

# Chromosome Conformation Capture on Chip (4C)

Meeting genomic neighbors

Marieke Simonis

The work presented in this thesis was performed at the department of Cell Biology at the Erasmus MC in Rotterdam.

cover photography: De Roode Optics

printed by: Gildeprint, Enschede

ISBN: 9789071382772

# Chromosome Conformation Capture on Chip (4C)

Meeting genomic neighbors

## Chromosome Conformation Capture on Chip (4C)

Genomische buren ontmoeten

Proefschrift

ter verkrijging van de graad van doctor aan de  
Erasmus Universiteit Rotterdam  
op gezag van de rector magnificus  
Prof.dr. S.W.J. Lamberts  
en volgens het besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op  
woensdag 17 december 2008 om 15.45 uur

door

*Maria Johanna Simonis*

geboren te Utrecht



## Promotiecommissie

Promotor: Prof.dr. F.G. Grosveld

Overige leden: Prof.dr. J.N.J. Philipsen  
Prof.dr. C.P. Verrijzer  
Dr. J.P.P. Meijerink

Copromotor: Dr. W.L. de Laat

*Na elke bocht ontdek je wat.*



# Contents

	Outline of this thesis	9
<b>Chapter 1</b>	Introduction to genome structure	11
<b>Chapter 2</b>	Genome-wide technologies	29
<b>Chapter 3</b>	FISH-eyed and genome-wide views on the spatial organization of gene expression	43
<b>Chapter 4</b>	Nuclear organization of active and inactive chromatin domains uncovered by 3C on chip (4C)	67
<b>Chapter 5</b>	An evaluation of 3C derived methods to capture DNA interactions	89
<b>Chapter 6</b>	High-resolution identification of chromosomal rearrangements by 4C technology	109
<b>Chapter 7</b>	A pilot study for Chromosome Conformation Capture <i>Se</i> -Quencing (4C-Q)	133
<b>Chapter 8</b>	General discussion and future directions	149
	Summary	160
	Samenvatting	163
	Samenvatting voor niet ingewijden	166
	Curriculum Vitae	168
	Dankwoord	170





## Outline of this thesis

This thesis describes the development of a novel technology, Chromosome Conformation Capture on Chip (4C), which can be used for two different applications; the investigation of the folded structure of chromosomes and the detection of genomic rearrangements.

The first three chapters serve as an introduction to the areas of research 4C technology can be implemented in and discuss the additive value of 4C for different subjects. In **Chapter 1** several aspects of genome biology are introduced. The emphasis is on the organization of the cell nucleus, packaging and folding of chromosomes, transcriptional regulation and variation in the genome sequence. **Chapter 2** describes commonly used methods in genome wide studies; microarray technology and massive parallel sequencing. An overview is presented of the different types of genome wide studies that have been performed. **Chapter 3** describes recent literature on the relation between spatial organization of the genome and regulation of gene expression. The impact of 4C on this field of research is discussed.

**Chapter 4** describes the development of 4C and its first application as a tool to investigate the spatial organization of the genome.

In **Chapter 5** technical aspects of 4C technology are discussed in detail. 4C is an adapted form of earlier developed chromosome conformation capture (3C). Other 3C derived methods are described and compared in this chapter.

**Chapter 6** describes the application of 4C to detect genomic rearrangements.

4C was adapted from a microarray based method to sequencing based 4C-Q. A pilot study to test the potential of 4C-Q is described in **Chapter 7**.

A general discussion and future directions are presented in **Chapter 8**.



# 1

## Introduction to genome structure

## Introduction to genome structure

All organisms ranging from prokaryotes such as bacteria to eukaryotes including plants animals and humans pass on traits to their offspring. The full construction plan of an organism, including the parental traits, is present in the single cell from which the entire organism develops. In the 19<sup>th</sup> century chromosomes were first seen in microscopy studies and much later they were found to contain the construction plan, also referred to as the genetic material. Initially it was believed that the genetic information was encoded in proteins, but in 1944 it was established that deoxyribonucleic acid (DNA) molecules perform this important task<sup>1</sup>. DNA is built up of four nucleotides, adenine (A), guanine (G), thymine (T) and cytosine (C). Watson and Crick uncovered the double helical (B-DNA) structure of DNA in the 1950's<sup>2</sup>. Due to selective pairing (adenine only pairs with thymine, and guanine only pairs with cytosine) the two helical strands are a template for each other and the DNA can be faithfully duplicated.

The human genome consists of 3 billion base pairs, divided over 23 chromosomes. Despite the availability of complete genome sequences ([www.ensembl.org](http://www.ensembl.org)), many aspects of genome function are still poorly understood. One of the important remaining questions is how the genome is spatially organized. When all the DNA from one human cell is linearly aligned (a paternal and a maternal copy of the genome), the strand is 2 meters, longer than an average human being. It is not yet known how the DNA folded and how the spatial organization relates to genomic functions, such as gene expression.

The linear structure of the genome, the sequence, is also still examined. The genome project resulted in a reference genome, but there is variation in the genome sequence, many small differences occur within the human population, the characterization of which only just started. Specific errors in the genome are also of interest, because they can lead to disease.

This chapter aims to provide an overview of mammalian genome function, with emphasis on the environment of the genome, the packaging and folding of the chromosomes, the process of transcription and the effects of variation and errors in the genome sequence.

### **The cell nucleus; an organized home for the genome**

Eukaryotic genomes are stored in the cell nucleus with a diameter of around 5 micrometer. The cell nucleus not only contains the 2 m of DNA, but is also rich in protein. The protein concentration is estimated to be 0.1 g/cm<sup>3</sup> on average in the nucleoplasm, and locally

even higher<sup>3</sup>. The proteins serve in packaging of the genome and in genomic processes such as transcription and DNA repair. Many proteins in the nucleus do not have a uniform spatial distribution, but are present in local accumulations called foci or bodies.

The largest and most well studied body is the nucleolus. In the nucleolus ribosomal DNA (rDNA) repeats come together, to be transcribed by RNA polymerase I (RNA PolI). In addition, the nuclear body contains proteins and RNA molecules needed to generate ribosomes. Hence, the nucleolus is considered to be a sub-compartment of the nucleus that is specialized in the process of ribosome generation<sup>4</sup>.

Other important nuclear processes have also been found to occur in specialized foci, most notably DNA repair and replication<sup>5,6</sup>. The protein complex involved in gene transcription, RNA Polymerase II (RNA PolII), has been shown to accumulate in foci called transcription factories<sup>7</sup>. Compared to other foci RNA PolII foci are small and are present in large numbers of up to 2000 per nucleus. From electron microscopy (EM) studies in HeLa cells it was estimated that each focus contains around 8 polymerases<sup>8</sup>. The RNA PolII foci are sites of active transcription, but genes can also be transcribed outside transcription factories<sup>7</sup>.

Proteins involved in splicing are also found at discrete sites in the nucleus, called speckles or splicing factor compartments (SFCs). The speckles are often not directly associated with the chromatin and sites of active splicing. It has been suggested that the speckles are a "storage compartment" for splicing factors. In this model the splicing factors leave a speckle to freely roam the nucleus for sites of transcription and engage in splicing there<sup>9</sup>.

Other protein bodies, such as the promyelocytic leukemia oncoprotein (PML) bodies and the Cajal bodies (CB) have very diverse protein contents and their function remains largely unresolved<sup>9,10</sup>.

A parallel is often drawn between the nuclear bodies and the organelles in the cytosol. The compartmentalization in organelles allows processes that require different environments to co-occur within one cell. Moreover, the restricted volume in organelles can increase the efficiency of the reactions inside the compartments. Likewise, in the nucleus the local accumulation of different proteins involved in a nuclear process, such as the assembly of ribosomes, may facilitate these processes. In the cellular organelles the membrane forms a physical barrier that allows a controlled in and out-flux of the compartments. The nuclear bodies do not have such a barrier; they are merely an accumulation of specific proteins. In fact, it has been shown that large dextrans can easily permeate the nuclear bodies, with exception of the most dense fibrillar part of the nucleolus. This suggests that the bodies are sponge-like structures<sup>3</sup>. The absence of a physical barrier raises the question of how the nuclear bodies are built and maintained.

