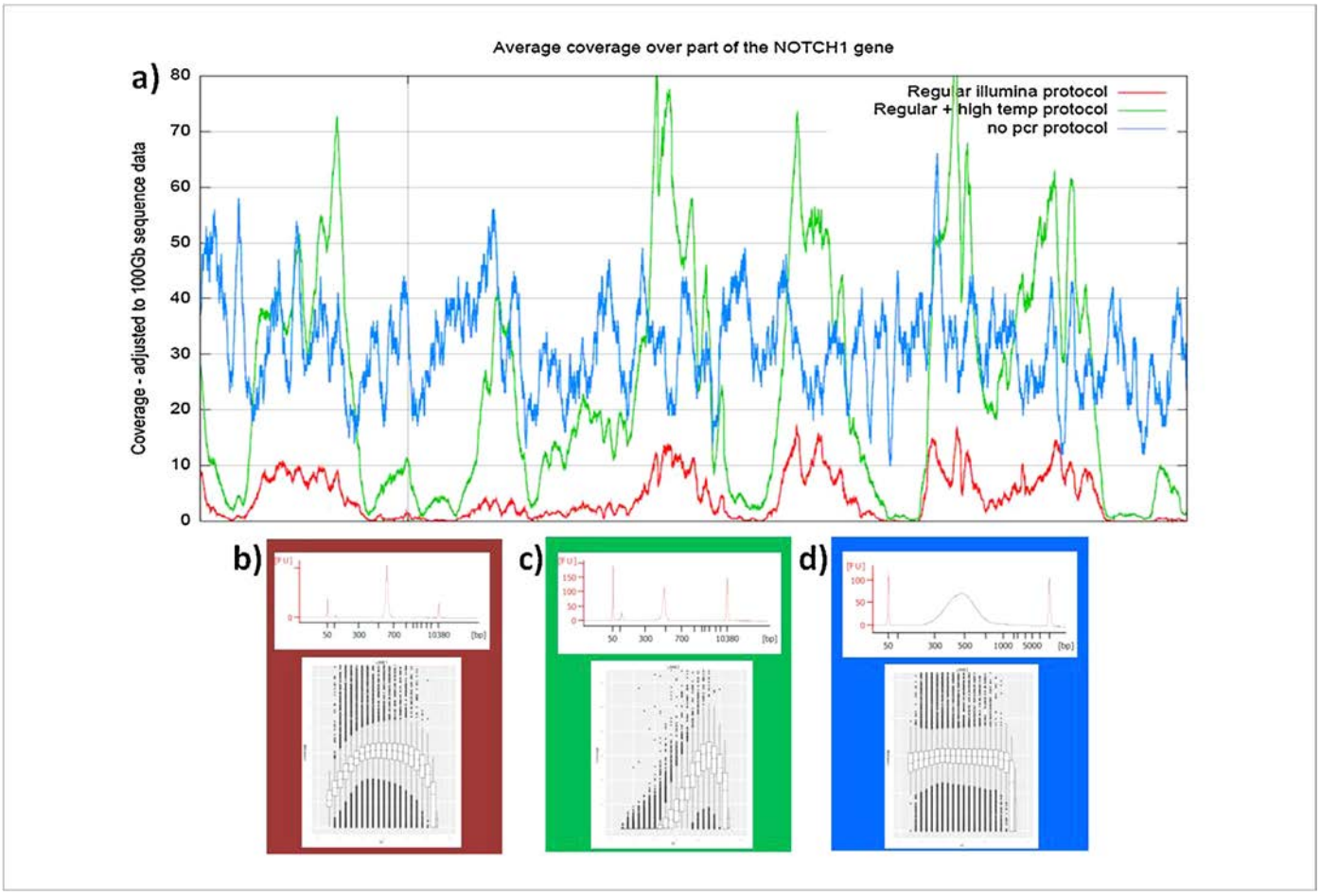

FIGURE 9

Sequence coverage of 30× whole genome sequencing across the NOTCH1 gene (panel **(a)**). Three different protocols were used, standard Illumina protocol in red, a modified version of the Illumina protocol that enhances GC-rich sequences in green, and a no PCR protocol in blue. Even coverage like in the blue protocol dramatically facilitates variant calling. In panel **(b)**, **(c)** and **(d)** are the Bioanalyzer traces and normalized GC representations for the three sample preparation protocols, corresponding to the coverage plots in panel (a).

but have many repetitive elements, duplications, similar sequences in different regions of the genome. This substantially complicates finding the best place to position a read. Exhaustive searches are computationally very expensive. A common feature of many aligners is the use shortcuts from exhaustive searching, by taking random decisions. With GEM a deterministic aligner was developed that explores the entire search space exhaustively while on the other hand being computationally very efficient and accurate [124].

