`

HIGH RESOLUTION OPTICAL DNA MAPPING

BY

MURAT BADAY

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biophysics and Computational Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Assistant Professor Charles M. Schroeder III, Chair
Professor Paul R. Selvin, Director of Research
Professor Munir H. Nayfeh
Assistant Professor Sua Myong
Assistant Professor Thomas E. Kuhlman

ABSTRACT

Many types of diseases including cancer and autism are associated with copy-number variations in the genome. Most of these variations could not be identified with existing sequencing and optical DNA mapping methods. We have developed Multi-color Super-resolution technique, with potential for high throughput and low cost, which can allow us to recognize more of these variations. Our technique has made 10–fold improvement in the resolution of optical DNA mapping. Using a 180 kb BAC clone as a model system, we resolved dense patterns from 108 fluorescent labels of two different colors representing two different sequence-motifs. Overall, a detailed DNA map with 100 bp resolution was achieved, which has the potential to reveal detailed information about genetic variance and to facilitate medical diagnosis of genetic disease.

For my parents, grandparents and my lovely wife,

ACKNOWLEDGEMENTS

Obtaining a PhD degree is only possible with guidance and support, therefore I recognize with deep appreciation all those acknowledged below and those whom I mistakenly forgot to acknowledge.

With insufficient gratitude I thank my parents, Latif and Salatin Baday and grandparents Resul and Gulevatin Baday who sacrificed so much for my education and who have constantly been my mentors and role models. They have continuously given me courage and guidance with endless love and constant support.

I have much appreciation for my wife, Rachel for her love, understanding, patience and constant support during my PhD.

I thank my parent's in law, Bessie and Mack McCormick for their motivation and support throughout my PhD and for reviewing my publications and thesis.

My siblings, Meryem, Sefer, and Fatih also inspired me with funny and encouraging conversations.

My dear children, Yusuf and Yunus cheered me up during my most busy and difficult times.

It is with significant gratitude I acknowledge my advisor, Paul R. Selvin for his great mentorship and support. He guided me with all his knowledge and experience in biophysics and he helped me to become an independent scientist.

Bob Clegg (Deceased) offered his mentorship and support during my PhD. He has given me exceptional advice for making decisions in my academic life. I always found great guidance from him regarding optics and microscopy.

Ming Xiao shared his in depth knowledge on Genomics and DNA mapping methods and for providing most of the DNA samples to test with our methods.

Ahmet Yildiz and Hamza Balci inspired me and introduced me to biophysics and Paul Selvin.

My labmates, Hyoengjun Kim, Sanghak Lee, and En Cai discussed and shared exciting new ideas and knowledge.

My old lab-mates helped me with technical and scientific issues with my research, Erdal Toprak and Comert Kural thought me the fundamentals of microcopy. Hamza Balci, Melikhan Tanyeri, and Hasan Yardimci gave me new insight and direction in my Research.

Sheyum Syed, Jeff Reifenberger, Ben Blehm, Mindy Hoffman, Paul Simonson, Pinghua Ge, Janet Sheung, Tobias Rosenkranz, Marco Tjioe, Yong Wang and Ruobing Zhang, all helped me during different stages of my research.

A special thanks to Ben Blehm, Mindy Hoffman, Pallavi Thiagarajan and Hannah DeBerg for their in-depth proofreading of my thesis and publications.

I have learned a lot of microlithography and nanolithography from Askin Kocabas, Fatih Degirmenci, and Taner Ozel. They also contributed by preparing several samples.

I had important discussions about research with UIUC alumni Burak Okumus, Salman Syed, Seongjin Park, Ruobo Zhou, Sinan Arslan, Rustem Khafizov, and Taejin Lance Min.

Cindy Dodds gave me essential administrative support and help.

My appreciation to Professors Paul Selvin, Bob Clegg, Charles Schroeder, Sua Myong, Munir Nayfeh and Thom Kuhlman for serving as my preliminary exam and defense committee members.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 Genomic Variations

More than 99.9 % of the human genome is identical across all people. Small differences in DNA lead to significant differences in human phenotypes. Two different people's DNA may differ one in every 1000 DNA bases. Genomic DNA differs because of mutations, mostly errors in copying DNA. Mutations in a sperm or egg can be passed on to offspring and future generations. All 46 chromosomes may have variations and these discrepancies vary from one chromosome to another. Until recently majority of DNA was thought to be junk and non-functioning[1]. It was also called non-coding DNA[2]. Research by the ENCODE project, which is supported by the National Human Genome Research Institute concluded that over 80% of DNA in human genomes has some significant function[3]. This non-coding DNA has now been discovered to be important in the genetic regulation of proteins such as transcription factors. Typically, variations in most of the genome would affect regulatory elements as well. Currently, it is not very well known how much these variations affect phenotype. There could be a wide range of phenotypic effects; physiological, psychological, and physical traits. Physical traits could be things like height, body shape, and eye and hair color; psychological could be being introverted or extroverted in personality, tendency for depression or attention disorders; physiological could be resistance to disease, low or high metabolism for digesting certain nutrients etc. However, genetic factors along with experience and learning play roles in intellectual, social and personal attributes. For example, genetic changes may not affect phenotype as in the example of the study, which showed that

removing about 0.1% of mouse genome did not cause any apparent changes in phenotype[4]. Despite this study, there may be unnoticed changes in phenotype of these mice, requiring more control experiments to ascertain. Many times, variations could be carried to following generations and may lead to susceptibility to diseases in offspring.

Biodiversity is the result of these genetic variations. Genetic diversity would help populations to adapt to environmental changes and is found in all species[5]. As an example, genetic diversity in farming due to selective breeding could help to produce more yield and desired products as well as high resistance to disease. It can also provide more resistance to disease in human being. Genetic variation may also lead to genetic polymorphism, which is defined as the presence of two or more different phenotypes in the same species. Even though genetic variation may enrich the quality of populations and extend life expectancy in the case of certain genetic polymorphisms, some variations may cause more problems in human life and lead to susceptibility to certain diseases. Since we have very complex traits, changes in genes may also result in malfunctioning in several different ways. Genetic risk factors are the effect of certain genes on the tendency to develop certain diseases. The likelihood of individuals having diseases and sharing certain similar traits is complex and still unknown because of the many unknown genetic factors. Despite the unknown factors, recent advances in sequencing technology and sequencing large populations have helped to identify genetic variations related to some traits, especially those associated with human health. Personalized medicine, the practice of medicine based on a person's genomic profile will be more popular as we discover more of the genome and the connection between genes and their relationship to traits.

Three major sources of genetic variation are mutations, gene flow form one population to another, and genetic exchange through sex. Mutations can be positive, negative, neutral, or a mix

as we mentioned above. They could be harmful, either leading to serious disease, improve certain traits or not have any affect at all. Mutations in non-productive cells may not be passed onto offspring, but may still cause disease in the person they appear in. Mutations in productive cells, like eggs and sperm, may result in big changes in phenotype. Even single point mutations may cause serious problems. There are several reasons why mutations occur. While cells divide, DNA might end up not being correctly copied and leading to mutations, one of the most common reasons/factors for mutations. External factors could also create mutations such as, exposure to radiation, chemicals, additives in food, and drug over dosage. A specific example, long time exposure to Sun, causes mutations in skin cells leading to skin cancer[6]. Another reason for genetic variations is gene flow, which is the migration of genetic elements from one population to another[7]. Gene flow may introduce new genes to other populations as in the case of pollen. Also, sex can create new genes through genetic shuffling between reproductive cells. It's a well-known fact that children can get different features from the maternal and fraternal sides. For example, the shape of child's head might resemble its father but have some other trait such as eye or hair color similar to its mother. But, it is also possible to end up with a completely new countenance not matching any parental features.

Genomic variations can vary from single nucleotide polymorphism (SNP) to abnormalities in chromosomes. Genome structural variations, changes in structure of an organism's genome, have several different forms, such as insertions, deletions, duplications, copy-number variants, inversions and translocations. These may range in size from a few hundred basepair to several megabases.

Single Nucleotide Polymorphism is a single nucleotide-A, G, C or T- change in the genome. Most common SNPs have only two alleles (two variants, each with a different nucleotide in a

specific spot). SNPs are the main reason for genetic diversity. Since each individual has different SNPs in their genome, it is used in DNA fingerprinting for forensic science. Synonymous and non-synonymous SNPs are types of SNPs in the coding region of the protein. Non-synonymous SNPs alter the protein sequence while synonymous don't. As in all types of genomic variations, SNPs in non-coding regions may affect transcription factors, RNA expression and gene silencing. The international HapMap (haplotype map) Project was organized to discover SNPs in populations around the world. The project started in fall 2002 and recent results were released with the Phase III dataset in spring 2009. HapMap 3 included 1184 individuals from 11 populations worldwide. Overall, 1.6 million common SNPs were genotyped in this project[8]. As of last year, the NCBI SNP data base listed over 180 million SNPs in humans. Many human diseases are associated with SNPs such as cancer, AIDS, sickle cell anemia, and cystic fibrosis[9–13]. The project also included research about Copy Number Variations. Some SNPs increase susceptibility to disease as in case of ApoE gene (Apolipoprotein E). Mutation in the ApoE is related to risk for Alzheimer disease[14]. SNPs in PEMT gene (phosphatidylethanolamine N-methyltransferase gene) leads to choline deficiency, associated with organ dysfunction[15]. In another study, post-menopausal women who have 1958A gene allele of MTHFR gene (Methylenetetrahydrofolate reductase) have 15 times more susceptibility to organ malfunction with choline deficiency[16]. All these studies relating nutrition to genomics are good examples of genomic variations which may play important roles in prevention of some diseases[17].

Chromosome abnormality in the large scale structure of the genome is easy to detect. It could be an abnormality in one or more chromosomes[18]. Karyotyping is the method to compare individual's chromosomes with normal chromosomes. Changes in chromosome number occur in cell division after meiosis or mitosis. These changes are irreversible. Studies show that 15-20 %

of human oocytes may have chromosome abnormalities[19].Changes in number of chromosomes is frequently referred to as Aneuploidy. In general, missing or additional chromosomes are responsible for defects during the embryo stage. Most of these irregularities result in miscarriage[20]. Nearly all spontaneous abortions are the known to be result of chromosomal abnormalities. There could be a missing chromosome or multiple chromosomes compared to regular diploid chromosome numbers. Over 60 diseases are related to chromosomal abnormalities, the most common ones are Klinefelter syndrome (47, XXY, extra X chromosome), Turner Syndrome (45, X, missing Y chromosome), Patau syndrome (Trisomy of chromosome 13). Another well-known aneuploidy is Down syndrome (Trisomy of chromosome 21). Some inherited syndromes may cause predisposition to cancers in addition to the direct aneuploidy in tumor cells[21]. There are many examples of aneuploidy in cancers such as leukemia which is associated with a copy increase of chromosome 8[22–24].

Deletions. Structural abnormalities in chromosomes have several forms as mentioned above. So far, we stated chromosomal level abnormalities such as the case of a lost or duplicate chromosome. There are other kinds of abnormalities related to particular portions of a chromosome. Deletions are one of the possible forms of structural abnormalities as in case of Jacobsen syndrome, where a terminal region of chromosome 11 which generally has a specific band 11q24.1 is missing. This syndrome is responsible for several physical abnormalities in the face, heart defects as wells as intellectual problems[25]. Deletions can occur from either the end or interior of a chromosome. There can be microdeletions of up to 5 kb regions of a chromosome which may include a few genes. In most of these chromosomal abnormalities, especially deletions, if it includes many genes, fetuses will spontaneously abort. Some short deletions in DNA sequences may alter protein synthesis and may result in the loss of important proteins. There are also deletions in telomere regions of the

genome which may subtly affect intellectual disability[26]. As we will point out in later parts of the chapter, it is hard to identify short deletions with current sequencing methods. Recently, new evidence of microdeletions effect on phenotype were discovered. A recent publication proposes that the 5q31 microdeletion leads to intellectual disability associated with changes in expression pattern on LRRTM2[27]. As mentioned earlier, several investigations have connected the link between chromosomal deletion and cancer. According to Schimenti et al., the loss of the gene NF1 in gliblastoma is observed in 28 percent of human breast cancers[28]. Another study shows deletions of the MTAP gene in a subgroup of endometrial cancer[29].

Inversions are another form of chromosomal abnormality. Inversion is the rotation of a gene 180 degrees before re-insertion at the same site. There are two types of inversion depending on the location of chromosomes: Paracentric and Pericentric Paracentric inversions. Paracentric inversions occurs in one arm of the chromosome does not include centromere. Pericentric covers centromeres and leaves breaks in each arm. Carriers of paracentric inversions may rarely give birth to phenotypic problems. On the other hand, occurrence of abnormalities is much higher in the case of pericentric inversions. Most inversions are balanced inversions, where the genetic sequence is not altered. It does not affect overall genetic material and therefore does not result in phenotypic abnormalities. Most well-known inversions occur in chromosomes 4 and 8. The inversion on chromosome 8 is 3.5 Mb and reported in 26% of health subjects. The one on chromosome 4 was observed in 12.5 % of controls and was about 6 Mb in size[30]. These two inversions do not have known phenotypic consequences. A recent study showed that a 4 year-old patient with developmental delay problems due to balanced inversions unlike other findings (inversion of 14(q21q23)[31]. Another study performed in Poland indicates that balanced chromosomal rearrangements may be associated with abnormal phenotypes[32]. Unbalanced inversion means that

sequences of proteins or heir copy numbers are changed. Unbalanced inversion which disrupts sequences of proteins may change the expression of several proteins and this causes significant mutation/deformity. Certain inversions are connected to numerous diseases, such as the factor VIII gene with hemophilia[33]; IDS gene with Hunter syndrome; and the emerin gene with Emery-Dreifuss muscular dystrophy. Problems may also occur in somatic cells, leading to cancer as in the case of other abnormalities. Thyroid cancers in patients who are exposed to Chernobyl radiation are linked to a 500 kb inversion on chromosome 10[34,35]. Current array based technologies can't locate many balanced inversions. They are estimated to be the most common inversion in the human genome. Many of these may be associated with diseases and disorders[36]. Currently, there are 914 inversions reported in the database. Most known inversions are in the range of 10 kb to 100 kb. Many may not have been discovered due to the resolution limit of current array base technologies. In fact, smaller sized inversions could be counted as copy number variations.

Translocations are another chromosome abnormality caused by relocation of parts of chromosomes. Similar to inversions, translocations could be balance or unbalanced depending on how changes in genome affect changes in genetic codes. There are two classifications of translocations, reciprocal and Robertsonian. Reciprocal are exchanges of DNA parts between nonhomologous chromosomes. This variation may occur in one in 500 newborns[37]. Most reciprocal translocations are balanced and do not result in abnormal changes. There may be hidden results of reciprocal translocations as well as inversions due to their indirect effect on phenotypes. Some scientists think that autism could be the result of such abnormalities in the genome. Robertsonian translocations are the rearrangement of two acrocentric chromosomes by joining together by their whole long arms. It is basically nonreciprocal translocation and it could include either homologous or nonhomologous chromosomes. In particular, Down syndrome is caused by

Robertsonian translocation of chromosome 21's long arm onto chromosome 14's long arm[38]. Robertsonian translocations could be the result of malformation of genes in eggs and that are effected during reproduction. Most common forms of this translocation are between chromosomes 14 and 15, 14 and 21, and 13 and 14. Several cancer types are known to be associated with translocations. Translocations are found in many diseases such as Mantle cell lymphoma[39] (t(11;14)(q13;q32)) and Schizophrenia[40] (t(1;11)(q42.1;q14.3)). Some translocations or deletions may show up in offspring but may also be passed on to following generations. Due to that fact, it is recommended to genetically test parents before getting pregnant.

Copy Number Variations (CNVs) are most common type of genomic variation. CNVs are changes in the genome to having multiple copies of one or more parts of the chromosome. Aneuploids may be considered large scale CNVs. CNVs are known to occur in 12% of the human genome. CNVs in a single chromosome may vary from 6 to 18% of the chromosome. Sizes of known CNVs may range from one kilobase to megabases[44]. Almost 3000 genes are known to have copy number variants which amounts to 10% of all genes in the genome. One or multiple genes may correspond to a copied DNA fragment. Thus, these genes may be expressed in multiple copies. Smaller repeats, less than 1 kb, are even possible but not identified easily with current sequencing methods. Copies may also correspond to non-coding regions of DNA which may be important to other regulation mechanisms in cells. It is now believed that CNVs are more frequent (at least three times) than SNPs because of the recent discovery of small size copy number variations[41]. Genomes of unrelated people may differ 0.4% due to copy number variations[42]. Deletions, translocations and inversions may also be called copy number variations. Large numbers of CNVs were discovered after the completion of human genome project[43,44] Recent studies show that 17 conditions of the nervous system may be affected by CNVs, including Alzheimer's and Parkinson

diseases. Having CNVs can either cause susceptibility or resistance to disease. Studies have shown that cancer cells have increased copy numbers, as in the case of EGFR in non-small cell lung cancer, which was found to have a higher copy using the FISH method[45]. Susceptibility to HIV infection is also associated with copy number of the CCL3L1 gene[46]. Many studies shows links between autism and copy number variations[47]. A study by Sebat et al. has utilized multiple detection methods to identify copy numbers in autistic patients. Their methods are comparative genome hybridization (CGH), paternity testing, fluorescence in situ hybridization, cytogenetic and microsatellite genotyping. Some of these methods will be explained later in the thesis. Overall, they have linked 10% percent of patients with sporadic autism[48]. A very recent study by Gai et al. included new variations which could cause autism[49]. Schizophrenia is also associated with copy number variants at 1q21, 15q11.2, 15q13.3, 16p11.2, 22q12, and Neurexin 1 loci according to several studies[50,51]

1.2 High Resolution Imaging by Using Fluorescent Microscopy

Fluorescent imaging microscopy has been an essential tool for biologists for many years, especially after the discovery of high yield fluorophores which can be used to tag DNA, RNA and biological molecules. Recent improvements in optical resolution in fluorescent microcopy have led to study of structures at a finer level of details than previously possible[52–62]. Before such advances in optical microscopy, electron microscopy has been used to inspect structural details. However, electron microscopy requires samples fixation that can change native state of the system under study. Furthermore, it is also difficult to label and image the target inside a cell with electron microscopy. For instance, it is not yet possible to label nucleotides in situ separately to resolve individual nucleotides and carry out sequencing. Hence, fluorescence microscopy is becoming

more popular in the life sciences[57], especially after the development of relatively easy methods for target labeling of proteins and nucleic acids even inside living cells.

In general terms, image resolution is a measure of the amount of information in reconstituting the image of an object. Therefore, higher resolution means being able to see more details in the image. In optical microcopy, resolution is determined by the smallest distance at which two closely spaced particles can be distinguished. Raleigh observed that resolving two lines in a diffraction grating is limited when the diffraction peak of the first line overlaps with the first minimum of the second. Two peaks become indistinguishable when distances between lines are closer than this. For a microscope system with a numerical aperture, NA, and observation wavelength lambda, the Raleigh limit is determined by

$$R = \frac{1.22*\lambda}{2*NA} \qquad [1]$$

Thus, the lower limit of resolution is approximately 200 nm for visible light (λ~500 nm), if using an objective with a numerical aperture of 1.45.

Especially in biology, there are very specific questions related to distances between two or more molecules. Answering such questions has led to an increased understanding of the mechanism of several molecular systems. An important factor in resolving such distances is localizing molecules with high precision. The localization error should be below the expected distances between two molecules[63]. The localization error for single molecules could be as low as 1.5nm which is achieved in our lab by using the technique called FIONA[64]. Last decade, new imaging techniques have emerged based on FIONA. Super resolution could be achieved by STORM and PALM techniques[52,55,59] which facilitate the localization of densely labeled photo switchable fluorophores. The fluorophores are sequentially activated and imaged. Super-resolution

image is reconstructed from localization of the fluorophores after repeated cycles of this process. Further improvements have been made to understand behavior macromolecules in vivo[65–69]. Super resolution optical fluctuation imaging (SOFI), Single-Molecule-Based Super-Resolution Images in the Presence of Multiple Fluorophores (PHILM), Direct stochastic optical reconstruction microscopy with standard fluorescent probes (dSTORM) which are able to resolve densely labeled fluorophores and locating them individually over long time period. All these techniques are able provide high resolution distribution of the interested particles in cell. FRET (Forster Resonance Energy Transfer) has been most common method to understand mechanisms of interaction between two or more molecules[70], but it is limited by 5-10 nm distance range. Several other methods has been used for co-localization of fluorescent probes within resolution range of 10-200 nm[53,56,71,72].

Single molecule-High Resolution Imaging with Photobleaching (SHRImP)[73] and Single-molecule High-Resolution Co-localization (SHREC)[74] have been combined to study high resolution imaging of DNA and RNA molecules. SHRImP can resolve adjacent fluorophores of the same color with 10nm resolution by using the quantal photobleaching behavior of single fluorescent dye molecules[73]. SHREC can measure distance between fluorophores in macromolecular systems with better than 10 nm resolution by using two chromatically different fluorophores. Fluorophores are imaged in separate channels by using a dual-view system and localizing individually with high accuracy. SHRImP can be extended to 3 or more dyes. Using more colors and the multiple fluorophores for each color will increase the number of resolved distances between individual molecules and provide more information about their structure. Because of their feasibility and potentials to improve further, we focused on combining these two techniques to work on our DNA mapping application.

1.3 Imaging Techniques

1.3.1 FIONA (Fluorescence Imaging with One Nanometer Accuracy)

Our lab has achieved nanometer accuracy in locating fluorescence dyes by combining total internal reflection microscopy and improving photon collection[64]. Total internal reflection microscopy is used to obtain high signal to noise ratio which is crucial for getting higher resolution. TIRF based on excitation and detection of fluorophores in a small volume of sample solution. When the incident light is totally reflected at the interface between water and glass, occurring evanescent wave illuminates and excites fluorophores in the solution that is adjacent to the interface. The evanescent field decays exponentially and typically penetrates about 100 nm into the sample (Figure 1.1). Snell's law is used to calculate depth of evanescent field. The critical angle is the lowest angle of incidence which total internal reflection starts. Critical angle can be calculated by equation [2] in which **n(2)** is the refractive index of solution or buffer where the sample is while **n(1)** is the refractive index of glass slide or coverslip. Total internal reflection occurs at angles greater than critical angle.

$$\theta c = \sin^{-1}\frac{n(2)}{n(1)} \qquad [2]$$

Despite no light propagates to other medium (which is buffer), small amount of light pass across the interface which is called evanescent field. Energy of evanescent field decreases by exponential in decay in the z direction (which is direction perpendicular to the interface) as shown in equation [3].

$$E(z) = E(0)e^{-z/d} \qquad [3]$$

E(0) is the energy at the surface, **E(z)** is the energy at distance **z** from the surface, **d** is the penetration depth. The penetration depth depends on the angle of incidence, refractive indices of the solution and the glass and wavelength of incidence light as shown in equation [4]

$$d = \frac{\lambda}{4\pi \times \sqrt{n(1)^2\ sin^2\ \theta(1) - n(2)^2}} \qquad [4]$$

By this method, fluorophores which are specifically tethered surface can be excited. Emission from non-targeted fluorophores in solution stray fluorescent, as well as Raman scattering from water that causes background is substantially decreased compared to epi-fluorescent microculminating in a much improved signal-to-noise.