The emerging view is that the bodies arise through a process termed self-organization; they are dynamic structures build from individual interactions between its components<sup>11</sup>. The dynamics of the nuclear bodies is twofold, both the structure as a whole and the in and out flux of each individual component is dynamic. An example of the dynamics of the structures is the finding that repair foci are only formed after DNA damage has occurred, even though repair proteins are always present in the nucleoplasm<sup>6</sup>. Similarly, the nucleoli are formed after mitosis, when rRNA transcription starts<sup>12</sup>. When rRNA transcription is inhibited the nucleoli disassemble<sup>13</sup>. The dynamics of the sub-units of the nuclear bodies has been demonstrated by fluorescent recovery after photobleaching (FRAP). There is a rapid exchange of proteins residing in the structures and proteins freely roaming the nucleoplasm<sup>14</sup>.

The process of self organization is possibly aided by the effect of macromolecular crowding, also called the excluded volume effect<sup>15,16</sup>. The space in which molecules can freely move (the reaction volume) is not only dependent on the concentration of the solutes, but also on the sizes of the molecules in the solution. The volume taken up by a molecule itself can not be entered by another molecule (the excluded volume). The minimal distance between two large molecules is much bigger than the minimal space between two small or a large and a small molecule. Thus, large molecules (such as the nuclear protein complexes) have a smaller space in which they can freely move. Therefore, when present in the same amount of molecules per volume, the effective concentration of large molecules is higher than the effective concentration of small molecules. By elevating the effective concentration, interaction frequencies can increase<sup>15</sup>. This effect is exploited in experiments, for example by the addition of polyethylene glycol to increase the efficiency of enzymes such as DNA ligase or DNA polymerase<sup>17</sup>. Several studies support the relevance of macromolecular crowding in the process of nuclear organization. New bodies appear in the nucleus after exogenous addition of DNA, protein, viral particles or oligonucleotides. Endogenous nuclear structures, the PML bodies, disassemble when the effect of molecular crowding is relieved by hypertonic swelling of the nucleus. Reassembly of the bodies can be established by relocating the nuclei to isotonic medium or by adding high molecular weight dextrans<sup>16,18</sup>.

Note that even if the assembly of bodies and foci relies on the effect of macromolecular crowding, this does not necessarily imply that the structures are non-functional. An alternative hypothesis for the construction of nuclear bodies is that they are assembled on the nuclear matrix (see below).

### **Packaging of DNA into chromatin**

The genome is not floating around freely in the nucleus; it is packaged by proteins. The combination of DNA and the associated proteins is called chromatin. Early EM studies have revealed the basic structure of the chromatin, the 10 nm fiber, which appears as “beads on a string”. The “beads” seen in EM experiments are nucleosomes. A nucleosome consists of ~146 bp of DNA wrapped around an octamer of histone proteins; H2A, H2B, H3 and H4, two of each. The nucleosomal histones form a globular structure, from which only the N-terminal tails of the histones protrude. The nucleosomes are connected to each other via a 29-43 bp stretch of linker DNA<sup>19</sup>. It has been postulated that wrapping the DNA around the histones is necessary to counteract the macromolecular crowding effect that would lead to intertwining of the DNA<sup>20</sup>. The effects of macromolecular crowding are only expected in eukaryotes, because the DNA concentration in the nuclei is higher than in prokaryotes (which store their genome in the cytoplasm). Prokaryotes indeed lack histones. The interaction of the positively charged histones with the negatively charged DNA was suggested to make nucleosome assembly energetically favorable over DNA self-association. Although this could be the evolutionary “raison d’être” of nucleosomes, it is clear that chromatin is also indispensable for the regulation of genomic processes, such as DNA repair and transcription<sup>21</sup>. Regulation at the level of chromatin is possible because the fiber can be modified in various ways: DNA can be methylated, nucleosomes can move and histones can be replaced or chemically modified.

Methylation of DNA occurs at CpG dinucleotides and is considered to be the most stable chromatin modification. DNA methylation is preserved during DNA replication and is therefore heritable<sup>22</sup>.

Packaging of the genome into nucleosomes needs to be controlled, because protein binding sites can become inaccessible if they are wrapped around histones. ATP dependent chromatin remodeling complexes can change the position of a nucleosome on the genome, by sliding the nucleosome along the DNA. There are four families of chromatin remodeling complexes, the SWI/SNF, ISWI, CHD and INO80-SWR1 family<sup>23</sup>. Different complexes can have slightly different modes of action. Possibly, this large diversity in remodeling complexes is necessary because remodeling is important for several very different nuclear processes<sup>24</sup>.

Not only the DNA, but also the histones in a nucleosome can be altered. There are alternative histone variants for the core histones that can be incorporated into the nucleosomes. Examples include CENPA<sup>25</sup>, a H3 variant found at centromeric regions and H3.3<sup>26</sup>, found at actively transcribed genes.

A large part of the nucleosomal diversity is achieved by covalent modification of the histone amino acid residues, mostly at the histone tails. The histones can be methylated, acetylated, phosphorylated, ubiquitinated and SUMO-ylated. Modifications of the histones can affect the function of nucleosomes in different ways. Covalent modification can alter the chemical properties of the histones. For example, acetylation changes the charge of the histones and thereby loosens the interaction with the DNA. Other modifications affect the interacting properties of nucleosomes by creating or destroying binding sites for chromatin associated proteins. The modifications are dynamic; both enzymes that add and those that remove the chemical adducts are present in the nucleus. The enzymes are targeted to the genome and the local balance between the two types of enzymes determines the final mode of modification<sup>27</sup>.

### Gene transcription

Ultimately, the main function of the genome is the storage of protein and RNA recipes; the genes. Controlling the transcriptional activity of genes is an intricate process that requires the coordination of a large amount of protein complexes and a variety of sequence elements in the genome itself.

Initiation of transcription is a cooperation of an estimated 40 proteins, with a combined mass of 2MDa<sup>28</sup>. The transcribing unit, RNA polymerase II (RNA PolII) does not recognize the DNA on its own, but requires general transcription factors, such as TFIIA,-B and -D, that recognize sequences in the core promoter. RNA PolII and the general transcription factors together form the pre-initiation complex (PIC) and their binding results in basal transcription *in vitro*, but *in vivo* additional proteins are needed. PIC formation can be inhibited or facilitated by a diverse set of co-factors, many of which act by chromatin modification.

The gene specificity of transcriptional regulation is largely determined by sequence specific transcription factors (TF's) that can affect transcription levels by recruiting or inhibiting general transcription factors and co-factors on a particular gene. Binding sites for TFs are not only found in the promoter region, but can be located up to a megabase away. The combination of TFs present in a cell ultimately determines which genes are expressed.

Modification of the chromatin fiber is an important factor in the process of transcription<sup>27</sup>. The nucleosomes form a barrier both for binding of protein complexes and for transcription elongation. Therefore chromatin remodeling enzymes are needed to move nucleosomes and allow efficient transcription. Histone modifying complexes are also



involved in gene regulation. Active genes are generally characterized by hyper-acetylated histones, tri-methylation of lysine residue 4 of H3 (H3K4) at the promoter, methylated H3K36 progressively towards their 3'end and methylated H3K79 throughout the transcribed region. Silent genes are associated with deacetylated histones and methylated H3K9, H4K20 and H3K27. Not every histone modification has a well described function, but at least some serve as binding sites for chromatin proteins. For example, methylated H3K9 can be bound by heterochromatin protein 1 (HP1), which is thought to form a tight chromatin structure. Methylated H3K27 is recognized by Polycomb (Pc), a transcriptional silencing complex.

Direct modification of the sequence by DNA methylation occurs in some exceptional cases of gene regulation, for example, X-chromosome inactivation in female mammals<sup>29</sup> and parent-of-origin-specific gene expression (imprinting)<sup>30</sup>.