Variant calling

After alignment, genomic variants are called from the consensus of aligned sequences. As implied by the alignment difficulties described above, the determination of true variants is not always straightforward. The development of two new variant calling tools/approaches was primarily motivated by our evaluation of alignment and variant calling tools, which revealed five important observations. Firstly, variant detection in enriched regions – even at high coverage – depends strongly on the chosen tools and their settings. Secondly, the tested variant callers do not seem to be well-trained to handle targeted NGS (T-NGS) data, leading to a significant fraction of false positive and negative SNV calls [14]. Thirdly, when analysing T-NGS sequences, the conventional (or one-step whole-genome) mapping and variant-calling approach is not optimal in terms of time and computer resources [14]. Fourthly, depending on sequencing platforms, up to 60% of relevant variants were undetected due to coverage gaps, or were misidentified [125]. Fifthly, when comparing two samples (e.g. healthy versus affected tissue), reliance on the derived variant lists instead of the primary BAM file will not allow detection of significant changes of allelic balance in heterozygous variants, unless the read counts per allele are included in the variant lists. The results of our studies showed that lack of evenness of coverage and inadequate enrichment and/or coverage can prevent the detection of real nucleotide variants, leading to higher false negative rates, particularly for heterozygotes [14]. These observations and analysis challenges emphasize that the costs, speed, and variant detection accuracy should be considerably adapted and improved in order to be useful in a diagnostic setting and affordable to a larger group of researchers. To address these challenges, we have developed two novel software tools, namely 'backmapping' and 'pibase'. The backmapping pipeline [14] (http://www.ikmb.uni-kiel.de/tngs-backmapping) is a two-step mapping approach, target-region mapping (to rescue false negative calls) with subsequent

read-backmapping of only the reads that cover a detected variant, to the whole genome. In an exome sequencing benchmark, our backmapping approach achieved two-fold faster mapping and increased sensitivity, compared to the conventional approach. The second novel software package, pibase (http://www.ikmb.uni-kiel.de/pibase), was developed to help distinguish between true variants, artifacts, and sequence coverage gaps. This is a tertiary *in silico* analysis tool, for interrogating BAM files, re-typing SNPs, and accurate comparison of variant data [125]. Using pibase, we demonstrated highly specific (99.97–100.00%) genotype calls with publicly available 1000 Genomes Project BAM files. We also reported that the false detection rate of variant differences in pairs of monozygotic twins is 10-fold lower using pibase's Fisher's exact test, than using a genotype-based comparison method [125]. As large-scale manual inspection of T-NGS or exome data is not feasible and prone to subjective errors, we recommend the two-step mapping approach; backmapping; [14] and validations with the pibase tool [125]. As pseudo-random technical artifacts may occur during library preparation, enrichment or sequencing, we also recommend, if cost allows, duplicating samples and validating computed variants by checking the concordance between these technical replicates.

RNA-Seq

The community opinion is that RNA profiling using arrays or sequencing is not robust and experiments carried out in different laboratories are not comparable and even experiments done in the same laboratory at different times or by different people are prone to variation. We proved that an RNA sequencing project can be distributed across several laboratories as long as a standard operating procedure is adhered to and the same lots of reagents are used. Having an RNA-Seq dataset with 500 samples in hand to work with has allowed the establishment of many different data analytical techniques for RNA-Seq [126,127].

DNA methylation software, MeDIP

DNA methylation occurring on cytosines in the context of the dinucleotide sequence CpG forms one of the multiple layers of epigenetic mechanisms which constitute the regulatory landscape of a cell. Aberrant DNA methylation changes have been detected in several diseases, particularly cancer and autoimmune diseases leading to the development of a large number of technologies for the analysis of DNA methylation patterns. Enrichment of the methylated fraction of the genome by Methylated DNA Immuno-Precipitation (MeDIP) and subsequent NGS sequencing provides a cost-effective alternative for genome-wide DNA methylation analysis and is not confounded as WGBS by the presence of hydroxymethylation as the antibodies are specific for the respective cytosine modifications. However, the exact nature of the immunoprecipitated products has never been investigated. A novel method for the analysis of the DNA methylation state of the immunoprecipitated population of molecules at single nucleotide resolution was developed [128]. A small part of the immunoprecipitated fraction was treated with sodium bisulfite thereby translating epigenetic differences into sequence differences which were then quantitatively read-out by pyrosequencing, demonstrating that MeDIP was highly specific for methylated molecules. The assay presents a rapid and cost-effective method for the quality control of MeDIP prior to NGS. As no complete bioinformatics solution for the processing of MeDIP-seq data was available, a complete pipeline for quality control, pre-processing, mapping, estimation of DNA methylation levels and calling of differentially methylated regions was developed for both single end and paired-end reads [129]. The freely available pipeline permits to streamline the data analysis process and enables researchers to concentrate on the biological interpretation of MeDIP-seq data.

ELSI aspects of next generation nucleic acid sequence data

Innovations in sequencing mean that more detailed and richer sequence information can be generated in larger quantities, at a cheaper cost and faster than ever before. One sample of DNA that is sequenced using NGS can provide information potentially on a number of conditions at once rather than separate tests having to be done for each condition. This raises a number of new ethical, legal and social issues (ELSI) as this technology starts to be more widely used in research and rolled out to the clinic. For example, the possibility of identifying an incidental finding of clinical significance that was not the object of the original research and diagnostic investigation increases, because of the quality and quantity of the information generated through NGS [130,131]. It is also increasingly difficult to guarantee the privacy of such information for the proband and their relatives because of the comprehensive nature of the data, the fact that DNA is a unique identifier and this technology is likely to be accessible to more people [132]. Obtaining consent for research is also difficult if data has a number of possible uses that cannot always be anticipated at the time of recruitment into research or when presenting for a particular condition [133]. One question which arises as a result of this is whether an individual should be informed of all of the potential health consequences connected with the variations in their genome and how this should be done. For this technology to have greater clinical application, it will also need to become more reliable and to be more easily replicated [134,135].

