Preparing and Sequencing Ultra-Long DNA Molecules from Single Chromosomes



David L.V. Bauer

St Edmund Hall

University of Oxford

Thesis Submitted in Partial Fulfilment of the Requirements
for the Degree of Doctor of Philosophy at the University of Oxford

Trinity Term 2013

Abstract

In this thesis, I describe the development of a single-molecule platform for analysing long DNA molecules that captures haplotype and large-scale structural variation (SV) in addition to DNA sequence. Current DNA sequencing methods cannot adequately examine haplotype and SV – both contribute to biological function and disease and are candidates for the location of "missing heritability" in the genome. Both haplotype and SV fundamentally relate to the structure of single chromosomes.

Using a lab-on-a-chip nanofluidic device, SV was analysed on stretched (> 2 Mb) DNA fragments. In order to integrate this larger-scale SV information with the base-by-base sequence of the molecule being analysed, the DNA molecule was amplified and sequenced. I developed algorithms to handle the unique features of sequence data from amplified single DNA molecules. I obtained sequence and genotyping data, confirmed successful isolation of single DNA fragments from the chip, and validated the barcoding method used to detect SV.

This lab-on-a-chip device for handling long DNAs can also serve as a 'reaction chamber' to answer more fundamental biological questions regarding chromosome structure as a whole. Using a microfluidic chip, I was able to provide the first direct images of DNA catenation within metaphase chromosomes and demonstrate that DNA catenation, in addition to proteins, plays a crucial role in metaphase chromosome architecture.

The fluidic platform can be adapted to future 'third-generation' single-molecule sequencing applications that interrogate single DNA molecules directly. I have demonstrated this potential in two ways: First, I used intercalating dyes to form an optical waveguide along DNA to improve single-molecule detection. Secondly, I engineered *E. coli* RNA Polymerase to detect single base translocation events along a DNA substrate. Such a polymerase could be used in future third-generation sequencing schemes based upon base-stepping motion or energy transfer to dye-modified nucleotides as the polymerase processes on a long DNA template.

Acknowledgements

I am immensely grateful to Achilles Kapanidis and the entire Kapanidis group for providing a welcoming second ‘home’ for my DPhil in Clinical Medicine in Condensed Matter Physics (!) over the past two years. I’ve loved masquerading as a biophysicist for every second.

I am indebted to Kalim Mir and Achilles Kapanidis for their guidance and supervision over the past four years, as well as to Keith Gull for his ever-valuable advice (whether re. career or allotment!) and Don Markwell for the meet ‘n mingles (no, I don’t want another soggy chip!). While I have acknowledged contributions to research at the end of each chapter, I would particularly like to thank Giuseppe Scozzafava, Johannes Hohlbein, and Tim Craggs for their help over the past four years and am fortunate to have them as friends.

While snowy winter months (and one week in May!) in Denmark were cold, I was very glad to be working with Rodolphe Marie and Kristian Rasmussen, whose good company and humour made the visits all the more worthwhile. Back in Oxford, I have been kept good company by my officemates-in crime: Louise Aigrain, Alexandra Tomescu, Marko Sustarsic, Nicole Robb and Anne Plochowietz. Thanks! I also owe a direct ‘thank you’ to those who contributed comments on various chapters of this thesis: Johannes Hohlbein, Timothy Craggs, and Nicole Robb. Lastly, I’ve been fortunate to cross paths in Oxford with Vivek Naranbhai (never enough chai!), Keiko Yata, and Susan Fotheringham. My thanks go to them for their continuing friendship over the years.

I would also like to acknowledge those who helped me to pay the bills: the Rhodes Trust, Nuffield Department of Medicine, St. Edmund Hall, and the US National Science Foundation. Speaking of bill paying, I cannot leave out my parents for their support, encouragement, and pride… whether it’s been researching brown bears at the Bronx Zoo or writing a thesis.

And, in ‘last author’ position: my wife, Anthea. Thank you for your love, your patience, and your impatience, too. (I hear it’s for my own good.) All my love, David.

Table of Contents

Chapters are paginated separately

Chapter 1	Introduction: The Unsequenced Genome
Chapter 2	Nanofluidics to Integrate Genome Structure and Sequence of a Single DNA Molecule
Chapter 3	Unfolding Chromosome Structure
Chapter 4	Using DNA as a Photonic Waveguide for Enhanced Fluorescence Imaging
Chapter 5	Engineering Polymerases using Unnatural Amino Acids
Chapter 6	Conclusion

Appendix A	Publications and Presentations
Appendix B	SV Discovered in NA18507
Appendix C	Materials and Methods

Chapter 1 – Introduction: The Unsequenced Genome

For much of the past fifty years, genomic analysis has largely made three fundamental assumptions[1-3]: (1) Each cell in the body contains an identical copy of the genome; (2) The two collections of alleles inherited act only in concert with one another, and not as distinct genotypes that each contribute to phenotype; and (3) The functional impact of genetic variation is rooted in the 'central dogma' of molecular biology – phenotypic change is effected through point mutations that affect protein coding and thus function.

While exceptions to each of these assumptions were always acknowledged (e.g. gametes, lymphocytes, and cancerous cells have different genomes; haplotype is important at the Major Histocompatibility Complex loci; chromosomal abnormalities such as trisomy 21 affect phenotype), high-throughput methods of genomic analysis still cannot adequately address (1) cell-to-cell heteterogeneity; (2) haplotype; and (3) structural variation larger than a few kilo-basepairs (Figure 1). The need to extend high-throughput genomics into these areas has been highlighted by the "missing heritability"

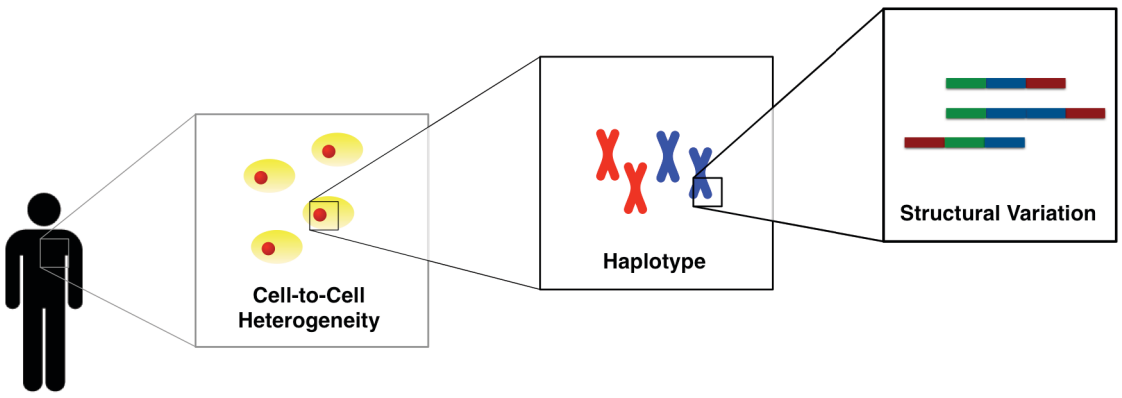


Figure 1. The full degree of human genomic variation is not fully addressed by current approaches.

problem: though thousands of personal genomes have been sequenced[4], the genetic basis of many inherited complex traits remains undetermined – from height to type 2 diabetes susceptibility[5]. Among the candidates for the origin of the missing heritability in the genome are a lack of haplotype-resolved data and a lack of high-resolution information on genomic structural variation. Consequently, there has been increased effort to develop methods to better understand the degree of variation in these three areas, its biological basis, and to develop methods to more easily identify this variation rapidly and accurately in order to correlate variation with phenotypic and clinical data[6,7].

Cell-to Cell Heterogeneity

Heterogeneity has not been widely considered as a candidate for the missing heritability, especially since it is largely viewed through the lens of cancer development – i.e. the accumulation of somatic mutations are considered to give rise to tumorigenesis[8]. Recent research has suggested that variation between the genomes of somatic cells occurs at a greater extent than has been considered previously: Abyzov and coworkers found that 30% of human skin cells contain unique somatic variants[9] and a number of studies have implicated somatic variation as a factor in neurological disorders[8]. The potential for multiple genotypes to be present within an individual limits the ability to draw biological conclusions from a sample that has been derived from one collection of cells (e.g. buccal swab) and may confound analysis of genomic data when assessing its role in phenotype or disease[1].

Haplotype

In diploid organisms like humans, one copy of each chromosome is inherited from each parent and thus the collection of alleles along one copy – a molecular haplotype – may be different from the collection along the other. Since each chromosome is maintained, transcribed and translated as a distinct molecular unit within the cell, haplotypic

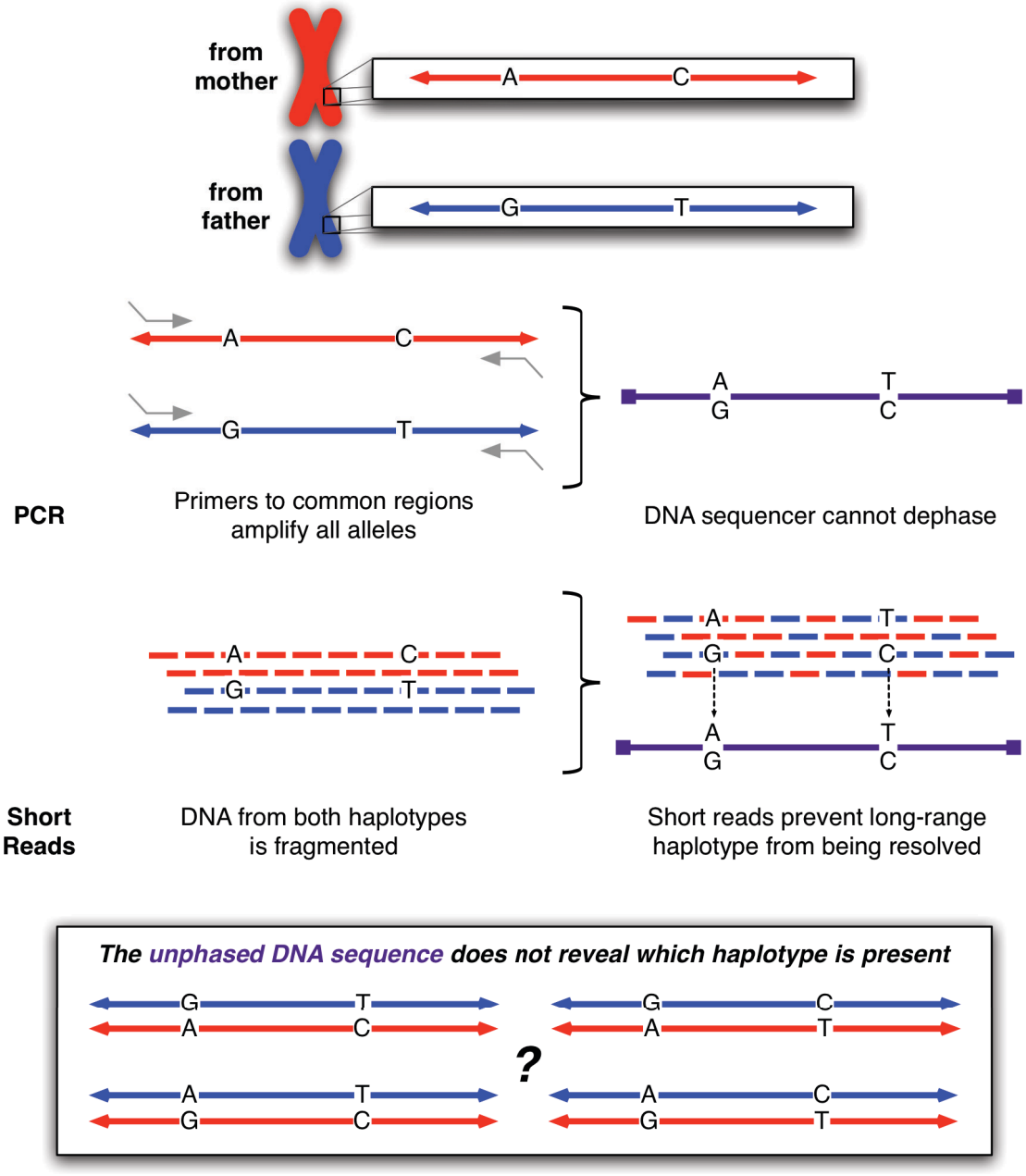


Figure 2. A simplified example of haplotype dephasing during DNA sequencing

information is fundamentally important to understanding and exploiting genomic information. DNA sequencing that is based on PCR amplification of a target or shotgun methods in which total-genomic DNA is fragmented *en masse*, disregards the fundamental biological organisation of the genome: chromosomes are not sequenced as a single unit and "phasing" of the alleles to determine the chromosome of origin is not readily accomplished, as illustrated in Figure 2.[2,10]

Large-Scale Structural Variation (SV)

The rearrangement, duplication, or deletion of large blocks of the genome (> 1 kb) was once considered to be confined to the genomes of cancerous cells, where structural variation is often extensive, and visible by microscopic techniques. This assumption was challenged by Lupski *et al.* in 1991 with the discovery that a ~1 Mbp duplication on chromosome 17 could give rise to an inherited genomic disorder, Charcot-Marie-Tooth disease type 1A[11]. Structural variation has now been found to occur 5-10 times more frequently than single-nucleotide polymorphisms (SNPs) in the genome[3,12].

SV arises from a number of molecular mechanisms, including non-allelic homologous recombination, retrotransposition, template switching of stalled replication forks, and the repair of double-strand DNA breaks via non-homologous end joining[12-15]. In some cancer genomes, extreme SV is the consequence of chromotrypsis in which the genome is 'shattered' leading to hundreds of simultaneous rearrangements[16].

'Unbalanced' SV that results in a change in copy number, such as duplications and deletions, can be detected using comparative genomic hybridisation[17,18] and by quantitation of DNA sequencing read mappings[19], but the method has limited resolution and does not provide full information about the genomic context of the variation[3]. As a result, important questions about a variant may be left unanswered, e.g. for a genomic

duplication: are copies in tandem, or scattered throughout the genome? Do they interrupt other coding genes? Are the copies placed under control of different regulatory elements – or do they interrupt the regulation of surrounding genes?

Analysing Haplotype and DNA Sequencing with Second-Generation DNA Sequencing

DNA sequencing can be used to access information regarding structural variation through end-sequence profiling[20] in which the ends of a clonal DNA fragment of a *known* size are sequenced and each read is aligned to the reference genome. If there is no structural variation is present, then the two reads will map to the reference genome with a distance between them that corresponds to size of the DNA fragment. Computational reconstruction[21] of read pairs from fosmid end-sequence profiling that map abnormally can then provide information about both balanced and unbalanced SV across the genome[22]. Haplotype information can similarly be reconstructed from the data[2,10].

The wide adoption of Solexa/Illumina second-generation DNA sequencing has provided new opportunities to examine SV and haplotype since each read produced is derived from a single molecule that has been amplified to form a cluster on a flowcell[23]. Both ends of the DNA can be sequenced to generate 'paired-end' reads that can analysed for abnormal read mappings[24] in the same manner as end-sequence profiling, using a variety of algorithms[25,26]. However, the size of structural variation and phased haplotype blocks that can be determined using paired-end reads is limited by the size of the DNA insert that can be prepared – typically < 5 kbp[27]. Consequently, the fosmid end-sequence profiling method has been used in conjunction with high-throughput

sequencing [28] and further refined to sequence the entire fosmid insert [29] to gain an increase in sensitivity.

Fosmid insert sequencing represents the current state-of-the-art for haplotype resolution and mapping of SV using DNA sequencing. The method formed the basis of the recently-published genome of the HeLa cancer cell line by the Shendure group[30], who identified that the cause of the HeLa tumour was a complex insertion of human papilloma virus (HPV) on chromosome 8 – but only on one of the haplotypes present. Notably, they found that HeLa cells did not contain any SNPs or disruptions in common oncogenes – underlining the importance of SV information when analysing genomic function and disease.

Despite the success of the fosmid insert sequencing method, there is considerable need for improvement. Used on its own, the method is still unable to resolve haplotype fully: the median size of the resolved blocks is 0.55 MBp. To complete their haplotype phasing, the Shendure group resorted to using population data sets and information from RNA sequencing of the HeLa cells. To complete mapping of SV, they combined their Illumina data with long-read data from third-generation SMRT sequencing[*]. On a practical level, the method itself is laborious since it requires the generation of fosmid libraries in *E. coli*[29]. This added cost in time and reagents may therefore limit the applicability of fosmid insert sequencing to large-scale projects instead of routine sequencing in smaller-scale projects or for routine medical diagnostic use.

[*] Third-generation DNA sequencing is discussed in Chapter 4.

Single-Molecule Methods to Address Haplotype, SV, and Cell-to-Cell Heterogeneity

Cell-to-cell heterogeneity, haplotype, and SV arise as a consequence of the fundamentally single-molecule nature of the genome within a cell; difficulty in analysing these three phenomena ultimately arises because genomics does not traditionally treat genetic material on a single-molecule level. In this thesis, I discuss work to extend single-molecule DNA analysis.

- In the second chapter, I describe the use of micro-/nano-fludic devices to isolate DNA molecules from single metaphase chromosomes, detect large-scale SV by barcoding and imaging a single DNA, and then sequence these single DNA molecules after amplification[31].
- In the third chapter, I use fluidics to examine the role of DNA catenation in the architecture of metaphase chromosomes – with implications for analysis of ultra-long fragments of DNA as well as for the biological understanding of chromosome structure[32].
- In the fourth chapter, I consider how single-molecule imaging of DNA-associated fluorophores could be enhanced by self-assembling a photonic wire waveguide along the DNA itself and demonstrate how this approach could be used in future third-generation DNA sequencing approaches.
- Lastly, in the fifth chapter, I use unnatural amino acids and click chemistry to site-specifically label RNA polymerase to detect single basepair motion on DNA.

References

1. Lupski, J. R. Genetics. Genome mosaicism--one human, multiple genomes. *Science* **341,** 358–359 (2013).
2. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
3. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
4. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
5. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).
6. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
7. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11,** 446–450 (2010).
8. Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic mutation, genomic variation, and neurological disease. *Science* **341,** 1237758 (2013).
9. Abyzov, A. *et al.* Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* **492,** 438–442 (2012).
10. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24,** i153–9 (2008).
11. Lupski, J. R. *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* **66,** 219–232 (1991).
12. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14,** 125–138 (2013).
13. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61,** 437–455 (2010).
14. Currall, B. B., Chiang, C., Talkowski, M. E. & Morton, C. C. Mechanisms for Structural Variation in the Human Genome. *Curr Genet Med Rep* **1,** 81–90 (2013).
15. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10,** 691–703 (2009).
16. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144,** 27–40 (2011).
17. Kallioniemi, A. *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258,** 818–821 (1992).
18. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20,** 207–211 (1998).
19. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470,** 59–65 (2011).
20. Volik, S. *et al.* End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad Sci USA* **100,** 7696–7701 (2003).
21. Raphael, B. J., Volik, S., Collins, C. & Pevzner, P. A. Reconstructing tumor genome architectures. *Bioinformatics* **19 Suppl 2,** ii162–71 (2003).
22. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).

23. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
24. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40,** 722–729 (2008).
25. Koboldt, D. C., Larson, D. E., Chen, K., Ding, L. & Wilson, R. K. Massively parallel sequencing approaches for characterization of structural variation. *Methods Mol Biol* **838,** 369–384 (2012).
26. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–681 (2009).
27. Schatz, M. C., Delcher, A. L. & Salzberg, S. L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20,** 1165–1173 (2010).
28. Kidd, J. M. *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat. Methods* **7,** 365–371 (2010).
29. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29,** 59–63 (2011).
30. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500,** 207–211 (2013).
31. Marie, R. *et al.* Integrated view of genome structure and sequence of a single DNA molecule in a nanofluidic device. *Proc Natl Acad Sci USA* **110,** 4893–4898 (2013).
32. Bauer, D. L. V., Marie, R., Rasmussen, K. H., Kristensen, A. & Mir, K. U. DNA catenation maintains structure of human metaphase chromosomes. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gks931

Chapter 2 – Nanofluidics to Integrate Genome Structure and Sequence of a Single DNA Molecule

Some of the material presented in this chapter has been published in:

> [1] Marie R, Pedersen JN*, **Bauer DLV***, Rasmussen KH, Yusuf M, Volpi EV, Flyvbjerg H, Kristensen A, and Mir KU. An Integrated Single Molecule View of Genome Structure and Sequence. *Proceedings of the National Academy of Sciences (USA),* 110, 4893–4898. (2013). * *denotes equal contribution*

Introduction

Difficulty in detecting large-scale genomic structural variation (SV) is fundamentally rooted in problems with scale: at the small end, the read lengths of DNA sequencing are too short when compared to the size of the structural variants under consideration. At the large end, cytogenetic techniques such as multicolour fluorescence in situ hybridisation mFISH[2] and spectral karyotyping[3] provide a large 'read length' – an image of the entire chromosome – but their resolution is too coarse to detect changes below ~0.5 Mb[4].

Optical Mapping

One approach to examine the genome at an intermediate scale is optical mapping of DNA. In this method, a DNA sample is deposited on a surface, stretched to display the DNA linearly, and a visual pattern 'barcode' is created on the DNA that reflects its underlying sequence. This pattern may be generated by a number of approaches: restriction enzymes that produce visible cuts in a DNA molecule at known (sequence-

specific) sites[5-7] as well as by multicolour approaches, including: direct hybridisation of fluorescent oligonucleotide/PNA/LNA probes[8], fluorescent tagging of bases by engineering DNA methyl transferases[9], and nick-translation with site-specific nickases either followed by hybridisation to labelled probes[10] or including direct incorporation

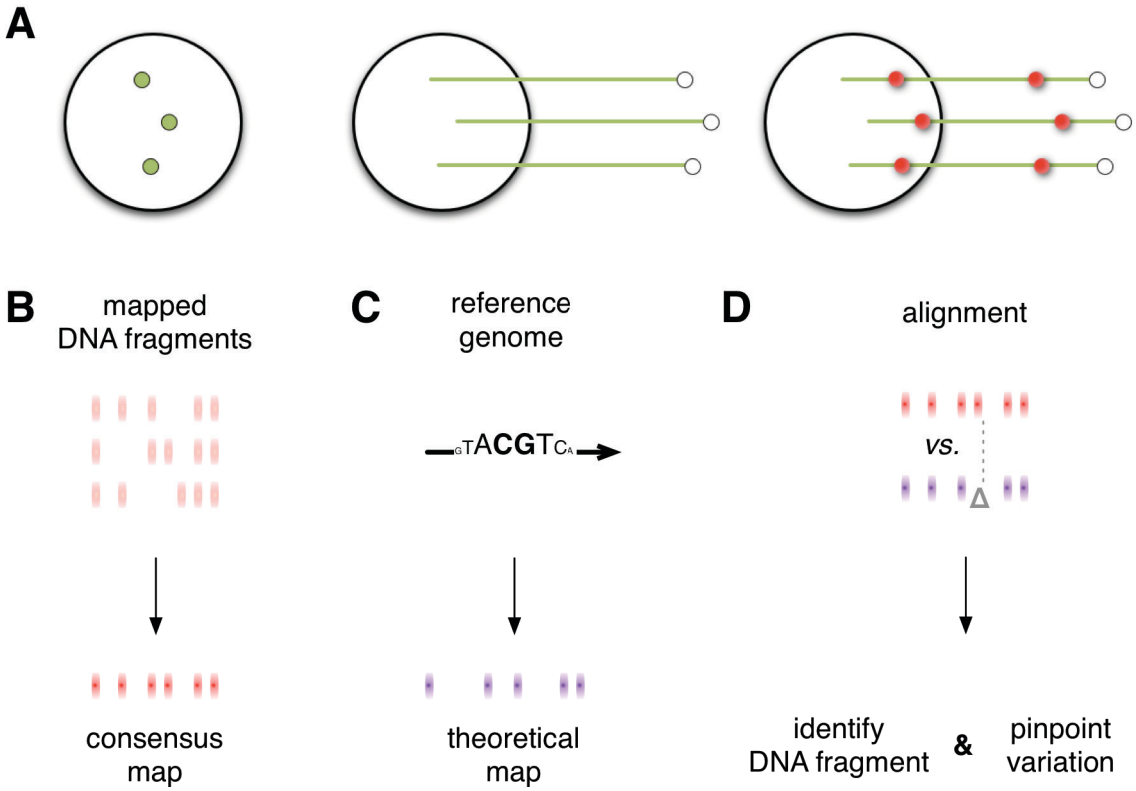


Figure 1. Schematic overview of optical mapping techniques. (**A**) A pattern is produced on stretched DNA that can be imaged using fluorescence microscopy, for example, by enzymatic labelling or hybridisation. (**B**) To account for inefficient labelling, multiple maps are aligned to produce a consensus map pattern that is analogous to a DNA 'barcode'. (**C**) The pattern expected based on the reference DNA sequence can be used to produce a 'theoretical map', which (**D**) is aligned to the consensus map. Since the barcode reflects the underlying sequence of the DNA molecule, it can serve to uniquely identify the molecule by comparing it to the barcode pattern expected from the reference genome sequence. Points of discordance (**Δ**) reflect a change in the underlying genome sequence.

of fluorescent dNTPs by polymerase during the translation step[11].

Optical mapping has found niche success in 'scaffolding' the assembly of reference sequences for organisms with complex genomes, such as the AT-rich genome of the malaria-causing *Plasmodium falciparum*[12], and the highly-repetitive mutiploid genome of maize[13]. More recently, it has facilitated *de novo* assembly of DNA sequencing reads of a large, repetitive cereal genome, *Aegilops tauschii*[14] and of the domestic goat[15]. The latter study used optical maps to *de novo* assemble short Illumina DNA sequencing reads into "super-scaffolds" that were more than 5 times larger than what was possible by paired-end fosmid sequencing alone.

While the genomics community has adopted optical mapping as a component of some large, complex projects, it does not use it as routine technique for the detection of SV – especially in medical diagnosis and clinical genome-wide association studies. Papers in the literature that report novel SV typically do not use optical mapping for validation and instead continue to rely on PCR and array-based methods[4,16,17]. Indeed, when optical mapping is employed, it is often not as a method in its own right, but as a validation tool for various novel DNA sequencing-based approaches to detect SV[18].

Part of the reason for this lack of embrace by the genomics community – despite commercialisation of optical mapping by companies such as BioNano Genomics and OpGen[19] – is that optical mapping approaches, at present, have a number of limitations that restrict the additional value they provide over conventional techniques. They:

1. Require averaging of multiple molecules since the enzymatic tagging methods have considerable false-positive and false-negative labelling rates[20,21],
2. Suffer from poor resolution due to uneven stretching of the DNA[6], and

3. Have relatively poor throughput as a consequence of repeated analysis cycles to build confidence in the data obtained[6,11].

It will be necessary to remove these limitations if optical mapping is to become a routine and useful tool in genomics analysis and diagnostics.

Application of Nanofluidics

Concerns regarding resolution and throughput have been addressed in part by placing DNA in nanofluidic channels instead of on a surface. The confinement of DNA in two dimensions to distances approaching the persistence length of the DNA polymer[22] results in the uniform stretching of the DNA along the channel to approximately 50% of its crystallographic length; confinement within a single dimension (z) produces similar effects, with the added benefit of simplified device fabrication[11,23]. These devices have been used to conduct optical mapping using restriction endonucleases[24] and fluorescent tagging[11,23].

Denaturation-Renaturation (DR) Mapping of DNA

As an alternative to enzyme-based tagging, DNA intercalating dyes have been used to produce sequence-specific patterns on stretched DNA. Melting of the DNA double helix after staining with YOYO-1 DNA intercalating dye generates a pattern related to local GC content of the DNA. This method relies upon two principles: (1) DNA melting (denaturation) is a function of the underlying GC content of the DNA[25,26], and (2) intercalating dye unbinds from a DNA strand upon DNA melting[27]. Therefore, partial denaturation of stained DNA results in GC-rich regions appearing brighter than AT-rich regions, since the former have retained more dye than the latter. These patterns, detected by fluorescence imaging of stretched DNA, reflect the underlying sequence composition and can function as a 'barcode' to identify the origin of a DNA

molecule from its image when stretched in a simple nanofluidic chip[27]. While this method avoids the complexities of enzyme-based tagging, the nanofluidic chip design employed by Reisner *et al.* was not able to stably stretch the DNA to the degree necessary for a high-resolution, unique map to be obtained from a single image of a single DNA molecule.

Integrating DR Mapping and Conventional Genomics

To make a single platform for optical mapping with broad applicability in genomics, a number of improvements can be considered: first, the system should be able to address haplotype and cell-to-cell variation in addition to the detection of SV. This ability is enabled by achieving a high degree of DNA stretching to avoid the need to average over multiple molecules. It will therefore have improved throughput and be able to address haplotype and cell-to-cell variation in addition to the detection of SV. Second, the platform must be able to handle long fragments of DNA without shearing them. Third, the system should allow tight integration with 'conventional' genomics analysis methods that provide complementary information such as DNA sequencing, genotyping and fluorescence in-situ hybridisation (FISH).

This third aspect is key: in contrast to conventional optical mapping in which DNA molecules are irrecoverably deposited on a surface, the use of a fluidic chip to perform DR mapping on unmodified DNA means that the *same* DNA molecule that is DR-mapped can subsequently be extracted from the fluidic chip, amplified, and sequenced — allowing us to 'zoom in' from the optical map to gain base-level information as well.

Challenges in Sequencing Amplified Single DNA Molecules

All DNA sequencing experiments contain inherent bias: the sequencing process itself is biased by GC content[28], and the mapping of short reads to the reference genome is confounded by repeat regions and areas of the genome with sequence homology, to which fewer reads can be uniquely mapped[29]. Unlike whole-genome sequencing data sets in which the entire genome should ideally be covered to some degree and these sequencing and mapping biases can be easily corrected, data from an isolated DNA fragment contains inherent sample bias[30] and should result in an overrepresentation of read mappings to the locus in the genome from which the fragment came. Identifying regions with an overrepresentation of reads (and thus the genomic origin of the DNA molecule) is therefore the first challenge in analysing data from single DNA molecules. The process is analogous to analysis methods for datasets derived from techniques such as chromatin immunoprecipitation – sequencing (ChIP-seq), in which the overrepresentation of reads mapping to a particular region is used to identify and quantitate the presence of specific DNA fragments[31].

Correcting for sample input bias in our case is complicated by the use of whole-genome multiple-displacement (MD) amplification, which uses phi29 DNA polymerase and random primers[32]. While whole-genome MD amplification of bulk samples with Φ29 showed that amplification bias against GC-rich regions could be corrected by employing similarly amplified controls[33], Fan and co-workers found that amplification bias was not reproducible when using single chromosomes as templates for amplification[34] and therefore could not be readily corrected for quantification.

New bioinformatics approaches are therefore necessary to examine DNA sequencing data derived from single molecule templates before being able to integrate it with complementary information from DR mapping.

Summary: Synthesis of the Project

In this chapter, I begin by presenting the development of a nanofludic chip to process whole metaphase chromosomes and carry out DR mapping to localise single DNA molecules within the genome at a resolution not previously achievable. We developed concepts and methods for using these localisations – and their discrepancies – to detect structural variation, which were then implemented by colleagues at the Technical University of Denmark (DTU) on samples from a Yoruban Nigerian individual (NA18507) and from an acute T-cell leukaemia cell line (Jurkat), a genome with instability at the cytogenetic level[1] that could not be mapped using the DR method.

In the remainder of the chapter, I present algorithms that I developed to handle DNA sequencing data obtained from amplified single molecules of DNA in order to integrate the DR map data with conventional genomic analysis. Using high-throughput sequencing, I aimed to: (1) confirm the origin of a single molecule of DNA from a single Jurkat chromosome that displayed significant structural variation at the kb-Mb level and could not be uniquely mapped to the genome as a result; (2) demonstrate the isolation of a haplotype from the DNA sequencing data; and (3), perform traditional variant discovery to find sequence-level variation in the sample in addition to the larger-scale structural variation. I conclude the chapter by attempting to place the results in the context of (limited) other reports in the literature of sequencing data derived from single cells and MD amplified DNA.

Results

Nanofluidic Chip for Chromosome Extraction and Mapping

A single lab-on-a-chip device was designed to combine DNA extraction, stretching, mapping, and recovery (Figure 2). Whole metaphase chromosomes were isolated from human cell culture and stained with YOYO-1 DNA intercalating dye; the chromosomes were loaded onto the chip and their compact nature was used a method to avoid shearing and fragmentation of DNA during pipetting and handling.

Chromosomal proteins were digested by proteinase K, DNA was partially denatured

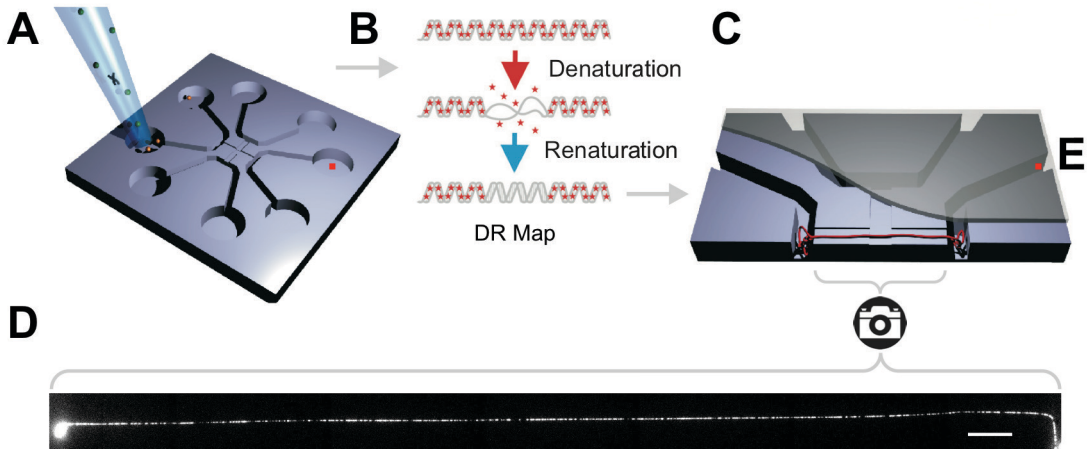


Figure 2. On-chip DNA extraction and single molecule mapping, and extraction. Stained metaphase chromosomes are delivered to the inlet port of the device, **(A)**, and proteins are digested. The DNA is patterned, **(B)**, by heating to partially 'D'enature it and rapidly cooling to 'R'e-anneal. The resulting pattern – a DR map – reflects the underlying GC content of the DNA at a given location. A single DNA molecule is then placed in a nanofluidic chamber to provide stable, high stretching, **(C)**, and imaged, **(D)**, using epifluorescence microscopy to produce a view of ~ 1.3 Mbp of DNA. The same DNA molecule can then be transferred to an extraction port, **(E)**, amplified, and analysed by conventional genomics methods, such as DNA sequencing. (Image adapted from Marie *et al.*, 2013, R. Marie, DTU)

and renatured to produce a DR pattern, and the DNA was stretched in the central nanofluidic portion of the chip. A DR map composed of light and dark regions could be observed over contiguous Mbp-lengths of each individual molecule. Light regions had been less denatured, and hence lost less intercalating dye due to higher CG content, whereas the reverse was true for the AT-rich dark regions.

A novel nanofluidic chip design contained a central 'nanoslit' which was 440 μm long, 50 μm wide, but only 85 nm high. The design allowed fluid to flow along the length of the DNA molecule in the nanoslit and the resulting hydrodynamic drag on the DNA molecule led to stretching beyond the level previously achieved by nanoscale confinement alone: the DNA molecule was stretched up to 98% of its crystallographic length (0.33 nm/bp)[35] at physiological salt concentrations, compared to ~50% previously[27]. As a result of the high degree of stretching, the DNA was held stably in place: longitudinal Brownian motion (and resultant blur) was completely eliminated. Overlapping fields-of-view could then be merged to form a single image of the entire DNA molecule in the centre of the chip without resorting to any image correction algorithms such as rescaling or morphing.

Our DR method was first evaluated by examining the well-characterised genome of a Yoruban African male (NA18507)[36-38]. DR maps of single molecules from NA18507 were compared with a theoretical DR map computed from the sequence of the human reference genome version 18 (hg18) using the *Bubblyhelix* algorithm[26]. Single DNA molecules were mapped to the genome with the algorithm used by Reisner *et al.*[27], though a single image from a single DNA molecule was used instead of an average map from 40 molecules. The algorithm returns the best alignment to the reference genome as well as a score than can be used to evaluate the confidence of the mapping[1].

For our NA18507 sample, all of the 21 molecules for which localisation to the human genome was attempted were successfully and uniquely mapped. This compares favourably with the statistics in a recent report in the literature of 'conventional' denaturation-renaturation mapping: mapping of barcodes from *Saccharomyces cerevisiae* DNA was successful on only 82% of selected individual molecules within a genome that is more than 100 times smaller than the human genome[39].

Analysis of Structural Variation by Sliding Window Analysis

Following localisation, SV can be detected by sequential comparison *in silico* of the molecule's experimental DR map to the reference 'theoretical' DR map computed from the reference genome. A 150-pixel window is taken from the DR map and matched to the theoretical map. As this window is 'slid' along the experimental DR map, the match

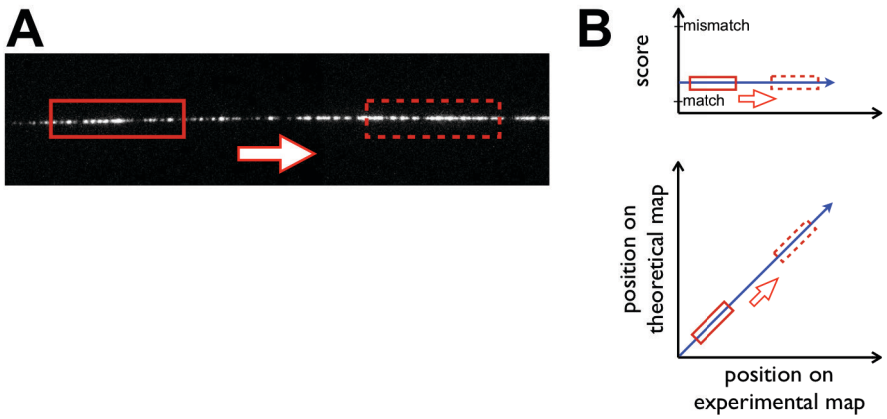


Figure 3. Sliding Window Analysis. (A) A barcode window (red rectangle) is moved along the DR map obtained experimentally for a given molecule. It is aligned to a theoretical DR map computed from the reference genome at each point. **(B)** When no structural variation is present, the score of the alignment remains good (top), and plotting the position on the experimental vs. theoretical maps produces a continuous line (bottom).

to the theoretical map should advance at the same rate with a continuously good match score. In an ideal situation where no structural variation is present, plotting the experimental map position vs. match score should produce a continuous line with a slope of 0; plotting the experimental position vs. the theoretical position should produce a continuous line with a slope of +1, as shown in Figure 3.

If structural variation is present in the sample that is not present in the reference genome, the alignment of the 150-pixel window no longer matches the theoretical reference map. This mismatch causes characteristic patterns: an increase in plots of score vs. position and discontinuities in the experimental vs. theoretical position plots as the window slides across the SV. Insertions and deletions can be detected directly. Inversions can be distinguished from insertions by comparing the experimental DR

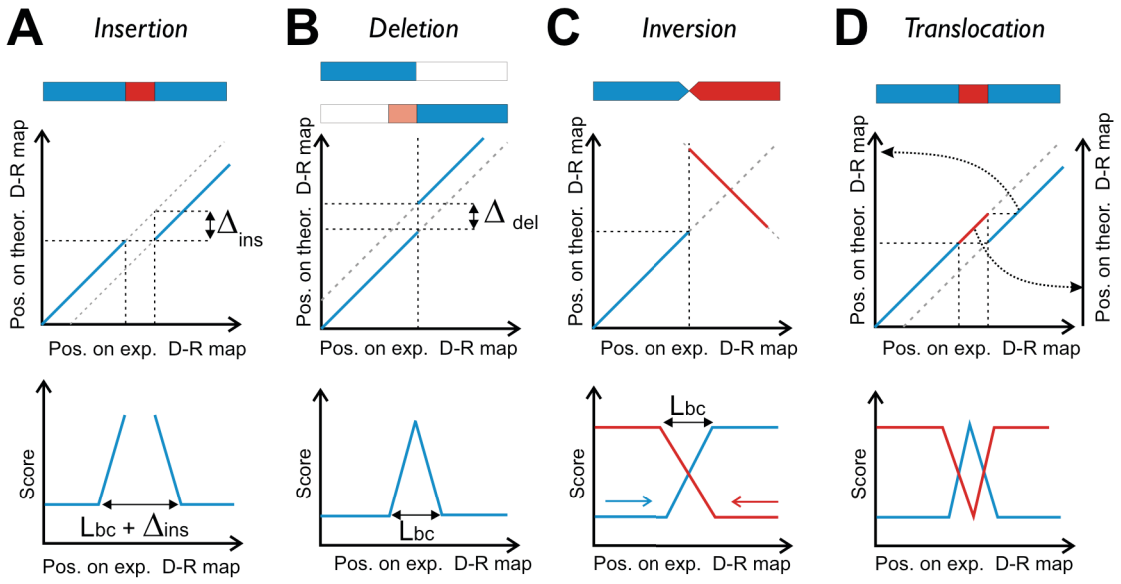


Figure 4. Detecting Structural Variation by Sliding Window Analysis of DR Maps. Structural variation produces characteristic patterns in position plots (top) and score plots (bottom) for **(A)** inversions, **(B)** deletions, **(C)** inversions, and **(D)** translocations. (Adapted from Marie *et al.*, 2013, R. Marie, DTU).

map at the point of mismatch to an inverted theoretical map, and translocations can similarly be detected by comparing the DR map at the point of mismatch to all other positions in the genome to find the best-matched alignment for that particular region.

This Sliding Window analysis method was applied to DR maps from our Yoruban African sample to detect structural variation. Across the 21 molecules, two revealed SV: an inversion on chromosome 16 around 22.4 Mbp and insertion on chr19 around 62.2 Mbp that appeared to be a translocation from chromosome 9 (Figure 5).

This inversion on chromosome 16 was previously detected in the NA18507 genome[40]

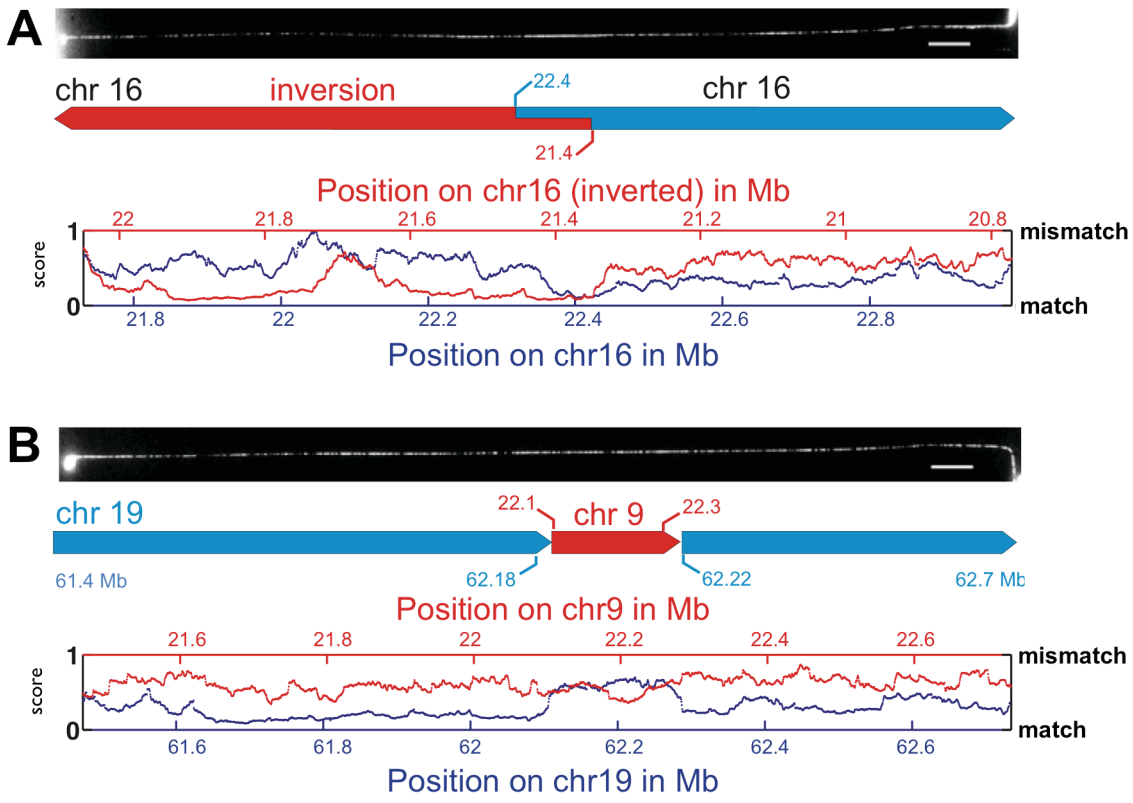


Figure 5. Detection of Structural Variation in NA18507. Sliding Window analysis detected **(A)** an inversion on chromosome 16 and **(B)** a 200 kbp translocation from chromosome 9 to chromosome 19. (From Marie *et al.*, 2013, R. Marie, DTU)

and has been found in other genomes as well[41-43]. The insertion on chromosome 19 occurs in a sub-telomeric region of the chromosome, which is prone to SV. While we are certain that a fragment of DNA inserted into chromosome 19 maps to chromosome 9 in the reference genome, we cannot be certain whether it is a duplication of this region of chromosome 9 or a true translocation without having a full map of the corresponding region to gain context. There is some evidence in the literature, however, for a duplication event: the region of chromosome 9 that we find duplicated/translocated in chromosome 19 has been reported in one study to have a gain in copy number in another Sub-Saharan population[44] and the ~0.20 Mbp insertion from chromosome 9 we observe closely fully encompasses a 0.12 Mbp copy-number gain reported in Asian populations in Singapore[45]. Further details and graphical maps are provided in the Appendix to this thesis.

Detecting Complex Structural Variation and Long-Range Repetitive Elements

We extended our DR mapping to the Jurkat genome, which has extensive structural variation at the microscopic level (> 2 Mbp) visible by mFISH karyotyping[1]. After constructing DR maps as described previously, we found striking, visual patterns that suggested the genome contained tandem duplications of blocks of genetic material scattered throughout the genome, as illustrated in Figure 6.