Not every gene requires the same protein complexes for its expression<sup>28</sup>. Mechanisms of transcriptional induction depend for example on the surrounding sequences and the chromatin the gene is embedded in. Induction of genes surrounded by heterochromatin requires special sequence elements, such as locus control regions and insulators. Locus control regions contain a combination of regulatory sequences that allow transcription of genes independent of the chromatin they are embedded in<sup>31</sup>. Insulators can serve as boundary elements that physically block the effect of neighboring heterochromatin<sup>32</sup>.

A different factor that could be involved in regulating gene transcription is the nuclear environment of a gene locus. Several studies have shown a relationship between the transcriptional status of investigated gene loci and the position in the nucleus at which they are found. For example, several genes were found to be localized near heterochromatin in their transcriptionally inactive state<sup>33</sup>. The relation between locus positioning and gene expression, and the insights gained with 4C on this subject are discussed in Chapter 3 of this thesis<sup>34</sup>.

### **Higher order structure of chromatin, how do chromosomes fold?**

During mitosis chromosomes form condensed structures, which can easily be discerned when viewed under a microscope. After cell division, the chromosomes decondense to tightly packed heterochromatin and more loosely packed euchromatin. Each chromosome takes up its own space, the chromosome territory<sup>35</sup>, but different chromosomes can also intermingle<sup>36,37</sup>. Extensive folding of the DNA is necessary in an interphase cell to fit the long strands in the nuclear volume. Despite this compacted structure genomic processes such as transcription can still occur in a controlled manner. Moreover, the DNA does not

become entangled (to a major degree) and can condense into ordered mitotic structures for a new round of cell division. These observations suggest the genome is folded in an organized manner. The progression of our understanding of higher order chromosomal folding has gone hand in hand with advances in technology<sup>38</sup>.

#### *Condensation of chromatin*

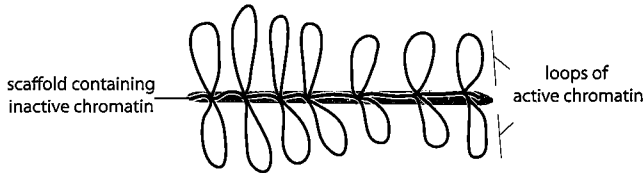
An early major technological advance was the development of electron microscopy in the 1950's. Using this technique it was first recognized that chromatin exists as fibers and that the basic structure is a 10 nm fibre<sup>39</sup>. *In vitro* experiments revealed that the nucleosomes fold into a 30 nm fibre when linker histone H1 is present<sup>40</sup>. The formation of this higher order structure is dependent on the histones and on salt concentrations, but the exact mechanisms of the process are not fully understood and its occurrence *in vivo* is still debated<sup>41</sup>.

Bruce and Belmont studied the unfolding of chromosomes during G1 and described that mitotic chromosomes unfold via 100-130 nm intermediates to 60-80 nm fibers, which locally decondense to 30 nm fibres<sup>42</sup>. The intermediates were called chromonema fibers. EM technology advanced and electron spectroscopic imaging (ESI) was developed, with which protein and nucleic acids containing structures can be discerned at EM resolution. Dhegani et al postulated based on ESI experiments, that thick chromatin fibers consist of a few 10 and 30 nm fibers that are crossing each other, rather than one highly compacted fibre<sup>38</sup>.

#### *The lampbrush model*

Sectioned nuclei can be used to study the thickness and the constituents of the fibers, but higher order structures can not be visualized, because they are severed. A different approach to studying chromosomal organization is to isolate intact chromosomes and visualize the chromatin fibers. One of the most famous examples of this type of experiment is the visualization of the chromosomes of amphibian oocytes<sup>43</sup>. These meiotically paired chromosomes are unusually thick and therefore easy to visualize. The meiotic chromosomes are configured into a structure that resembles a lampbrush (**Fig. 1.1**). They have a condensed linear axis, or scaffold from which large chromatin loops extend that are actively transcribed. These lampbrush chromosomes have served as an important model for chromosomal organization, and the finding initiated quests both for a scaffold and for looped structures in mammalian chromosomes. In mammalian metaphase chromosomes both a scaffold and loops can be visualized. 30-90 kb loops

are then seen extending from a rigid core of the metaphase chromosomes, but only after depletion of the histones<sup>44</sup>. Visualizing the organization of interphase chromosomes has been more challenging.



**Figure 1.1 Schematic representation of a lampbrush chromosome.**

### *Scaffold structures in the interphase nucleus*

The nuclear scaffold or nuclear matrix is a subject of great controversy with strong opinions in favour<sup>45</sup> and against<sup>46</sup> its existence. Different methods have been developed that show a fibrillar structure that remains after removing the majority of the chromatin and free proteins in the nucleus<sup>47</sup>. This structure was called the nuclear matrix. It is only seen after rigorous removal of all the other constituents of the nucleus. The nucleus is a tightly balanced environment. All the different extraction methods involve changing the ionic strength in the (remnants of) the nucleus. Moreover, by extracting the majority of the chromatin the largest source of anions in the nucleus (the DNA) is removed, altering the chemical balance *per se*. These types of alterations are known to change the binding properties of proteins and could lead to their aggregation into a fibrillar structure<sup>46,48</sup>. In favor of the nuclear matrix model, very different extraction methods result in a similar mesh-like ultra-structure in the nuclear remnants<sup>45,46</sup>. However, despite many efforts visualization of a matrix structure in untreated cells has not been achieved. Until a fibrillar component is identified and visualized, the nuclear matrix is likely to remain controversial. A structure that is clearly visible in the nuclear lamina. The nuclear lamina is a filamentous structure that is attached to the inner membrane of the nucleus and provides structural support. Lamin proteins have been shown to interact with chromatin and it has been proposed that the lamina may serve as an anchorage point for the chromatin. This can not be a static anchorage, because the chromatin-lamin associations are dynamic<sup>49</sup>.

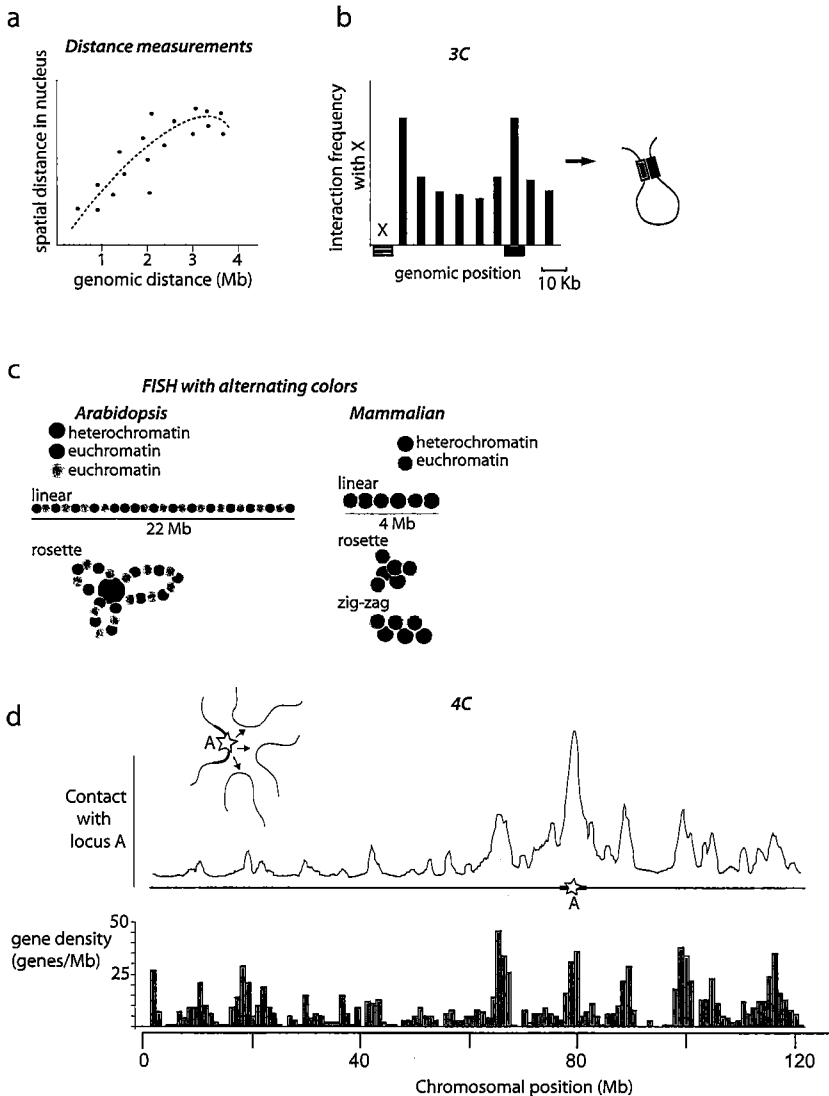
### *Chromatin loops*

The detection of loops in interphase nuclei has also been a challenge. How do you measure the presence of loops, when you can not visualize them directly? One way to tackle this problem is to measure the spatial distance in the nucleus between different