Perspectives – what remains to be done? What are future challenges?

READNA has made major contributions to many aspects of nucleic acid analysis and its applications. Many of the developments reached maturity already during the course of READNA and are now used in different contexts. We have shown many concepts that can now be translated for research and application. Interesting developments are in nanopore sequencing, single cell nucleic acid applications and the work on devices for handling and analyzing long DNA molecules, and in NGS data analysis. However, there are many aspects that do not yet have a final solution. Extending the length of DNA sequences is critical. The more we refine the tools for genome analysis, the more we find that structures of genomes are more diverse than we thought. We are discovering that the concept of a single reference for a species is not entirely correct. Technologies to analyze long DNA molecules at high resolution need further development. In general, technologies to carry out nucleic acid analyses at very low concentration still require further development. In the ultimate limit this applies to single cell analysis at DNA, RNA, DNA modifications and DNA protein interaction level. Understanding the functioning of the genome requires such techniques, as stochastic effects of

doing these analyses with DNA from many cells results in a loss of molecular context. A further opportunity that would be very beneficial and would advance the integration of these technologies into routine is the development of fully integrated cartridge-based systems in which procedures can be done far more reliably, reproducibly and with a dramatically reduced risk of contamination. Sequencing with nanopores is very enticing. Many aspects of nanopore sequencing still require resolution, such as the robustness of the pores, software for deconvoluting signal and software to deal with long reads. However, the community-based tool development initiated by Oxford Nanopore Technologies for their MinION nucleic acid sequencer is very innovative. 4th generation sequencing for which READNA has provided a first proof of concept still has a great need for further development. Currently, this technology can be used for *in situ* RNA analysis, but evidently techniques for measuring somatic mutations on the level of the DNA *in situ* of many different targets simultaneously would be of interest and being able to determine the interaction of regulatory elements *in situ* would allow studying genome regulation at unprecedented resolution [1]. There is also a great need for standardization and benchmarking of 2nd generation sequencing and data analysis for this technology to achieve robustness and get the deserved uptake into clinical practice. Care needs to be taken that good standards of data analysis are installed in common practice. An eye needs to be kept on computational efficiency of analysis and storage as large-scale projects are initiated.

Conclusion

READNA has contributed substantially to many aspects of the revolution that nucleic acid sequencing has experienced in recent years, has opened many new lines of work and provided proofs-of-concepts. The body of work stands as a lasting legacy.

Acknowledgements

The review was written and funded through the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. [HEALTH-F4-2008-201418] entitled READNA. We would like to thank Professor Sunney Xie, Professor Stephan Beck, Professor Wilhelm Ansorge and Professor George Church, distinguished members of the READNA Scientific Advisory Board for their critical review of the READNA project. Finally we would like to thank Dr. Tomasz Dylag project officer at the European Commission for his support throughout the duration of the project.