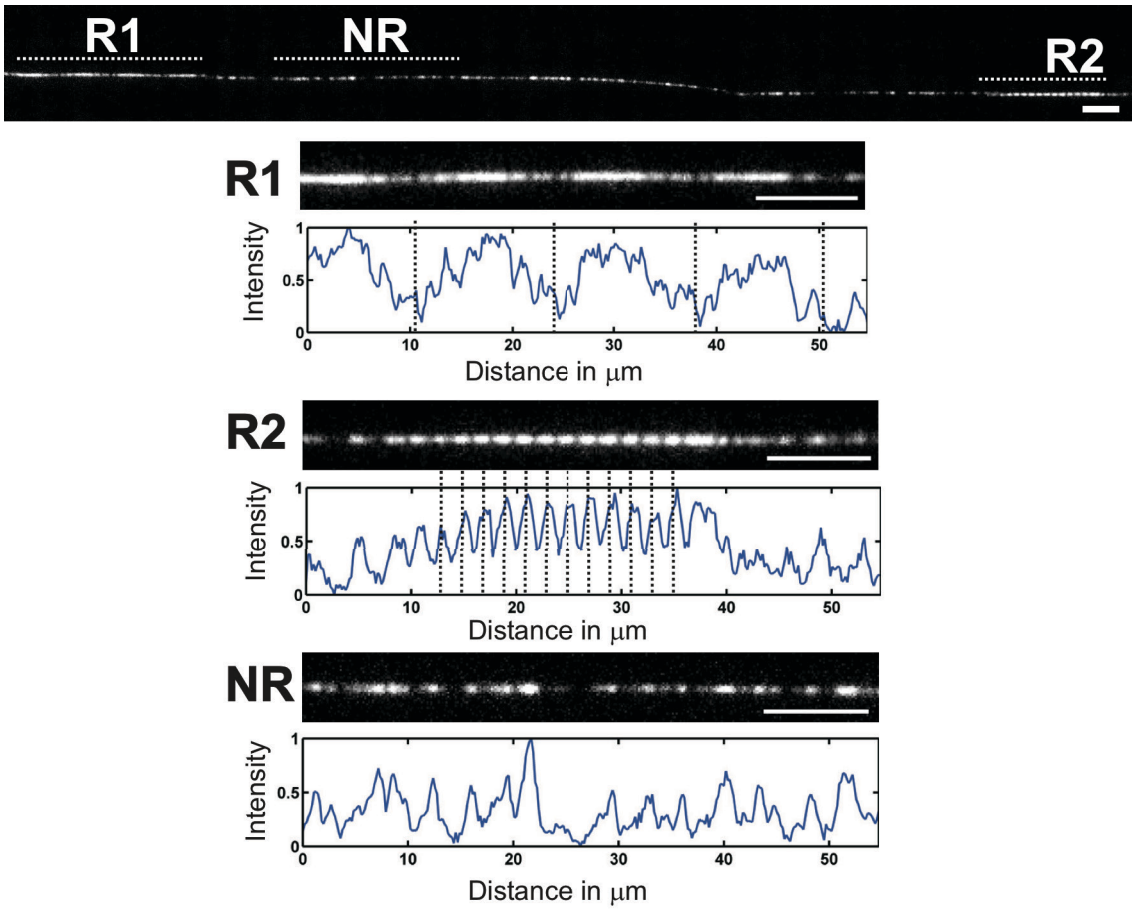


Figure 6. Tandem Duplications in the Jurkat Genome. A DR map of a single molecule from the Jurkat genome shows two repeat features **(R1, R2)**. A nonrepeat **(NR)** element is shown for comparison. The periodicity of the repeat features is apparent from their DR map (dotted lines), with a repeat unit lengths of approximately 40.4 ± 0.5 kbp for **R1** and 6.3 ± 0.1 kbp for **R2**. Scale bar: 10 μm. (R. Marie, DTU)

Notably, unique barcodes from the region between R1 and R2 failed to find significant match to any chromosomes in the hg18 reference assembly, which suggested that the structural varion observed at microscopic (cytogenetic) and sub-microscopic (DR map) levels extends further to the sequence level as well.

To analyse the molecule at the sequence level, we used our laboratory-on-a-chip device to recover a DR-mapped Jurkat DNA molecule, amplified it using multiple-

displacement (MD) amplification, and then used an aliquot of this material to carry out next-generation DNA sequencing with the aim of identifying the origin of the molecule within the genome as well as any sequence-level variation, such as SNPs. More broadly, I aimed to characterise the data obtained from an MD-amplified DNA fragment after DR mapping, compare it to other available examples in the literature, and develop a general set of bioinformatics protocols to handle similar data in the future.

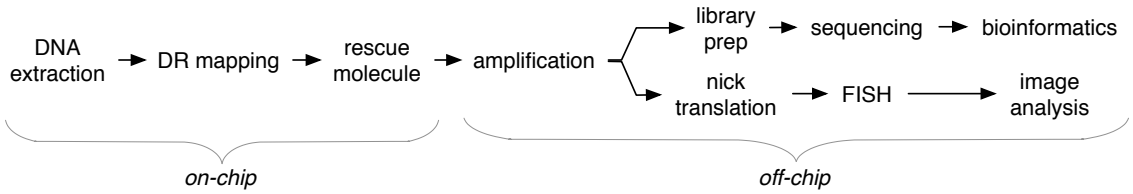


Figure 7. Workflow to integrate DR mapping with conventional genomics analysis.

Sequencing of single DNA molecules after DR Mapping

From the aliquot of amplified DNA delivered to Genomics Services (WTCHG, Oxford), 9.9 million 51 bp paired-end reads were generated by the Illumina Genome Analyser IIx platform. These reads were mapped to the hg19 human reference genome using *stampy*[46]. 1.5 million reads mapped successfully to the human genome – accounting for only approximately 15% of the total reads.

DNA Sequencing Reads Are Contaminated by Environmental Samples. To determine the origin of the remaining 85% of the sequencing reads, a random sample of 10,000 non-mapping reads was BLAST searched to the entire NCBI nucleotide collection. About 25% of the reads were from beta proteobacteria, widely found in the environment (eg. soil, saliva), indicating possible amplification of contaminants, potentially contained within the Qiagen amplification kit, as suggested by Blainey and

Quake[32]. A very small number of reads (~1%) mapped to the wheat genome – a concurrent project at the sequencing centre (D. Buck, WTCHG, personal communication) – suggesting potential cross-contamination during the library prep and pooling steps.

It is important to note that the majority of reads (~53%) had no match *at all* to the NCBI nucleotide collection, suggesting they were the result of nonspecific "auto-amplification" of primer-dimers during whole-genome amplification by the Φ29 DNA polymerase contained in the Qiagen Repli-G kit. Regardless, reads mapping to the human genome (77 GBp of sequence) were easily identifiable and used for subsequent analysis.

Identifying the Chromosomal Origin of an Amplified Single DNA Molecule

While the DR map of an isolated DNA fragment can be used to determine the origin of

hg19 Aligned	15.26%
Unknown	53.49%
Delftia	23.89%
Stenotrophomonas	2.02%
Propionibacterium	1.30%
Tritucum	1.20%
Varivorax	0.33%
Pinus	0.25%
Ralsontia	0.25%
Acidovorax	0.18%
Uncultured	0.16%
Prseudoxanthomonas	0.15%
Ricinus	0.11%
Xanthomonas	0.09%
Other	1.32%
Total (excl Other)	98.68%

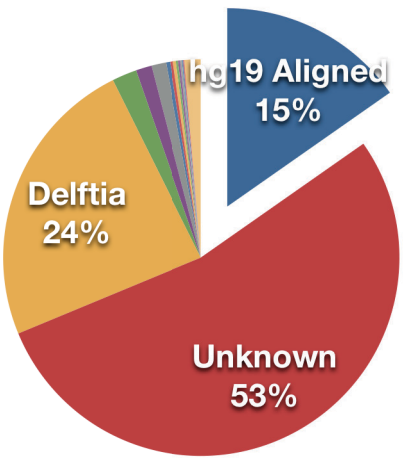


Figure 8. Read Contamination Analysis. Origin of reads from single molecule DNA amplification and sequencing were obtained by finding the best match of a subset of reads (n=10,000) to the entire NCBI BLAST nucleotide collection.

the fragment in the genome, I aimed to used the sequencing data independently to determine the origin of the molecule with the aim of (a) validating the DR mapping approach and (b) providing a method of analysing fragments that could not be DR mapped due to extensive mosaicism, as was the case for samples from the Jurkat cell line. I adopted three approaches to determining the origin of an amplified single molecule based on the mappings of the DNA sequencing reads to the reference genome:

1. Simple counting of number of reads mapped to each chromosome;
2. Accounting for biases introduced by Illumina sequencing and mapping by calculating the expected number of reads that should map to each chromosome and comparing that to the actual number reads from the sample; and
3. Examining the clustering of read mappings to account for (and exploit) biases introduced by MD amplification of long, single DNA molecules.

1. Origin Identification - Analysis of Raw Read Counts. As an initial survey, the raw number of reads that mapped to each chromosome in the reference human genome (hg19) was counted and weighted according to the length of each chromosome. The results, shown in Figure 9A, do not provide a clear indication of the origin of the molecule: reads map to all chromosomes, with a broad distribution in the number of reads. To determine whether the reads were spread across each chromosome or localised within them, I plotted the number of reads mapping to a rolling 2 Mbp bin along each chromosome. Reads were found to map to loci within each chromosome (Figure 9B), with varying read depth.

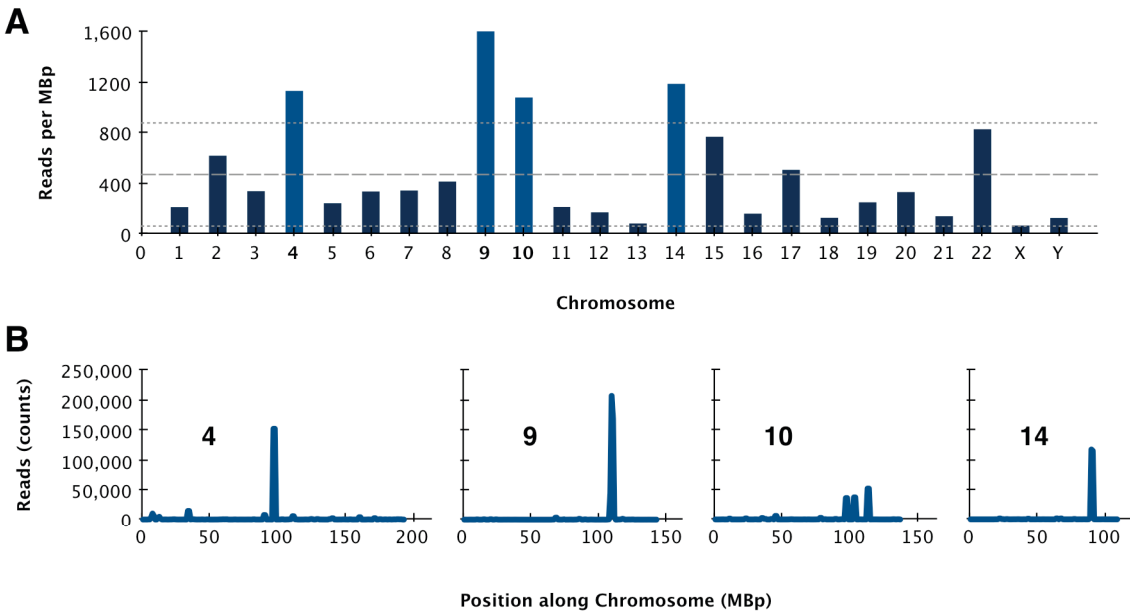


Figure 9. Analysis of raw read counts. After mapping reads to the human reference genome (hg19), the number of reads from the amplified DNA molecule that mapped to each chromosome was counted and weighted according to chromosome length, **(A)**. Dashed lines indicate the mean number of reads across all chromosomes and the standard deviation. **(B)** Using a rolling 2 Mbp window, the location of the reads along the chromosomes showing significant hits (**4**, **9**, **10**, and **14**) reveals that reads are not spread along the entire chromosome, but map to particular loci.

While the counting of raw DNA reads does not account for any systematic error introduced by the process of Illumina sequencing and basecalling, nor in the mapping of the resultant reads to the human genome, the presence of multiple peaks (*vs.* a single peak expected from a single, contiguous DNA fragment) gives an indication that the sample contains contaminating fragments of human DNA – in addition to the previously-discovered environmental contamination.

2. Origin Identification - Accounting for Sequencing and Mapping Bias. In order to determine the fraction of reads expected per chromosome and account for varying mappability, I simulated 10 million mate pairs (20 million reads) from the hg19 human

genome using *MetaSim*[47], taking empirically determined error profiles of the Illumina sequencer into account. These reads were then mapped back to the hg19 reference genome using *bowtie*[48], and each chromosome's percentage of the total mapped reads was recorded as the "expected" percentage.

The expected percentage was then compared with the actual percentage of reads each chromosome received in the data set from a single amplified DNA molecule; a plot of the actual vs. expected percentage of mapped reads is shown in Figure 10.

The plot reveals a significant over-representation of reads mapping to chromosome 9, which suggested it was a likely candidate chromosome for the origin of the single DNA

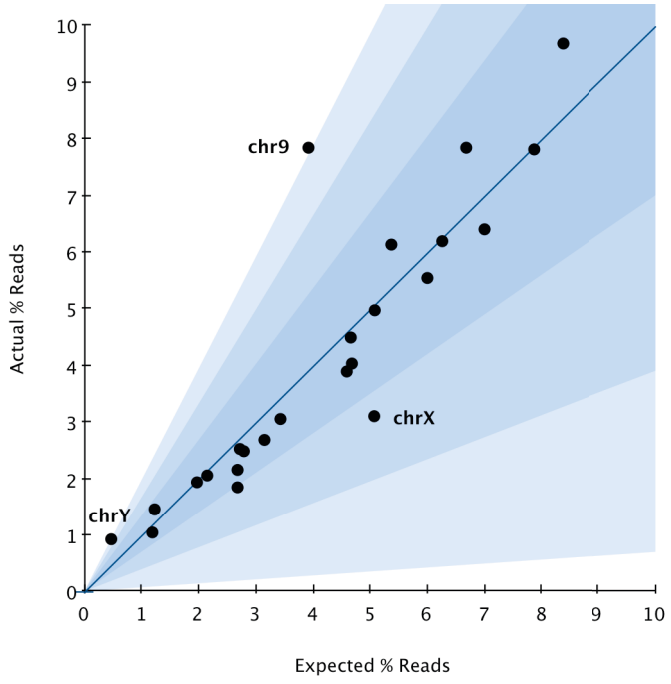


Figure 10. Identification of chromosome-of-origin. The number of reads mapped to each chromosome is compared to the expected values for a whole-genomic sample with an 2nXXYY genotype. The blue line is the mean slope (actual/expected) and the blue gradient represents standard deviation (± 1σ, 2σ, 3σ)

fragment. Interestingly, only approximately 3% of the reads mapped to chromosome X – compared to an expected ~5% if the genotype were XX, therefore indicating that contaminating DNA fragments might be derived from an organism with a 2nX genotype – or some multiple thereof. Subsequent multicolour FISH karyotyping of the Jurkat cell line (D. Moralli, Cytogenetics Core, WTCHG) indicated that this prediction is correct: Jurkat cells are generally tetraploid with a 4nXXYYY genotype[1].

After identifying chromosome 9 as a likely candidate, I then attempted to determine the specific region that the strand originated from. Since I expected amplification to introduce bias in the quantitative results (e.g. number of reads mapping to a specific position), I instead looked at the "footprint" of the reads. This analysis was conducted to determine the locations along each chromosome where the sequencing coverage of the genome was high *and* spread out over a longer region in order to exclude short contaminating fragments that might have been deeply sequenced. A single peak was observed, on chromosome 9, centred around 110 Mbp (Figure 11).

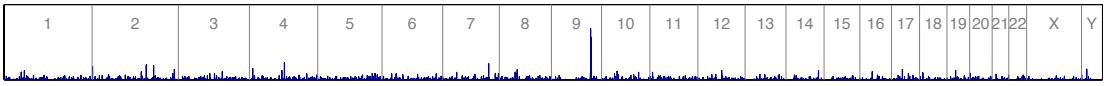


Figure 11. Footprint analysis of well-covered genomic regions with both distributed *and* high coverage.

3. Origin Identification - Reads cluster as a result of bias introduced by MD amplification. Once chromosome 9 was identified as a likely candidate for the origin of the DNA molecule, I looked in greater detail at the single peak in read mappings that occurred on chromosome 9 around 110 Mbp (Figure 9B and Figure 11). Closer examination of the distribution of reads within this peak was made by dividing the reference genome into 1 Kbp bins and counting the number of reads that mapped within

each bin. When using large (1-2 Mbp) bins previously, the data along chromosome 9 appeared as a single peak. However, "zooming in" with 1 Kbp bins to the regions of interest exposed discrete clusters of well-covered genomic regions, which I termed "islands". These islands of reads were on average approximately 20 Kbp wide, as shown in Figure 12.

Notably, these islands did not exist in isolation. Rather, they clustered together into an "archipelago" that formed the 110 Mbp peak – the archipelago spanned 2-3 Mbp, consistent with the physical observation of the ultra-long DNA molecule during DR mapping in the nanofluidic chip, which was at least 1.3 Mbp long.

To visualise this clustering and compare it to other positions in the genome, I developed "read density maps" on which each 1 kbp bin was represented by a square and the number of reads it contained by the degree of shading. Only on chromosome 9 did I observe a clustering of these islands, which I termed an "archipelago"– other minor peaks on the footprint map were consistently composed of single, isolated

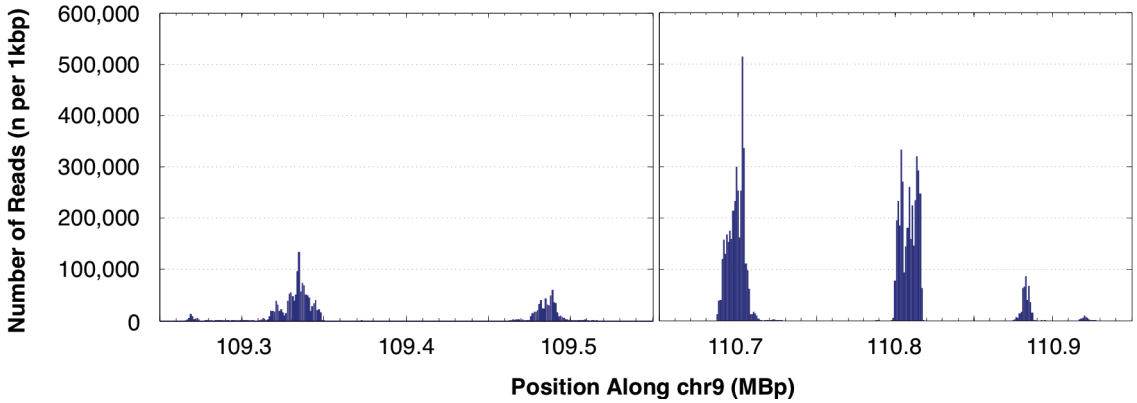


Figure 12. Clustering of Read Mappings. Representative clusters of DNA sequencing reads ("Islands") after mapping reads from an MD-amplified single DNA fragment to the hg19 reference genome using *stampy*.

islands (Figure 13B and C). This "archipelago" pattern is consistent with the read density patterns of other long single-molecule templates amplified by multiple-displacement-amplification that have been reported in the literature[34].

I constructed a read density map of these previously reported results to provide a direct visual comparison (Figure 13D) by obtaining the raw sequencing reads from the NCBI SRA (accession no. SRR077303) and remapping them to the hg19 reference genome using *bowtie*[48], with an insert size window of 0-1000 bp.

Perhaps not coincidentally, the width of the islands in both cases is on the same order as the processivity of the Φ29 DNA polymerase used in the MDA protocol[49] and the apparent size of the whole-genome amplified products when visualised by gel electrophoresis. Lastly, in order to confirm that this hit was not due to varying "mappability" of the genome due to gaps in the reference assembly or repeat regions, read density maps were constructed for the simulated total genomic DNA previously generated using *MetaSim*[47]; no archipelago pattern is present (Figure 13E).

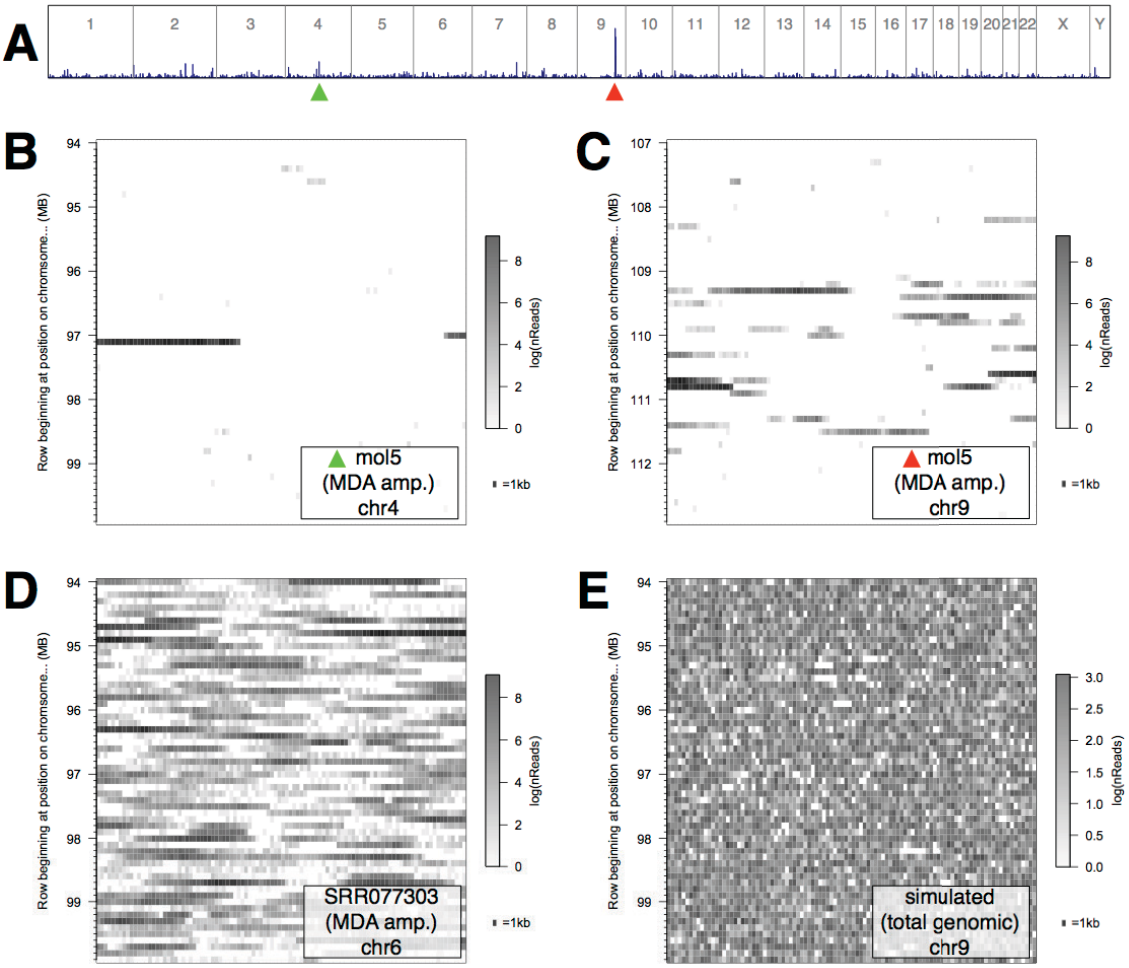


Figure 13. Sequencing of a Rescued Single DNA Molecule. After mapping sequencing reads to the hg19 reference genome, I constructed a "footprint" map, **(A)**, to visualise regions of the genome which had a high degree of coverage by sequencing reads. A single peak was observed on chromosome 9 centred at 110 Mbp (red ▲); other minor peaks (e.g. green ▲) were also observed. I then constructed read density maps, **(B-E)**, to "zoom-in" on regions of interest: each 1 kbp of the reference genome was represented as a single block increasing in ascending order from left to right, with the number of reads mapping to each block represented by its shading. Each row contains 100 blocks, representing 100 kbp of the reference genome beginning at position indicated on the vertical axis. Reads putatively from the rescued DNA molecule, **(C)**, clustered together in islands, which uniquely formed an archipelago on chromosome 9 around 110 Mbp. This archipelago pattern was not present anywhere else in the genome, for example, **(B)**, on chromosome 4 at 97 Mbp. Moreover, this archipelago effect is characteristic of single molecule templates amplified by multiple displacement amplification, such as reads from single amplified

chromosomes[32], **(D)**, and is not due to varying mappability of the human genome, as we do not observe this pattern by mapping simulated reads, **(E)**, from total genomic DNA.

Automating Archipelago Detection within Sequencing Data

I exploited the clustering of read islands to develop an automated algorithm for determining the origin of sequenced single DNA molecules, and to estimate the length of the molecule that was captured in the nanofluidic chip. The algorithm groups nearby islands into clusters, termed 'bigIslands', and then "hops" to nearby islands to determine if they are part of a connected archipelago. By applying the method to the data from our single, amplified DNA molecule, the significant hit to chromosome 9 can be easily identified: it is the only hit that is long and composed of a large number of islands, as shown below.

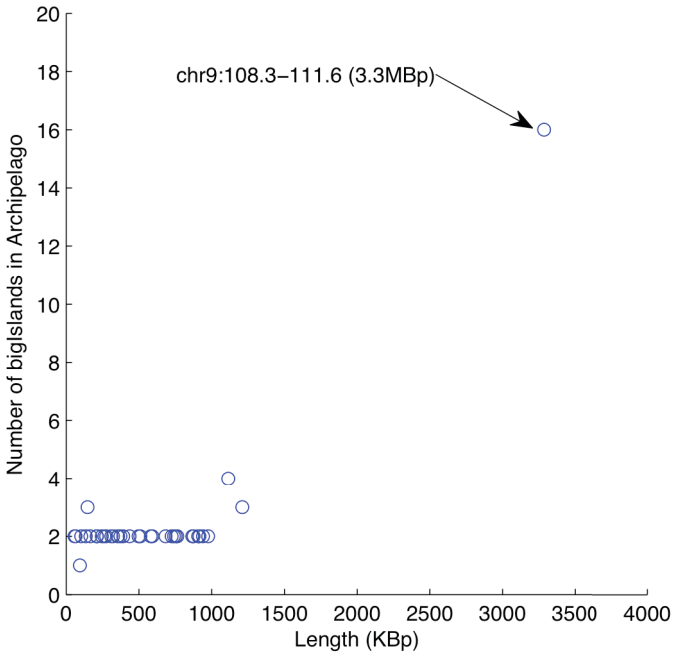


Figure 14. Automated archipelago detection for locating the chromosome of origin of single DNA molecules.

Chr	Start (Kbp)	End (Kbp)	Length (Kbp)	Number of Biglslands
9	108285	111569	3284	16
13	92984	94194	1210	3
1	55859	56974	1115	4
3	16841	17817	976	2
1	111601	112542	941	2
3	100794	101710	916	2
11	55912	56818	906	2
11	23598	24473	875	2
3	87517	88385	868	2

Table 1. Output from automated archipelago detection algorithm.

The output of the algorithm also provides a method to estimate the length of the long DNA strand using the edges of the archipelago, plus half the mean distance between the islands. In the case of the archipelago on chromosome 9, its extent is 3.5 ± 0.1 Mbp –

consistent with the visual inspection of the DNA molecule spanning the nanofluidic portion of the chip and beyond—from 108.3 to 111.6 Mbp.

Heterozygosity Analysis and Haplotype Confirmation

To further confirm that the sequencing data we examined arose from the isolation of a single DNA molecule, I sought to analyse the rate of homozygosity within the data putatively from the rescued DNA molecule; sequencing data arising from a single DNA fragment isolated from a single chromosome should constitute a single haplotype and therefore not contain any heterozygous base calls. I analysed heterozygosity in three ways:

1. Identifying SNPs within sequencing data, their rate of heterozygosity, and comparing this rate to other population samples in the literature;
2. Examining the rate of concordance between multiple sequencing reads at all positions in the genome that were well-covered by DNA sequencing data, and comparing this rate to the rates of concordance reported in the literature for MDA-amplified single chromosomes; and
3. Comparing genotyping data from total Jurkat cell DNA to the same positions covered by DNA sequencing from the MDA-amplified single DNA fragment.

All methods support the successful isolation and sequencing of a single DNA fragment.

1. Heterozygosity - Identifying SNPs in Sequencing Data from MD-amplfied DNA fragment. I performed standard variant calling within the archipelago of reads along chromosome 9: a total of 49 SNPs were identified along the region (chr9:108.3-111.6

Mbp) using *samtools*[76] to compare to the hg19 reference genome. A total of 49 SNPs were detected. From these, 37/49 SNPs (76%) were covered at a depth greater than 10 reads; of these, 23/49 SNPs had a mapping quality score greater than 50. Of the set of

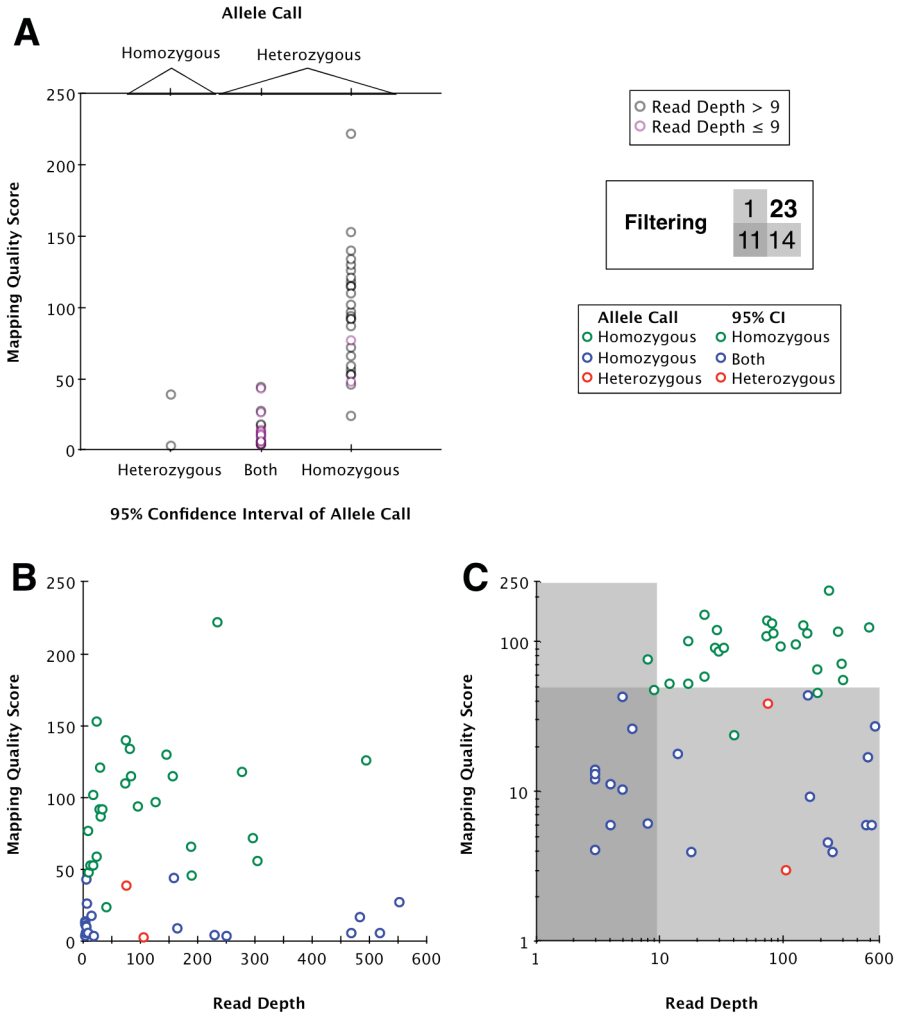


Figure 15. SNP calls from an amplified single DNA molecule. (A) comparing the allele call vs. the mapping quality score (actual and the 95% confidence interval) shows that high-quality calls are almost exclusively heterozygous. **(B)** The read depth and mapping quality for each group of allele calls (legend top-right) can also be compared. **(C)** standard filtering of the data for read depth > 9 and mapping quality score > 50 ($p_{miscall} < 10^{-5}$) gives 23 high-quality SNPs in comparison to the hg19 reference genome.

26/49 SNPs rejected by filtering, 23/26 were called as homozygous by *samtools* but within a 95% confidence interval could have been either homozygous or heterozygous. Only 2/26 rejected SNPs were called as heterozygous, whereas all of the well-covered, high quality SNPs (23/49) were consistently called homozygous, as shown in Figure 15.

To make a conservative comparison to the rate of SNP heterozygosity in other samples and populations, I considered the two low-quality SNPs that were called heterozygous. If they were assumed to be heterozygous (hetSNPs), then the overall hetSNP rate over the extracted DNA molecule 'archipelago' would be 0.04. This rate is substantially lower than the average human hetSNP rate, which has been reported[77] as 0.263 ± 0.171, though it varies considerably for genome position. To attempt a more detailed comparison, the collection of SNPs reported by the 1000 Genomes Project and HapMap within the bounds of the single molecule archipelago were downloaded and the average hetSNP rate was calculated over the archipelago region of the isolated single DNA fragment, as shown in Table 2.

The results were inconclusive: there is a large amount of variability in the rate of

	SNPs	SNPs with Het determined	Mean rate(HetSNP)	SD	Median
HapMap	10428	8478	0.009	0.056	0
HapMap CEU pop MAF > 0.10	3011	2461	0.006	0.049	0
HapMap ASW pop MAF > 0.05	80	30	0.342	0.157	0.4
1000 Genomes	4053	3303	0.131	0.146	0.08

Table 2. Data obtained from dbSNP build 134 for the rate of SNP heterozygocity within the DNA fragment archipelago

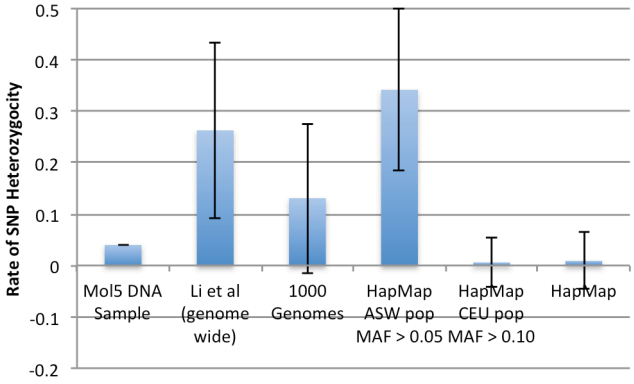


Figure 16. Comparison of the rate of SNP heterozygocity (HetSNP rate). The rate of the isolated, MDA-DNA fragment (Mol5 DNA Sample) is compared to published data for genome-wide (Li *et al.*, 2007) and SNPs within the MolFive archipelago from 1000 Genomes and specific HapMap populations (ASW and CEU) and overall.

hetSNPs between data sets, as shown in Figure 16.

While the lack of heterozygous SNPs among those selected after quality filtering, supports the isolation of a haplotype, I decided to pursue additional methods to examine the heterozygosity of the sample given the small number of SNPs (n=49) and the variability of population hetSNP rates.

2. Heterozygosity - Rate of concordance between multiple DNA sequencing reads.

To extend the analysis beyond a small set of SNPs and capture an overview of the level of heterozygous positions within the entire 'archipelago', I determined the degree of heterozygosity for *all* positions in the sequencing data by calculating the degree of discordance amongst the reads (i.e. > 25% discordance between reads covering the same position) at varying read depth.

While it is reassuring to note that at highly-covered bases (≥ 80), no positions are potentially heterozygous, I sought to compare the degree of heterozygosity to other data sets, in particular to one subjected to similar fluidic handling and multiple-displacement amplification.

Overall, 99.97% of all well-covered positions were homozygous in our sample – comparable to the rate of homozygosity present in sequencing data previously reported for physically isolated single DNA molecules[34] (99.99%, NCBI Accession Number SRR077303). This comparison suggests that our rate of heterozygosity in the data is

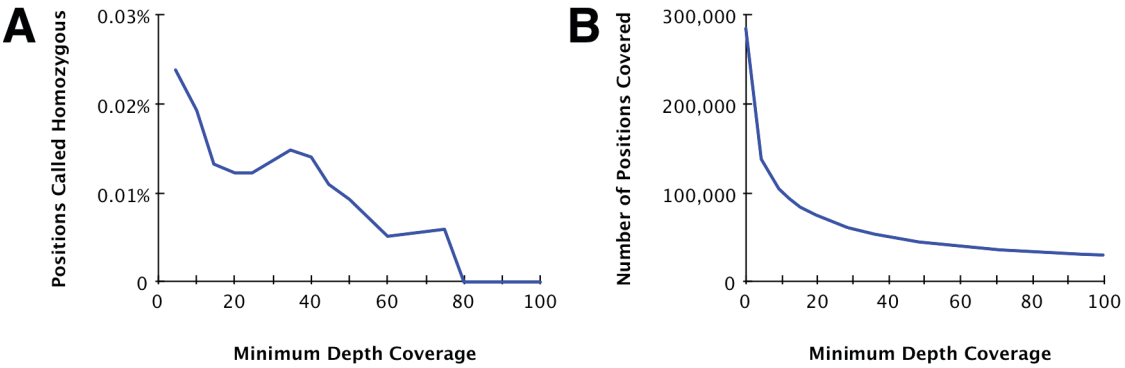


Figure 17. Effect of Depth Cutoff on coverage and heterozygosity. (A) Fraction of reads in the DNA fragment archipelago called as heterozygous at different depth cutoff levels. The number of covered bases **(B)** is shown as a comparison.

	MDA fragment, chr9:108285000-111569000			Fan et al. 2009:SRR077303, chr6		
Region Length	3,284,000			171,115,067		
Num Positions Covered	285,158	8.68%	of archipelago	56,150,473	32.81%	of chromosome
Num Positions Well-Covered (depth)	116,925	3.56%	of archipelago d>9	6,839,782	4.00%	of chromosome d>19
Num Positions Homozygous	116,893	99.97%	of well-covered positions	6,839,343	99.99%	of well-covered positions
Num Positions Heterozygous	32	0.03%	of well-covered positions	439	0.01%	of well-covered positions

Table 3. Read discordance analysis to determine the degree of heterozygosity in sample. Comparison between MD-amplified single DNA fragment over archipelago region and data set from Fan *et al.*, 2011 for an MD-amplified single chromosome.

consistent with the isolation of a single haplotype.

3. Heterozygosity - Comparison between isolated DNA fragment data and total Jurkat DNA. Total DNA was isolated from Jurkat cells in mammalian tissue culture and genotyping was carried out using the Illumina Infinium assay using a Human Omni1-Quad (v1.0) chip and genotyping SNP calls were made using Illumina software. I then cross-matched these genotyping calls from total Jurkat DNA with well-covered

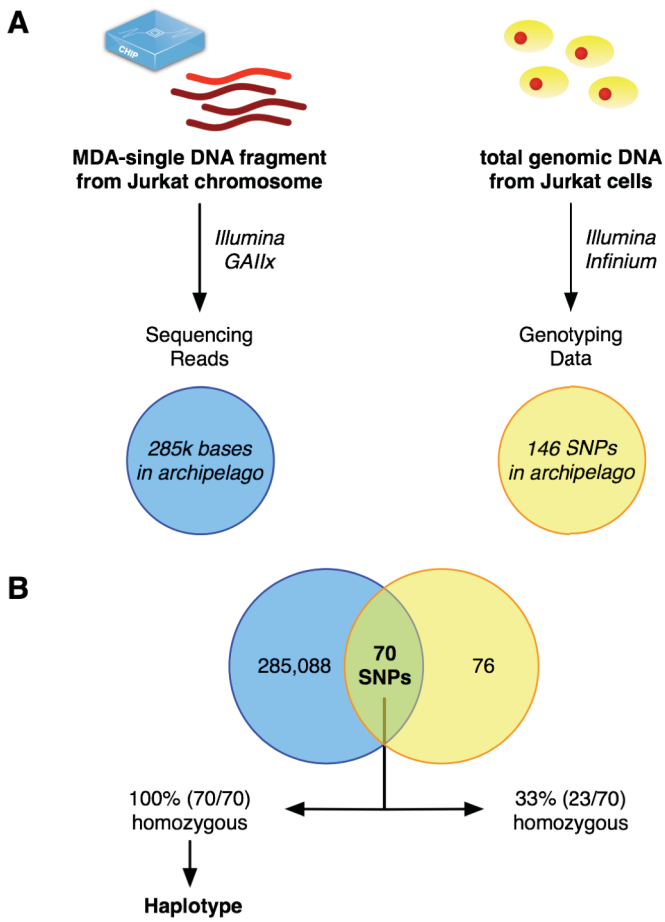


Figure 18. Comparing the Rate of Heterozygosity: Same positions, same cell line. An outline of the procedure used to produce comparable data from MDA-amplified DNA fragment and total Jurkat genomic DNA, **(A)**. Comparison of the overlapping SNPs in each data set, **(B)**, suggests that a single haplotype has been isolated: 100% of positions are called as homozygous vs. 33% from Jurkat DNA.

(read depth > 9) positions in the sequencing data of the MDA-amplified DNA fragment (Figure 18). While all 70/70 relevant SNPs were called as homozygous from the sequencing data of the rescued DNA molecule, genotyping of bulk Jurkat DNA yielded just 23/70 homozygous positions, strongly supporting the successful isolation of a single haplotype.

Discussion

Using DR maps and sliding window analysis, we have demonstrated how SV can be detected in single molecules of DNA in order to provide information about genome structure at the kbp-Mbp scale. The use of a novel nanofluidic device also allows us to combine this information with bp-kbp information from DNA sequencing.

For our Jurkat-derived DNA fragment that displayed extreme SV and could not be mapped to the genome using DR mapping alone, I was able to confirm its origin to be chromosome 9q31 – specifically chromosome 9:108.3-111.6 Mbp – and to perform traditional SNP discovery to gain information about bp-level variation as well. In addition, I have explored the data for what it reveals about the amplification mechanics of a single molecule template and the bias introduced. This information was then used as the basis for an automated detection algorithm for determining the origins of isolated Mbp-length DNA fragments, which will be applicable to future work.

Comparison of Sequencing Data from MD-amplified DNA Fragment to Literature

To put our sequencing results into the context of other related datasets, I carried out a meta-analysis of sequencing data from two publications which involved whole-genome

amplification from single chromosomes (Fan *et al.*)[34] and from single cells (Navin *et al.*)[50].

Environmental Contamination and MDA. I first examined the degrees of

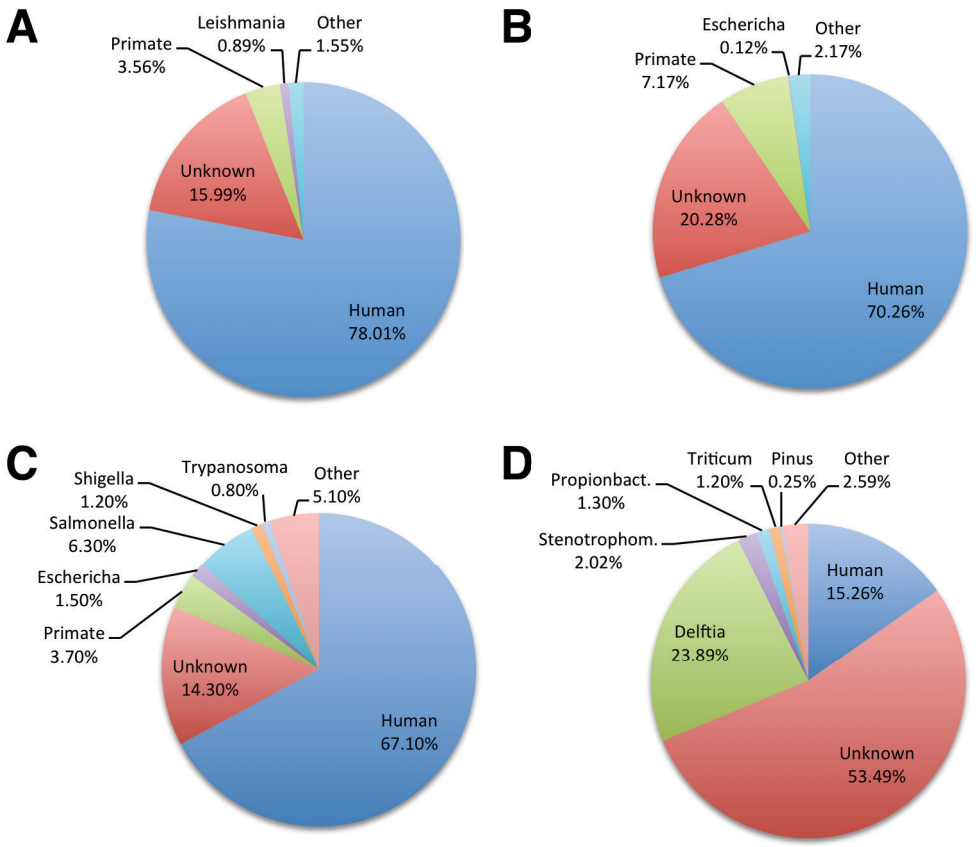


Figure 19. Comparison of contamination among NGS sequencing results. I compared libraries derived from **(A)** whole-genome amplified material (Fan *et al.*), **(B)** Navin *et al.*, as well as from **(C)** clonal DNA (Kitzman *et al.*), and **(D)** our MDA-amplified sample after isolation via nanofluidics. A random sample of reads was BLAST searched to the NCBI nucleotide collection and the first word of the top result – typically the genus name – was counted. Therefore reads reporting as primate or mammalian are should be considered likely "human" reads as well. Our results have a larger proportion of unknown reads, likely the result of nonspecific amplification, and contamination by environmental organisms (e.g. Delftia). Sample size for A: n=10,038; B: n=4,000; C: n=1,000; and D: n=10,000.

contamination in each data set, and included data from clonally-derived non-whole-genome-amplified DNA (Kitzman *et al.*)[51] to observe any impact the amplification process might have. As described above, a random sample of reads was selected and used to search the entire NCBI nucleotide collection. The results are presented in Figure 19.

Aside from environmental contamination likely due to amplification kit contamination or sample handling, our amplified sample contains a large proportion of reads (53%) that are unknown, and likely a result of non-specific amplification of the random primers. Since each primer has many partial complements in the primer pool, they are able to form duplexes to prime the phi29 DNA polymerase used in whole-genome amplification. It has been reported that reducing the volume of the phi29 DNA polymerase amplification reaction can improve amplification efficiency[52]. Since we have only a single copy of our template, any increase in volume results in a proportional decrease in concentration. At fixed primer concentrations, any decrease in the template concentration will give the primer dimerization reaction a more favourable equilibrium over the template-priming hybridization.

Read Mapping Performance from MDA samples. A more detailed comparison was then made of the alignment performance of the reads from Fan *et al.* and Navin *et al.* without taking paired-end constraints into account.

In each case, a substantial proportion of reads aligned to multiple positions in the genome, while large numbers of identical reads were identified and removed from the aligned data set – likely due to homologous and repetitive portions of the genome. The high number of failed alignments in the data set from the isolated DNA fragment are the result of contamination and primer-dimers as discussed above.

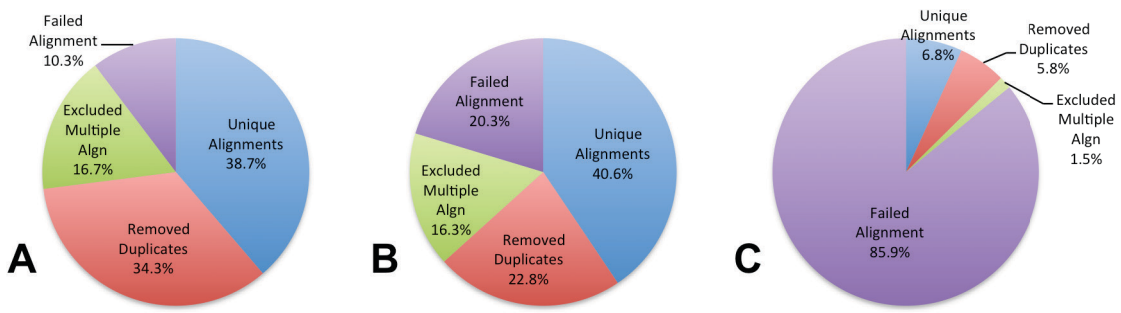


Figure 20. Performance of *non*-paired-end read alignments. Reads were aligned with *bowtie* to hg19 for whole-genome amplified samples of **(A)** single chromosome from Fan *et al.*, **(B)** single cell from Navin *et al.*, and **(C)** MolFive. Reads which mapped to more than one location on the genome ("excluded multiple algn") and which were identical to other reads likely due to PCR steps in library ("removed duplicates") were counted separately from unique alignments. Sample size for A: $n=29.1\times10^6$; B: $n=9.5\times10^6$ and; C: $n=9.9\times10^6$

To compare our original paired-end data, the reads from Fan *et al.* were then re-mapped to the hg19 genome assembly taking paired-end information into account: the authors report a gel-selected library size of 300-500 bp, which would correspond to an insert size of at least 180-380 bp, assuming a total adaptor length of at most 120 bp – no mention is made of any deviation from the standard Illumina library prep. I attempted to first map using *bowtie* with an insert size window of 250-550 bp and of 100-700 bp. Both mapping attempts resulted in low percentages of successful alignments (2% and 25%) compared to the non-paired mapping results obtained.

