4.1 READNA Final publishable summary report

I) Executive summary

DNA analysis methods are tools to advance on the goal of understanding the information encrypted in our genomes. The overall objective of READNA (REvolutionary Approaches and Devices for Nucleic Acid analysis) was to develop a toolbox of nucleic acid analysis methods that will enable effective and economic deployment for the good of society. READNA has made substantial headway on several different fronts during the 54 months of the project. In several lines developments and findings have led to key publications and have contributed to the understanding and advancement of nucleic acid analysis methodology. Efforts of the READNA consortium are substantially contributing to the advancement on the 1000 $ genome. READNA has also worked on the dissemination of its knowhow and on uniting the research community involved in nucleic acid technology development and application. The READNA project has produced over 100 published or accepted papers, more than 160 presentations at conferences and more than 20 patents or applications. Several of the developments have or will be reaching the market soon and several spin-off companies have been created as a result of READNA activities.

WP1 (**Near Term Innovations: Ancillary elements for 2nd generation DNA sequencers**) has evaluated target enrichment methods as well as developing a novel enrichment methodology entitled Selectors. The Selector method has been commercialised under the spin-off HaloGenomics (acquired by Agilent). WP2 (**Near Term Innovations: Improvement and extension of existing methods**) has been involved in the development of facile sequencing methods such as DASH and ribo-PCR MS. WP3 (**Fluorescence-based Single Molecule Sequencing**) has extensively evolved DNA handling methodologies in nanofluidic devices, resulting in a new EU consortium - Cell-O-Matic which is implementing READNA developments. WP4 (**Nanopore Sequencing)** has developed a method for the distinction of the different nucleosides with nanopores opening up the possibility of commercialising a nanopore sequencer. When released this Nanosequencer will revolutionise the way DNA sequencing is performed. WP5 (**New Genotyping Challenges**) has developed methods towards *in situ* sequencing in cells and tissue that have reached proof of concept stage and a publication has been submitted. A digital RCA approach pursued in this WP (UU & Olink) has been fully developed and will be commercially exploited by the spin-off q-Linea. Methods for targeted multiplex methylation profiling have also been developed and exploited. A unique approach to localized mutation detection in histology sections has been developed and in addition the approach is being optimized further to allow detection of oncogenic mutations directly in formalin-fixed paraffin embedded tissues.

In addition READNA has organised 4 large workshops, 4 training workshops and contributed to several international conferences. More than 40 post-docs have worked on the project and 11 students obtained their PhDs through READNA.

Full details of READNA can be found at www.cng.fr/READNA.

II) A summary description of project context and objectives

Project Context

Methods for DNA analysis have evolved dramatically in the last 20 years and 2nd generation sequencing technologies emerged in 2005. These technologies although cheaper and faster than Sanger sequencing are still expensive and slow to perform whole genome sequencing in a study of epidemiological scale with thousands of samples. At the outset of READNA 2nd generation sequencing could achieve resequencing of a human genome for $100,000. However, the costs of reagents and DNA amplification suggest that these technologies will never reach the low costs required for a $1000 genome. The goal of being able to sequence whole genomes 100-times faster than with these new 2nd generation sequencers at 1 % of the cost will be primordial. The only technologies likely to meet the $1000 goal are single molecule approaches. In addition, high-throughput, accuracy and high information-content are additional important goals for future sequencing technologies. Resequencing complete genomes will require another paradigm shift and a third generation of sequencing technology and new analytical methods. Research into disease underlying genetic variants could be satisfied with a sequencer capable of delivering 98 % of a human genome sequence with accuracy better than 99.9 % in a day at a cost of 1000 $ per sample analysed. Until this device is available enriching the 5 % of the genome containing all genes and regulatory elements constitutes a great step forward for disease genetics.

In disease diagnostics the sensitivity of genotyping is limited and the technology does not capture all disease relevant information, illustrated in the screening of all newborns to detect prevalent cystic fibrosis (CF) mutations in which rare disease causing mutations are missed. In complex disorders the situation is similar: Susceptibilities for breast cancer due to mutations in BRCA1 and BRCA2 is only incompletely captured. In Crohn's disease the CARD15 gene has three main disease associated variants but approximately 25% of the gene-associated risk relates to rare variants. To resequence CF, BRCA1/BRCA2 or CARD15 in hundreds of thousands of diagnostic samples is currently still way beyond the capacity of diagnostic laboratories and economically not feasible.

Screening for somatic mutations is becoming an important diagnostic in oncology. Certain forms of cancer are accompanied by mutations in genes such as *KRAS* (small cell lung cancer) or *APC* (colorectal cancer). Mutations in these genes are usually only picked up once a DNA sample from a biopsy of a tumour is analysed. However, tumours shed cells into the bloodstream that make up only a very small proportion of cells next to lymphocytes. Nucleic acid analysis methods do not have the specificity to genotype or sequence DNA molecules from cells shed from tumours that are present in a huge background of non-mutated cells. A simple, functioning and operational method to do this would revolutionize preventative cancer screening.

Understanding the medical role of pathogen variability is considered a premise to a rational use of anti-infective therapies. Identification of pathogens and scoring of pathogenicity characteristics is an entirely different problem in terms of nucleic acid analysis again. Here it is important to identify quickly and accurately a subtype of, for example, a bacterial strain in order to combat disease at a very early stage. Rapid analysis would be of great benefit.

The complete identification of the genetic risk components will be a pre-requisite in complex diseases to understand the interaction with triggering environmental risk factors. The analysis of haplotypes is more informative than the sum of the genotypes of all individual polymorphisms. In most nucleic acid analysis approaches haplotypes are not captured and can only be estimated statistically. Molecular methods for haplotype analysis are of particular benefit in areas of the genome where the sequence diversity and polymorphism rate is very high, such as the MHC on chromosome 6. In addition, if haplotypes could be determined over long stretches of several tens of kilobases or more would be of great value.

Copy number variations (CNVs) have become important since insight was gained that genomes are far more copy number variant than single nucleotide polymorphic. Array-based methods are predominantly used to map them but arrays only can capture what is represented on them. More flexible and efficient methods for CNV analysis currently do not exist but would be of great benefit for research and diagnostics.

DNA methylation can be used as a pointer as to which parts of the genome are active and DNA methylation profiles of different cell types vary strongly. Changes in DNA methylation from a normal state are strong indicators of deleterious effects and are frequently observed in cancerous tissue. Current methods for DNA methylation analysis rely on bisulphite conversion of DNA or enrichment of the methylated or unmethylated compartment of a sample followed by array analysis. The difficulty with this is that arrays only represent what they contain. All sequences that are not represented escape analysis. Approaches for genome-wide and unbiased DNA methylation analysis are needed to further this field.

Array-based procedures for genome-wide transcript profiling have been in existence for many years. However, one of the major defaults of all array-based methods is that only what is represented on the array can be detected and the dynamic detection range is very limited. RNA transcripts can vary in abundance by more than six orders of magnitude. Here 2nd and 3rd generation sequencers could lead to substantial improvements of the measurement of these important biological molecules and could remove some of the limitations of current technology.

Objectives

The READNA consortium is built around complementary skills stretching from technology development to clinical research and informatics in order to provide the competence for resolving the outlined problems. Affordable tools for genetic studies will underpin the understanding of the causes of the molecular basis of disease. Early detection of disease onset will reduce suffering and will allow guided treatment decisions. These factors will result in a reduction of healthcare cost and dramatically improve quality of life. READNA joins partners who are and have been implicated in the development and distribution of different generations of nucleic acid analysis methods and instrumentation which will extend the range of nucleic acid applications and make them more accessible to the generic system user. The READNA consortium joins the forefront of European researchers from nucleic acid chemistry, molecular biology, nanopore technology, microfabrication all through to partners with proven expertise in high-throughput nucleic acid analysis and disease genetics. Certain partners have close ties to the clinic with experience of nucleic acid problems that present a real medical need and who can also secure access to clinical material.

WP1 (Near Term Innovations: Ancillary elements for 2nd generation DNA sequencers) will aim to provide a benchmark of targeted enrichment methods and will develop and implement concepts such as MegaPlex PCR and Selectors. These methods will be standardized, be applied for targeted studies of genomic variability and the verification of rare variants with the aim of making them ready for widespread deployment. Standards and formats for data analysis from 2nd generation sequencing experiments will be a second focal point of this WP. Several project partners will be implicated in international efforts to this end. In addition the WP set out to define an objective and comprehensive set of reporting metrics for enrichment and sequencing methods. A component was added to examine the ethical and societal impact of high-resolution DNA analysis methods with the aim of providing a detailed report and a publication in a high impact journal.

WP2 (Near Term Innovations: Improvement and extension of existing methods) will focus on the completion and stabilisation of the DNA analysis methods that had reached a proof of principle at the outset of the project. The DASH (dynamic allele-specific hybridisation) technique for genotyping will be evolved towards sequencing in an array format. Datta arrays will be developed for individual picolitre reactions. ribo-PCR MS will be optimized to become a simple re-sequencing protocol that requires only genomic DNA, primers and a PCR mastermix. Post cycling treatment will consists of a simple treatment with NaOH and desalting before mass spectrometric analysis. Dedicated software for the interpretation of results will be developed. In a second mass spectrometry implementation a very economic, high resolution HLA typing protocol for HLA-A, HLA-B and HLA-DRB1 based on the analysis of microhaplotypes will be developed. A third task will focus on the development of a system for sample characterisation that uses DNA, RNA and protein profiles analyzed by mass spectrometry.