regions on a chromosome (**Fig. 1.2a**). If chromosomal organization is random, the average spatial distance is expected to be larger between regions that are further away from each other on the genome. If the DNA has a looped structure, the average spatial distance between the chromosomal segments at the base of the loop is smaller than it would be in a random situation. Using FISH and 3D microscopy, two studies measured the average spatial distance of multiple pairs of sequences that each had a different genomic distance between them. Plotting the spatial distance of all the pairs against their genomic distance showed a slightly curved line. Modeling of these data can predict the existence and the size of loops. One study reported the existence of giant ~3 Mb loops<sup>50</sup>, whereas a second study predicted the existence of ~120 kb<sup>51</sup>. These studies assume an ordered looped structure, with loops of comparable sizes across the chromosome. A similar approach was employed recently for a detailed analysis of the structure of the 2 Mb gene locus of the immunoglobulin heavy-chain<sup>52</sup>.

Specific relatively small loops (~10-600 kb) in selected regions of chromosomes can be investigated using a biochemical assay, called Chromosome Conformation Capture (3C) (**Fig. 1.2b**)<sup>53,54</sup>. In this assay interaction frequencies (rather than distances) between a “bait” restriction fragment and other selected DNA fragments are measured (for detailed description of the technique see Chapter 5). Similar to the experiments described above, if an increase in interaction frequencies is seen with increasing genomic distance from the bait, this demonstrates the presence of a loop in the chromatin. Using 3C, specific loops have been found within gene loci (described in detail in Chapter 3)<sup>54</sup>. The discovery of these structures has elicited hypotheses about a function of specific looped structures in the regulation of gene expression, rather than them just being a way to store the genome in an orderly fashion. Some loops were described to be connected to the nuclear matrix and the authors suggest that sequestering of a gene to this structure could be a mechanism of transcriptional silencing<sup>55</sup>. In several gene loci the contacts established at the base of the loop are functionally relevant. For example, in the  $\beta$ -globin locus physical contacts are established between the gene promoter of the active gene and the regulatory elements spread over 200 kb<sup>54</sup>. Tissue specific transcription factors<sup>56</sup> and a ubiquitous factor, CTCF (CCCTC-binding factor), are essential for the formation of these loops<sup>57</sup>. CTCF is also involved in loop formation in the *Igf2/H19* locus, where the loops are essential for establishing parent of origin specific gene expression (imprinting)<sup>58</sup>. CTCF binds at many sites across the genome and is for example found at borders of gene clusters<sup>59</sup> and at edges of chromatin domains<sup>60</sup>. Through these and other findings CTCF is regarded an important candidate for spatial organizer of mammalian genomes, that could establish a

chromosomal structure that is strongly related to gene expression. Interestingly, CTCF was shown to have a binding profile that overlaps with cohesin<sup>61</sup>, a ring like structure that is known to tie sister chromatids after S-phase. This CTCF-cohesin interaction could be a link between metaphase and interphase chromosomal organization.



**Figure 1.2** Schematic representation of different methods to analyze the folded structure of DNA. (see text).

In lampbrush chromosomes the loop structures are actively transcribed. There are many examples in literature that describe active regions having a looped conformation. Some large gene dense loci have been shown to be, on average, further away from their chromosome territory when they are active compared to their inactive state<sup>62</sup>. This repositioning has been termed “looping out of the chromosome territory”. However, moving away from the core of the chromosome territory does not necessarily mean that an actual loop is formed and these type of descriptions are therefore very different from for example 3C experiments. In general, genomic regions decondense when they are active or preparing for transcription<sup>62,63</sup>.

Visualization of chromosomal structure has also been approached by labeling long stretches of chromosomes using alternating fluorescent colors (**Fig. 1.2c**). FISH (fluorescent in situ hybridization) studies in *Arabidopsis* showed that their chromosomes are organized in rosette like structures, with a core of heterochromatin (the chromocentre) and the euchromatin surrounding in sometimes visible looped structures<sup>64,65</sup>. This rosette model still shows resemblance to the lampbrush model of a condensed core with protruding, active loops. In the *Arabidopsis* genome only a few large heterochromatin blocks are present per chromosome. If rosette structures would exist in human and mouse genomes, they would be much smaller than in *Arabidopsis*, because these mammalian genomes have shorter blocks of hetero- and euchromatin<sup>65</sup>. The positioning of hetero- and euchromatic blocks was studied in mammalian cells for a 4 Mb region. The two types of chromatin were labeled in alternative colors. This analysis showed that the heterochromatic blocks indeed fold towards each other, but not only rosette like structures were found. Zig-zag structures were also frequently observed<sup>66</sup>.

3C-on chip (4C) technology (chapter 4,5), is a genome wide method derived from 3C<sup>37</sup>. 4C identifies all the DNA fragments in the genome that are frequently in spatial vicinity to a selected locus (**Fig. 1.2d**). 4C was applied on two different active gene loci and these gene loci were found to frequently contact other active parts across the chromosome. An inactive chromosomal region was found to contact silent parts on the same chromosome<sup>37</sup>. The finding that contacts are made across tens of megabases, and the large number of regions that contact each locus, do not agree with highly ordered small rosette structures. The data rather suggest that sub-regions of chromosomes (e.g. gene clusters and gene deserts) behave as separate units that each fold towards preferred co-localization partners, with a strong tendency for units with similar characteristics to find each other in the nucleus, especially within a chromosome territory. The high number of interacting regions found makes it unlikely that all the contacts are made in one nucleus

and suggest that in a population of cells the units of the chromosomes form many different combinations of contacts, within the defined set. By performing 4C on multiple sites on the same chromosome (or making a chromosomal walk) and in depth modeling of the data, new insights in chromosomal folding may be obtained. The structures or forces that induce this organization also need to be defined.

### *Biophysical forces in the nucleus*

A model opposing the scaffold model has also been postulated, stating that nuclear organization is mainly the result of a combination of biophysical forces<sup>67</sup>. The combination of slow moving chromatin and freely diffusion proteins results in visco-elastic phase separation. This effect is also seen in polymer chemistry, where it is shown that asymmetric kinetic properties lead to phase separation, ending in a network of slow moving polymer and a pool of vast moving solubles. The phase separation becomes more pronounced when the slow moving polymers self-associate, which may be the case for chromatin, especially heterochromatin. The self association of the chromatin can be attenuated by the effect of macromolecular crowding, as was described above for the nuclear bodies.

The shape of the chromatin can also be influenced by other factors that act on a local scale. Not the whole genome consists of the B-DNA described by Watson and Crick. Mainly in repetitive areas different structures can occur, such as cruciforms or quadruplexes<sup>68</sup>. The helical structure can also be affected by the process of transcription; the unwinding of the double helix leads to supercoiling 3' and 5' of the transcribed area<sup>69</sup>. Effects of supercoiling have mainly been studied in bacteria and much less in eukaryotes.