Widening the insert size window to 0-1000 bp captured a much larger number of reads (72%), nearly all of which were captured by an insert size window of 0-200 bp. Fan *et al.* report that the median insert size as determined from paired-end mappings was 75 bp, though they do not attempt to explain the discrepancy in their paper.

An illustration of the different alignment results is presented below (Figure 21).

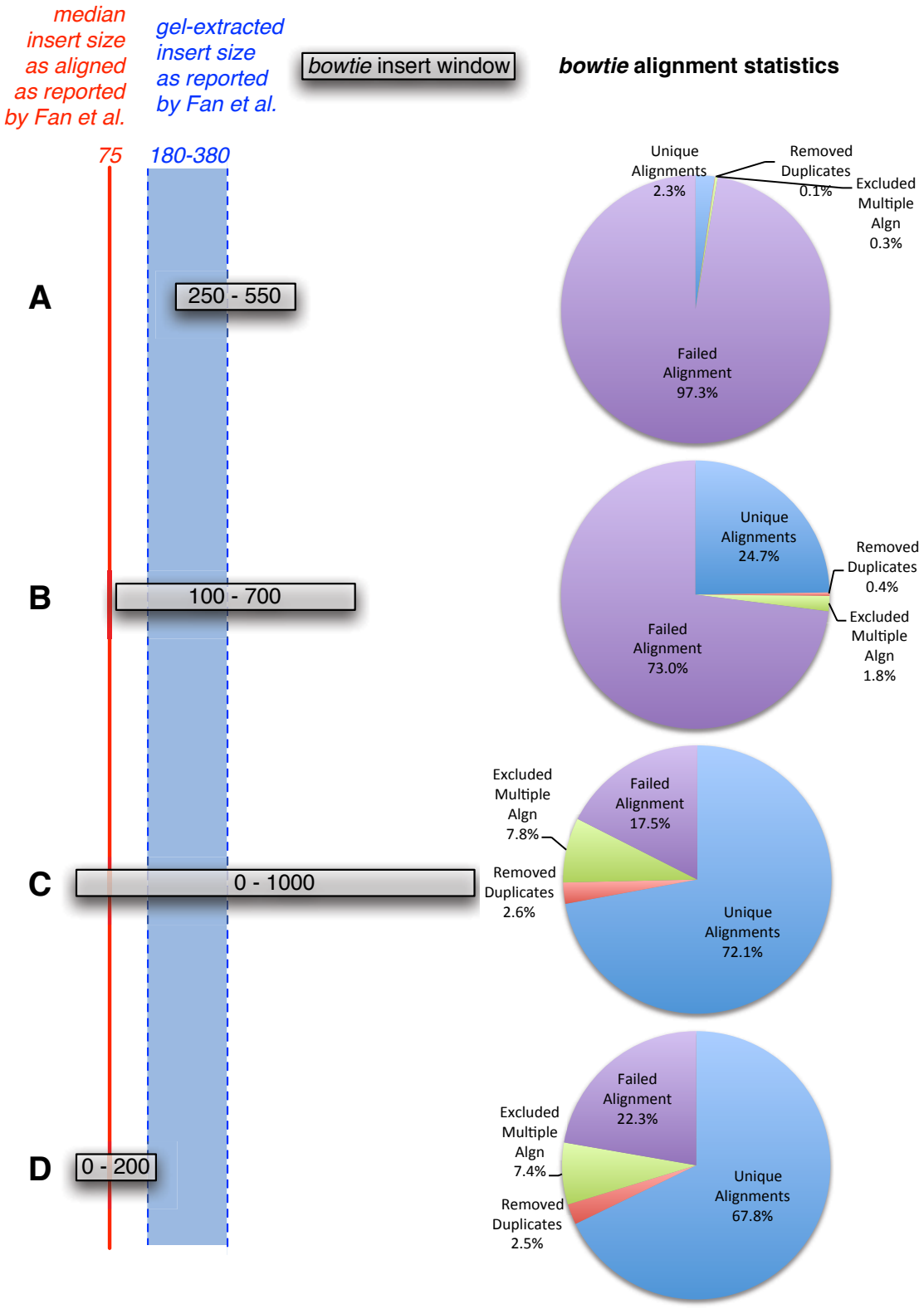


Figure 21. Comparison of paired-end alignment results using *bowtie* with varying insert size windows on sequencing reads from Fan *et al.* (NCBI SRA Accession Number SRR077303). The gel-selected library size was reported as 300-500 bp (minus 120 bp sequencing adaptor length = 180-380 bp insert), but the median

insert length as determined from the alignment was reported as 75 bp. The same variation is borne out in the paired-end alignment I performed: Paired-end alignments constrained to an insert window above the median **(A)** and **(B)** resulted in low numbers of unique alignments to the hg19 reference genome. Widening the window **(C)** resulted in a large number of reads aligned, the majority of which **(D)** were from mate pairs with an insert size below 200 bp.

One possible explanation for the mismatch between the median insert size and the gel-selected insert size is the tendency of PCR to preferentially amplify smaller fragments at short extension times, coupled with the imprecision of gel size selection arising from poor separation of bands at high DNA concentrations and inaccurate band excision. This explanation has been acknowledged as a likely cause (S. Quake, personal communication). Regardless, our own MolFive data did not show any deviation from the gel-selected insert size (500 bp) performed by Genomics Services (WTCHG, Oxford) during Illumina library preparation.

Baseline Contamination with Human Genomic DNA. A final comparison was made using the footprint analysis previously carried out for isolated, MDA-amplified DNA fragment and the MDA-amplified whole chromosome data from Fan *et al.*, shown in Figure 22.

The footprint of the reads shows successful isolation of chromosome 6 by Fan *et al.*, and also reveals that the amplified MolFive sample has higher baseline contamination with human DNA along the genome. Aside from contamination due to handling, one possible origin of this difference is in the on-chip handling procedure (R. Marie, DTU): chromosomes are digested using proteinase K prior to isolating separate strands. Once digested, the DNA can be much more easily broken and therefore the solution on-chip

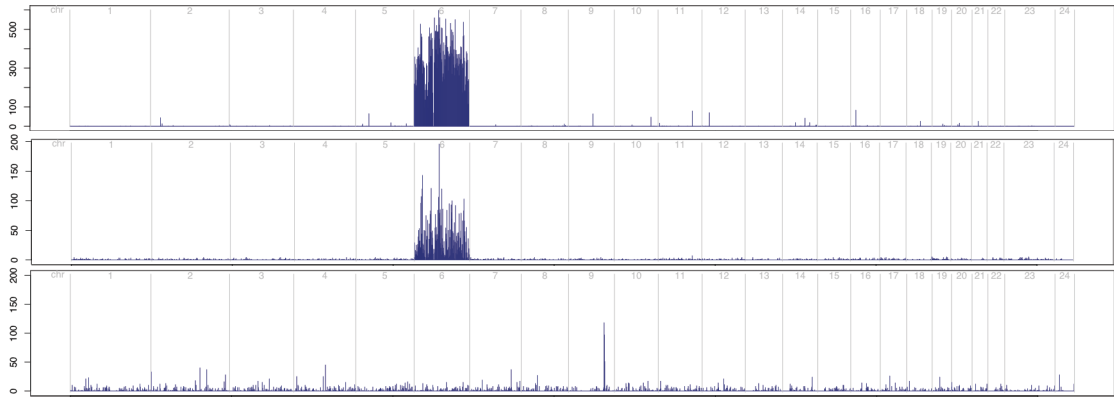


Figure 22. Footprint analysis. Comparison of reads from Fan *et al.* mapped with a (top) 0-200 bp window and (middle) 250-550 bp window for comparison to (bottom) the read footprint from a single DNA fragment. The "footprint" is the number of 1 Kbp bins per 1 Mbp that contain ≥ 16 reads. Note that the vertical grey lines represent chromosomal boundaries.

may contain smaller fragments of DNA. In contrast, Fan *et al.* preserve the compacted chromosomes until after separation from each other, giving less opportunity for cross-contamination.

Potential Improvements to MD Amplification

These results highlight the significant room for improvement in both sample handling and amplification in future experiments – despite repeated attempts, amplification of rescued single DNA fragments often failed to amplify or amplified mostly contaminating DNA instead. To address this issue in the future, a number of changes can be introduced into the amplification protocol: first, amplification can be carried out on-chip, where the total volume of the reaction will be significantly smaller and there will be less opportunity for contamination during handling steps. Second, the amplification reaction was carried out using a primer concentration optimised for bulk samples; further refinement of the amplification conditions will therefore also likely

increase the efficiency of the reaction. Third, environmental contamination can be reduced by the avoidance of commercial MD amplification kits, which have been reported to be the source of much contamination[32], and user-introduced contamination will also be reduced by minimised handling as a consequence of on-chip amplification.

Conclusion

Integrating Genome Structure and Sequence. While much of this chapter has dealt with the details of DNA sequence analysis, it is important to keep the context of the work in mind: this work demonstrates a comprehensive workflow involving DNA extraction, mapping and sequencing of ultra-long DNA molecules to examine structural variation from the microscopic to basepair level (Figure 23).

In addition to addressing SV by mapping, the use of fluidics to isolate single fragments of DNA consequently gives haplotype-phased information. Earlier, I discussed the possibility of performing post-mapping amplification on-chip; pre-mapping steps could also be incorporated, particularly for the handling, and sorting of single cells and lysis to obtain chromosomal DNA. Integrating these processing steps would allow the micro-/nano-fluidic method presented in this chapter to access information regarding cell-to-cell heterogeneity in addition to haplotype and SV.

An improved system would also increase the lengths of the DNA fragments that can be isolated on-chip. While we can readily obtain fragments > 1 Mbp in length, an ideal system would be able to 'thread' a chromosome into nanofluidic slit and be able to map it end-to-end. Our ability to obtain even-longer DNA fragments is hampered by DNA catenation in the metaphase chromosome samples, which I explore in Chapter 3 of this thesis.

While we have dealt with the integration of our platform with second-generation DNA sequencing, the platform for handling whole chromosomes would be useful to future 'third-generation' single molecule approaches as well. For example, one could imagine the use of a similar chip-based system to 'deliver' DNA to a single nanopore device

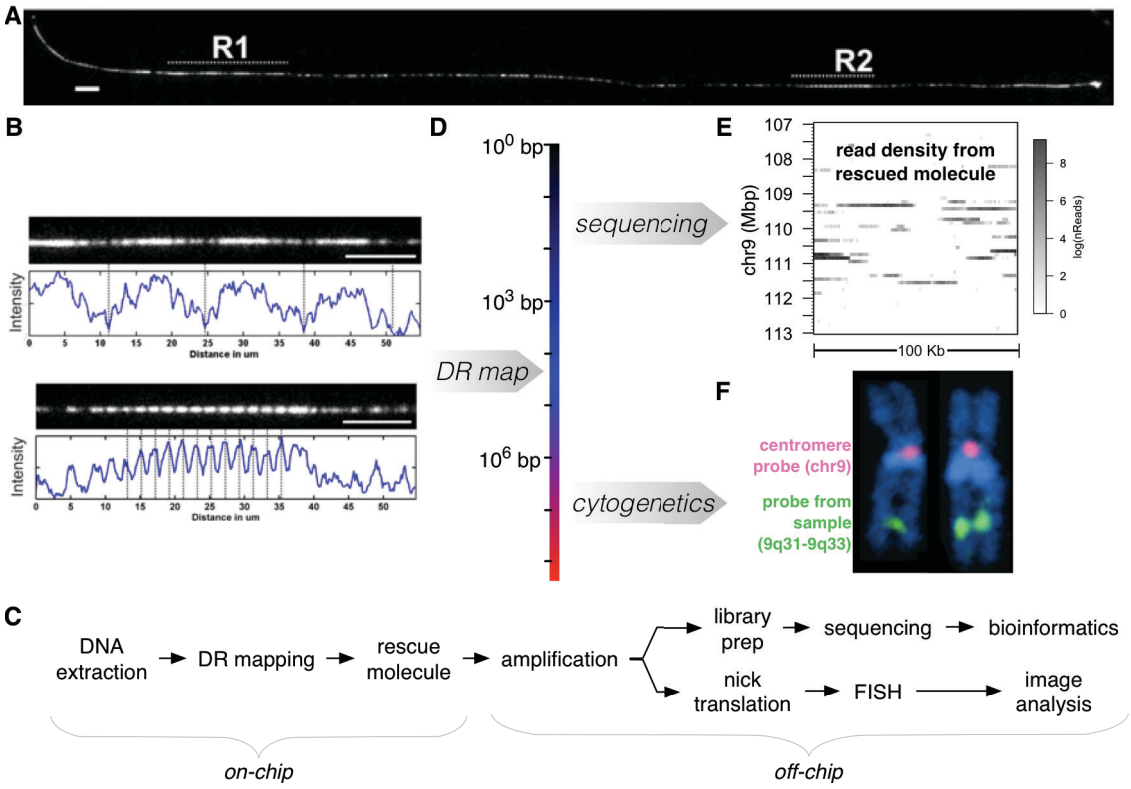


Figure 23. Workflow of integrated DR mapping and conventional genomics. (A) DR map of a single molecule from the Jurkat genome with two repeat features **(R1, R2)**. Unique barcodes from the region between **R1** and **R2** failed to find significant match to any chromosomes in hg18, but reveals **(B)** that long-range repetitive elements are present. (Scale bar: 10 µm.) **(C)** A molecule is rescued and amplified, and sequencing and FISH analysis provides complementary information **(D)**. **(E)** Sequencing reads are mapped to the reference genome to identify the origin of the rescued molecule as well as SNPs and short insertions/deletions, and **(F)** FISH provides a bird's-eye chromosomal view and confirmation of the origin identified by sequencing. (**A & B** are data from R. Marie, DTU; **F** is from M. Yusuf, Core Cytogenetics, WTCHG)

positioned at the end of a nanofluidic channel[53].

The linear display of the stretched DNA molecule in the nanoslit would also provide an opportunity to integrate third-generation sequencing methods based on single-molecule fluorescence measurements[54,55]. Multiple sequencing reactions could be initiated along the stretched DNA: the fluorescence readout at the reaction site would provide base-level sequence information, while structural context is provided by the observed location of the reaction on the stretched DNA[56]. Many of fluorescence-based approaches remain in the proof-of-concept stage[57], and there is substantial room for improvement to single-molecule methods, for example in regards to the sensitivity of single-molecule fluorescence detection and to the ability to site-specifically modify proteins, such as polymerases, that underpin many proposed third-generation DNA sequencing technologies.

I discuss improvements to imaging in Chapter 4 and improvements to site-specific protein modification in Chapter 5 of this thesis with the view that they could form the basis of enhancements to the platform presented in this chapter for the analysis of genome structure and sequence.

Contributions

This work presented in this chapter is the result of a very close 4-year collaboration between the Mir group at the Wellcome Trust Centre for Human Genetics (WTCHG, Oxford) and the Marie and Flyvbjerg groups at the at DTU Nanotech, Technical University of Denmark (DTU) and was part of a collaborative European Union FP7 project, READNA.

The overall workflow for mapping and sequencing was conceived by R. Marie (DTU) and K. Mir (WTCHG). The nanofluidic chip design and the DR mapping process were developed and carried out by R. Marie, J.N. Pedersen, and H. Flyvbjerg (DTU). Results from this portion of the project are presented in this chapter in order to provide context to my work. R. Marie and I jointly developed the idea of sliding window / rolling barcode method for detecting SV, which R. Marie extended and then implemented on genomic DNA samples. The analysis of the context of the SV and its relationship to published data that is presented in this chapter is my own work.

The work presented relating to the DNA sequencing of isolated single DNA molecules, the development of algorithms to analyse the data, and analysis of sequencing and genotyping data is my own.

Aspects of the genomics work were contributed by others, specifically amplification (D. Bauer, G. Scozzafava, and K. Mir, WTCHG), and the actual Illlumina library prep and sequencing was performed by Genomics Services (WTCHG), while Illumina genotyping was performed by Samantha Knight and Elham Sadighi Akha (WTCHG). Cytogenetic analysis and mFISH were performed by M. Yusuf, D. Moralli, and E. Volpi (WTCHG).

References

1. Marie, R. *et al.* Integrated view of genome structure and sequence of a single DNA molecule in a nanofluidic device. *Proc Natl Acad Sci USA* **110,** 4893–4898 (2013).
2. Yusuf, M. *et al.* Combining M-FISH and Quantum Dot technology for fast chromosomal assignment of transgenic insertions. *BMC Biotechnol.* **11,** 121 (2011).
3. Schröck, E. *et al.* Multicolor spectral karyotyping of human chromosomes. *Science* **273,** 494–497 (1996).
4. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nat. Genet.* **39,** S7–15 (2007).
5. Schwartz, D. C. *et al.* Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science* **262,** 110–114 (1993).
6. Lin, J. *et al.* Whole-genome shotgun optical mapping of Deinococcus radiodurans. *Science* **285,** 1558–1562 (1999).
7. Teague, B. *et al.* High-resolution human genome structure by single-molecule analysis. *Proc Natl Acad Sci U S A* **107,** 10848–10853 (2010).
8. Chan, T. R., Hilgraf, R., Sharpless, K. B. & Fokin, V. V. Polytriazoles as copper(I)-stabilizing ligands in catalysis. *Org. Lett.* **6,** 2853–2855 (2004).
9. Neely, R. K., Dedecker, P., Hotta, J. & Urbanavičiūtė, G. DNA fluorocode: a single molecule, optical map of DNA with nanometre resolution. *Chemical ...* (2010).
10. Xiao, M. *et al.* Rapid DNA mapping by fluorescent single molecule detection. *Nucleic Acids Res.* **35,** e16 (2007).
11. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30,** 771–776 (2012).
12. Lai, Z. *et al.* A shotgun optical map of the entire Plasmodium falciparum genome. *Nat. Genet.* **23,** 309–313 (1999).
13. Zhou, S. *et al.* A single molecule scaffold for the maize genome. *PLoS Genet.* **5,** e1000711 (2009).
14. Hastie, A. R. *et al.* Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. *PLoS ONE* **8,** e55864 (2013).
15. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat. Biotechnol.* **31,** 135–141 (2013).
16. Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet.* **24,** 238–245 (2008).
17. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500,** 207–211 (2013).
18. van Heesch, S. *et al.* Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. *BMC Genomics* **14,** 257 (2013).
19. Mak, H. C. Goat genome sequence by optical mapping. *Nat. Biotechnol.* **31,** 123 (2013).
20. Neely, R. K., Deen, J. & Hofkens, J. Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers* **95,** 298–311 (2011).
21. Neely, R. *et al.* DNA fluorocode: A single molecule, optical map of DNA with nanometre resolution. *Chem. Sci.* **1,** 453–460 (2010).
22. Tegenfeldt, J. O. *et al.* From the Cover: The dynamics of genomic-length DNA

molecules in 100-nm channels. *Proc Natl Acad Sci USA* **101,** 10979–10983 (2004).
23. Jo, K. *et al.* A single-molecule barcoding system using nanoslits for DNA analysis. *Proc Natl Acad Sci USA* **104,** 2673–2678 (2007).
24. Riehn, R. *et al.* Restriction mapping in nanofluidic devices. *Proc Natl Acad Sci USA* **102,** 10012–10016 (2005).
25. Tøstesen, E. Partly melted DNA conformations obtained with a probability peak finding method. *Phys Rev E Stat Nonlin Soft Matter Phys* **71,** 061922 (2005).
26. Tøstesen, E. A stitch in time: efficient computation of genomic DNA melting bubbles. *Algorithms Mol Biol* **3,** 10 (2008).
27. Reisner, W. *et al.* Single-molecule denaturation mapping of DNA in nanofluidic channels. *Proc Natl Acad Sci USA* **107,** 13294–13299 (2010).
28. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** e105 (2008).
29. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27,** 66–75 (2009).
30. Cheung, M.-S., Down, T. A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.* **39,** e103 (2011).
31. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4,** 651–657 (2007).
32. Blainey, P. C. & Quake, S. R. Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res.* **39,** e19 (2011).
33. Bredel, M. *et al.* Amplification of whole tumor genomes and gene-by-gene mapping of genomic aberrations from limited sources of fresh-frozen and paraffin-embedded DNA. *J Mol Diagn* **7,** 171–182 (2005).
34. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* (2011). doi:10.1038/nbt.1739
35. Mandelkern, M., Dattagupta, N. & Crothers, D. M. Conversion of B DNA between solution and fiber conformations. *Proc Natl Acad Sci USA* **78,** 4294–4298 (1981).
36. Li, Y. *et al.* Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat. Biotechnol.* **29,** 723–730 (2011).
37. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19,** 1527–1541 (2009).
38. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
39. Welch, R. L., Sladek, R., Dewar, K. & Reisner, W. W. Denaturation mapping of Saccharomyces cerevisiae. *Lab Chip* **12,** 3314–3321 (2012).
40. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).
41. Tüzün, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37,** 727–732 (2005).
42. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
43. Kidd, J. M. *et al.* Characterization of missing human genome sequences and

copy-number polymorphic insertions. *Nat. Methods* **7,** 365–371 (2010).
44. Vogler, C. *et al.* Microarray-based maps of copy-number variant regions in European and sub-Saharan populations. *PLoS ONE* **5,** e15246 (2010).
45. Xu, H. *et al.* SgD-CNV, a database for common and rare copy number variants in three Asian populations. *Hum. Mutat.* **32,** 1341–1349 (2011).
46. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21,** 936–939 (2011).
47. Richter, D. C., Ott, F., Auch, A. F., Schmid, R. & Huson, D. H. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS ONE* **3,** e3373 (2008).
48. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
49. Blanco, L. *et al.* Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **264,** 8935–8940 (1989).
50. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).
51. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29,** 59–63 (2011).
52. Marcy, Y. *et al.* Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet.* **3,** 1702–1708 (2007).
53. Schneider, G. F. & Dekker, C. DNA sequencing with nanopores. *Nat. Biotechnol.* **30,** 326–328 (2012).
54. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).
55. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–109 (2008).
56. Beechem, J. M. *et al.* Single Molecule Real-time DNA Sequencing using FRET-based reagents. (2011). at <http://tools.lifetechnologies.com/content/sfs/posters/cms_091831.pdf>
57. Hohlbein, J., Gryte, K., Heilemann, M. & Kapanidis, A. N. Surfing on a new wave of single-molecule fluorescence methods. *Phys Biol* **7,** 031001 (2010).

Chapter 3 – Unfolding Chromosome Structure

The work presented in this chapter has been published in:

> [1] **Bauer DLV**, Marie R, Rasmussen KH, Kristensen A, and Mir KU. DNA Catenation Maintains Structure of Human Metaphase Chromosomes. *Nucleic Acids Research*, 40, 11428-11434 (2012).

Introduction

In the previous chapter, I demonstrated that we could integrate information about the genomic structure and sequence of a single DNA molecule using a nanofluidic device, coupled with next-generation DNA sequencing. While we have been able to extract ultra-long lengths of DNA (> 1Mbp) from chromosomes, we would ideally wish to extract and process much larger pieces of chromosomal material – and ultimately be able to map and sequence a single chromosome end-to-end.

In previous experiments[2], it was found to be challenging to "unpack" the metaphase chromosomes and attempt to thread them into the nanofluidic channel: even after protease K (proK) digestion to remove chromosomal proteins, the DNA molecules that were extracted from chromosomes were tangled. DNA could only be "threaded" into the nanochannels of the silicon chip after exerting high forces on the DNA, which resulted in fragmentation of the chromosome into pieces smaller than 10 Mbp[2].

The observation of tangled DNA within chromosomes suggested that we might be observing DNA catenation within the nanofluidic chamber of our chip. In addition to the technological interest of being able to untangle chromosomes for mapping or sequencing using the methods presented in the previous chapter, these observations led

us to the realisation that the fluidic devices we were using could act as reaction chambers in which chromosomes could be manipulated and observed to test more fundamental questions about the role that DNA catenation plays in the structure of metaphase chromosomes.

The Structure of Metaphase Chromosomes

Complete and equal transmission of DNA to daughter cells is crucial during mitosis. Within each chromosome, two dimensions of organization are at play: condensation along the axes ensures the entire chromatid, end-to-end, is kept together, while the tight association of sister chromatids until anaphase, termed sister chromatid cohesion (SCC), ensures that each daughter cell receives only one copy[3]. Two mechanisms are known to play a role in SCC: DNA catenation, which physically interlocks (catenates) DNA across the sister chromatids[4]; and protein linkages through the cohesin complex, which physically tether the sister chromatids to one another[5,6].

Both linkages must be resolved to complete mitosis: cohesin is specifically cleaved by separase at anaphase onset[7]; topoisomerase II (Topo II) in yeast and its human analogue (topoisomerase IIα, Topo IIα) is required to complete DNA decatenation[8].

Unlike separase, which irreversibly removes cohesin-mediated SCC, Topo II can both introduce and remove DNA catenation involved in SCC[9]. The enzyme does strongly bias its actions towards decatenation through ATP hydrolysis[10] and by acting on severely bent DNA[11] – shown by mathematical models to be a key feature of knotted and catenated DNA[12,13]. This mechanism for preferential DNA decatenation is supported by the crystal structure of yeast Topo II, in which bound DNA is bent by 150º [14]. While this preference towards decatenation is able to prevent Topo II from extensively intertwining newly-replicated DNA, it alone is not sufficient to ensure the

complete decatenation necessary for sister chromatid separation[9]. Through stochastic action alone, a few DNA intertwines would necessarily remain[10] and these would need to be completely removed to complete mitosis[9]. Complete decatenation of sister chromatids by Topo II is therefore driven by external factors.

Recent experiments have demonstrated a complex choreography between condensin, cohesin, and DNA catenation: electrophoresis of 5 kb-long yeast centromeric plasmids has been used to show that positive supercoiling generated by condensin and induced by kinetochore attachment can drive Topo II to fully decatenate DNA before anaphase[15], though the use of longer plasmids (26 kb) and linear "minichromosomes" (42 kb) indicated that centromeric catenation was still present at metaphase and maintained by cohesin[16]. This work was able to reconcile previous results from yeast model systems with results from immunolabeled fixed mammalian cells, which showed that DNA catenation was present at metaphase and into anaphase[17-19] and could only be resolved after cohesin cleavage[20]. Whether the persistence of catenation is due to direct blocking of Topo II activity by cohesin or a consequence of the physical proximity of the chromatids promoted by cohesin and stochastic catenation-decatenation by Topo II has not yet been determined[21].

DNA catenation has also been implicated in chromosome condensation: knockdown of the condensin complex only delays condensation[22], whereas Topo II is essential[23]. *In vitro* micromanipulation experiments – in which the ends of chromosomes are aspirated into pipette tips and subjected to axial stretching forces to measure chromosome elasticity – have shown that while nuclease treatment completely "dissolves" chromosomes, protease treatment only relaxes them[24,25]. These results suggest that DNA forms the 'backbone' of the chromosome to which proteins are attached, and not *vice versa*. In subsequent experiments, Topo IIα was shown to relax mitotic

chromosomes, indicating that DNA catenation has a structural role in chromosome architecture. The changes in chromosome structure, however, could not be observed visually[26].

One of the biggest challenges in elucidating the extent, structure, and function of DNA catenation is the fact that DNA catenation itself "has never been observed directly within mitotic chromosomes," as Farcas *et al.* have noted[16]. Microscopy allows for obtaining visual images of DNA, but direct staining of DNA typically produces high background from dense chromatid bodies and observing fine structure is difficult[17]. Indirect methods have included the use of fluorescent histone fusion proteins in live cells[27] and antibody labelling of DNA-associated proteins in fixed cells[17,19] that were shown to be co-local to DNA by counterstaining with antibodies to bromodeoxyuridine, which had been incorporated into the DNA[28].

Synthesis of Project

In order to contribute to the understanding of metaphase chromosome structure – for both the better understanding of chromosome biology and for a better understanding of how to 'unpack' chromosomes for sequencing, I devised a series of experiments using our existing microfluidic lab-on-a-chip device and developed an integrated image processing pipeline to facilitate observation of DNA catenation and quantifying its effect on chromosome structure. Fluidic systems have long been used to handle metaphase chromosomes[29] and for more complex manipulations, such as on-chip cell lysis and chromosome extraction[30] and isolation for single-chromosome sequencing[31]. In order to combine the benefits of microscopy, biochemistry and micromanipulation techniques, I have used a microfluidic device[2] with a central reaction chamber that allows for real-time observation, temperature control, reagent exchange, and gentle

fluidic manipulation in order to tease apart chromosomes and examine their structure in more detail while chromosomes remain free and unfixed in solution.

Results & Discussion

Observation of DNA catenation between sister chromatids

After mounting the microfluidic chip in its holder on an inverted epifluorescence microscope (Figure 1A) and passivation of the surface with bovine serum albumin (BSA), a single chromosome was positioned within the central microfluidic reaction chamber using fluid flow. Chromosomes were stained with a DNA intercalating dye prior to loading onto the microfluidic chip and appeared as bright, dense masses of fluorescent DNA floating freely in solution (Figure 1B). To digest chromosomal proteins, proteinase K (proK) was continuously flushed around the chamber and allowed to diffuse inwards so that the chromosome could be digested without being dislodged (Figure 1C). Chromosomes rapidly expanded during digestion: after filling the height of the reaction chamber, they spread out horizontally – analogous to inflating a balloon in a thin rectangular box – but did not adhere to the 'ceiling' and 'floor' of the microfluidic chip and could be moved with fluid flow. Digestion was complete within 20 minutes and chromosomes possessed a canonical 'X' shape, with close association of sister chromatids at the centromere.

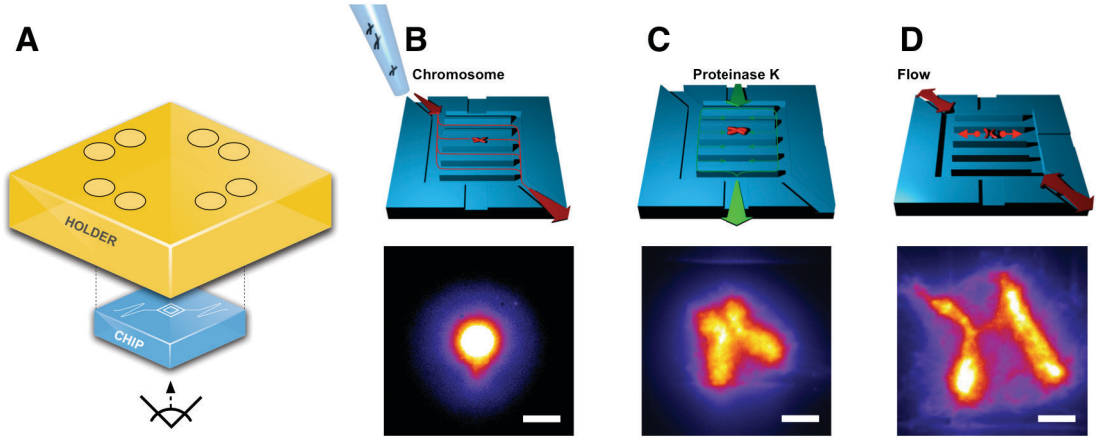


Figure 1. Direct Observation of DNA catenation between sister chromatids. (A) Experimental setup: a custom fluidic chip is mounted in a holder on an inverted fluorescence microscope. **(B)** Native, unfixed human mitotic chromosomes are loaded onto the chip and a single chromosome is positioned in the central reaction chamber. **(C)** Chromosomal proteins are digested; chromosomes expand and exhibit a canonical 'X' shape. Back-and-forth fluid flow **(D)** separates sister chromatids, revealing the DNA catenating the sister chromatids, which is concentrated at the centromere. Scale bars, 10 μm; chromosome images shown are false-coloured by intensity.

Fluidic manipulation was then used to pry the sister chromatids apart to examine the structure of the DNA tethering them. After stopping proK flow, gentle back-and-forth fluid flow was introduced in the reaction chamber. As a result, the two chromatid bodies moved apart, permitting visualization of DNA catenation that persisted between the chromatids without being obstructed by the bright fluorescence of the main mass of DNA (Figure 1D). This DNA catenation between the sister chromatids was concentrated at the centromeres, though some was present along the arms, particularly at their ends.

My observation of centromere-concentrated DNA catenation has implications for what forces might drive the decatenation of sister chromatids to ensure correct separation and complete mitosis. As discussed earlier, on its own, Topo II will both catenate and decatenate DNA, so some external 'driver' must impose directionality on Topo II to fully remove the DNA catenation.

In vivo evidence has long supported the dependence of Topo II on an external factor to complete decatenation. In prokaryotes, negative supercoiling by DNA gyrase induces Topo IV (the *E. coli* analogue of eukaryotic Topo II) to fully decatenate DNA[32], explaining previous observations that defects in gyrase lead to the failure of the nucleoid to decatenate[33]. Recent work in Topo II-deficient yeast with centromeric plasmids[15] has led to the proposal of a similar supercoiling-driven decatenation process might operate in eukaryotes, in which spindle attachment induces condensin to positively supercoil DNA, which is then fully decatenated by Topo II. This mechanism remains an open question: work with longer (and perhaps more biologically-relevant) yeast minichromosomes has not shown the same condensin-dependent behaviour, though similarly, the attachment of spindles was key to inducing DNA decatenation[16].

This observation of centromere-concentrated DNA catenation, therefore, might suggest the involvement of the most obvious driving force during mitosis: the spindle force that pulls the sister chromatids apart. Recent discussions in the literature, however, have assumed that DNA catenation is spread evenly along the length of the chromosome[16,21] and it has therefore been argued that the spindle force cannot be an efficient driver of decatenation since its force would be high at the centromeres and would diminish along the length of the chromatid arms[21]. In contrast, my observation of centromere-concentrated catenation would suggest that it is premature to rule out spindle forces as a factor in driving the decatenation of sister chromatid DNA.

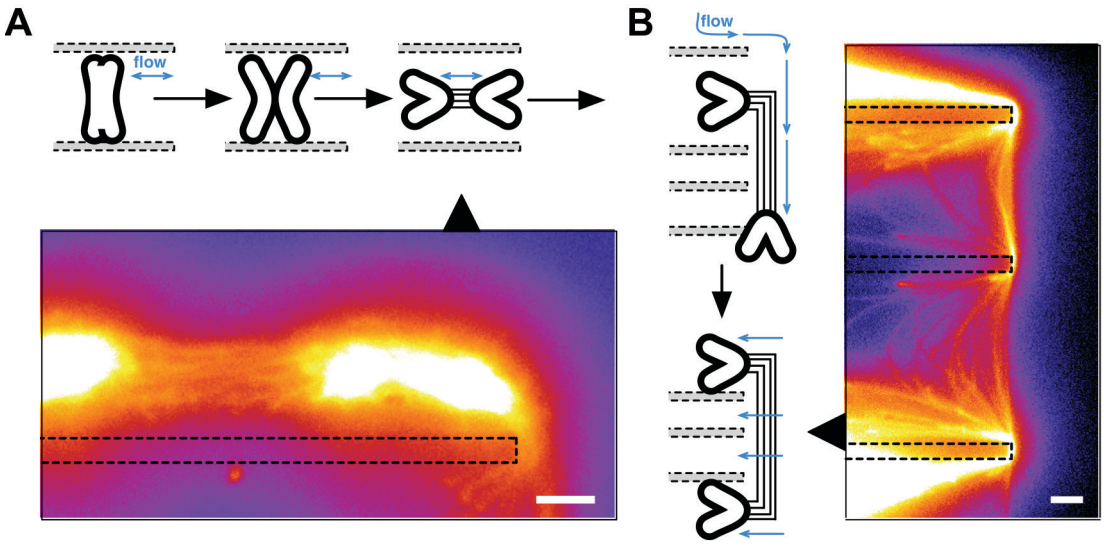


Figure 2. DNA catenation at centromeres is composed of discrete fibres. Sister chromatids are further separated with fluid flow (schematic) and the structure of the DNA catenation between the chromatids becomes visible, **(A)**. The DNA catenation is robust and continues to tether the chromatids even after entrapping the chromatids on edges within the microfluidic reaction chamber (**B**, dotted lines). Scale bars, 10 μm.

I then sought to examine the structure of the catenating DNA in more detail. Continued fluid flow allowed further separation of the sister chromatids (Figure 2): no longer constrained by chromosomal proteins, the arms of each chromatid folded outwards, leaving the centromeric regions at the centre of the mass of DNA. The fluid flow forced the bodies of the chromatids apart, pulling apart the centromere and giving a direct view of the DNA fibres catenating the two sister chromatids at the centromere (Figure 2a). The presence of fibres was confirmed by trapping the chromatids on the edges of the reaction chamber to stretch them further apart (Figure 2B). Observation under fluid flow showed the dynamics of the catenating DNA, which was organised as discrete fibres.

Overall, I observe a relatively large amount of DNA catenating the sister chromatids. While past work in yeast with short model DNAs showed that there was little DNA catenation at metaphase[34], immunofluorescence data from mammalian cells was contradictory: DNA catenation was shown to be present at metaphase and diminish as cells progress through anaphase[17,28] as a result of Topo II action after cohesin cleavage[20]. Recent work has reconciled these results with yeast models by using longer yeast minichromosomes and demonstrating that longer DNAs have greater amounts of DNA catenation present at metaphase[16]. Though we cannot exclude the possibility that the large amount of DNA we observed is being 'pulled along' by a smaller number of actual DNA catenanes, our observation of a mass of DNA catenation – as opposed to a few single fibres – is in line with these recent results that imply larger natural chromosomes will have substantial amounts of DNA catenation at metaphase[21]. In addition to the amount of DNA we observe, its organisation has implications for the current debate over how Topo II is regulated at the centromere. While Topo II must decatenate DNA to complete mitosis, centromeric cohesin prevents premature Topo II-mediated DNA decatenation at the centromere prior to anaphase. Two models have been proposed for the mechanism of Topo II inhibition by cohesin[16]: the first model proposes that the presence of cohesin directly blocks or inhibits the activity of Topo II, which is then prevented from decatenating sister chromatids. The second model proposes that Topo II remains fully active, but that cohesin serves to keep the sister chromatids in close proximity at the centromere. This proximity would then *facilitate* the action of Topo II, which would randomly catenate (and decatenate) the sisters while they are nearby. The result of this stochastic action of Topo II would be a mass of stochastically inter-catenated DNA and we would expect the structures to rapidly

become entangled *in vivo* and to observe a 'snapshot' of this disordered catenation after proK digestion.

It is noteworthy, therefore, that the DNA catenates I observe are maintained as discrete fibres at the centromere, despite being in close proximity to each other; our observations are consistent with the first model in which Topo II is not active at the centromere *in vivo*. This model would therefore support the hypothesis that cohesin can actively block Topo II activity – or at least that Topo II is co-regulated with cohesin maintenance at the centromere.

DNA catenation maintains overall shape and structure of metaphase chromosomes

After examining the structure of DNA catenation at the centromeres of metaphase chromosomes, I explored how DNA catenation affected their overall structure. I disrupted DNA catenation *in vitro* with Topo IIα before loading chromosomes into the microfluidic reaction chamber, digesting away chromosomal proteins, and quantitating differences in chromosome shape and size using an image processing routine.

While native chromosomes maintained a distinct 'X' shape after proK digestion (Figure 3A), Topo IIα-treated chromosomes exhibited a significant alteration in structure: chromatids no longer had distinct axes, distal arm separation was abolished, and the overall shape of the chromosome was more globular (Figure 3B). This alteration in morphology is the result of the stochastic decatenation (and catenation) action of Topo II during *in vitro* treatment: disordering the catenation along the arms of the chromatids removed topological constraints that provided the axial definition of the arms, and some catenation was introduced between the sister chromatids, bringing the arms together.

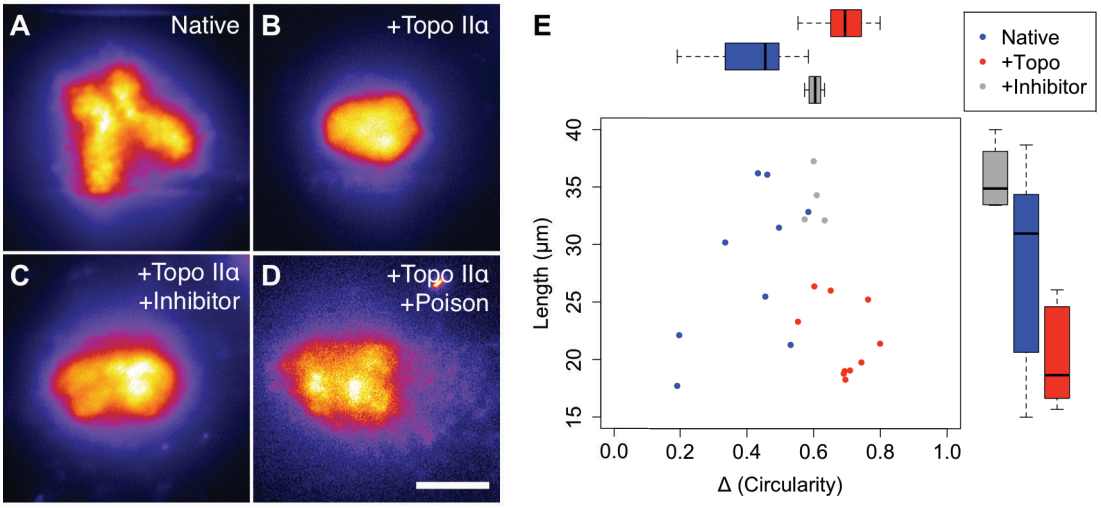


Figure 3. Disruption of DNA catenation alters the overall shape and structure of metaphase chromosomes. Native chromosomes **(A)** after proteinase K digestion on-chip exhibit typical morphology with compact axes and well-separated arms. Topo IIα-pretreated chromosomes **(B)** lose their axial definition and appear more circular. A catalytic inhibitor, dexrazoxane, partially inhibits Topo IIα and results in an intermediate morphology, **(C)**. In contrast, the presence of a poisonous inhibitor, etoposide, causes the Topo IIα to act as an endonuclease. No intact chromosomes are observed and any large DNA-containing objects **(D)** disintegrate upon proteinase K digestion. Morphometry analysis **(E)** of microscopy data quantitates the effects shown in **(A-C)** in terms of the chromosomal circularity and length and shows distinct effects of DNA catenation on gross chromosome structure. Scale bar, 20 µm.

I then modulated Topo IIα activity *in vitro* using two different types of specific inhibitors. In order to catenate and decatenate DNA, Topo IIα must generate a double-strand break (DSB) in one DNA molecule, weave the second one through the cut, and then re-ligate the first molecule. 'Standard' catalytic Topo IIα inhibitors prevent the generation of the initial double-strand break, while 'poisonous' inhibitors stabilise the

DSB-Topo IIα intermediate and prevent DNA re-ligation, effectively rendering Topo IIα an endonuclease[35]. Chromosomes were pre-treated with Topo IIα that was partially inhibited by a standard catalytic inhibitor, dexrazoxane (ICRF-187), and separately with Topo IIα in the presence of a poisonous inhibitor, etoposide (VP-16).

Chromosomes pre-treated with partially-inhibited Topo IIα in the presence of dexrazoxane had a shape in between native and Topo IIα-treated chromosomes: chromatid axes were visible and slightly elongated, but the ends of the arms were not well separated, with the chromosome appearing more as an 'H' than an 'X' shape and overall slightly round (Figure 3C). In contrast, Topo IIα treatment in the presence of etoposide resulted in few intact chromosomes. Those that appeared partially intact were subjected to on-chip proK digestion to remove chromosomal proteins (Figure 3D), but quickly released many small DNA fragments and disintegrated, consistent with the conversion of Topo IIα to nuclease-like activity.

In order to quantitate the differences we observed in chromosome shape after disrupting DNA catenation, I developed an automated image processing pipeline: grayscale images were converted to binary black-and-white images using automatic thresholding, any internal dark 'holes' missed by the thresholding were filled, and morphometry analysis was used to determine chromosome length and roundness. This pipeline was applied to timelapse images of chromosome digestions for native, Topo IIα-treated, and Topo IIα+dexrazoxane-treated chromosomes. Chromosomes that had been pretreated with Topo IIα+etoposide disintegrated and left no intact chromosomes to analyse.

After proK digestion, native chromosomes (n=9) had an average length of 28 μm, while Topo IIα-treated chromosomes (n=10) were shorter by 6 μm, with an average length of 22 μm. Topo IIα-treated chromosomes also had a greater degree of roundness than native chromosomes: 0.69 vs. 0.41. The intermediate morphology of chromosomes

pretreated with partially-inhibited Topo IIα (n=4) was also apparent: chromatid axes were slightly elongated, with an average length of 34 μm as a result of their arms remaining associated, and an intermediate roundness of 0.60. Plotting roundness vs. length for each chromosome after proK digestion (Figure 3E) revealed separate populations for each type, with catenation-disrupted chromosomes having a distinct, altered morphology as compared to native chromosomes.

Overall, these results highlight the importance of DNA catenation in higher-level chromosome structure. In particular, catenation plays a role *within* each individual chromatid, maintaining the definition of the arms while the chromosomes are condensed; this role for DNA catenation is in addition to its function at the centromere, serving to maintain sister chromatid cohesion. While there is general agreement that chromosome compaction – whether by topological changes in DNA or protein linkages – leads to decatenation between chromosomes and sister chromatids outside of the centromeres[36], results from small model DNAs in yeast have led to a model proposing supercoiling as being the primary method of DNA compaction[37]. This mechanism has yet to be reconciled with studies of larger natural chromosomes. In addition to our observations, others have observed that DNA catenation maintains chromosome rigidity[26] and that Topo II inhibition *in vivo* at metaphase causes chromosomes to visibly decondense[27]. It is possible that further studies using larger model DNAs may reveal that catenation *within* each chromatid is more extensive than previously thought, similar to recent results regarding inter-chromatid DNA catenation[16]. It also remains unclear whether DNA catenation is simply a side effect of the actions of other compacting forces such as condensin, or whether it has a distinct role in compaction on its own, in addition to any role in tethering sister chromatids.

This dual role for DNA catenation would appear to be a fundamental contradiction since catenation must be *removed* in order to disjoin sister chromatids and complete mitosis, while simultaneously *maintained* to preserve tight chromatid packing and the structure of the arms. How can both of these processes be carried out? One possible answer is that condensin blocks Topo II activity within the arms, either directly or by inducing changes in the DNA (e.g. conformation) that disfavour Topo II binding or decatenation. This mechanism would be somewhat surprising since the presence of condensin does not block Topo II activity earlier during chromosome compaction. Another possibility is that Topo II is differentially regulated at the centromere and the arms, similar to how cohesin is preserved at the centromeres by the localised actions of protein phosphatase 2A (PP2A) – which, perhaps not coincidentally, associates with Topo IIα during interphase[38] and regulates the chromosomal association of condensin II[39]. Such differential regulation would be consistent with our observations of fibrous centromeric DNA catenation, as we have discussed above. It remains to be determined whether the process of decatenation that takes place to disjoin the chromatids arises from a "wave" of supercoiling and decatenation[15], or as a result of spindle forces[16]. Our observations, however, suggest that any process that induces DNA decatenation in chromosomes at anaphase would necessarily be localised to the centromeres to ensure full disjunction of sister chromatids while maintaining intra-chromatid compaction until two separate daughter cells are formed as mitosis ends.