In addition to TIRF microscopy, innovations in photon collection are crucial to obtaining nanometer precision. Photon collection is greatly improved by the ability to extend the number of photons a dye will emit before it photobleaches. Photobleaching is irreversible chemical damage of a fluorophore. The localization of a fluorescent dye is inversely proportional to square root of total photon number. The collection of more photons corresponds to a more precise localization. To localize dyes, we fit the emission pattern of a single dye by a Gaussian function. For a given numerical aperture (NA), the width of the Gaussian fit for single spot is determined by equation [1]. We get a diffraction limit 223 nm, when we use 1.4 NA microscope and 532 nm green laser (Figure 1.3). We can achieve nanometer accuracy by localizing position of the dye according with precision given in equation.

$$\sigma_\mu = \sqrt{\frac{s^2}{N} + \frac{a^2/12}{N} + \frac{8\pi s^4 b^2}{a^2 N^2}} \qquad [5]$$

In this equation, **s** is the standard deviation of the Gaussian fit, **a** is the pixel size of CCD camera, b is the standard deviation of background, **N** is the total number of photons collected by the CCD

camera. The last term under the square root is background noise. High signal to noise improves the precision by minimizing this term. Collecting more photons and having long lasting dyes is achieved by using an oxygen scavenging systems.

1.3.2 Single-Molecule High-Resolution Imaging with Photobleaching (SHRImP)

SHRImP was developed in our lab[73]. It is based using two identical fluorophores which has distances between them as small as 10 nm. Conventional light microscopy is limited in resolution to about 250 nm because of the diffraction-limited of light. Fluorophores within diffraction limit appear as a single spot. FRET technique cannot measure distances over 10 nm. With the SHRImP technique, we can bridge the gap between distances resolvable by FRET and distances resolvable by conventional light microscopy.

In SHRImP, the two dyes are too close enough that their intensity profile overlaps. The dyes photobleach stochastically (see Figure 1.4). To perform SHRImP, first, we fit and localize the remaining dye after the other dye has photobleached and then subtract its intensity from the combined intensity profile of two dyes. Then, the second dye are fit and localized with high accuracy. By measuring distance between each dye, two dyes can be resolved on the order of 10 nm. Three or more dyes could be resolved with same method with extended analysis. Gordon et al. resolved two identical dyes which are attached to short DNA pieces at distances of 10, 17, and 51 nm. Balci et al[75] used the SHRImP to measure the inter-head distance (30 nm) of myosin VI molecules that are labeled with GFPs on the actin-binding of domains agreement with hand-over-hand model.

Another method, NAMLS (nanometer-localized multiple single-molecule) microscopy independently developed by Qu et at based on same principle. They resolved distances between

five identically labeled fluorophores on DNA. Later on, multiple methods based on blinking of fluorophores developed to resolve hundreds of dyes within diffraction limit as mentioned above. Main advantage of these methods is using a single color of fluorophores which is very convenient with single laser line of fluorescence microscopy. Besides, not involving chromatic aberration is another advantage of the SHRImP method due to using single color imaging.

1.3.3 Single-Molecule High Resolution Co-localization (SHREC)

SHREC is another technique to measure distances between dyes within the diffraction limit. It is based on FIONA and uses two chromatically different fluorescence probes. SHREC uses two different laser colors to excite the sample. The emission from spectrally distinct dyes is split by using a dual- view apparatus which contains a dichroic, two different emission filters, and related optics (Figure 1.5 and Figure 1.6). The split image can be viewed in a CCD camera in two channels. Mapping between the channels is achieved by using a nano-hole pattern and an advanced mapping algorithm developed by our group. To measure distances with SHREC, spots in each channel are fit and mapped onto each other. After measuring distances between mapped spots, we can get resolution on the order of 10 nm.

Decreasing the mapping error is a major challenge in SHREC experiments. In the original SHREC experiments, Churchman et al[74], used fluorescent beads which have emit light in both green and red channels. They used a piezo nanonstage to move beads by 0.5 μM intervals across the field of view. After fitting intensities and finding x and y positions in both channel, they calculated mapping functions and found a mapping error using the target registration error calculation (Equation 3). We achieve smaller mapping errors by using nanoholes which is designed with chromium and silver on coverglass. 1.4 μm spaced nanonholes were made using Focused

Ion Beam (FIB). We pass white light through the nanoholes and then, it splits into both channels by a dichroic filter. After fitting spots in both channels by Gaussian fit, we calculate a mapping function by comparing corresponding spots in both channels.

Functions f and g are calculated using a locally weighted mean method in equation 3 [74].

$$f(x,y)=\frac{\sum_{i=1}^{N} W\left\{\frac{\left[(x-x_i)^2+(y-y_i)^2\right]^{1/2}}{R_n}\right\} P_i(x,y)}{\sum_{i=1}^{N} W\{\frac{\left[(x-x_i)^2+(y-y_i)^2\right]^{1/2}}{R_n}\}} \quad [3]$$

with $W_i(R)=1-3R^2+2R^3 (0\leq R\leq 1)$, $W_i(R)=0 (R>1)$ where $R=[(x-xi)^2+(y-y_i)^2]^{1/2}/R_n$

The Mapping error is determined by Target Registration Error (TRE) calculation. The best result we have obtained is 2.3 nm, 30% smaller than the error obtained in the original SHREC experiments. Since most our application needs to resolve distances on the order of 10 nm, 2.3 nm mapping error indicates that SHREC is a very feasible for our application.

1.4 Figures for Chapter I

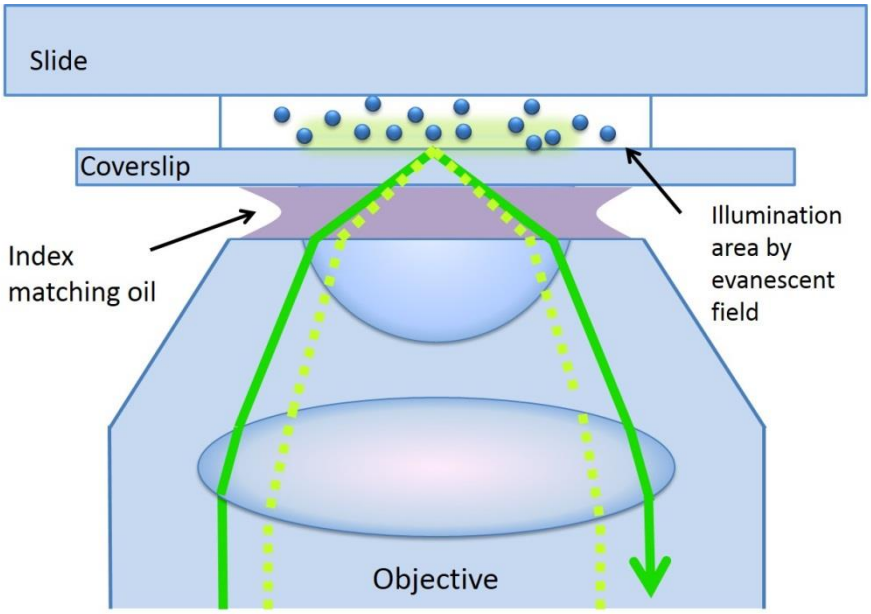


Figure 1.1: Objective type TIRF microscopy, the laser beam is reflected at the interface between the coverslip and sample (sample is illuminated about average of 100 nm in z-depth.)

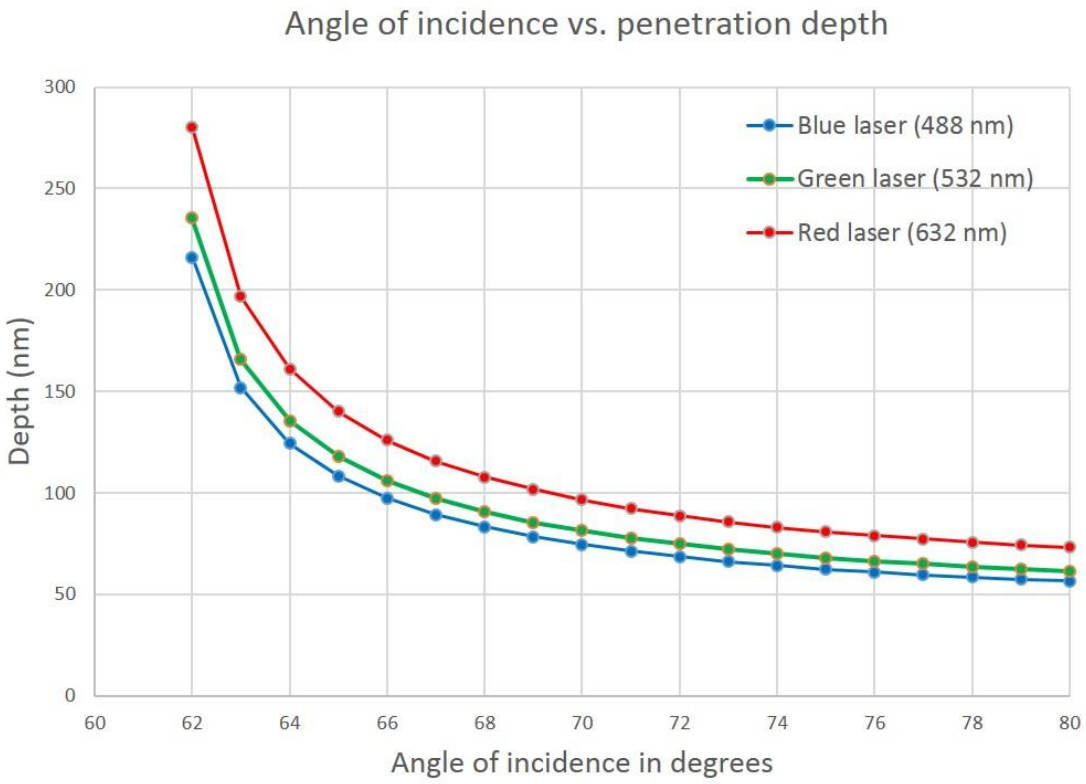


Figure 1.2: Graph for angle of incidence vs. penetration depth. Current Objective TIRF set up ideally gives about 100 nm penetration depth because it is hard to obtain higher angles. Lower penetration depth is desired to get high signal to noise ratio.

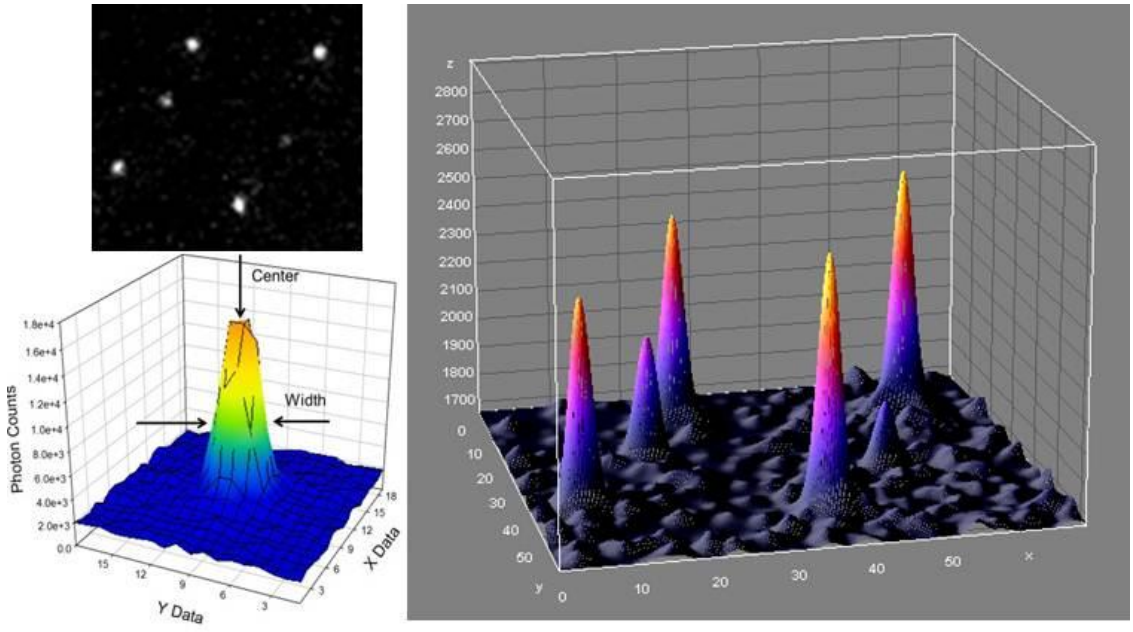


Figure 1.3: Top Left. CCD image of several Cy3 molecules. **Right**. Gaussian profiles of all spots which can easily be distinguished from each other. **Bottom Left.** Gaussian profile of single molecule. The width is ~250 nm, but the center can be located to within ± 1.5 nm accuracy.

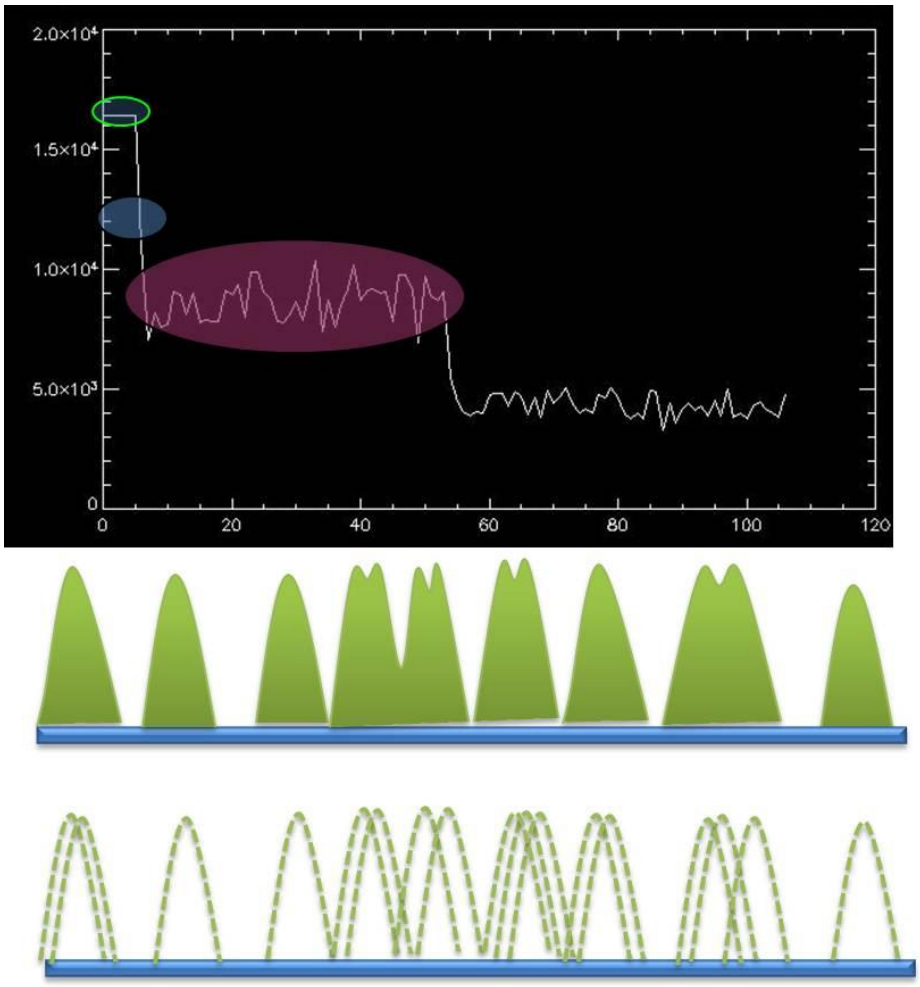


Figure 1.4: Top. A plot of total integrated intensity versus time for three closely spaced Cy3 molecules, showing a three-step photobleaching behavior. **Middle**. Several close spaced dyes. Some of them are within diffraction limit. It is hard to resolve with regular diffraction limited microscopy. **Bottom**. By using SHRImP method, we could locate each individual dyes with high precision.

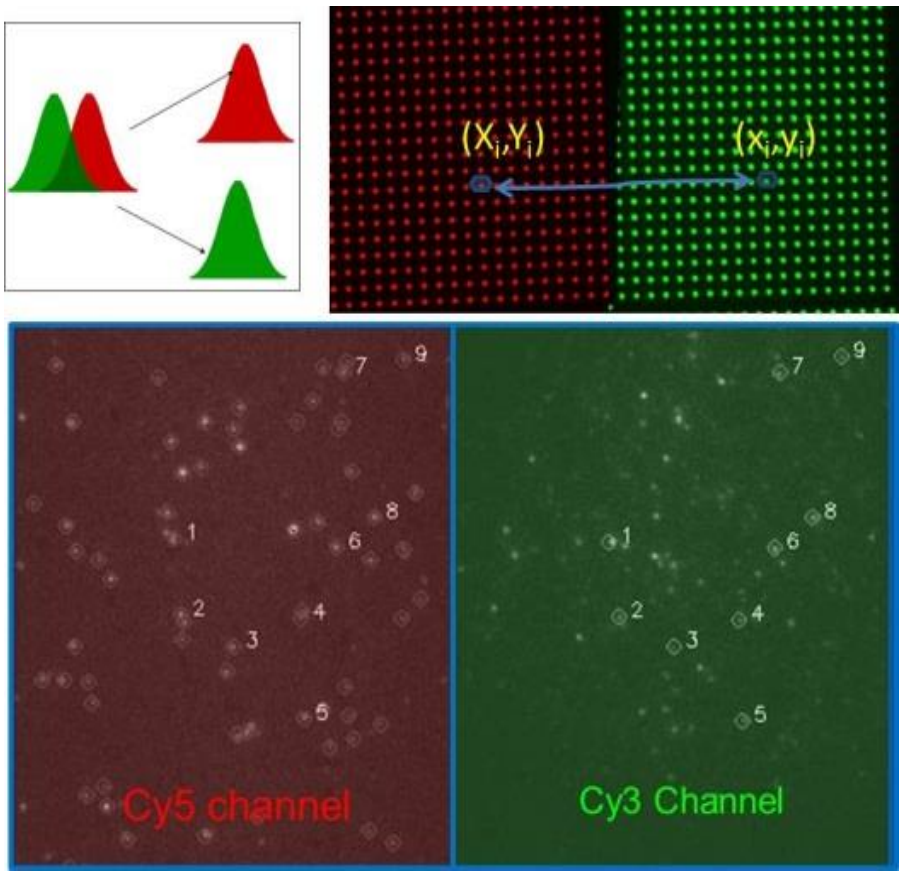


Figure 1.5: In the SHREC technique, dyes of two different colors are split into two channels and localized by using FIONA. **Top left:** Illustration of green and red channel separation with dichroic. **Top right:** Image of nanoholes with both green and red channels. Nanoholes is used as fiduciary marker. **Bottom**: Actual image from DNA which is labeled by Cy3 and Cy5 dyes, then chromatically separated into two channels. Our autospot finder code finds co-localized spots in both channels.

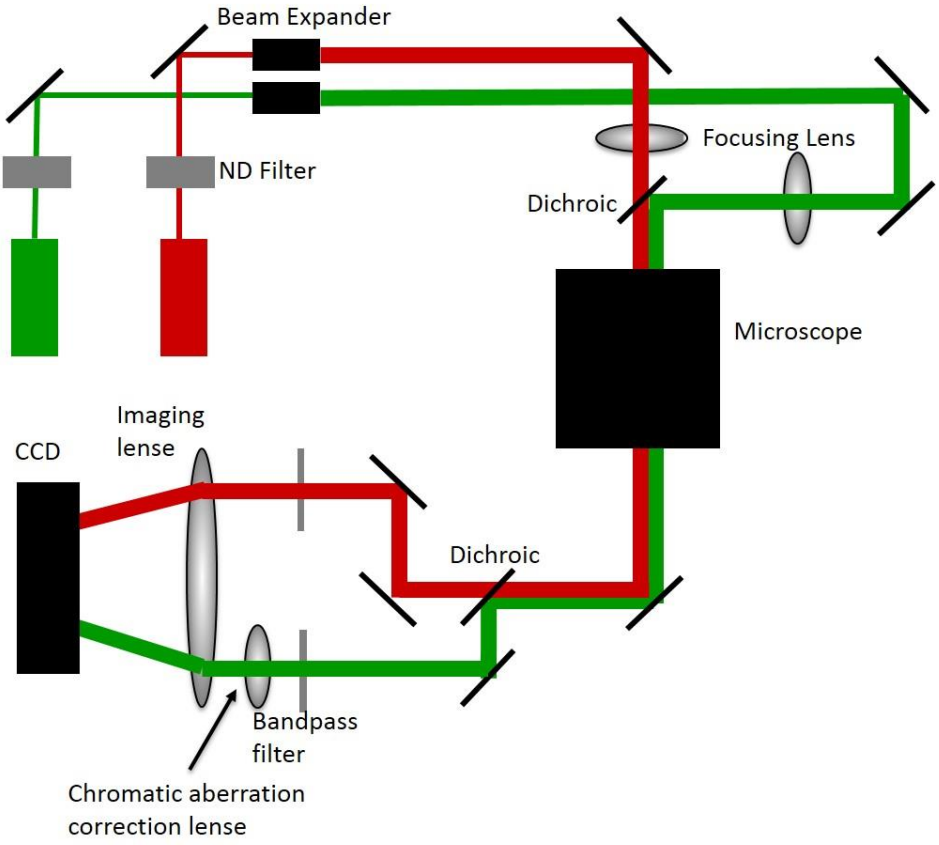


Figure 1.6: SHREC set up (two-color FIONA setup)

CHAPTER II

HIGH YIELD DNA STRETCHING FOR HIGH RESOLUTION SINGLE-MOLECULE IMAGING

Having accurately stretched and fixed DNA is very important for obtaining high resolution imaging with DNA and RNA. Both optical DNA mapping and DNA-Protein interaction studies need a feasible technique that will provide stretching of DNA to its near-contour length and fix it on the surface, so that fluctuations don't cause errors. We have developed a new stretching technique with an easy exchange buffer system, which yields uniformly stretched, high density of DNA on surface. Using chloroquine and functionalized primers, our technique stretched DNA to over 90% of contour length and achieved three times more surface coverage than comparable methods. Three fluorophores within the "standard" resolution limit on stretched lambda DNA were resolved by applying SHRImP (Stochastic High-Resolution Imaging by Photobleaching). This method could be used determine structural variations of long DNA molecules, to profile of the alternative splicing variants of RNA, and to understand mechanisms of DNA/RNA-protein interaction.

2.1 Overview of Stretching Methods

Several new DNA mounting methods were developed last decade to study DNA-Protein interaction studies[76]. Both DNA-protein studies and optical DNA mapping for genomic studies need optimized DNA stretching systems. Such systems should give high yields so reasonable statistics could easily be obtained. On the other hand, stretched DNA should be close to its contour length so that localization of dyes or proteins on the DNA would have minimal fluctuations and positions would map accurately to locations on the DNA. In addition, an easy buffer exchange system is essential to run experiments with several sets of trials on a single sample. Recent

techniques don't have all these critical features together. Our target application DNAs are longer than 1kb. We need to stretch DNA fully to its contour length and fixed it on the surface so that fluctuations don't cause errors when localizing them. We have tried several existing methods and optimized them for our systems of interest. DNA curtains are made of patterns of barriers by using nanolithography[77,78] and based on diffusion of DNAs on lipid-bilayer coated surface. Thousands of DNAs can get stretched and aligned on chromium barriers which are fabricated by using e-beam lithography. After Cr barriers are fabricated on the glass surface, the surface is coated with biotinylated lipid bilayers. DNA with one end attached to a biotin-neutradavidin-biotin complex can easily diffuse over the lipid bilayer without sticking to it. The DNA cannot go the beyond the Cr barriers, hence is stretched (Figure 2.1). However, there are several drawbacks to this technique. Stretching DNA less than its contour length and fluctuations during flow cause limitation for getting super resolution imaging. Another issue is overlapping of DNA strands. This can be minimized by using triangular shaped barriers, but even with this feature, approximately 50% of DNA strands overlap.