WP3 (Fluorescence-based Single Molecule Sequencing) will aim to prepare the stage for development of fluorescence-based 3rd generation sequencing methodology and long-range DNA analysis. Different sequencing chemistries will be developed using chemical ligation methods and FRET-based donor-acceptor systems on DNA polymerases. The resolution of optical detection will be improved to go below the diffraction limit of light. Microfluidic devices will be developed together with methods for DNA handling that will allow record size DNA molecules to be manipulated and enzymatic reactions such as restriction or hybridisation reactions to be carried out inside the nanochannels. The ultimately goal will be to use a developed sequencing chemistry to obtain sequence information in an ultra-long range context (within microfluidic devices). Two main directions will be investigated, one direction being chemical ligation-based sequencing which could become a cost-effective, approach for when a large number of molecules need to be enumerated such as in digital transcriptomics. The second direction is real-time sequencing and its application on long-stretched molecules.

WP4 (Nanopore Sequencing) will develop the use of nanopore-based measurements for single molecule nucleic acid sequence analysis. Detection techniques will be developed that allow for the discrimination of all four nucleotides and the variants methylcytosine and hydroxymethlycytosine as individual bases after exonuclease cleavage, and when present in the nanopore as an intact strand of DNA. Methods of control of DNA translocation will also be developed, one of which includes the successful coupling of exonuclease enzymes to a nanopore. Key developments will be made towards the creation of large stable arrays of

nanopores by the insertion of protein-based nanopores into similarly sized apertures fabricated in solid materials for clinical application. In addition alternative membrane materials will be investigated including silicon nitride and monoatomically thin graphene monolayers. Efforts will be made on droplet-interface bilayers towards the goal of possibly achieving single cell DNA sequencing. Finally measurement of static RNA strands inside nanopore proteins will be investigated to provide similar degrees of resolution and discrimination compared to DNA, including RNA base modifications, moving a step closer to direct analysis of RNA using nanopore devices. All of the work highlighted above aims to deliver a transformational approach to sequencing, one in which rapid, long-read direct analysis is implemented in highly scalable electronic read-out devices, suitable for a clinical setting as well as high-throughput research facilities.

WP5 (<u>New Genotyping Challenges</u>) will develop methods for nucleic acid analyses for which currently no good, economic methods exists. In addition the WP is devoted to solving genotyping and epigenotyping challenges, where the performance of current technology is not satisfactory. Existing nucleic acids analysis methods will be adapted for DNA methylation analysis and copy number variation analysis to deliver very precise quantitative results. Methods will be developed to analyse mutant nucleic acid molecules under challenging conditions such as *in situ* in histological sections or in a high background of wildtype molecules. The technologies will serve for genome-wide-, targeted multiplexed-, and clinical analysis needs. The efforts towards *in situ* sequencing in cells and tissue have reached a proof of concept stage and have been submitted for publication. The digital RCA approach pursued in this WP (UU & Olink), initially for biodefense applications will be commercially exploited if the technique is brought to fruition. Methods for the development of a multiplex typing procedure for rare somatic variants in circulation and for targeted multiplex methylation profiling will be developed and exploited. A unique approach to localized mutation detection in histology sections will be developed and if successful the technology will be further adapted to allow detection of oncogenic mutations directly in formalin-fixed paraffin embedded tissues. The WP aims to take certain technologies and apply them to diagnostic applications (e.g. *in situ* method will be developed for *KRAS* testing in biopsies).

III) A description of the main S&T results/foregrounds

WP1: Near Term Innovations: Ancillary elements for 2nd generation DNA sequencers

WP1 was focused on peripheral elements that will impact on the near-term improved utility of 2nd generation nucleic acid sequencing. Areas of specific focus included: i) enrichment by 'MegaPlex PCR', which was extended to a fundamental discovery project which revealed that DNA strands in many genome regions are far more difficult to separate than generally realised, impacting upon sequence depth uniformity and many other parameters, with differential severity across sequence regions and DNA samples; ii) enrichment by the 'Selector' technique, which has now been highly optimised and validated as a superior approach, and recently commercialised; iii) benchmarking and optimisation of 'Hybridisation-Based DNA Enrichment' methods, along with technological enhancements and adjunct analysis software; iv) 'Rare Mutation Enrichment', to recover rare somatic mutations from clinical DNA samples; v) 'Amplified Single-Molecule Arrays as Sequencing Substrate', representing an extension to the Selector activity; vi) 'Data Complexity Reduction', which evolved into an effort to define an objective and comprehensive set of reporting metrics for enrichment and sequencing methods; and, vii) 'Definition of ethical standards' relating to the use of 1st, 2nd and 3rd generation DNA analyzers.

The major achievements of WP1 are summarized below.

i) Fundamental insight into why multiplex assays vary in efficiency due to 'TUF' DNA

Work towards a surface-based high-multiplex PCR amplification system, useful for DNA enrichment (amongst other things) showed great potential, but was difficult to progress at a speed competitive with alternative systems due to the lack of a convenient system for arraying mixed oligo features at high and tunable density. Therefore, we instead focused on evidence emerging from work investigating why certain genome regions react with massively different efficiency across many DNA analysis platforms, including sequencing.

Extending initial results by use of an extensive series of methods (e.g., quantitative PCR, Paralogous Ratio Tests, whole genome amplification, high-throughput SNP genotyping, genome partitioning, next generation sequencing, genome informatics, Southern Blotting, and advanced statistics) and exploiting native and cloned DNAs from a range of species pre-processed by several enzymatic and physical means, we proved that the human genome contains many non-melting (and hence non-assayable) sequences – ultimately termed 'Thermodynamically Ultra-Fastened (TUF) DNA'. This was recently published in BMC Genomics (*Veal et al., 2012*[1]), representing a major discovery regarding the nature of genomic DNA and how it behaves upon analysis by any modern technology that requires DNA strand separation (i.e., most current technologies).

Contrary to common understanding, we showed that many human genome sequences resist denaturation, even up to 130 degrees Celsius, by cooperative stabilisation between closely positioned C+G elements. This holds adjacent sequences in close proximity in procedures where they are supposed to separate, so allowing them to quickly reanneal, and hence preventing their efficient analysis. This causes differential assay signal strengths across the

genome of 10-100 fold (given differing proximities to TUF elements), and also from one sample to another (given differing degrees of nicking and fragmentation). In some samples, this impacts 10-20% of the genome, especially regions enriched for genes and functional elements. Fortunately, we also showed that the problem can be largely overcome by fragmenting input DNAs by physical or enzymatic means.

TUF therefore causes major non-uniformity of assay signal strengths in many modern day technologies, such as quantitative PCR, Paralogous Ratio Tests, SNP genotyping, CNV scoring, whole-genome amplification, 2nd generation DNA sequencing, and so on. A detailed human TUF map now needs to be constructed, and pilot studies into how this might be achieved have now been initiated with other READNA Partners, exploiting knowhow and technologies also developed in this project.

ii) Optimisation of a superior amplification based enrichment method, based on Selectors.

Selectors provide a strategy for targeted DNA enrichment by multiplex-amplification with high specificity. In the Selector procedure, a large number of 'Selectors' (synthetic DNA molecules) are used in solution to convert large sets of sequence-specific subselections of a DNA sample into circularized form, allowing efficient enrichment and parallel amplification using a universal amplification system.

The initial step of the selector procedure is restriction digestion of eight aliquots of a DNA sample, using different combinations of restriction enzymes in each tube. The restriction digested DNA is then pooled and denatured. A selector probe library is then added to the pooled DNA. The library is designed such that each base-pair in the targeted genes are present in multiple overlapping fragments that are selected and amplified by different, independent selector probes. In an exhaustive design each base-pair can be targeted by up to 16 different selector probes, one for each strand in eight different restriction digestions. This probe redundancy provides selector assays with very high target recovery and sequencing coverage. Other advantages with the selector technology are the high specificity inherent to amplification based enrichments techniques, and low requirements on sample DNA input. The protocol is simple and easily automatable and generates a sequencing library ready for loading on a NGS instrument. Combined, these qualities make selectors very promising for diagnostic applications.

After developing the basic protocol (*Johansson et al., 2011*[2]), we have been optimizing selectors for DNA extracted from formalin-fixed paraffin-embedded (FFPE) tissue. This is the routine sample source for tumor tissue, particularly if the tumor is not resected, but only sampled as a biopsy. FFPE DNA is damaged and degraded which makes it difficult to work with, and the damages may generate false positive mutations. We have adapted the selector protocol to target very short fragments, and also to tag all target DNA molecules with unique identifier sequences, to ensure that the mutation analysis is based on different target molecules. By analyzing sequences generated from both polarities of the target DNA, sequencing artifacts can be avoided (Moens *et al.,* in preparation)

iii) Benchmarking and optimizing a pipeline around 'Hybridisation-Based DNA Enrichment'

Direct physical enrichment of target genome regions by sequence-specific hybridization and selection, on surfaces or in solution, is a widely used but challenging reaction concept. Generally, they are limited in their enrichment power and their evenness of sequence recovery over different genomic regions. Cost is also a limiting factor. All of this was carefully reviewed, and then published (*Mertes et al., 2011*[3]). We also carefully devised a series of objective and transparent reporting metrics that would allow enrichment based sequencing studies to be fairly compared. This will be published shortly. And to advance the field in practical terms, we undertook a series of benchmarking and method improvement studies to produce better ways for both enriching genome sequences and analyzing the sequence data thereby produced.

The benchmarking studies encompassed long-range PCR approaches, the droplet-based PCR technology from RainDance (MA, USA), the Agilent SureSelect technology, the Nimblegen arrays, and the Febit HybSelect Biochips, to target multiple human genes and 100's-1000s of kbp. We also evaluated the leading commercial whole-exome re-sequencing approaches (solution versus solid phase), using three different NGS technology platforms. We considered various enrichment measures and SNP concordance, to cross-compare the different technologies. Our findings revealed divergent enrichment efficiencies, and guided us in specifying an optimal whole-exome protocol. Moreover, we tested the applicability of whole-genome amplified (WGA) material for NGS and demonstrated comparable sequence performance between non-amplified and preamplified DNA samples and between different indexing strategies (*Elsharawy et al., 2012*[27]).