Conclusion

By combining our microfluidics platform with simple image analysis, we were able to explore chromosome structure and provide a direct observation of DNA catenation and its organisation. In particular, we observe key roles for DNA catenation, maintaining chromatid compaction and promoting sister chromatid cohesion, and advance the

discussion involving current models for the mechanism of sister chromatid decatenation and the regulation of Topo II. Aside from the aspects of biology that we have explored, the method presented here will extend the existing toolkit for studying the molecular organisation of chromosomes beyond conventional methods such as the electrophoresis of plasmids, indirect cell labeling, and micropipette manipulation.

Lastly, these results have implications for our attempts to unpack chromosomes for end-to-end mapping of structural variation: while metaphase chromosomes are convenient for the delivery of intact chromosomes, their extensive catenation presents unique challenges. Based on our observations, we know that we cannot release catenation by Topo IIα treatment prior to the removal of proteinaceous linkages that hold the DNA strands in close proximity and drive the decatenated / catenated equilibrium of Topo IIα away from decatenation and simply introduce stochastic catenanes.

Perhaps further optimisation of the protein digestion protocol – with a more specific and inactivatable protease – coupled with Topo II treatment might allow the simplification of chromosome catenation and facilitate the loading of entire chromosomes (or at least fragments even longer than the 1-2 Mb currently routinely achieved) into nanofluidics for analysis of structural variation using optical mapping methods or, in the future, for direct sequencing of the stretched DNA by third-generation DNA sequencing methods, discussed in Chapter 4.

Contributions

The nanofluidic chips used in this chapter were designed by K. Rasmussen and R. Marie at the Technical University of Denmark (DTU), who first observed masses of DNA catenation while attempting to extract DNA from chromosomes on-chip. The method I used for proteinase K digestion and timelapse imaging was based on a modified version of their protocol. Four of the timelapse images of native chromosomes that I analysed were acquired by K. Rasmussen. I have included them to increase the statistics shown here. Metaphase chromosomes were isolated by J. Lange (DTU) and the 3D schematics of chromosome digestion in Figure 1B-D were rendered by R. Marie.

References

1. Bauer, D. L. V., Marie, R., Rasmussen, K. H., Kristensen, A. & Mir, K. U. DNA catenation maintains structure of human metaphase chromosomes. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gks931
2. Rasmussen, K. H. *et al.* A device for extraction, manipulation and stretching of DNA from single human chromosomes. *Lab Chip* (2011). doi:10.1039/c0lc00603c
3. Nasmyth, K. Segregating sister genomes: the molecular biology of chromosome separation. *Science* **297,** 559–565 (2002).
4. Murray, A. W. & Szostak, J. W. Chromosome segregation in mitosis and meiosis. *Annu. Rev. Cell Biol.* **1,** 289–315 (1985).
5. Michaelis, C., Ciosk, R. & Nasmyth, K. Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. *Cell* **91,** 35–45 (1997).
6. Nasmyth, K. & Haering, C. H. Cohesin: its roles and mechanisms. *Annu. Rev. Genet.* **43,** 525–558 (2009).
7. Uhlmann, F., Wernic, D., Poupart, M. A., Koonin, E. V. & Nasmyth, K. Cleavage of cohesin by the CD clan protease separin triggers anaphase in yeast. *Cell* **103,** 375–386 (2000).
8. Holm, C., Stearns, T. & Botstein, D. DNA topoisomerase II must act at mitosis to prevent nondisjunction and chromosome breakage. *Mol Cell Biol* **9,** 159–168 (1989).
9. Hardy, C. D., Crisona, N. J., Stone, M. D. & Cozzarelli, N. R. Disentangling DNA during replication: a tale of two strands. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **359,** 39–47 (2004).
10. Rybenkov, V. V., Ullsperger, C., Vologodskii, A. V. & Cozzarelli, N. R. Simplification of DNA topology below equilibrium values by type II topoisomerases. *Science* **277,** 690–693 (1997).
11. Vologodskii, A. V. *et al.* Mechanism of topology simplification by type II DNA topoisomerases. *Proc Natl Acad Sci USA* **98,** 3045–3049 (2001).
12. Burnier, Y., Weber, C., Flammini, A. & Stasiak, A. Local selection rules that can determine specific pathways of DNA unknotting by type II DNA topoisomerases. *Nucleic Acids Res.* **35,** 5223–5231 (2007).
13. Buck, G. R. & Zechiedrich, E. L. DNA disentangling by type-2 topoisomerases. *J. Mol. Biol.* **340,** 933–939 (2004).
14. Dong, K. C. & Berger, J. M. Structural basis for gate-DNA recognition and bending by type IIA topoisomerases. *Nature* **450,** 1201–1205 (2007).
15. Baxter, J. *et al.* Positive supercoiling of mitotic DNA drives decatenation by topoisomerase II in eukaryotes. *Science* **331,** 1328–1332 (2011).
16. Farcas, A.-M., Uluocak, P., Helmhart, W. & Nasmyth, K. Cohesin's concatenation of sister DNAs maintains their intertwining. *Mol. Cell* **44,** 97–107 (2011).
17. Baumann, C., Körner, R., Hofmann, K. & Nigg, E. A. PICH, a centromere-associated SNF2 family ATPase, is regulated by Plk1 and required for the spindle checkpoint. *Cell* **128,** 101–114 (2007).
18. Chan, K.-L., North, P. S. & Hickson, I. D. BLM is required for faithful chromosome segregation and its localization defines a class of ultrafine anaphase bridges. *EMBO J.* **26,** 3397–3409 (2007).
19. Wang, L. H.-C., Schwarzbraun, T., Speicher, M. R. & Nigg, E. A. Persistence of

DNA threads in human anaphase cells suggests late completion of sister chromatid decatenation. *Chromosoma* **117,** 123–135 (2008).
20. Wang, L. H.-C., Mayer, B., Stemmann, O. & Nigg, E. A. Centromere DNA decatenation depends on cohesin removal and is required for mammalian cell division. *J. Cell. Sci.* **123,** 806–813 (2010).
21. Aragón, L. A double lock on sister chromatids by cohesin. *Mol. Cell* **44,** 5–6 (2011).
22. Hirano, T. Condensins: organizing and segregating the genome. *Curr. Biol.* **15,** R265–75 (2005).
23. Wang, J. C. Cellular roles of DNA topoisomerases: a molecular perspective. *Nat. Rev. Mol. Cell Biol.* **3,** 430–440 (2002).
24. Pope, L. H., Xiong, C. & Marko, J. F. Proteolysis of mitotic chromosomes induces gradual and anisotropic decondensation correlated with a reduction of elastic modulus and structural sensitivity to rarely cutting restriction enzymes. *Mol Biol Cell* **17,** 104–113 (2006).
25. Marko, J. F. Micromechanical studies of mitotic chromosomes. *Chromosome Res* **16,** 469–497 (2008).
26. Kawamura, R. *et al.* Mitotic chromosomes are constrained by topoisomerase II-sensitive DNA entanglements. *J Cell Biol* **188,** 653–663 (2010).
27. Oliveira, R. A., Hamilton, R. S., Pauli, A., Davis, I. & Nasmyth, K. Cohesin cleavage and Cdk inhibition trigger formation of daughter nuclei. *Nat. Cell Biol.* **12,** 185–192 (2010).
28. Chan, K.-L., North, P. S. & Hickson, I. D. BLM is required for faithful chromosome segregation and its localization defines a class of ultrafine anaphase bridges. *EMBO Journal* **26,** 3397–3409 (2007).
29. Young, B. D., Ferguson-Smith, M. A., Sillar, R. & Boyd, E. High-resolution analysis of human peripheral lymphocyte chromosomes by flow cytometry. *Proc Natl Acad Sci USA* **78,** 7727–7731 (1981).
30. Prinz, C., Tegenfeldt, J. O., Austin, R. H., Cox, E. C. & Sturm, J. C. Bacterial chromosome extraction and isolation. *Lab Chip* **2,** 207–212 (2002).
31. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29,** 51–57 (2011).
32. Zechiedrich, E. L., Khodursky, A. B. & Cozzarelli, N. R. Topoisomerase IV, not gyrase, decatenates products of site-specific recombination in Escherichia coli. *Genes Dev.* **11,** 2580–2592 (1997).
33. Kato, J., Nishimura, Y. & Suzuki, H. Escherichia coli parA is an allele of the gyrB gene. *Mol. Gen. Genet.* **217,** 178–181 (1989).
34. Koshland, D. & Hartwell, L. H. The structure of sister minichromosome DNA before anaphase in Saccharomyces cerevisiae. *Science* **238,** 1713–1716 (1987).
35. Jensen, L. H. *et al.* A three-dimensional quantitative structure-activity relationship study of the inhibition of the ATPase activity and the strand passing catalytic activity of topoisomerase IIalpha by substituted purine analogs. *Mol Pharmacol* **70,** 1503–1513 (2006).
36. Cuylen, S. & Haering, C. H. Deciphering condensin action during chromosome segregation. *Trends Cell Biol.* **21,** 552–559 (2011).
37. Baxter, J. & Aragón, L. A model for chromosome condensation based on the interplay between condensin and topoisomerase II. *Trends Genet.* **28,** 110–117 (2012).
38. Escargueil, A. E. & Larsen, A. K. Mitosis-specific MPM-2 phosphorylation of DNA topoisomerase IIalpha is regulated directly by protein phosphatase 2A.

Biochem J **403,** 235–242 (2007).
39. Takemoto, A. *et al.* The chromosomal association of condensin II is regulated by a noncatalytic function of PP2A. *Nat. Struct. Mol. Biol.* **16,** 1302–1308 (2009).

Chapter 4 – Using DNA as a Photonic Waveguide for Enhanced Fluorescence Imaging

Introduction

In the preceding chapters, I have shown that a platform for handling long DNA molecules and mapping structural variation (SV) can be integrated with current DNA sequencing technologies. The lab-on-a-chip device used, however, offers the possibility of integrating the DNA handling, stretching, and linear display steps with future "third-generation" DNA sequencing methods based on the analysis of single DNA molecules. In particular, the existing optical methods for SV detection can be coupled with sequencing approaches based on single-molecule fluorescence. This chapter introduces single-molecule (and, more specifically, single-fluorophore) techniques and discusses their use in existing third-generation DNA sequencing methods. I then present the design, modelling, and characterisation of DNA-based "photonic wires" to enhance imaging of fluorophores associated with DNA. I demonstrate that a waveguide can be assembled along double-stranded DNA in order to confine fluorescence illumination and reduce background signal — a key challenge for third-generation DNA sequencing methods.

Single-Molecule Fluorescence: Detection and Energy Transfer

The ability to detect single fluorescent molecules at room temperature in solution was first demonstrated over 20 years ago[1]. Subsequent application to the monitoring of biological reactions using total internal reflection fluorescence (TIRF) microscopy[2]

allowed many events to be observed in parallel. Single molecule imaging and spectroscopy have been applied across biology over the past 15 years to study systems ranging from molecular motor motion to biopolymer folding[3,4]. The advent of Förster Resonance Energy Transfer (FRET) measurements between single dye molecules (smFRET)[5] and subsequent utilisation with TIRF microscopy[6] has paved the way for sensitive analysis for populations of single molecules[7,8] – and led to the first published account 10 years ago of DNA sequence determination from single, surface-immobilised DNA molecules[9].

Described over 65 years ago[10], FRET is a physical effect in which energy from an excited 'donor' chromophore can be transferred non-radiatively to a second 'acceptor' chromophore as a result of a dipole-dipole interaction between the two[11]. As shown[12] in Equation 1, the efficiency of this energy transfer, E, is highly dependent on the distance, R, between the two chromophore dipole moments and a characteristic constant for the pair, their Förster radius, R_0:

$$E = \frac{1}{1 + (R/R_0)^6} \qquad \textbf{Equation 1}$$

The Förster radius, R_0, depends on the photophysical properties of the fluorophores and their environment: specifically the refractive index of the solution, η, the quantum yield of the donor, Q_0, the relative orientations of the two dyes, κ^2, and the degree of spectral overlap between the donor emission and the acceptor absorbance, J:

$$R_0 = \sqrt[6]{\frac{9ln(10)}{128\pi^5 N_A}\ \eta^{-4}\ Q_0\ \kappa^2\ J(\lambda)} \qquad \textbf{Equation 2}$$

where N_A is Avogadro's constant and the spectral overlap, J, is calculated as:

$$J(\lambda) = \int f_D(\lambda)\ \epsilon_A(\lambda)\ \lambda^4\ d\lambda$$ **Equation 3**

in which f_D is the normalized fluorescence emission spectrum of the donor and ε_A is the absorbance spectrum of the acceptor, measured as a function of its molar extinction coefficient over the relevant wavelengths, λ.

The process of FRET has formed the basis for a number of proposed "third-generation" DNA sequencing approaches based on the analysis of single DNA molecules.

Third-Generation DNA Sequencing

First-generation Sanger and second-generation Solexa/Illumina DNA sequencing approaches are essentially 'ensemble' assays of multiple copies of the same DNA molecule. So-called "third-generation" DNA sequencing approaches, in contrast, analyse single DNA molecules at a time[13]. While three present and past commercial approaches to single-molecule DNA sequencing (from Pacific Biosciences, Helicos, and Life Technologies) are discussed below, only the SMRT sequencing method from Pacific Biosciences is commercially available at present.

Pacific Biosciences (SMRT Sequencing)[14]. DNA sequencing is conducted in real-time by immobilising polymerase-DNA complexes[15] in the bottom of a zero-mode waveguide that confines fluorescence illumination to the small volume around the polymerase and thus allows the use of high concentrations of fluorescence species in solution[16]. Nucleotides are modified with additional phosphate groups (pentaphosphate vs. triphosphate) and a fluorescent dye is attached to the terminal phosphate. The dye group is therefore cleaved from the growing DNA strand following nucleotide incorporation by a Φ29 DNA polymerase mutant[17]. Since the incorporation of the nucleotide by the polymerase is slower than the diffusion rate, prolonged fluorescence peaks are observed as a nucleotide is incorporated, producing a distinct colour-coded

fluorescence signal that corresponds to the identity of the incorporated base[14]. While the accuracy of each individual read is less than what is presently achievable by Illumina sequencing, its ability to achieve longer read lengths (> 1000 bp) has found applicability in the scaffolding of genomes during sequence assembly[18,19] and in the sequencing of GC-rich genomes[20].

Furthermore, the single-molecule nature of the SMRT assay also allowed the kinetics of individual nucleotide incorporation events to be analysed, revealing that nucleotide incorporation opposite a methylated or hydroxymethlated template base is followed by a characteristic delay in subsequent nucleotide incorporation[21]. Measurement of this delay was used to detect genome-wide methylation in *E. coli* by SMRT sequencing[22] and was recently optimised by applying polymerase enzyme kinetic models in order to provide confidence statistics of methylation detection[23].

Helicos Biosciences[24]. The Quake group published the first demonstration of single-molecule DNA sequencing in 2003[9]. Their approach – similar to the reversible terminator chemistry employed in 'second-generation' Solexa/Illumina sequencing[25] – immobilised single DNA templates on a surface and sequentially added dye-labelled nucleotides whose subsequent extension by a polymerase was blocked by a cleavable termination group[26]. Imaging of the fluorescence signal by total internal reflection fluorescence (TIRF) microscopy and cleavage of the termination group following each nucleotide addition cycle allowed the sequence of the DNA template to be decoded[24]. While Helicos demonstrated this approach could be used to sequence single molecules of RNA directly[27] and quantify gene expression levels without amplification[28], the system did not prove commercially viable.

Life Technologies (Starlight) [29]. Life Technologies developed a real-time sequencing approach that monitored the incorporation of terminal phosphate-labelled nucleotides

by a polymerase[8,29]. Based on an approach earlier put forward by VisiGen Biotechnologies, a DNA polymerase was modified to carry a fluorescent quantum dot (QD) that acted as a donor for FRET to incoming dye-labelled nucleotides. The use of FRET effectively confined excitation of the dye-labelled nucleotides to those bound to the polymerase, allowing confined illumination (like Pacific Biosciences' SMRT) without the need for an engineered physical substrate, potentially allowing the sequencing reaction to be carried out on elongated single molecules of DNA. Despite apparent initial success[29], Life Technologies has decided not to pursue further development of the technology.

Optics for Single Molecule DNA Sequencing

The fundamental challenge facing both real-time fluorescence-based DNA sequencing

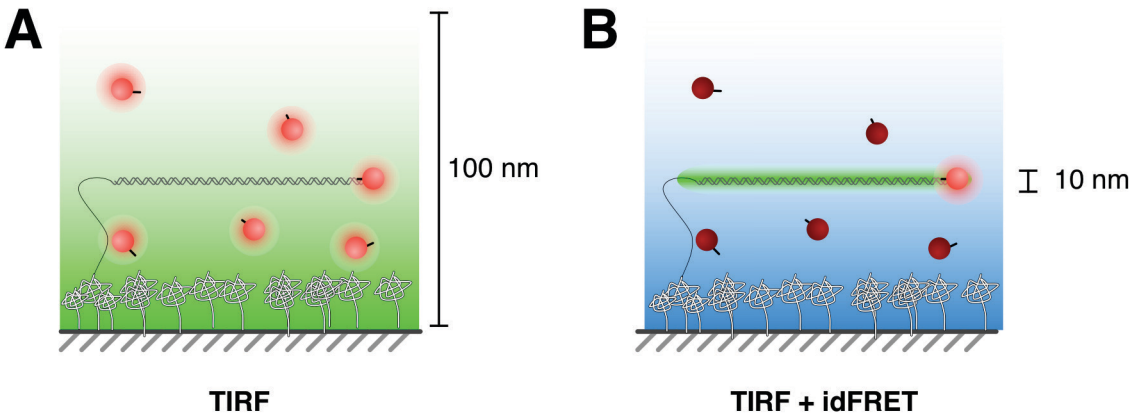


Figure 1. Illustration of illumination confinement by TIRF *vs.* TIRF in combination with intercalating-dye FRET. The use of TIRF alone **(A)** produces a decaying evanescent field of illumination that penetrates ~100 nm into the sample across the entire surface. When high concentrations of fluorescent species are present (>100 nM), considerable background fluorescence signal is present. Combining TIRF illumination with FRET, **(B)**, from a DNA intercalating dye to a DNA-associated fluorophore effectively confines illumination to a volume around the DNA with a ~10 nm radius.

technologies is the need to maximise the signal of a successful base incorporation event against background fluorescence. Dye-labelled nucleotides, in particular, are a challenge since it is necessary to balance the need for low fluorescence background against the need to maintain a nucleotide concentration in the micromolar range to drive the polymerisation reaction that underpins the sequencing process (i.e. [dNTP] > K_m of the polymerase). SMRT sequencing from Pacific Biosciences and the discontinued Starlight approach from Life Technologies each attempt to circumvent the fluorescence background problem by engineering an element of their system to confine illumination and minimise the presence of background signal: SMRT relies upon a nanofabricated zero-mode waveguide[15], while Starlight requires a polymerase engineered to carry a QD[29].

An ideal waveguide would be able to confine illumination to the area around the DNA being analysed, without the requirement of engineering a substrate or a polymerase. Here, we present one approach that meets this criterion by self-assembling a waveguide along the DNA.

We propose the use of dyes that specifically intercalate into double-stranded DNA (dsDNA). By exciting these intercalating dyes and exploiting FRET from these dyes to nearby fluorophores, only those within ~10 nm of the dsDNA would be illuminated, reducing background fluorescence signal substantially below what is achievable using TIRF[3] (Figure 1).

Dimeric DNA intercalators, such as the YOYO-1 dye used for DR mapping in Chapter 2, are particularly attractive since they have a high selectivity for dsDNA, show dramatic fluorescence enhancement upon DNA binding (typically more than 1000-fold), and have a relatively high quantum yield[30] (Table 1). These properties have

permitted their use in single-molecule fluorescence assays[26] and a variety of these dyes are available with varying fluorescence emission wavelengths (Figure 2).

While some of these intercalators, notably TOTO-1, show a strong preference for binding AT vs. GC-rich DNA, this effect is not as pronounced for others[31]. NMR studies of DNA-intercalator complexes in solution have shown that while the double helix is unwound approximately 60º at the site of intercalation, basepairs on either side of the intercalator retain canonical B-DNA geometry, with 3.4 Å spacing between base pairs and no net bend introduced in the DNA[32] (PDB: 108D). Despite the local structural deformation of the DNA, monomeric derivatives of these intercalators (e.g.

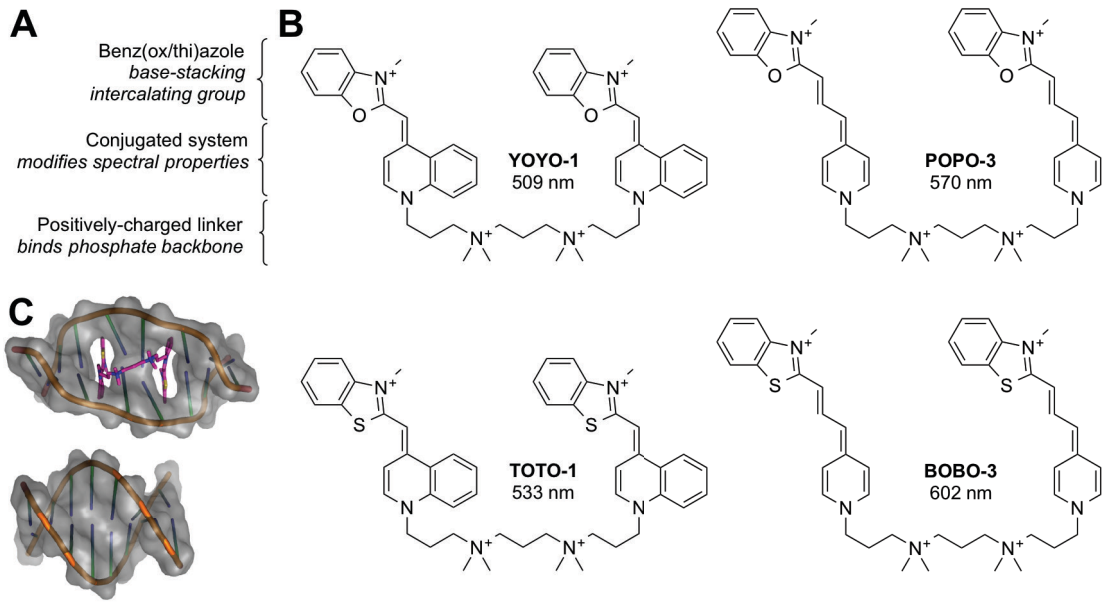


Figure 2. Structure of Dimeric Cyanine DNA-Intercalating Dyes. Dimeric intercalators contain three chemical domains, **(A)**. The structure and fluorescence emission maxima are shown for four dyes **(B)**: YOYO-1, TOTO-1, POPO-3 and BOBO-3. The structure **(C)** of the DNA–TOTO-1 complex (PDB: 108D) shows the local unwinding of the double helix and consequent lengthening as a result of intercalation as compared to the structure of standard B-DNA.

SYBR Green I) are routinely used in real-time quantitative PCR (qPCR) reactions[33], and therefore it is possible that intercalators can be used as waveguides in applications where 'correct' functioning of DNA-protein interactions must be maintained — including for DNA sequencing.

DNA-Based Photonic Wires

Wagner and Lindsay used FRET as a method of transferring energy over a linear distance nearly 20 years ago by constructing a photonic 'wire' from a chain of three prophoryn rings connecting an 'input' donor dye and an 'output' acceptor dye[34]. A number of groups subsequently extended this approach to organic fluorophores arranged on a DNA scaffold, as either an analogous chain of identical chromophores undergoing energy transfer from one to another (termed 'homo-FRET') to carry energy between an input dye and an output dye[35], or as a series of different flurorophores placed in order of increasing emission wavelength so as to induce a 'cascade' of FRET from one end of the wire to the other[36,37]. Further experimentation by the Tinnefeld group demonstrated more complex behaviour, including switchable transfer paths[38,39] and explored heterogeneity of the energy transfer process using single molecule methods[40,41]. Meanwhile, the Albinsson group proposed the use of a monomeric DNA intercalator, YO-PRO-1, to form self-assembled photonic wires between a Pacific Blue 'input' dye and a Cy3 'output' dye[42] and later extended their approach to demonstrate switchable output[43].

Synthesis of Project

In this chapter, I present work to design and characterise photonic wires that would be suitable for single-molecule fluorescence applications by employing dimeric intercalating dyes and donor and acceptor fluorophores that are typically used[3,7] in

Dye	λ_{abs} (nm)	ε ($cm^{-1}M^{-1}$)	λ_{em} (nm)	λ_{abs} - λ_{ex} (nm)	Φ	Absorption (- - -) and Emission (—) Spectra
Alexa 488	495	71,000	519	24	0.92	
YOYO-1	491	98,900	509	18	0.52	
POPO-3	534	146,400	570	36	0.46	
BOBO-3	570	147,800	602	32	0.39	
ATTO 647N	644	150,000	669	25	0.65	

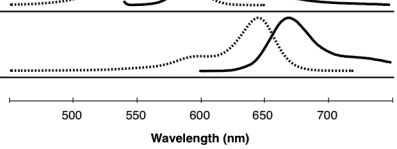


Table 1. Spectral properties of fluorescent dyes for single-molecule photonic wire constructs. Normalised absorbance and fluorescence spectra are shown.

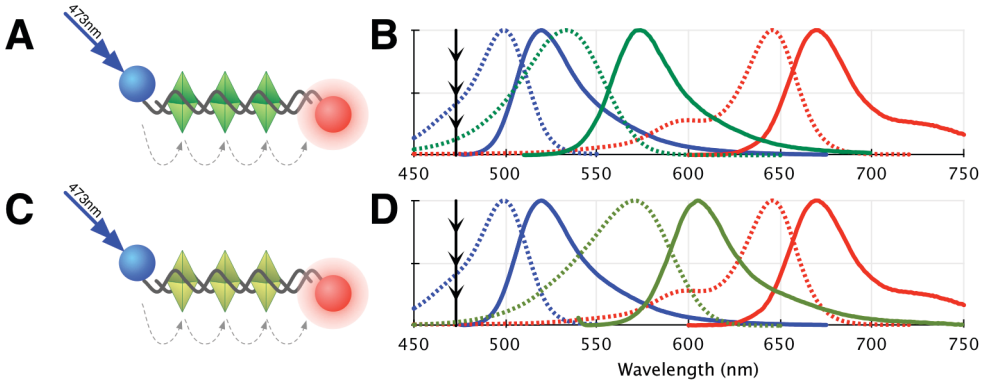
single-molecule assays (Table 1). I conclude by demonstrating how such constructs could be employed in third generation DNA sequencing applications through a series of proof-of-principle experiments.

Results

Designing a 'Point-to-Point' Photonic Wire

As a first design principle, we ensured that the fluorophores employed were compatible with existing imaging setups[44,45] for single-molecule fluorescence consisting of laser illumination at blue (473/488 nm), green (532/561 nm), and red (633/647 nm) wavelengths. As illustrated in Figure 3, the photonic wire construct consisted of a pair of complementary 33-mer DNA oligonucelotides labelled at their 3' ends: one with a single Alexa 488 dye as an 'input' for the wire, and the other with a single ATTO 647N dye as an 'output' for the wire. Photonic wires were then formed by annealing the two strands and incubating the dsDNA with an intercalating dye: either POPO-3 or BOBO-

3. Given the predicted R_0 values for each of the possible FRET processes in the wire of 33 bp (Figure 3), energy transfer into the wire from Alexa 488, within the wire between the intercalating dyes ("homo-FRET"), and out of the wire to ATTO 647N is predicted to be high. Energy transfer directly from Alexa 488 to ATTO 647N in the absence of



E

Donor	Acceptor	R_0 (Å)
Alexa 488	ATTO 647N	51
Alexa 488	POPO-3	71
Alexa 488	BOBO-3	68
POPO-3	POPO-3	52
POPO-3	ATTO 647N	58
BOBO-3	BOBO-3	55
BOBO-3	ATTO 647N	63

Figure 3. Photonic wire constructs and predicted R_0 values for dye pairs assuming free dye rotation. A schematic representation of a POPO-3–based wire, **(A)**, and the normalised absorbance and fluorescence emission spectra, **(B)**, of the constituent fluorophores. A wire containing BOBO-3 is shown in **(C)**, along with the corresponding spectra, **(D)**. Note the red-shift of the BOBO-3 absorbance and emission relative to POPO-3. **(E)**, The Förster radii, R_0, for FRET between each dye in the constructs were calculated using **Equation 2**, assuming $\kappa^2 = ⅔$. The validity of this assumption is discussed in the text.

any intercalator should be close to 0 given the approximate length of the DNA duplex in the absence of intercalating dye (33 bp ≈ 11 nm).

Modelling of the Point-to-Point Photonic Wire

I present a model of the photonic wire based on Förster theory[10] to give a qualitative prediction of the energy transfer efficiency of a photonic wire as a function of the spacing between the intercalating dyes. I begin by considering the overall energy transfer efficiency of the photonic wire as a result of multiple, discrete energy transfer events between the constituent fluorophores of the wire:

- Energy is injected into one end of the wire by exciting the blue fluorophore (Alexa 488) at 473 nm.
    I. Energy is transferred by FRET to a green intercalating dye (POPO-3 or BOBO-3)
    II. Energy is transferred along the wire by *multiple* homo-FRET steps between the green intercalating dyes, until it reaches the end of the wire
    III. Energy is transferred by FRET from the green intercalating dye to the red fluorophore
- Energy is ejected from the other end of the wire by the red dye (ATTO 647N) by fluorescence or non-radiative relaxation of the excited state.

I modelled the expected energy transfer efficiency of the wire to be the product of the FRET efficiencies of these three overall processes, written as:

$$E_{wire} = E_I \times E_{II} \times E_{III}$$

$$E_{wire} = E_{B\to G} \times (E_{G\to G})_{net} \times E_{G\to R}$$

Equation 4

The energy transfer processes (I) and (III) are single-step events and their efficiencies can be readily calculated from Equation 1 since we assume* that the terminal blue and red dyes are free to rotate in space (i.e. do not base-stack with the DNA), so we can consider $\kappa^2 = ⅔$ for Equation 2. The efficiency of the second energy transfer process (II), E_{II}, is the net result of multiple homo-FRET events between the intercalating dyes, $(E_{G \rightarrow G})_{net}$. In order to determine the efficiency of the wire overall, it is necessary therefore to consider the homo-FRET process.

At each energy step during homo-FRET, there is a certain probability that the transfer will be successful: the FRET efficiency of each step, E_{step}. I therefore considered the net FRET efficiency along the wire, $(E_{G \rightarrow G})_{net}$. To simplify notation on the follow pages, I refer to $(E_{G \rightarrow G})_{net}$ as E_{net}:

$$(E_{G \rightarrow G})_{net} = E_{net} = (E_{step})^{N_{steps}} \qquad \textbf{Equation 5}$$

I then modelled: (1) expected values for the number of energy transfer steps, N_{steps}, and (2) the FRET efficiency of each step, E_{step}, in order to predict the net energy transfer of the wire.

(1) Modelling the Number of Energy Transfer Steps. As discussed above, the energy transfer along the photonic wire is composed of multiple steps as energy is transferred from one intercalating dye to the next. In order to determine the number of steps, it is necessary to consider the path that the energy transfer is expected to take: if energy transfer 'cascaded' uni-directionally along the wire, then we could consider the total number of steps to depend simply on the number of dyes in the wire: $N_{steps} = N_{dyes} - 1$.

This 'cascade' model, however, is not reflective of the system since the homo-FRET process is not uni-directional: there is an equal probability that energy from an

* Neither Alexa 488 nor ATTO 647N are expected to base-stack with the DNA to the extent that the cyanine dyes (e.g. Cy3, Cy5) do. For a discussion, see [46] and later in this chapter.

intercalating dye in the centre of the wire will be transferred 'forwards' to the next intercalating dye as there is that the energy will be transferred 'backwards' to the previous intercalating dye. As a result of this bidirectional transfer, the number of energy transfer steps in the wire is expected to be considerably *higher* than the number of dyes.

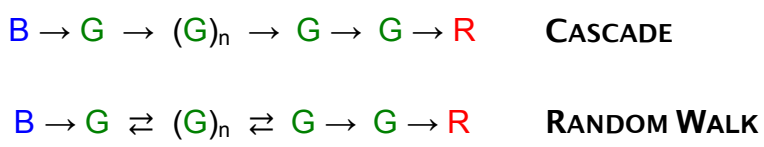


Figure 4. Energy transfer in a photonic wire: cascade vs. random walk.

To determine the distribution of expected values for N_{steps}, the path of energy transfer through the wire can be modelled as a random walk: the excitation energy of an electron in the Alexa 488 dye is modelled as a particle, an exciton, that moves in discrete steps from one dye to an adjacent one as a result of FRET. The exciton is modelled as having an equal probability of moving 'forward' or 'backward' along the wire, as illustrated in Figure 4.* Using Matlab to model the random walk, I simulated the passage of 10,000 excitons through photonic wires composed of a given number of intercalating dyes and counted the number of steps it took each exciton to exit the wire, N_{steps}. From the simulations, I determined the mean number of steps, N_{steps}, it would take an exciton to exit the wire (Figure 5B) as well as the distribution of these values for all the exciton paths (Figure 5C) for wires composed of 3 to 10 intercalating dyes.

* Note that the exciton is assumed to move at every step since the probability that the exciton is not transferred is considered separately as part of E_{step} in Equation 5.

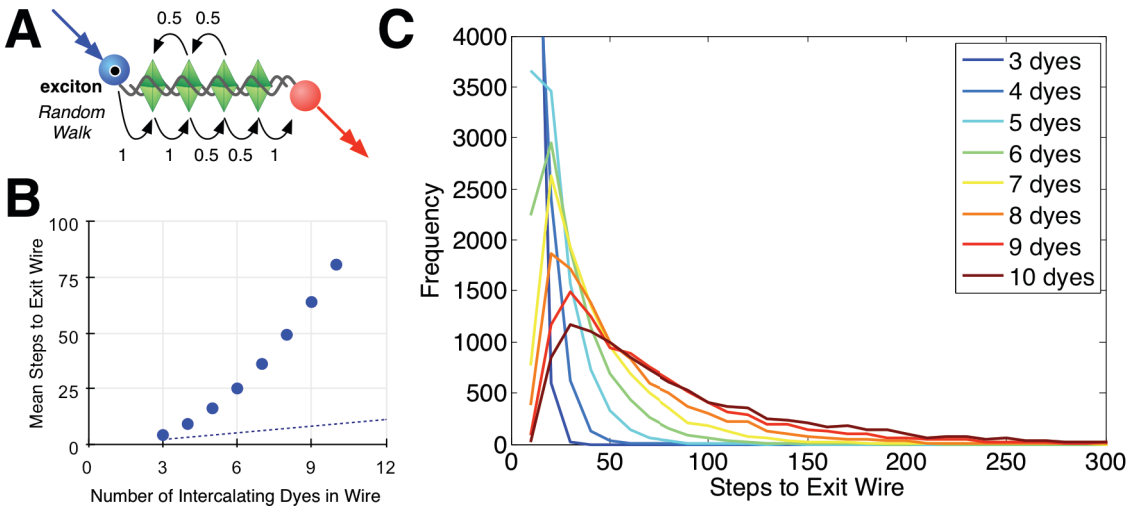


Figure 5. Simulating the path of 10,000 excitons in a photonic wire. (A) Illumination of the wire at 473 nm generates an exciton through the wire by FRET, modelled as a random walk. Simulating this process shows that increasing the number of intercalating dyes results in a greater number of mean steps to exit the wire, **(B, ●)** than if exciton travelled unidirectionally though the wire in a cascade (dashed line). Increasing the number of dyes also broadens distribution in the number of steps, **(C)**.

From the random walk simulations, we can now relate determine N_{steps} for a given N_{dyes}. We now need to determine the number of intercalating dyes that are in the photonic wire constructs. The length of our dsDNA construct is fixed (33 bp) and we know that each DNA intercalating dye encloses 2 bp of DNA from structural studies of single intercalator-DNA complexes[32]. What is not precisely known is the spacing *between* each dimeric intercalator along the DNA when the DNA is saturated with intercalating dye, as illustrated in Figure 6.

Given the fixed length of the DNA in the wire, the spacing between each intercalator, $S_{intercalator}$, will determine the total number of intercalators that will make up the photonic wire, N_{dyes}, which I calculated as:

$$N_{dyes} = \lfloor (L_{DNA} - 1) \ / \ (S_{intercalator} + 2) \rfloor \qquad \textbf{Equation 6}$$

Dimeric intercalating dyes such as POPO-3 and BOBO-3 are reported [47,48] to saturate DNA at ratios of 1 dye per 3-4 bp, corresponding to a values of $S_{intercalator}$ of 1 or 2 bp. In this case, the total number of dyes in the wire, N_{dyes}, would be 10 or 8. From the simulations of the random walk, it can be assumed that the mean number of steps, N_{steps}, it takes for an exciton to exit the wire to be 81 for 1 bp spacing, or 49 steps for 2 bp spacing (Figure 5B). I will use these values of N_{steps} with Equation 5 to predict values for the net energy transfer of the wire, E_{net} – but first need to determine the FRET efficiency of each step, E_{step}.

(2) Modelling the Efficiency of homo-FRET between Intercalating Dyes. While FRET is often thought of in terms of energy transfer between two *different* dyes with different absorbance and emission spectra, it can also occur between two dyes that are chemically and spectrally the same, provided that their absorbance and emission spectra

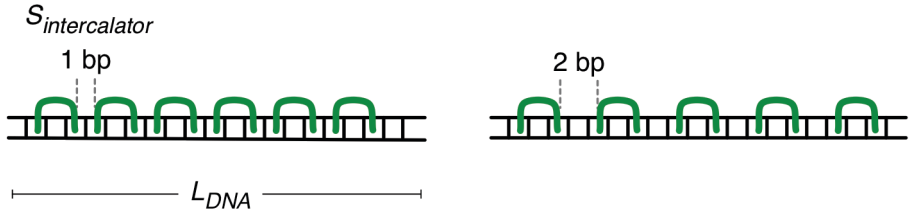


Figure 6. Spacing between DNA Intercalators. Modelling intercalator spacing along dsDNA (black) of a fixed length in basepairs, L_{DNA}: Each intercalator (green) encloses 2 bp DNA. Spacings between the intercalators, $S_{intercalator}$, of 1 bp and 2 bp are shown.

overlap (Equation 3). This process, termed homo-FRET, is treated in exactly the same manner as a FRET process between two different dyes.

The parameters chosen above for modelling the photonic wire will affect two aspects of the FRET equations: (1) the basepair spacing between the intercalating dyes, $S_{intercalator}$, determines the distance in angstroms, R, between the two intercalating dyes; (2) since the intercalators both stably base-stack with the dsDNA, their relative orientation to one another is not random. Consequently, the dye orientation factor, κ^2, is not equal to ⅔ (Equation 2) and we must determine κ^2 from the geometry of the DNA[46] for each given spacing of the intercalting dyes, $S_{intercalator}$. As a result, the characteristic Förster radius, R_0, for homo-FRET between a pair of intercalating dyes varies based on their spacing as a consequence of the change in orientation factor.

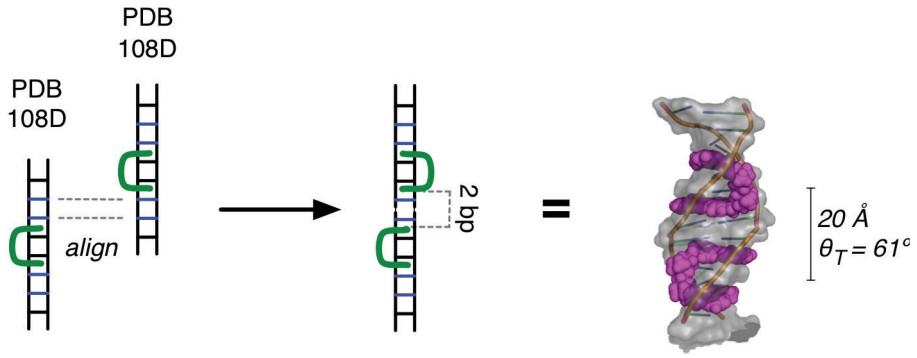


Figure 7. Modelling the geometry between 2 intercalating dyes. A single intercalator:DNA complex structure was determined by NMR (PDB: 108D). The bases proximal to the intercalator have canonical B-DNA geometry (blue) and these are used to align 2 identical structures to one another to generate a 3D model of a 2(intercalator):DNA complex at 2 bp spacing. The emission dipoles of the 2 intercalating dyes lie in parallel planes that are perpendicular to the axis of the DNA. The 2 emission dipoles lie at a relative angle of θ_T to one another.

I constructed a model for the geometry between two intercalating dyes based on available high-resolution structures of DNA-bound dimeric intercalating dye[32]. As shown in Figure 7, I concatenated 2 identical PDB structures of the intercalator-DNA complex (PDB: 108D) by aligning the two basepairs adjacent to the intercalator, which have canonical B-DNA geometry and are unperturbed by the intercalator itself. The overall geometry of the dsDNA remained unbent, even after constructing an assembly of > 8 tandem subunits, suggesting that alignment method approximated the known structure of intercalated DNA well[31]. The resulting model contained 2 bp spacing between each intercalator dimer with a distance of 20 Å between the mean chromophore positions*. The two intercalating dyes lie in parallel planes along the length of the DNA with a mean dihedral angle between their emission dipoles, θ_T, of 61º.

The homo-FRET efficiency between two intercalating dyes at 2 bp spacing can now be calculated using the measurements from the 3D structural model. The dye orientation factor κ^2 for two fluorophores lying in two parallel planes whose emission dipoles are rotated relative to one another by an angle, θ_T, can be calculated as[46]:

$$\kappa^2 = cos^2\theta_T \qquad \textbf{Equation 7}$$

Using the resulting value of κ^2 = 0.23, we can recalculate the R_0 for homo-FRET between a given intercalators using Equation 2. For two POPO-3 molecules at 2 bp spacing, the calculated energy transfer efficiency of a single homo-FRET step, E_{step}, is 0.9905. In combination with the value we determined earlier for N_{steps} from the random walk simulations (Figure 5), we can use Equation 5 to find E_{net}: $(0.9905)^{49} = 0.63$.

* The dimers were modelled as a single chromophoric units. The absorption coefficients of the monomers of the dimeric cyanine dyes are not 50% less than their counterparts ($\varepsilon_{dimer} \neq 2\varepsilon_{monomer}$), suggesting that the dimers behave – at least partially – as single chromophores with their own quantum properties[59].

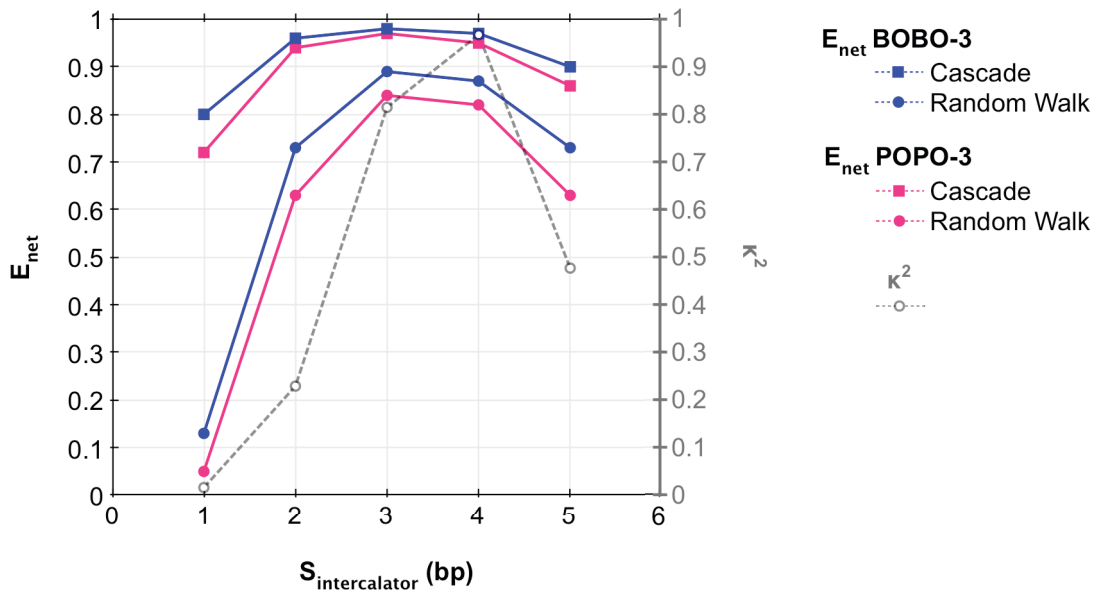


Figure 8. Predicted net energy transfer (E_{net}) of photonic wire at specified spacing of intercalating dyes ($S_{intercalator}$). Data is shown for BOBO-3 and POPO-3 intercalating dyes in which the energy transfer process follows either a unidirectional 'cascade' or a random walk. At saturation with intercalating dyes (low $S_{intercalator}$), the degree of energy transfer is predicted to be low, largely as a result of the dye orientation factor, κ^2, which based on the geometry of the intercalator:DNA complex and is the same for both POPO-3 and BOBO-3.

I then sought to predict E_{net} for intercalating dyes at different spacings: since the B-DNA geometry is known to be unperturbed in between intercalator binding sites[32], I modelled homo-FRET between POPO-3 at varying intercalator spacings by adjusting the distance between the two fluorophores by 3.32 Å/bp and their relative angles by 35.9 °/bp.* I performed the same calculations to determine BOBO-3 homo-FRET as well. The results are presented in Figure 8.

* The model I have employed assumes uniform spacing of the dyes – a reasonable assumption for cases where the DNA is saturating the intercalator, such as with 1 or 2 bp spacing. I have calculated the predicted E_{net} for larger intercalator spacings for comparison, though it is likely that there will be heterogeneity in the intercalator distributions in these scenarios.

At saturating concentrations of intercalating dye (and consequently low intercalating dye spacing), the net FRET efficiency becomes very sensitive to the relative dye orientations: Most strikingly, the modelling predicts that the transfer efficiency of a wire saturated with intercalating dye will be low.

Lastly, we can return to our equation for the overall efficiency of the wire, noting that $E_{net} = (E_{G \rightarrow G})_{net}$:

$$E_{wire} = E_I \times E_{II} \times E_{III}$$

Equation 4

$$E_{wire} = E_{B \rightarrow G} \times (E_{G \rightarrow G})_{net} \times E_{G \rightarrow R}$$

The values of $E_{B \rightarrow G}$ and $E_{G \rightarrow R}$ are expected to be close to 1 and therefore not greatly affect the value of E_{wire}, which is determines mostly by the homo-FRET term, $(E_{G \rightarrow G})_{net}$.