### **Changes in the genome sequence in health and disease**

The sequences of the genomes of any two humans are believed to be 99.9 % identical<sup>70</sup>. The small differences in sequence can cause alterations in phenotype or susceptibility to disease. Several types of sequence variation can be found in the genome. Natural genetic variations of a single base are present in 1 in 300 basepairs in the human genome and are called single nucleotide polymorphisms (SNPs)<sup>71</sup>. ~3% of the human genome consists of microsatellites; sequences containing a variable number of 1-6 bp repeats<sup>72</sup>. Minisatellite repeats are slightly larger, 6-100 bp. The human genome contains an estimated 150,000 minisatellites of which about 20% have been shown to be variable in length within the population<sup>73</sup>. Another important source of variation is the copy number variants (CNVs), stretches of sequence larger than 1kb, present in a varying number of copies<sup>74</sup>.

Intermediate sized structural variants (mainly deletions and small inversions 8-50 kb in length) have only recently been shown to contribute to genomic variation<sup>75</sup>.

In addition to all these sequence changes of relatively small size large parts of chromosomes, several megabases or more, can change position in the genome, for example, large deletions, translocations and inversion can occur and are referred to as genomic rearrangements. The frequency with which this type of structural variation occurs naturally within the human population has not been investigated thoroughly. Genomic rearrangements can affect gene expression by changing the number of copies of a gene, by creating a fusion gene consisting of parts of two genes located on either side of a breakpoint, or by repositioning a regulatory element. When an enhancer of an active gene is repositioned next to a gene that is normally off, this gene can become aberrantly expressed. Many genetic diseases are caused by one of these mechanisms; therefore the characterization of the breakpoints of large structural rearrangements can lead to the discovery of disease associated genes. Characterization of different translocations has recently led to the identification of genes involved schizophrenia<sup>76</sup> and dyslexia<sup>77</sup>, syndromes for which the genetic components were previously unknown. Genomic rearrangements have also been identified frequently in cancer samples, mainly in hematological tumors. Only a few genomic rearrangements have been described in solid tumors<sup>78</sup>, because their detection is more difficult in these samples for practical reasons. Metaphase spreads used to visualize rearrangements are more difficult to obtain from solid tumors. In addition, solid tumors are often heterogeneous and have a more complex karyotype. 4C technology described in chapter 6 does not depend on the isolation of metaphase chromosomes and can detect translocations even when present in only a small subpopulation of the cells. 4C will likely contribute to a more thorough description of genomic rearrangements in different types of tumors and in genetic disease.



## References

- <sup>1</sup> O. T. Avery, C. M. Macleod, and M. McCarty, *J. Exp. Med.* **79**, 137 (1944).
- <sup>2</sup> J. D. Watson and F. H. Crick, *Nature* **171** (4356), 737 (1953).
- <sup>3</sup> K. E. Handwerker, J. A. Cordero, and J. G. Gall, *Molecular biology of the cell* **16** (1), 202 (2005).
- <sup>4</sup> F. M. Boisvert, S. van Koningsbruggen, J. Navascues et al., *Nature reviews* **8** (7), 574 (2007); M. Carmo-Fonseca, L. Mendes-Soares, and I. Campos, *Nature cell biology* **2** (6), E107 (2000).
- <sup>5</sup> R. Berezney, D. D. Dubey, and J. A. Huberman, *Chromosoma* **108** (8), 471 (2000).
- <sup>6</sup> G. Dellaire and D. P. Bazett-Jones, *Cell cycle (Georgetown, Tex)* **6** (15), 1864 (2007).
- <sup>7</sup> F. J. Iborra, A. Pombo, D. A. Jackson et al., *Journal of cell science* **109** ( Pt 6), 1427 (1996).
- <sup>8</sup> D. A. Jackson, F. J. Iborra, E. M. Manders et al., *Molecular biology of the cell* **9** (6), 1523 (1998).
- <sup>9</sup> M. Dundr and T. Misteli, *The Biochemical journal* **356** (Pt 2), 297 (2001).
- <sup>10</sup> R. Bernardi and P. P. Pandolfi, *Nature reviews* **8** (12), 1006 (2007).
- <sup>11</sup> T. Misteli, *The Journal of cell biology* **155** (2), 181 (2001).
- <sup>12</sup> M. Dundr, T. Misteli, and M. O. Olson, *The Journal of cell biology* **150** (3), 433 (2000); M. O. Olson, M. Dundr, and A. Szebeni, *Trends in cell biology* **10** (5), 189 (2000).
- <sup>13</sup> M. Oakes, Y. Nogi, M. W. Clark et al., *Molecular and cellular biology* **13** (4), 2441 (1993).
- <sup>14</sup> R. D. Phair and T. Misteli, *Nature* **404** (6778), 604 (2000).
- <sup>15</sup> R. J. Ellis, *Trends in biochemical sciences* **26** (10), 597 (2001).
- <sup>16</sup> R. Hancock, *Biology of the cell / under the auspices of the European Cell Biology Organization* **96** (8), 595 (2004).
- <sup>17</sup> S. B. Zimmerman, *Biochimica et biophysica acta* **1216** (2), 175 (1993).
- <sup>18</sup> R. Hancock, *Journal of structural biology* **146** (3), 281 (2004).
- <sup>19</sup> K. Luger, A. W. Mader, R. K. Richmond et al., *Nature* **389** (6648), 251 (1997).
- <sup>20</sup> A. Minsky, R. Ghirlando, and Z. Reich, *Journal of theoretical biology* **188** (3), 379 (1997).
- <sup>21</sup> A. Groth, W. Rocha, A. Verreault et al., *Cell* **128** (4), 721 (2007).
- <sup>22</sup> S. Beck and V. K. Rakyán, *Trends Genet* **24** (5), 231 (2008); H. Nagase and S. Ghosh, *The FEBS journal* **275** (8), 1617 (2008).
- <sup>23</sup> V. K. Gangaraju and B. Bartholomew, *Mutation research* **618** (1-2), 3 (2007).

- <sup>24</sup> L. Mohrmann and C. P. Verrijzer, *Biochimica et biophysica acta* **1681** (2-3), 59 (2005).
- <sup>25</sup> B. E. Black and E. A. Bassett, *Current opinion in cell biology* **20** (1), 91 (2008).
- <sup>26</sup> Y. Mito, J. G. Henikoff, and S. Henikoff, *Nature genetics* **37** (10), 1090 (2005).
- <sup>27</sup> T. Kouzarides, *Cell* **128** (4), 693 (2007); O. J. Rando, *Current opinion in genetics & development* **17** (2), 94 (2007); O. J. Rando, *Trends Genet* **23** (2), 67 (2007).
- <sup>28</sup> E. Martinez, *Plant molecular biology* **50** (6), 925 (2002).
- <sup>29</sup> E. Heard, *Current opinion in cell biology* **16** (3), 247 (2004).
- <sup>30</sup> K. Delaval and R. Feil, *Current opinion in genetics & development* **14** (2), 188 (2004).
- <sup>31</sup> F. Grosveld, G. B. van Assendelft, D. R. Greaves et al., *Cell* **51** (6), 975 (1987).
- <sup>32</sup> R. K. Maeda and F. Karch, *Current opinion in genetics & development* **17** (5), 394 (2007).
- <sup>33</sup> T. Ragozcy, A. Telling, T. Sawado et al., *Chromosome Res* **11** (5), 513 (2003); R. R. Williams, V. Azuara, P. Perry et al., *Journal of cell science* **119** (Pt 1), 132 (2006).
- <sup>34</sup> M. Simonis and W. de Laat, *Biochimica et biophysica acta* (2008).
- <sup>35</sup> P. Lichter, T. Cremer, J. Borden et al., *Human genetics* **80** (3), 224 (1988).
- <sup>36</sup> M. R. Branco and A. Pombo, *PLoS biology* **4** (5), e138 (2006).
- <sup>37</sup> M. Simonis, P. Klous, E. Splinter et al., *Nature genetics* **38** (11), 1348 (2006).
- <sup>38</sup> H. Dehghani, G. Dellaire, and D. P. Bazett-Jones, *Micron* **36** (2), 95 (2005).
- <sup>39</sup> H. Ris and D. F. Kubiak, *Annual review of genetics* **4**, 263 (1970); S. L. Wolfe and P. G. Martin, *Experimental cell research* **50** (1), 140 (1968).
- <sup>40</sup> F. Thoma, T. Koller, and A. Klug, *The Journal of cell biology* **83** (2 Pt 1), 403 (1979).
- <sup>41</sup> D. J. Tremethick, *Cell* **128** (4), 651 (2007); C. Wu, A. Bassett, and A. Travers, *EMBO reports* **8** (12), 1129 (2007).
- <sup>42</sup> A. S. Belmont and K. Bruce, *The Journal of cell biology* **127** (2), 287 (1994).
- <sup>43</sup> A. E. Mirsky and H. Ris, *The Journal of general physiology* **34** (5), 475 (1951).
- <sup>44</sup> J. R. Paulson and U. K. Laemmli, *Cell* **12** (3), 817 (1977).
- <sup>45</sup> J. Nickerson, *Journal of cell science* **114** (Pt 3), 463 (2001).
- <sup>46</sup> T. Pederson, *Journal of molecular biology* **277** (2), 147 (1998).
- <sup>47</sup> R. Berezney and D. S. Coffey, *Biochemical and biophysical research communications* **60** (4), 1410 (1974).
- <sup>48</sup> L. M. Neri, B. M. Riederer, A. Valmori et al., *J Histochem Cytochem* **45** (10), 1317 (1997).
- <sup>49</sup> T. Dechat, K. Pflieger, K. Sengupta et al., *Genes & development* **22** (7), 832 (2008).