Considering all relevant factors, solution phase methods were found to work best. These were then successfully adapted and improved in terms of reducing enrichment costs significantly by first producing our own sets of mass-amplified capture probes. To circumvent contamination of sample, RNA/DNA hybrid capture selections were adopted, with subsequent RNA digestion of the bait library. Final system performance is comparable to that achieved by the Agilent SureSelect technology, but with far lower reagent cost (up to 10x reduction) and greater convenience.

To then handle the typical data analysis challenges associated with the use of advanced sequencing methods (e.g., read alignment, single nucleotide variants detection, copy number variation calling, etc.), we also developed two software solutions. To address the mapping issues, we have developed the "Backmapping pipeline", which is a novel time-saving two-step mapping approach (*Elsharawy et al., 2012*[28]). The second software package, namely pibase (*Forster et al., 2012*[29]), is applicable to diploid and haploid genome, exome, or targeted enrichment data. In test cases, this has been shown to identify positions of sequence heterozygosity rapidly and with extremely good specificity.

We established a survey of all of the technological advances that were being worked on in READNA from the point of view of their ethical, legal and societal impact. A report and publication was prepared detailing the issues that are anticipated from the availability of ever faster and accurate nucleic acid analysis technology.

WP2: Near Term Innovations: Improvement and extension of existing methods

WP2 focused on nucleic acid analysis methods that had passed their proof-of-principle. Certain of these methods were further improved and extended to make them widely usable, particularly for clinical applications that require informative but simplified methods for routine diagnosis and disease monitoring. The major achievements of WP2 are summarized below and three major developments have been brought to fruition.

i) ribo-PCR MS sequencing

The first major achievement of WP2 has been the development of a novel, simplified version for DNA sequencing entitled **ribo-PCR MS sequencing**. The ribo-PCR MS sequencing method had already been developed, tested and published prior to the start of the READNA project (*Mauger et al., 2007*[4]). The ribo-sequencing method although successful for sequencing the hypervariable region 1 (HV1) of mitochondrial DNA in 22 individuals was far from optimal. It was a cumbersome **4 step process** (PCR reaction, ribo-extension reaction, cleavage and MS analysis) not adapted to rapid resequencing of DNA samples. The READNA project has successfully simplified the technique to a **3 step process**, single-tube process (combined PCR/ribo-extension reaction, cleavage and MS analysis). Genomic DNA is mixed with a ribo-PCR cocktail, containing two locus specific primers, three regular nucleotides and one ribo-nucleotide and a novel ribo-incorporating DNA polymerase (supplied by Roche Diagnostics). After thermal cycling, products are sequence-specifically cleaved with sodium hydroxide at sites of NTP incorporation. The resultant fragments are analysed by MALDI-TOF mass spectrometry. Dedicated software to use with this procedure was developed. The entire integration provides a streamlined, extremely low cost solution and process from DNA to result. The simplified technique has been tested and validated on the SLC01B1 locus and the NOS1 gene on a total of 100 DNA samples that show 100% concordance with the reference DNA. Implementation of novel target regions is easy from this point. The pipeline is suitable for mass screening approaches. This system was developed in a cooperation of the two **CEA** partners – **CNG** and **LIST**. We are currently trying to identify a commercial partner for this product.

ii) Dynamic Allele-Specific Hybridisation (DASH)

The second major achievement of WP2 is the development of a new DNA analysis method entitled **Array based Dynamic Allele-Specific Hybridisation (Array-DASH)** – intended to facilitate widespread DNA diagnostics in myriad settings. The latest version of this technology enables very flexible and highly multiplexed DNA fingerprinting, genotyping, scanning and resequencing, or simultaneous combinations thereof, with the ability to detect ultra-low abundance mutations in mutated and mixed samples – all at low cost and using standard run conditions.

The core DASH reaction principle was developed previously, and publications are available that describe its successful use for high throughput SNP genotyping (*Jobs et al., 2003*[5]*, Prince et al., 2001*[6]*),* and other papers thereby referenced). READNA provided the necessary resources to transfer DASH to an array-based format.

DASH exploits the proven diagnostic utility of melt-curve analysis, wherein duplex DNA is detected by fluorescence, and a simple temperature ramp is used to dynamically heat and

thereby denature one or many target molecules away from sequence-specific probes attached to a solid support. The subsequent melting curves are then analysed to detect sequence variants, and virtually all sequence alternatives in all contexts can be fully and semi-quantitatively resolved. Using the capacity and diversity of probes that can be tiled on modern arrays, this method delivers unprecedented sensitivity (detects allelic variants down to at least 1% representation), virtually error-free detection of *de novo* and known short range mutations (single- and multi-base alleles, as well as indels), and impressive robustness against extremes of C+G content and secondary structure.

The development of Array-DASH involved a 3-party team including the **University of Leicester** (concept development and performing the experiments), the SME **FlexGen** (producing suitable oligonucleotide arrays) and the SME **Genewave** (providing the 'HybLive' device for simultaneous heating and monitoring of microarrays in real time). The three groups established that their combination of chemistry, arrays, and dynamic reader are suitable for Array-DASH implementation, and alternatives were also explored. Arrays with tens of thousands of features are now being examined routinely by this technology, in principle allowing many thousands of contiguous or dispersed bases or stretches of DNA to be examined, after amplification by a range of methods.

Many assay parameters were empirically optimized including: different array providers; buffer and heating variables; surface oligo spacers; surface oligo lengths; required target concentration ranges; viable target lengths; ways to generate suitable targets from genomic DNA (e.g., PCR with conversion to ssDNA, selectors (with partner UU)); ability to detect and quantify low level DNA representations in mixtures (e.g., specific sequence alternatives at 1:99 ratios or less, and different genomes by means of multiple repeat element based probes that vary between species); and robustness against high C+G content and highly folded structures.

To analyse the large number of melt curves produced by the Array-DASH method, we created software that is generic in nature and therefore suitable for use upon any target region(s) of interest. This allows tuning of analysis parameters for research settings, or the use of preset optimized variables when interrogating specific targets in diagnostics scenarios – including the ability to compare against data from one or more pre-established benchmarks experiments.

Pilot studies exploring several challenging real-world applications, not least: prenatal testing of fetal genomes in maternal plasma; detection of low abundance clonal mutations in cancer biopsies; detecting rare drug resistance variants in HIV infections, and; purity testing of harvested crops were carried out.

iii) Risk profiling of multifactorial diseases by MALDI mass spectrometry

The third major achievement of WP2 is the development of a new method of risk profiling multifactorial diseases using MALDI mass spectrometry. As with achievements i & ii described above, protocols using MALDI mass spectrometry for SNP genotyping were already successfully developed prior to the READNA project (*Sauer S et al., 2006*[7]). The READNA project has further optimized these protocols to simultaneously analyse SNPs, CNVs and transcripts of target markers, as well as proteins, to better characterise the risk profile

of a patient. Nucleic-acid based methods such as PCR (multiplex) and primer extension in combination with MALDI mass spectrometry have been pushed further. Work involved selecting target markers for disease or bacteria detection and standardising the protocols to ensure high data quality and throughput. The protocols for simultaneous analysis of SNPs, CNVs and transcripts of target markers and proteins on a single device with read-out by mass spectrometry have been successfully applied in a pilot study for differential analysis of inflammatory bowel diseases (*i.e.* Crohn's disease and ulcerative colitis). Interestingly the approach was also used for other important diagnostic applications such as for the classification and identification of bacteria. These kinds of applications have become extremely popular in recent years as the influence of the environment on the host has been recognized. Suitable tools for this analysis will be invaluable both in research and diagnostics. Finally the partner involved has also set up a new mass spectrometry assay to unambiguously screen natural products with health-beneficial effects and a publication was generated as a result (*Weidner et al., 2012*[8]) with a primary biological focus was on anti-diabetic and anti-inflammatory properties.

WP3: Fluorescence-based Single Molecule Sequencing

At the outset the stated aims of the workpackage were (1) to develop massively scalable techniques that reduce costs and increase throughput of sequencing and (2) to add value to sequence reads by providing long-range contextual information. The product of the research was (1) a proof-of-principle for novel sequencing biochemistry and (2) the acquisition of sequence information in its true genomic context. Some products of the work are summarized below:

- We have fabricated and tested new device designs that enable ultra-long DNA molecules to be visualized and reagents to be exchanged over them.
- We have extracted ultra-long DNA from single cells and chromosomes.
- We have explored the fundamentals of novel FRET-based polymerase sequencing approaches and Click Chemistry-based, enzyme free sequencing approach.
- We have developed methods to passivate the interior walls of nanofluidic chips.
- We have conducted enzymatic reactions on DNA molecules stretched via nanofluidics.
- We have shown means for extracting sequence information in a long range context.
- We have demonstrated methods for nanometric imaging and localization.
- We have developed effective means for mapping megabase lengths of DNA, determining their identity by comparing to reference maps.
- We have collected DNA after it has been mapped in a nanofluidic chip, amplified it and conducted next generation sequencing and FISH on the product.

In short, we have achieved our grand aim of setting sequence information in its long-range context

Establishing feasibility for future generations of sequencing biochemistry

Sequencing methods in use today rely on reading sequence by making a complementary copy of the target The first aim of the work-package was to explore new ways to expand this

basic concept, particularly by developing approaches that could be applied at the single molecule level and involved optical detection which is well understood and is amenable to massive parallelization.