We devoted much effort to applying DNA curtains to our project but ultimately found the obstacles presented were too large to warrant continued application of this technique. In particular, there are several obstacles to getting good surface chemistry and run stretching on curtains. First, the modified quartz surface has a lot of unwanted residues from the nanolithography process that makes it difficult to achieve a uniform lipid bilayer. Also, the generation of the Cr 100 nm features with an Raith electron beam is technically difficult and time consuming (Figure 2.2). Furthermore, once a curtain device is fabricated, keeping and reusing it is very difficult since cleaning procedures are likely to remove patterns or causes gaps in the design.

Another stretching technique involves chemically tethering the DNA on the PEG surface and pulling on a large bead attached to the other end by flow. Magnetic beads with anti-digoxigenin are used to stretch DNA[79]. They labeled DNA at one end with biotin so that they could attach DNA to surface with a biotin-streptavidin linkage. The other end of the DNA was labeled with digoxidenin, allowing the attachment of a magnetic bead coated with anti-digoxigenin. They used a 2.8 μm small earth rare magnet (NdFeB) to apply a magnetic stretching force. While beads were moved with the magnet inside of flow chamber and away from surface, they apply a constant flow of fluid to stretch DNA. We tried bead stretching both with magnetic and non-magnetic beads. Tethering beads to DNA is challenging part given that the tethers break easily during flow. The large size of the bead is also an obstacle for high resolution imaging due to tilting labeled DNA at some uncertain angle. We need to stretch multiple DNAs simultaneously for our imaging purposes. We could not get the uniform, multiplexing by bead stretching our experiments would require.

Another technique were developed by Washizu et al. who performed electrostatic stretching of DNA over a decade ago [80,81]. Later on many groups worked on similar techniques. They used vacuum evaporated parallel strip electrodes and attached DNA either with biotin or non-specific to the electrodes. The electrodes were made of aluminum and etched onto the glass surface. It is difficult to obtain single-molecule resolution with this method. We created a similar circuit on the glass surface by using microlithography techniques and applied a 150 V/cm Electric Field to stretch DNA. There was much background from electrodes as well as not able to get uniform fully stretched lambda DNAs.

The Austin group in Princeton University has studied DNA stretching in nanochannels by applying electric fields. They have performed several studies of the effect of a nanochannels's width on the contour length and radius of gyration of DNA. They made channels with cross section

as small as 10 nm by 50 nm[82]. Images are taken by epi-fluorescence. The major drawback of this system is that imaging could not be done with by TIRF microscopy. Thus, we cannot use our high resolution imaging techniques. Even after extensive chip modification performed, we could not get the nanochip system to work with TIRF microscopy. Newer version of the method designed by Reisner et al.[83,84] which can potentially be used with TIRF microscopy, but we could not get high yield stretching due to artifacts while designing chips on fuse silica.

Functionalizing of glass surface with two different chemicals according to Braslavsky et al.[85] and Kartalov et al.[86] is another feasible method which provides high yield stretching. Polymers; Poly(allyamine) (PAII) and Poly(acrylic acid) (PAcr) coats surface with plus and minus charges sequentially. Since DNA backbone is minus charged, DNA is stretched when it is sandwich on glass surface Even though this method is easy, the major drawback in this method is the difficulty to control uniform stretching and that leads to under-stretched or over-stretched DNAs. Due to its feasibility, we have worked with this method to implementation of our multi-color DNA mapping method.

We tried several existing methods and optimized them for our systems of interest, but none appear satisfactory. We finally utilized a recently developed chloroquine stretching technique[87] that allowed us to increase the stretching rate close to its contour length and perform high resolution imaging, which can resolve distances as low as 10 nm.

2.2 New Stretching Method by Using Chloroquine

There have been several DNA stretching studies tethering one end of DNA and applying a constant flow to stretch it. The main challenge of flow stretching is the fluctuation at the free end of the DNA. To fix the problem, we tethered both ends of DNA with biotin on a PEG (poly-

ethyleneglycol) surface. Functionalizing a surface with PEG is essential to get uniform stretching and remove impediments to the stretching process.[70] Before loading DNA on the surface, 0.5 mg/ml neutradavidin is flowed, incubated, and then washed with T50 buffer. After getting biotin-neutradavidin binding, double labeled biotinylated DNA in T50 buffer with chloroquine is flowed in at a rate of 100 ug/min (Figure 2.3). Chloroquine is an intercalating drug that extends the contour length of double stranded DNA by inserting itself in between DNA's base pairs. Lambda DNA was used to test the technique. DNAs were labeled with Sytox green DNA stain and imaged by using TIRF microscopy (Figure 2.4). We get very uniform fully stretched lambda DNA and observed around 30 percent of DNA fully stretched.

Later, we improved the technique further by using lambda DNA biotinylated at only one end. The other end which has a 12 nt overhang is left unlabeled. After attaching lambda DNA on the PEG surface at one end, we flowed in 12 bp biotinylated primers (complementary to the overhang) in the presence of 100 μM chloroquine at a flow rate of 100 μl/min. Hybridizing lambda DNA and labeled primers stretches and fixes DNA at both ends on the surface very uniformly. We could routinely make samples with more than 60 percent of DNA stretched (Figure 2.4).

We decided to test whether high-resolution distance measurements can be performed on well-stretched DNA molecules. To prepare a long DNA substrate containing multiple site specific flurophores, we first amplified a 288 bp DNA template from a SNAP-Tag plasmid (Figure 2.6-a). After melting the double strands of the PCR product, Cy3 labeled primers at specific locations were hybridized (Figure 2.6-b). Primers were then extended by using DNA polymerase to make double stranded DNA (Figure 2.6-c). One primer was designed with a 12 nt overhang which is complementary to the overhang present in lambda DNA. We finally annealed the labeled DNA substrate to lambda DNA molecules (Figure 2.6-d). After annealing, we were left with a long piece

of DNA with three different Cy3 fluorophores located either 94 bp, 172bp, or 266 bp distances from each other. By using the SHRImP technique on stretched DNA molecules, we then measured the distances between Cy3 molecules to be 31, 62, and 92 nm, in good agreement with the expected distances of 32nm (94 bp), 58nm (172 bp), and 91nm (266 bp) (Figure 2.7).

2.3 Discussion

Recent improvements of super-resolution methods have opened a new era on studies of biomolecules. Especially, the implementation of the FIONA technique[64] has led to discoveries of protein mechanisms. Proteins can be localized with 1 nm accuracy and their interaction with other molecules including DNA and RNA can easily be tracked, understood and modeled [75,88,89]. The FIONA method has led to new high resolution methods such as resolving of multiple dyes which are chromatically same[73] or different[74]. Multi-color DNA mapping methods based on high resolution methods are performed in genomic studies[90] and could potentially reveal further detailed information about the genome[91]. Many groups have been using DNA stretching methods to be able to understand proteins' interaction with DNA[92–96]. For better measurements of interactions with proteins, it is important to have good DNA mounting methods which have low DNA fluctuation rates. Most of the methods mentioned above have high DNA fluctuation rates. In addition, an easy buffer exchange method is crucial for some of these studies to learn the effect of different buffer conditions on the proteins.

We have used lambda DNA to test our method. Flow rates and other buffer conditions may need to be varied for different sizes of DNA. For each DNA length, control measurements could be made to determine stretching rates. We demonstrated that 30 nm resolution on stretched DNA could be achieved by applying the SHRImP method. Different DNAs could be used for this experiment, but each DNA should be modified to have a specific overhang that is complimentary

to the flowed primer. Chloroquine concentrations and flow speeds could be adjusted to get good stretching rates in different buffer conditions.

Our method will provide more ideal environment for studies of protein-DNA interactions as well as DNA mapping methods. We can stretch DNA close to its contour length and have higher coverage rates compared to other methods. The method can be used with high resolution microscopy as we demonstrated. It has the potential to surpass current barriers in several genomic and protein-DNA studies.

2.4 Materials

Flow Stretching Experiment: PEG slides were prepared according to protocol [70]. Chambers were assembled as shown in Figure 2.3. PEG coated coverslips were rinsed with milli-Q water and dried with nitrogen before assembling flow chambers. . Holes in the cover slide were drilled at the marked points, 0.75 cm away from the center of coverslips using a ¾-mm diamond drill bit. Double-sided tape was placed onto the slide along long edges, just next to the drilled holes and the coverslip was positioned on top of the glass slide. Both ends of the coverslip were sealed using epoxy. Drilled holes are a size that tubing (WEICO ETT-28) can fit in. After inserting tubes into the hole, it was sealed with epoxy resin.

Neutravidin (0.2 mg/ml final in T50 buffer) were flowed into the channel and incubated for 10 minutes. Blocking buffer (20 mM Tris pH;7.5, 50mM NaCl, 2mM EDTA, 0.2 mg/ml BSA) in open eppendorf tubes (1.5 ml) was degassed under vacuum for half an hour right before experiment to remove air bubbles. A syringe (5 ml, NORM-JECT) and syringe needle (B-D 26G-3/8) were used with the tubing. After assembling the syringe system, blocking buffer was flowed to remove excess neutravidin and incubated in the chamber for 5 minutes. Blocking buffer was flowed with a 150 ul/min rate using a pump (Harvard Apparatus, PhD2000). Chloroquine was

added into 1 mL blocking buffer to get 100 µM final concentration. Biotinylated lambda DNA (final concentration 3 pM) and chloroquine in blocking buffer were flowed into chambers for 8 minutes at100 ul/min. Next, 15 nM Sytox Green (Invitrogen) in blocking buffer was flowed for 5 minutes at 100 ul/min to remove excess chloroquine and to image bound DNA. 10 nM biotinylated primers within blocking buffer were flowed in at a rate of 100 ul/min for 5 min. Stretched and fixed DNA on the surface was imaged with Sytox Green buffer and PCD, PCA[97] and Trolox[98] (Sigma-Aldrich, St Louis, MO) de-oxygenation system.

Microscopy: Lambda DNA samples were imaged using total internal reflection fluorescence (TIRF) microscopy, implemented on an Olympus IX-71 microscope (Olympus America Inc) with PlanApo objective (Olympus 100x1.45 NA, oil). Cy3 labels were excited with 532 nm laser (Crysta Laser, Reno, Nv). Lambda DNA backbone stained with Sytox Green was excited using 488 nm blue laser (Melles Griot, Argon-Ion). Laser lines were set to have TIR and entered the microscope from the side port. The emitted photons were collected through a triple band filter cube which has an emission filter z488/532/635m and polychroic mirror z488/532/633rpc (Chroma Technology). The image was recorded by a back-illuminated, frame-transfer charge-coupled device (CCD) detector IxonEM (Andor Technology, Belfast, North Ireland). All DNA samples were imaged with a 0.3 s frame rate.

Image analysis: Cy3-Cy3-Cy3 labeled Lambda DNA samples were excited by a 532 nm laser. Double and triple photobleaching steps were searched to find SHRImP spots. While going through each qualified spot custom code, each picked spot was checked individually and filtered with certain FIONA error. Accuracy of spots less 5 nm were counted in. After locating SHRImP qualified spots, photobleaching frame number was entered manually. Distances between dyes are calculated with IDL code written according to Gordon *et al.*[73]

2.5 Figures for Chapter II

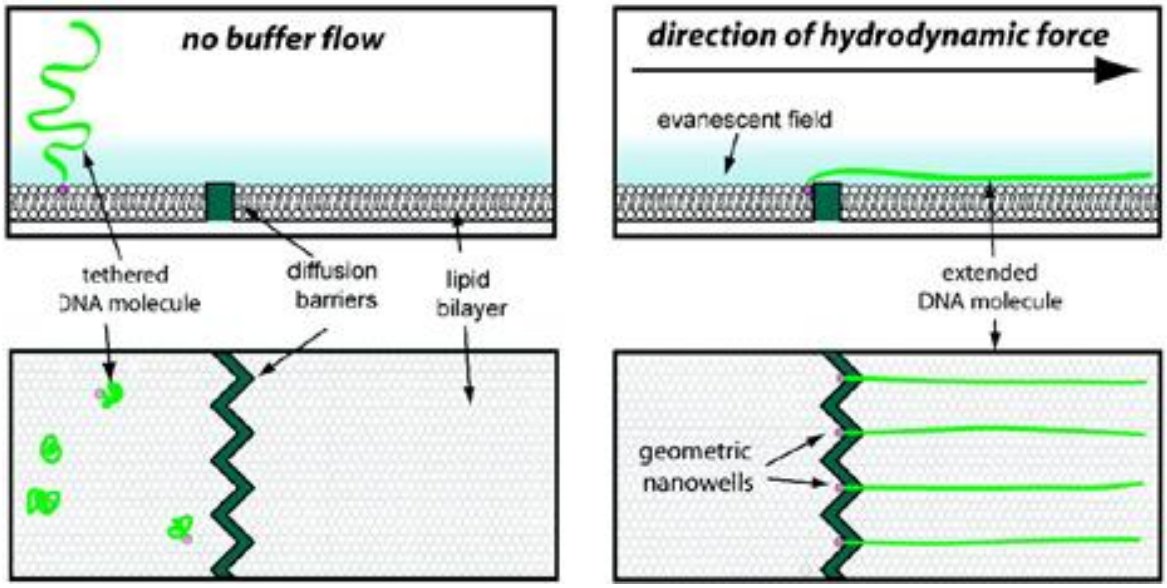


Figure 2.1: Design strategy for DNA curtains. (Adapted with permission from Fazio et al. Langmuir, 2008, 24 (19), pp 11293–11299. Copyright © 2008 American Chemical Society).

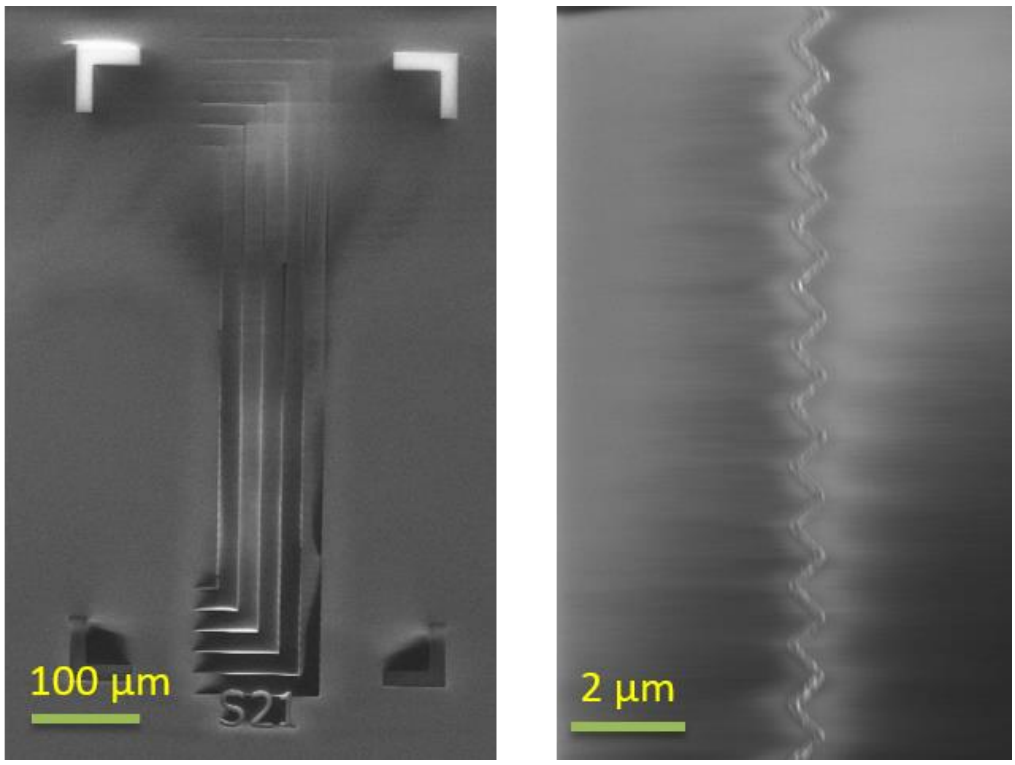


Figure 2.2: SEM image of barriers.

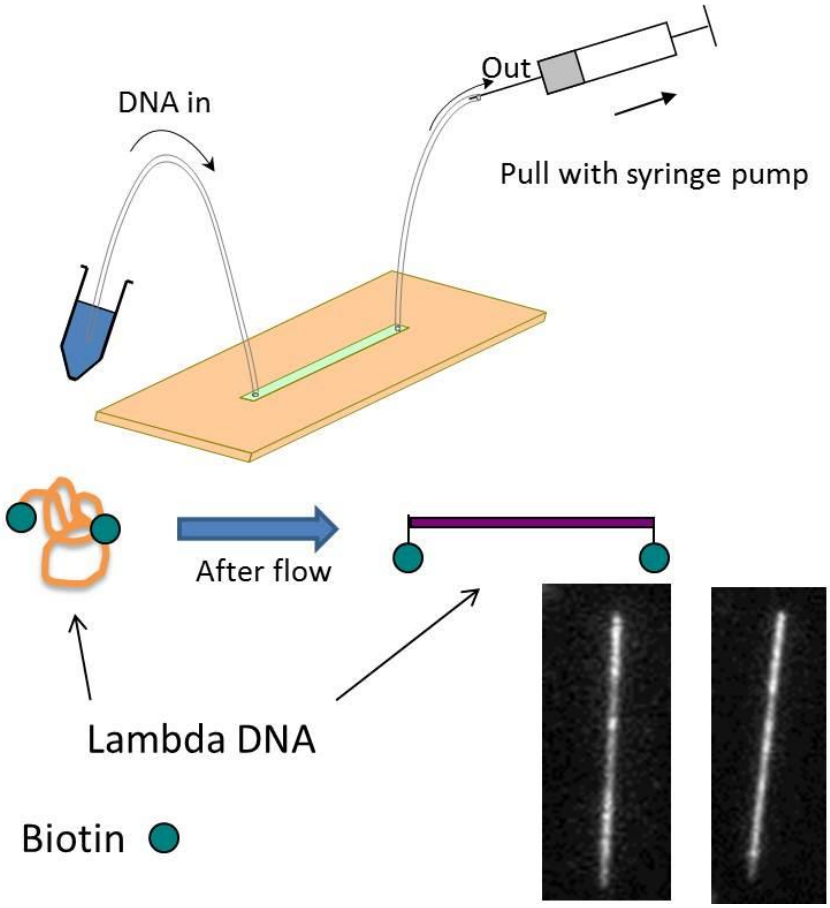


Figure 2.3: Top. Illustration of flow stretching experiment. **Bottom**. Tethered DNA at both ends labeled with biotin can be stretched out easily.

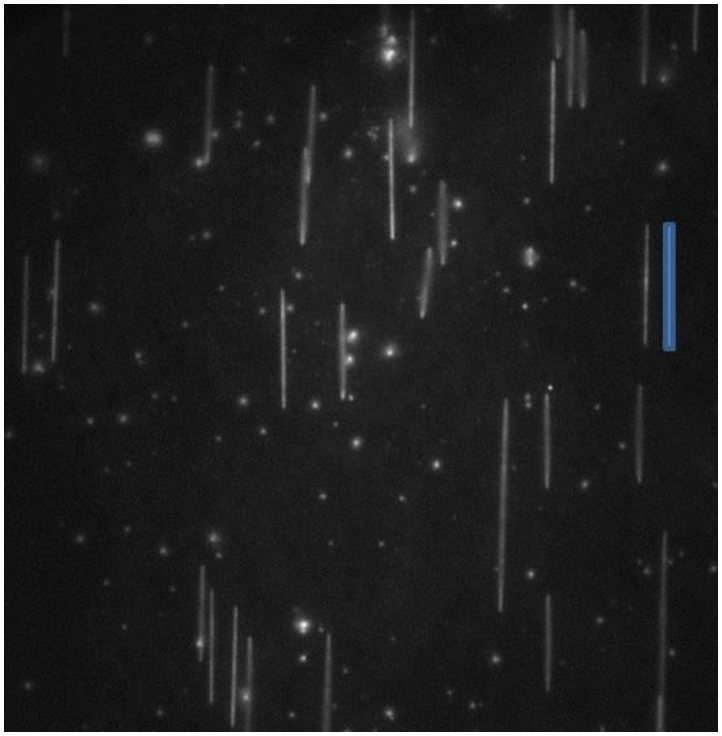


Figure 2.4: Image of double ended biotinylated DNA flow stretched by using 1.00 uM Chloroquine.

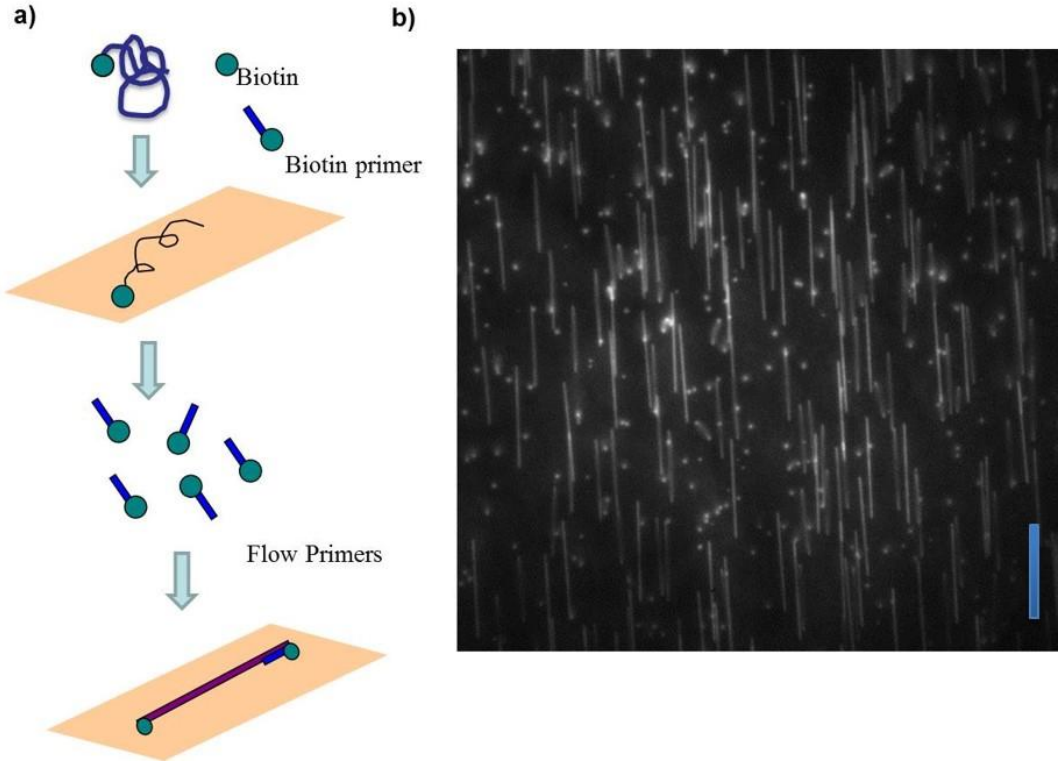


Figure 2.5: a) Experiment Design: one end biotinylated DNAs were tethered first on PEG surface by incubating 10 minutes. Mixture of chloroquine and biotinylated primers was flowed into chamber. The primer is the complementary start for overhang of lambda DNA. Primer (100 nM, phosphorylated) (containing 100 uM Chloroquine). Flow at 100uL/min for ~10 min. b) Stretched DNA is imaged by using Sytox Green DNA stain.