References

[1] Mir KU. Sequencing genomes: from individuals to populations. Brief Funct Genomic Proteomic 2009;8(5):367–78.
[2] Parkhomchuk D, Amstislavskiy V, Soldatov A, Ogryzko V. Use of high throughput sequencing to observe genome dynamics at a single cell level. Proc Natl Acad Sci U S A 2009;106(49):20830–35.
[3] McGinn S, Gut IG. DNA sequencing – spanning the generations. New Biotechnol 2013;30(4):366–72.
[4] Veal CD, Xu H, Reekie K, Free R, Hardwick RJ, McVey D, et al. Automated design of paralogue ratio test assays for the accurate and rapid typing of copy number variation. Bioinformatics 2013;29(16):1997–2003.
[5] Veal CD, Freeman PJ, Jacobs K, Lancaster O, Jamain S, Leboyer M. A mechanistic basis for amplification differences between samples and between genome regions. BMC Genomics 2012;13:455.
[6] Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. Nucleic Acids Res 2005;33(8):e71.
[7] Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M, et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci U S A 2007;104(22):9387–92.
[8] Johansson H, Isaksson M, Sorqvist EF, Roos F, Stenberg J, Sjoblom T, et al. Targeted resequencing of candidate genes using selector probes. Nucleic Acids Res 2011;39(2):e8.
[9] Zieba A, Grannas K, Soderberg O, Gullberg M, Nilsson M, Landegren U. Molecular tools for companion diagnostics. Nat Biotechnol 2012;29(6):634–40.
[10] Moens LN, Falk-Sorqvist E, Ljungstrom V, Mattsson J, Sundstrom M, La Fleur L, et al. HaloPlex targeted resequencing for mutation detection in clinical formalin-fixed paraffin-embedded (FFPE) tumor samples. Mol Diagn 2015;17(6):729–39.
[11] Meuzelaar LS, Lancaster O, Pasche JP, Kopal G, Brookes AJ. MegaPlex PCR: a strategy for multiplex amplification. Nat Methods 2007;4(10):835–7.
[12] Querfurth R, Fischer A, Schweiger MR, Lehrach H, Mertes F. Creation and application of immortalized bait libraries for targeted enrichment and next-generation sequencing. Biotechniques 2012;52(6):375–80.
[13] Mir KU, Qi H, Salata O, Scozzafava G. Sequencing by Cyclic Ligation and Cleavage (CycLiC) directly on a microarray captured template. Nucleic Acids Res 2009;37(1):e5.
[14] Elsharawy A, Forster M, Schracke N, Keller A, Thomsen I, Petersen BS, et al. Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing. BMC Genomics 2012;13:417.
[15] ElSharawy A, Warner J, Olson J, Forster M, Schilhabel MB, Link DR, et al. Accurate variant detection across non-amplified and whole genome amplified DNA using targeted next generation sequencing. BMC Genomics 2012;13:500.
[16] Mertes F, Elsharawy A, Sauer S, van Helvoort JM, van der Zaag PJ, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics 2011;10(6):374–86.
[17] Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R. Exome sequencing: dual role as a discovery and diagnostic tool. Ann Neurol 2012;71(1):5–14.
[18] Fairfield H, Gilbert GJ, Barter M, Corrigan RR, Curtain M, Ding Y, et al. Mutation discovery in mice by whole exome sequencing. Genome Biol 2011;12(9):R86.
[19] Melum E, May S, Schilhabel MB, Thomsen I, Karlsen TH, Rosenstiel P, et al. SNP discovery performance of two second-generation sequencing platforms in the NOD2 gene region. Hum Mutat 2010;31(7):875–85.
[20] Nilsson, M., et al., Target Enrichment Sequencing Descriptors (TESD) – descriptors for reporting target enrichment in DNA sequencing experiments, manuscript in preparation.
[21] Mauger F, Jaunay O, Chamblain V, Reichert F, Bauer K, Gut IG, et al. SNP genotyping using alkali cleavage of RNA/DNA chimeras and MALDI time-of-flight mass spectrometry. Nucleic Acids Res 2006;34(3):e18.
[22] Mauger F, Bauer K, Calloway CD, Semhoun J, Nishimoto T, Myers TW, et al. DNA sequencing by MALDI-TOF MS using alkali cleavage of RNA/DNA chimeras. Nucleic Acids Res 2007;35(8):e62.
[23] Mauger F, Bauer K, Semhoun J, Myers TW, Gelfand DH, Gut IG. Ribo-polymerase chain reaction – a facile method for the preparation of chimeric RNA/DNA applied to DNA sequencing. Hum Mutat 2012;33(6):1010–5.
[24] Mauger F, Gelfand DH, Gupta A, Bodepudi V, Will SG, Bauer K, et al. High-specificity single-tube multiplex genotyping using Ribo-PAP PCR, tag primers, alkali cleavage of RNA/DNA chimeras and MALDI-TOF MS. Hum Mutat 2013;34(1):266–73.
[25] Sauer S, Freiwald A, Maier T, Kube M, Reinhardt R, Kostrzewa M, et al. Classification and identification of bacteria by mass spectrometry and computational analysis. PLoS One 2008;3(7):e2843.
[26] Freiwald A, Sauer S. Phylogenetic classification and identification of bacteria by mass spectrometry. Nat Protoc 2009;4(5):732–42.
[27] Freiwald A, Mao L, Kodelja V, Kliem M, Schuldt D, Schreiber S, et al. Differential analysis of Crohn's disease and ulcerative colitis by mass spectrometry. Inflamm Bowel Dis 2011;17(4):1051–2.
[28] Sauer S, Kliem M. Mass spectrometry tools for the classification and identification of bacteria. Nat Rev Microbiol 2010;8(1):74–82.
[29] Kliem M, Sauer S. The essence on mass spectrometry based microbial diagnostics. Curr Opin Microbiol 2012;15(3):397–402.
[30] Howell WM, Jobs M, Gyllensten U, Brookes AJ. Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. Nat Biotechnol 1999;17(1):87–8.
[31] Jobs M, Howell WM, Stromqvist L, Mayr T, Brookes AJ. DASH-2: flexible, low-cost, and high-throughput SNP genotyping by dynamic allele-specific hybridization on membrane arrays. Genome Res 2003;13(5):916–24.
[32] Rasmussen KH, Marie R, Lange JM, Svendsen WE, Kristensen A, Mir KU. A device for extraction, manipulation and stretching of DNA from single human chromosomes. Lab Chip 2011;11(8):1431–3.
[33] Bauer DL, Marie R, Rasmussen KH, Kristensen A, Mir KU. DNA catenation maintains structure of human metaphase chromosomes. Nucleic Acids Res 2012;40(22):11428–34.
[34] Eriksen J, Thilsted AH, Marie R, Luscher CJ, Nielsen LB, Svendsen WE, et al. Dynamic in situ chromosome immobilisation and DNA extraction using