Ligation has the potential to read several bases per cycle. Early in the project, we completed the development of a ligation based sequencing biochemistry (*Mir et al., 2009*[9]) and explored approaches for reading more than one base per cycle by coding multiple bases. We also advanced the ligation approach by conducting basic research on ligation by Click Chemistry so that a rapid, enzyme-free sequencing approach can be developed. Experiments have shown that ligation specificity sufficient for correct base-calling can be achieved (***Figure 1***).

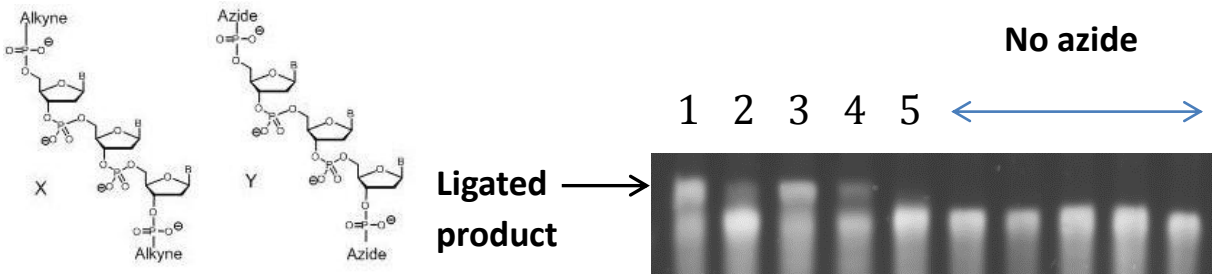

Figure 1*: Specifiity of Click Ligation. Lanes1 and 2 perfect match and single base mismatch respectively using template 1; lanes 3 and 4 perfect match and single base mismatch respectively; lane 5 azide template, no incoming oligo; lanes 6-10 no azide controls.*

Polymerase-based sequencing has the potential for rapid continuous sequencing that can be followed in real-time without requiring reagent exchange. We performed fundamental studies on dissecting the finger opening-closing action of a DNA polymerase utilizing FRET (*Santos et al., 2010*[10]) followed the incorporation of a number of bases in real-time (***Figure 2***).

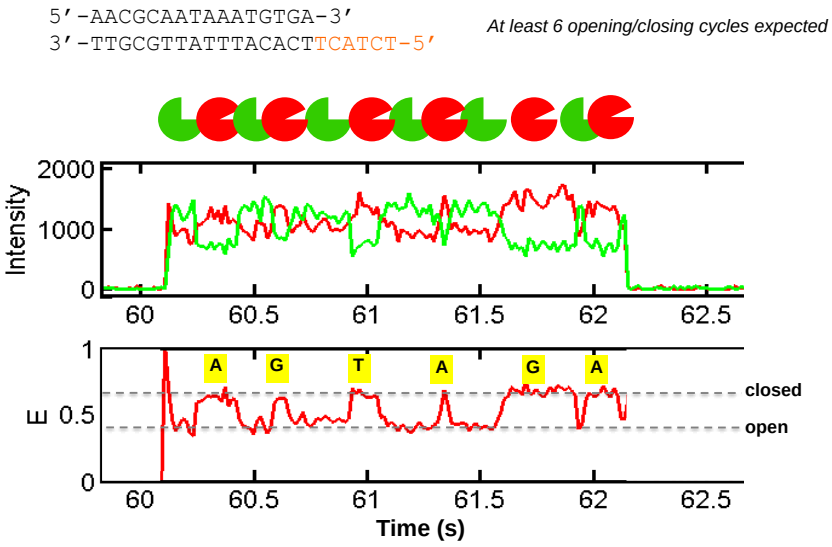

Figure 2*: DNA Polymerase Fingers opening-closing observed by FRET between a donor and acceptor pair on the Polymerase.*

We have utilized FRET-based polymerase fingers opening-closing for sequencing by a nucleotide limiting method. We have also conducted fundamental studies into two real-time sequencing schemes: modulation of the fluorescence of labels on a polymerase by dark quenchers (*Lereste et al., 2012*[11]) attached to the nucleotide and; FRET between a DNA stain intercalated in a DNA duplex with base-specific labels on the nucleotide. This latter approach seamlessly fuses with the single molecule display of mega-base lengths of DNA in nanochannels which are visualized via DNA stains as described below.

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

No sequencing method to date has produced a complete faithfully-re-assembled genome. Current sequencing methods produce sequencing reads whose genomic location of origin is not known *a priori,* rather reads are mapped to a reference genome, thereby losing any individual-specific organization of a personal genome. We aimed to add value to sequencing by reading sequence within its long-range context so that a complete long-range genome sequence, incorporating structural variation and haplotype information could be obtained. For this we had proposed to extract DNA in up to whole chromosome lengths, handle and display such unusually long genomic DNA and obtain sequence information along single ultra-long DNA molecules at multiple locations and to develop methods that allow such locations to be detected at sub-nanometric resolution and precision.

Extracting DNA

The complete context of a sequence read should ultimately be within a single whole chromosome. However, DNA extraction and handling techniques tend to break chromosomal DNA into smaller pieces. We delivered whole chromosome lengths of DNA within their robust metaphase packaging into a novel micro-/nanofluidic chip and extracted DNA without subjecting it to the shearing forces that would cause it to break up (*Rasmussen et al., 2011*[12]), ***Figure 3*** shows the result of a DNA molecule being pulled out of a bundle of DNA from a single metaphase chromosome.

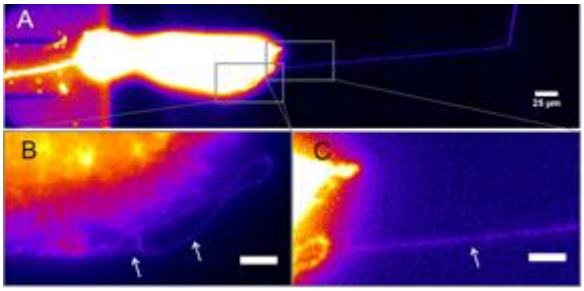

Figure 3*: Chromosmal DNA extracted on Chip. A, mass of extracted DNA entering a nanoslit and several Kbp length of DNA being pulled out. B, C details.*

Displaying chromosome lengths of DNA in a single field of view

We developed a novel nanofluidic chip design in which the nanocannels fold into a back and forth configuration and showed that a whole chromosome length of DNA could be made to enter and be displayed for imaging (***Figure 4***).

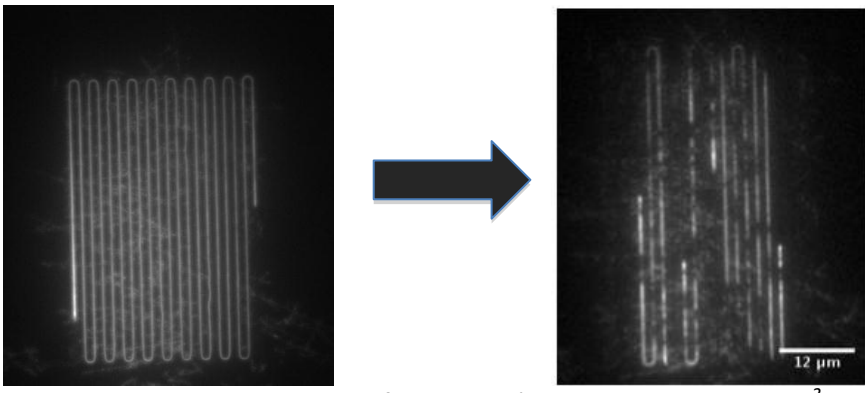

***Figure 4**: Whole chromosome of S. pombe (5.7Mbp in 250 x 250 nm^2 meander channels imaged using a fluorescence microscope (right). Pattern after partial denaturation (GC rich regions are bright as they retain the DNA stain; AT rich regions are dark to loss of DNA stain. The DNA molecule is stained with YOYO-1.*

Single Molecule Mapping , Molecule Rescue and Next Generation Sequencing

As an alternative to sequencing *in situ* on stretched molecules, we developed means to physically map megabase-lengths of DNA by Denaturation Mapping (*Reisner et al., 2010*[13]). We found that an experimentally obtained map could be compared to an *in silico* generated map to robustly identify the genomic origin of any molecule analysed. Moreover, we developed a means to reconcile the long range view of genome sequence with the gross cytogenetic view and the short-range base sequence view. We did this after imaging the map of a DNA molecule stretched in the device, by rescuing the molecule from the chip, performing whole genome amplification and making a nick translated probe for FISH and preparing a library for sequencing (***Figure 5***). Hence the Illumina sequencing reads obtained could be put into a mega-base scale and whole chromosomal context (*Marie et al. PNAS, accepted*). Moreover, sequence-based maps should be able to guide *de novo* assembly of short-read sequences, thereby helping to assemble a complete individual human genome for the first time.