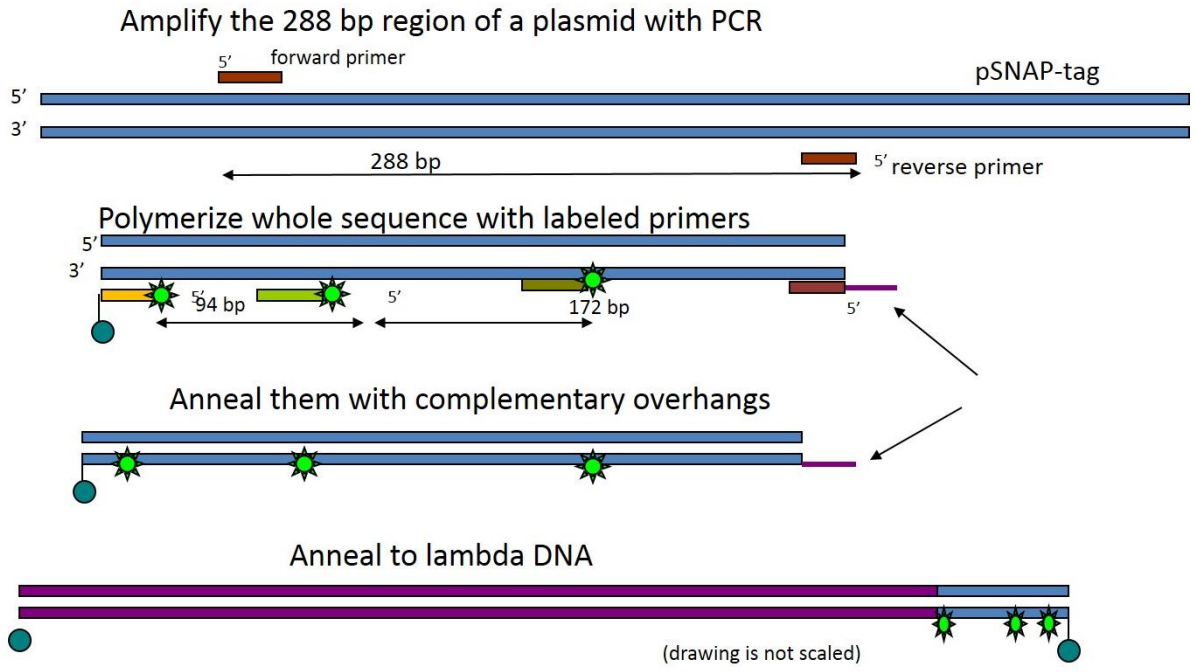


Figure 2.6: Constructing a DNA sample for Chloroquine stretching technique.

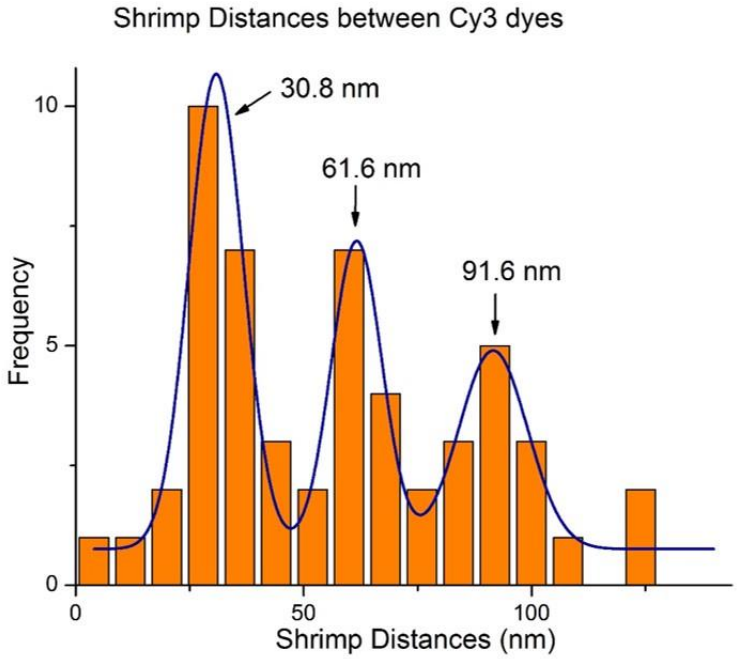


Figure 2.7: The imaging results after labeled DNA molecules were stretched and linearized by new stretching technique with chloroquine.

CHAPTER III

MULTI-COLOR SUPER-RESOLUTION DNA IMAGING*

3.1 Introduction

Structural variations in the genome play an important role in human health and common diseases as we mentioned in introduction chapter[99,100]. In general, these variations are defined as being longer than 500 bp[43,101]. Despite their importance, most genome-wide approaches for detecting copy number variations (CNVs) are indirect, depending on signal intensity differences between samples and controls to predict regions of variation. They therefore provide limited quantitative signal and positional information, and cannot detect balanced events such as inversions and translocations. Non-uniform sensitivity, specificity, and probe density of these platforms often lead to conflicting results even with identical samples[102,103]. This qualitative measurement requires further validation by low throughput detection methods, such as PCR and FISH.

Optical mapping of long single DNA molecules by using gaps generated by digestion with restriction endonucleases as landmark sites[104], or using fluorescent tags bound to specific sequence-motif sites[105,106] have provided a new way for comparatively rapid and direct whole-genome characterization and visualization for structural variation studies. However, due to the optical nature of the mapping, they are limited in their ability to resolve motifs that are closer than about ~1.5 kbp.

This generally requires selecting sequence-motif frequencies of at least 10 kb per site to avoid significant portions of genome with stretches of unresolvable sites (<1.5 kb) being left out. This not only reduces the ability to detect smaller structural variations, but also results in lower information density. Improving resolution would have clear advantages.

In this study, we describe a DNA mapping method with 100 bp resolution, based on localization of multiple sequence-motifs. We achieve this by two closely-related super-resolution techniques, both of which have been shown to have 10 nm resolution on DNA (and other samples)[73,74,107]. In particular, they are two-color SHRImP (Single molecule-High Resolution Imaging with Photobleaching), and two-color SHREC (Single-molecule High-Resolution Co-localization). As both SHRImP and SHREC can be extended to three or more dyes, using more colors will increase the number of resolved distances between individual molecules. Here, both techniques are corrected for chromatic aberration to get the relative distances between different colored dyes. We successfully generated two-color sequence-motif maps of 180 kb BAC clones (GCTGAGG and GCTCTTC) at 100 bp resolution.

3.2 Demonstration of One and Two Color SHRImP and SHREC Imaging

A 741 bp dsDNA template was constructed by PCR with one cy3-labeled primer (Figure 3.1-a) at the 5' end. Two other cy3-fluorophores were introduced at specific locations 94 bp and 172 bp from the 5' end by nick-labeling[108,109]. After stretching and linearizing DNA on a glass surface, we applied SHRImP. Each dye was photobleached and the measured distances between dyes was determined. The measured and expected distances were 27 nm and 32 nm, for the 94 bp, 61 nm and 58 nm for the 172 bp, and 95 nm and 90 nm for the 266 bp, which are in good agreement (Figure 3.1-b). The results also demonstrate that three fluorophores of the same color can be imaged with SHRImP simultaneously.

To test the feasibility of simultaneously using SHRImP and SHREC, a 741 bp-dsDNA model system was constructed with a Cy5-labeled PCR primer at the 5' end and two Cy3 molecules, at positions 32 nm (94 bp) and 58 nm (172 bp) from the Cy5 (Figure 3.2-a). The positions of the dyes were localized using a dual-view imaging system as described in the methods. The distances between the two Cy3-Cy5 pairs were determined to be 34 nm (32 nm-94 bp expected) and 88 nm (90 nm-266 bp expected). The distance between the Cy3-Cy3 pair was 56 nm (58 nm-172 bp expected) (Figure 3.2-b). This demonstrated the combination of SHRImP and SHREC can resolve distances smaller than the diffraction limit between multiple fluorophores of different colors.

3.3 One-Color Super-Resolution DNA Mapping

Lambda DNA (48.5 kb) was nick-labeled using Nb.BbvCI and Tamra-ddUTP, which has seven nicking sites along the 48.5 kb lambda DNA. In a previous study by Xiao *et al*., four sites were resolved, but the two nicking sites clustered at location B (Figure 3.3-a) and three nicking sites clustered at location C (Figure 3.3-a) could not be resolved due to their close proximity within the diffraction limit. Using conventional fluorescence, the seven nicking sites are reduced to four resolvable locations and distances are measured as in Figure 3.3-c with 1.47 μm, 3.27 μm and 4.27 μm respectively, which are in good agreement with expected distances. Using SHRImP, the clustered sites at B and C can be clearly resolved. Figure 3.3-d shows the distances between the two nicking sites at the B location to be 104 nm, which agrees well with 108 nm (318 bp) predicted distance. The distances between three clustered nicking sites at location C are also resolved to be 101, 202 and 312 nm which are in close agreement with expected distances of 102, 208 and 310 nm. (Figure 3.3-c)

3.4 Two-Color Super-Resolution DNA Mapping

Nicking sequence-motifs for Nb.BsmI and Nb.BbvCI occur on average every 2 kb per sequence-motif across human genome. With the limited resolution of regular fluorescent microscopy, most of the sites will not be resolved. Here we applied DNA-mapping with two-color super-resolution techniques and constructed Nb.BsmI and Nb.BbvCI sequence-motif map of a 180 kb BAC clone. Nb.BsmI has 71 recognition sequences (GCATTC) and Nb.BbvCI has 35 recognition sequences (GCTGAGG) across the 180 kb BAC clone.

In Figure 3.4, Nb.BsmI sites are nick-labeled with the green dye Tamra, while Nb.BbvCI sites are nick-labeled with the red dye Cy5. The DNA backbone is stained with YOYO-1. Three-color images were generated by using sequential excitation of Tamra (at 532 nm), Cy5 (at 642 nm), and YoYo-1 (at 488 nm). A few typical overlaid DNA fragments are shown in Figure 3.4-a. The distances between each neighboring spot of the same color are then calculated separately by using SHRImP analysis (Figure 3.4-b-d). To correlate the red and green channel with minimal chromatic aberration, Tamra and Cy5 labeled sites were analyzed together by using SHREC analysis. We modified the original SHREC technique[74] by using alternating excitation of green and red channels instead of DualView apparatus. This provided full view of imaging area rather than half as in DualView case. For this, both Tamra and Cy5 channels were merged after making chromatic aberration correction by using nanoholes as fiduciary marker, which are 100 nm in diameter and 1.5 μm apart (Figure 3.4-e). Patterned nanoholes all over the illumination area gave much better correlation between channels than using beads as fiduciary marker. The color spatial correlation function was created for each image frame based on the nano-hole fiduciary; this color correlation function has 5 nm resolution. By using the color correlation function, all Cy5 spots were mapped to the Tamra channel. A true two-color super-resolution image was created with

minimum chromatic aberration. Each DNA fragment was then mapped to the BAC clone reference sequences.

The nicking sites based on the Genbank reference sequences (Al662781.4) are shown in the top graph of Figure 3.5-a as the reference map. The distribution of nicking sites derived from each DNA fragment (Figure 3.4-a.) is mapped to the reference map by the searching algorithm described later in the chapter. All the nicking sites of all DNA fragments are plotted as the histogram along DNA backbone shown in the bottom graph of Figure 3.5-a. The histogram is created with bin sizes of 200 nm (diffraction limit) by using over 1000 DNA fragments and its range covers the whole BAC length. Experimentally localized nick–labels for Tamra and Cy5 dyes are shown in green and red, respectively. The experimental map agrees well with the reference sequence map (Figure 3.5-a). The peak height of each individual peak correlates well with the density of the nicking sites. More dense regions have higher peaks than other regions. Figure 3.5-b shows two different regions analyzed with two-color SHRImP analysis. One region covers from 44 kb (15 μm) to 54 kb (18.5 μm). There are five Nb.BsmI sites labeled with Tamra and six Nb.BbvCI sites labeled with Cy5. The closest distance measured between the same color (same sequence-motif) and different colors (different sequence-motif) are 134 bp (46 nm), 313 bp (106 nm) respectively. These results in the inset region show that 100 bp resolution can be achieved with the technique. The other region shown in Figure 3.5-b covers from 19 kb to 24 kb, in this region, all four Nb.BsmI and three Nb.BbvCI nicking sites were resolved at their expected locations.

Out of a total 105 Nb.BsmI and Nb.BbvCI sites, the super-resolution map resolves 91 sites compared to 65 sites with regular DNA mapping. Some of the sites could not be resolved due to the 30 nm resolution limitation as well as having more than two dyes within the diffraction limit.

Table 1 shows the complete super-resolution map. The full super-resolution two color sequence-motif map with SHREC analysis are also shown in the Figure 3.6.

3.5 Discussion

Despite recent advances in next-generation sequencing technologies, *de novo* genome assembly, structural variant and haplotype analysis using "short read" shotgun sequencing, remain challenging. Consequently, most medical re-sequencing projects rely on mapping the sequencing data to the reference human genome sequence to identify sequences and variants of clinical relevance[110]. One approach to address the sequence assembly challenge is optical mapping. Optical mapping has been used to construct ordered restriction sites for whole genomes, and has proven to be useful in providing scaffolds for shotgun sequence assembly and validation[104,111–116]. However, the information content and mapping capabilities are limited by low resolution and use of only a single restriction enzyme[113]. Additionally, it is also limited by the low throughput, imprecise DNA length measurement, and high error rates.

The resolution of DNA mapping is traditionally limited by the optical resolution (diffraction limit). Small fragments, or neighboring motif sites below 2kb, are hard to measure[106,109,113]. This results in increased false negatives. Consequently, lower density sequence-motif recognizing enzymes, such as a nicking endonuclease of 9 kb per nicking[117] site or a restriction endonuclease[113] of 20 kb per cutting site were used in part to avoid the complication of the resolution limitation. This not only decreases the information content and increases the frequency of desert regions (genomic region without sequence-motifs), but also limits the ability of detecting smaller genomic features.

We have shown here a multi-color super-resolution DNA mapping method, which provides more detailed DNA-sequence information. Single-color SHRImP can measure up to three dyes[73,107] within ≈ 30 nm (100 bp), significantly below the diffraction limit. The two color SHRImP method is shown to perform as efficiently as single color SHRImP. To correlate the two color channels for precise distance measurement, we developed a modified SHREC procedure with a nano-hole fiduciary marker to minimize the chromatic aberration. In this way, the resolution between different colors reaches 30 nm (100 bp). Super-resolution DNA mapping provides significantly higher uniqueness when compared to existing optical DNA mapping technologies for a given molecule length because more dense sequence-motif information can be obtained. This not only helps in *de novo* sequence assembly and physical map generation, since smaller contigs and less overlap between molecules is needed; it also reduces the current requirement for sample preparation of extremely long DNA molecules. Moreover, the method can generate sequence-motif maps for more damaged DNA sample, such as paraffin embedding (FFPE) samples.

In our two-color nick-labeling scheme, the high specificity for sequence recognition is determined by both the enzymatic nicking reaction and the fluorescent nucleotide incorporation reaction. More colors could be incorporated with additional nicking-endonucleases or in combination with other DNA labeling schemes (e.g. polyamides, Bis-PNA, methyltransferase[118–120]). Improved labeling technologies together with the advance of multi-color super-resolution imaging techniques can provide a DNA sequence-motif map of unprecedented detail. This map can be used to resolve smaller genetic variations over long distances, helping to resolve haplotypes, and can approach sequencing resolution.

3.6 Methods

3.6.1 Preparation of the glass coverslips and DNA mounting

Glass surface were functionalized with two different polyelectrolytes according to earlier studies.[86,106] 22mm × 30 mm coverslips were sonicated first in acetone for 30 min then in 1M KOH for 30 min. Between sonication steps, coverslips were rinsed with MilliQ water. After final drying with Nitrogen, coverslips were kept under Argon plasma cleaner for 5 minutes. Polymers Poly(acrylic acid) (PAcr) and Poly(allylamine) (PAll) (Sigma-Aldrich, St Louis, MO) PAII and PAcr were dissolved at 2 mg/mL in desterilized water and filtered through 0.22 nm filter. Cleaned coverslips were incubated in positive (PAII), negative (PAcr) and positive (PAII) polyelectrolytes consecutively about 30 minutes on shaker 150 rpm at 35^{o}C. Then, coated coverslips were kept in high purity water at room temperature.

Coverslips were placed on glass slide (1"x3") and pipet 4 μl DNA on the edge. DNA concentration is 10 pM and has oxygen scavenger system made of T50 buffer (10 mM Tris, pH 8.0, 50 mM NaCl) with PCD, PCA[97] and Trolox[98] (Sigma-Aldrich, St Louis, MO). The DNA was stretched on the coverslip surface by capillary action which causes strong flow. Coverslips were sealed with nail polish to prevent drying solution. DNA backbone is imaged with YOYO intercalating dye at 300:1 base-pair to dye ratio.

3.6.2 Microscopy

Short DNA and Lambda DNA samples were imaged by using total internal reflection fluorescence microscope (TIRF) which is built on an Olympus IX-71 microscope (Olympus America Inc) with PlanApo objective (Olympus 100x1.45 NA, oil). While imaging test-DNA in Figure 3.2-a for combined SHREC and SHRImP method, the Cy3 labels were excited with 532

nm laser (Crysta Laser ,Reno, Nv) and Cy5 is excited by 633 nm laser (Crysta/Melles Griot, Helium-Neon). DualView Apparatus (MSMI-DV-CC, Optical Insights, AZ) was used to split emission into two-color channels. Lambda DNA backbone stained with YOYO-1 was excited using 488 nm blue laser (Melles Griot, Argon-Ion). All laser lines were set to have TIRF and entered the microscopy from same port. The emitted photons were collected through triple band filter cube which has emission filter z488/532/635m and polychroic mirror z488/532/633rpc (both from Chroma Technology). The image was recorded by a back-illuminated, frame-transfer charge-coupled device (CCD) detector IxonEM (Andor Technology, Belfast, North Ireland). All DNA samples were imaged in 0.5 s frame rate. Nikon Eclipse Ti inverted microscope system was used to image the BAC DNA. Eclipse Ti has a built-in multi-color laser TIRF set up in which alternating lasers are controlled by AOTF. By using custom-made macro code, images were taken with 488, 532 and 641 nm lines sequentially, and then stage moved in 50 micron intervals spanning several regions in the same sample. Same filter sets and CCD camera is used to image BAC DNA as well.

3.6.3 DNA Sample Preparation

While preparing SHRImP test DNA, two PCR primers, one labeled at the 5' end with cy3 and the other one phosphorylated at the 5' end were used. The DNA was amplified, creating a product with one 5' end labeled with cy3 and the other end phosphorylated. Single- stranded DNA molecules were then produce d with the 5' end labeled with Cy3 by using lambda exonuclease enzyme to remove one strand of the PCR product. Two oligos, both with Cy3 at their 5' end, were hybridized at different locations on the single strand, resulting in a DNA labeled with fluorophores with either 94 bp or 266 bp apart. Then, by filling in the gaps between oligos using a polymerase enzyme, a double stranded DNA with cy3 fluorophores at specific locations was afforded. The

SHREC-SHRImP DNA sample was prepared in a similar way to the SHRImP sample explained in the paper, replacing one Cy3 labeled primer with Cy5 labeled primer

We obtained BAC clone in LB slabs from the BACPAC Resource Center at the Children's Hospital Oakland Research Institute (http://bacpac.chori.org/) from the BAC library CHORI-501. BAC DNA sample (cho501-1H2, GenBank Al662781.4) used in the study were prepared using Qiagen's Large-Construct Kit. The DNA sample was quantified using Nanodrop 1000 (Thermal Fisher Scientific) and their quality assessed using pulsed-field gel electrophoresis. One milligram BAC DNA was linearized with 2 U of NotI and nicked with 0.5 U nicking endonuclease Nt.BbVcI and Nt.BsmI (New England BioLabs, NEB) at 37 ^{o}C for 2 hours in NEB Buffer 3. The resultant DNA fragments were labeled with 25 nM Cy3 acyclo-dUTP and Cy5 acyclo-dCTP (Perkin Elmer) and Vent (exo-) (NEB) for 1 hour at 72 ^{o}C. The backbone of above fluorescently tagged DNA (5 ng/uL) was stained with YOYO-1 (3 nM; Invitrogen). Lambda DNA sample was nicked with Nt.BbVcI and labeled with 25 nM Cy3 acyclo-dUTP.

3.6.4 Image Analysis

Cy3-Cy3-Cy3 labeled DNA samples (Figure 3.1-a) were excited by 532 nm laser. Double and photobleaching steps were searched to find SHRImP spots. After locating SHRImP qualified spots, photobleaching frame number was entered manually. Distances between dyes are calculated with IDL code written according to Gordon *et al*.[73] Cy5-Cy3-Cy3 DNA samples (Figure 3.2-a) are imaged with addition of DualView apparatus to separate Cy3 and Cy5 emissions. Each spot in green channel were searched for double photobleaching steps. Distances between Cy3-Cy3 in this sample were calculated with the SHRImP code. Before and after running each experiment, nanoholes as fiduciary markers were used to get mapping function between two channels for

SHREC calculations. Nanoholes were 100 nm in diameter and 1.5 micron apart from each other in both x and y direction. Mapping functions in MATLAB were used to register and map channels[121,122]. When analyzing SHREC data, mapping functions with mapping error less than 5 nm were used to calculate distances between Cy3-Cy5 dyes.

Experiments with lambda DNA were done by using 488 nm and 532 nm excitations. Backbone was labeled by YOYO-1 stain and restriction sites were labeled with Tamra fluorophore. After overlapping YOYO-1 labeled backbone and Tamra labeled restriction sites channels in ImageJ, spots on lambda DNA were picked by Point Picker Plug-in. For each lambda DNA, distances between dyes were calculated and searched for double photobleaching spots to identify qualified SHRImP spots. Location of each SHRImP spot were recognized according the other non-SHRImP dyes on the same DNA and its distance to DNA ends. Non-SHRImP and SHRImP distance were all calculated using custom made code in IDL.

BAC DNA with YOYO-1, Tamra and Cy5 channels were overlaid by using ImageJ. Later on, point picker (imageJ plug-in) was used to pick spots in Tamra and Cy5 channel which are on DNA backbone. By using IDL custom code, each picked spots were checked individually and filtered with certain FIONA error. Picked spots were checked to see whether they are SHRImP spot or not by looking at count trace over time. Distance between each spot were calculated and results were printed in an excel sheet. Data in each column in an excel sheet represented consecutive distances between neighboring spots for every DNA fragment. Search algorithm (which is developed in Mathlab) was run to match the consecutive distance list of each DNA fragment on the actual DNA reference list. If the DNA fragment were located on full BAC DNA with the search algorithm, new positions of restriction sites were assigned to their corresponding locations on full BCA DNA. If DNA is overstretched or under-stretched and there is any non-

specific label on the list, restriction sites on the DNA fragment were excluded from all output list. Final output list from several experiments was plotted in a histogram in Origin Graphing Software.