However, I chose to model the FRET efficiencies of the single-step transfers for completeness: $E_{B \rightarrow G}$ and $E_{G \rightarrow R}$ for FRET from Alexa 488 to an intercalator and from an intercalator to ATTO 647N were calculated by assuming they were free to rotate relative to the intercalating dye and were attached by a flexible linker to the 3' end of the DNA. The centroid of the dye accessible volume was modelled using the 'nanopositioning system'[49] and this centroid position was used to calculate the distance from the dye to the to the position of the dye and determine the FRET efficiencies. These values were then used to calculate a predicted value of E_{wire} for wires containing POPO-3 and BOBO-3, each at spacings of 1 bp and 2 bp, as shown in Table 2.

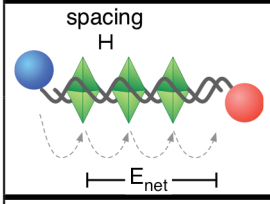
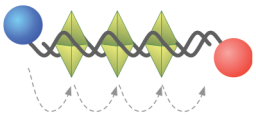
Modelled Wire Construct		Spacing	E_{net}	E_{wire}
	POPO-3	1 bp	0.05	0.05
		2 bp	0.63	0.62
	BOBO-3	1 bp	0.13	0.13
		2 bp	0.73	0.72

Table 2. Predicted photonic wire transfer efficiencies for photonic wire constructs saturated by POPO-3 or BOBO-3 intercalating dye at a spacing of 1 and 2 bp.

Fluorescence Measurements of Point-to-Point Photonic Wires

Ensemble fluorescence emission spectra were taken of POPO-3-containing photonic wire constructs under excitation of the blue, green, and red dyes. The number of photons emitted in blue, green, and red detection channels was corrected for the photon detection efficiency in each channel and then corrected for cross-talk of fluorescence emission of a given fluorophore into the others' detection channels[45].

For each photonic wire construct, we can consider the fluorescence of a particular dye at a given excitation and emission wavelength, which we denote as:

$$\text{construct}\,\mathbf{F}^{*}\,{}^{\textbf{emission}}_{\textbf{excitation}}$$

where **F*** represents the crosstalk-corrected photon counts of the excited dye that are emitted in the specified emission channel.

To illustrate: the full photonic wire contains blue, green and red dyes and is denoted as 'BGR'. After exciting the blue dye (B_{exc} = 473 nm), we can measure the crosstalk-corrected photon count, $\mathbf{F}^*$, of emission in the blue channel (B_{em} = 512 nm). This value is represented by:

$$ {}^{\mathrm{BGR}}\mathbf{F}^{*}{}^{\mathbf{B_{em}}}_{\mathbf{B_{exc}}} $$

I excited the blue dye at 473 nm to 'inject' energy into the full photonic wire construct and measured the fluorescence in three emission channels: blue emission (B_{em} = 525 nm), green emission (G_{em} = 560 nm), and red emission (R_{em} = 658 nm). I repeated the process for three other constructs that were missing the blue, red, or both terminal dyes, as shown in Table 3.

One further correction step is required since excitation of the wire at 473 nm induces direct excitation of the green intercalating dye in addition to the 'intended' excitation of

Exciting Blue Dye…

Construct	Emission: Blue	Emission: Green	Emission: Red
BGR	${}^{\mathrm{BGR}}\mathbf{F}^{*}{}^{\mathbf{B_{em}}}_{\mathbf{B_{exc}}}$	${}^{\mathrm{BGR}}\mathbf{F}^{*}{}^{\mathbf{G_{em}}}_{\mathbf{B_{exc}}}$	${}^{\mathrm{BGR}}\mathbf{F}^{*}{}^{\mathbf{R_{em}}}_{\mathbf{B_{exc}}}$
GR	${}^{\mathrm{GR}}\mathbf{F}^{*}{}^{\mathbf{B_{em}}}_{\mathbf{B_{exc}}}$	${}^{\mathrm{GR}}\mathbf{F}^{*}{}^{\mathbf{G_{em}}}_{\mathbf{B_{exc}}}$	${}^{\mathrm{GR}}\mathbf{F}^{*}{}^{\mathbf{R_{em}}}_{\mathbf{B_{exc}}}$
BG	${}^{\mathrm{BG}}\mathbf{F}^{*}{}^{\mathbf{B_{em}}}_{\mathbf{B_{exc}}}$	${}^{\mathrm{BG}}\mathbf{F}^{*}{}^{\mathbf{G_{em}}}_{\mathbf{B_{exc}}}$	${}^{\mathrm{BG}}\mathbf{F}^{*}{}^{\mathbf{R_{em}}}_{\mathbf{B_{exc}}}$
G	${}^{\mathrm{G}}\mathbf{F}^{*}{}^{\mathbf{B_{em}}}_{\mathbf{B_{exc}}}$	${}^{\mathrm{G}}\mathbf{F}^{*}{}^{\mathbf{G_{em}}}_{\mathbf{B_{exc}}}$	${}^{\mathrm{G}}\mathbf{F}^{*}{}^{\mathbf{R_{em}}}_{\mathbf{B_{exc}}}$

Table 3. Measured fluorescence for photonic wire constructs under blue excitation for emission in blue, green, and red channels.

the blue dye. I corrected the measured fluorescence of photonic wire construct (BGR) under blue excitation in each emission channel by subtracting the fluorescence of the partial (GR) construct under blue excitation. The photon counts before correction for direct excitation are represented by **F*** and after correction by **F**:

$$ {}^{\mathrm{BGR}}\mathbf{F}^{\mathrm{emission}}_{\mathbf{B_{exc}}} = {}^{\mathrm{BGR}}\mathbf{F}^{*\mathrm{emission}}_{\mathbf{B_{exc}}} - {}^{\mathrm{GR}}\mathbf{F}^{*\mathrm{emission}}_{\mathbf{B_{exc}}} \qquad \textbf{Equation 8} $$

$$ {}^{\mathrm{BGR}}\mathbf{F}^{\mathbf{B_{em}}}_{\mathbf{B_{exc}}} = {}^{\mathrm{BGR}}\mathbf{F}^{*\mathbf{B_{em}}}_{\mathbf{B_{exc}}} - {}^{\mathrm{GR}}\mathbf{F}^{*\mathbf{B_{em}}}_{\mathbf{B_{exc}}} $$

$$ {}^{\mathrm{BGR}}\mathbf{F}^{\mathbf{G_{em}}}_{\mathbf{B_{exc}}} = {}^{\mathrm{BGR}}\mathbf{F}^{*\mathbf{G_{em}}}_{\mathbf{B_{exc}}} - {}^{\mathrm{GR}}\mathbf{F}^{*\mathbf{G_{em}}}_{\mathbf{B_{exc}}} $$

$$ {}^{\mathrm{BGR}}\mathbf{F}^{\mathbf{R_{em}}}_{\mathbf{B_{exc}}} = {}^{\mathrm{BGR}}\mathbf{F}^{*\mathbf{R_{em}}}_{\mathbf{B_{exc}}} - {}^{\mathrm{GR}}\mathbf{F}^{*\mathbf{R_{em}}}_{\mathbf{B_{exc}}} $$

Similarly, the fluorescence of the partial (BG) construct under blue excitation is corrected by subtracting the fluorescence of the partial (G) construct under blue excitation:

$$ {}^{\mathrm{BG}}\mathbf{F}^{\mathrm{emission}}_{\mathbf{B_{exc}}} = {}^{\mathrm{BG}}\mathbf{F}^{*\mathrm{emission}}_{\mathbf{B_{exc}}} - {}^{\mathrm{G}}\mathbf{F}^{*\mathrm{emission}}_{\mathbf{B_{exc}}} \qquad \textbf{Equation 9} $$

$$ {}^{\mathrm{BG}}\mathbf{F}^{\mathbf{B_{em}}}_{\mathbf{B_{exc}}} = {}^{\mathrm{BG}}\mathbf{F}^{*\mathbf{B_{em}}}_{\mathbf{B_{exc}}} - {}^{\mathrm{G}}\mathbf{F}^{*\mathbf{B_{em}}}_{\mathbf{B_{exc}}} $$

$$ {}^{\mathrm{BG}}\mathbf{F}^{\mathbf{G_{em}}}_{\mathbf{B_{exc}}} = {}^{\mathrm{BG}}\mathbf{F}^{*\mathbf{G_{em}}}_{\mathbf{B_{exc}}} - {}^{\mathrm{G}}\mathbf{F}^{*\mathbf{G_{em}}}_{\mathbf{B_{exc}}} $$

$$ {}^{\mathrm{BG}}\mathbf{F}^{\mathbf{R_{em}}}_{\mathbf{B_{exc}}} = {}^{\mathrm{BG}}\mathbf{F}^{*\mathbf{R_{em}}}_{\mathbf{B_{exc}}} - {}^{\mathrm{G}}\mathbf{F}^{*\mathbf{R_{em}}}_{\mathbf{B_{exc}}} $$

No direct excitation of ATTO 647N is observed at 473 nm and therefore only the above corrections for direct excitation of the green dyes are necessary.

From the direct-excitation-corrected photon counts, we can now calculate the overall efficiency of FRET from the blue dye to the red dye observed for the full photonic wire construct (BGR):

$$ {}^{\mathrm{BGR}}\mathbf{E}_{\mathbf{B}\rightarrow\mathbf{R}} = \frac{{}^{\mathrm{BGR}}\mathbf{F}^{\mathbf{R}_{em}}_{\mathbf{B}_{exc}}}{{}^{\mathrm{BGR}}\mathbf{F}^{\mathbf{B}_{em}}_{\mathbf{B}_{exc}} + {}^{\mathrm{BGR}}\mathbf{F}^{\mathbf{G}_{em}}_{\mathbf{B}_{exc}} + {}^{\mathrm{BGR}}\mathbf{F}^{\mathbf{R}_{em}}_{\mathbf{B}_{exc}}} $$ **Equation 10**

We can calculate ${}^{\mathrm{BGR}}\mathbf{E}_{\mathbf{B}\rightarrow\mathbf{R}}$ using the measured photon counts for photonic wires containing POPO-3 as the green intercalating dye: POPO-3 ${}^{\mathrm{BGR}}\mathbf{E}_{\mathbf{B}\rightarrow\mathbf{R}} = 0.02$.

Measurements were also made for wires containing BOBO-3 as the green intercalating dye and subjected to the same corrections and calculations: BOBO-3 ${}^{\mathrm{BGR}}\mathbf{E}_{\mathbf{B}\rightarrow\mathbf{R}} = 0.04$.

We can now compare these measured values to the predicted transfer efficiency of the wire:

$$ (E_{wire})_{\mathrm{predicted}} \stackrel{?}{=} {}^{\mathrm{BGR}}\mathbf{E}_{\mathbf{B}\rightarrow\mathbf{R}} $$ **Equation 11**

A comparison of the values is shown in Table 4. We observe energy transfer efficiencies that are qualitatively in line with the modelled performance of the photonic wire at saturating concentrations of DNA intecalating dye: The level of energy transfer is low, and the effiency (${}^{\mathrm{BGR}}\mathbf{E}_{\mathbf{B}\rightarrow\mathbf{R}}$) of the wire containing BOBO-3 is higher than the efficiency of the same construct containing POPO-3.

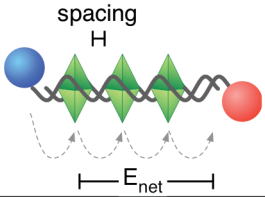
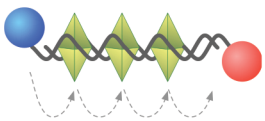
Modelled Wire Construct		Spacing	E_{net}	*Predicted* E_{wire}	*Actual* ${}^{\mathrm{BGR}}\mathbf{E}_{\mathbf{B}\rightarrow\mathbf{R}}$
	POPO-3	1 bp	0.05	0.03	**0.02**
		2 bp	0.63	0.41	
	BOBO-3	1 bp	0.13	0.09	**0.04**
		2 bp	0.73	0.47	

Table 4. Predicted *vs.* Actual FRET efficiencies of photonic wire constructs.

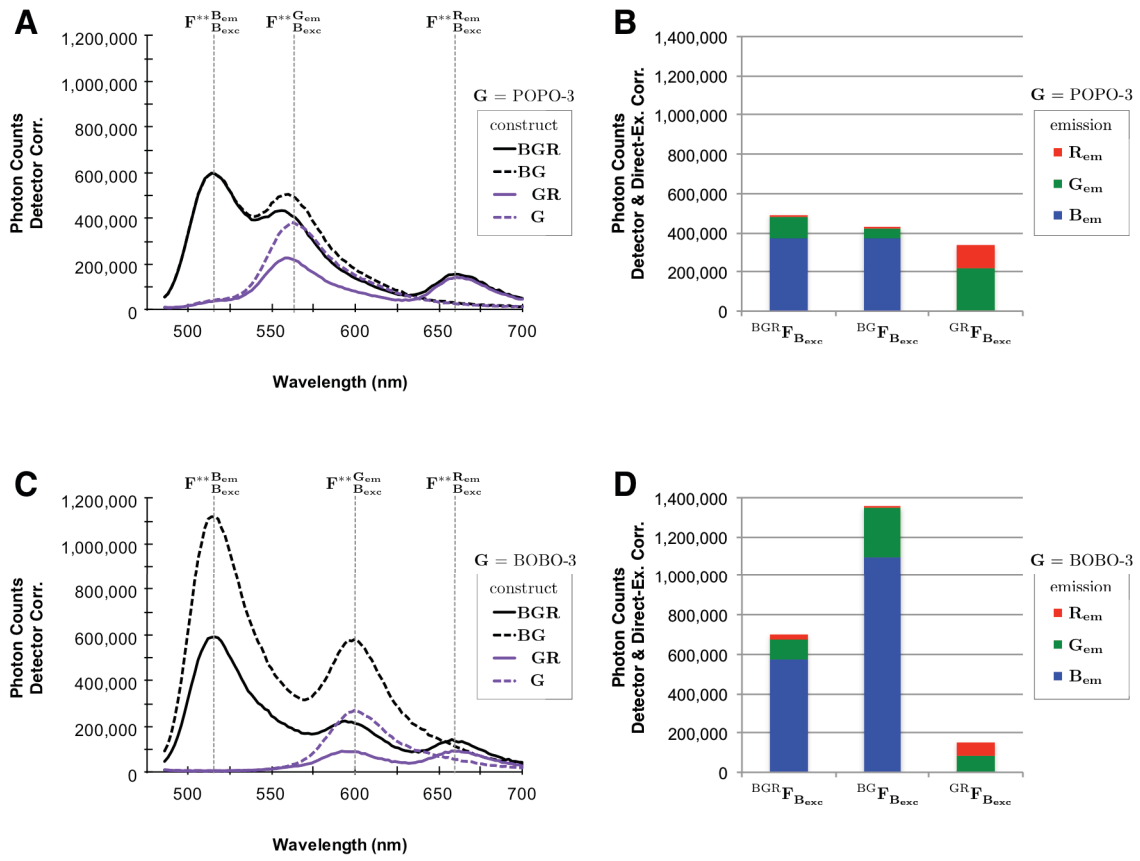


Figure 9. Inefficient energy transfer of the photonic wire is due to poor energy transfer from Alexa 488 to Intercalating Dye. Four wire constructs containing POPO-3 were excited at 473 nm and their fluorescence spectra before crosstalk correction are shown, **(A)**. The photon counts (F**) of each channel are corrected for crosstalk and direct excitation and shown after correction in **(B)**. Spectra from constructs containing BOBO-3 are shown in **(C)**,and the photon counts following crosstalk and direct excitation are shown in **(D)**. In both **(B & D)**, the majority of the photons emitted by the BGR and BG constructs are in the blue channel, indicating low transfer efficiency from Alexa 488 to the intercalating dye. In contrast, the photon counts of the GR construct suggest that the energy transfer from the intercalator to ATTO 647N is more efficient.

Evaluating The Fluorescence of Point-to-Point Photonic Wires

While I observed qualitative agreement between the expected and actual low FRET efficiencies of the photonic wires overall, examining the data in more detail yields substantial disagreement with the model. In particular, the distribution of the photon counts between the emission channels for each construct (Figure 9) in unexpected.

In the model that I employed earlier, the low energy transfer efficiency of the wire was due to the multiple homo-FRET steps between the green intercalating dyes. Consequently, we would expect the majority of the photons to be emitted in the green emission channel.

However, the majority of the photons emitted by the BGR photonic wire construct (Figure 9B and D, ${}^{BGR}\mathbf{F}_{B_{exc}}$) are in the blue channel – i.e. emitted by Alexa 488. The effect is the same, regardless of the intercalator — and regardless of the presence of the ATTO 647N dye, as indicated by the photon counts of the BG photonic wire construct (Figure 9B and D, ${}^{BG}\mathbf{F}_{B_{exc}}$). The high emission in the blue channel suggests, therefore, that the energy transfer process from Alexa 488 *to* the intercalating dye is inefficient.

It is possible that this poor efficiency is caused by a property of either of Alexa 488 or of the intercalating dye. However, the energy transfer process *from* the intercalating dye to ATTO 647N is relatively efficient, as reflected in the photon counts of the GR construct (Figure 9B and D, ${}^{GR}\mathbf{F}_{B_{exc}}$), suggesting that the low energy transfer is an effect of Alexa 488.

In my model, I assumed that the Alexa 488 and ATTO 647N dyes were unconstrained in space and not base-stacking or forming stable associations with the DNA when I calculated the R_0 between them and the intercalators and assigned $\kappa^2 = ⅔$. It is possible that the Alexa 488 *does* associate with the DNA, perhaps end-stacking on the dsDNA[50].

It is also possible to imagine that positive charge of the dimeric intercalators (Figure 2) affects electrostatic interactions of the negatively-charged Alexa 488 [51]. While the positive charges of the intercalator might interact directly with the DNA (instead of Alexa 488), the intercalator could also act to effectively shield the negative charge of the DNA from Alexa 488, consequently allowing Alexa 488 to form more stable interactions with the DNA itself.

It is also worth briefly noting a few key limitations to the model of FRET that I have used to provide a qualitative framework for my data. In addition to my assumption regarding the orientation factors of the terminally-attached dyes, I have only considered FRET as discrete events occurring between adjacent dyes. Modelling multiple FRET pathways as competing energy transfer processes with multiple *rates* of energy transfer would allow a more complex energy transfer landscape to be considered[12]. I have also disregarded alternate binding modes of the intercalating dye that might affect the quantum yield, Φ, of the dye[52].

Lastly, substantial complications arise when considering energy transfer at distances below 20 Å; at close intercalator spacings of 1 or 2 bp, we are considering distances of 17-20 Å. At this small scale, complications may arise, for example from the breakdown of inherent assumptions of Förster theory[10] since the emission of the fluorescent dyes may no longer modeled as an 'ideal' dipole[53,54]. Competing energy transfer mechanisms such as Dexter transfer[55] may also quench fluorescence: quenching has been shown to be highly efficient between fluorophores attached on double-stranded DNA at a spacing of 2 bp[56].

While it may be challenging to dissect the exact mechanism of energy transfer in the point-to-point photonic wire constructs and obtain efficient transfer from Alexa 488 to the intercalating dye, the relatively efficient transfer from directly-excited intercalating

dye to ATTO 647N that was observed in GR constructs lacking Alexa 488 suggested that a simpler photonic wire construct could be considered.

Simplified 'Feeder' Photonic Wire Construct for Third-Generation DNA Sequencing

A waveguide scheme can be considered in which an intercalating dye is excited directly and energy transfer occurs from the intercalators – collectively acting as a FRET donor – to a single acceptor dye. Excitation of the intercalating dye, therefore can be used effectively as a mechanism of exciting the DNA-attached acceptor, instead of direct illumination and excitation.

This scheme would also be more directly applicable to use along stretched genomic DNA molecules where the extra step to introduce an 'input' dye would be undesirable. Given the challenges posed by undesirable direct excitation of FRET acceptors, an ideal scheme would incorporate an intercalator and acceptor pair with as much spectral separation as possible, while maintaining efficient energy transfer.

As illustrated in Figure 10, the blue-absorbing DNA intercalator, YOYO-1, and the red-absorbing ATTO 647N are an ideal dye pair for use as a waveguide for fluorescence imaging since the dyes are spectrally separate: ATTO 647N is not directly excited to an appreciable extent at 488 nm (1.1% of maximium absorbance at 473 nm and 1.3% at 488 nm) excitation) and YOYO-1 fluorescence emission effectively does not 'bleedthrough' into the wavelengths at which ATTO 647N fluoresces (bleedthrough is < 1% above 682 nm). The degree of FRET is still predicted to be large, with an R_0 of 48 Å, assuming free dye rotation.

The use of YOYO-1 as a FRET donor has been mentioned in passing in the literature: FRET was used to image Alexa 647-labelled nucleotides along stretched, YOYO-1 stained, DNA molecules with lengths >100 kbp during DNA optical mapping[57]. No

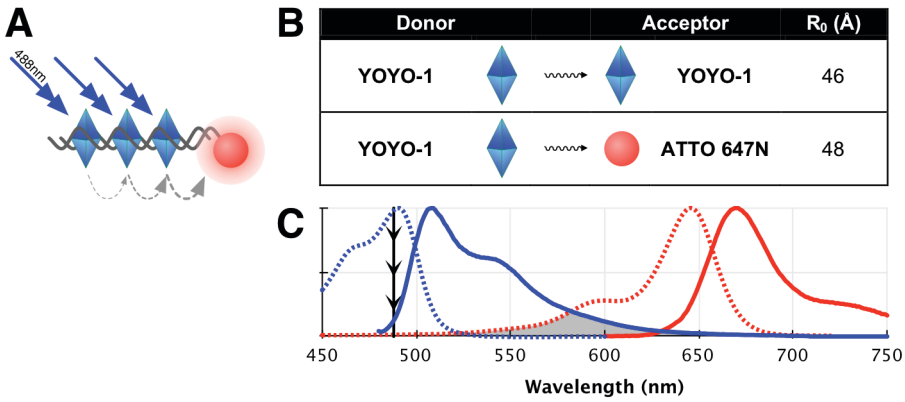


Figure 10. A 'feeder' photonic wire for single-molecule imaging. (A) Double stranded DNA is stained with YOYO-1 intercalating dye and excited along its length at 488 nm. FRET from YOYO-1 to a terminal ATTO 647N dye is predicted to be highly efficient **(B)**, as is homo-FRET from YOYO-1 to itself, despite an apparently small spectral overlap, **(C)**, of the normalised absorbance and emission spectra.

bleedthrough emission from the YOYO-1 stained DNA was visible in the Alexa 647 image.

Characterisation of Surface-Deposited Feeder Photonic Wires

To examine the performance of the YOYO-1 / ATTO 647N photonic wires, two oligonucleotides, one of which was end-labelled with ATTO 647N, were annealed to form a 20 bp dsDNA construct. I formed photonic wires by incubating the DNA with YOYO-1, which were then deposited onto a glass slide and allowed to air-dry. I imaged the spots using a laser microarray scanner using three detection schemes: YOYO-1 emission under YOYO-1 excitation ('DD'), ATTO 647N emission as a result of FRET under YOYO-1 excitation ('DA') and ATTO 647N emission under direct excitation of ATTO 647N ('AA'). Quantitation of the fluorescence intensities and calculation of the energy transfer efficiency, E, according to Equation 12 suggests that FRET from YOYO-1 to ATTO 647N is efficient ($E > ⅔$).

$$E = \frac{DD}{DD + DA}$$ **Equation 12**

As a control, I imaged the same dsDNA sequence that had been stained with YOYO-1 but that did not contain the red ATTO 647N fluorophore. No FRET signal was observed, and minimal bleedthrough signal from YOYO-1 is present in the ATTO 647N detection channel at 660 nm (Figure 11A). Quantitation of red fluorescence emission of the wire and control constructs under excitation of YOYO-1 (488 nm) or of ATTO 647N directly (633 nm) suggests that less than 10% of the fluorescence originates as 'bleedthrough' crosstalk fluorescence from the multiple YOYO-1 chromophores present (Figure 11B). These results confirm that YOYO-1 and ATTO 647N can be used as a fluorophore pair for FRET.

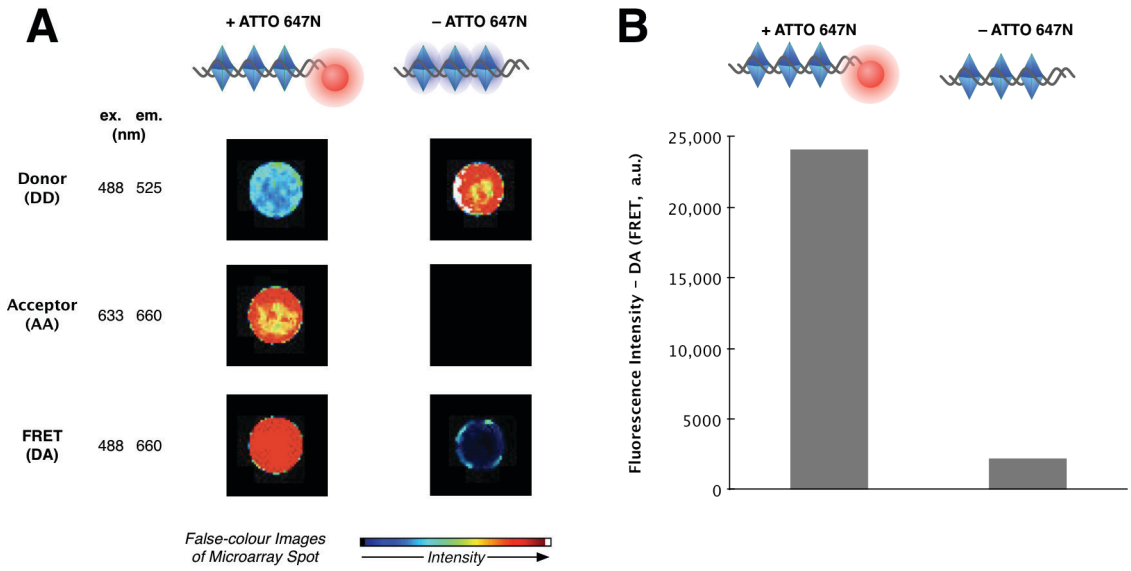


Figure 11. Fluorescence of surface-deposited YOYO-1 / ATTO 647N feeder photonic wires. (A) False-colour images of fluorescence emission show that energy transfer from YOYO-1 to ATTO 647N is efficient. **(B)** Fluorescence emission of the constructs at 660 nm under excitation at 488 nm as a result of FRET.

I then extended the assay to the single-molecule level by immobilising the feeder photonic wire constructs via a biotin-neutravadin linkage to a clean glass coverslip that had been passivated with polyethylene glycol (PEG)[58], as illustrated in Figure 12. The complexes were imaged with TIRF from a 488 nm laser and the resulting fluorescence image was split into two emission channels corresponding to fluorescence emission by

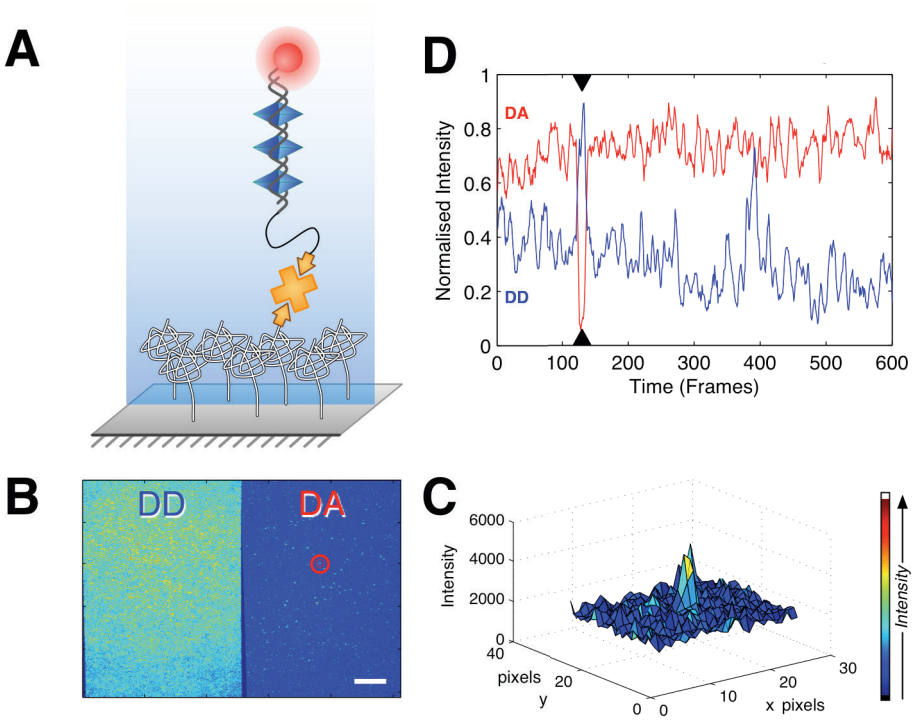


Figure 12. Single-molecule FRET imaging of photonic wire constructs. (A) DNA was immobilised by a biotin-Neutravadin linkage (yellow →X←) and imaged by TIRF using a 488 nm laser. The fluorescence image was split into two emission channels, **(B),** donor emission ('DD') and acceptor FRET emission ('DA'). The DA channel shows spots (red ○) that correspond to fluorescence of single molecules, **(C)**. Time trace, **(D)**, of fluorescence intensity from the DD and DA channels is representative of the 81 traces analysed and shows blinking of ATTO 647N and anticorrelated increase in YOYO-1 emission (▼), characteristic of smFRET. 1 frame = 100 ms. Scale bar: 5 μm

YOYO-1 ('DD') and by ATTO 647N as a result of FRET ('DA').

While there is substantial background fluorescence in the DD channel (perhaps as a result of the strong affinity of the positively-charged YOYO-1 for glass[59]), fluorescent molecules are clearly visible in DA channel. These molecules exhibit characteristic behaviour of single ATTO 647N fluorophores: timetraces of fluorescence intensity show single-step blinking – short-lived, reversible loss of fluorescence emission[60] – consistently among the 81 traces analysed.

Most interestingly, comparison of intensity timetraces of the single fluorophores in the 'DA' image to timetraces of the corresponding region of the 'DD' images shows that these blinking events are often accompanied by an anti-correlated increase in the YOYO-1 fluorescence (Figure 12D). This behaviour is characteristic of a pair of single molecules involved in FRET, and suggests that we can detect smFRET events between YOYO-1 and ATTO 647N, despite the high background in the DA channel. The detection of smFRET suggests it is possible for a photonic wire construct to be used as a method of confining illumination on a conventional single-molecule TIRF microscope.

Primer Extension on Photonic Wire Constructs: Ensemble and Single-Molecule Assays

As a final proof-of-principle for the use of intercalating dyes as waveguides for DNA sequencing reactions, I used a primer extension assay (Figure 13) to confirm (1) whether polymerisation to extend a primed DNA strand could be performed in the presence in intercalating dye, and (2) whether a dye-labelled nucleotide could be incorporated during the reaction.

I first carried out four ensemble primer extension reactions to extend a 20bp dsDNA template with a 5' overhang by a single base using the Klenow Fragment of DNA Polymerase I with removed exonuclease activity. Two of the four reactions were

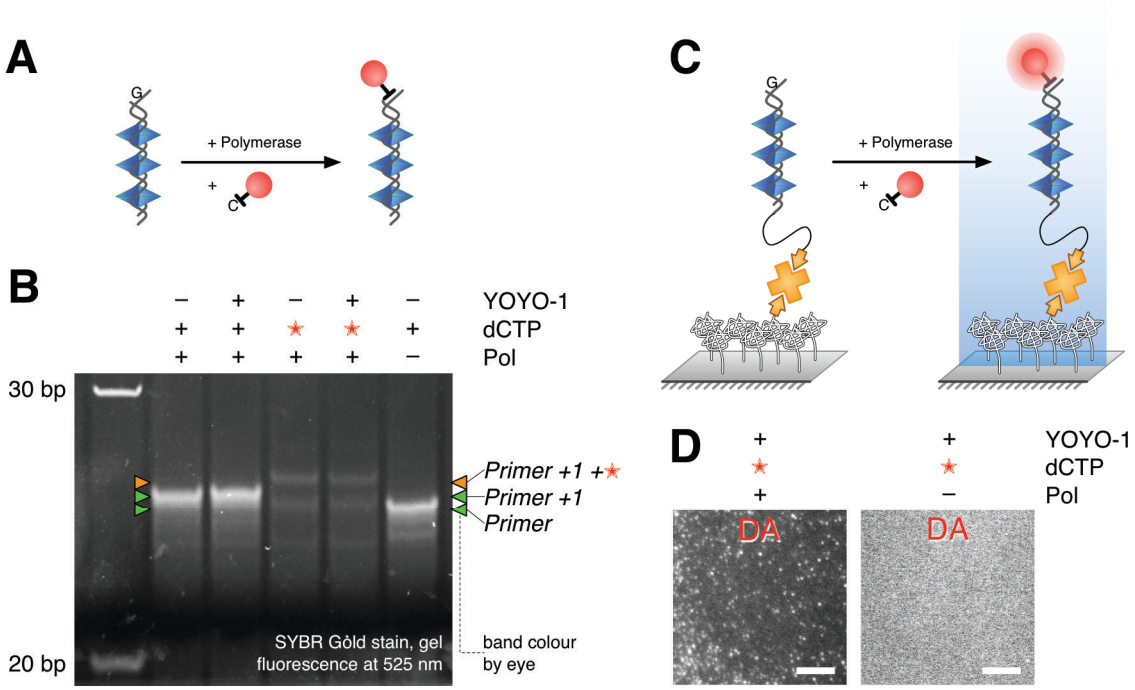


Figure 13. DNA Polymerisation on Photonic Wire: Ensemble and Single-molecule Assays. An ensemble primer extension reaction, **(A)**, was used to determine whether extension of a primed DNA strands could be carried out in the presence of YOYO-1 and if ATTO 647N-labelled nucleotides could be incorporated during the reaction. Separation of products by PAGE, **(B)**, confirms successful primer extension. FRET from the gel stain to ATTO 647N-containing bands cause them to appear orange under UV illumination (band colour indicated by orange and green ◂). The primer extension assay was performed at a single-molecule level, **(C)**, and single-molecules are visible as a result of FRET from YOYO-1 to successfully-incorporated ATTO 647N-dCTP, **(D)**. No fluorescent molecules are visible in the negative control reaction. Scale bar: 5 μm.

supplemented with YOYO-1 and two were not; one each of these reactions contained the unlabelled complementary nucleotide (dCTP), while the other two contained ATTO 647N-labelled dCTP instead. A fifth control reaction contained dCTP, but no polymerase.

The products of five reactions were separated by PAGE and the gel was stained and imaged (Figure 13B). Primer extension was successful in the presence of YOYO-1 and ATTO 647N-dCTP was incorporated by the polymerase; bands incorporating ATTO 647N-dCTP run slower than the bands incorporating unlabelled dCTP due to the presence of the dye group. These bands also appear orange by eye as a result of in-gel FRET from the SYBR Gold gel stain to the ATTO 647N. As a consequence of FRET, the ATTO 647N-containing bands appear much fainter on the 525 nm fluorescence image of the gel shown in Figure 13B.

Lastly, I carried out a modified version of the primer extension at the single-molecule level: surface-immobilised dsDNA templates containing YOYO-1 (Figure 13C) were incubated with ATTO 647N-labelled dCTP in the presence of polymerase, and in the absence of polymerase as a control for nonspecific sticking of the nucleotides. ATTO 647N fluorescence was imaged using TIRF under YOYO-1 excitation at 488 nm to obtain a FRET 'DA' image of each reaction (Figure 13D); single fluorescent molecules are visible in the primer extension reaction containing ATTO 647N-dCTP and polymerase, which are not present in the negative control when polymerase is omitted from the reaction.

The results of the ensemble and single-molecule primer extension assay indicate that it is possible to detect smFRET from YOYO-1 to ATTO 647N and that the scheme could be used to create a self-assembled waveguide to enhance imaging of DNA-associated fluorophores, for example during single-molecule sequencing of a DNA strand.

Conclusion

In this chapter, I have explored various aspects of DNAs containing intercalating dyes. While the processes of energy transfer within these photonic wire assemblies are highly complex, the qualitative model that I developed and the initial experiments that I carried out allowed us to characterise design parameters for photonic wires suitable for single-molecule experiments. I found that while energy transfer was not efficient from Alexa 488 to an intercalating dye, energy transfer *was* efficient to ATTO 647N when directly exciting the intercalating dye, suggesting that 'feeder' type photonic wires involving intercalating dye donors and a terminal acceptor would transfer energy efficiently.

I explored whether this photonic wire could be used as a tool to benefit single-molecule assays. The desire to work at higher concentrations of fluorescent species is not just limited to DNA sequencing applications: the "concentration barrier" presents a general challenge to those who use single-molecule fluorescence to study biological systems[61]. I found that YOYO-1 intercalating dye can be used at the single-molecule level to excite ATTO 647N via FRET, potentially enabling a new approach to work with higher concentrations of fluorophores in solution.

Implementation of this system as an effective waveguide to confine illumination during single-molecule assays – perhaps in future third-generation DNA sequencing – will require further optimisation, e.g. to minimise photocleavage of DNA during long observation times [48,62] and to and maximise kinetics of DNA-bound enzymes, such as polymerases, in the presence of intercalating dyes[33]. The work presented in this chapter provides results to encourage this future work.

Contributions

This project presents work done with the Kapanidis (Condensed Matter Physics, CMP, Oxford) and Mir groups (WTCHG, Oxford). The 'point-to-point' photonic wire concept has been examined by others in the Kapanidis group in the past, and I benefitted from examining their past data and reports on process, in particular those from M. Heilemann and T. Cordes (CMP). I benefitted from discussions with A. Kapanidis, J. Hohlbein (CMP and Wageningen University, The Netherlands), and A. Plochowietz (CMP) in the development and evaluation of the FRET model for the photonic wire .

'Feeder' photonic wire experiments were performed in the Mir group and K. Mir proposed the use of intercalating dyes as FRET donors during single-molecule sequencing. I received assistance from J. Hohlbein (CMP) to implement slide passivation and worked with V. Prabu (WTCHG) to optimise reaction conditions for the primer extension assay.

References

1. Brooks Shera, E., Seitzinger, N. K., Davis, L. M., Keller, R. A. & Soper, S. A. Detection of single fluorescent molecules. *Chemical Physics Letters* **174,** 553–557 (1990).
2. Funatsu, T., Harada, Y., Tokunaga, M., Saito, K. & Yanagida, T. Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution. *Nature* **374,** 555–559 (1995).
3. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5,** 507–516 (2008).
4. Walter, N. G., Huang, C.-Y., Manzo, A. J. & Sobhy, M. A. Do-it-yourself guide: how to use the modern single-molecule toolkit. *Nat. Methods* **5,** 475–489 (2008).
5. Ha, T. *et al.* Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc Natl Acad Sci USA* **93,** 6264–6268 (1996).
6. Zhuang, X. *et al.* A single-molecule study of RNA catalysis and folding. *Science* **288,** 2048–2051 (2000).
7. Hohlbein, J., Craggs, T. D. & Cordes, T. Alternating-laser excitation: single-molecule FRET and beyond. *Chem Soc Rev* (2013). doi:10.1039/c3cs60233h
8. Hohlbein, J., Gryte, K., Heilemann, M. & Kapanidis, A. N. Surfing on a new wave of single-molecule fluorescence methods. *Phys Biol* **7,** 031001 (2010).
9. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci USA* **100,** 3960–3964 (2003).
10. Förster, T. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.* **437,** 55–75 (1948).
11. Stryer, L. & Haugland, R. P. Energy transfer: a spectroscopic ruler. *Proc Natl Acad Sci USA* **58,** 719–726 (1967).
12. Lakowicz, J. R. *Principles of fluorescence spectroscopy*. (New York : Springer, 2006).
13. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19,** R227–40 (2010).
14. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).
15. Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc Natl Acad Sci USA* **105,** 1176–1181 (2008).
16. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–686 (2003).
17. Korlach, J. *et al.* Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. *Nucleosides Nucleotides Nucleic Acids* **27,** 1072–1083 (2008).
18. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500,** 207–211 (2013).
19. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. *Genome Biol.* **14,** 405 (2013).
20. Shin, S. C. *et al.* Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. *PLoS ONE* **8,** e68824 (2013).
21. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-

molecule, real-time sequencing. *Nat. Methods* **7,** 461–465 (2010).
22. Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat. Biotechnol.* **30,** 1232–1239 (2012).
23. Schadt, E. E. *et al.* Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res.* **23,** 129–141 (2013).
24. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–109 (2008).
25. Klenerman, D. & Solexa Ltd. Polynucleotide sequencing. (2000).
26. Bowers, J. *et al.* Virtual terminator nucleotides for next-generation DNA sequencing. *Nat. Methods* **6,** 593–595 (2009).
27. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461,** 814–818 (2009).
28. Lipson, D. *et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.* **27,** 652–658 (2009).
29. Beechem, J. M. *et al.* Single Molecule Real-time DNA Sequencing using FRET-based reagents. (2011). at <http://tools.lifetechnologies.com/content/sfs/posters/cms_091831.pdf>
30. Rye, H. S. *et al.* Stable fluorescent complexes of double-stranded DNA with bis-intercalating asymmetric cyanine dyes: properties and applications. *Nucleic Acids Res.* **20,** 2803–2812 (1992).
31. Netzel, T. L., Nafisi, K., Zhao, M., Lenhard, J. R. & Johnson, I. Base-Content Dependence of Emission Enhancements, Quantum Yields, and Lifetimes for Cyanine Dyes Bound to Double-Strand DNA: Photophysical Properties of Monomeric and Bichromomphoric DNA Stains. *J. Phys. Chem.* **99,** 17936–17947 (1995).
32. Spielmann, H. P., Wemmer, D. E. & Jacobsen, J. P. Solution structure of a DNA complex with the fluorescent bis-intercalator TOTO determined by NMR spectroscopy. *Biochemistry* **34,** 8542–8553 (1995).
33. Zipper, H., Brunner, H., Bernhagen, J. & Vitzthum, F. Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications. *Nucleic Acids Res.* **32,** e103 (2004).
34. Wagner, R. W. & Lindsey, J. S. A molecular photonic wire. *J Am Chem Soc* **116,** 9759–9760 (1994).
35. Vyawahare, S., Eyal, S., Mathews, K. & Quake, S. Nanometer-scale Fluorescence Resonance Optical Waveguides. *Nano Lett* **4,** 1035–1039 (2004).
36. Heilemann, M. *et al.* Multistep energy transfer in single molecular photonic wires. *J Am Chem Soc* **126,** 6514–6515 (2004).
37. Tinnefeld, P., Heilemann, M. & Sauer, M. Design of molecular photonic wires based on multistep electronic excitation transfer. *Chemphyschem* **6,** 217–222 (2005).
38. Tinnefeld, P. & Sauer, M. Molecular Optical Switches and Waveguides. *Optik & Photonik* **2,** 45–48 (2007).
39. Stein, I., Steinhauer, C. & Tinnefeld, P. Single-Molecule Four-Color FRET Visualizes Energy-Transfer Paths on DNA Origami. *J Am Chem Soc* **133,** 4193–4195 (2011).
40. Heilemann, M., Kasper, R., Tinnefeld, P. & Sauer, M. Dissecting and reducing the heterogeneity of excited-state energy transport in DNA-based photonic wires. *J Am Chem Soc* **128,** 16864–16875 (2006).
41. Sánchez-Mosteiro, G. *et al.* DNA-based molecular wires: multiple emission

pathways of individual constructs. *J Phys Chem B* **110,** 26349–26353 (2006).
42. Hannestad, J. K., Sandin, P. & Albinsson, B. Self-assembled DNA photonic wire. *Nucleic Acids Symp Ser (Oxf)* 685 (2008). doi:10.1093/nass/nrn346
43. Hannestad, J., Gerrard, S., Brown, T. & Albinsson, B. Self-Assembled DNA-Based Fluorescence Waveguide with Selectable Output. *Small* **7,** 3178–3185 (2011).
44. Kapanidis, A. N. *et al.* Fluorescence-aided molecule sorting: analysis of structure and interactions by alternating-laser excitation of single molecules. *Proc Natl Acad Sci USA* **101,** 8936–8941 (2004).
45. Lee, N. K. *et al.* Three-color alternating-laser excitation of single molecules: monitoring multiple interactions and distances. *Biophys. J.* **92,** 303–312 (2007).
46. Iqbal, A. *et al.* Orientation dependence in fluorescent energy transfer between Cy3 and Cy5 terminally attached to double-stranded nucleic acids. *Proc Natl Acad Sci USA* **105,** 11176–11181 (2008).
47. Günther, K., Mertig, M. & Seidel, R. Mechanical and structural properties of YOYO-1 complexed DNA. *Nucleic Acids Res.* **38,** 6526–6532 (2010).
48. Tycon, M. A., Dial, C. F., Faison, K., Melvin, W. & Fecko, C. J. Quantification of dye-mediated photodamage during single-molecule DNA imaging. *Anal. Biochem.* **426,** 13–21 (2012).
49. Muschielok, A. *et al.* A nano-positioning system for macromolecular structural analysis. *Nat. Methods* **5,** 965–971 (2008).
50. Iqbal, A., Wang, L., Thompson, K. C., Lilley, D. M. J. & Norman, D. G. The structure of cyanine 5 terminally attached to double-stranded DNA: implications for FRET studies. *Biochemistry* **47,** 7857–7862 (2008).
51. Panchuk-Voloshina, N. *et al.* Alexa dyes, a series of new fluorescent dyes that yield exceptionally bright, photostable conjugates. *J. Histochem. Cytochem.* **47,** 1179–1188 (1999).
52. Paik, D. H. & Perkins, T. T. Dynamics and Multiple Stable Binding Modes of DNA Intercalators Revealed by Single-Molecule Force Spectroscopy. *Angew Chem Int Ed Engl* (2011). doi:10.1002/anie.201108162
53. Muñoz-Losa, A., Curutchet, C., Krueger, B. P., Hartsell, L. R. & Mennucci, B. Fretting about FRET: failure of the ideal dipole approximation. *Biophys. J.* **96,** 4779–4788 (2009).
54. Hirschfeld, V., Paulsen, H. & Hübner, C. G. The Spectroscopic Ruler Revisited at 77 K. *Phys. Chem. Chem. Phys.* – (2013). doi:10.1039/C3CP51106E
55. Dexter, D. L. A Theory of Sensitized Luminescence in Solids. *J. Chem. Phys.* **21,** 836 (1953).
56. Cordes, T. *et al.* Sensing DNA opening in transcription using quenchable Förster resonance energy transfer. *Biochemistry* **49,** 9171–9180 (2010).
57. Jo, K. *et al.* A single-molecule barcoding system using nanoslits for DNA analysis. *Proc Natl Acad Sci USA* **104,** 2673–2678 (2007).
58. Selvin, P. R. & Ha, T. *Single-molecule techniques : a laboratory manual*. (Cold Spring Harbor, N.Y. : Cold Spring Harbor Laboratory Press, 2008).
59. *Dimeric Cyanine Nucleic Acid Stains*. (Molecular Probes, Inc., 2000). at <http://probes.invitrogen.com/media/pis/mp03600.pdf>
60. Kapanidis, A. & Weiss, S. Fluorescent probes and bioconjugation chemistries for single-molecule …. *J. Chem. Phys.* (2002).
61. Loveland, A. B., Habuchi, S., Walter, J. C. & van Oijen, A. M. A general approach to break the concentration barrier in single-molecule imaging. *Nat. Methods* **9,** 987–992 (2012).

62. Akerman, B. & Tuite, E. Single- and double-strand photocleavage of DNA by YO, YOYO and TOTO. *Nucleic Acids Res.* **24,** 1080–1090 (1996).