- <sup>50</sup> R. K. Sachs, G. van den Engh, B. Trask et al., *Proceedings of the National Academy of Sciences of the United States of America* **92** (7), 2710 (1995).
- <sup>51</sup> C. Munkel, R. Eils, S. Dietzel et al., *Journal of molecular biology* **285** (3), 1053 (1999).
- <sup>52</sup> S. Jhunjhunwala, M. C. van Zelm, M. M. Peak et al., *Cell* **133** (2), 265 (2008).
- <sup>53</sup> J. Dekker, K. Rippe, M. Dekker et al., *Science (New York, N.Y)* **295** (5558), 1306 (2002).
- <sup>54</sup> B. Tolhuis, R. J. Palstra, E. Splinter et al., *Molecular cell* **10** (6), 1453 (2002).
- <sup>55</sup> S. Galande, P. K. Purbey, D. Notani et al., *Current opinion in genetics & development* **17** (5), 408 (2007).
- <sup>56</sup> R. Drissen, R. J. Palstra, N. Gillemans et al., *Genes & development* **18** (20), 2485 (2004).
- <sup>57</sup> E. Splinter, H. Heath, J. Kooren et al., *Genes & development* **20** (17), 2349 (2006).
- <sup>58</sup> S. Kurukuti, V. K. Tiwari, G. Tavoosidana et al., *Proceedings of the National Academy of Sciences of the United States of America* **103** (28), 10684 (2006).
- <sup>59</sup> T. H. Kim, Z. K. Abdullaev, A. D. Smith et al., *Cell* **128** (6), 1231 (2007).
- <sup>60</sup> L. Guelen, L. Pagie, E. Brassat et al., *Nature* **453** (7197), 948 (2008).
- <sup>61</sup> V. Parelho, S. Hadjur, M. Spivakov et al., *Cell* **132** (3), 422 (2008); K. S. Wendt, K. Yoshida, T. Itoh et al., *Nature* **451** (7180), 796 (2008).
- <sup>62</sup> E. V. Volpi, E. Chevret, T. Jones et al., *Journal of cell science* **113** (Pt 9), 1565 (2000).
- <sup>63</sup> W. G. Muller, D. Rieder, G. Kreth et al., *Molecular and cellular biology* **24** (21), 9359 (2004); T. Tumber, G. Sudlow, and A. S. Belmont, *The Journal of cell biology* **145** (7), 1341 (1999).
- <sup>64</sup> P. Fransz, J. H. De Jong, M. Lysak et al., *Proceedings of the National Academy of Sciences of the United States of America* **99** (22), 14584 (2002).
- <sup>65</sup> R. van Driel and P. Fransz, *Experimental cell research* **296** (1), 86 (2004).
- <sup>66</sup> L. S. Shopland, C. R. Lynch, K. A. Peterson et al., *The Journal of cell biology* **174** (1), 27 (2006).
- <sup>67</sup> F. J. Iborra, *Theoretical biology & medical modelling* **4**, 15 (2007).
- <sup>68</sup> E. B. Jagelska, V. Brazda, P. Pecinka et al., *The Biochemical journal* **412** (1), 57 (2008); A. Verma, K. Halder, R. Halder et al., *Journal of medicinal chemistry* (2008).
- <sup>69</sup> C. J. Dorman, *Science progress* **89** (Pt 3-4), 151 (2006); A. Travers and G. Muskhelishvili, *Nat Rev Microbiol* **3** (2), 157 (2005).
- <sup>70</sup> M. Przeworski, R. R. Hudson, and A. Di Rienzo, *Trends Genet* **16** (7), 296 (2000); D. E. Reich, S. F. Schaffner, M. J. Daly et al., *Nature genetics* **32** (1), 135 (2002).

Chapter 1

- <sup>71</sup> L. Kruglyak and D. A. Nickerson, *Nature genetics* **27** (3), 234 (2001).
- <sup>72</sup> E. S. Lander, L. M. Linton, B. Birren et al., *Nature* **409** (6822), 860 (2001).
- <sup>73</sup> K. Naslund, P. Saetre, J. von Salome et al., *Genomics* **85** (1), 24 (2005).
- <sup>74</sup> J. Sebat, B. Lakshmi, J. Troge et al., *Science (New York, N.Y)* **305** (5683), 525 (2004).
- <sup>75</sup> J. O. Korbelt, A. E. Urban, J. P. Affourtit et al., *Science (New York, N.Y)* **318** (5849), 420 (2007); E. Tuzun, A. J. Sharp, J. A. Bailey et al., *Nature genetics* **37** (7), 727 (2005).
- <sup>76</sup> J. K. Millar, S. Christie, C. A. Semple et al., *Genomics* **67** (1), 69 (2000).
- <sup>77</sup> M. Taipale, N. Kaminen, J. Nopola-Hemmi et al., *Proceedings of the National Academy of Sciences of the United States of America* **100** (20), 11553 (2003).
- <sup>78</sup> G. Attard, J. Clark, L. Ambrosine et al., *British journal of cancer* **99** (2), 314 (2008).

# 2

## Genome-wide technologies

## Genome-wide technologies

At the end of the 20<sup>th</sup> century the way genomes are studied started to change. The establishment of fully sequenced genomes opened doors for genome-wide studies. Microarray technology has been especially influential and more recently large scale sequencing has boosted the development of novel genomic methods.

This chapter introduces microarray and sequencing technologies and provides a brief overview of the genomic methods in which they have been applied.

### **Microarray technology**

Traditionally DNA sequences and RNA species are measured using a labeled probe complementary to the sequence of interest. This concept was first modified to be used on a large scale in 1995, when the Brown lab created the first microarray<sup>1</sup>. In microarray technology probes are attached or spotted to a glass slide in an orderly fashion, such that each "spot" on the array contains a large excess of a specific probe. A fluorescently labeled RNA or DNA sample is hybridized to the microarray and, after stringent washing, the fluorescence hybridized to each spot is measured using a sensitive scanner. The fluorescence is a measure for the abundance of a specific sequence in the sample. On the first microarray the expression levels of 45 genes were measured simultaneously, using PCR products as probes. The technology was initially received skeptically, but has made some giant leaps forward. Instead of 45 genes, all known human genes are now being analyzed on a single array. Moreover, the PCR products have been replaced by more controllably synthesized oligo-nucleotide probes. Through these and other advances in the technique and standardization of data-analysis, microarray expression analysis has become a widely used tool in research and is starting to be applied in diagnostics<sup>2-5</sup>.