3.6.5 Image Processing for Distance Calculations on BAC DNA

The IDL code retrieves coordinates of spots from text file. These spots were previously picked by the point picker which is an ImageJ plug-in. After displaying DNA backbone image, retrieved spots are printed on their locations on the backbone image. Then, locations for beginning and end of stretched DNA are asked to be clicked on the image display. By using DNA end points, slopes and origins of each DNA segment are calculated. Program searches through a line on a DNA and listing corresponding spots to each DNA line. Later, 2D-Gaussian-fits are performed for all spots for all frames to get their coordinates over the time. Certain number of frames is averaged to remove the fluctuations effect in dyes' emissions. If the point is non-SHRImP spot, first 12 frames of the spot are averaged. If the point is SHRImP point, position is averaged from beginning frame till before photobleaching occurs. Averaging excludes off-values which may be due to bad photon collection in any single frame and bad fit results. After listing all the spots for each DNA segment, calculation of distances between each neighboring spots are made by using their averaged positions. While going through each spot on a DNA, some spots can be excluded by looking at its raw image's quality and trace over the time. During the spot check, SHRImP spots are detected by looking at their photobleaching behavior. If the spot is a SHRImP spot which has gone through double photobleaching, SHRImP calculation is made according to previous study[73] and the measured SHRImP distance is added to the DNA segment's distance list. Distances of each DNA segment is printed in a column of an excel sheet.

3.6.6 Search Algorithm

The best match for the measured data is found by moving the measured span across the template by small increments and checking how closely the points line up. The distance between two measured points is used to determine a range in which any template points will be considered possible matches with the second measured point. If a template point is within this range, the template point is added as a found point and the potential position of the next point after is updated in accordance to this new starting point, and the process is repeated to find the next point. This allows for a cascade of found points when the correct location is found as each successive point will be in the correct range. If no point is found within range of the current length, the next length is added and a new range is produced, and this continues until a point is found that falls within range or all the lengths have been used up. This allows for the program to skip over inaccurate data from erroneous measurements. If there are multiple points in the range the program chooses the one closest to the measured point they are being matched to. If matching points are found, the algorithm then goes back to check if the first measured point lines up with a template point using the distance between the first matching point and the first measured point to determine the range that a match can be in.

This scanning method is repeated multiple times, moving the starting point farther down the template until the entire length of the template has been tested. While the algorithm is traversing the template it remembers the best match it has found so far. A match is considered better than another if more points line up with the template. If two different matches have the same number of points that line up, the match where the points are closest to the template is chosen.

3.7 Figures and Tables for Chapter III

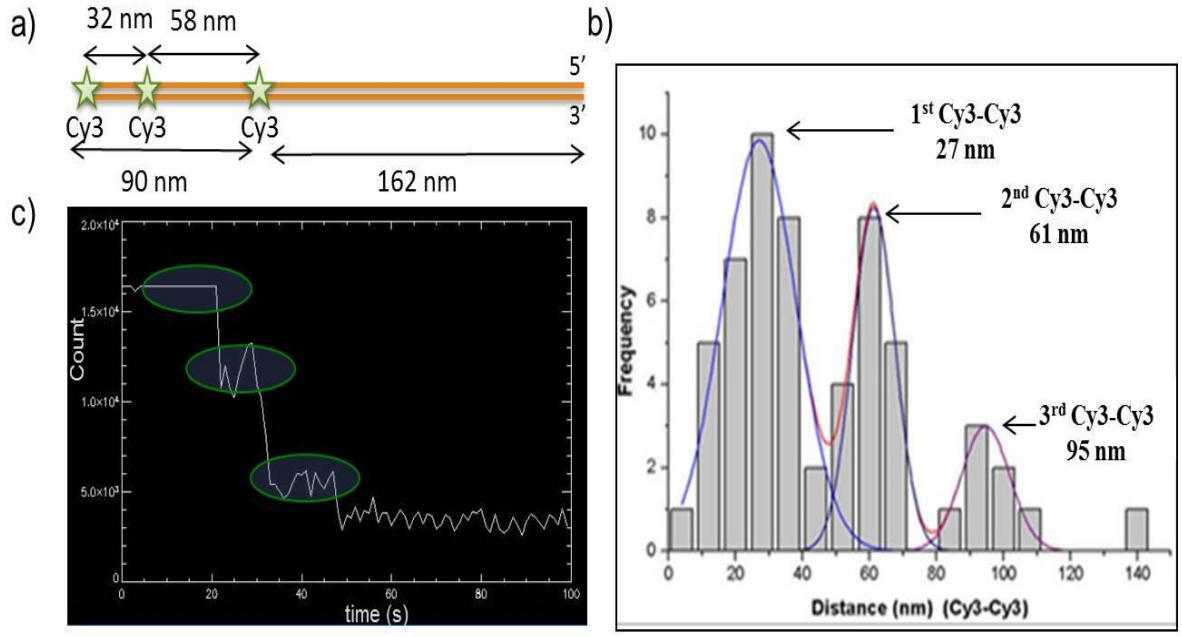


Figure 3.1: (a) Diagram of a DNA molecule with 3 Cy3 dyes to test the SHRImP method (b) Histogram of SHRImP distances between Cy3 dyes shows that the 3 distances between Cy3 pairs are 27 nm, 61 and 95 nm, which is in good agreement with expected distances (c) Three step photobleaching of the sample.

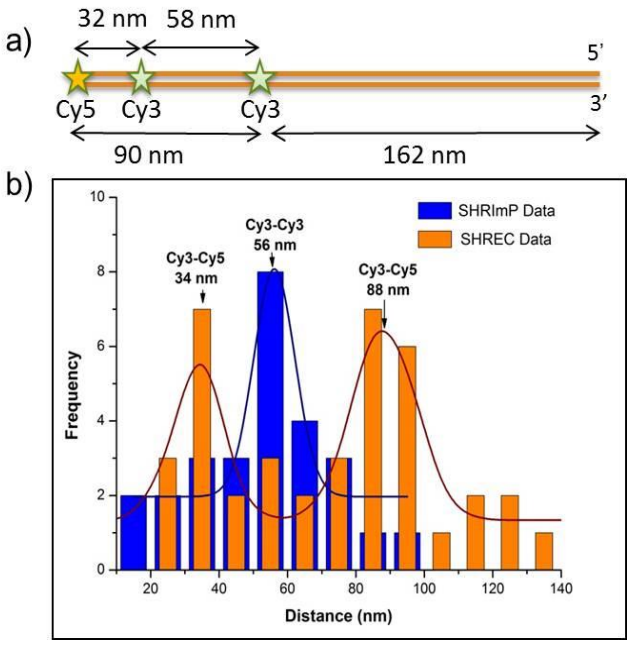


Figure 3.2: (a) Diagram of a DNA molecule which has Cy5-Cy3-Cy3 dyes on it.(b) Combined histogram of SHREC (red) and SHRImP (blue) analysis indicates that SHRImP distance between two Cy3 dye molecules is 56 nm, and two SHREC distances between Cy5-Cy3 molecules is 34 nm and 88 nm.

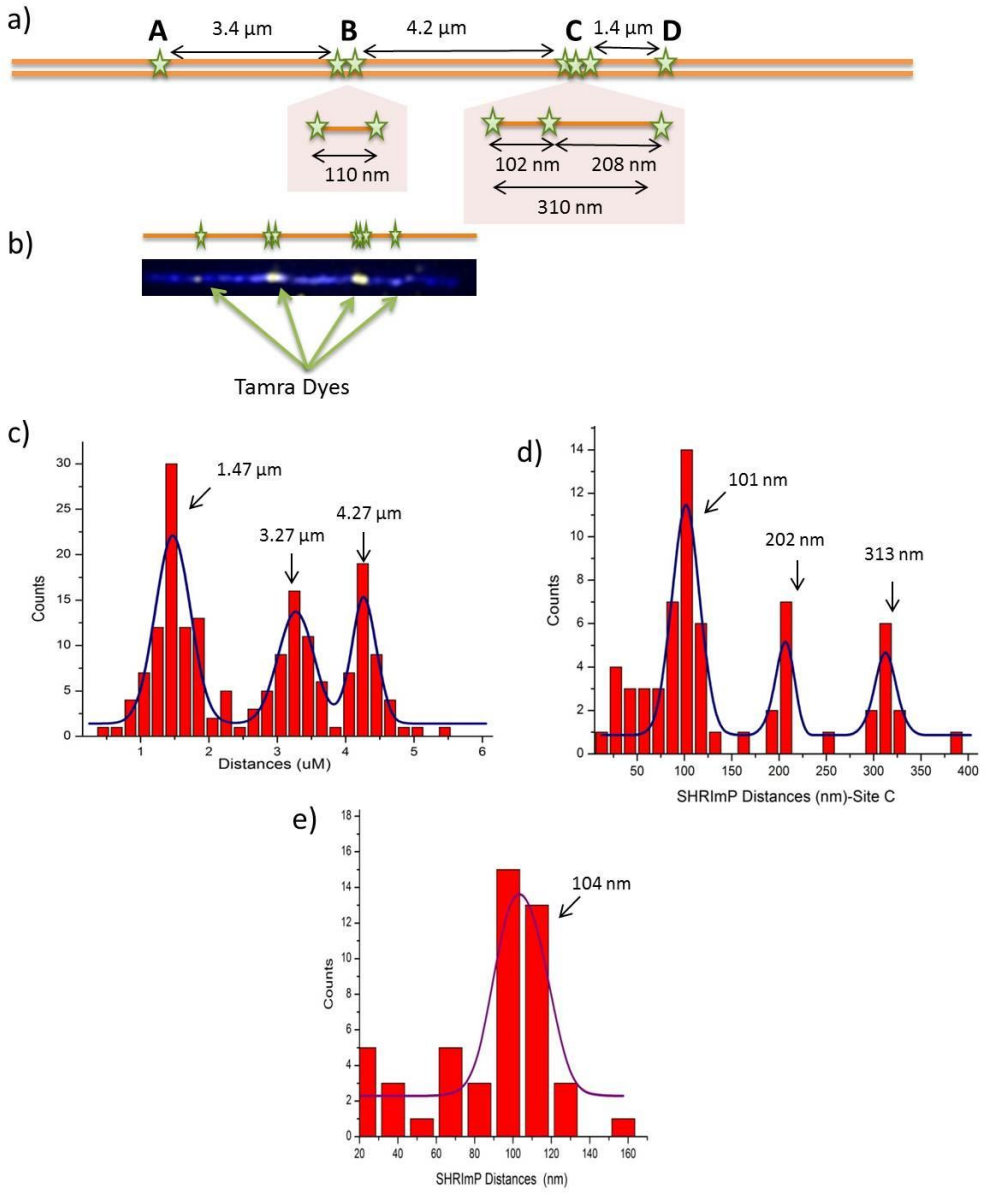


Figure 3.3: Resolving the Lambda DNA nicking pattern. (a) Lambda DNA nicking pattern by the enzyme Nb.BbvCI. Green stars show the expected location of labeled nicked sites. Site-B has two dyes and Site-C has three dyes in diffraction limited spots (b) an image of Tamra labeled nicking sites on YOYO stained Lambda DNA backbone. (c) Histogram of non-SHRImP distances. Peaks are at 1.47 µm, 3.27 µm and 4.27 µm which are close to the expected distances. (d) Histogram of SHRImP distances for site-C which has three peaks at 101 nm, 202 nm and 312 nm, in good agreement with the expected distances of 104 nm, 209 nm and 313 nm. (e) Histogram of SHRImP distances for site-B which has a peak at 104 nm that is also close to the expected distance of 110 nm.

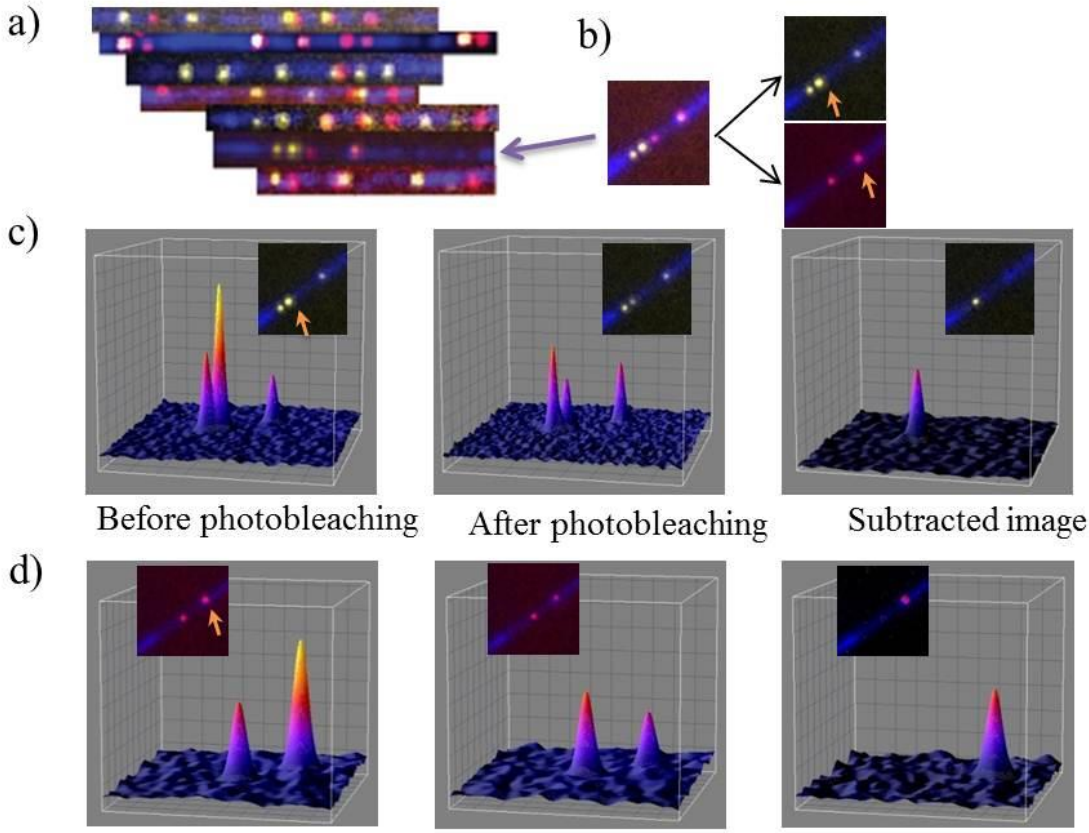


Figure 3.4: Two color super-resolution imaging of DNA molecules (a) Diffraction limited imaging of BAC DNA fragments. DNA backbone is shown in blue (YOYO-1 staining). Nb.BsmI nicking sites are in yellow and Nb.BbvCI sites are in red. DNA fragments are rotated to be aligned with all others based on the nicking pattern. (b) Image of a single DNA fragment (the fragment on the figure 4a shown with an arrow). The image on the left is the superposition of the green and red images on the right. Excitations in the green and red channels are done alternatively. Arrows shows SHRImP spots in each channel (c) Gaussian profiles of spots in the green channel on the DNA fragment. Colors in the profiles shows varying intensities.(blue to yellow is low to high intensities. The graph includes Gaussian profiles of the the SHRImP spot(shown with arrow) before and after photobleaching (first and second graph). Third graph is the Gaussian profile of the second dye on the SHRImP spot. It is found by subtracting spot's intensity profile after photobleaching from that of before photobleaching (d) Gaussian profiles of dyes on the red channel on the same DNA fragment. Spot shown with arrow is a SHRImP spot. Gaussian profiles of the spot with before and after photobleaching are shown in first two graphs. Third one is Gaussian profile of the second dye on the SHRImP spot.

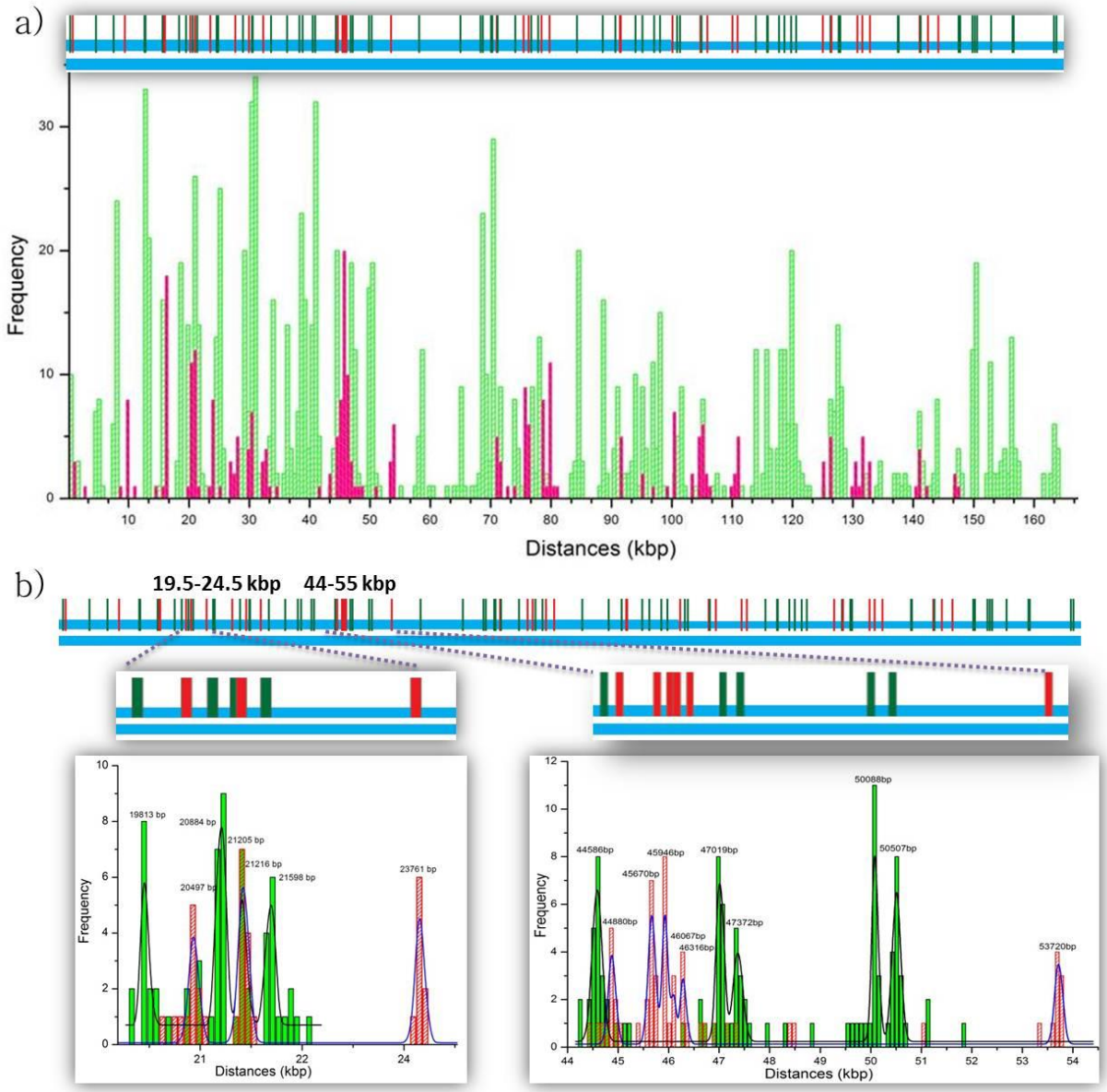


Figure 3.5: BAC DNA Mapping with two-color SHRImP analysis. (a) Locations of experimentally determined nicking sites on BAC DNA are shown above the histogram. DNA was nicked by Nb.BsmI at 71 locations (labeled by Tamra dyes) and by Nb.BbvCI at 34 locations (labeled by Cy5 dyes). The histogram is created with bin sizes of 200 nm. Experimentally located nicking sites for Tamra and Cy5 dyes are shown with green and red bars respectively. (b) The histogram on the left shows the region from 19.5 kb to 24.5 kb, in which four Tamra and three cy5 labeled sites are resolved as expected. The histogram on the right has 11 nicking sites and covers an 11 kb region from 44 kb to 55 kb. Five Tamra and sixCy5 labeled sites are resolved with two-color SHRImP in expected locations.

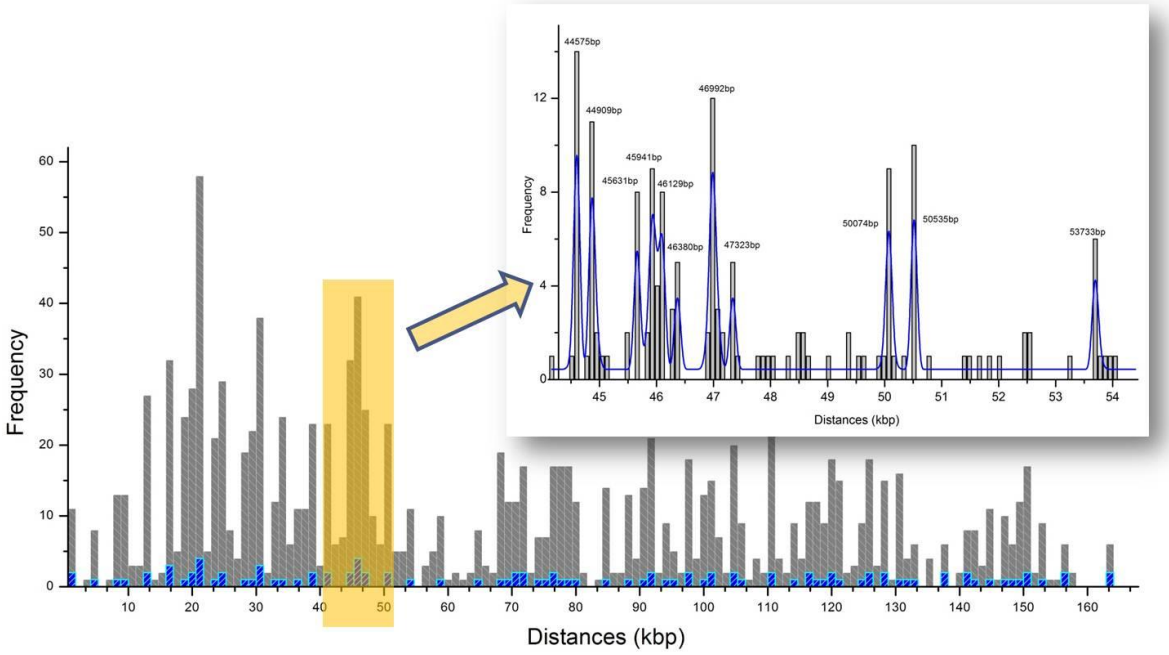


Figure 3.6: SHRImP/SHREC combined analysis of DNA mapping. Histogram ranges from 0-165 kbp (0-60 μm) which is the full BAC DNA length. Histogram's bin size is 200 nm. Main image has columns in two different colors. Dark gray is the histogram of experimental distances of restriction sites. Blue columns are the histogram of actual restriction list that includes both green and red labeled sites. Zoomed graph is the sub-region of main graph from 44 kbp to 55 kbp with binning size of 100bp (30 nm). Peaks are fit with multi-gaussian function to get mean values. All expected 11 restriction sites in that region is resolved as shown in zoomed histogram.