Chapter 5 – Engineering Polymerases using Unnatural Amino Acids

Introduction

In the previous chapters, I have discussed the possibility of performing third-generation DNA sequencing on stretched DNA using methods based on single-molecule fluorescence assays. Like most DNA sequencing approaches, these rely upon engineered polymerases to modify activity and, crucially, to bind the polymerase to elements of the fluorescence imaging system – such as an attached fluorophore for FRET[1] or a zero-mode waveguide[2].

Beyond DNA sequencing, single-molecule imaging and FRET can act as useful tool to explore protein structure and dynamics, both on short and long DNA[3]. In these assays, FRET is exploited as a precise molecular ruler, instead of a mechanism of conveying energy or exciting fluorophores as a 'waveguide'. Given the sensitivity of FRET to changes in distance (Equation 1, Chapter 4), FRET can be used to detect basepair-level changes in distance[4,5]. Successfully implementing these assays often hinges on the ability to incorporate a fluorophores at a particular location on a protein-of-interest[6]. Conventional labelling approaches used for FRET, summarised below, put substantial constraints on options for site-specific incorporation of fluorescent dyes. Two parallel avenues of research over the past ten years have now opened the possibility of near-universal site-specific labelling of proteins using orthogonal chemistries: we now have the ability to generate tRNA/synthetase pairs to incorporate a wide variety of unnatural amino acids into proteins during *in vivo* translation, and can take advantage of the

development of specific, orthogonal labelling chemistries – notably click chemistry – whose reactive groups can be incorporated into an unnatural amino acid and do not cross-react with the functional groups present on native amino acids during labelling[7]. I discuss these developments after summarising conventional labelling approaches, below, and then apply them to a model nucleic acid processing enzyme, *E. coli* RNA Polymerase.

Conventional Labelling Approaches for Single Molecule Fluorescence

Single-molecule studies of DNA-interacting proteins have been facilitated by the relative ease with which fluorescent dyes can be coupled to oligonucleotides[8]. Amino- or thiol-modified bases are incorporated into a DNA/RNA oligonucleotide during solid-phase oligonucleotide and are subsequently labelled using activated ester- or maleimide-functionalised fluorescent dyes, respectively (Figure 1A). These oligonucleotides can then be used either directly as short DNA constructs, or incorporated into larger DNAs by ligation[3] or as labelled primers for generating longer constructs by PCR[9].

Site-specific modification of proteins is generally more complex[8]: the most general approach is the removal of surface-accessible cysteine residues and the introduction of a single cysteine residue (or the preservation of a single native residue) at the desired labelling location, followed by reaction with maleimide derivatives of fluorescent dyes (Figure 1B).

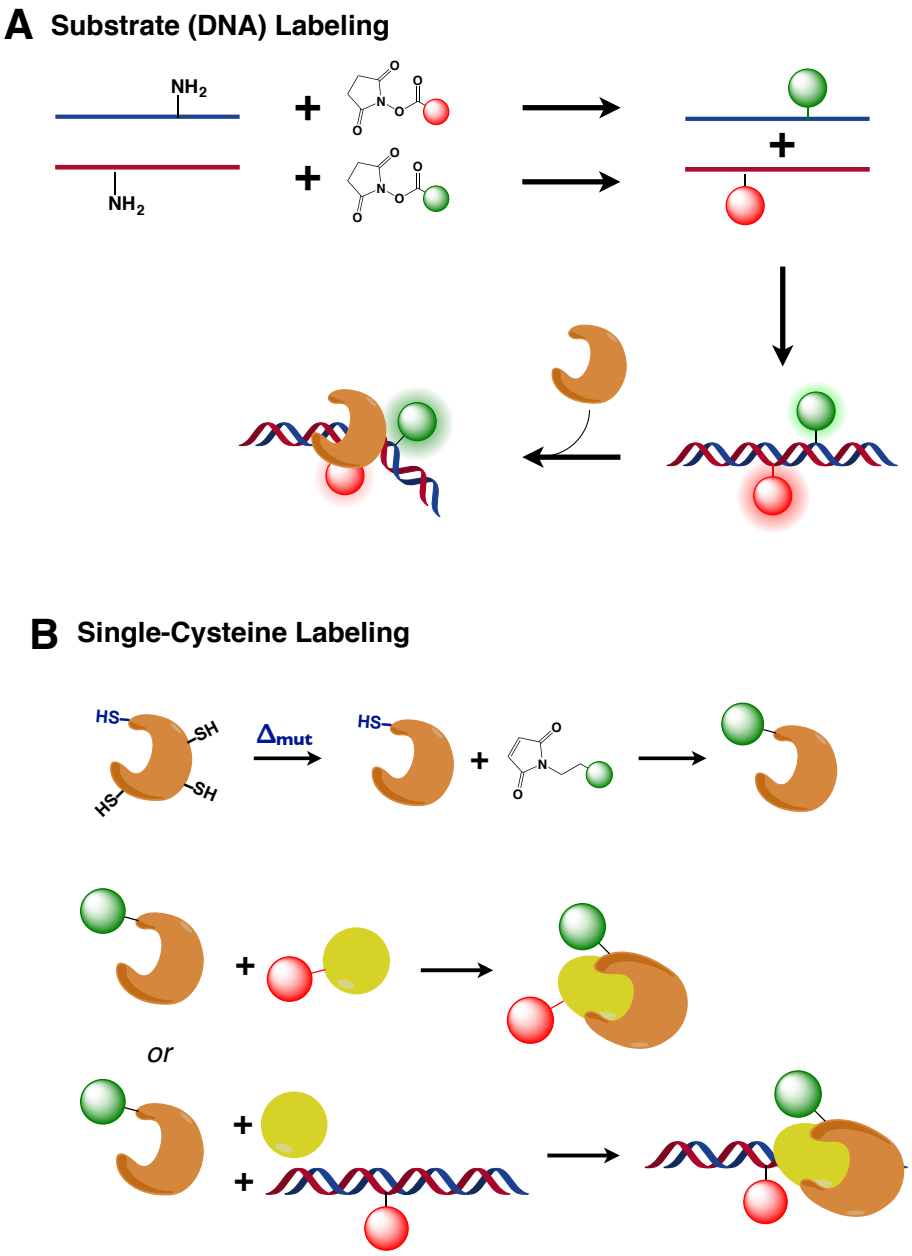


Figure 1. Conventional site-specific labelling strategies for single-molecule FRET assays. In a simple approach, **(A)**, substrate DNAs can be labelled by incorporation of amino-modified nucleobases during oligonucleotide synthesis and labelling with NHS-esters of fluorescent dyes. Changes in DNA conformation can be examined upon the addition of an interacting protein. Alternatively, proteins themselves can be labelled, **(B)**, at particular sites by removing surface-exposed cysteine residues by site-directed mutagenesis and leaving (or adding) a single cysteine residue, whose thiol group can be specifically labelled using maleimide-coupled dyes. A second labelled subunit or other interacting protein can be added and conformational changes can be monitored by FRET. Alternatively, if the protein interacts with DNA, FRET can be measured between a labelled protein and DNA.

While single- cysteine labelling has found applicability to *in vitro* studies of a variety of nucleic acid-processing enzymes[6,10,11], the approach suffers from the fact that it is not generally applicable: many proteins possess large numbers of cysteine residues, some of which play key functional roles. In multisubunit proteins, smaller subunits may often be labelled using this approach, while larger ones cannot[12].

Incorporation of Unnatural Amino Acids into Proteins

The incorporation of unnatural amino acids (UnAAs)[*] has been used as a tool for exploring protein structure since the 1990s when the Schultz group demonstrated that UnAAs could be incorporated into proteins during *in vitro* translation in order to probe structure[13]. In their initial approach, the codon for the amino acid they wished to substitute was mutagenized to the amber stop codon, TAG, and the *in vitro* translation reaction was supplemented with a synthetic $tRNA_{CAU}$ that had been previously coupled with an unnatural amino acid of interest[14]. The presence of synthetic $tRNA_{CAU}$ was able to efficiently suppress innate termination by native release factors for the amber stop codon, thus allowing for the UnAA to be incorporated at the desired site and translation of the protein to continue.

Other groups sought to incorporate UnAAs *in vivo* by exploiting the tolerance of innate *E. coli* tRNA synthetases to charge tRNAs with amino acid analogues having similar structures to their natural counterparts – valine and methionine, in particular[15,16]. Notably, the Bertozzi group demonstrated that this approach could be used to site-specifically incorporate a methionine analogue, azidohomoalanine, suitable for site-specific labelling[17]. In addition to the need to remove endogenous methionines from a protein of interest, this approach required the use of *E. coli* strains that were not capable

[*] Unnatural amino acids are often referred to as 'UAA's in the literature. I have used 'UnAA' to avoid confusion with the (unrelated) ochre stop codon.

of synthesising methionine themselves and the use of methionine-depleted culture media (supplemented with azidohomoalanine) during protein expression.

The Schultz group successfully extended their amber-suppression approach to *in vivo* protein expression by cloning naturally-occurring genes for a $tRNA_{CAU}$ and its corresponding synthetase from the archaebacterium *M. jannaschii* into *E. coli*[18]. Directed evolution of the tRNA/synthetase pair in *E. coli* led to a general approach for the development of machinery capable of specifically incorporating a wide variety of unnatural amino acids suitable for chemical ligation, including the azide-containing 4-azidophenylalanine[19,20]. The Schultz group devised a general method for UnAA incorporation into a protein of interest[21]: a plasmid containing the protein of interest is mutagenized to contain the amber codon at the desired UnAA incorporation site; this plasmid is then co-transformed into *E. coli* with a second plasmid that encodes the machinery (tRNA/synthetase) for the desired UnAA; protein expression is then carried out, supplementing the culture media with the UnAA prior to induction. Notably, this method allows different UnAAs to be incorporated into the same protein location simply by co-transforming the plasmid encoding the protein of interest with a different plasmid for the desired UnAA machinery, as shown in Figure 2. Lastly, the inclusion of an affinity purification tag C-terminal to the amber stop codon allows any truncated protein that did not incorporate the UnAA to be removed during the standard affinity purification step of the protein.

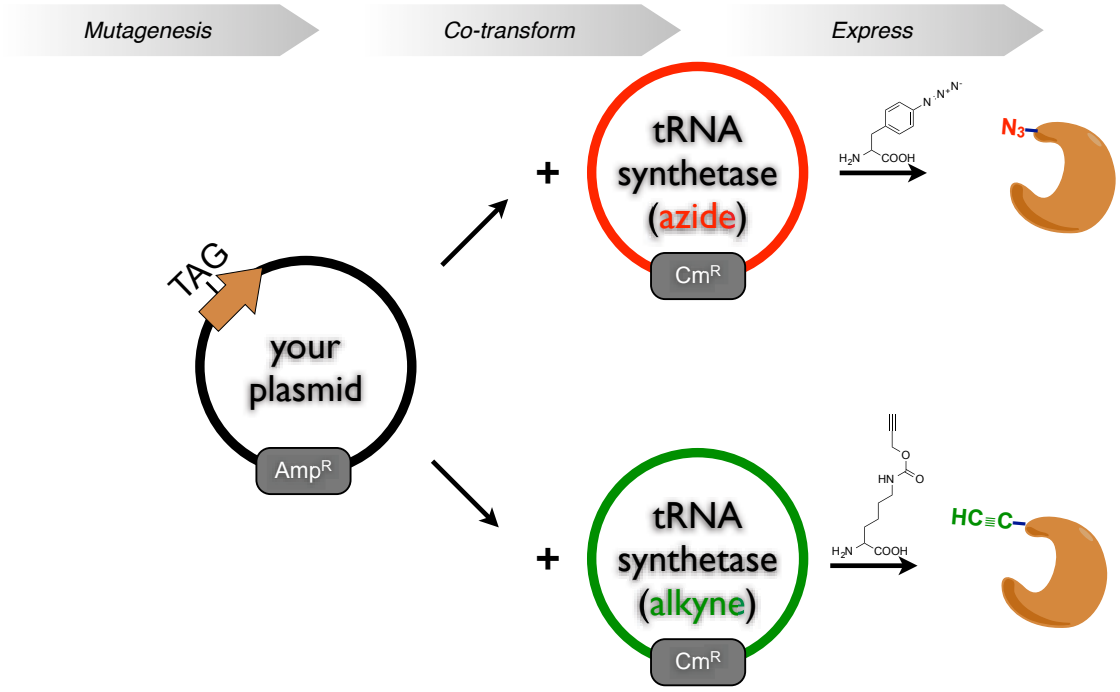


Figure 2. Incorporation of UnAAs into a protein of interest. A plasmid encoding a protein of interest ('your plasmid') is mutagenized to contain the amber stop codon (TAG) at the desired site for UnAA incorporation. *E. coli* cells are then cotransformed with a second plasmid that encodes the tRNA/synthetase pair necessary to incorporate the desired UnAA; double selection with different antibiotics encoded in each plasmid (e.g. Amp and Cm) ensures all *E. coli* contain both plasmids. Proteins are expressed in media that has been supplemented with the desired UnAA, producing a protein of interest carrying a UnAA at a specific site. Different UnAAs (e.g. an azide or an alkyne) can be incorporated into the same protein construct by co-transforming with a different plasmid encoding the UnAA machinery.

The exact arrangement, copy number, and promoter control of the tRNA and synthetase genes have been further optimised by the Schultz lab[21,22] and other groups[23,24], as well as the exact media composition and protein expression conditions[24,25].

Despite these optimisations, the overall process for UnAA incorporation has remained the same and it has found wide applicability in functional and structural protein studies[7] – including those that have employed single-molecule fluorescence techniques[26-28] –

and has been expanded beyond *E. coli* to eukaryotes, from yeast[29] to higher organisms including *C. elegans*[30], *Drosophila melanogaster*[31], and mammalian cells in culture[32].

Orthogonal Labelling Chemistries

The ability to incorporate unnatural amino acids into proteins has opened the possibility of introducing reactive groups into recombinant proteins that are not found in naturally occurring amino acids, thus allowing protein labelling that is completely orthogonal to reactions with natural amino acid side chains. A number of groups have exploited UnAA-based approaches to incorporate fluorescent dyes by a number of methods, including oxime ligation, Staudinger ligation, and click chemistry[26,33,34]. Each method requires a matched UnAA and dye reactive group to achieve coupling, and has distinct advantages and disadvantages, as discussed below and summarised in Table 1.

Oxime Ligation. Ketone functional groups do not naturally occur within proteins and can be incorporated as an unnatural amino acid (4-acetylphenylalanine, AcF) during *in vivo* expression in *E. coli*[35]. Reaction of the ketone-containing protein with hydrazide, or more optimally, hydroxylamine-functionalised dye leads to a stable oxime linkage between the two[26]. Both AcF and the hydroxylamine/hydrazide derivatives of the cyanine, Alexa, and Atto series dyes are commercially available. While this reaction is specific, oxime ligation is ideally performed at low pH and higher temperatures. Even at pH 4 and 65º, the reaction gives poor labelling yields, particularly for proteins larger than 100 amino acids[8]. These labelling conditions are not well-tolerated by most proteins, and therefore the applicability of oxime ligation is limited.

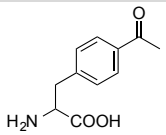
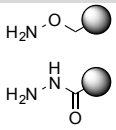
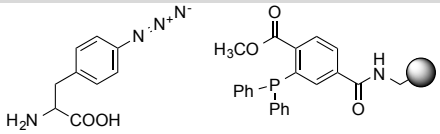
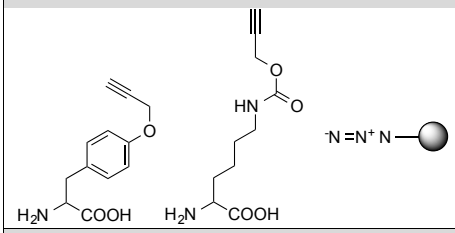
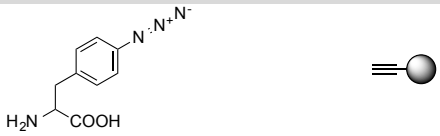
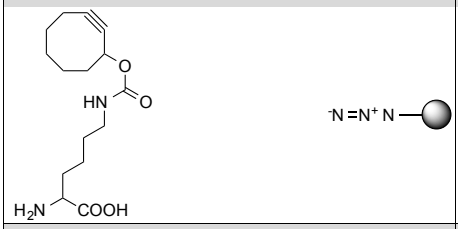
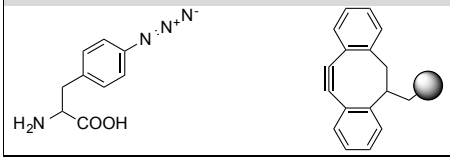
Reaction type	UnAA Reactive Group	Dye Reactive Group	Conditions	Commercial Availability
Oxime Ligation	**Ketone**	**Hydroxylamine or Hydrazide**	✖ low pH (4) ✖ high temp. (65º) Slow, refolding req[d]	✔ UnAA: yes ✔ Dyes: cyanine, Alexa, and Atto series
				
Staudinger Ligation	**Azide**	**Phosphine**	High concentrations Med. Temp. (37º)	✔ UnAA: yes ✖ Dyes: DyLight only
				
Cu-assisted Click	**Alkyne**	**Azide**	✖ Cu-catalyst ✖ TBTA, Reducing conditions, Oxygen radical generation	✖ UnAA: no, but easy synthesis ✔ Dyes: cyanine, Alexa, and Atto series
				
	Azide	**Alkyne**	✖ *–as above–*	✔ UnAA: yes ✔ Dyes: cyanine, Alexa, and Atto series
				
Cu-free Click	**Strained Alkyne**	**Azide**	✔ No Cu-catalyst ✖ UnAA not efficiently incorporated ✖ Slow [3+2] cyclo-addition (no e^- withdrawing gps.)	✖ UnAA: no ✔ Dyes: cyanine, Alexa, and Atto series
				
	Azide	**Enhanced-Strained Alkyne**	✔ No Cu-catalyst	✔ UnAA: yes ✔ Dyes: Cy3/Cy5, Alexa series
				

Table 1. Evaluation of UnAA strategies for site-specific incorporation of fluorescent dyes for single-molecule experiments. Incorporating an azide into a protein of interest and reacting a strained-alkyne derivative of a fluorescent dye avoids harsh reaction conditions of other methods. Both the UnAA and dye derivatives are commercially available.

Staudinger Ligation. Azide functional groups can be reacted with high specificity with triarylphosphine-functionalised dyes, as first demonstrated by the Bertozzi group, who labelled azide-functionalised glycans on mammalian cell surfaces in culture[33]. They extended their approach to protein labelling by modifying azidohomoalanine-containing proteins that had been produced using the methionine depletion approach for UnAA incorporation, described above[17]. Meanwhile, the Schultz group demonstrated that 4-azidophenylalanine (AzPhe) could be incorporated by amber suppression[19], which they envisioned as a tool for protein crosslinking studies. To date, only DyLight dyes are commercially available as phosphine derivatives; DyLight has not been characterised for single molecule fluorescence experiments. While phosphine derivatives of the highly fluorescent dyes typically used in single molecule experiments[8] are not commercially available, their synthesis has been described in the literature[36] and they have been employed to label single subunits of *E. coli* RNA polymerase with >90% yield and specificity[28]. Aside from the lack of commercial availability, the applicability of Staudinger ligation is somewhat limited by oxidative degradation of the phosphine group and its slow reaction rate as compared to other azide coupling reactions[37], necessitating that reactions are typically performed at 37ºC for more than 12 hours[28,36].

CuAAC (Cu-catalyzed Azide-Alkyne Cycloaddition). Since Sharpless highlighted azide-alkyne Huisgen cycloaddition as an ideal 'click chemistry' coupling reaction[38] and demonstrated azide-alkyne cycloaddition could be performed in aqueous solutions at room temperature with a copper(I) catalyst[39], 'click' chemistry itself has become synonymous with this reaction (Table 1), which yields an exceptionally-stable triazole linkage. For orthogonal labelling of biomolecules where targets are typically at low concentrations in aqueous solutions, CuAAC reactions are catalysed by the addition of

Cu(II), typically as $CuSO_4$, a reducing agent to convert Cu(II) to Cu(I), such as sodium ascorbate, and a ligand to stabilise Cu(I), typically tris(benzyltriazolyl)methyl amine (TBTA)[34].

Both azide- and alkyne-containing UnAAs can be incorporated into proteins by amber suppression: in addition to the previously-noted azidophenylaline[19], alkyne derivatives of phenylalanine[40] and lysine[27] have been used to label proteins for single-molecule fluorescence studies with commercially-available azide- and alkyne-functionalised Alexa Fluor dyes[27]. In each case, optimisation of reducing conditions and copper concentrations for the given construct was required[27]. The conditions of the Cu-catalysed reaction are not ideal for protein labelling, however[41,42]: excessive reducing conditions may disrupt protein structure, the copper(I)-stabilising ligand TBTA may cross-react with proteins of interest, and the copper itself is cytotoxic and may also induce protein aggregation *in vitro*.

Cu-free Click. As an alternative to the slow Staudinger ligation and the bioincompatible Cu-catalised click reaction, the Bertozzi group developed strain-promoted azide-alkyne cycloaddition, in which the reaction is driven by the opening of a strained cyclooctyne ring instead of a copper catalyst[41]. Subsequently, the Lemke group successfully engineered a tRNA/synthetase pair to incorporate a cyclooctyne-functionalised UnAA into *E. coli* proteins[43].

The reaction of a simple strained cyclooctyne with an azide, however, is relatively slow and labelled protein yields when using the UnAA approached are further reduced by the difficulty with which the synthetase is able to charge the tRNA with the bulky cyclooctyne group. Strain-promoted click reactions can be substantially improved by the addition of strong electron withdrawing groups[44] or by further enhance ring strain through the addition of benzyl groups[45] and a partially-endogenous amide[46,47]. Both

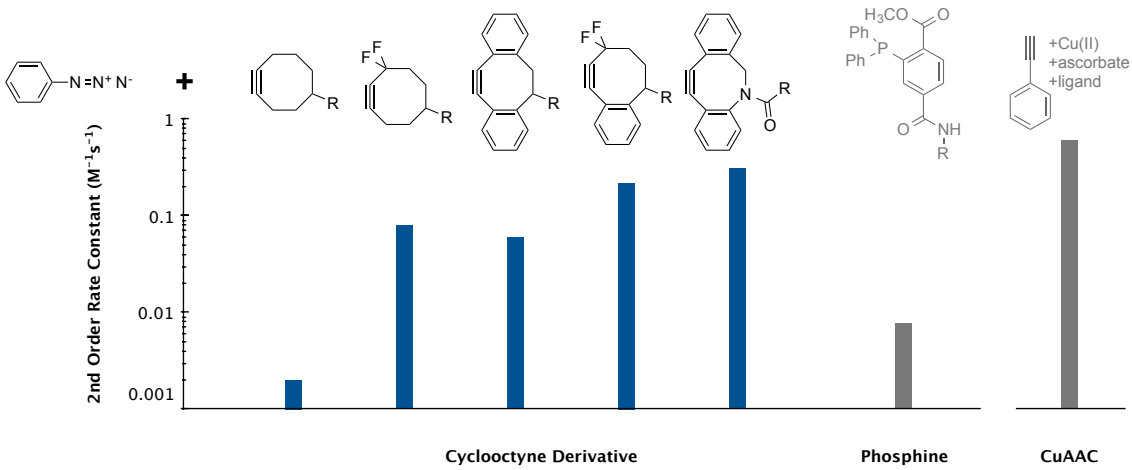


Figure 3. Increasing the reaction rate of strained cyclooctynes. Chemical modification of cyclooctyne derivatives (blue bars) with electron-withdrawing or strain-enhancing groups substantially increases their rate azide coupling. The second order rate constants of difluro- and dibenzyl-cyclooctynes are nearly 2 orders of magnitude greater than unmodified cyclooctynes, and an order of magnitude greater than the fastest reported Staudinger ligation (Phosphine, grey bar). Incorporation of an amide bond with the nitrogen forming part of the 8-membered cyclic alkyne further increases the rate, nearly approaching first order rates achieved with CuAAC at conditions for bioorthogonal labelling (CuAAC, grey bar). (Compiled using data from Sletten *et al.*, 2010; Sletten *et al.*, 2011; Soellner *et al.*, 2006; and Presolski *et al.*, 2010.)

approaches are capable of increasing the reaction rate considerably[48,49], as illustrated in Figure 3, which was compiled using data from the literature[48-51].

While the addition of these substituent groups gives considerably high yields for azide coupling, these bulky and electron-dense groups preclude incorporation as part of an unnatural amino acid by a tRNA synthetase[43]. Therefore, the benefits of Cu-free click ligation can only be realised when dyes are functionalised *in vitro* with an enhanced strained-alkyne and coupled to a protein containing an azide-containg UnAA. This approach is readily implemented in the lab: both the readily-incorporated UnAA azidophenylalanine[19] and the enhanced strained-alkyne derivatives of Cy3, Cy5 (amidodibenzylcyclooctyne) and the Alexa Fluor dye series (dibenzylcyclooctyne) are

commercially available. We aimed to apply this strategy to site-specifically label *E. coli* RNA polymerase, introduced below.

E. coli RNA Polymerase – a model nucleic acid processing-enzyme

According to the central dogma of molecular biology, RNA polymerase (RNAP) stands at the head of cellular life processes as the first step in converting genetic information into protein and eventual actions (and reactions) for living organisms. In *E. coli*[52,53], the RNAP enzyme consists of four distinct subunits: α, β, β', and ω which assemble to form core RNAP as $\alpha_2\beta\beta'\omega$. Transcription initiation factors, such as σ^{54} and σ^{70}, are recruited to form the RNAP holoenzyme (e.g. $\alpha_2\beta\beta'\omega\sigma^{70}$) which then binds to promoter DNA to begin the transcription cycle, as shown in Figure 4.

RNAP binding to promoter DNA forms a 'closed' DNA-RNAP complex in which the DNA remains double stranded. Melting of the DNA to form the transcription bubble[55,56] results in the formation of an 'open' DNA-RNAP complex that can begin to transcribe RNA. In this initial transcribing complex, RNAP remains bound to the promoter DNA[10] until it escapes and proceeds to elongation[57]. The σ subunit dissociates from the transcription complex at some point during elongation[9]. In response to termination signals (such as the formation of hairpin structures in the transcribed RNA during intrinsic termination[58]), transcription pauses, which can then lead to termination and dissociation of the RNA product and the RNAP enzyme[59].

While this scheme is presented largely as a unidirectional cycle, nearly all of the steps are reversible to a certain degree: initial transcription can repeatedly cycle to produce a collection of abortive transcripts[60,61] before proceeding to elongation and the generation of a mature RNA; pausing of transcription does not necessarily result in termination, and an elongation complex can reform[53].

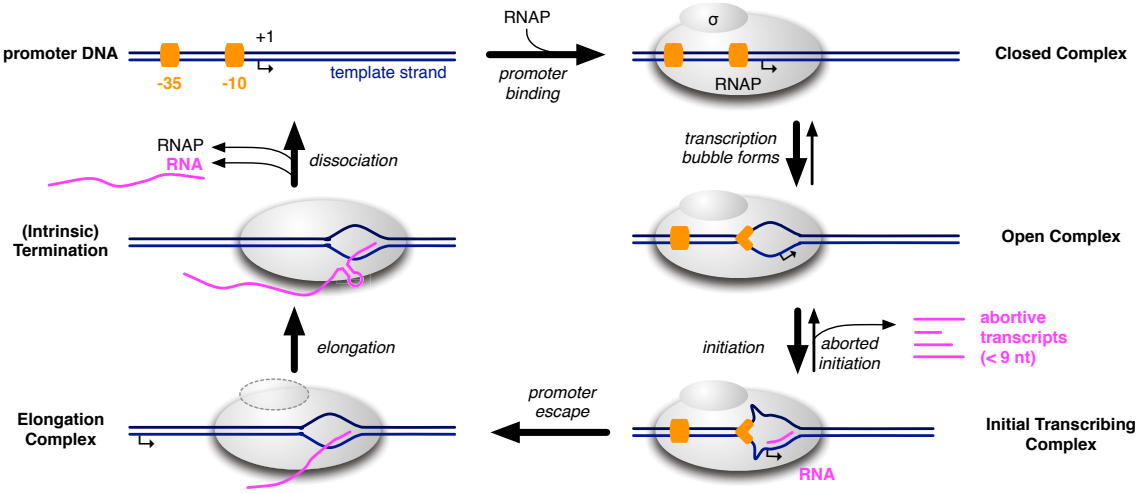


Figure 4. Transcription cycle of *E. coli* RNA Polymerase. A schematic summary of transcription in *E. coli*, adapted from: [53,54].

These events can act as regulation points for transcription: both pausing[62,63] and abortive initiation[64] mechanisms are thought to be key in regulating gene expression. Fully understanding transcription, therefore, requires an understanding of these complex pathways and the behaviour of RNAP throughout each step of the transcription cycle.

The complexity of RNAP behaviour, however, makes dissecting mechanisms of different aspects of the transcription cycle exceptionally difficult using conventional ensemble methods[12]. Single molecule techniques – notably optical tweezers – have been used to study the motions of RNAP during various steps in the transcription cycle[65-67], but fluorescence techniques, such as smFRET on TIRF that offers simultaneous observation of a large number of single molecules, provide the greatest opportunity to view rare states/events and obtain statistics regarding heterogeneous behaviours in populations of biomolecules[68].

While smFRET has been used to observe abortive initiation and the transition to elongation by labelling the 'leading edge' of the σ^{70} subunit[57], this approach has been limited by available labelling positions on RNAP (single-cysteine mutants of the α, β, and β' subunits cannot be made) and consequently required the use of multiple DNA constructs to observe transcription at various stages. Continued observation of RNAP during elongation is limited since the labelled σ^{70} subunit can dissociate from the RNAP core[9], thus leaving the transcribing RNAP without a fluorescent label.

Ideally, the entire transcription cycle could be observed continuously and monitored at near-basepair resolution using smFRET. To achieve this aim, however, new labelling positions must be considered at the 'extreme' leading edge of the polymerase to measure downstream DNA motion and polymerase/DNA translocation. The use of UnAA-based labelling methods would enable access to these new constructs, whose dye distances would allow the examination of complex behaviours of RNAP that have been previously difficult to examine using smFRET, such as promoter-proximal pausing[62,63] and translocational dynamics of RNAP during elongation.

More broadly, RNAP is an excellent model of a typical (and challenging) nucleic acid-processing enzyme for single-molecule fluorescence experiments.

- Structurally, it is complex (multisubunit with varying composition), and large (~400 kDa).
- Functionally, it must process signals from other proteins (e.g. sigma factors, elongation factors, rho) and in the DNA (promoter elements) and, consequently, transcribed RNA (hairpin).
- Dynamically, its behaviour is complex, with substantial conformational dynamics present in both the DNA[54,69] and the protein, as illustrated by crystal

structures of RNAP trapped in different states[70], and in-solution smFRET experiments[28].

Consequently, the implementation and optimisation of UnAA-based strategies for site-specific modification of *E. coli* RNAP will be useful for the study of a wide class of other enzymes.

Project Summary

Within the past year, nearly all elements of the UnAA-based labelling system have become commercially available, including the unnatural amino acids themselves, the plasmids encoding appropriate tRNA/synthetase pairs, and fluorescent dyes suitable for single-molecule fluorescence experiments that already include compatible reactive groups. I developed an optimised UnAA protein production workflow for multisubunit proteins encoded within single protein expression vectors. The workflow comprised optimised mutagenesis, validation by DNA sequencing, optimised protein expression, labelling, and purification. Using this approach, I incorporated 4-azidophenylalanine into RNAP at its extreme leading edge and labelled it using a strained alkyne-functionalised Cy5 dye; the labelled protein was characterised by SDS-PAGE and gel band mass spectrometry and assayed using bulk in-vitro transcription reactions and single-molecule FRET experiments. Lastly, a construct was designed to measure single-basepair steps of RNAP along a DNA template by smFRET from a Cy3B dye attached to downstream DNA to the Cy5 on the extreme leading edge of RNAP.

Results and Discussion

Analysis and Modelling of Labelling Positions on *E. Coli* RNA Polymerase

In order to select a site for labelling that would be close enough to downstream DNA to detect its motion at basepair resolution, I initially evaluated 4 positions at which unnatural amino acids had been introduced into *E. coli* RNAP and were previously reported to be functional after mutagenesis – two positions were on the β subunit and two on the β' subunit[71]. I then sought to compare the positions to the crystal structure of the RNAP elongation complex. While a crystal structure of a complete *E. coli* RNAP complex was not available, I used a hybrid model produced in the Kapanidis group (D. Duchi, Oxford) based on *T. thermophilus*. I aligned the *T. thermophilus* protein sequence for *rpoB* (β) and *rpoC* (β') gene products with those of *E. coli* using NCBI *blastp*[72]. To evaluate their suitability for detecting DNA motion, the 'nanopositioning system' (NPS) [73] was used to determine the accessible volume of fluorescent dyes attached by a linker to amino acid Cβ atoms, and to calculate the spatial centre of the dye-accessible volume as approximated by a single-kernel 3D Gaussian. PyMol was used to determine distances to the centre of dye-accessible-volume clouds for dyes attached at positions +15 and +20 on the DNA, as shown in Table 2.

Subunit	*E. coli* aa	*T. therm.* aa	Å from +15	Å from +20	In Clamp?
β	Glu106*	Ile101*	66.1	94.5	No
β	Asp222†	Arg209†	50.5	61.8	No
β'	Gly150	Ala140	31.6	54.3	Yes
β'	Asp284	Ala559	56.6	86.0	Yes
β'	Ser1183	Ser1300	43.8	16.7	No

Table 2. Potential Labelling Positions for Measuring Downstream DNA Motion in RNAP. Amino acid residues of *E. coli* RNAP that had been previously reported to be successful when mutagenized to a UnAA were evaluated by molecular modelling using a RNAP-DNA elongation complex model based on *T. thermophilus* RNAP. The homologous amino acid for *T. thermophilus* was determined by protein sequence alignment and the distance from that position to the +15 and +20 positions on the downstream DNA during elongation was determined using molecular modelling. The clamp domain of RNAP is mobile, and the position of each aa was considered.

* *E. coli* Glu106 has no direct homologue on *T. thermophilus*. The nearest aa, Ile101, is actually homologous to V103. Glu106 is in a 7aa range that is not present in *T. thermophilus*; the crystal structure is also missing the next 8aa, suggesting a disordered region.

† *E. coli* Asp222 is two aa beyond a region of *blastp* alignment (last aligned is Ile220=Leu207). In *T. thermophilus.* Arg209 is at the end of an α-helix, but would likely be out of the helix in *E. coli* due its isoleucine sidechain at codon 220.

While position β'150 had distances to downstream DNA within the linear range of common FRET dye pairs for single molecule measurements[74], it is also located on the clamp of RNAP, which exhibits considerable intra-protein motion[28] and would thus not provide a stable reference point on the protein for protein-DNA distance measurements. An alternate leading edge position outside of the mobile region, β'1183, was then considered since RNAP was reported to be tolerant of UnAA-labelling at this position (R. Ebright, personal communication to A. Kapanidis). Molecular modelling revealed the position was ideal and extended so far along the downstream DNA that it extended

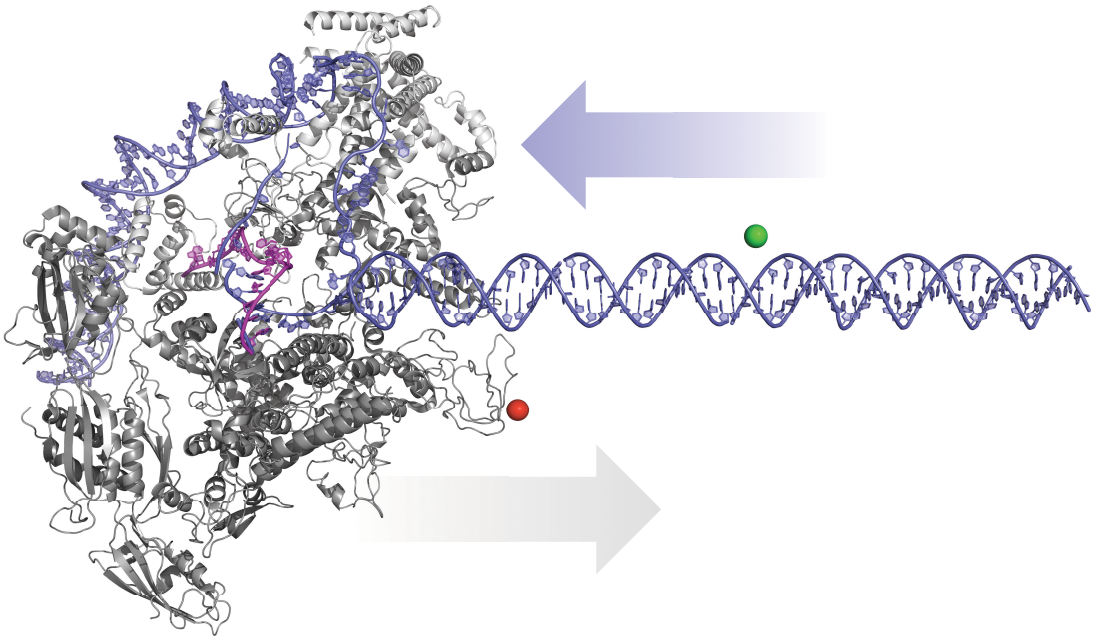


Figure 5. UnAA Labeling of RNAP Measures Downstream DNA distances. The centroid position of dye accessible volumes for Cy5 (red ●) on β'1183 and Cy3B (green ●) on downstream DNA at +36 was obtained using molecular modelling. Before transcription, the dyes are separated by 79.5 Å. Transcription of the DNA by RNAP to produce RNA (pink) results in a decrease in the protein-downstream DNA distance as a result of relative motion of the protein (grey arrow) and downstream DNA (blue arrow). (PDB file of hybrid model, discussed in Materials and Methods, D. Duchi & D. Bauer)

beyond approximately 16 bp of downstream DNA. This position at the 'extreme' leading edge of RNAP will enable the maximum possible distances to downstream to be measured using FRET (Figure 5).

Based on the results from molecular modelling, I selected the β'1183 position for mutagenesis with the aim of incorporating an unnatural amino acid for site-specific labelling.

Single-Step Mutagenesis of a Long Plasmid (15 kb) Encoding Multiple Genes

Conventional site-directed mutagenesis involving large plasmids or genes relies upon subcloning smaller portions for mutagenesis and then re-cloning the mutagenized portion back into the larger vector. I sought to avoid these step and instead optimise mutagenesis conditions for long templates and validate the resultant plasmids using DNA sequencing.

A single 15 kbp plasmid, pVS10, contains 4 genes (rpoA, rpoB, rpoC, rpoZ) under the control of a T7 promoter[75]. These genes that encode the four subunits of *E. coli* RNAP (α, β, β', and ω) with a C-terminal 6xHis tag appended to rpoC (β') for purification of all four RNAP subunits by Ni-NTA affinity chromatography under native conditions, as shown in Figure 6.

In order to incorporate an unnatural amino acid into position β'1183, the serine codon TCG would have to be replaced by the amber stop codon, TAG. Since the affinity purification tag (6xHis) is located at the C-terminus of the same subunit, any truncated

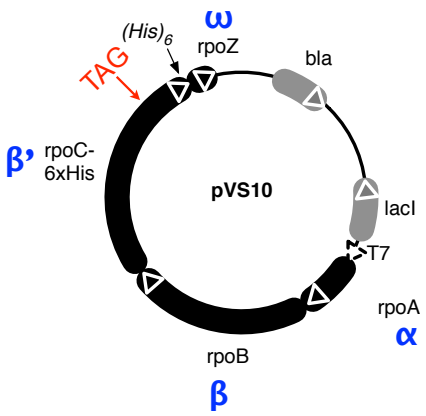


Figure 6. pVS10 plasmid (15 kbp) encodes all 4 E. Coli RNAP subunits on genes rpoA, rpoB, rpoC, and rpoZ. Purification of core RNAP enzyme is accomplished via a 6xHis tag fused to the C-terminus of rpoC. Incorporation of unnatural amino acids requires mutagenesis of codon 1183 of the rpoC gene to TAG.

protein produced as a result of a failure to incorporate the unnatural amino acid will be removed without additional purification steps.

While mutagenesis of long plasmids is challenging[76], I sought to optimise yields by exploiting high-fidelity *PfuUltra* DNA polymerase and adding DMSO to improve processivity[77]. I then explored the parameter space of the mutagenesis and transformation protocol: I varied (1) the primer:template ratio by increasing the amount of template plasmid in the mutagenesis reaction, (2) the amount of DNA electroporated into competent cells, and (3) the amount of culture plated onto antibiotic plates for selection. The colonies were counted after 16-hour incubation at 37º.

PCR Reaction	ng pVS10	ng each Primer	purified by	ng DNA recovered	ng DNA electroporated	frac. Of culture plated	colonies after 16h	
A	5	120	EtOH precip	4130	83	0.1	0	
						0.9	0	
					826	0.1	0	
						0.9	0	
B	10	120	EtOH precip	4995	100	0.1	0	
						0.9	0	
					999	0.1	0	
						0.9	0	
C	31	120	QiaPrep	170	3	0.1	0	
						0.9	0	
					3	0.1	0	
						0.9	0	
D	46	120	EtOH precip	4230	85	0.1	0	
						0.9	0	
					846	arc	–	
							–	
E	91.5	120	EtOH precip	5225	105	0.1	0	
						0.9	24	*> picked 5*
					1045	0.1	1	*> picked 1*
						0.9	11	*> picked 8*

Table 3. Results of site-directed mutagenesis of pVS10 plasmid. In five mutagenesis reactions, **A-E**, the amount of template was varied, along with the amount of DNA electroporated into competent cells and the fraction of the resultant culture that was plated for selection. Colonies were counted after incubation at 37º for 16 hours.

A total of 14 colonies were picked from the three successful transformations and were used to inoculate overnight cultures for plasmid minipreps. Successful mutagenesis of codon β'1183 to TAG would also introduce an additional cut site for the BamHI restriction enzyme; restriction digest of the putative pVS10:β'1183tag plasmid followed by agarose gel electrophoresis could therefore confirm mutagenesis, as illustrated by the *in silico* digest, shown in Figure 7.

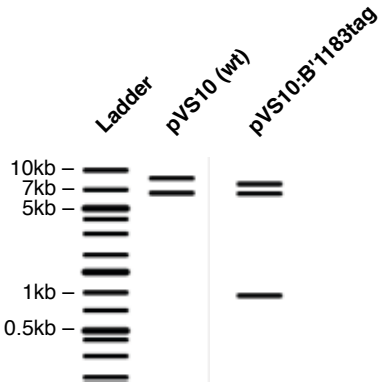


Figure 7. *In Silico* digest with BamHI of pVS10 (wt) plasmid and pVS10 plasmid following mutagenesis of codon β'1183 to TAG. The addition of a BamHI cut site results in a decrease in the size of the largest DNA fragment (8.6 to 7.7 kbp) and the appearance of a new fragment at 0.9 kbp.

Plasmid DNA isolated from each of the 14 colonies was then digested with BamHI restriction enzyme. Of the 14 colonies, 11 showed a downshift of the 8.6 kbp band to 7.7 kbps and the characteristic appearance of a new band at 0.9 kbp, suggesting they contained the TAG mutation (Figure 8).

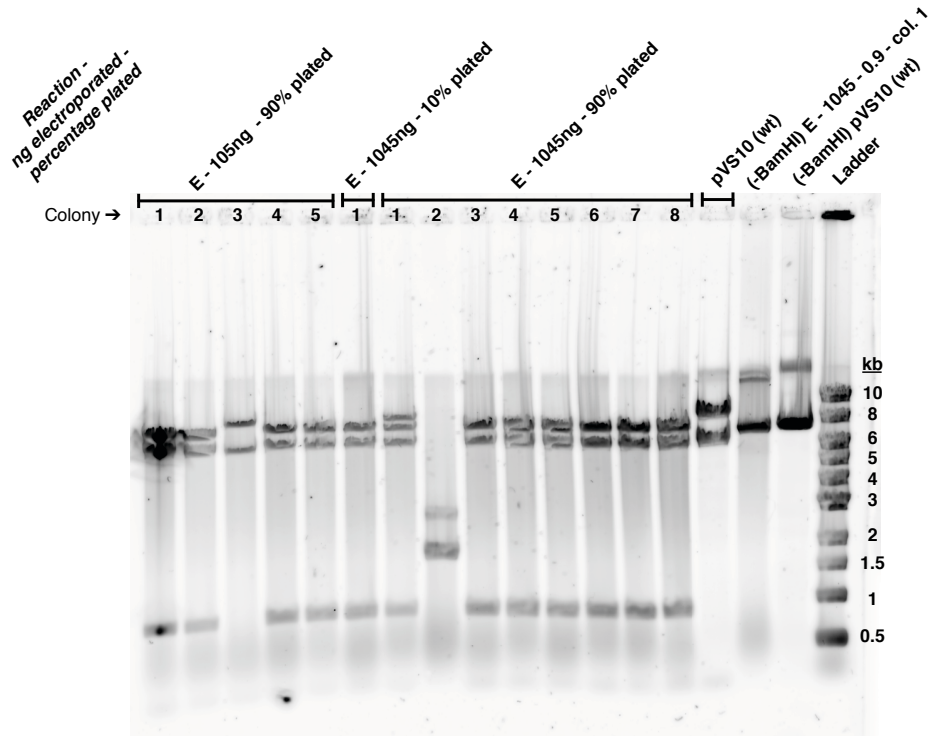


Figure 8. BamHI digest of mutagenised pVS10 plasmids. Colonies from the three successful transformation reactions (see **Table 3**) were miniprepped and plasmid DNAs were digested with BamHI and compared to a wt pVS10 plasmid digest by agarose gel electrophoresis. A total of 11/14 colonies were positive for the introduction of an addition TAG site, suggesting successful mutagenesis. (1 % agarose-TAE, 640ng plasmid loaded on each lane, 40min staining with SYBR Gold at 1:10,000.)

In order to confirm that the mutagenesis was successful and that I did not introduce any unintended mutations, I designed tiling primers for DNA sequencing the entire coding region of the pVS10 plasmid. I then selected the plasmid from a single colony (E-1045ng-90%-col.4) and sent it for Sanger DNA sequencing. The overlapping primers covered all of the four coding genes on the plasmid.

Analysis of the sequence data confirmed: that QuickChange mutagenesis was successful for substitution of TAG at codon β'1183; that mutagenesis did not introduce any other mutations in the coding region of pVS10 plasmid; and that 4 silent mutations

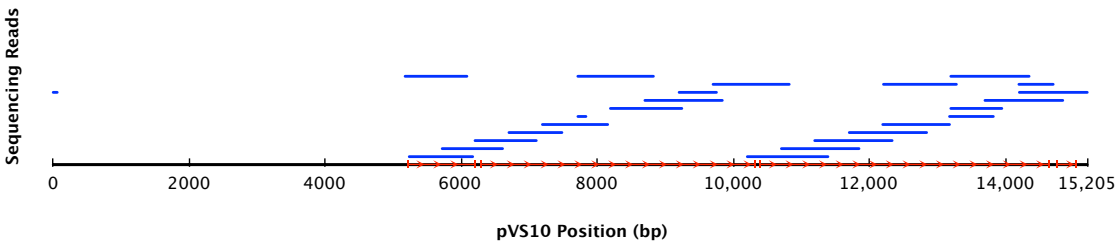


Figure 9. Sequencing of pVS10. Tiling primers were used to produce overlapping sequencing reads (blue) that fully covered the coding region of pVS10 (red).

were also present. I resequenced portions of the pVS10 (wt) plasmid and confirmed that these silent mutations were originally present.