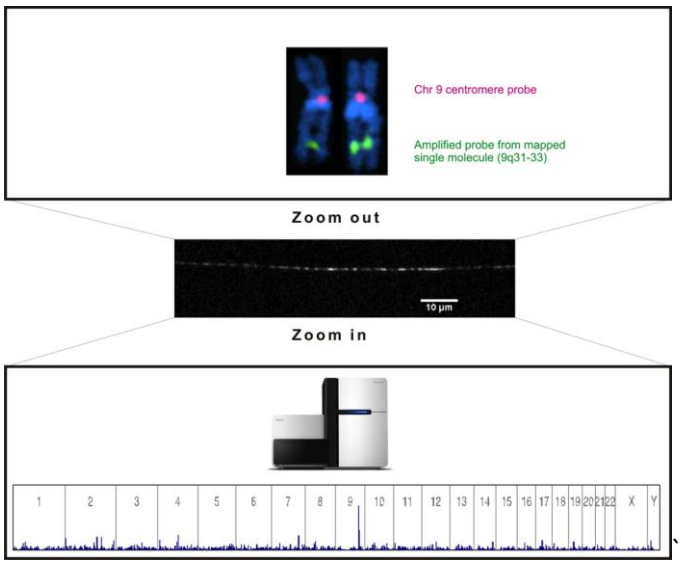

***Figure 5**: Mapping, sequencing and FISH of a single chromosomal DNA molecule. Centre, Section of experimental D-R map obtained from a single molecule from the Jurkat genome; the molecule was collected and amplified; Top, the molecule was mapped to 9q31-33 by FISH; Bottom, Illumina sequencing was performed and the read footprint localized the molecule to chromosome 9.*

Towards **sequencing** ***in situ*** **on stretched DNA**

We developed means to carry out enzymatic reaction on DNA stretched in nanochannels, so that sequencing reads can be obtained in their long-range context. This involved fundmental studies. This involved several innovations. Firstly, we needed to develop a highly effective lipid-based method for passivating nanochannels (***Figure 6***) so that enzymes would not stick to the interior walls of the channels (*Persson et al., 2012*[14]).

a) **b)**

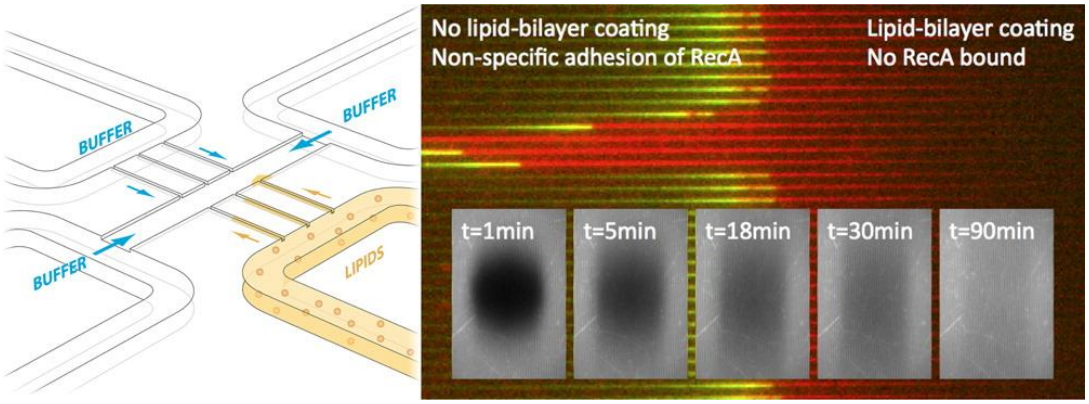

Figure 6*: Lipid passivation of micro- and nanochannels. (a) Schematic overview of the device. Four microchannels are used to bring in reagents to the nanofluidic structures in the center. In the illustrated scenario the right microchannel contains lipid vesicles and is coated with a lipid bilayer (LBL) that spreads against a fluid flow into the nanochannels and the slit. (b) LBL prevents non-specific sticking of recA protein in nanochannels (b inset) FRAP demonstrates the fluidity of the LBL in the nanochannels. Solid line: time dependence of the fluorescence of the center of a photobleached spot (10 µm radius) in an array of 150 × 110 nm2 nanochannels, coated with a fluorescent LBL. The four images are recorded at times indicated by the arrows along the axis.*

Secondly, we developed means to exchange reagents inside nanochannels while the stretched DNA remained in place. Thirdly we developed reaction and imaging conditions to detect the incorporation of fluorescent nucleotides into DNA stretched in the nanochannels (***Figure 7***).

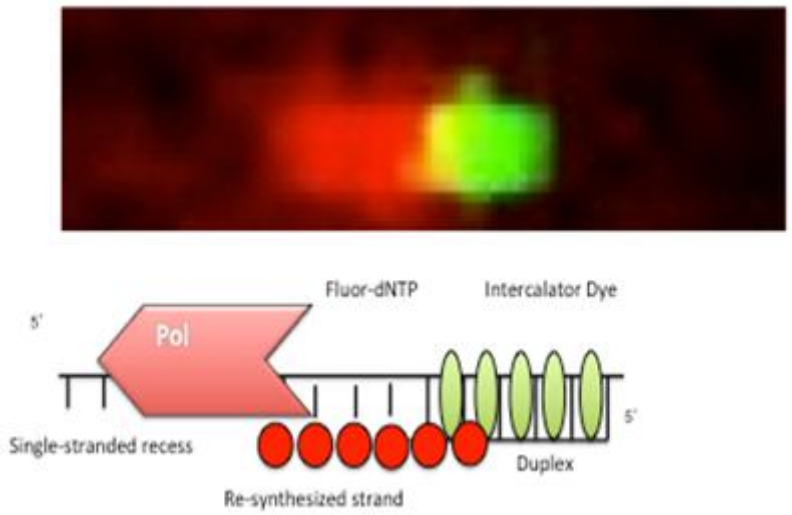

Figure 7*: Template-direct DNA synthesis via polymerase on the lambda phage genome stretched in nanochannels.*

This forms the basis for conducting either the sequencing biochemistries that we have explored in this work package or methods developed by others. In particular, because the DNA molecules are stained with an interacting dye to permit their imaging, using the intercalator as a FRET donor to nucleotides bearing base-specific labels is an attractive way forward. To this end we have demonstrated the detection of nucleotide incorporation into a stretched molecule via FRET from intercalator dye as donor. These accomplishments provide a firm basis for commercial development of stepwise or real-time sequencing on DNA stretched in nanochannels. We have also explored various means for nanometric resolution and localization. ***Figure 8*** shows the resolution of signals from a dense field which will be needed to resolve sequence information at a plurality of sites along stretched DNA.

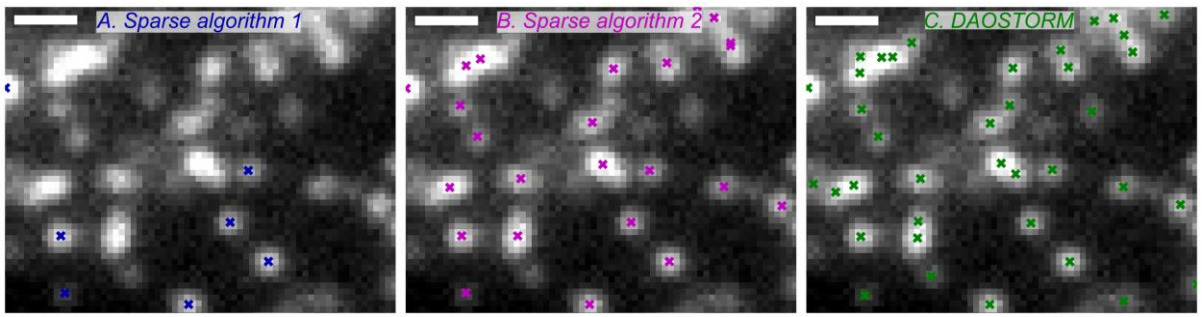

Figure 8*: Visual comparison of algorithmic performance on a single image. Each super-resolution image analysis algorithm was applied to a single image of fixed COS-7 cells with Alexa647-stained microtubules under photoswitching conditions. Crosses show localizations for each algorithm. Scale bar*

<u>WP4: Nanopore Sequencing</u>

WP4 was designed to work towards achieving two overarching objectives. The first objective was to investigate techniques that would reduce the overall cost and increase the speed at which DNA sequencing can be performed. The second objective was to develop the technological underpinnings for a small-scale/portable device for analysing specific clinically

relevant sequences, where 'portable' envisioned the use of a device outside of the classical centralized laboratory. Three consortium members, Oxford Nanopore Technologies, University of Oxford and Delft University of Technology, worked together on three key projects designed to achieve these objectives.

The first 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 that would pass sequentially through a nanopore detector appropriately modified for recognition of nucleosides. The second project focused on 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. The third project aimed to explore the utility of direct sensing enabled by nanopores for the measurement of modifications on DNA, on the basis that these modifications are critical to proper functioning of the genome and such analysis may be vital for true clinical assessment of disease or drug response.

Overall, the Workpackage has been highly successful in achieving many of its goals whilst working towards the governing objectives. The funding also enabled profitable collaboration between the two academic groups and the SME Oxford Nanopore Technologies, whereby the academic groups were able to take highly significant scientific steps whilst the company was able to leverage its relationship to expand its efforts, make remarkable technical progress and become closer to achieving commercially viable products. Progress in each of the project areas is summarized below:

Project 1: A Novel Sequencing Chemistry

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, originally conceived in the mid-1980s. However, the original designs suffered from the requirement to label every base as well as 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 measure small molecules directly, and would enable nanometer scale localisation of the enzyme next to the detector through the coupling of the two proteins components together.

The first goal was to attach cyclic adaptors to the interior of a protein nanopore based on the extensive work by Prof Hagan Bayley (University of Oxford) on the use of non-covalently bound cyclodextrins to enable detection of a wide range of molecules, including monophosphate nucleosides. This important first step was achieved after a thorough investigation of the method and site of attachment of the cyclodextrin adaptor at a single unique point within the pore. In order to achieve recognition of the bases passing through the protein pore α-Hemolysin, a particular mutant was found to be needed. This work led to a key publication in 2009 (*Clarke et al., 2009*[15]).

Considerable refinement of nanopore-based measurement techniques were 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 further distinguishing methylcytosine (mC) and hydroxymethylcytosine (hmC). However these techniques typically use non-physiological conditions and therefore are incompatible with most enzyme activity. Therefore a follow on goal was to discover suitable enzymes with required activity under nanopore measurement conditions, and once this was achieved, to attach the enzyme to the pore appropriately whilst retaining activity of both enzyme and pore.