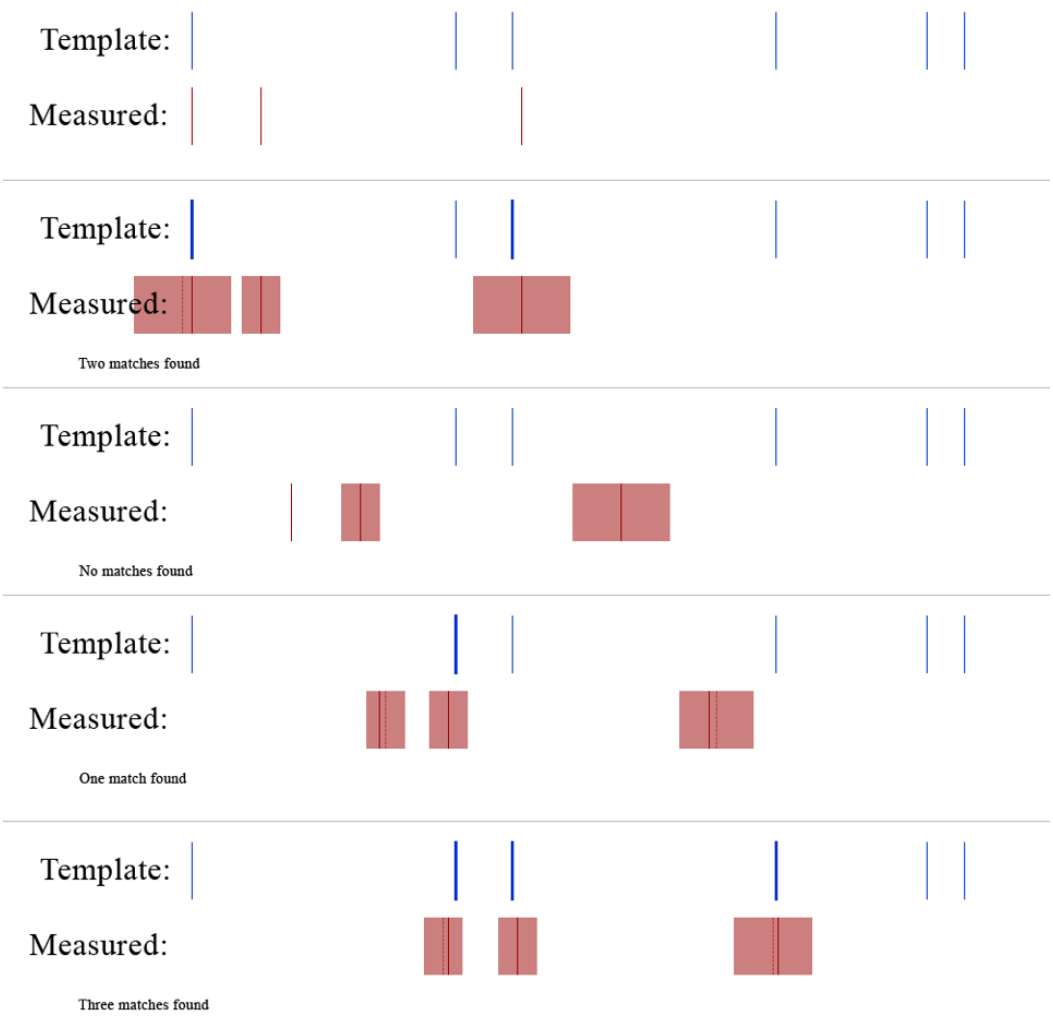


Figure 3.7: An illustration of the algorithm

Row 1: The template and the starting position of the measured data before a scan is run.

Row 2: The first scan. The distance between measured points one and two determines the range around point two that a match can be in. No match is found, so the distance between measured points one and three determines the range around point three that a match can be in. A template point is within the range, so it is recorded as a found point and shown as thick here. Point three is the last point, so the first point needs to be checked. The distance between points one and three is again used to determine the range that a match can be found in, but the center of this range is based on the template position of the first found point (point three) minus the distance between measured points three and one. This center is shown slightly to the left of measured point one as a lighter

Figure 3.7: (Cont.)

dotted line. A template point falls within this range and it is saved as the start of the found points. This point is also shown in bold.

Row 3: A later scan where the measured data has been shifted over to the right, having already gone through a number of previous scans. Once again the distance between measured points one and two are used to determine the range around point two to look for template points. No match is found, so the distance between measured points one and three determines the range around point three that a match can be in. Once again, there is no template point that falls within that range. The program has hit the end of the measured points without finding any match, so it does not bother checking the starting point.

Row 4: Another later scan farther to the right determines the range around measured point two that matches can be found in based on the distance between measured points one and two. This time a match is found and recorded. It is shown in bold in the figure. The distance between measured points two and three is used to find the range for a match for measured point three, but the center of this range is determined by adding the distance between measured points two and three to the template point that was found to match measured point two. This means that measured point three is not in the center of its own range of potential matches. No matches are found for measured point three. Since the algorithm has reached the end of the measured points and has found at least one matching point, it goes back to check if there is a match for measured point one. The distance between measured point one and the first point with a match, measured point two, is used to determine the range that a match can be in, and the center of that range is determined by subtracting that distance from the location of the template point that matches measured point two. No point is found within this range.

Row 5: Once again the scan takes place even farther to the right. The range for matching points for measured point two is based on the distance between measured points one and two, and centered on measured point two. A template point is found within this range and recorded. It is shown as bold in the figure. The range for measured point three is determined by the distance between measured points two and three, and the center of that range is determined by adding that distance to the position of the template point that matches measured point two. A template point is found in this range, recorded, and shown as bold in the figure. Having reached the end of the measured data, and having recorded at least one match, the algorithm goes back to check if

Figure 3.7: (Cont.)

measured point one matches a template point. The range that is searched is determined by the distance between the first measured point and the first measured point that matches a template point, in this case measured point two. This distance is subtracted from the position of the template point that matches measured point two to determine the center of this range. A template is found in this range and recorded as the starting point of the matching points. This point is shown in bold. This is the best match found so far, with all three points lining up with the template and very little offset from any of their matching points. This sequence of matching points is saved as the best match found so far along with other relevant data.

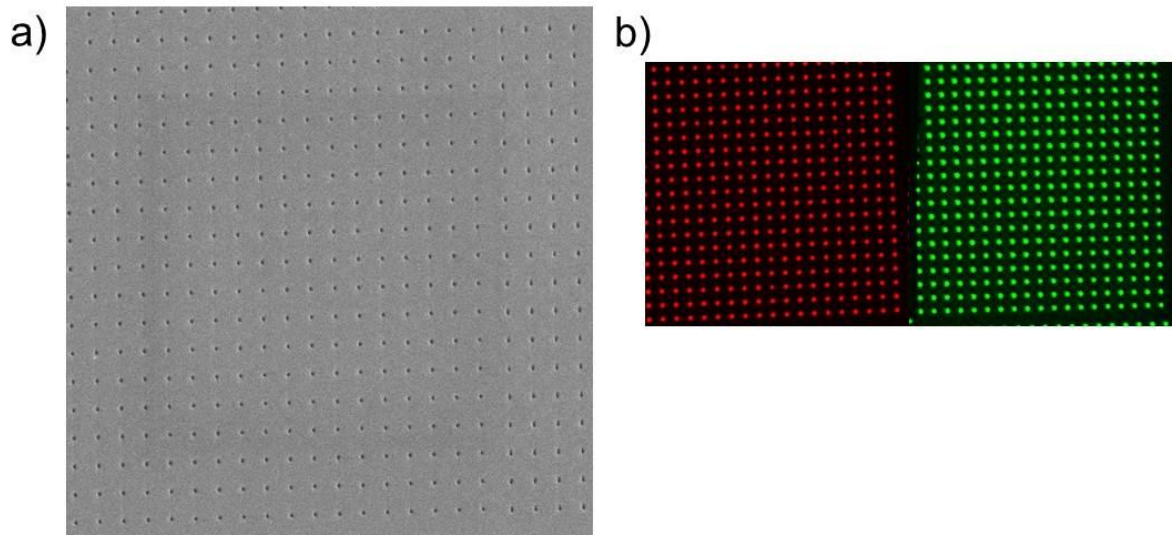


Figure 3.8: Nanoholes as fiduciary markers are made by using Focus Ion Beam (FIB). They are 100 nm in diameter and 1.5 µm apart from each other. First, thermal evaporator is used to deposit ~15 nm Cr and ~140 nm Ag (silver) on coverslips. Then, nanoholes image map which is drawn by CorelDRAW are patterned on coverslip by FIB. a) SEM image of nanoholes. b) TIRF image of nanoholes in green and red channels is taken by using Dualview apparatus to image SHREC sample in figure 2 of the paper. The white light was shed from the top, and only nanoholes are the region where the light can pass. And because of its broad wavelength and Dual-view optics, the transmitted light is separated according to wavelength.

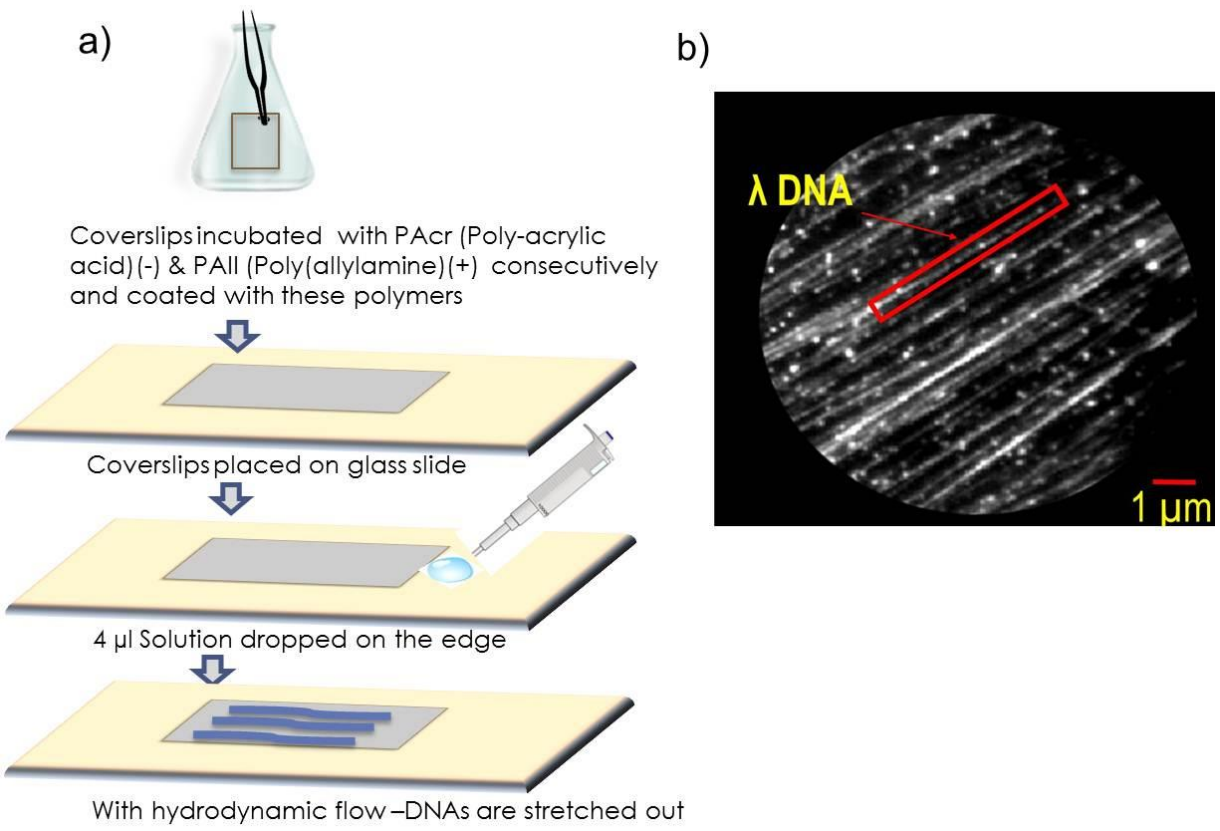


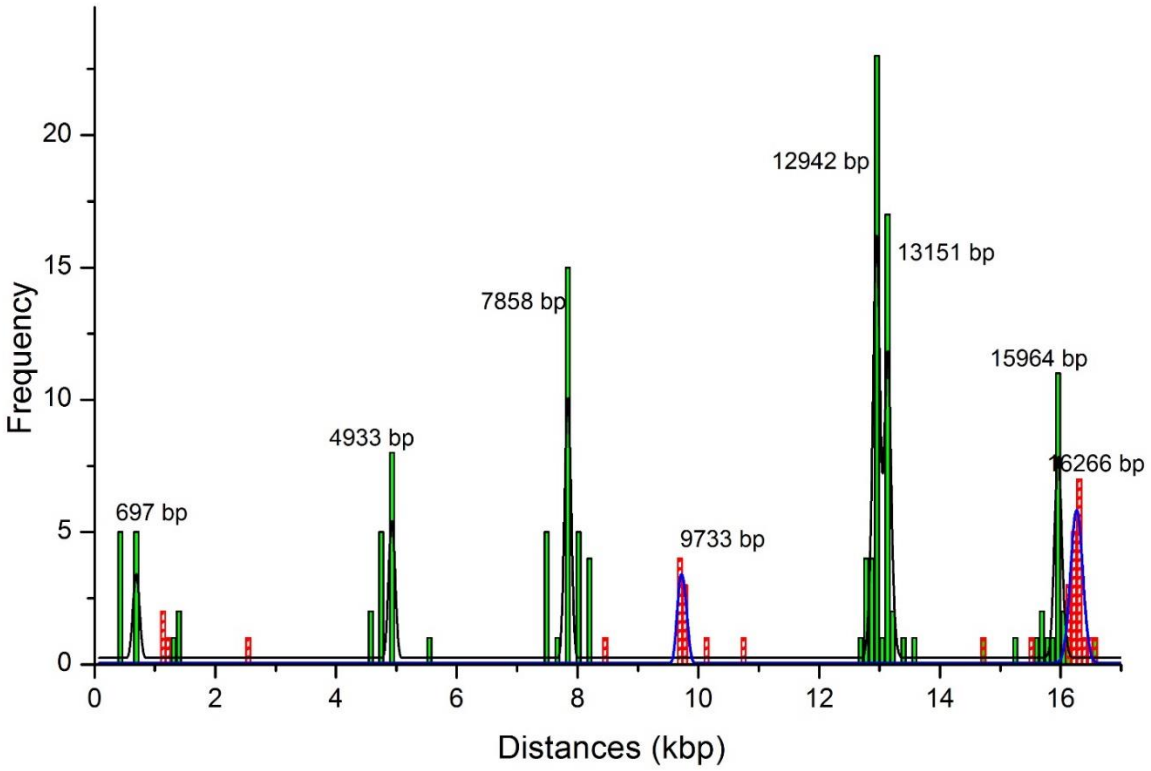
Figure 3.9: a) Illustration of stretching method b) Stretched lambda DNA image with YoYo on PAcr-PAII coated coverslip.

Expected	Experiment	Expected	Experiment	Expected	Experiment	Expected	Experiment
680	697	36474	36465	76895	76897	119948	119988
1092		38516	38496	78033	78043	120642	120007
4934	4933	39052	39082	78597	78573	125064	130683
7857	7858	40734	40744	79929	79897	126378	126378
9696	9733	41154	41130	84447		126438	131577
12953	12942	44579	44586	88710	88686	127661	127657
13111	13151	44892	44880	90791	90729	127928	127876
15921	15964	45652	45670	91563		130765	
16188	16266	45925	45946	91611		131581	
16323	20497	46059	46067	94113	94137	132842	
18648	18646	46324	46316	95268	95281	137529	
19782	19815	47009	47019	97195	97182	137545	
20482		47367	47372	98169	98131	141034	140969
20874	20884	50060	50088	101018		141204	141189
21185	21203	50507	50507	101476	101502	144114	144072
21275	21216	53723	53720	100239	100242	142420	
21622	21602	58383	58400	104866	104850	147481	
23780	23761	65194	65186	105036	105005	147759	147792
24826	24801	68517	68510	105996	105891	149787	149793
25096	25087	68893	68931	110148	110174	150168	150178
27917	27838	70246	70375	111001	126426	150546	150569
29187	29184	70454	70516	113998	113999	152863	152884
30200	30188	71221	71139	115926		156393	
30687	30665	71352	71338	115966		156591	156405
30804	30806	74253	74281	117839	117880	163235	163221
32533	32567	75632	75679	118725	118776	163642	163537
33832	33820	76443	76479	119837	119794		

Table 1. Restriction sites of Nb.BsmI and Nb.BbvcI enzymes on BAC DNA are located with 2D-SHRImP method. The table also includes expected distances of each site. Nb.BsmI and Nb.BbvcI are shown with green and red fonts respectively.

APPENDIX-A

MORE DETAILED HISTOGRAMS FOR BAC ANALYSIS



a

Figure A.1: BAC DNA two-color SHRIMP results. Zoomed in region of the full histogram from 0 to 16 bp.

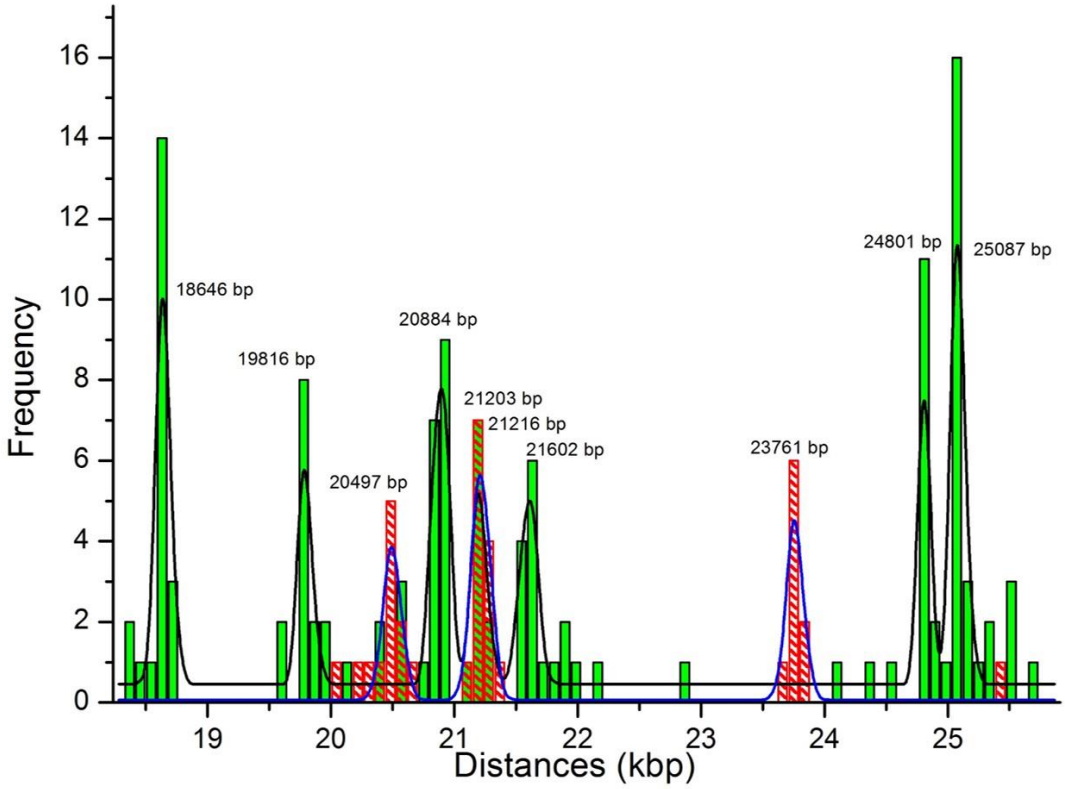


Figure A.2: BAC DNA two-color SHRIMP results. Zoomed in region of the full histogram from 17 to 26 bp.

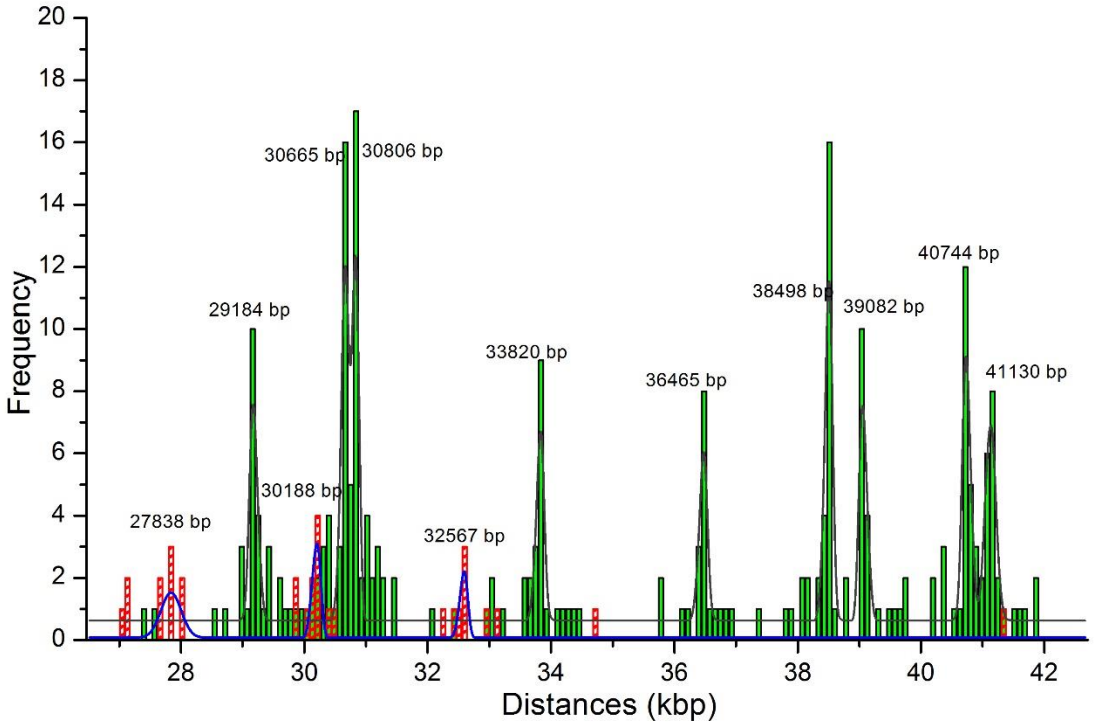


Figure A.3: BAC DNA two-color SHRIMP results. Zoomed in region of the full histogram from 27 to 42 bp.