Generation of *E. Coli* Strains for Protein Production

As discussed previously, a plasmid containing the protein of interest (pVS10:β'1183tag) must be co-transformed into a suitable *E. coli* strain with a second plasmid containing the necessary machinery (tRNA/synthetase) for incorporating a particular unnatural amino acid in order to produce a final strain for UnAA protein production. A number of considerations must be taken into account in selecting each component.

Plasmid containing the protein of interest. Like many commonly-used plasmids for protein expression, pVS10 is a pET derivative and – like the majority of canonical protein expression vectors – carries genes for protein expression under the control of a T7 promoter and also confers ampicillin resistance[75]. In developing a general method for UnAA-incorporation into proteins, I aimed to allow the use of existing in-house plasmids with these 'typical' characteristics and wished to avoid subcloning, if possible. This desire is in conflict with the need to minimise any 'leaky' expression – even in the cases of nontoxic protein –in order to increase yield during UnAA

incorporation[23]. For UnAA application, therefore, genes for proteins-of-interest are typically subcloned into pBAD vectors[25] in which they are under the control of the arabinose promoter[78].

Suitable Host Strain. To create a more general route for UnAA incorporation into proteins of interest and avoid subcloning steps, I selected BL21-AI cells, which contain the T7 RNA polymerase gene under the control of the araBAD promoter. Consequently, transcriptions of genes under the control of the T7 promoter are effectively repressed until induction with arabinose.

Plasmid encoding tRNA/synthetase pair. A number of plasmids containing a tRNA/synthetase pair for the incorporation of 4-azidophenylalanine are described in the literature, notably derivatives of pEVOL (Cm^R) from the Schultz group[22] and derivatives of pDULE (Tet^R) and pDULE2 (Sm^R) from the Mehl group[23,24]. The pDULE pCNF RS and pDULE2 pCNF RS plasmids are unique in that their tRNA/synthetase pair is capable of incorporating both alkyne- and azide-functionalised phenylalanine. Strains carrying one of these plasmids will therefore be able to incorporate one or the other by supplementing media prior to induction with the desired UnAA – without the need to generate a new strain[23]. In order to select for co-transformed cells that contain both plasmids, each must encode a different antibiotic resistance marker that is distinct from any resistance also carried by the host strain. A

Plasmid	Resistance
pET	Amp
pUC19	Amp
pBR322	Amp + Tet
pBAD	Tet

Host	Resistance
BL21	–
BL21(DE3)	–
BL21-AI	Tet
DH10B	Sm
TOP10	Sm
DH5α	–

UAA Plasmid	Resistance
pEVOL	Cm
pDULE	Tet
pDULE2	Sm

Table 4. Summary of antibiotic resistance in selected protein expression plasmids, *E. coli* strains, and plasmids encoding UnAA machinery.

brief list of antibiotic resistance for common protein expression plasmids, host strains, and UnAA plasmids is presented in Table 4. Since my chosen host strain, BL21-AI, is resistant to tetracycline, I chose pDULE2 pCNF RS as the plasmid to encode the UnAA-incorporating machinery. I then generated two strains by co-transforming chemically competent BL21-AI cells:

- pVS10:β'1183tag and pDULE2 pCNF RS
- pVS10 (wt) and pDULE2 pCNF RS

Following double antibiotic selection on LB-agar plates, I picked a single colony, and grew the culture overnight to generate glycerol stocks of the *E. coli* strain. I confirmed successful co-transformation by agarose gel electrophoresis of miniprepped plasmid DNA. Bands corresponding to both pVS10 and pDULE2 derivatives were observed, as shown in Figure 10C.

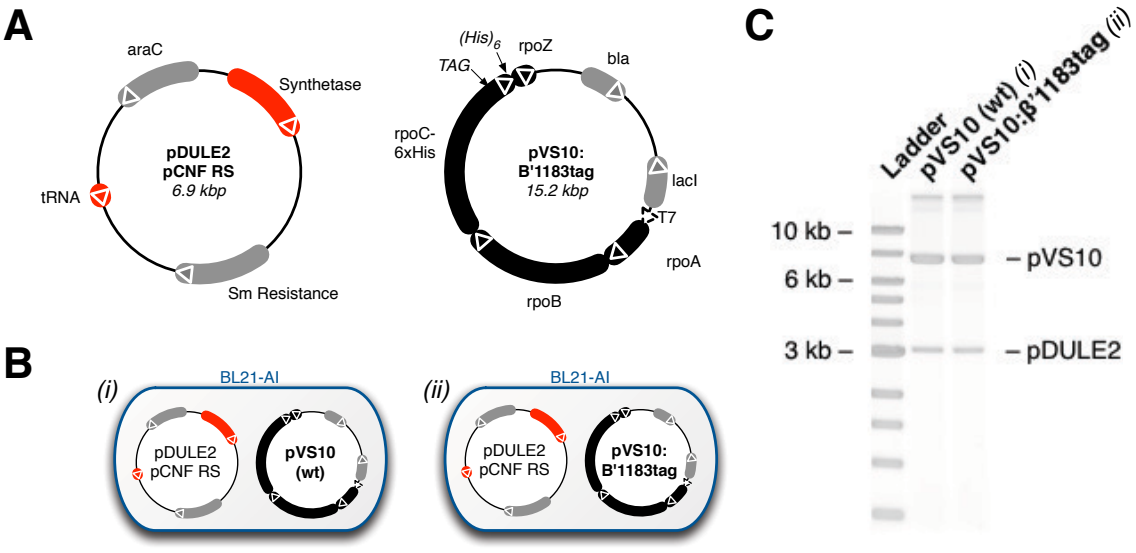


Figure 10. Plasmids, strains, and cotransformation. (A) Maps of plasmids encoding UnAA machinery and protein of interest. **(B)** BL21-AI cells are co-transformed with plasmids. **(C)** Gel electrophoresis confirms successful cotransformation.

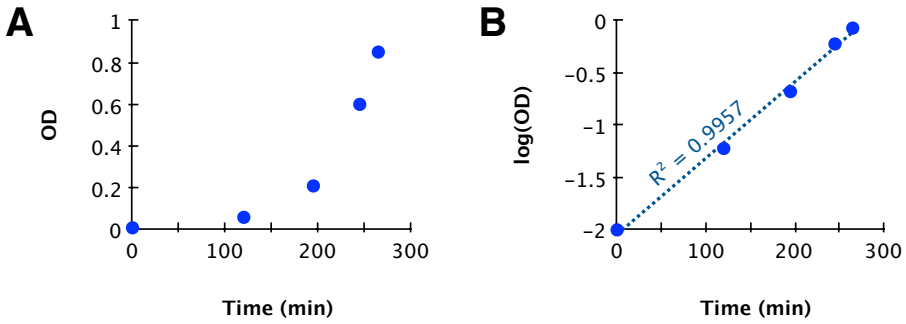


Figure 11. Growth curves of co-transformed *E. coli* are normal. Despite the presence of plasmids encoding unnatural amino acid machinery and RNAP, *E. coli* cultures show typical logarithmic growth curves, **(A)**. Logarithmic fit to data, **(B)**, gives an R^2 value of 0.9957.

The Schultz group uses pre-induction growth curves of the co-transformed *E. coli* strain as an indicator for full repression of the UnAA tRNA and synthetase[22]; growth curves of the *E. coli* BL-21 AI strain containing pVS10:β'1183tag and pDULE2 pCNF RS suggest that the plasmids do not interfere with growth (Figure 11).

Protein Expression, Labelling, and Purification

The optimum media composition for UnAA-containing protein expression remains unclear and the use of standard media such as LB[21], 2xYT[21], and Terrific Broth[43] is commonplace. However, it has been clearly shown[23,24] that the creation of custom culture media (laboriously, from yeast extract, individual amino acids, and individual trace metals) and the use of autoinduction[79] greatly enhances protein yield. Autoinduction relies upon the preferential metabolism of glucose by *E. coli*, which results in the suppression of proteins under control of the lac operon until the glucose is consumed and metabolism of lactose in the culture media begins. The strains I generated add additional complexity: the pDULE2 plasmid and host T7 RNAP is under

the control of an arabinose-inducible promoter, but the RNAP-encoding genes within pVS10 are still under the control of the lac operon. Simultaneous induction by lactose and arabinose is therefore required.

Instead of producing custom media, I modified a commercially-available autoinduction media designed for use with IPTG-inducible plasmids (OvernightXPRESS Instant Media, Novagen). I supplemented the OvernightXPRESS media with arabinose to concentrations suggested in the literature[24] for custom autoinduction media used with pDULE2, under the assumption[*] that induction triggered by consumption of glucose in media with available lactose *or* arabinose would also work with a mixture of lactose *and* arabinose[79]. To avoid any hindrance to initial *E. coli* growth, I opted not to supplement the media with any UnAA initially[25].

I carried out protein expression from my BL21-AI strain harbouring the pVS10:β'1183tag plasmid in this modified autoinduction media, supplementing with azidophenylalanine 40 minutes after initial inoculation; cells were harvested after 18 hours. Full details are given in the Materials & Methods section of the thesis. Standard protein purification for His-tagged RNAP[75] was integrated with Cy5 labelling, which was carried out at room temperature for 1 hour using a strained alkyne derivative of Cy5 following initial purification on a NiNTA column as illustrated in Figure 12.

Both an unlabelled aliquot of protein and the Cy5-labelled aliquot were then purified (separately) on heparin-sepharose, eluting with a salt gradient while monitoring absorbance at 280nm (for protein) and at 650nm (for Cy5).

[*] This assumption is not totally naïve since 'direct' co-induction with IPTG and arabinose is routine.

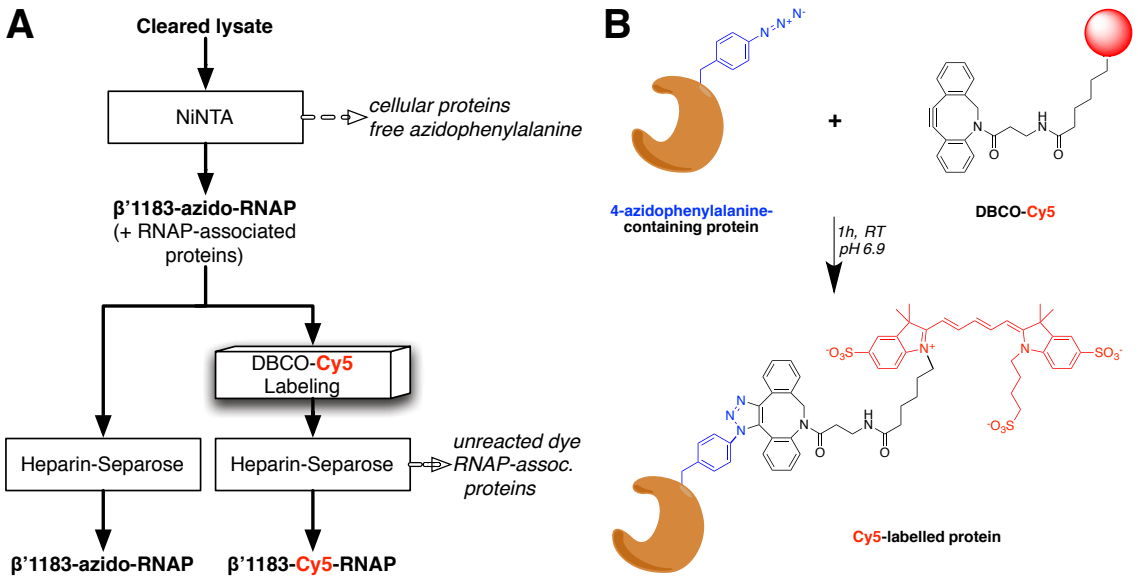


Figure 12. Integrated purification and labelling of azide-containing RNAP with DBCO-Cy5. Purification, **(A)**, is carried out using standard methods for His-tagged RNAP: crude purification carried out on a NiNTA column to remove most cellular proteins, and – crucially – any free azidophenylalanine. The crude protein (β'1183-azido-RNAP) is split into two aliquots, and prior to separate purification on heparin-sepharose, one aliquot is labelled, **(B)**, with DBCO-Cy5.

Characterisation of Protein Labelling

I first considered the chromatogram from the heparin-sepharose purification of the unlabelled protein (Figure 13A): two main A_{280} peaks were observed and these fractions were analysed by SDS-PAGE (Figure 13B). The fraction corresponding to the second peak clearly contained RNAP, with bands for β',β, and α clearly visible. I then considered the chromatogram from the Cy5-labelled fraction (Figure 13C): again, two main A_{280} peaks were observed, however the second peak coincided with a peak on the A_{650} trace (Cy5 absorbance) as well – suggesting the RNAP was also labelled. I analysed four fractions from the peak by SDS-PAGE (Figure 13D), and – prior to silver staining – I imaged the gel for Cy5 fluorescence (Figure 13E). A single, main

fluorescent band was observed, corresponding to the β' subunit. Quantitation of the fluorescence intensity (Figure 13F) suggested a labelling specificity of 87% for the two

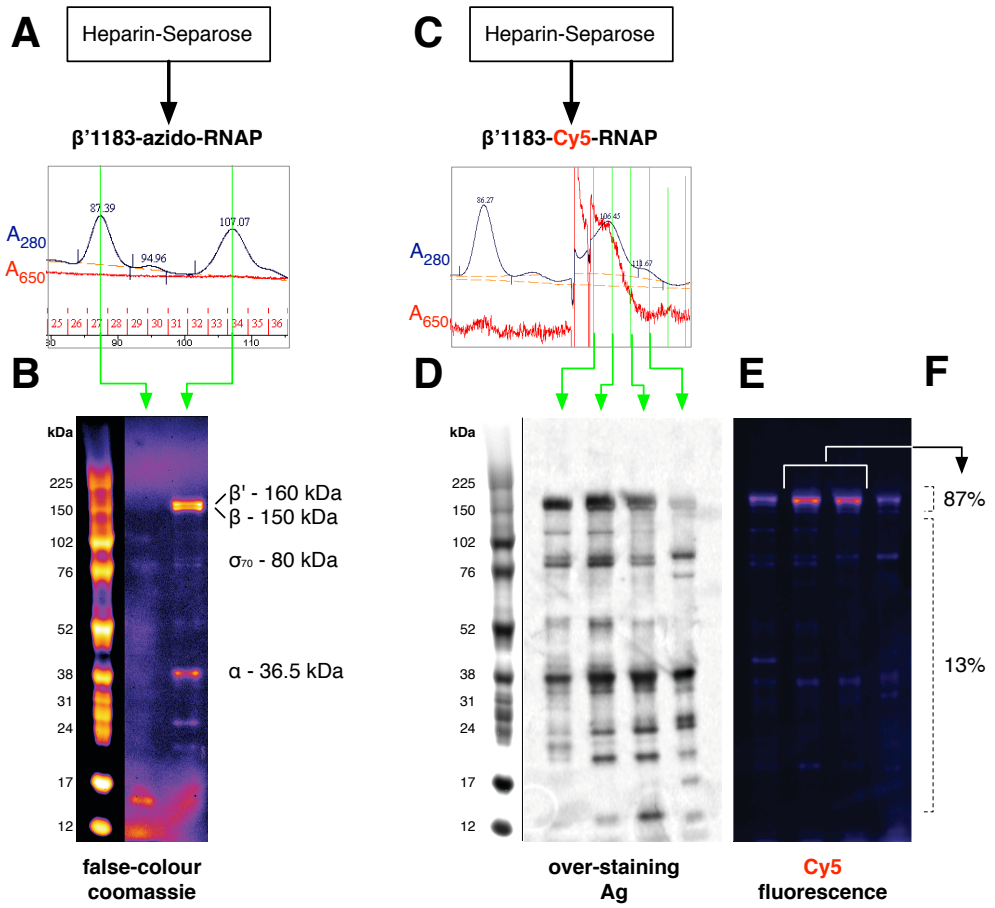


Figure 13. Characterisation of DBCO-Cy5 labelling of β'1183-azido-RNAP by SDS-PAGE. Chromatogram **(A)** from heparin-sepharose purification of unlabelled β'1183-azido-RNAP showed 2 main peaks, the second of which contained RNAP when analysed **(B)** by SDS PAGE. The chromatogram from the labelled aliquot **(C)** showed a coincident increase at Cy5 absorbance (A_{650}) at the second RNAP peak. Over-staining fractions with silver stain **(D)** showed a number of contaminating proteins, later identified by mass spectrometry as the ribosomal subunits, catabolite activator protein (CAP), and lysozyme. Prior to silver staining, the gel was imaged for Cy5 fluorescence **(E)**, which showed the β' band to be clearly fluorescent. Quantitation of the fluorescence **(F)** indicated a labelling specificity of 87% for the 2 main RNAP-containing fractions. Note that the ladder shown in **D** has been transposed from **B** for reference and that the images shown in **B** and **E** are false-coloured to show fainter bands more clearly.

main RNAP-containing fractions. This value is in line with the ~90% labelling specificity reported for other azide-containing RNAP subunits reported in the literature[28].

The two main fractions containing β'1183-Cy5-RNAP were dialysed into storage buffer; UV-visible absorption measurements at 280 nm and 650 nm indicated that the RNAP was approximately 50% labelled. This yield is likely a result of the reduction of some of the azidophenylalanine to the amine form by the reducing agents present during the protein purification (Materials and Methods) and could be increased by avoiding reducing agents during pre-labelling lysis and purification steps (D. Grohmann, personal communication). The 50% labelling yield, however, was sufficient enough to proceed with characterisation of the β'1183-Cy5-RNAP activity by bulk *in vitro* transcription assay and single-molecule FRET open complex formation assay.

Activity Assay: Bulk *in vitro* transcription

The initial transcribing complexes of *E. coli* RNA polymerase can abort transcription and revert to open complexes, releasing short 'abortive' RNA transcripts in the process[60]. As illustrated in Figure 14, repeated cycling between the open complex and initial transcription produces a characteristic set of abortive transcripts for a given promoter that can be visualized by initiating transcription with a radiolabelled dinucleotide, separating the products by denaturing PAGE, and exposing the gel to x-ray film[61,69]. In addition to providing a 'yes-or-no' answer as to whether β'1183-Cy5-RNAP is able to transcribe DNA, the abortive transcription assay provides information about the behaviour of the polymerase within the transcription cycle – key if we intend to use the labelled RNAP in mechanistic studies.

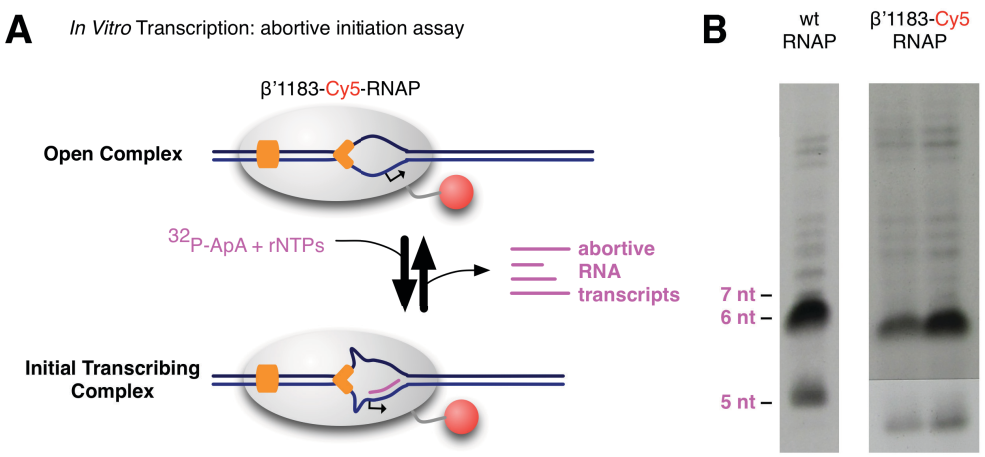


Figure 14. Abortive transcripts produced by β'1183-Cy5-RNAP are indistinguishable from those of wt RNAP. The *in vitro* transcription assay for abortive initiation **(A)** uses a ^{32}P-labelled dinucleotide to initiate transcription at a lacCons+2 promoter, which is fed by unlabelled NTPs. Abortive cycling between the open complex and the initial transcribing complex produces a characteristic pattern of abortive transcription products **(B)**. The pattern produced by the two purified fractions of β'1183-Cy5-RNAP is the same as that produced by wt RNAP, shown for reference. This result suggests that UnAA incorporation and protein labelling does not affect activity. Note that the break in the gel in **B** is the result of overlapping x-ray films used to cover the large area of the gel.

An *in vitro* transcription assay was carried out by A. Tomescu (Oxford) with both purified fractions of β'1183-Cy5-RNAP. The pattern and quantity of the products (Figure 14B) is from the labelled protein is characteristic of normally-functioning RNAP, suggesting that the labelled protein is suitable for mechanistic studies.

Activity Assay: smFRET Open Complex Formation

To examine the behaviour of β'1183-Cy5-RNAP, we then conducted a single-molecule assay to examine the dynamics of open complex formation by labelled β'1183-Cy5-RNAP as compared to the dynamics of complexes formed by wt RNAP. Single-molecule FRET experiments performed using TIRF microscopy allow the dynamics of

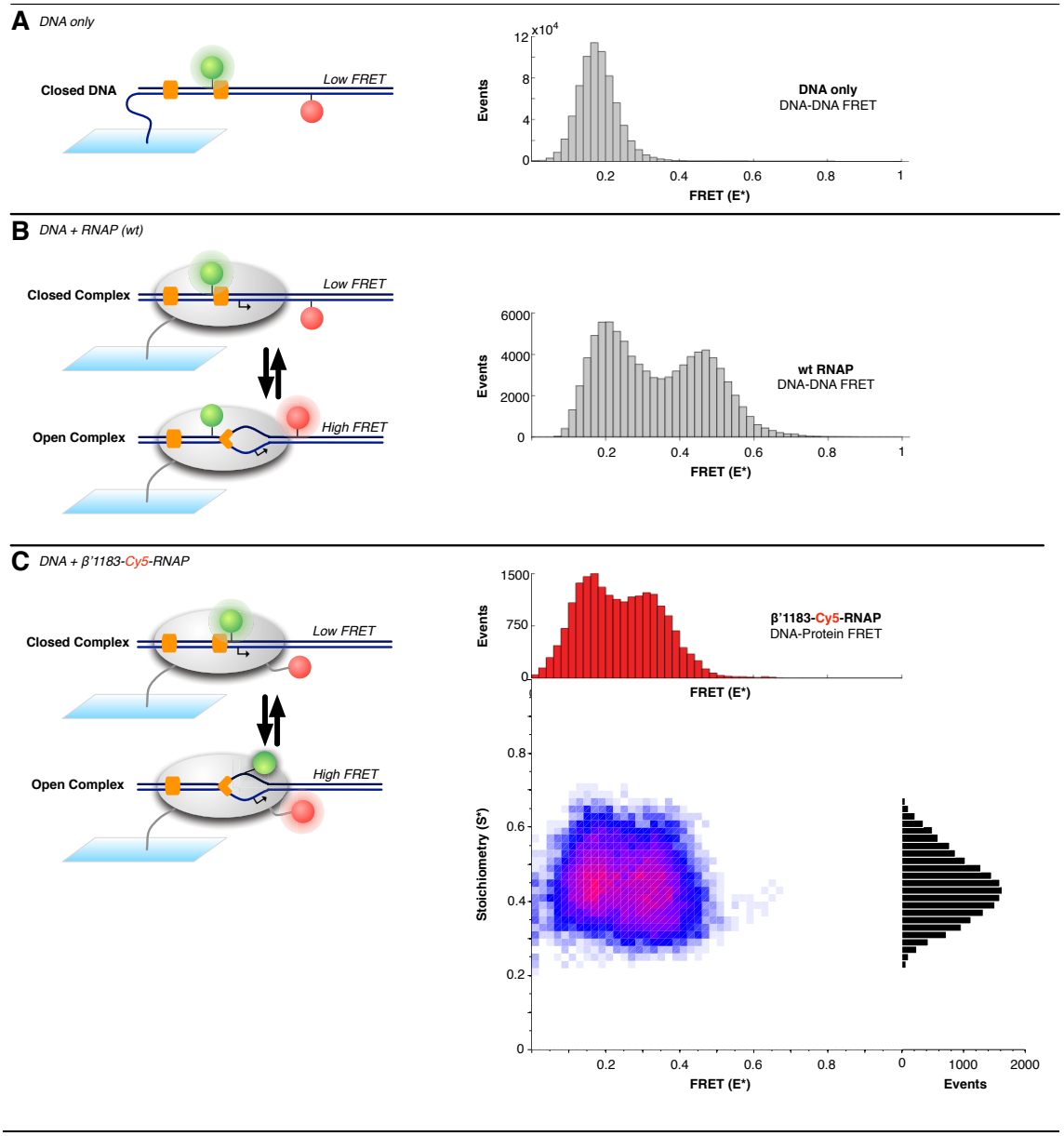


Figure 15. Labelled β'1183-Cy5-RNAP forms open transcription complexes. These complexes show the same dynamic behaviour as those formed by wt RNAP. Single molecules of double-stranded promoter DNA are immobilised on a glass surface, (**A**, left column). The DNA is doubly labelled with Cy3B and Atto647N and imaged by TIRF under ALEX to measure the degree of FRET between the two dyes. A histogram of the FRET values from the single molecules is shown (**A**, right column), showing that the DNA, in its double-stranded 'closed' form, exists in a single conformation (i.e. its double-stranded 'closed' form) with a mean FRET value of ~0.2. RNAP open complexes are formed **(B)** with the same doubly-labelled promoter DNA and the RNAP is immobilised on the surface by a C-terminal His tag on the β'

single surface-immobilised molecules to be observed over time. While open complex formation has been observed for DNA-immobilised complexes[69], it had not been observed for complexes immobilised by attaching the RNAP to the surface.

An assay for open complex formation was developed by D. Duchi (Oxford) that immobilised RNAP by capturing a C-terminal 6xHis tag on the β' subunit[80] with an anti-His antibody that had been attached to a modified glass surface[81]. While promoter DNA alone exists as a single, closed state (Figure 15A), the RNAP-bound DNA was found to exist in both the open and closed states *while still remaining bound to the RNAP* (Figure 15B). Previous experiments on DNA-immobilised complexes assigned this low-FRET peak to unbound (but surface-immobilised) DNA since they had no direct observable for RNAP binding events and consequently were unable to distinguish between 1) free, unbound DNA, and 2) RNAP-bound DNA in the closed conformation[69].

The single-molecule assay for immobilised-RNAP open complex formation was modified for the labelled β'1183-Cy5-RNAP by changing the DNA labelling scheme

---

subunit[80,81]. The complexes show both a low-FRET 'closed' state and a high-FRET 'open' state, as indicated by the two FRET peaks centred at ~0.2 and ~0.45. Open complexes are formed with labelled β'1183-Cy5-RNAP **(C)** and promoter DNA that was singly-labelled with Cy3B were also surface-immobilised. We are able to detect stable complexes of the β'1183-Cy5-RNAP and Cy3B-promoter DNA (stoichiometry = ~0.5). The complexes exhibit the same distribution between open and closed states as observed in B, with slightly different FRET values for the two states (~0.15 and ~0.3) as a result of the different labelling positions and different FRET scheme (DNA-DNA vs. DNA-RNAP). smFRET experiments were performed by D. Duchi and L. Fernandez (Oxford).

(Figure 15): instead of a doubly-labelled promoter DNA, the promoter was singly-labelled with Cy3B. Changes in FRET from this dye to the Cy5 on the extreme leading edge as a result of transcription bubble formation was used to monitor open complex formation and any dynamics that were present. Additionally, the use of alternating-laser excitation (ALEX) gave a stoichiometry measurement for each single molecule observed, which allowed us to be certain that both dyes were present in the complex – and thus that both RNAP and DNA were present, regardless of their FRET state. Using this assay, we determined that open complexes could stably be formed by β'1183-Cy5-RNAP with promoter DNA and they exhibited the same distribution between the open and closed states as had been observed with wt RNAP.

In addition to confirming previous results regarding the open-closed dynamics of RNAP-promoter complexes (D. Duchi, Oxford), these results further show that the β'1183-Cy5-RNAP is not compromised by labelling and can be used in future smFRET experiments.

Monitoring Downstream DNA Motion from Initiation to Elongation

The labelled β'1183-Cy5-RNAP allows us to consider a number of different experiments that were previously not possible due to labelling limitations. First, we now have a fixed reference point on the core of RNAP. Previous labelling of the σ^{70} subunit did not allow continued observation throughout elongation since it dissociates from the core enzyme as elongation proceeds. Secondly, the location of the label on the 'extreme' leading edge of RNA polymerase should now allow monitoring of downstream DNA motion over a considerable number of basepair steps by the polymerase, allowing a more complete picture of the transcription cycle (Figure 16A).

To set the stage for future experiments, I modelled the path of a dye as transcription progressed that was attached to the nontemplate DNA strand of the lacUV5 promoter, 36bp downstream from the transcription start site. Using the *in silico* model of RNAP elongation complex structure described above (Figure 16B), I determined the distance between the Cy5 on the RNAP at β'1183 and a Cy3B dye on the downstream DNA at each basepair-step (Figure 16C).

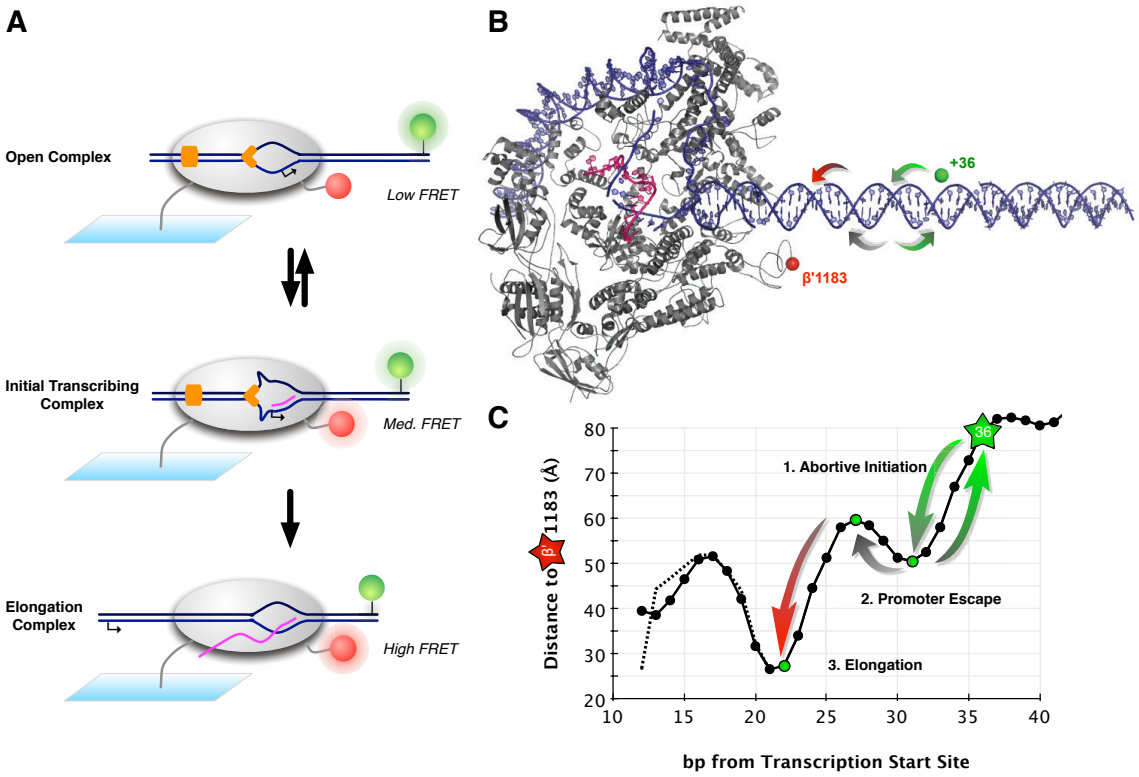


Figure 16. Monitoring downstream DNA motion using extreme-leading-edge FRET. The label of β'1183-Cy5-RNAP is ideally placed to monitor FRET from a Cy3B dye positioned on downstream DNA, schematically shown in **(A)**. As transcription proceeds, downstream DNA is drawn into the RNAP, bringing the downstream Cy3B closer to the Cy5 on the core of the RNAP and producing an increase in FRET. The actual situation **(B)** is somewhat more complex since the RNA polymerase tracks the helical pitch of the DNA[82]. Molecular modelling gives the predicted distances **(C)** between the Cy5 label on the RNAP and the Cy3B position at each bp-step as transcription advances (solid line). Accounting for the steric hindrance of the RNAP core on the dye position as it is drawn into the downstream DNA channel of the RNAP alters the predicted dye centroid positions near the transcription start site (dotted line).

The distance changes predicted by the model are not linear and reflect the fact that RNAP tracks the helical pitch of the DNA[82]. The location of the Cy3B dye at position +36 allows both abortive initiation and elongation to be observed in real-time since the distances are within the linear range of FRET between Cy3B and Cy5. Elongation steps are detected with high sensitivity (Δ_{FRET} = 0.1-0.2) since they occur when the helical track of the DNA places the dye motion along the dye-DNA axis, rather than perpendicular to it. In addition to evaluating the suitability of the downstream labelling point, the model also can be used to further design the sequence of the DNA template. For example, the template can be designed so that the polymerase can be stalled at particular points in the transcription cycle by 'starving' the RNAP of a particular NTP and inducing a pause in transcription[9]. In addition to allowing the characterisation of stalled complexes by confocal smFRET methods, such an approach would allow the polymerase to be 'stepped' through the transcription cycle while under observation using TIRF, permitting the dynamics of DNA translocation during elongation to be observed by single-molecule fluorescence for the first time.

Conclusion

The work presented in this chapter forms the basis for ongoing work to characterise the behaviour of *E. coli* RNA polymerase at the single-molecule level. The novel labelling position now allows examination of the motion of downstream DNA relative to a fixed point on the RNAP and opens up new avenues for examining transcription.

The flexibility offered by a fixed label on RNAP will allow questions to be addressed that were not previously possible using single-molecule methods. In combination with more advanced single molecule fluorescence techniques, such as 3-colour ALEX[83] and switchable FRET[74], more elaborate assays can now be considered to examine multiple

parameters, such as the presence – and even conformation – of other RNAP subunits or interacting partners, such as repressors and activators to be examined in parallel with the assays already described in this chapter. These new experimental avenues are made possible by the flexibility offered by UnAA technology to label a large, multisubunit protein in a site-specific manner.

More generally, this chapter has discussed a number of improvements to the UnAA protein modification and expression protocol: single-step mutagenesis conditions for long plasmids, strain selection for UnAA production using existing protein expression vectors, and modification of standard autoinduction media. The possibility of further optimisation of the Cu-free click labelling reactions remains; it is likely that more ideal conditions and reagents will be determined as more groups adopt this approach for labelling proteins for single-molecule studies. Taken together, the workflow presented in this chapter for mutagenesis, strain generation, and protein expression should prove useful for any group that wishes to implement UnAA-modification on proteins using existing in-house plasmids based on common protein expression vectors.

Contributions

I benefitted from discussions on mutagenesis and protein purification with T. Craggs, N. Robb, L. Aigrain, and D. Duchi (CMP, Oxford), as well as from instruction on UnAA methods at the EU P-CUBE Consortium Protein Engineering Workshop, led by E. Lemke and S. Tyagi (EMBL, Heidelberg, Germany). The pVS10 plasmids and original purification protocols were provided by the Artsimovitch group (Ohio State University, USA), while information about viable labelling positions on *E. coli* RNA polymerase was provided by R. Ebright (Rutgers University, New Jersey, USA) to A.

Kapanidis. Ryan Mehl (Oregon State University, USA) suggested the consideration of autoinduction and provided the pDULE2 plasmids, while Dina Grohmann (TU Braunschweig, Germany) provided advice regarding Copper-free click-chemistry labelling.

Mass spectrometry of excised gel bands was performed by H. Kramer (DPAG, Oxford), who also guided my analysis of the data. In vitro transcription reactions were performed by A. Tomescu (CMP) and single-molecule experiments with RNAP were performed by D. Duchi and L. Fernandez (CMP).

References

1. Beechem, J. M. *et al.* Single Molecule Real-time DNA Sequencing using FRET-based reagents. (2011). at <http://tools.lifetechnologies.com/content/sfs/posters/cms_091831.pdf>
2. Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc Natl Acad Sci USA* **105,** 1176–1181 (2008).
3. van Oijen, A. M. *et al.* Single-molecule kinetics of lambda exonuclease reveal base dependence and dynamic disorder. *Science* **301,** 1235–1238 (2003).
4. Stryer, L. & Haugland, R. P. Energy transfer: a spectroscopic ruler. *Proc Natl Acad Sci USA* **58,** 719–726 (1967).
5. Hohlbein, J., Gryte, K., Heilemann, M. & Kapanidis, A. N. Surfing on a new wave of single-molecule fluorescence methods. *Phys Biol* **7,** 031001 (2010).
6. Santoso, Y. *et al.* Conformational transitions in DNA polymerase I revealed by single-molecule FRET. *Proc Natl Acad Sci USA* **107,** 715–720 (2010).
7. Bertozzi, C. R. A decade of bioorthogonal chemistry. *Acc. Chem. Res.* **44,** 651–653 (2011).
8. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5,** 507–516 (2008).
9. Kapanidis, A. N. *et al.* Retention of transcription initiation factor sigma70 in transcription elongation: single-molecule analysis. *Mol. Cell* **20,** 347–356 (2005).
10. Kapanidis, A. N. *et al.* Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science* **314,** 1144–1147 (2006).
11. Pinkney, J. N. M. *et al.* Capturing reaction paths and intermediates in Cre-loxP recombination using single-molecule fluorescence. *Proc Natl Acad Sci USA* **109,** 20871–20876 (2012).
12. Mukhopadhyay, J. *et al.* Translocation of sigma(70) with RNA polymerase during transcription: fluorescence resonance energy transfer assay for movement relative to DNA. *Cell* **106,** 453–463 (2001).
13. Mendel, D. *et al.* Probing protein stability with unnatural amino acids. *Science* **256,** 1798–1802 (1992).
14. Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C. & Schultz, P. G. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science* **244,** 182–188 (1989).
15. Döring, V. *et al.* Enlarging the amino acid set of Escherichia coli by infiltration of the valine coding pathway. *Science* **292,** 501–504 (2001).
16. Kiick, K. L., van Hest, J. C. M. & Tirrell, D. A. Expanding the Scope of Protein Biosynthesis by Altering the Methionyl-tRNA Synthetase Activity of a Bacterial Expression Host. *Angew Chem Int Ed Engl* **39,** 2148–2152 (2000).
17. Kiick, K. L., Saxon, E., Tirrell, D. A. & Bertozzi, C. R. Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proc Natl Acad Sci USA* **99,** 19–24 (2002).
18. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. Expanding the genetic code of Escherichia coli. *Science* **292,** 498–500 (2001).
19. Chin, J. W. *et al.* Addition of p-azido-L-phenylalanine to the genetic code of Escherichia coli. *J Am Chem Soc* **124,** 9026–9027 (2002).
20. Xie, J. & Schultz, P. G. Adding amino acids to the genetic repertoire. *Curr Opin*

Chem Biol **9,** 548–554 (2005).
21. Ryu, Y. & Schultz, P. G. Efficient incorporation of unnatural amino acids into proteins in Escherichia coli. *Nat. Methods* **3,** 263–265 (2006).
22. Young, T. S., Ahmad, I., Yin, J. A. & Schultz, P. G. An enhanced system for unnatural amino acid mutagenesis in E. coli. *J. Mol. Biol.* **395,** 361–374 (2010).
23. Hammill, J. T., Miyake-Stoner, S., Hazen, J. L., Jackson, J. C. & Mehl, R. A. Preparation of site-specifically labeled fluorinated proteins for 19F-NMR structural characterization. *Nat Protoc* **2,** 2601–2607 (2007).
24. Peeler, J. C. & Mehl, R. A. Site-specific incorporation of unnatural amino acids as probes for protein conformational changes. *Methods Mol Biol* **794,** 125–134 (2012).
25. Tyagi, S. & Lemke, E. A. Genetically encoded click chemistry for single-molecule FRET of proteins. *Methods Cell Biol.* **113,** 169–187 (2013).
26. Brustad, E. M., Lemke, E. A., Schultz, P. G. & Deniz, A. A. A general and efficient method for the site-specific dual-labeling of proteins for single molecule fluorescence resonance energy transfer. *J Am Chem Soc* **130,** 17664–17665 (2008).
27. Milles, S. *et al.* Click strategies for single-molecule protein fluorescence. *J Am Chem Soc* **134,** 5187–5195 (2012).
28. Chakraborty, A. *et al.* Opening and closing of the bacterial RNA polymerase clamp. *Science* **337,** 591–595 (2012).
29. Chin, J. W. *et al.* An expanded eukaryotic genetic code. *Science* **301,** 964–967 (2003).
30. Greiss, S. & Chin, J. W. Expanding the genetic code of an animal. *J Am Chem Soc* **133,** 14196–14199 (2011).
31. Bianco, A., Townsley, F. M., Greiss, S., Lang, K. & Chin, J. W. Expanding the genetic code of Drosophila melanogaster. *Nat. Chem. Biol.* **8,** 748–750 (2012).
32. Mukai, T. *et al.* Adding l-lysine derivatives to the genetic code of mammalian cells with engineered pyrrolysyl-tRNA synthetases. *Biochem. Biophys. Res. Commun.* **371,** 818–822 (2008).
33. Saxon, E. & Bertozzi, C. R. Cell surface engineering by a modified Staudinger reaction. *Science* **287,** 2007–2010 (2000).
34. Hein, J. E. & Fokin, V. V. Copper-catalyzed azide-alkyne cycloaddition (CuAAC) and beyond: new reactivity of copper(I) acetylides. *Chem Soc Rev* **39,** 1302–1315 (2010).
35. Wang, L., Zhang, Z., Brock, A. & Schultz, P. G. Addition of the keto functional group to the genetic code of Escherichia coli. *Proc Natl Acad Sci USA* **100,** 56–61 (2003).
36. Chakraborty, A., Wang, D., Ebright, Y. W. & Ebright, R. H. Azide-specific labeling of biomolecules by Staudinger-Bertozzi ligation phosphine derivatives of fluorescent probes suitable for single-molecule fluorescence spectroscopy. *Meth. Enzymol.* **472,** 19–30 (2010).
37. Lin, F. L., Hoyt, H. M., van Halbeek, H., Bergman, R. G. & Bertozzi, C. R. Mechanistic investigation of the staudinger ligation. *J Am Chem Soc* **127,** 2686–2695 (2005).
38. Kolb, H. C., Finn, M. G. & Sharpless, K. B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew Chem Int Ed Engl* **40,** 2004–2021 (2001).
39. Rostovtsev, V. V., Green, L. G., Fokin, V. V. & Sharpless, K. B. A stepwise huisgen cycloaddition process: copper(I)-catalyzed regioselective ‘ligation’ of

azides and terminal alkynes. *Angew Chem Int Ed Engl* **41,** 2596–2599 (2002).
40. Deiters, A. & Schultz, P. G. In vivo incorporation of an alkyne into proteins in Escherichia coli. *Bioorg. Med. Chem. Lett.* **15,** 1521–1524 (2005).
41. Agard, N. J., Prescher, J. A. & Bertozzi, C. R. A strain-promoted [3 + 2] azide-alkyne cycloaddition for covalent modification of biomolecules in living systems. *J Am Chem Soc* **126,** 15046–15047 (2004).
42. Wang, A., Nairn, N. W., Marelli, M. & Grabstein, K. Protein Engineering with Non-Natural Amino Acids. *Protein Engineering with Non-Natural Amino Acids, Protein Engineering, Prof. Pravin Kaumaya (Ed.)* (2012).
43. Plass, T., Milles, S., Koehler, C., Schultz, C. & Lemke, E. A. Genetically encoded copper-free click chemistry. *Angew Chem Int Ed Engl* **50,** 3878–3881 (2011).
44. Baskin, J. M. *et al.* Copper-free click chemistry for dynamic in vivo imaging. *Proc Natl Acad Sci USA* **104,** 16793–16797 (2007).
45. Ning, X., Guo, J., Wolfert, M. A. & Boons, G.-J. Visualizing metabolically labeled glycoconjugates of living cells by copper-free and fast huisgen cycloadditions. *Angew Chem Int Ed Engl* **47,** 2253–2255 (2008).
46. Kuzmin, A., Poloukhtine, A., Wolfert, M. A. & Popik, V. V. Surface functionalization using catalyst-free azide-alkyne cycloaddition. *Bioconjug Chem* **21,** 2076–2085 (2010).
47. Debets, M. F. *et al.* Aza-dibenzocyclooctynes for fast and efficient enzyme PEGylation via copper-free (3+2) cycloaddition. *Chem Commun (Camb)* **46,** 97–99 (2010).
48. Sletten, E. M., Nakamura, H., Jewett, J. C. & Bertozzi, C. R. Difluorobenzocyclooctyne: synthesis, reactivity, and stabilization by beta-cyclodextrin. *J Am Chem Soc* **132,** 11799–11805 (2010).
49. Sletten, E. M. & Bertozzi, C. R. From mechanism to mouse: a tale of two bioorthogonal reactions. *Acc. Chem. Res.* **44,** 666–676 (2011).
50. Soellner, M. B., Nilsson, B. L. & Raines, R. T. Reaction mechanism and kinetics of the traceless Staudinger ligation. *J Am Chem Soc* **128,** 8820–8828 (2006).
51. Presolski, S. I., Hong, V., Cho, S.-H. & Finn, M. G. Tailored ligand acceleration of the Cu-catalyzed azide-alkyne cycloaddition reaction: practical and mechanistic implications. *J Am Chem Soc* **132,** 14570–14576 (2010).
52. Darst, S. A. Bacterial RNA polymerase. *Curr. Opin. Struct. Biol.* **11,** 155–162 (2001).
53. Buc, H. & Strick, T. *RNA polymerases as molecular motors*. (Cambridge, UK : RSC Pub, 2009).
54. Robb, N. C. *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.* **425,** 875–885 (2013).
55. Feklistov, A. & Darst, S. A. Structural basis for promoter-10 element recognition by the bacterial RNA polymerase σ subunit. *Cell* **147,** 1257–1269 (2011).
56. Zhang, Y. *et al.* Structural basis of transcription initiation. *Science* **338,** 1076–1080 (2012).
57. Margeat, E. *et al.* Direct observation of abortive initiation and promoter escape within single immobilized transcription complexes. *Biophys. J.* **90,** 1419–1431 (2006).
58. Weixlbaumer, A., Leon, K., Landick, R. & Darst, S. A. Structural basis of transcriptional pausing in bacteria. *Cell* **152,** 431–441 (2013).