Many enzymes and mutants were screened before one group was selected for attachment. Novel chemical coupling techniques were developed enabling the efficient placement of the preferred enzymes such that the active site was ~1nm 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 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 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 the nanopore after digestion, it was likely that the efficiency of binding and/or detection by the cyclodextrin-modified pore was too weak. Therefore, despite the promising initial technical progress, further work would be required to significantly improve the detection efficiency before optimisation of sequencing might be able to begin.

Project 2: High Density Arrays of Nanopores

Despite the failure to complete the sequencing chemistry as originally envisioned in Project 1, the capability of nanopores to recognize and discriminate DNA bases became clear. This gave the consortium members confidence to continue to develop technologies that would enable a highly scalable array of nanopores to increase the analysis power of any future device.

One approach used was to integrate protein nanopores into apertures drilled into solid state materials, with the prospect, once realized, of creating incredibly robust arrays. To achieve this, mutant forms of alpha 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 2-4nm diameter fabricated in silicon nitride using an applied potential to create an electrophoretic force. Once localised, electrical properties of the bionanopore hybrid were analysed and shown to have similar behaviour to hemolysin in lipid bilayer systems. Functionality of the hemolysin was finally probed by adding ssDNA polymer templates to the cis side of the bionanopore hybrid and demonstrating 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. This work was published in Nature Nanotechnology in December 2010 (*Hall et al., 2010*[16])

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 sub-cellular activity, such as amplification or transcription. Although beyond the original scope of the Project, 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. (*Villar et al., 2011*[17])

Independently from this Workpackage, but 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 nanopore arrays led ultimately to the unveiling at The American Society of Human Genetics in November 2012 of the company's prototype nanopore array instrumentation, designed to scale from the small, portable, USB-connectivity enabled device (MinION ™) through to the networked nodes using server-type installations for ultra-powerful analysis (GridION ™).

Project 3: Nucleic Acid Modification Analysis

With the unique ability to directly sense chemical structure or composition, nanopore detection has the potential to provide DNA modification analysis, and significant effort went towards the realization of this potential. Further work on mutating alpha hemolysin ultimately showed that individual bases on intact strands of DNA could be resolved and discriminated when held stationary inside the pore. Through painstaking optimisation, it was demonstrated that not only could the 'standard' four bases be determined, but also mC and 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, and may shed important new light on the origins of human disease and drug response. (*Wallace et al., 2010*[18]).

This work on DNA was then subsequently followed by an extension to the technique to analyse RNA, and it was further demonstrated that there too, nanopores have the unique ability to resolve 'normal' and 'modified' RNA bases on a static strand. (*Ayub and Bayley, 2012*[19]).

WP5: New Genotyping Challenges

The work in this WP was devoted to solving genotyping and epi-genotyping challenges, where the performance of current technology is not satisfactory. Existing nucleic acids analysis methods will be adapted for DNA methylation analysis and copy number variation analysis which have to deliver very precise quantitative results. Methods have also been developed to analyse mutant nucleic acid molecules under challenging conditions such as *in situ* in histological sections or in a high background of wildtype molecules. The technologies will serve for genome-wide-, targeted multiplexed-, and clinical analysis needs.

Focus areas have been: "**CNV analysis using oligo-based dynamic array-CGH**" aiming to optimize protocols by which oligo-based array CGH can be conducted upon arrays; "**Multiplex targeted copy-number assays**" exploiting a simple electrophoretic readout of selector probes ; "**Precise copy-number assays using amplified single-molecule detection**", a padlock probe-based digital quantification approach for precise quantification of DNA copies; "**Somatic Mutation in high background**", aiming to develop a blood-based test for screening for somatic mutations associated with cancer based on Pyrophosphorolysis-Activated Polymerization (PAP); "***In situ* genotyping**", a padlock probe-based approach to detect mutations in tissue sections; "***In situ* sequencing**", a "fourth generation" sequencing technique that allows sequencing of RNA in fixed cells and tissue sections; "**Haplotypes**", aiming to develop a method for molecular haplotyping ; "**Genome-wide DNA methylation analysis**", aiming to develop a novel approach for genome-wide unbiased differential DNA methylation analysis; "**Multiplex targeted epigenetic assays**" using a selector-based approach for analysis of a large set of diagnostically relevant CpG methylation sites.

The efforts towards *in situ* sequencing in cells and tissue have reached a proof-of-concept stage. The work in the WP has been well integrated and the SME Olink has actively collaborated with several partners in the consortium. Olink and UU have spun out the SME Q-linea (www.qlinea.com) that commercialises the digital RCA approach pursued in this WP, initially for biodefence applications. We have developed a multiplex typing procedure for rare somatic variants in circulation and for targeted multiplex methylation profiling. A unique approach to localized mutation detection in histology sections has been developed and published. The work in this WP aims to take the technologies to diagnostic applications.

The major achievements of WP5 are summarized below.

i) Precise quantification using amplified single molecule detection.

This task aimed to develop the homogenous amplified single-molecule detection (HASMD) format (*Jarvius et al., 2006*[20]) started in the FP6 project MOLTOOLS for application in routine diagnostic copy-number analysis. The ASMD assay is initiated by specific padlock probing of the target molecules, followed by amplification of reacted probes. Reacted probes are then converted from individual target recognition events to the formation of fluorescent micrometer-sized DNA molecules which are detectable and countable in a newly developed optical instrument. The specificity of the padlock probes allows for discrimination of single-nucleotide variants which in combination with ASMD, results in an analytical sensitivity similar to the best PCR-based approaches. The ASMD assay is less sensitive to many known inhibitors in the sample matrices. Furthermore, the ASMD approach is perfect for multiplexing, has high throughput and combines nucleic acid and protein analysis on the same platform. The digitally amplified single-molecule detection instrument prototype was developed by SME partner Olink and UU developed padlock probe assays for the system.

The instrument is a detector module capable of continuously measuring a new sample every two minutes. The instrument design is based on technologies used in confocal fluorescence microscopy and flow cytometry, but utilizes high-power solid-state lasers, line illumination, line detectors, and high optical efficiency components. The dedicated instrument enables an

increased sampling rate of about 50x compared to the confocal microscope (*Göransson et al., 2012*[21]). Multiplexing is possible with the present three laser setup. The samples are pumped through a microfluidic CD containing multiple channels, and a channel is switched by just turning the CD. Initial work on developing the digital analysis system was evaluated on a dilution series of bacterial DNA to establish limits of detection and quantitative precision. The precision of the instrument was less than 1.7 % (CV), determined in the form of a RCA product dilution series.

We have applied the padlock probe based ASMD approach for non-invasive prenatal diagnostics. Our aim was to precisely diagnose trisomy 21 and 18 in first trimester samples. The approach is based on the enrichment of cell-free fetal (ccf) DNA with methylation-sensitive restriction enzymes. Fifteen *Hpa*II sites on each chromosome were selected that are hypermethylated in fetal DNA. Fetal DNA was enriched upon restriction digestion due to *Hpa*II's inability to digest methylated DNA. Padlock probes were hybridized and ligated to its complementary target and were amplified by RCA. Trisomies were determined by calculating the ratio of RCPs received from chromosome 21 and 18. Using this approach we could detect as low as 300 genomic equivalents which would make this technique suitable for detection of ccf DNA in non-invasive prenatal diagnostics. Furthermore, trisomic chorionic villus samples (CVS) could be discriminated from normal disomic CVS. The methylation sensitive restriction digestion was shown to enrich for fetal DNA over maternal DNA by a factor of 5. This methodology is moving to clinical validation.

ii. In situ genotyping and sequencing.

The aim of this activity was to develop a technique that will enable *in situ* genotyping and sequencing of transcripts in fixed cells and tissue. The approach is based on target-primed RCA of padlock probes that we initially developed for genotyping of mitochondrial DNA. We have now adapted this protocol for detection of cDNA synthesized *in situ*. We have found conditions that allow us to detect beta-actin transcripts with up to 30 % efficiency in fixed cells. We have also been able to detect a single-nucleotide difference between the alpha- and beta-actin transcripts in fresh-frozen mouse tissue sections, thus demonstrating that it is possible to detect a single-nucleotide variation between two transcript sequences, which is the basic requirement for *in situ* genotyping reactions. Finally, we were able to detected four different transcripts *in situ* in multiplex and detected *KRAS* mutations in cell-lines (*Larsson et al., 2010*[22]). Our *in situ* genotyping approach utilizes the target strand as a primer for the localized RCA and it is thus important that there is a free 3'end close to the padlock probe binding site. We have therefore developed a strategy to cut the target strand in a site specific manner using a combination of MutY endonuclease and AP-lyase activities (*Howell et al., 2010*[23]).

We have applied the *in situ* genotyping approach for detection of the seven most common mutations in codons 12 and 13 of the *KRAS* transcript in fresh frozen tissue sections and we have adapted this technology to FFPE material, which is the most important source of clinical material for diagnostics. The *KRAS* mutations are important diagnostic markers of EGFR inhibitor treatment response, and this assay format should have several advantages to current PCR-based diagnostics, e.g., being able to detect mutations also in tissue specimens with very low *KRAS* mutation positive cell content. We have applied the above described *in*

situ KRAS mutation detection assay, as well as assays targeting EGFR and TP53 mutations, in a large series of clinical samples with previously know *KRAS* mutations. We have also applied the *KRAS* mutation assay on 41 prospective cases where the *KRAS* mutation status was determined after our *in situ* analysis. In both series of samples the concordance was 100% with the IVD approved pyrosequencing assay. This work has been submitted for publication (*Grundberg et al submitted*).