Initially microarrays were mainly used to measure gene expression levels, but essentially any mixture of nucleic acids can be analyzed, provided that informative microarray probes are designed. If an entire genome is sequenced probes can be designed that cover whole or specific parts of genomes. On these tilling arrays BACs (Bacterial Artificial Chromosomes) or oligos are used as probes. Their size and their spacing on the genome determine the resolution of the microarray. Tilling arrays can be used to study a large variety of genome characteristics, such as the binding sites of a specific protein.

Microarray technology is still advancing, for example, the density of the spots is increasing; several million different probes can now be placed on one array. However, due to recent

rapid developments the use of large scale sequencing is becoming increasingly popular in genomic analyses. In the future, sequencing may replace microarray technology in at least some applications.

### Microarrays versus sequencing analyses

There are several reasons why sequencing analysis is becoming increasingly popular over the use of microarrays (**table 2.1**). Most importantly, sequencing is in some regards less biased and generates more detailed data. In microarray technology only sequences complementary to the probes that are used are measured. With sequencing technology every DNA or RNA fragment present in a sample is analyzed individually. Thus, provided that enough sequences are analyzed, nothing is missed. In microarray analysis quantification is achieved by measuring the hybridization to a probe sequence. When a sample is sequenced the exact nucleotide compositions are retrieved. This could not only make the analysis more quantitative, but importantly also allows a much more in depth analysis of nucleic acid samples and makes it possible to detect (unexpected) changes in the analyzed sequences.

However, microarray technology still has the advantage of the head start (**table 2.1**). It is in a more advanced stage of development, and not unimportantly, the data analysis of many microarray applications is standardized, such that relevant information is more easily obtained from the datasets. Sequencing data sets are much larger than array datasets and data processing requires more computational power. The fact that the use of probes on microarrays predefines what is measured makes their data analysis more comprehensive. To position sequencing data on the genome BLAST (Basic Local Alignment Search Tool) analyses need to be performed. Lastly, microarrays are also still the least expensive method for many applications.

**Table 2.1. Current differences between microarray and sequencing analyses**

	Microarray	Sequencing
sequences measured	only complementary to probes	all
measures	hybridization to probe	basepair sequence
stage of development	advanced	novel
standardized analysis	for some applications	no
costs	less expensive	more expensive

### Sequencing technology

Because sequencing analysis gives detailed information about changes in the genome it has a lot of potential for diagnostics, but conventional sequence analysis is too time consuming and too costly for clinical applications. Therefore the aim is to enhance the

technology such that sequencing of a complete genome can be done in a single fast and affordable experiment. This goal has been given the working title “the \$1000 genome”. Much technical progress was made to achieve this target.

Sequencing analysis traditionally includes two important steps. The analyzed mixture is first split, such that each DNA fragment in the sample is measured individually. Second, the isolated DNA needs to be amplified before it is sequenced, because sequencing of a single molecule is not (yet) possible. In conventional sequencing analysis DNA samples were split by cloning the target sequences into vectors and transforming bacteria. This way each bacterial colony contained one DNA sequence that was amplified by growing the bacteria. The entire human genome was sequenced using this method, which took many years and cost around a billion dollars.

Recently, new technologies were developed that yield of a large number of sequence reads (500,000-60 million) per experiment without bacterial transformation. These methods are collectively referred to as next generation sequencing or massive parallel sequencing. With this generation of sequencing methods a whole mammalian genome can be sequenced in a time span of weeks and at price that is 200 fold lower than that of the human sequencing project. Different next generation methodologies were developed in parallel, but they all include the same principle steps. The DNA samples are fragmented, adapters are ligated to the DNA fragments and the product is attached to a surface, thereby isolating each fragment. Each individual DNA fragment is then amplified and sequenced<sup>6</sup>. Each method has its own advantages and disadvantages. Differences are mainly found in read length and the number of sequence reads that is obtained (reviewed in [7]).

The illumina/solexa system uses methodology that is in many ways similar to microarray technology. The mixture of sequences containing adapters is adhered to a glass slide. Each fragment is amplified in an on-glass-slide PCR reaction, such that each individual sequence will create its own “spot”. Each individual spot is sequenced by measuring the incorporation of fluorescent dyes, basepair by basepair. The detection of the fluorophores at the spots occurs in a manner similar to the scanning of a microarray ([www.illumina.com](http://www.illumina.com))<sup>7</sup>.

In the Roche (454) GS FLX sequence system individual sequences are attached to primer coated beads and amplified by emulsion PCR. In this type of PCR each bead is contained in a small oil droplet that also holds the necessary PCR reaction components. After amplifications each droplet will contain up to one million copies of a unique DNA



fragment. The PCR products are subsequently positioned on a picotiter plate and analyzed by pyrosequencing ([www.454.com](http://www.454.com))<sup>8, 7</sup>.

In the Applied Biosystems SOLiD sequencing system DNA fragments are captured on beads that are attached to glass slides. Intricate ligation mediated sequencing is applied to determine the base constituents (for details see

[http://marketing.appliedbiosystems.com/images/Product/Solid\\_Knowledge/flash/102207/solid.html](http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/flash/102207/solid.html))<sup>7</sup>.

Although next generation sequencing was only recently introduced, more advanced sequencing technologies have already emerged. Helicos BioSciences' launched the HeliScope ([www.helicosbio.com](http://www.helicosbio.com)), which allows the detection of single fluorescently labelled nucleotides. The Helicos system is very similar to the illumina\solexa technology, with the advantage that the sequences no longer need to be amplified before they can be sequenced.

Sequencing methods generally rely on incorporation of a labeled nucleotide followed by abortion of elongation and identification of the incorporated base. The repeated abortion steps slow down the sequencing process and increase the costs of the reactants. Pacific Biosciences has developed an innovative sequencing system that measures nucleotide incorporation real time ([www.pacificbiosciences.com](http://www.pacificbiosciences.com)). In this single molecule real time DNA sequencing technology (SMRT™) the polymerases instead of the DNA sequences are attached to a surface. Each polymerase is confined within an ultra small measurement volume, which makes it possible to detect the incorporated nucleotide above the background of all the labeled nucleotides in the solution.

The novel sequencing technologies have only recently become available and their capabilities and incapacities have yet to be defined in detail. However it is clear that with all the technological advancements the \$1000 genome is in sight. This does not only have applications in the clinic, but will also provide new opportunities for research. Possibly the biggest challenge lying ahead is to come towards a comprehensive and informative analysis of the rapidly expanding and accumulating datasets.

### **Genome-wide expression analysis**

The genome wide application that has been used most extensively is the analysis of gene expression. Searching for 'microarray expression analysis' in the PubMed database results in almost 20,000 publications ([www.ncbi.nlm.nih.gov/sites/entrez](http://www.ncbi.nlm.nih.gov/sites/entrez)). The large datasets obtained in gene expression studies have been used in very different ways.

It is possible to focus on individual genes in the data and determine which of them are alternatively expressed between samples. This can reveal tissue specific genes or changes in activity upon (experimental) manipulation of the system<sup>9,10</sup>.

The size of the datasets provides the opportunity to not only focus on the few genes with aberrant expression levels, but rather to investigate the behavior of groups of genes. Genes can be clustered according to their expression across experiments. An example of this is a cell-cycle experiment in which expression patterns in consecutive phases of the cell cycle are analyzed. Genes can be clustered according to their transcriptional activity throughout the cell cycle and classified as such<sup>11</sup>.

In addition to clustering the genes, different samples within an experiment can be compared and clustered based on their expression profiles. This has been used extensively in cancer research. Patients with the same type of cancer can have very different expression profiles (in the tumor). Studying the gene expression of different clusters of patients can give insight in the different pathways involved in the pathogenesis<sup>12</sup>. Ultimately the different profiles will be linked with clinical outcome and gene expression analysis can be used as a diagnostic tool.