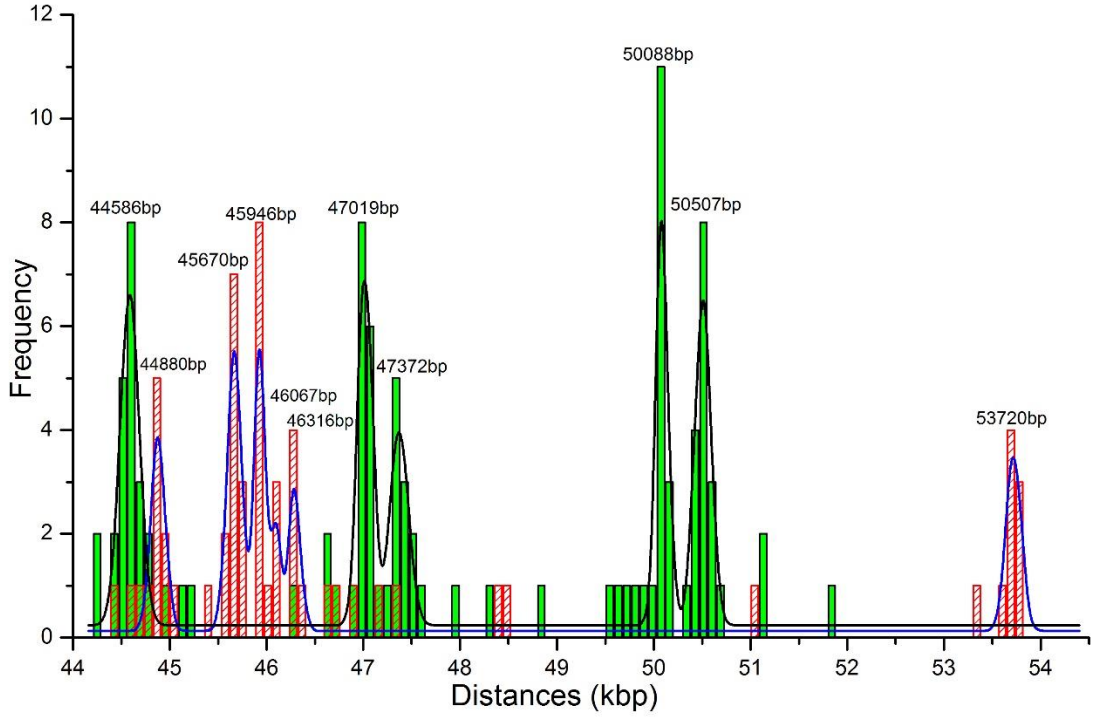


Figure A.4: BAC DNA two-color SHRIMP results. Zoomed in region of the full histogram from 44 to 54 bp.

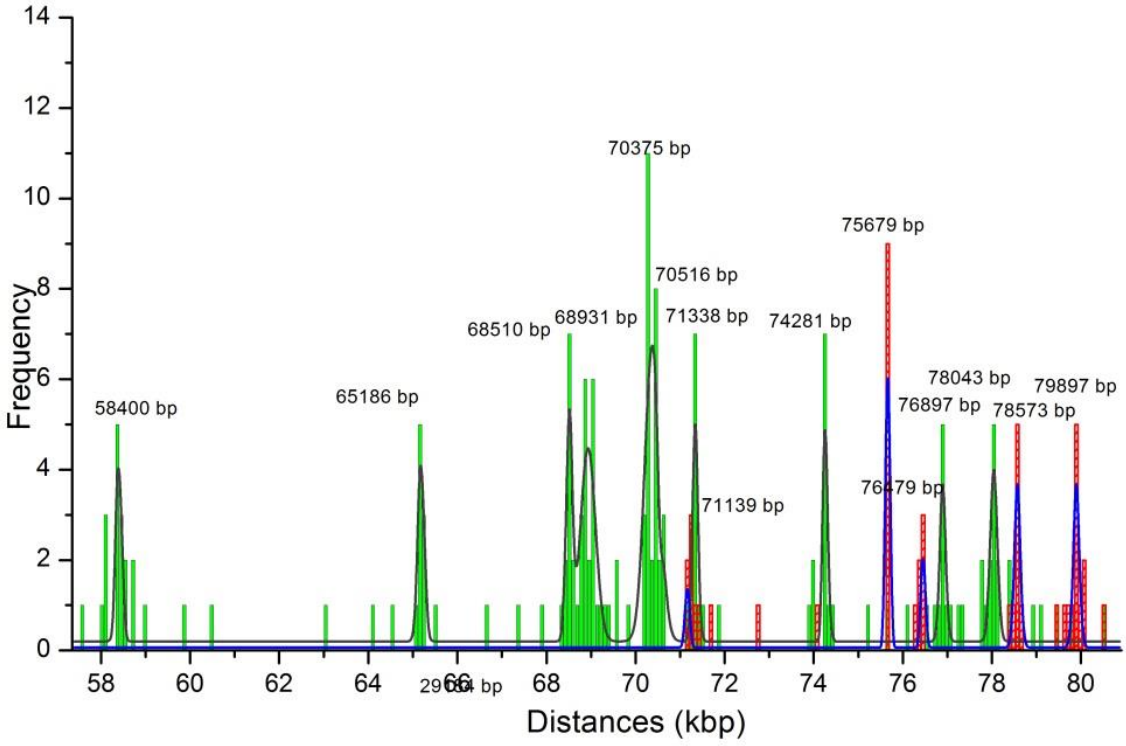


Figure A.5: BAC DNA two-color SHRIMP results. Zoomed in region of the full histogram from 57 to 80 bp.

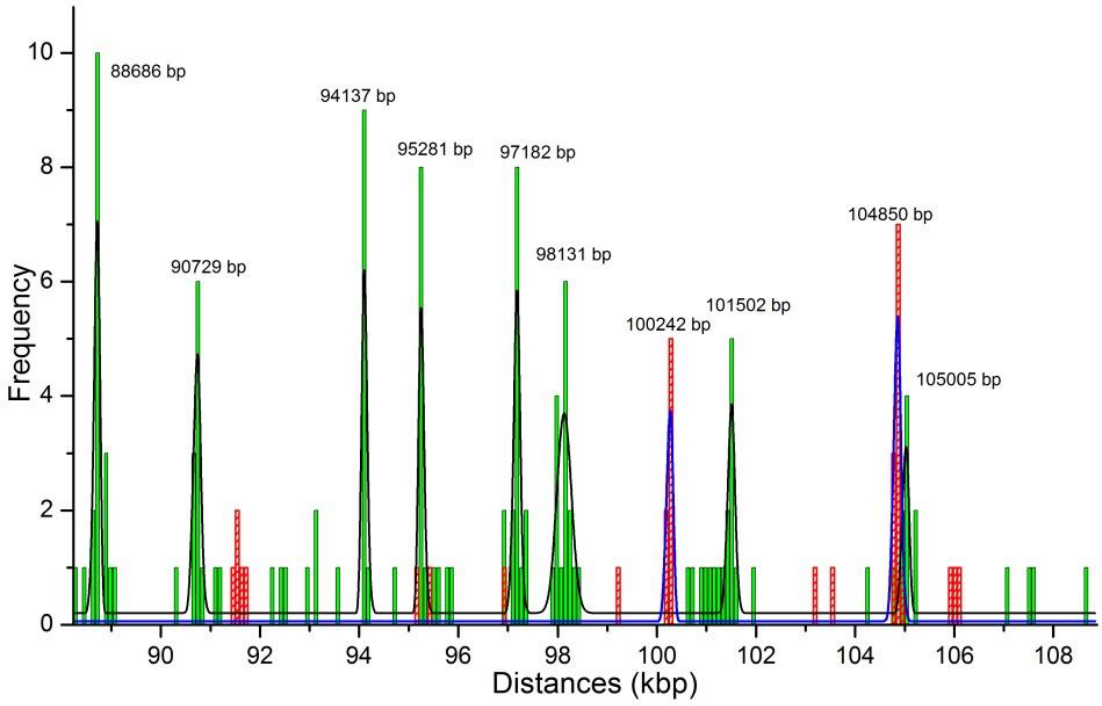


Figure A.6: BAC DNA two-color SHRIMP results. Zoomed in region of the full histogram from 88 to 108 bp.

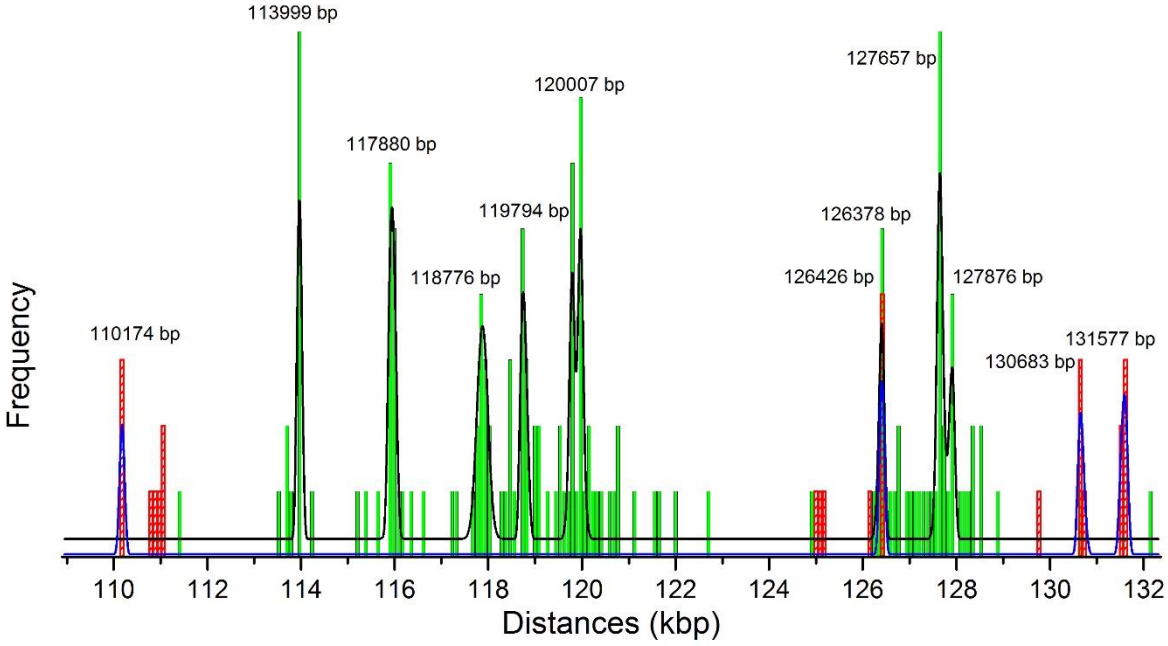


Figure A.7: BAC DNA two-color SHRIMP results. Zoomed in region of the full histogram from 110 to 132 bp.

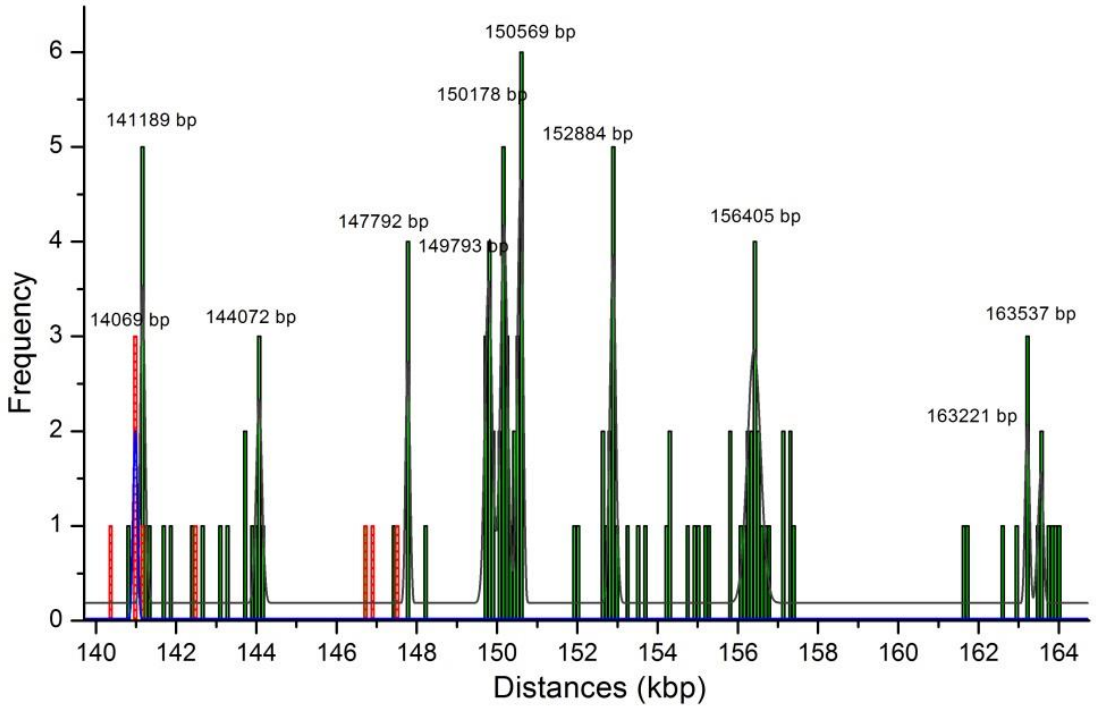


Figure A.8: BAC DNA two-color SHRIMP results. Zoomed in region of the full histogram from 140 to 165 bp.

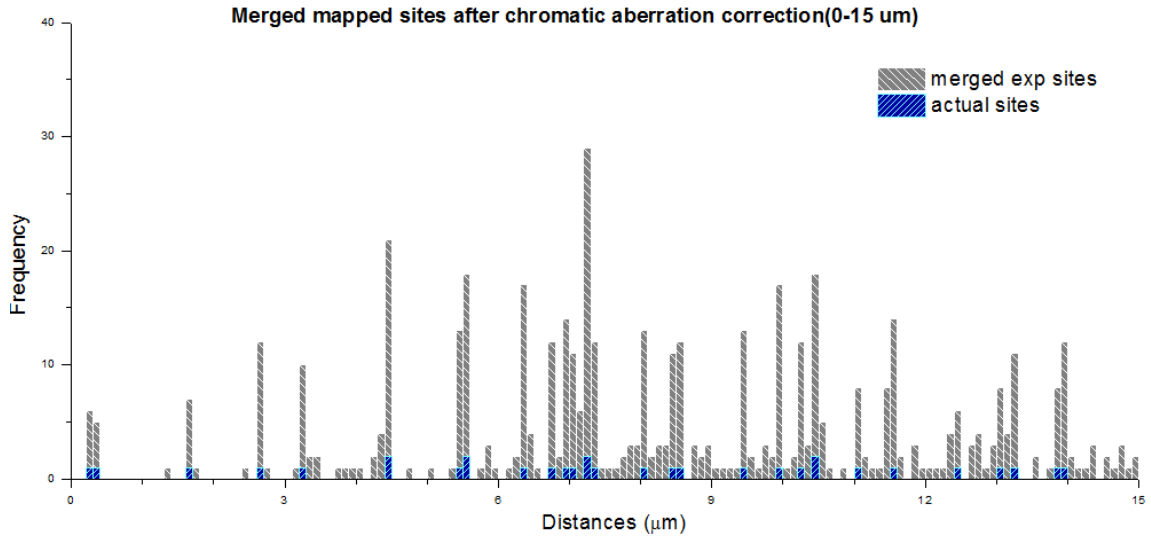


Figure A.9: Merged mapped sites after chromatic aberration correction range from range from 0 µm to 15 µm (0 kb to 44 kb).

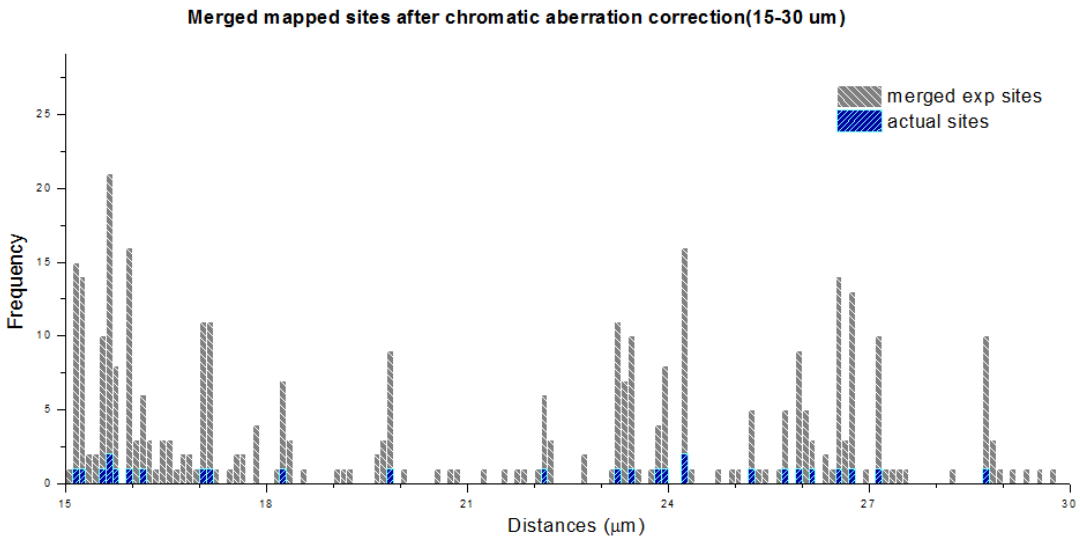


Figure A.10: Merged mapped sites after chromatic aberration correction range from 15 µm to 30 µm (44 kb to 88 kb)

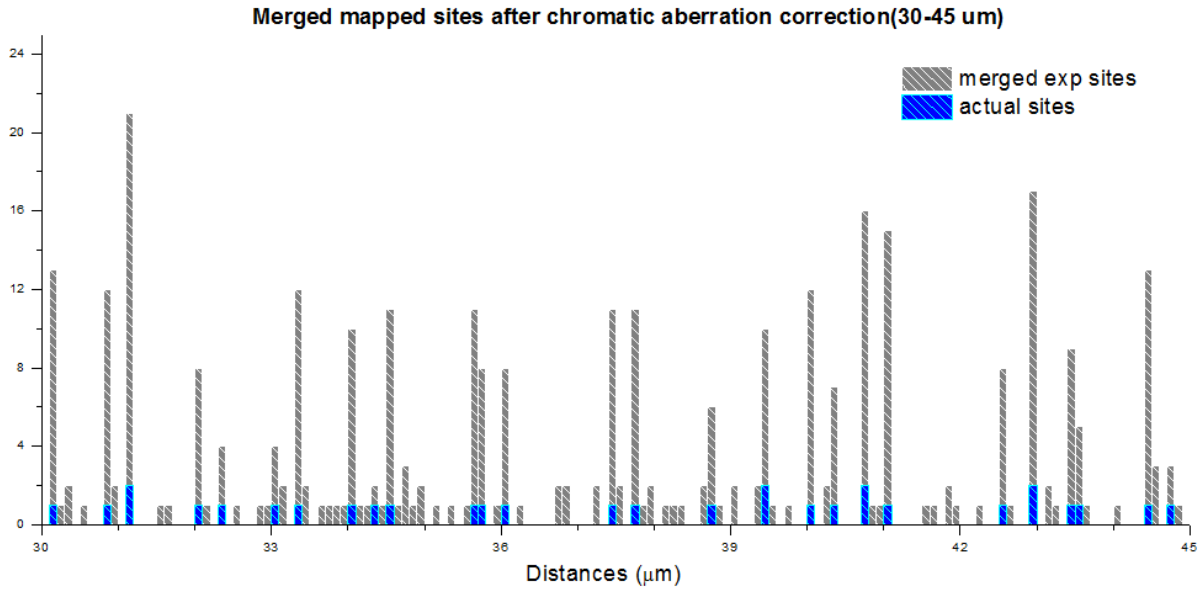


Figure A.11: Merged mapped sites after chromatic aberration correction range from range from 30 µm to 45 µm (88 kb to 132 kb).

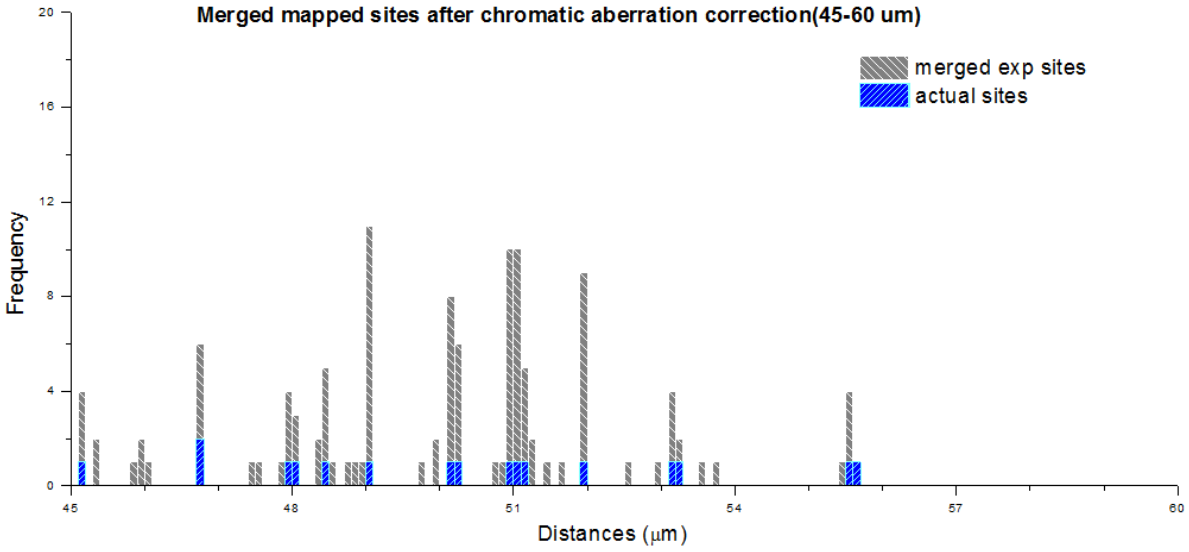


Figure A.12: Merged mapped sites after chromatic aberration correction range from range from 45 µm to 56 µm (132 kb to 165 kb).

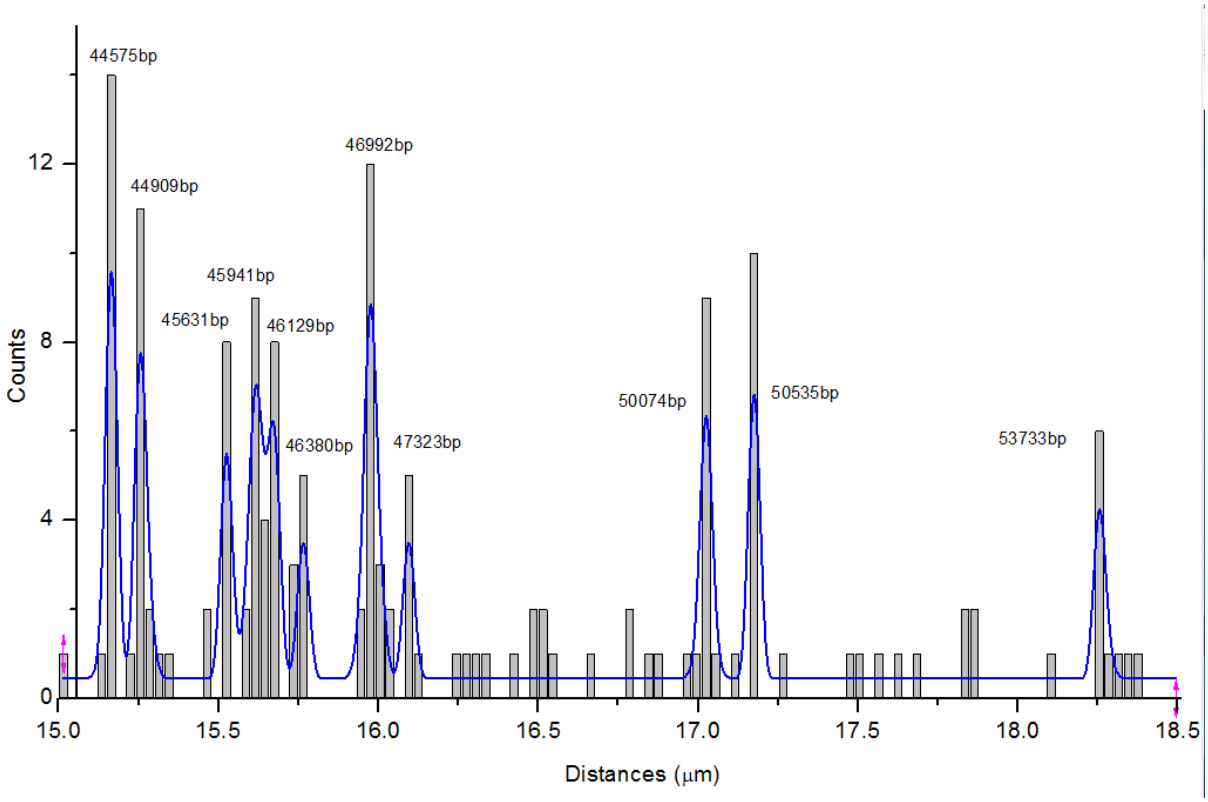


Figure A.13: A Subset region (range from 15 μm to 18 μm) of full merged data.