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 will introduce cDNA sequence into the RCA products. The RCA products were then sequenced by ligation, essentially according to *Shendure et al., 2005*[24]. We have successfully sequenced four bases in the human and mouse beta-actin transcript (*ACTB*) in cell lines and also *HER2* and *ACTB* transcripts in a HER2 positive breast cancer tissue. We have also utilized *in situ* sequencing 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*. 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 consider it reliable. The staining patterns are clearly not random. Pair-wise analyses of transcript expression showed that genes like *CTSL2*, *EPCAM* and *MUC1* were co-expressed with *HER2* exhibiting a high extent of spatial correlation. We also observed that *VIM* expression by large showed a reverse staining pattern compared to that of *HER2*. Co-expression of *CTSL2*, *EPCAM*, *MUC1* and *HER2* is expected in the cancer cells, while *VIM* is expected to be expressed in the stromal cells, and cancer cells undergoing epithelial-mesenchymal transition (EMT). Histologic correlation to the H&E staining showed that the *HER2* expression indeed was localized mainly to the cancer cell compartment, while *VIM* stained infiltrating lymphocytes and other components of the stroma. We also found some other staining patterns that do not directly mirror the *HER2* or *VIM* staining, such as *CD68*, a macrophage marker that only stains a subset of the stromal cells. Some of the genes, such as the proliferation markers *KI-67*, *CCNB1 (*cyclin B1), *CCND1* (cyclin D1) and *BIRC5* (survivin), show different expression levels in different parts of the cancer compartment as well as in stroma. A manuscript describing the *in situ* sequencing approach is in the process of publication (*Ke et al.*). 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.

Cytosine methylation in CpG islands in gene promoters is a hallmark of gene activity. Differential DNA methylation can be celltype-specific. Aberrant methylation is strongly correlated with deregulation, in particular present in cancer. In this workpackage we established pipelines for partial and whole genome DNA methylation analysis. MeDIPseq, RBS and whole genome bisulphite sequencing was established on two sites (CEA-CNG and PCB-CNAG). The major achievements consisted of stabilizing and standardizing the

laboratory protocols, including process control standards and establishing quality control measures for the entire pipeline. Products were analysed using 2nd generation nucleic acid sequencers. Critically, we also established data analysis pipelines that take account issues of the reduced sequence complexity of bisulfite converted DNA, the doubling of the size of the reference genome because of the conversion and the tracking of the experimental controls. Computational efficiency of the alignment is critical for bisulfite converted DNA. These developments greatly benefitted from the development of new alignment software also carried out in READNA.

Another development of this WP was based on the work done on ribo-PCR in WP2. We extended that approach to a single-tube multiplex genotyping assay. Allele-distinction is achieved using a PAP process developed with new, 3'blocked primers and a suitable polymerase system. This is the only single-tube, single-step multiplex genotyping assay to date. We applied to several different relevant markers, such as the pharmacogenomic markers for statin-use in SLCO1B1 (*Mauger et al, 2012*[25]). We further developed the single-tube procedure for the identification of ten different *KRAS* variants. We could show that the procedure worked, however, it did not quite reach the expectations of specificity and sensitivity that we had set for ourselves at the outset. Our target had been to be able to discern 1 in a million mutant variants. The assay works and could be applied to biopsy material, but is not sufficient for detection in circulating DNA straight from plasma.

IV) The potential impact (including the socio-economic impact and the wider societal implications of the project so far) and the main dissemination activities and exploitation of results.

WP1: Near Term Innovations: Ancillary elements for 2nd generation DNA sequencers

The potential impact of the advances made in WP1 are primarily on improving the speed, accuracy and cost efficiency of genetics research, and commercialization opportunities associated with this. Furthermore, the discovery that certain (TUF) genome regions are largely refractory to analysis, probably means that such domains have been under-represented in the content of commercial platforms and so hidden from modern omics analysis. Equally, the weak signals that would have been generated from such regions (from sequencing or any other forms of analysis) are more likely to include errors. Therefore, one impact of this work could be that previously hidden or misinterpreted gene-disease relationships could start to be more fully and faithfully revealed, with possible relevance to healthcare and drug development.

Our work on the benchmarking, standardization and definition of accurate nomenclature of enrichment technologies will have far-reaching impact in the scientific community, provides a solid framework for other researchers to base their studies on and to report results correctly.

The main commercial impact of WP1 relates to the Selector technology. It is now commercially available as the HaloPlex kit from Agilent. Agilent acquired the technology by acquiring the company Halo Genomics (www.halogenomics.com; previously Olink Genomics), which is a spinout company from partners UU and OLINK with currently 12 employees. The Halo Genomics site in Uppsala is now an R&D site within Agilent's genomics division.

Finally, the ethical dimensions of this work need to be elaborated, as issues of DNA sequencing impinge upon donor confidentiality, privacy and individual protection. Improved sequencing capabilities naturally increase the need to address these issues properly. It is generally assumed by researchers that information generated by sequencing projects will not be used for negative means – but current public opinion is confused with topics such as the creation of a genetic ID card. Strict guidelines are therefore required to ensure that certain companies (insurance and employers) do not take advantage of the information publicly available to avoid genetic and racial discrimination.

WP2: Near Term Innovations: Improvement and extension of existing methods

i) ribo-PCR MS sequencing

The ribo-PCR MS sequencing method has potential for sequencing 10s of thousands of short fragments of 500bp. The basis of this is published (*Mauger et al., 2012*[26]). In addition one of the inventors of ribo-PCR MS sequencing obtained her PhD thesis on the development and optimization work done on the technique (Mauger F. ***Development of a method for DNA***

analysis using cleavage of RNA/DNA chimera and MALDI-TOF mass spectrometry). The number of potential applications is numerous and is summarized in ***figure 9***.

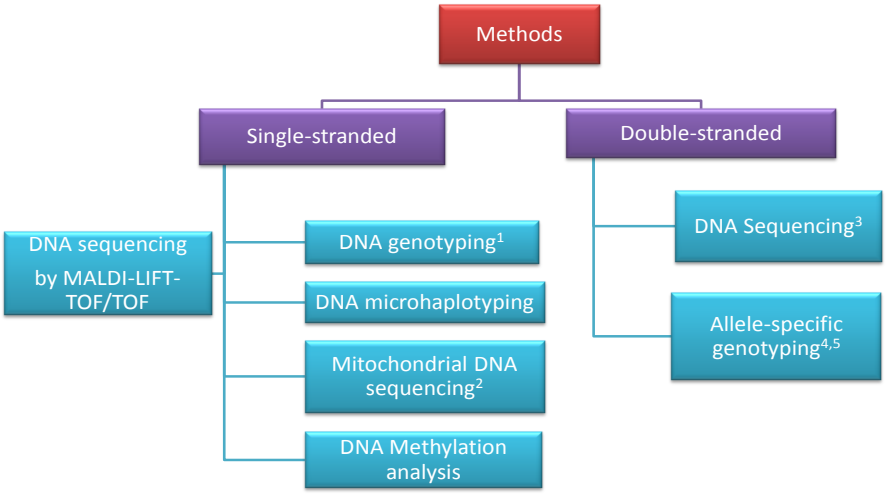

Figure 9*: Schema of ribo-PCR technique and associated applications*

The ribo-PCR concept was also drawn further in WP5 with a single-tube, multiplex genotyping assay. A patent for this technology has been granted. It has commercial potential as a stand-alone test of a low number of variants. Efforts are currently underway to identify commercial opportunities.

ii) Dynamic Allele-Specific Hybridisation (DASH)

The DASH technique should impact many significant diagnostics areas, given its robustness, its speed and ease of operation (two hour process, simple component steps, standard run conditions, and automated data analysis), its low running cost, its semi-quantitative nature, and its far greater fidelity (ability to detect trace species in mixed and mutated DNAs) compared to all other similarly convenient methods. Examples include: rapid detection and genetic characterization of pathogens (e.g., monitoring environments for disease causing bacteria, and detecting low level drug resistance viral mutations in HIV patients); mixed species purity analysis by fingerprinting complex genomes via repetitive element analysis; detecting negatively prognostic mutant clones in cancer biopsies; non-invasive prenatal diagnostics upon trace amounts of fetal DNA present in maternal serum (e.g., for paternity testing by resolving complex HLA variants in the 1-10% abundance range, and Mendelian gene mutation screening in cases with a risk of inherited disease); and many more situations. Therefore, to disseminate and promote this method, and to seek deployment partnerships in various domains of use, we have presented our progress at numerous meetings and workshops (poster and oral formats) and patented the core inventions in a range of countries ("Detection of nucleic acid polymorphism" and "High Multiplex nucleic acid amplification").

iii) Risk profiling of multifactorial diseases by MALDI mass spectrometry

The optimized protocols have shown that mass spectrometry can analyse SNPs, CNVs, transcripts of target markers and proteins simultaneously to allow effective risk profiling of multifactorial diseases. The described integrated mass spectrometry approach turned out to be in particular useful for microbial diagnostics applications. This methodology is revolutionising routine diagnostics laboratories, which are presently largely replacing more expensive automated biochemical test panels by cost efficient and facile mass spectrometry methods. In addition risk profiling in inflammatory barrier diseases have all been shown to be feasible and may also be taken into the diagnostic arena. In general, for broad application of described or alternative methods the further discovery of powerful diagnostic genetic markers for complex diseases will be important. The mass spectrometry based detection of markers from several layers of biological organization - as demonstrated in this project - was very helpful to provide highly informative diagnostic read-out.

WP3: Fluorescence-based Single Molecule Sequencing

Dissemination

The dissemination of the work has involved a number of publications with several remaining; numerous presentations have been made at international conferences including significant impact in the United States, including the 2012 NHGRI Advanced Sequencing technologies workshop.