localized poly(N-isopropylacrylamide) phase transition. Biomicrofluidics 2011;5(3): 31101–311014.
[35] Marie R, Kristensen A. Nanofluidic devices towards single DNA molecule sequence mapping. J Biophotonics 2012;5(8):673–86.
[36] Marie R, Pedersen JN, Bauer DL, Rasmussen KH, Yusuf M, Volpi E, et al. Integrated view of genome structure and sequence of a single DNA molecule in a nanofluidic device. Proc Natl Acad Sci U S A 2013;110(13):4893–8.
[37] Freitag C, Noble C, Fritzsche J, Persson F, Reiter-Schad M, Nilsson AN, et al. Visualizing the entire DNA from a chromosome in a single frame. Biomicrofluidics 2015;9:044114.
[38] Reisner W, Larsen NB, Silahtaroglu A, Kristensen A, Tommerup N, Tegenfeldt JO. Single-molecule denaturation mapping of DNA in nanofluidic channels. Proc Natl Acad Sci U S A 2010;107(30):13294–99.
[39] Thamdrup LH, Pedersen JN, Flyvbjerg H, Larsen NB, Kristensen A. Nanoimprinted polymer chips for light induced local heating of liquids in micro- and nanochannels. Proc SPIE 2010;7764. 77640I-77640I-13.
[40] Thamdrup LH, Larsen NB, Kristensen A. Light-induced local heating for thermophoretic manipulation of DNA in polymer micro- and nanochannels. Nano Lett 2010;10(3):826–32.
[41] Pedersen JN, Luscher CJ, Marie R, Thamdrup LH, Kristensen A, Flyvbjerg H. Thermophoretic forces on DNA measeured with a single-molecule spring balance. Phys Rev Lett 2014;113:268301.
[42] Mikkelsen MB, Reisner W, Flyvbjerg H, Kristensen A. Pressure-driven DNA in nanogroove arrays: complex dynamics leads to length- and topology-dependent separation. Nano Lett 2011;11(4):1598–602.
[43] Nyberg LK, Persson F, Berg J, Bergstrom J, Fransson E, Olsson L, et al. A single-step competitive binding assay for mapping of single DNA molecules. Biochem Biophys Res Commun 2012;417(1):404–8.
[44] Persson F, Fritzsche J, Mir KU, Modesti M, Westerlund F, Tegenfeldt JO. Lipid-based passivation in nanofluidics. Nano Lett 2012;12(5):2260–5.
[45] Lymperopoulos K, Crawford R, Torella JP, Heilemann M, Hwang LC, Holden SJ, et al. Single-molecule DNA biosensors for protein and ligand detection. Angew Chem Int Ed Engl 2010;49(7):1316–20.
[46] Persson F, Westerlund F, Tegenfeldt JO. Fluorescence nanoscopy of single DNA molecules by using stimulated emission depletion (STED). Angew Chem Int Ed Engl 2011;50(24):5581–3.
[47] Mortensen KI, Churchman LS, Spudich JA, Flyvbjerg H. Optimized localization analysis for single-molecule tracking and super-resolution microscopy. Nat Methods 2010;7(5):377–81.
[48] Persson F, Westerlund F, Tegenfeldt JO. Fluorescence microscopy of nanochannel-confined DNA. Methods Mol Biol 2011;783:159–79.
[49] Hannestad JK, Brune R, Czolkos I, Jesorka A, El-Sagheer AH, Brown T, et al. Kinetics of diffusion-mediated DNA hybridization in lipid monolayer films determined by single-molecule fluorescence spectroscopy. ACS Nano 2013;7(1):308–15.
[50] Persson F, Tegenfeldt JO. DNA in nanochannels – directly visualizing genomic information. Chem Soc Rev 2010;39(3):985–99.
[51] Werner E, Persson F, Westerlund F, Tegenfeldt JO, Mehlig B. Orientational correlations in confined DNA. Phys Rev E Stat Nonlin Soft Matter Phys 2012;86(4 Pt 1):041802.
[52] Yusuf M, Bauer DL, Lipinski DM, MacLaren RE, Wade-Martins R, Mir KU, et al. Combining M-FISH and Quantum Dot technology for fast chromosomal assignment of transgenic insertions. BMC Biotechnol 2011;11:121.
[53] Hohlbein J, Gryte K, Heilemann M, Kapanidis AN. Surfing on a new wave of single-molecule fluorescence methods. Phys Biol 2010;7(3):031001.
[54] Richardson JA, Gerowska M, Shelbourne M, French D, Brown T. Six-colour HyBeacon probes for multiplex genetic analysis. Chembiochem 2010;11(18):2530–3.
[55] Rindermann JJ, Akhtman Y, Richardson J, Brown T, Lagoudakis PG. Gauging the flexibility of fluorescent markers for the interpretation of fluorescence resonance energy transfer. J Am Chem Soc 2011;133(2):279–85.
[56] Hohlbein J, Aigrain L, Craggs TD, Bermek O, Potapova O, Shoolizadeh P, et al. Conformational landscapes of DNA polymerase I and mutator derivatives establish fidelity checkpoints for nucleotide insertion. Nat Commun 2013;4:2131.
[57] Santoso Y, Joyce CM, Potapova O, Le Reste L, Hohlbein J, Torella JP, et al. Conformational transitions in DNA polymerase I revealed by single-molecule FRET. Proc Natl Acad Sci U S A 2010;107(2):715–20.
[58] Santoso Y, Torella JP, Kapanidis AN. Characterizing single-molecule FRET dynamics with probability distribution analysis. Chemphyschem 2010;11(10):2209–19.
[59] Torella JP, Holden SJ, Santoso Y, Hohlbein J, Kapanidis AN. Identifying molecular dynamics in single-molecule FRET experiments with burst variance analysis. Biophys J 2011;100(6):1568–77.
[60] Le Reste L, Hohlbein J, Gryte K, Kapanidis AN. Characterization of dark quencher chromophores as nonfluorescent acceptors for single-molecule FRET. Biophys J 2012;102(11):2658–68.
[61] Holden SJ, Uphoff S, Hohlbein J, Yadin D, Le Reste L, Britton OJ, et al. Defining the limits of single-molecule FRET resolution in TIRF microscopy. Biophys J 2010;99(9):3102–11.
[62] Uphoff S, Gryte K, Evans G, Kapanidis AN. Improved temporal resolution and linked hidden Markov modeling for switchable single-molecule FRET. Chemphyschem 2011;12(3):571–9.
[63] Uphoff S, Holden SJ, Le Reste L, Periz J, van de Linde S, Heilemann M. Monitoring multiple distances within a single molecule using switchable FRET. Nat Methods 2010;7(10):831–6.