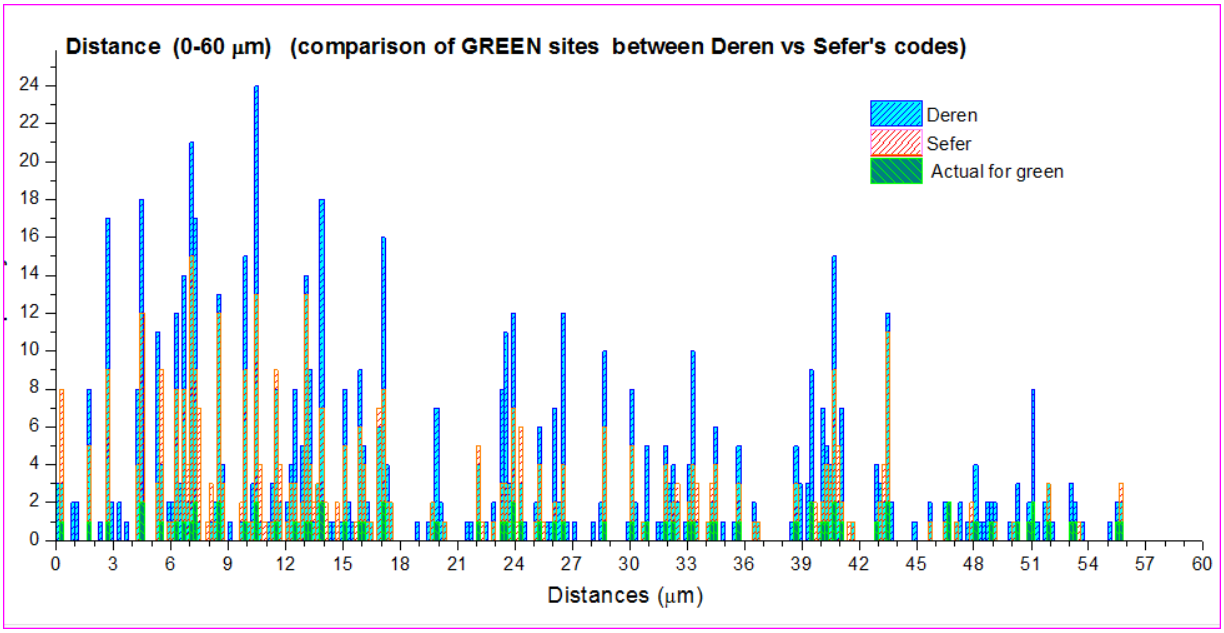


Figure A.14: Comparison of two different search algorithms on same data set. Both results agree each other meaning that any of programs are reliable to be used for analysis.

REFERENCES

1. Khajavinia, A. & Makalowski, W. What is 'junk' DNA, and what is it worth? *Sci. Am.* **296,** 104 (2007).
2. Elgar, G. & Vavouri, T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.* **24,** 344–352 (2008).
3. Pennisi, E. ENCODE Project Writes Eulogy for Junk DNA. *Science* **337,** 1159–1161 (2012).
4. Nóbrega, M. A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E. M. Megabase deletions of gene deserts result in viable mice. *Nature* **431,** 988–993 (2004).
5. Nevo, E. Evolution of genome-phenome diversity under environmental stress. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 6233–6240 (2001).
6. Koh, H. K., Geller, A. C., Miller, D. R., Grossbart, T. A. & Lew, R. A. Prevention and early detection strategies for melanoma and skin cancer. Current status. *Arch Dermatol* **132,** 436–443 (1996).
7. Rosas, F., Quesada, M., Lobo, J. A. & Sork, V. L. Effects of habitat fragmentation on pollen flow and genetic diversity of the endangered tropical tree Swietenia humilis (Meliaceae). *Biological Conservation* **144,** 3082–3088 (2011).
8. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
9. Hamosh, A. *et al.* Cystic fibrosis patients bearing both the common missense mutation, Gly->Asp at codon 551 and the ?F508 mutation are clinically indistinguishable from ?F508 homozygotes, except for decreased risk of meconium ileus. *Am J Hum Genet* **51,** 245–250 (1992).
10. INGRAM, V. M. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature* **178,** 792–794 (1956).
11. Bonilla, C. *et al.* Using genetic proxies for lifecourse sun exposure to assess the causal relationship of sun exposure with circulating vitamin D and prostate cancer risk. *Cancer Epidemiol. Biomarkers Prev.* (2013). doi:10.1158/1055-9965.EPI-12-1248
12. Mondal, P. *et al.* Comprehensive SNP scan of DNA repair and DNA damage response genes reveal multiple susceptibility Loci conferring risk to tobacco associated leukoplakia and oral cancer. *PLoS ONE* **8,** e56952 (2013).
13. Duggal, P. *et al.* Genome-wide association study of spontaneous resolution of hepatitis C virus infection: data from multiple cohorts. *Ann. Intern. Med.* **158,** 235–245 (2013).
14. Wolf, A. B., Caselli, R. J., Reiman, E. M. & Valla, J. APOE and neuroenergetics: an emerging paradigm in Alzheimer's disease. *Neurobiol. Aging* **34,** 1007–1017 (2013).
15. Da Costa, K.-A. *et al.* Common genetic polymorphisms affect the human requirement for the nutrient choline. *FASEB J.* **20,** 1336–1344 (2006).
16. Kohlmeier, M., Da Costa, K.-A., Fischer, L. M. & Zeisel, S. H. Genetic variation of folate-mediated one-carbon transfer pathway predicts susceptibility to choline deficiency in humans. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 16025–16030 (2005).
17. Zeisel, S. H. Nutritional genomics: defining the dietary requirement and effects of choline. *J. Nutr.* **141,** 531–534 (2011).
18. Compton, D. A. Mechanisms of aneuploidy. *Curr. Opin. Cell Biol.* **23,** 109–113 (2011).
19. Pellestor, F., Andréo, B., Anahory, T. & Hamamah, S. The occurrence of aneuploidy in human: lessons from the cytogenetic studies of human oocytes. *Eur J Med Genet* **49,** 103–116 (2006).
20. Sullivan, A. E., Silver, R. M., LaCoursiere, D. Y., Porter, T. F. & Branch, D. W. Recurrent fetal aneuploidy and recurrent miscarriage. *Obstet Gynecol* **104,** 784–788 (2004).
21. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nat Rev Genet* **13,** 189–203 (2012).

22. Paulsson, K. & Johansson, B. Trisomy 8 as the sole chromosomal aberration in acute myeloid leukemia and myelodysplastic syndromes. *Pathologie Biologie* **55,** 37–48 (2007).
23. Maurici, D. *et al.* Frequency and Implications of Chromosome 8 and 12 Gains in Ewing Sarcoma. *Cancer Genetics and Cytogenetics* **100,** 106–110 (1998).
24. Ozery-Flato, M., Linhart, C., Trakhtenbrot, L., Izraeli, S. & Shamir, R. Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy. *Genome Biology* **12,** R61 (2011).
25. Jacobsen, P. *et al.* An (11;21) translocation in four generations with chromosome 11 abnormalities in the offspring. A clinical, cytogenetical, and gene marker study. *Hum. Hered.* **23,** 568–585 (1973).
26. Ruiz Esparza-Garrido, R. *et al.* Duplication of the Miller-Dieker Critical Region in a Patient with a Subtelomeric Unbalanced Translocation t(10;17)(p15.3;p13.3). *Mol Syndromol* **3,** 82–88 (2012).
27. Kleffmann, W. *et al.* 5q31 Microdeletions: Definition of a Critical Region and Analysis of LRRTM2, a Candidate Gene for Intellectual Disability. *Mol Syndromol* **3,** 68–75 (2012).
28. Wallace, M. D. *et al.* Comparative Oncogenomics Implicates the Neurofibromin 1 Gene (NF1) as a Breast Cancer Driver. *Genetics* **192,** 385–396 (2012).
29. Wong, Y. F., Chung, T. K., Cheung, T. H., Nobori, T. & Chang, A. M. MTAP gene deletion in endometrial cancer. *Gynecol. Obstet. Invest.* **45,** 272–276 (1998).
30. Giglio, S. *et al.* Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am. J. Hum. Genet.* **71,** 276–285 (2002).
31. Jiang, Y.-H. *et al.* De novo and complex imbalanced chromosomal rearrangements revealed by array CGH in a patient with an abnormal phenotype and apparently 'balanced' paracentric inversion of 14(q21q23). *Am. J. Med. Genet. A* **146A,** 1986–1993 (2008).
32. Borg, K. *et al.* [Balanced chromosomal rearrangements resulting in intellectual disability. An analysis of 22 cases with application of CGH and FISH methods]. *Med Wieku Rozwoj* **13,** 81–93 (2009).
33. Lakich, D., Kazazian, H. H., Jr, Antonarakis, S. E. & Gitschier, J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat. Genet.* **5,** 236–241 (1993).
34. Nikiforov, Y. E., Koshoffer, A., Nikiforova, M., Stringer, J. & Fagin, J. A. Chromosomal breakpoint positions suggest a direct role for radiation in inducing illegitimate recombination between the ELE1 and RET genes in radiation-induced thyroid carcinomas. *Oncogene* **18,** 6330–6334 (1999).
35. Nikiforova, M. N. *et al.* Proximity of chromosomal loci that participate in radiation-induced rearrangements in human cells. *Science* **290,** 138–141 (2000).
36. Feuk, L. Inversion variants in the human genome: role in disease and genome architecture. *Genome Medicine* **2,** 11 (2010).
37. Mackie Ogilvie, C. & Scriven, P. N. Meiotic outcomes in reciprocal translocation carriers ascertained in 3-day human embryos. *Eur. J. Hum. Genet.* **10,** 801–806 (2002).
38. Choices, N. H. S. Down Syndrome - Causes - NHS Choices. (2013). at <http://www.nhs.uk/Conditions/Downs-syndrome/Pages/Causes.aspx>
39. Li, J. Y. *et al.* Detection of translocation t(11;14)(q13;q32) in mantle cell lymphoma by fluorescence in situ hybridization. *Am. J. Pathol.* **154,** 1449–1452 (1999).
40. Semple, C. A., Devon, R. S., Le Hellard, S. & Porteous, D. J. Identification of genes from a schizophrenia-linked translocation breakpoint region. *Genomics* **73,** 123–126 (2001).
41. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61,** 437–455 (2010).
42. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).
43. Sebat, J. *et al.* Large-Scale copy number polymorphism. *Science* **305,** 525–528 (2004).

44. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36,** 949–951 (2004).
45. Cappuzzo, F. *et al.* Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer. *J. Natl. Cancer Inst.* **97,** 643–655 (2005).
46. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307,** 1434–1440 (2005).
47. Costa, S. *et al.* DNA repair polymorphisms might contribute differentially on familial and sporadic breast cancer susceptibility: a study on a Portuguese population. *Breast Cancer Res. Treat.* **103,** 209–217 (2007).
48. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).
49. Gai, X. *et al.* Rare structural variation of synapse and neurotransmission genes in autism. *Mol. Psychiatry* **17,** 402–411 (2012).
50. Cook, E. H., Jr & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455,** 919–923 (2008).
51. St Clair, D. Copy number variation and schizophrenia. *Schizophr Bull* **35,** 9–12 (2009).
52. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods* **3,** 793–796 (2006).
53. Huang, B., Jones, S. A., Brandenburg, B. & Zhuang, X. Whole-cell 3D STORM reveals interactions between cellular structures with nanometer-scale resolution. *Nat. Methods* **5,** 1047–1052 (2008).
54. Zhuang, X. Nano-imaging with Storm. *Nat Photonics* **3,** 365–367 (2009).
55. Betzig, E. *et al.* Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science* **313,** 1642–1645 (2006).
56. Shroff, H. *et al.* Dual-color superresolution imaging of genetically expressed probes within individual adhesion complexes. *PNAS* **104,** 20308–20313 (2007).
57. Lippincott-Schwartz, J. & Patterson, G. H. Development and Use of Fluorescent Protein Markers in Living Cells. *Science* **300,** 87–91 (2003).
58. Garini, Y., Vermolen, B. J. & Young, I. T. From micro to nano: recent advances in high-resolution microscopy. *Curr. Opin. Biotechnol.* **16,** 3–12 (2005).
59. Hess, S. T., Girirajan, T. P. K. & Mason, M. D. Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy. *Biophysical Journal* **91,** 4258–4272 (2006).
60. Bock, H. *et al.* Two-color far-field fluorescence nanoscopy based on photoswitchable emitters. *Applied Physics B* **88,** 161–165 (2007).
61. Egner, A. *et al.* Fluorescence nanoscopy in whole cells by asynchronous localization of photoswitching emitters. *Biophys. J.* **93,** 3285–3290 (2007).
62. Fölling, J. *et al.* Photochromic Rhodamines Provide Nanoscopy with Optical Sectioning. *Angewandte Chemie International Edition* **46,** 6266–6270 (2007).
63. Lacoste, T. D. *et al.* Ultrahigh-resolution multicolor colocalization of single fluorescent probes. *PNAS* **97,** 9461–9466 (2000).
64. Yildiz, A. *et al.* Myosin V Walks Hand-Over-Hand: Single Fluorophore Imaging with 1.5-nm Localization. *Science* **300,** 2061–2065 (2003).
65. Dertinger, T. *et al.* Superresolution optical fluctuation imaging (SOFI). *Adv. Exp. Med. Biol.* **733,** 17–21 (2012).
66. Linde, S. van de *et al.* Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nature Protocols* **6,** 991–1009 (2011).
67. Simonson, P. D., Rothenberg, E. & Selvin, P. R. Single-Molecule-Based Super-Resolution Images in the Presence of Multiple Fluorophores. *Nano Lett.* **11,** 5090–5096 (2011).

68. Gustafsson, M. G. L. Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 13081–13086 (2005).
69. Donnert, G. *et al.* Macromolecular-scale resolution in biological fluorescence microscopy. *PNAS* **103,** 11440–11445 (2006).
70. Joo, C. & Ha, T. Single-Molecule FRET with Total Internal Reflection Microscopy. *Cold Spring Harb Protoc* **2012,** pdb.top072058 (2012).
71. Dani, A., Huang, B., Bergan, J., Dulac, C. & Zhuang, X. Superresolution imaging of chemical synapses in the brain. *Neuron* **68,** 843–856 (2010).
72. Bates, M., Huang, B., Dempsey, G. T. & Zhuang, X. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science* **317,** 1749–1753 (2007).
73. Gordon, M. P., Ha, T. & Selvin, P. R. Single-molecule high-resolution imaging with photobleaching. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 6462 –6465 (2004).
74. Churchman, L. S., Ökten, Z., Rock, R. S., Dawson, J. F. & Spudich, J. A. Single molecule high-resolution colocalization of Cy3 and Cy5 attached to macromolecules measures intramolecular distances through time. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 1419 –1423 (2005).
75. Balci, H., Ha, T., Sweeney, H. L. & Selvin, P. R. Interhead distance measurements in myosin VI via SHRImP support a simplified hand-over-hand model. *Biophys. J.* **89,** 413–417 (2005).
76. Herrick, J. & Bensimon, A. Single molecule analysis of DNA replication. *Biochimie* **81,** 859–871 (1999).
77. Fazio, T. A., Visnapuu, M., Greene, E. C. & Wind, S. J. Fabrication of Nanoscale 'Curtain Rods' for DNA Curtains Using Nanoimprint Lithography. *J Vac Sci Technol A* **27,** 3095–3098 (2009).
78. Visnapuu, M.-L., Fazio, T., Wind, S. & Greene, E. C. Parallel arrays of geometric nanowells for assembling curtains of DNA with controlled lateral dispersion. *Langmuir* **24,** 11293–11299 (2008).
79. Kim, S., Blainey, P. C., Schroeder, C. M. & Xie, X. S. Multiplexed single-molecule assay for enzymatic activity on flow-stretched DNA. *Nat Meth* **4,** 397–399 (2007).
80. Washizu, M., Kurosawa, O., Arai, I., Suzuki, S. & Shimamoto, N. Applications of electrostatic stretch-and-positioning of DNA. *IEEE Transactions on Industry Applications* **31,** 447–456 (1995).
81. Kabata, H. *et al.* Visualization of single molecules of RNA polymerase sliding along DNA. *Science* **262,** 1561 –1563 (1993).
82. Cao, H. *et al.* Fabrication of 10 nm enclosed nanofluidic channels. *Applied Physics Letters* **81,** 174 (2002).
83. Reisner, W., Larsen, N. B., Flyvbjerg, H., Tegenfeldt, J. O. & Kristensen, A. Directed self-organization of single DNA molecules in a nanoslit via embedded nanopit arrays. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 79–84 (2009).
84. Reisner, W. *et al.* Single-molecule denaturation mapping of DNA in nanofluidic channels. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 13294–13299 (2010).
85. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences* **100,** 3960 –3964 (2003).
86. Kartalov, E. P., Unger, M. A. & Quake, S. R. Polyelectrolyte surface interface for single-molecule fluorescence studies of DNA polymerase. *BioTechniques* **34,** 505–510 (2003).
87. Yardimci, H., Loveland, A. B., Van Oijen, A. M. & Walter, J. C. Single-molecule analysis of DNA replication in Xenopus egg extracts. *Methods* **57,** 179–186 (2012).
88. Yildiz, A., Tomishige, M., Vale, R. D. & Selvin, P. R. Kinesin walks hand-over-hand. *Science* **303,** 676–678 (2004).
89. Kural, C. *et al.* Tracking melanosomes inside a cell to study molecular motors and their interaction. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 5378–5382 (2007).

90. Baday, M. *et al.* Multicolor Super-Resolution DNA Imaging for Genetic Analysis. *Nano Lett.* **12,** 3861–3866 (2012).
91. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology* **30,** 771–776 (2012).
92. Han, Y.-W. *et al.* Direct observation of DNA rotation during branch migration of Holliday junction DNA by Escherichia coli RuvA-RuvB protein complex. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 11544–11548 (2006).
93. Van der Heijden, T. *et al.* Homologous recombination in real time: DNA strand exchange by RecA. *Mol. Cell* **30,** 530–538 (2008).
94. Spies, M. *et al.* A molecular throttle: the recombination hotspot chi controls DNA translocation by the RecBCD helicase. *Cell* **114,** 647–654 (2003).
95. Prasad, T. K. *et al.* A DNA-translocating Snf2 molecular motor: Saccharomyces cerevisiae Rdh54 displays processive translocation and extrudes DNA loops. *J. Mol. Biol.* **369,** 940–953 (2007).
96. Quessada-Vial, A. & Oijen, A. M. van. How DNA-repair proteins find their targets. *PNAS* **109,** 18243–18244 (2012).
97. Aitken, C. E., Marshall, R. A. & Puglisi, J. D. An Oxygen Scavenging System for Improvement of Dye Stability in Single-Molecule Fluorescence Experiments. *Biophys J* **94,** 1826–1835 (2008).
98. Rasnik, I., McKinney, S. A. & Ha, T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat Meth* **3,** 891–893 (2006).
99. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307,** 1434–1440 (2005).
100. Hollox, E. J. *et al.* Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.* **40,** 23–25 (2008).
101. Wong, K. K. *et al.* A Comprehensive Analysis of Common Copy-Number Variations in the Human Genome. *Am J Hum Genet* **80,** 91–104 (2007).
102. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics* **38,** 75–81 (2005).
103. Locke, D. P. *et al.* Linkage Disequilibrium and Heritability of Copy-Number Polymorphisms within Duplicated Regions of the Human Genome. *Am J Hum Genet* **79,** 275–290 (2006).
104. Jing, J. *et al.* Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *PNAS* **95,** 8046–8051 (1998).
105. Xiao, M. *et al.* Determination of haplotypes from single DNA molecules: a method for single-molecule barcoding. *Hum. Mutat.* **28,** 913–921 (2007).
106. Xiao, M. *et al.* Rapid DNA mapping by fluorescent single molecule detection. *Nucl. Acids Res.* **35,** e16–e16 (2007).
107. Qu, X., Wu, D., Mets, L. & Scherer, N. F. Nanometer-Localized Multiple Single-Molecule Fluorescence Microscopy. *PNAS* **101,** 11298–11303 (2004).
108. Morgan, R. D., Calvet, C., Demeter, M., Agra, R. & Kong, H. Characterization of the specific DNA nicking activity of restriction endonuclease N.BstNBI. *Biol. Chem.* **381,** 1123–1125 (2000).
109. Das, S. K. *et al.* Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res* **38,** e177 (2010).
110. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
111. Zhou, S. *et al.* Validation of rice genome sequence by optical mapping. *BMC Genomics* **8,** 278 (2007).
112. Zhou, S. *et al.* A Single Molecule Scaffold for the Maize Genome. *PLoS Genet* **5,** e1000711 (2009).
113. Teague, B. *et al.* High-resolution human genome structure by single-molecule analysis. *PNAS* **107,** 10848–10853 (2010).

114. Wu, C., Schramm, T. M., Zhou, S., Schwartz, D. C. & Talaat, A. M. Optical mapping of the Mycobacterium avium subspecies paratuberculosis genome. *BMC Genomics* **10,** 25 (2009).
115. Latreille, P. *et al.* Optical mapping as a routine tool for bacterial genome sequence finishing. *BMC Genomics* **8,** 321 (2007).
116. Church, D. M. *et al.* Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. *PLoS Biol* **7,** e1000112 (2009).
117. Das, S. K. *et al.* Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Research* (2010). doi:10.1093/nar/gkq673
118. Dervan, P. B. & Bürli, R. W. Sequence-specific DNA recognition by polyamides. *Curr Opin Chem Biol* **3,** 688–693 (1999).
119. Chan, E. Y. *et al.* DNA mapping using microfluidic stretching and single-molecule detection of fluorescent site-specific tags. *Genome Res.* **14,** 1137–1146 (2004).
120. Neely, R. K. *et al.* DNA fluorocode: A single molecule, optical map of DNA with nanometre resolution. *Chem. Sci.* **1,** 453–460 (2010).
121. Goshtasby, A. Image registration by local approximation methods. *Image Vision Comput.* **6,** 255–261 (1988).
122. Goshtasby, A. Piecewise linear mapping functions for image registration. *Pattern Recognition* **19,** 459–466 (1986).