Impact

The impact of the workpackage is largely on two patents have been applied for, one concerning extracting, stretching and rescuing DNA and the second on displaying ultra-long DNA in meandering on which sequencing can be conducted in situ. A second EU project, Cell-O-Matic has been initiated where the denaturation mapping and rescue of molecule will be taken forward. Cell-O-Matic involves several READNA partners (DTU, UOXF, Phillips, Oxford Nanopores) and a company, Genotype2Phenotype currently with two full-time employees has been been set up at the Oxford Science Park to exploit some of the outcomes of the micro-nanofluidic work in READNA and which is a partner in the Cell-O-Matic project.

Applications and exploitation routes

The long-range view of the genome that we have enabled in this WP has applications as a genomic tool. Specifically, the analysis of haplotypes has applications from genealogical, medical genetics and tissue typing for transplantation. The long-range view also enables the detection of structural variation. Furthermore, the single molecule mapping methods we have developed have applications in microbial identification, which will have its most potent application to identification of hospital infections, applications which we are beginning to move forward.

The sequencing biochemistries we have been developing could be applied in a next generation of technologies from current vendors with whom we are in discussions for their further development. Also combined with technology for extracting, handling and imaging

ultra-long DNA we provide a package of entirely new sequencing methods. The precise single molecule methods involving ALEX and FRET we have developed have applications for mechanistic studies of the enzymes involved in molecular biology and has potential application in molecular diagnostics. The algorithimic superresolution methods we established have potential for applications to a wide range of areas in biology and will also have potential impact in molecular diagnostics.

We are actively engaged in paving a path to exploitation through start-up, licensing and cooperation with companies.

WP4: Nanopore Sequencing

The potential impact of the advances made in WP4 is difficult to truly assess. Further effort is required to deliver commercial products, however the potential uncovered for nanopore analysis of nucleic acids using a highly scalable, cost effective, rapid, accurate device may lead to a widescale disruption of all existing processes and workflows used in genetic research. Once adopted beyond the research setting into the applied clinical setting, the implications could be incredibly far reaching, not only enabling a new generation of sequencing device capable of almost instantaneous result but also one that is so cost effective as to enable broad implementation in many aspects of healthcare as well as other industries. As with the other genetic techniques developed by this consortium, such technical advances come with ethical implications for management of genetic information and use by society for improved treatment of diseases and improved lifestyles by society members.

Primary dissemination activities have come from publications of the results of the work, as well as presentations at public symposia and conferences. Oxford Nanopore continues to push forward its developments for commercial exploitation of nanopore technology, and its growth from <20 staff at the initial conception of READNA to >120 today may be indicative of its success so far in that regard.

WP5: New Genotyping Challenges

The main commercial impact of WP5 thus far relates to padlock probe technology. To ensure optimal commercial outcome of the diverse potential project results from READNA building on IP owned or controlled by partner Olink, different pathways towards commercialization were explored. Methods for *in situ* analysis, such as padlock probing *in situ*, were kept within Olink while other aspects were evaluated for licensing, partnership or other means towards commercialization.

ASMD: Olink early-on established a spin-out company, Qlinea AB (www.qlinea.com), fully dedicated to the commercialization of the ASMD-platform. All rights for the ASMD-method were however retained within Olink. Olink's work within READNA was focused on molecular optimization and design of microfabrication-structures and then implemented on the instrument developed at Qlinea. In 2011 Olink made a licensing agreement with Qlinea allowing Qlinea to fully exploit the ASMD-technique on their platform. The licensing agreement allowed Olink to put the work performed within READNA in a commercial setting

capable of bringing the results generated within READNA to the market. Qlinea aims to put an MDx product, utilizing ASMD, for hospital based measurement of sepsis in 2015. Qlinea has currently 12 employees.

Padlock *in situ*: The results within READNA formed the basis for further work that lead to a patent application by researchers at partner UU, acquired and filed by partner Olink. The patent applications allow Olink to develop a completely new product line with great potential. The *in situ* mutation detection technology is completely unique on the market and enables molecular pathology at a much greater resolution that will address tumor tissue heterogeneity that is probably an important factor for selecting the right combination of molecularly targeted drugs. This is currently largely an entirely unmet clinical need for a technology, such as the *in situ* mutation detection, that can find minority populations of mutated cancer cells, and that can determine in what constellation mutations exist in cancer cell sub-clones in the tumor. Olink have now initiated work with the aim to commercialize the outcome of the results obtained within the READNA project.

The analytical pipelines for DNA methylation analysis that were established at the CEA-CNG and the PCB-CNAG allowed the integration into other EU FP7-funded projects, such as the European project of the IHEC, BLUEPRINT and IBD-CHARACTER. It has also made the genome centers running these pipelines attractive collaboration partners for high-profile research projects.

The flag-ribo-PAP genotyping system that was developed has commercial potential for high-volume genotype screening applications. IP for this technology has been granted and the CEA is currently looking for a commercial partner for this system.

WP6 : Training and dissemination

This WP ensured that the project has been disseminated in an efficient manner. It organised 4 large workshops, 4 training workshops and 10 mobility awards. In addition the WP has diffused project results to a large scale audience using scientific publications, presentations at international conferences and by the project website. The WP has produced a project flyer, press releases and a video that promote project activities.

References

1. Veal, CD et al. A mechanistic basis for amplification differences between samples and between genome regions. BMC Genomics 13(455), 2012.
2. Johansson, H., et al. Targeted resequencing of candidate genes using Selector Probes. Nucl. Acids Res. 39 (2):e8, 2011.
3. Mertes, F et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Briefings in Functional Genomics 10(6):374-86, 2011.
4. Mauger, F et al. DNA sequencing by MALDI-TOF MS using alkali cleavage of RNA/DNA chimeras. Nucleic Acids Research 35(8):e62, 2007.
5. Jobs, M et al. DASH-2: flexible, low-cost, and high-throughput SNP genotyping by dynamic allele-specific hybridization on membrane arrays. Genome Research 13(5):916-24, 2003.
6. Prince, JA et al. Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation. Genome Research 11(1):152-62, 2001.

7. Sauer, S et al. Single-nucleotide polymorphisms: analysis by mass spectrometry. Nature Protocols 1(4): 1761-71, 2006.
8. Weidner, C et al. Amorfrutins are potent antidiabetic dietary natural products. PNAS 109(19): 7257-7262, 2012.
9. Mir, KU. Sequencing by Cyclic Ligation and Cleavage (CycLiC) directly on a microarray captured template. Nucleic Acids Research 37(1):e5, 2009.
10. Santoso, Y et al. Conformational transitions in DNA polymerase I revealed by single-molecule FRET. PNAS 12;107(2):715-20, 2010.
11. Lereste, L et al. Characterization of Dark Quencher Chromophores as Nonfluorescent Acceptors for Single-Molecule FRET. Biophysical Journal 102(11):2658-68, 2012.
12. Rasmussen, KH et al. A device for extraction, manipulation and stretching of DNA from single human chromosomes. Lab on a Chip 11(8):1431-3, 2011.
13. Reisner, W et al. Single-molecule denaturation mapping of DNA in Nanofluidic Channels. PNAS 107(30): 13294–13299, 2010.
14. Persson, F et al. Lipid-Based Passivation in Nanofluidics. Nanoletters 12(5):2260-5, 2012.
15. Clarke, J et al. Continuous base identification for single-molecule nanopore DNA sequencing. Nature Nanotechnology 4(4):265-70, 2009.
16. Hall AR et al. Hybrid pore formation by directed insertion of α-haemolysin into solid-state nanopores. Nature Nanotechnology 5(12):874-7, 2010.
17. Villar, G et al. Formation of droplet networks that function in aqueous environments. Nature Nanotechnology 6(12):803-8, 2011.
18. Wallace, EV et al. Identification of epigenetic DNA modifications with a protein nanopore. Chemical Communications 46(43):8195-7, 2010.
19. Ayub, M et Bayley, H. Individual RNA base recognition in immobilized oligonucleotides using a protein nanopore. Nanoletters 12(11):5637-43, 2012.
20. Jarvius, J et al. Digital quantification using amplified single-molecule detection. Nature Methods 3(9):725-7, 2006.
21. Göransson J et al. Rapid identification of bio-molecules applied for detection of biosecurity agents using rolling circle amplification. PLoS One 7(2):e31068, 2012.
22. Larsson, C et al. In situ detection and genotyping of individual mRNA molecules. Nature Methods 7(5):395-7, 2010.
23. Howell WM et al. Glycosylases and AP-cleaving enzymes as a general tool for probe-directed cleavage of ssDNA targets. Nucleic Acids Research 38(7):e99, 2010.
24. Shendure, J et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309(5741):1728-32, 2005.
25. Mauger, F et al. High Specificity Single-Tube multiplex Genotyping Using ribo-PAP PCR, Tag Primers, Alkali Cleavage of RNA/DNA Chimeras and MALDI-TOF MS. Human Mutation 34(1):266-73, 2012.
26. Mauger, F et al. Ribo-polymerase chain reaction-a facile method for the preparation of chimeric RNA/DNA applied to DNA sequencing. Human Mutation 33(6):1010-5, 2012.
27. Elsharawy, A et al. Accurate variant detection across non-amplified and whole genome amplified DNA using targeted next generation sequencing. BMC Genomics 13:500, 2012.
28. Elsharawy, A et al. Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing. BMC Genomics 13:417, 2012.
29. Forster, M et al. From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. Nucleic Acids Research 41(1):e16, 2013.

V) The address of the project public website, if applicable as well as relevant contact details.

www.cng.fr/READNA