[64] Santoso Y, Kapanidis AN. Probing biomolecular structures and dynamics of single molecules using in-gel alternating-laser excitation. Anal Chem 2009;81(23):9561–70.
[65] Cordes T, Santoso Y, Tomescu AI, Gryte K, Hwang LC, Camara B, et al. Sensing DNA opening in transcription using quenchable Forster resonance energy transfer. Biochemistry 2010;49(43):9171–80.
[66] Finan K, Torella JP, Kapanidis AN, Cook PR. T7 RNA polymerase functions in vitro without clustering. PLoS One 2012;7(7):e40207.
[67] Pinkney JN, Zawadzki P, Mazuryk J, Arciszewska LK, Sherratt DJ, Kapanidis AN. Capturing reaction paths and intermediates in Cre-loxP recombination using single-molecule fluorescence. Proc Natl Acad Sci U S A 2012;109(51):20871–76.
[68] Robb NC, Cordes T, Hwang LC, Gryte K, Duchi D, Craggs TD, et al. The transcription bubble of the RNA polymerase-promoter open complex exhibits conformational heterogeneity and millisecond-scale dynamics: implications for transcription start-site selection. J Mol Biol 2013;425(5):875–85.
[69] Uphoff S, Reyes-Lamothe R, Garza de Leon F, Sherratt DJ, Kapanidis AN. Single-molecule DNA repair in live bacteria. Proc Natl Acad Sci U S A 2013;110(20):8063–8.
[70] Holden SJ, Uphoff S, Kapanidis AN. DAOSTORM: an algorithm for high-density super-resolution microscopy. Nat Methods 2011;8(4):279–80.
[71] Crawford R, Erben CM, Periz J, Hall LM, Brown T, Turberfield AJ, et al. Non-covalent single transcription factor encapsulation inside a DNA cage. Angew Chem Int Ed Engl 2013;52(8):2284–8.
[72] Crawford R, Kelly DJ, Kapanidis AN. A protein biosensor that relies on bending of single DNA molecules. Chemphyschem 2012;13(4):918–22.
[73] Ding F, Manosas M, Spiering MM, Benkovic SJ, Bensimon D, Allemand JF. Single-molecule mechanical identification and sequencing. Nat Methods 2012;9(4):367–72.
[74] Ranasinghe RT, Brown T. Ultrasensitive fluorescence-based methods for nucleic acid detection: towards amplification-free genetic analysis. Chem Commun (Camb) 2011;47(13):3717–35.
[75] El-Sagheer AH, Cheong VV, Brown T. Rapid chemical ligation of oligonucleotides by the Diels-Alder reaction. Org Biomol Chem 2011;9(1):232–5.
[76] El-Sagheer AH, Brown T. Synthesis and polymerase chain reaction amplification of DNA strands containing an unnatural triazole linkage. J Am Chem Soc 2009;131(11):3958–64.
[77] Shelbourne M, Chen X, Brown T, El-Sagheer AH. Fast copper-free click DNA ligation by the ring-strain promoted alkyne-azide cycloaddition reaction. Chem Commun (Camb) 2011;47(22):6257–9.
[78] El-Sagheer AH, Brown T. Click chemistry with DNA. Chem Soc Rev 2010;39(4):1388–405.
[79] Dierckx A, Diner P, El-Sagheer AH, Kumar JD, Brown T, Grotli M, et al. Characterization of photophysical and base-mimicking properties of a novel fluorescent adenine analogue in DNA. Nucleic Acids Res 2011;39(10):4513–24.
[80] El-Sagheer AH, Brown T. Efficient RNA synthesis by in vitro transcription of a triazole-modified DNA template. Chem Commun (Camb) 2011;47(44):12057–58.
[81] El-Sagheer AH, Sanzone AP, Gao R, Tavassoli A, Brown T. Biocompatible artificial DNA linker that is read through by DNA polymerases and is functional in *Escherichia coli*. Proc Natl Acad Sci U S A 2011;108(28):11338–43.
[82] Dallmann A, El-Sagheer AH, Dehmel L, Mugge C, Griesinger C, Ernsting NP. Structure and dynamics of triazole-linked DNA: biocompatibility explained. Chemistry 2011;17(52):14714–17.
[83] El-Sagheer AH, Brown T. New strategy for the synthesis of chemically modified RNA constructs exemplified by hairpin and hammerhead ribozymes. Proc Natl Acad Sci U S A 2010;107(35):15329–34.
[84] Shelbourne M, Brown Jr T, El-Sagheer AH, Brown T. Fast and efficient DNA crosslinking and multiple orthogonal labelling by copper-free click chemistry. Chem Commun (Camb) 2012;48(91):11184–86.
[85] El-Sagheer AH, Brown T. Click nucleic acid ligation: applications in biology and nanotechnology. Acc Chem Res 2012;45(8):1258–67.
[86] Japrung D, Henricus M, Li Q, Maglia G, Bayley H. Urea facilitates the translocation of single-stranded DNA and RNA through the alpha-hemolysin nanopore. Biophys J 2010;98(9):1856–63.
[87] Rotem D, Jayasinghe L, Salichou M, Bayley H. Protein detection by nanopores equipped with aptamers. J Am Chem Soc 2012;134(5):2781–7.
[88] Sapra KT, Bayley H. Lipid-coated hydrogel shapes as components of electrical circuits and mechanical devices. Sci Rep 2012;2:848.
[89] Schneider GF, Dekker C. DNA sequencing with nanopores. Nat Biotechnol 2012;30(4):326–8.
[90] Schneider GF, Kowalczyk SW, Calado VE, Pandraud G, Zandbergen HW, Vandersypen LM, et al. DNA translocation through graphene nanopores. Nano Lett 2010;10(8):3163–7.
[91] Dekker C, Kowalczyk S, Kapinos L, Lim RY. In vitro measurements of single-molecule transport across an individual biomimetic nuclear pore complex. Biophys J 2011;100(3):521a.
[92] Hall AR, Keegstra JM, Duch MC, Hersam MC, Dekker C. Translocation of single-wall carbon nanotubes through solid-state nanopores. Nano Lett 2011;11(6):2446–50.
[93] Hall AR, van Dorp S, Lemay SG, Dekker C. Electrophoretic force on a protein-coated DNA molecule in a solid-state nanopore. Nano Lett 2009;9(12):4441–5.
[94] Kowalczyk SW, Blosser TR, Dekker C. Biomimetic nanopores: learning from and about nature. Trends Biotechnol 2011;29(12):607–14.
[95] Kowalczyk SW, Dekker C. Measurement of the docking time of a DNA molecule onto a solid-state nanopore. Nano Lett 2012;12(8):4159–63.