59. Larson, M. H., Greenleaf, W. J., Landick, R. & Block, S. M. Applied force reveals mechanistic and energetic details of transcription termination. *Cell* **132,** 971–982 (2008).
60. Carpousis, A. J. & Gralla, J. D. Cycling of ribonucleic acid polymerase to produce oligonucleotides during initiation in vitro at the lac UV5 promoter. *Biochemistry* **19,** 3245–3253 (1980).
61. Hsu, L. M. Monitoring abortive initiation. *Methods* **47,** 25–36 (2009).
62. Brodolin, K., Zenkin, N., Mustaev, A., Mamaeva, D. & Heumann, H. The sigma 70 subunit of RNA polymerase induces lacUV5 promoter-proximal pausing of transcription. *Nat. Struct. Mol. Biol.* **11,** 551–557 (2004).
63. Perdue, S. A. & Roberts, J. W. A backtrack-inducing sequence is an essential component of Escherichia coli σ(70)-dependent promoter-proximal pausing. *Mol. Microbiol.* **78,** 636–650 (2010).
64. Hsu, L. M. Promoter clearance and escape in prokaryotes. *Biochim. Biophys. Acta* **1577,** 191–207 (2002).
65. Yin, H. *et al.* Transcription against an applied force. *Science* **270,** 1653–1657 (1995).
66. Forde, N. R., Izhaky, D., Woodcock, G. R., Wuite, G. J. L. & Bustamante, C. Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by Escherichia coli RNA polymerase. *Proc Natl Acad Sci USA* **99,** 11682–11687 (2002).
67. Abbondanzieri, E. A., Greenleaf, W. J., Shaevitz, J. W., Landick, R. & Block, S. M. Direct observation of base-pair stepping by RNA polymerase. *Nature* **438,** 460–465 (2005).
68. Hohlbein, J., Craggs, T. D. & Cordes, T. Alternating-laser excitation: single-molecule FRET and beyond. *Chem Soc Rev* (2013). doi:10.1039/c3cs60233h
69. Cordes, T. *et al.* Sensing DNA opening in transcription using quenchable Förster resonance energy transfer. *Biochemistry* **49,** 9171–9180 (2010).
70. Darst, S. A. *et al.* Conformational flexibility of bacterial RNA polymerase. *Proc Natl Acad Sci USA* **99,** 4296–4301 (2002).
71. Wang, D. Ensemble Fluorescence Resonance Energy Transfer Analysis of RNA Polymerase Clamp Conformation.
72. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
73. Muschielok, A. *et al.* A nano-positioning system for macromolecular structural analysis. *Nat. Methods* **5,** 965–971 (2008).
74. Uphoff, S. *et al.* Monitoring multiple distances within a single molecule using switchable FRET. *Nat. Methods* **7,** 831–836 (2010).
75. Belogurov, G. A. *et al.* Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol. Cell* **26,** 117–129 (2007).
76. *QuikChange II XL Site-Directed Mutagenesis Kit*. (Agilent Technologies).
77. *Guidelines for PCR Optimization with Phusion High-Fidelity DNA Polymerase*. (New England BioLabs). at <https://www.neb.com/protocols/2012/06/01/guidelines-for-pcr-optimization-with-phusion-high-fidelity-dna-polymerase>
78. Guzman, L. M., Belin, D., Carson, M. J. & Beckwith, J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* **177,** 4121–4130 (1995).
79. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41,** 207–234 (2005).

80. Anthony, L. C., Artsimovitch, I., Svetlov, V., Landick, R. & Burgess, R. R. Rapid purification of His(6)-tagged Bacillus subtilis core RNA polymerase. *Protein Expr. Purif.* **19,** 350–354 (2000).
81. Jain, A. *et al.* Probing cellular protein complexes using single-molecule pull-down. *Nature* **473,** 484–488 (2011).
82. Harada, Y. *et al.* Direct observation of DNA rotation during transcription by Escherichia coli RNA polymerase. *Nature* **409,** 113–115 (2001).
83. Lee, N. K. *et al.* Three-color alternating-laser excitation of single molecules: monitoring multiple interactions and distances. *Biophys. J.* **92,** 303–312 (2007).

Chapter 6 – Conclusion

I have presented a broad range of experiments that examine different aspects of DNA on the single molecule level – from genomic structural variation at the nano-scale, to chromosomal architecture at the micro-scale. I have explored how DNA imaging could be improved using novel optical materials created using the DNA itself – and how the study of interactions between DNA and large, multi-subunit proteins could be improved through the use of unnatural amino acids and click chemistry.

We have shown that it is possible to view structural variation in the genome by imaging ultra-long DNA fragments of DNA that are stretched within a nanofluidic chip. In a single image acquisition of less than 5 seconds, we obtain a 'birds-eye-view' of the genome at ~1 kbp resolution from Denaturation-Renaturation mapping. The perspective of genomic structure that we obtain constitutes a haplotype: it is derived from a single chromosomal piece of DNA. I demonstrated that base-level sequence information could be obtained from the *same* single DNA molecule after extracting it from the nanofluidic chip, amplifying, and sequencing it. I used the sequencing data to confirm successful isolation of single DNA fragments from the chip and validate the barcoding method used to detect SV.

Work is on-going to develop a 'second generation' of this platform that allows the entire process – from single-cell handling before mapping, to DNA sequencing library preparation – to be performed on a single chip made out of injection-moulded plastic. The algorithms I developed to process sequencing data derived from single, amplified fragments of DNA will form the foundation of the data analysis pipeline for future data sets from this improved nanofluidic chip.

The fluidic devices can also serve as a 'reaction chamber' for the examination of chromosomes and DNA while confined within the device. After removing chromosomal proteins, we are able to gently manipulate the mass of DNA and can directly observe DNA catenation, tethering sister chromatids. This perspective – the first direct image of catenation on native, unfixed metaphase chromosomes – allowed us to alter the DNA catenation state of the chromosomes and observe the effect it had on their shape when confined within the microfluidic chip. I found that DNA catenation played a substantial role in determining chromosome architecture – both between sister chromatids and along the axis of each chromatid. These results highlight the challenges that will be faced by any third-generation DNA sequencing method that attempts to sequence chromosomes directly: the complex topological structure of the genome must be taken into account when developing sample preparation methods.

In the second part of this thesis, I have considered two improvements to single-molecule fluorescence assays. First, I have demonstrated that DNA intercalating dyes can transfer energy along a DNA molecule and effectively confine illumination to DNA-bound fluorophores. I also show, in principle, how they could be incorporated into a 'third-generation' DNA sequencing approach based on the detection of dye-labelled nucleotide incorporation by a DNA polymerase. In the near-term, this work provides insight into the process of complex energy transfer in multi-chromophore systems.

Second, I have used unnatural amino acid technology to engineer an *E. coli* RNA polymerase with a fluorophore at a position that allows us to monitor the motion of DNA relative to a single polymerase molecule. We will be able to resolve single-basepair steps of the polymerase while it initiates transcription and then transitions to elongation. While perhaps a useful technique for engineering enzymes for DNA

sequencing methods in the future, the use of unnatural amino acids has produced a new reagent for the study of prokaryotic transcription and the demonstrates that unnatural amino acids can be broadly applied in ultra-sensitive analysis of nucleic acids and mechanistic studies of nucleic acid polymerases.

In this thesis I have presented a broad variety of work, drawing on techniques from molecular biology, protein chemistry, optics and mathematical modelling. The results all serve to highlight the value of interdisciplinary collaboration between the physical and biological sciences. Work across multiple disciplines can contribute to solving challenges in genomics – and in the process yield new important results for each respective discipline as well.

Appendix A – Publications and Presentations

Peer-Reviewed Publications

Marie R, Pedersen JN*, **Bauer DLV***, Rasmussen KH, Yusuf M, Volpi EV, Flyvbjerg H, Kristensen A, and Mir KU. An Integrated Single Molecule View of Genome Structure and Sequence. *Proceedings of the National Academy of Sciences (USA),* 110, 4893–4898. (2013). [*equal contribution]

Bauer DLV, Marie R, Rasmussen KH, Kristensen A, and Mir KU. DNA Catenation Maintains Structure of Human Metaphase Chromosomes. *Nucleic Acids Research*, 40, 11428-11434 (2012).

Yusuf M, **Bauer DLV**, Lipinski DM, MacLaren RE, Wade-Martins R, Mir K, and Volpi EV. Combining M-FISH and Quantum Dot Technology for Fast Chromosomal Assignment of Transgenic Insertions. *BMC Biotechnology*, 11:121. (2011).

Conference Talks & Posters

Bauer DLV et al. Nanofluidics to Enhance Single Molecule DNA Imaging: Detecting Genomic Structural Variation in Humans. *19th International Workshop on Single Molecule Spectroscopy – PicoQuant* (2013). [Talk won 'Best Student' prize]

Bauer DLV, Duchi D, Tomescu A, Fernandez L, and Kapanidis AN. Cutting-Edge Meets Extreme Leading-Edge: Unnatural Amino Acid Labels Measure Downstream DNA Motion During Transcription. *RSC Nucleic Acids Forum*, (2013). [Poster]

Bauer DLV, Mir KU and Kapanidis AN. Single Molecule Studies of DNA-based Photonic Wires. *RSC Nucleic Acids Forum*, (2012). [Poster won 2nd Place Prize]

Bauer DLV and Mir KU. Towards Single Molecule DNA Sequencing, *Wellcome Trust Leena Peltonen School of Human Genomics*, (2010). [Poster & Talk]

Appendix B – SV Discovered in NA18507

SV was detected on 3 DNA fragments (A, B, and C) from the Yoruban African sample we analysed using the Sliding Window method on DR Maps. I undertook a comprehensive review of the variants we identified using this approach using data from the Database of Genomic Variation (DGVa) viewed with the ENSEMBL Genome Browser. Genome coordinates of SV breakpoints from DR mapping were based on the hg18 genome assembly; they were converted to hg19 coordinates using the NCBI Genome Remapping Service as required.

A summary table is presented below.

Context and supporting evidence for SV discovered by DR Mapping of NA18507

	Type	Chr	Start	End (MBp)	Length	ENSEMBL Gene Annotation	DGVa Results
A	Inversion	16	21.4[a]	22.4	1	mid-arm, genic w/cDNA mostly unknown fcn	**Reported**: Tuzun 2005, Korbell 2007, Kidd 2008 & 2010, Pang 2010
B	Deletion	19	62.18	62.22	0.04	Sub-telomeric region, intergenic	**Novel**, but flanking deletions [1000 Genomes 2010; Cooper 2011 Nat Gen]
	Insertion	19	62.18	–	0.20	"	**Novel**, but flanking insertions [Kidd 2008, Teague 2010]
	(Translocation)[b]	9	22.1	22.3	0.20	mid-arm, incl. 3' end (20kb) of CDKN2B asRNA. Nearby: MTAP is 0.2Mb upstream	**Novel**, but encompassing CNVs reported in Sub-saharan population [Vogler 2010] and encompassing 0.12 Mbp CN gain in Asian population [Xu 2011 Hum Mutat]
C	Insertion	7	42.49	–	0.007	mid-arm, inter-genic	**Novel**, but within span of reported insert [Kidd 2008], and many small insertions nearby from other studies [Levy 2007; Bentley 2008; McKernan 2009]

[a] the inversion start extends beyond the DR map image; start is approximate
[b] while we are certain that a fragment of DNA inserted into chr19 maps to chr9 in the reference genome, we cannot be certain whether it is a duplication of this region of chr9 or a true translocation without mapping the corresponding region of chr9 to gain context

Remapping* SV positions (MBp) from hg18 (NCBI36) to hg19 (GRCh37) genome assembly

	Type	Chr	Start hg18	End hg18	Start hg19	End hg19
A	Inversion	16	21.4	22.4	21.5	22.5
B	Deletion	19	62.18	62.22	57.49	57.53
	Insertion	19	62.18	–	57.49	–
	(Translocation)[b]	9	22.1	22.3	22.1	22.3
C	Insertion	7	42.49	–	42.52	–

* remapping was performed using the NCBI Genome Remapping Service (http://www.ncbi.nlm.nih.gov/genome/tools/remap)

Appendix C – Materials and Methods

Chapter 2

As noted in the chapter itself, the materials & methods have already been published[1], in part.

Bioinformatics and Calculations

Manipulation of sequencing data was performed as described below using a variety of software, including *bowtie* (version 0.12.7), *samtools* (version 0.1.16, including *bcftools*) on a 2.93 GHz Intel Quad Core i7 iMac with 16 GB RAM running Mac OS X 10.6.8 (Apple, Cupertino, CA, USA). Pre-built indices for *bowtie* (hg19) were downloaded from the software's SourceForge page (http://bowtie-bio.sourceforge.net/). Analysis of the results was further carried out using custom bash scripts in combination with Matlab r2010a (Mathworks, Natick, MA, USA), R 64-bit version (R Development Core Team), and Excel 2011 for Mac (Microsoft, Redmond, WA, USA). Unless otherwise noted, standard deviations were calculated using either Matlab's STD function or Excel's STDEVP function and represented as value ± standard deviation; significant figures are accurately represented.

Read Contamination Analysis

A random sampling of reads was obtained by sampling the FASTQ files of raw sequence reads using a bash script to read every n^{th} sequence such that the total number of sequences sampled was the desired number. For example:

```
awk '( (NR+2)%4==0 || (NR+3)%4==0 ) && ( (NR+2)%156000==0 || (NR+3)%156000==0)'
 file.fastq | sed 's/@/>/g' > toBlast.fasta
```

The resulting sample of reads was then fed into a custom Matlab program that submitted them individually as MEGABLAST queries to the NCBI (NIH, Bethesda, MD, USA). The first word of the top result was recorded and the frequency of these words was counted for the sample of reads. The program was written to automatically generate a pie chart and data table to present the results and was later optimised by rewriting the entire workflow to run locally.

Raw Read Distribution Analysis

The raw number of reads mapped to each chromosome was counted manually with a bash script and then divided by the length of the chromosome to calculate the weighted number of reads mapped to each chromosome:

```
#!/bin/bash
for c in {1..22}
do
  echo 'Summing chromosome '${c}
  samtools view -c mol9.bam 'chr'${c} >> outSum
done

c="X"
echo 'Summing chromosome '${c}
samtools view -c mol9.bam 'chr'${c} >> outSum

c="Y"
echo 'Summing chromosome '${c}
samtools view -c mol9.bam 'chr'${c} >> outSum
```

To identify the locations within each chromosome to which the reads were mapping, the number of reads mapping to a rolling 2 MBp window along each chromosome were counted using a custom bash script:

```
#!/bin/bash
for c in {1..22}
do
  echo 'Summing chromosome '${c}
  samtools view -c mol9.bam 'chr'${c}':1-1000000' >> out${c}.txt
  for i in {2..300}
  do
          let startPos=${i}-1
          let endPos=${i}+1
          samtools view -c mol9.bam 'chr'${c}':'${startPos}'000000-
'${endPos}'000000' >> out${c}.txt
  done
done
```

```
c="X"
echo 'Summing chromosome '${c}
for i in {2..300}
do
  let startPos=${i}-1
  let endPos=${i}+1
  samtools view -c mol9.bam 'chr'${c}':'${startPos}'000000-'${endPos}'000000' >>
out${c}.txt
done

c="Y"
echo 'Summing chromosome '${c}
for i in {2..300}
do
  let startPos=${i}-1
  let endPos=${i}+1
  samtools view -c mol9.bam 'chr'${c}':'${startPos}'000000-'${endPos}'000000' >>
out${c}.txt
done
```

Expected Read Counts per Chromosome

The hg19 human genome reference assembly was obtained as individual chromosome files from the UCSC Genome Browser (Santa Cruz, CA, USA) and used by Metasim to simulate 1 million mate-pairs corresponding to 2 million 51bp Illumina sequence reads. Empirical error profiles for 80 bp were used as supplied by the authors (www-ab.informatik.uni-tuebingen.de/software/metasim) except with the length trimmed to 51 bp. The reads were then aligned using *bowtie* with the command:

```
/Applications/seq/bowtie/bowtie --chunkmbs 256 -f -S -p8 -m 1 --best --strata -I
  $insertMin -X $insertMax $reference -1 $readsForward  -2 $readsReverse  |
  samtools view -S -b - | samtools sort - ${readsForward}_bowtie.bam
```

The number of reads per chromosome was obtained using *samtools*' IDXSTATS command and expected vs. actual reads were tabulated in Excel and plotted using OmniGraphSketcher.

Footprint Analysis - Mapping to Specific Chromosomal Regions

Using a custom bash script, the whole genome was subdivided into 1 MBp bins, and each bin was further divided into 1,000 1-KBp regions. The number of these smaller regions with more than 51% average coverage was counted per larger 1 MBp bin using *samtools* and the output recorded and plotted using Matlab. A comprehensive R script

was later developed to generate the footprint results presented in Chapter 3 in a much shorter amount of time.

Visualisation of Read Clustering

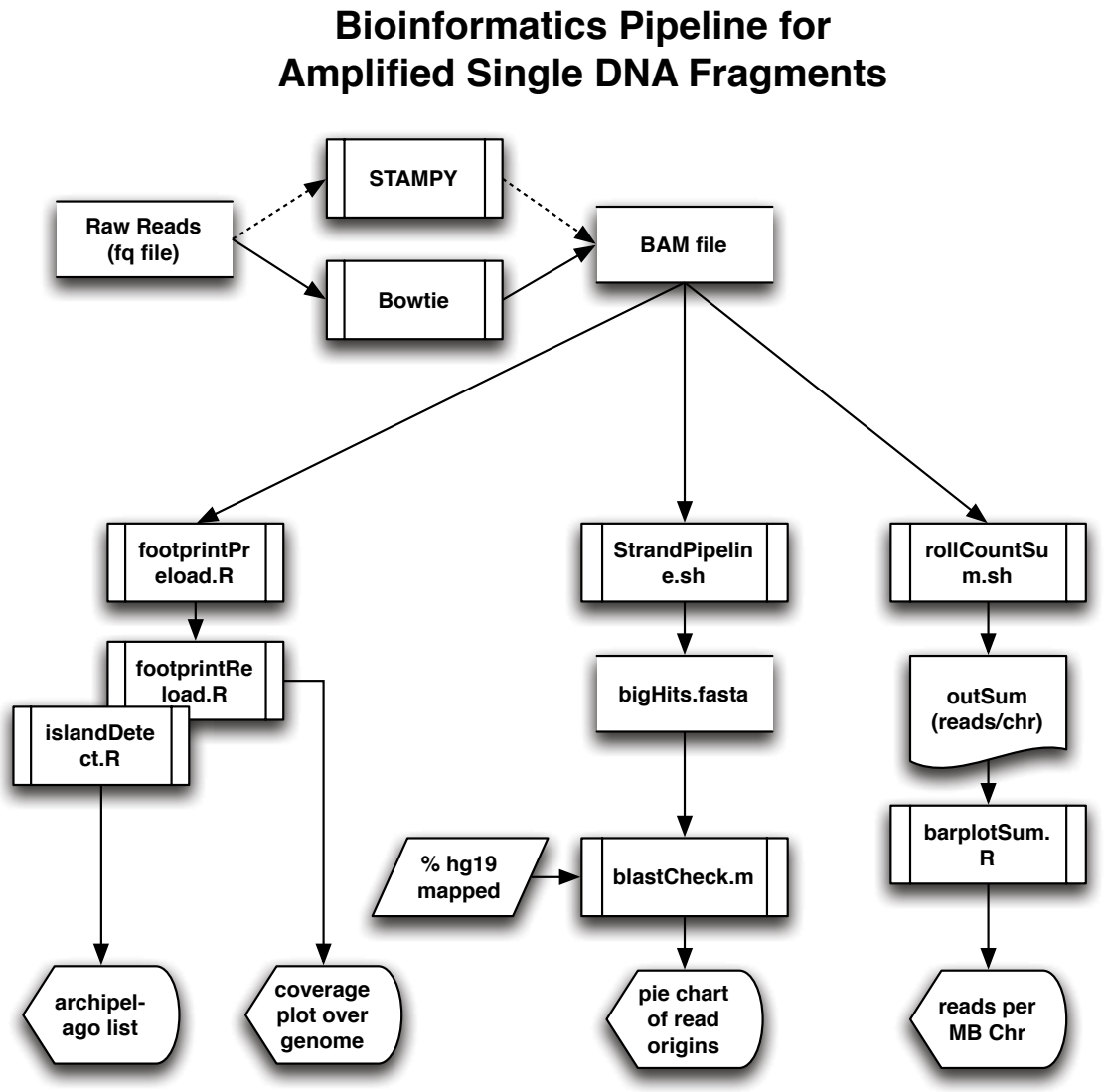
The first base position of each read mapped to chromosome 9 was recorded and the positions binned into 1 KBp bins using Matlab. Discrete clusters of reads, termed "islands", were observed and representative islands were plotted as genome position in MBp vs. read strength (number of reads × read length). Additional visualization to confirm the results was performed using the Integrative Genome Viewer[2] (version 1.5.64, Broad Institute, Cambridge, MA, USA).

Island and Archipelago Detection

Clusters of nearby (<20 KBp away) well-covered regions (1 KBp regions with >51% coverage) were grouped into islands. The locations of islands along each chromosome were plotted with Matlab, as shown for chromsomes 4 and 9. Large islands ('bigIslands', longer than 5 KBp) were selected for further analysis and bigIslands within 1MBp of one another were then collected together as a single archipelago using a custom algorithm written in Matlab. The observed archipelagos in a given data set were reported along with their overall length and constituent number of bigIslands to determine whether the results were significant or not. The same algorithm was later rewritten in R.

Overall Workflow Implemented

The overall workflow discussed above is shown schematically, below:



Variant Calling & Heterozygocity Calculations

A standard *samtools* / *bcftools* variant calling pipeline was used to find variants along the observed MolFive archipelago using the command:

```
samtools mpileup -r 'chr9:108273000-111803000' -uf 'chr9.fa' mol5.bam |
  /Applications/seq/samtools/bcftools/bcftools view -vcg - > mol5CHR9.vcf
```

The resulting table was imported into Microsoft Excel for further analysis; the AF1 (Allele Frequency 1) field was used to determine whether a read was being called as homozygous. SNPs reported in the MolFive archipelago for the 1000 Genomes data set and in various HapMap data sets with a given MAF were downloaded from dbSNP,

build 134 (NCBI). The results were then analysed in Excel to determine the mean rate of HetSNP occurrence among SNPs with determined heterozygocity. Plots of heterzygocity (read concordance) among all base positions at a given depth cutoff and plots of the number of positions covered at given depth cutoffs were made in Matlab using a table generated by *samtools*' MPILEUP command.

Automated quantitation of discordance between reads was obtained using the command:

```
samtools mpileup -f chr9.fa -r 'chr9:108285000-111569000' mol9.bam | awk
'{if($4>19) {split($5,x,","); split($5,y,"."); nummatch=length(x)+length(y)-2;
```

Heterozygocity Comparison Using Genotyping Data

Genotyping was carried out by Sam Knight (WTCHG, Oxford) using the Illumina Infinium assay on a Human Omni1-Quad (v1.0) chip. Genotyping calls were made using Illumina software (Illumina, Cambridge, UK). I cross-matched these calls with well- covered (depth >19 reads) positions in the sequencing data of the single DNA molecules using a custom Matlab / bash script:

```
countNull = 0;
for i = 1:length(posList)
    pos = posList(i);
    text = ['/opt/local/bin/samtools mpileup -f ''/ReferenceSequence/chr9.fa'' -r
''chr9:' int2str(pos) '-' int2str(pos) ''' /sequencingData.bam | awk ''$4>9''  >>
/genotypingSequencingIntersect.txt' ];
    result = system(text);
    if result == 0
        countNull = countNull + 1;
    end
    disp(result);
end
```

The results from this script were then inputted into Microsoft Excel and examined for the degree of heterozygosity of the basecalls.

Nanofluidic Chip fabrication, Handling, DR Mapping, and SV Detection

Details regarding these methods are presented in our publication[1] and supplement.

Chapter 3

As noted in the chapter itself, the materials & methods section here has been published[3].

Solutions and reagents

Chromosomes were isolated from mammalian cell culture (Jurkat cell line, DSMZ, Germany) based on a modified polyamine isolation protocol[4,5]. YOYO-1 DNA intercalating dye was used without further purification (Life Technologies, Naerum, Denmark). Chemicals and buffers were purchased from Sigma Aldrich (Copenhagen, Denmark) and any water used was purchased nuclease-free unless otherwise noted. Buffer solutions for chip-handling work were prepared on the day of use from degassed 0.5x TBE buffer. A working buffer solution for chip handling (BwT) was composed of 3% 2-mercaptoethanol (BME), 0.5% Triton-X and 0.5x degassed Tris-borate EDTA. For chromosome pre-treatment, fresh solutions (described below) were prepared and care was taken to avoid inadvertent nuclease contamination. Stock 1M DTT was diluted to 10 mM working concentration. Human topoisomerase IIα solution (Topo IIα), dexrazoxane, and etoposide were purchased from Sigma Aldrich (UK). Topo IIα was used as supplied and the inhibitors were dissolved in DMSO at 30 mg/mL prior to use.

Lab-on-a-chip device

The fluidic chip was fabricated in silicon as previously described[6]; the chip contained a chromosome inlet channel, leading to a central reaction chamber, and a chromosome outlet channel. Reagents could be flushed around this chamber from separate inlets and allowed to diffuse inwards to the chromosome without dislodging it. The surface of the interior channels of the chip was coated with a bovine serum albumin solution (BwT+BSA: 0.1 mg/mL BSA in BwT) prior to chromosome loading. The entire chip was mounted in a custom holder on a Nikon TE-2000 microscope that contained wells

for loading solutions and air pressure was used to drive fluid flow from these wells. A cartridge heater (Omega Engineering, UK) set at 37 °C was used to maintain temperature during digestion.

Chromosome pre-treatment, staining, and digestions

To observe native chromosomes, staining was performed by supplementing chromosome suspensions from cell culture with YOYO-1 dye to 1 µM and incubating in the dark for 1 hour at room temperature, followed by 1 hour at 50ºC. Chromosomes were diluted in BwT+BSA to approximately 10^5 chromosomes per ml, then loaded onto the microfluidic chip and positioned in the central reaction chamber by fluid flow. ProK (0.1 mg/mL in BwT+BSA buffer supplemented with YOYO-1 to 1 µM) was continuously flushed around the chamber from where it diffused into the reaction chamber and digested chromosomal proteins.

To pre-treat chromosomes with Topo IIα, unstained chromosomes were diluted in Topo II reaction buffer with a final concentration of 50 mM Tris at pH 8, 10 mM $MgCl_2$, 120 mM KCl, and 0.5 mM DTT. When used, inhibitors were added to a final concentration of 2.6 mg/ml. The reactions were then incubated for 45 minutes at room temperature, after which the reaction was supplemented with ATP to 0.5 mM and with 0.5 µl Topo IIα solution. The reaction was mixed by gently pipetting up and down, incubated for 30 minutes at 37 ºC, and then heated to 65 ºC to inactivate the Topo IIα enzyme. Chromosomes were then stained, loaded, and digested as described above.

Image acquisition and processing pipeline

Imaging was conducted on a custom Nikon TE-2000 inverted fluorescence microscope using metal halide lamp illumination and appropriate excitation and emission filters for YOYO-1 dye (B2E/C filter, Nikon Instruments, Amstelveen, Netherlands). Image

acquisition was conducted using MetaMorph 7 (Molecular Devices, Sunnyvale, CA, USA), which controlled a Cascade II:512 back-illuminated EMCCD camera (Photometrics, Tuscon, AZ, USA), and an electromechanical illumination shutter (Prior Scientific, Jena, Germany). Time-lapse images of each chromosome digestion were acquired using a macro which sampled a set of 5 frames every 30 seconds and averaged each set in Matlab (Mathworks, Natick, MA, USA) to reduce image jitter and real-time images of DNA catenation were frame-averaged to produce clearer static images for publication. ImageJ and Fiji were used to apply automatic thresholding and extract feature length and roundness during timelapse reactions. Roundness was calculated as $4\pi(area/perimeter^2)$. A custom ImageJ Java macro was used to carry out this morphometry analysis:

```
setAutoThreshold("Default");
setAutoThreshold("Default dark");
setAutoThreshold("Default dark");
run("Convert to Mask", "calculate black");
setAutoThreshold("Default dark");
run("Convert to Mask", "calculate black");
run("Open", "stack");
run("Fill Holes", "stack");
run("Analyze Particles...", "size=100-Infinity circularity=0.00-1.00 show=Nothing
display clear summarize stack");
saveAs("Text", "/DataFolder/file.xls");
```

The output from each run was put into Excel & plotted vs. time. The morphology of the chromosomes after 20 minutes' digestion was used to compare different chromosome samples.

Chapter 4

Modelling

Initial calculations to determine the R_0 between dye pairs were made using Microsoft Excel. Random walk simulations were carried out using Matlab using the following code:

```
% Random walk of exciton modelled for a wire of 3-10 Intercalating Dyes
% Determine where the exciton will go

hold off
close
ylim([0 400])

minDyes = 3;
maxDyes = 10;

results = zeros(maxDyes-minDyes,3);

for numDyes = minDyes:maxDyes;
    numExperiments = 10;
    stepsToExitMatrix = zeros(1,numExperiments);

    for j = 1:numExperiments
        numWalks = 10000;
        stepsToExit = zeros(1,numWalks);

        for i=1:numWalks
            %disp (['i = ' int2str(i)])
            excitonPosition = 1;
            excitonSteps = 0;
            while excitonPosition ~= numDyes
                if excitonPosition == 1
                    newPosition = 2;
                    excitonPosition = newPosition;
                elseif excitonPosition > 1
                    if rand > 0.5
                        newPosition = excitonPosition - 1;
                    else
                        newPosition = excitonPosition + 1;
                    end
                    excitonPosition = newPosition;
                end
                %disp ([int2str(excitonSteps) ' :: ' int2str(excitonPosition)])
                excitonSteps = excitonSteps + 1;
            end
            stepsToExit(i) = excitonSteps;
        end

        stepsToExitMatrix(j) = mean(stepsToExit);
    end

    map = colormap(jet(1+maxDyes-minDyes));

    plot(histc(stepsToExit,0:10:1000), 'LineWidth', 2,'Color', map(numDyes-
minDyes+1,:))
    hold on

    meanStepsToExitMatrix = mean(stepsToExitMatrix);
    stdvStepsToExitMatrix = std(stepsToExitMatrix);
    disp(sprintf('For %i dyes, %2.2f +/- %1.2f steps to exit',numDyes,
meanStepsToExitMatrix,stdvStepsToExitMatrix ))
    results(numDyes,:) = [numDyes, meanStepsToExitMatrix,stdvStepsToExitMatrix];

end

ylim([0 4000])
xlim([0 30])
```

the results from the simulations were fed back into Microsoft Excel, which was used to perform the calculations described in the text. 3D structural modelling of the DNA-intercalator complexes was performed using PyMol and by aligning two identical models of PBB 108D as described in the text, using the align command for the selected basepairs, as described in the text. The nanopositioning system[7] of Muschielok et al.

was used to model the position of Alexa 488 and ATTO 647N using the FRETnps Matlab package and the "Model satellite position prior" program, assuming Linker length = 20 Å; diameter = 4.5 Å; dye diameter = 7 Å. The resulting 3D Gaussian approximation of dye position was imported into PyMol as a pseudoatom using the command:

```
pseudoatom DYE, pos=[x,y,z], vdw=3
```

This position was used to determine distances within the model using PyMol.

Photonic Wire Formation and Measurement

Two complementary DNA oligonucleotides, PWT7-top and PWT7-bottom were designed to contain a 3' C6 Amino modification. PWT7-top contained a 5' biotin as well. 5 nmol of each oligonucleotide was labelled with a 10-fold excess of amine-reactive NHS ester dye derivative: PWT7-top with Alexa 488 and PWT7-bottom with ATTO 647N.

Oligo	**5'-Sequence-3'**
PWT7-top	GAACTGATAAATTAATACGACTCACTATAGGGA
PWT7-bottom	TCCCTATAGTGAGTCGTATTAATTTATCAGTTC

The reaction was ethanol precipitated to remove excess free dye and the products separated by denaturing PAGE. Gel bands corresponding to full-length, labelled DNA were excised and DNA was eluted from the gel by crush-soaking in TE buffer for 18 hours, and re-precipitating the DNA. Similarly, unlabelled DNA was purified using the same method. Photonic wire constructs described in the text were formed using the appropriate labelled or unlabelled oligonucleotides. DNA was annealed at 1µM concentration in AB buffer (TE+250 mM NaCl) and photonic wire constructs were formed by diluting DNA to 100 nM and incubating with POPO-3 or BOBO-3 dye (Life

Technologies, UK) at a ratio of 1 dye : 4bp (for 100 µl reaction: 8.25 µl POPO-3 dissolved in DMSO at 10 µM) for 1 hour in the dark, at room temperature. Control fluorescence spectra of DNA constructs indicated that DMSO did not affect the spectral properties of the Alexa 488 and ATTO 647N. Fluorescence spectra were taken of the dyes using a PTI fluorimeter and corrected for detector efficiency. Further corrections, quantitation, and calculations described in the text were carried out in Microsoft Excel.

Feeder Photonic Wires

Photonic wires were formed as described above from annealed DNA, but with YOYO-1 intercalating dye and constructs that contained a single ATTO 647N terminal label. For DNA microarray experiments, photonic wire constructs were spotted onto a glass slide, which was dried in air and imaged using a ProScanArray laser microarray scanner (Perkin Elmer, Waltham, MA, USA) by exciting YOYO-1 at 488 nm and ATTO 647N at 633 nm. Images for emission in both channels were recorded at each stage, and the intensities of the images were quantitated using the ProScanArray software. For single molecule measurements, the constructs were diluted to 1 nM and deposited on a clean, PEGylated coverslip by a biotin-streptavidin linkage using well-established protocols[8]. The coverslides were imaged by TIRF on a Nikon TE-2000 inverted fluorescence microscope with laser illumination at 488 nm. Images from a cooled EMCCD camera (Photometrics Cascade) were recorded with MetaMorph 6.0 (Molecular Devices Inc., Sunnyvale, CA, USA) software. Primer extension assays were carried out in solution using pre-annealed DNA constructs (containing a 5 bp single-stranded overhang containing a single 'G' as the next templating base) at 1 µM and in the presence or absence of YOYO-1 at 20 µM. The reactions contained either unlabelled dCTP or ATTO 647N-dCTP (Jena Biosciences, Germany) at 5 µM. After the addition of 5 units of Klenow Fragment (3'→5' exo-) (New England Biolabs, UK), the reaction was

incubated at 37º for 30 minutes. The products were separated on a 20% native polyacrylamide gel, which was then stained for 30 minutes with SYBR Gold at a 1:10,000 dilution in distilled water. The fluorescence of the gel at 525 nm was then photographed while on a UV transilluminator. Single-molecule primer extension reactions were carried out by immobilising unlabelled duplexes as previously described, then incubating the surface with 100 μM unlabelled dTTP in T50 buffer (Tris pH 7.9 + 50 mM NaCl) for 15 minutes to further passivate it. The surface was then incubated with the primer extension reaction mix, as described above, but with ATTO dCTP at 1 nM. After washing the wells extensively with T50 buffer, the surface was imaged as described above. A control reaction was carried out using identical conditions, but omitting the polymerase from the reaction. Fluorescence time traces of intensity were plotted in Matlab after overlaying DD and DA images.

Chapter 5

Molecular Modelling

As a basis for my work, I used an *in silico* model of RNAP during initial transcription that had been previously developed by the Kapanidis Group (D. Duchi, Oxford). This model was constructed using crystal structures for *T. Therm.* that included upstream DNA[9] and the downstream end of the transcription bubble (downstream fork-junction DNA)[10]. I modelled further downstream DNA using 3D-DART[11] and overlapping bases were aligned onto the end of the transcription bubble using PyMol to create a final model that incorporated downstream DNA to a position of +60.

Site-Directed Mutagenesis

Primers for QuikChange II site-directed mutagenesis to replace B'1183 codon with TAG were designed using Agilent QuikChange® Primer Design Program (http://www.genomics.agilent.com/primerDesignProgram.jsp)

Primer Name	Primer-Template Duplex	Length	Tm (ºC)
pVS10tag1183_top	5'-atcaccccggtagacggt**tag**gatccgtacgaagagatg-3'	39	78.6
	||||||||||||||||| ||||||||||||||||||		
pVS10	…caatagtggggccatctgccatcgctaggcatgcttctctactaa…		
plasmid	…gttatcaccccggtagacggtagcgatccgtacgaagagatgatt…		
	||||||||||||||||| ||||||||||||||||||		
pVS10tag1183 bottom	3'-tagtggggccatctgcca**atc**ctaggcatgcttctctac-5'	39	78.6

Reactions were set up according to the manufacturer's protocol in a total volume of 50 µl and supplemented with 3 µl DMSO. The amount of plasmid template was varied from 5 to 91.5 ng while keeping the primer concentration at 125 ng/reaction. Thermal cycling was carried out using the following program:

Program (approx. 12hrs):
- 95º for 1 minute

Cycle 18 times:
- 95º 50s
- 60º 50s
- 68º for 38min (2.5 min/kb * 15.2 kb)

- 68º for 7 min

The reactions were the supplemented with 10 units DpnI restriction enzyme and were gently mixed by pipetting up and down. The reaction was incubated for 37º and the products were ethanol precipitated, taken up in TE Buffer (10 mM Tris HCl, pH 7.5; 1 mM EDTA), and the concentration was determined by absorbance at A_{260}.

The products of the mutagenesis reaction were used to transform ElectroMAX DH5α-E Cells (Invitrogen). Cells were thawed on wet ice, gently mixed, and 20 µl aliquots were added to 1 µl of DNA solution at specified amounts (ng). The mixture was pipetted into a chilled 0.1cm cuvette and electroporated using a BioRad GenePulser® II, on the

"Ec1" preset. 1 ml of warm SOC medium was quickly added and transferred to a 15ml tube, which was incubated 37°C for 1 hour while shaking at 225–250 rpm. 100 µl of the culture was then plated on a warmed LB-amp plate and the remaining fraction was concentrated by briefly spinning down, removing ~800 µl supernatant and plating the cells after resuspending them in the remaining ~100 µl. Plates were incubated for 18 hours at 37 ºC and colonies were picked for overnight cultures and miniprep isolation using a Qiagen Miniprep spin kit. To test for successful mutagenesis, 640 ng plasmid DNA was digested with 20 units of BamHI (New England Biolabs) for 45 minutes at 37 ºC. The reactions were then supplemented with 2 µl NEB 6x Loading Buffer, loaded onto a 0.7% TAE-agarose gel and electrophoresed at 100V, 400 mA for 80 minutes. The gel was stained with SYBR Gold, diluted 1:10,000 in MQ water, for 30 minutes and imaged on a laser gel scanner (Perkin Elmer). Quantitation of fluorescence intensity carried out using ImageJ. DNA from further overnight cultures was sequenced on an ABI 3730 DNA Analyzer by SourceBiosciences (Oxford, UK) using tiling primers, below:

Oligo	**5'-Sequence-3'**
pVS10seq1	AACAATTCCCCTCTAGAAATAATTTTG
pVS10seq2	TATGTGCCGGCTTCTACCCG
pVS10seq3	GCCTGGAAAACTGGCCACC
pVS10seq4	GCGAAATTCCGCTCATGACA
pVS10seq5	CATCGCAGGTAAAGTGGTTGC
pVS10seq6	GGAAAACCAGTTCCGCGTTG
pVS10seq7	GGATGAAGAAGGCCACTTCGTAGA
pVS10seq8	TGAACTGGCGCTTGGTCAGA
pVS10seq9	GAAGAAAGACCTGTCTGAAGAACTGC
pVS10seq10	CGACAAGATCAACGCCATGC
pVS10seq11	AAGTCTGATGACGTGAACGGTCG

pVS10seq12	CGGCGTTGAAGTGACCCAG
pVS10seq13	ATGGTGGTCGTTTCGCGACT
pVS10seq14	TGAACCGTGCACCGACTCTG
pVS10seq15	ACTCCATCGTCAACCAGGCG
pVS10seq16	CAATCACCGCGAACTTCCGT
pVS10seq17	GTCCATCGGTGAACCGGGTA
pVS10seq18	CAGGTACCGATATGCCAGCG
pVS10seq19	TCGTTCGTCAGATGCTGCGT
pVS10seq20	ACGTAAAAACCCGCTTCGGC

Reads were aligned to the pVS10 plasmid reference sequence using NCBI *blast* and coverage statistics were analysed using Microsoft Excel and plotted using OmniGraphSketcher.

Cotransformation, Protein Expression, Labelling, and Purification

BL21-AI chemically competent cells (Life Technologies) were cotransformed with 50 ng pVS10 and 50 ng pDULE2 pCNF RS plasmid following the manufacturer's standard heat-shock protocol. I plated 50 µl (10%) of the transformation reaction onto LB-agar plates that had been supplemented with spectinomycin at 100 µg/ml and carbenicillin at 100 µg/ml. The plates were incubated overnight and colonies picked for overnight culture in LB containing spectinomycin at 100 µg/ml and carbenicillin at 100 µg/ml. Cultures were used to generate glycerol stocks, perform minipreps to verify plasmid content, and as starter cultures for protein expression.

Commercially-available Novagen OvernightXPRESS Instant Media with supplemented arabinose and antibiotics: 500 ml media was supplemented with arabinose at 3.3 mM, carbenicillin to 100 µg/ml and spectinomycin to 100 µg/ml and inoculated with a starter culture. 40 minutes after inoculation, the culture was supplemented with 103 mg 4-azidophenylalanine and left overnight at 37º, shaking at 220 rpm. The protocol for

RNAP purification is described in the literature[12]. Briefly, cells were harvested after 17.7 hours and lysed. Cleared lysate was incubated with Ni-NTA slurry, washed, and eluted with imidazole. A 2 ml aliquot of protein at 8.1 µM (by A_{280} with an excinction coefficient of 198500 $M^{-1}cm^{-1}$) was supplemented with 16.25 µl of Cy5-DBCO (Jena Biosciences, Germany) in DMSO at 10 mM. The reaction was incubated in the dark at room temperature for 1 hour, and the crude reaction was loaded onto a 5 ml Heparin-Sepharose column (GE Heathcare, UK) and purified on an ÄKTA FPLC (GE Healthcare, UK) by eluting with a salt gradient as described[12]. Fractions were dialysed into RNAP storage buffer and stored at -20 ºC. An unlabelled aliquot of protein was subsequently purified in the same manner and purified fractions were analysed by SDS PAGE with staining by silver (Silver Stain Xpress, Life Technologies, UK) or Coomassie. Gel bands were excised and analysed by trypsin digest and sent for LC/MS analysis. MASCOT software (Matrix Science, UK) was used to identify the proteins in each band.

In vitro transcription Assay

Abortive transcription assays were carried out by A. Tomescu (Oxford) using the method described by Cordes *et al.*[13] and earlier by Hsu[14]. Briefly, DNA-RNAP open complex was formed using a double-stranded wt lacCons+2 promoter

Wild-type lacCons+2

```
                                    ↱
5'AGGCTTGACACTTTATGCTTCGGCTCGTATAATGTGTGGAATTGTGAGAGCGGATAACAATTTC
3'TCCGAACTGTGAAATACGAAGCCGAGCATATTACACACCTTAACACTCTCGCCTATTGTTAAAG
```

and challenging with heparin to isolate RNA open complexes. Radiolabelled ApA dinucleodtide was used to initiate transcription, which was also supplemented with unlabelled ATP, CTP, GTP, and TTP. The resulting abortive RNA transcripts were separated by denaturing PAGE and imaged by exposing X-ray film to the DNA. Overlapping films were used to cover the large area of the gel.

Single Molecule FRET Assay

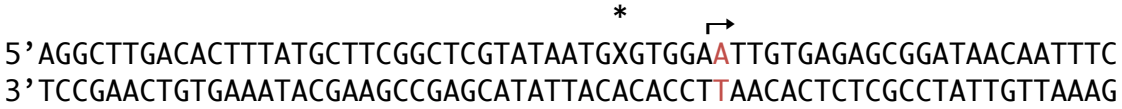
Single-molecule experiments were carried out by D. Duchi and L. Fernandez (Oxford) on a custom-built TIRF microscope, incorporating alternating-laser excitation (ALEX) of green (532nm, 0.43mW) and red (637nm, 0.16mW) lasers, as described previously[13]. Fluorescence emission is separated into two channels and imaged on an EMCCD camera (iXon+, BI-887, Andor) at 100ms (ALEX). Data was processed using in-house software (SENECA, K. Gryte, Oxford) to produce statistics for the FRET (E*) and stoichiometry (S*) values of all molecules in a field-of-view over all frames of each acquisition movie. Microscope coverslips were prepared for single-molecule imaging as described elsewhere[15]. Briefly: coverslips were cleaned, and then PEGylated, incorporating biotin groups sparsely on the surface by inclusion of PEG-biotin. Neutravadin was deposited on the surface, ready for deposition of biotin-labelled species for smFRET TIRF imaging.

For DNA-only experiments, wt lacCons+2(-29/+25) promoter DNA containing a biotin at the 5' end of nontemplate strand was internally labelled with Cy3b and Atto647N at (-15/+15) positions, respectively. The DNA was diluted to 20 pM and deposited on the surface by incubating for 5 minutes. For DNA-RNAP experiments, open complexes were formed on wt lacCons+2(-29/+25) promoter DNA which was internally labelled with Cy3b and Atto647N at (-15/+15) positions for wt RNAP experiments or at (-5) with Cy3B for β'1183-Cy5-RNAP experiments, as described above. The DNA sequences used were:

(-15/+15) Labelled lacCons+2



(-5) Labelled lacCons+2



Complex immobilisation was performed using the SiMPull assay[16]. Briefly: a Penta·His antibody (Qiagen) was deposited on a clean, PEGylated glass coverslide, excess antibody was washed away, and complexes were incubated on the surface until sufficient density was observed by TIRF microscopy.

References

1. Marie, R. *et al.* Integrated view of genome structure and sequence of a single DNA molecule in a nanofluidic device. *Proc Natl Acad Sci USA* **110,** 4893–4898 (2013).
2. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).
3. Bauer, D. L. V., Marie, R., Rasmussen, K. H., Kristensen, A. & Mir, K. U. DNA catenation maintains structure of human metaphase chromosomes. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gks931
4. Young, B. D., Ferguson-Smith, M. A., Sillar, R. & Boyd, E. High-resolution analysis of human peripheral lymphocyte chromosomes by flow cytometry. *Proc Natl Acad Sci USA* **78,** 7727–7731 (1981).
5. Cram, L. S., Bell, C. S. & Fawcett, J. J. Chromosome sorting and genomics. *Methods Cell Sci* **24,** 27–35 (2002).
6. Rasmussen, K. H. *et al.* A device for extraction, manipulation and stretching of DNA from single human chromosomes. *Lab Chip* (2011). doi:10.1039/c0lc00603c
7. Muschielok, A. *et al.* A nano-positioning system for macromolecular structural analysis. *Nat. Methods* **5,** 965–971 (2008).
8. Selvin, P. R. & Ha, T. *Single-molecule techniques : a laboratory manual*. (Cold Spring Harbor, N.Y. : Cold Spring Harbor Laboratory Press, 2008).
9. Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O. & Darst, S. A. Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science* **296,** 1285–1290 (2002).
10. Zhang, Y. *et al.* Structural basis of transcription initiation. *Science* **338,** 1076–1080 (2012).
11. van Dijk, M. & Bonvin, A. M. J. J. 3D-DART: a DNA structure modelling server. *Nucleic Acids Res.* **37,** W235–9 (2009).
12. Anthony, L. C., Artsimovitch, I., Svetlov, V., Landick, R. & Burgess, R. R. Rapid purification of His(6)-tagged Bacillus subtilis core RNA polymerase. *Protein Expr. Purif.* **19,** 350–354 (2000).
13. Cordes, T. *et al.* Sensing DNA opening in transcription using quenchable Förster resonance energy transfer. *Biochemistry* **49,** 9171–9180 (2010).
14. Hsu, L. M. Monitoring abortive initiation. *Methods* **47,** 25–36 (2009).
15. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5,** 507–516 (2008).
16. Jain, A. *et al.* Probing cellular protein complexes using single-molecule pull-down. *Nature* **473,** 484–488 (2011).