[96] Kowalczyk SW, Grosberg AY, Rabin Y, Dekker C. Modeling the conductance and DNA blockade of solid-state nanopores. Nanotechnology 2011;22(31):3151–61.
[97] Kowalczyk SW, Hall AR, Dekker C. Detection of local protein structures along DNA using solid-state nanopores. Nano Lett 2010;10(1):324–8.
[98] Kowalczyk SW, Kapinos L, Blosser TR, Magalhaes T, van Nies P, Lim RY, et al. Single-molecule transport across an individual biomimetic nuclear pore complex. Nat Nanotechnol 2011;6(7):433–8.
[99] Kowalczyk SW, Wells DB, Aksimentiev A, Dekker C. Slowing down DNA translocation through a nanopore in lithium chloride. Nano Lett 2012;12(2):1038–44.
[100] Smeets RM, Kowalczyk SW, Hall AR, Dekker NH, Dekker C. Translocation of RecA-coated double-stranded DNA through solid-state nanopores. Nano Lett 2009;9(9):3089–96.
[101] Song B, Schneider GF, Xu Q, Pandraud G, Dekker C, Zandbergen H. Atomic-scale electron-beam sculpting of near-defect-free graphene nanostructures. Nano Lett 2011;11(6):2247–50.
[102] Wallace EV, Stoddart D, Heron AJ, Mikhailova E, Maglia G, Donohoe TJ, et al. Identification of epigenetic DNA modifications with a protein nanopore. Chem Commun (Camb) 2010;46(43):8195–7.
[103] Ayub M, Bayley H. Individual RNA base recognition in immobilized oligonucleotides using a protein nanopore. Nano Lett 2012;12(11):5637–43.
[104] Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 2009;4(4):265–70.
[105] Hansen AS, Thalhammer A, El-Sagheer AH, Brown T, Schofield CJ. Improved synthesis of 5-hydroxymethyl-2′-deoxycytidine phosphoramidite using a 2′-deoxyuridine to 2′-deoxycytidine conversion without temporary protecting groups. Bioorg Med Chem Lett 2011;21(4):1181–4.
[106] Stoddart D, Heron AJ, Klingelhoefer J, Mikhailova E, Maglia G, Bayley H. Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore. Nano Lett 2010;10(9):3633–7.
[107] Stoddart D, Maglia G, Mikhailova E, Heron AJ, Bayley H. Multiple base-recognition sites in a biological nanopore: two heads are better than one. Angew Chem Int Ed Engl 2010;49(3):556–9.
[108] Thalhammer A, Hansen AS, El-Sagheer AH, Brown T, Schofield CJ. Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. Chem Commun (Camb) 2011;47(18):5325–7.
[109] Hall AR, Scott A, Rotem D, Mehta KK, Bayley H, Dekker C. Hybrid pore formation by directed insertion of alpha-haemolysin into solid-state nanopores. Nat Nanotechnol 2010;5(12):874–7.
[110] Villar G, Heron AJ, Bayley H. Formation of droplet networks that function in aqueous environments. Nat Nanotechnol 2011;6(12):803–8.
[111] Conze T, Goransson J, Razzaghian HR, Ericsson O, Oberg D, Akusjarvi G, et al. Single molecule analysis of combinatorial splicing. Nucleic Acids Res 2010;38(16):e163.
[112] Howell WM, Grundberg I, Faryna M, Landegren U, Nilsson M. Glycosylases and AP-cleaving enzymes as a general tool for probe-directed cleavage of ssDNA targets. Nucleic Acids Res 2010;38(7):e99.
[113] Larsson C, Grundberg I, Soderberg O, Nilsson M. In situ detection and genotyping of individual mRNA molecules. Nat Methods 2010;7(5):395–7.
[114] Grundberg I, Kiflemariam S, Mignardi M, Imgenberg-Kreuz J, Edlund K, Micke P. In situ mutation detection and visualization of intratumor heterogeneity for cancer research and diagnostics. Oncotarget 2013;4(12):2407–18.
[115] Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 2005;309(5741):1728–32.
[116] Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wahlby C, et al. In situ sequencing for RNA analysis in preserved tissue and cells. Nat Methods 2013;10(9):857–60.
[117] Goransson J, Wahlby C, Isaksson M, Howell WM, Jarvius J, Nilsson M. A single molecule array for digital targeted molecular analyses. Nucleic Acids Res 2009;37(1):e7.
[118] Goransson J, Ke R, Nong RY, Howell WM, Karman A, Grawe J, et al. Rapid identification of bio-molecules applied for detection of biosecurity agents using rolling circle amplification. PLoS One 2012;7(2):e31068.
[119] Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res 2009;37(18):e123.
[120] Weibrecht I, Lundin E, Kiflemariam S, Mignardi M, Grundberg I, Larsson C, et al. In situ detection of individual mRNA molecules and protein complexes or post-translational modifications using padlock probes combined with the in situ proximity ligation assay. Nat Protoc 2013;8(2):355–72.
[121] Darmanis S, Nong RY, Vanelid J, Siegbahn A, Ericsson O, Fredriksson S, et al. ProteinSeq: high-performance proteomic analyses by proximity ligation and next generation sequencing. PLoS One 2011;6(9):e25583.
[122] Nong RY, Gu J, Darmanis S, Kamali-Moghaddam M, Landegren U. DNA-assisted protein detection technologies. Expert Rev Proteomics 2012;9(1):21–32.
[123] Ke R, Nong RY, Fredriksson S, Landegren U, Nilsson M. Improving precision of proximity ligation assay by amplified single molecule detection. PLoS One 2013;8(7):p. e69813.
[124] Marco-Sola S, Sammeth M, Guigo R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 2012;9(12):1185–8.
[125] Forster M, Forster P, Elsharawy A, Hemmrich G, Kreck B, Wittig M, et al. From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. Nucleic Acids Res 2013;41(1):e16.
[126] Lappalainen T, Sammeth M, Friedlander MR, 't Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013;501(7468):506–11.
[127] 't Hoen PA, Friedlander MR, Almlof J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol 2013;31(11):1015–22.
[128] Sengenes J, Daunay A, Charles MA, Tost J. Quality control and single nucleotide resolution analysis of methylated DNA immunoprecipitation products. Anal Biochem 2010;407(1):141–3.
[129] Huang J, Renault V, Sengenes J, Touleimat N, Michel S, Lathrop M, et al. MeQA: a pipeline for MeDIP-seq data quality assessment and analysis. Bioinformatics 2012;28(4):587–8.
[130] Knoppers BM, Zawati MH, Kirby ES. Sampling populations of humans across the world: ELSI issues. Annu Rev Genomics Hum Genet 2012;13:395–413.
[131] van El CG, Cornel MC, Borry P, Hastings RJ, Fellmann F, Hodgson SV, et al. Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. Eur J Hum Genet 2013;21 Suppl 1:S1–5.
[132] Chadwick R. Personal genomes: no bad news? Bioethics 2011;25(2):62–5.
[133] Peppercorn J, Shapira I, Deshields T, Kroetz D, Friedman P, Spears P, et al. Ethical aspects of participation in the database of genotypes and phenotypes of the National Center for Biotechnology Information: the Cancer and Leukemia Group B Experience. Cancer 2012;118(20):5060–8.
[134] Tabor HK, Berkman BE, Hull SC, Bamshad MJ. Genomics really gets personal: how exome and whole genome sequencing challenge the ethical framework of human genetics research. Am J Med Genet A 2011;155A(12):2916–24.
[135] Desai AN, Jere A. Next-generation sequencing: ready for the clinics? Clin Genet 2012;81(6):503–10.
[136] Westerlund F, Persson F, Kristensen A, Tegenfeldt JO. Fluorescence enhancement of single DNA molecules confined in Si/SiO_2 nanochannels. Lab Chip 2010;10(16):2049–51.
[137] Puente XS, Bea S, Valdes-Mas R, Villamor N, Gutierrez-Abril J, Martin-Subero JI, et al. Non-recurrent mutations in chronic lymphocytic leukaemia. Nature 2015. doi: 10.1038/nature14666.
[138] Kulis M, Heath S, Bibikova M, Queiros AC, Navarro A, Clot G, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 2012;44(11):1236–42.
[139] Kulis M, Merkel A, Heath S, Queiros AC, Schuyler RP, Castellano G, et al. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nat Genet 2015;47(7):746–56.