Diagnostic purposes of gene expression analysis can also be achieved by approaching the data in a different order; by first grouping the patients according to clinical variables and then investigating the gene expression data, searching for genes that together allow discrimination between the different patient groups. The clinical variables of other patients can subsequently be predicted based on the expression of the discriminating set of genes, rather than total gene expression profiles. This has been used most successfully for breast cancer, where expression analysis has become an important method in tumor classification<sup>13</sup>.

Sequencing technology will allow an even more elaborate analysis of gene expression profiles. If the exact base pair sequence of expressed genes is measured, splice variants, small mutations and novel fusion genes created by structural genomic rearrangements can be detected.

### **Genome-wide analysis of chromatin**

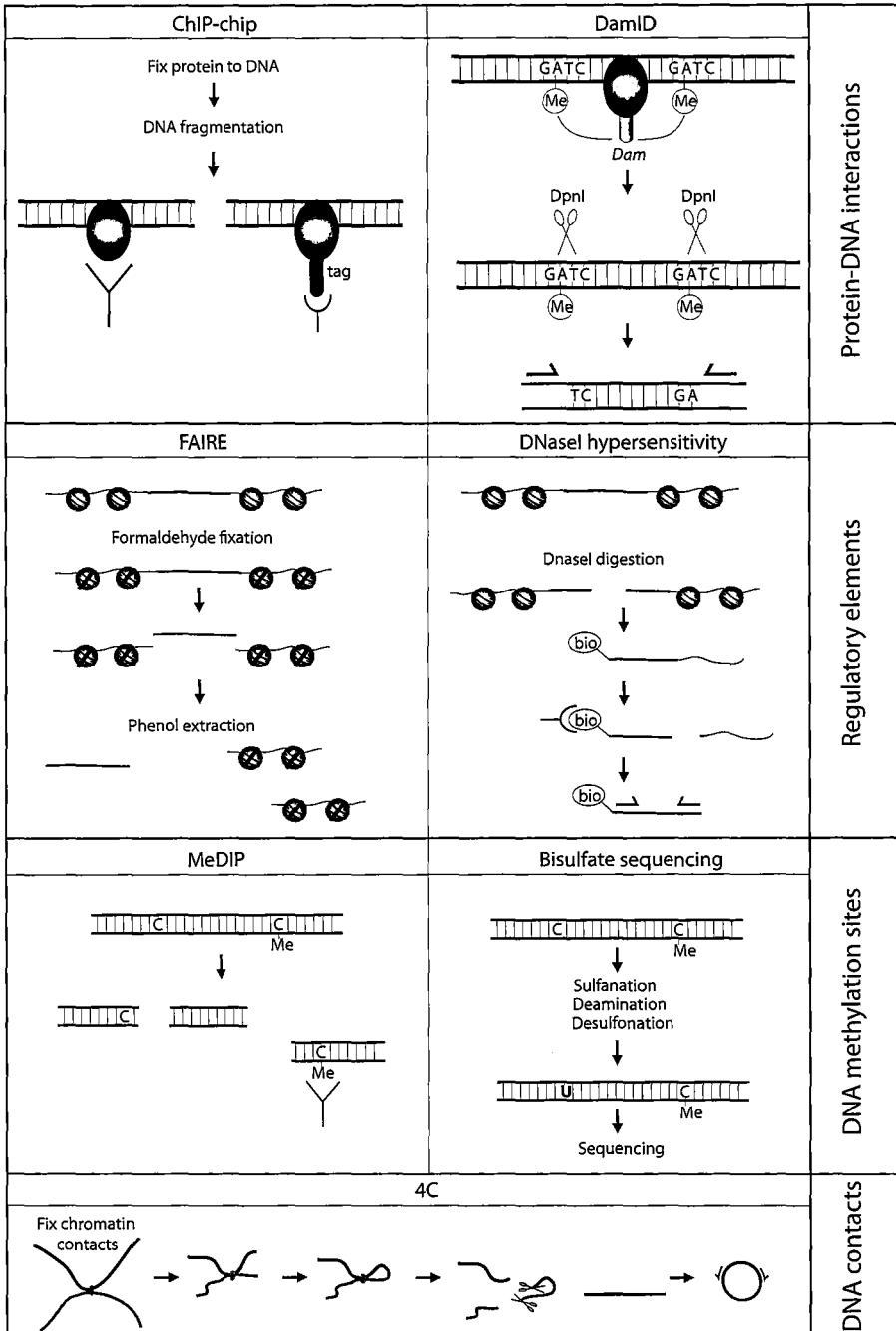
In addition to levels of gene expression many aspects of genome structure have been analyzed on a genome-wide scale. The technical challenge in genome wide methods is to extract the sequences of interest from a complex sample containing the entire genome (**Fig. 2.1**). The captured sequences are analyzed by microarray or sequencing analysis.

Protein-DNA interactions can be captured in two principle ways. Firstly, sequences bound by a specific protein can be obtained by selectively retrieving the protein of interest and the binding sites from the nucleus. To achieve this protein-DNA contacts are stabilized by formaldehyde cross-links, the DNA is fractionated and the protein of interest and bound sequences are pulled down using a specific antibody. This approach is called chromatin immune precipitation on chip (ChIP-chip)<sup>14</sup>. DamID is based on a different principle; the protein of interest is not extracted from the nucleus but is modified such that it leaves chemical marks at the positions of the genome it binds to. The studied protein is fused with a Dam enzyme (DNA adenine methyltransferase) that methylates adenine residues of the DNA<sup>15</sup>. To retrieve the affected sequences the genome is digested with a restriction enzyme that only cleaves the methylated sites. By ligation of adapters and subsequent PCR the sites contacted by the protein of interest are selectively amplified. Protein-DNA interaction measuring techniques can be applied to study diverse aspects of chromatin structure, for example the binding sites of transcription factors, the positions of histone variants and the locations of histone modifications across the genome<sup>14,16,17</sup>.

Two methods were employed to find regulatory elements in the genome (e.g. promoters, enhancers and silencers), both were based on the idea that sequences bound by regulatory proteins generally lack nucleosomes. In formaldehyde assisted isolation of regulatory elements (FAIRE), nucleosomes are fixed to the DNA using formaldehyde and the genome is fragmented<sup>18</sup>. In a subsequent phenol extraction regulatory elements end up in the upper aqueous phase and can be isolated. In a different approach, regulatory elements have been characterized making use of their increased sensitivity to DNase I<sup>19,20</sup>, which is likely also the result of a lack of histones. In this strategy the genome is treated with (a low amount of) DNase I and digested sites are obtained by: ligation of biotinylated adapters, trimming of the DNA fragments, purification of biotinylated DNA, addition of a second adapter and PCR amplification.

Methylation of cytosine residues in the genome can be measured either by methylated DNA immunoprecipitation (MeDIP), using an antibody against methylated sites, or by bisulfite sequencing. In the latter several chemical steps convert unmethylated cytosine to uracil, whereas the methylated cytosines are unaffected. The changes in sequence can be measured, for example by high throughput sequencing<sup>21</sup>.

The folded structure of the genome can be analyzed with 4C (chromosome conformation capture on chip), as described in chapter 4<sup>22</sup>. In 4C, co-localizing DNA is cross-linked with formaldehyde, digested and cross-linked fragments are ligated. Thus, in a pool of analyzed cells each DNA fragment is ligated to a set of sequences it co-localized with. For a selected



**Figure 2.1** Schematic representations of genome-wide assays to measure different elements of chromatin structure. (see text)

fragment this entire set can be analyzed by trimming and circularizing the ligation products, after which an inverse PCR with primers on the selected fragment amplifies all the co-localization partners. Because the amplified parts are always located at the restriction site of the enzyme that is used, custom 4C microarray probes can be designed and the entire genome can be spotted on a single microarray (a whole mammalian genome tiling array covers 36 microarrays). The technical details of this technology are discussed in chapter 5 of this thesis.

Together the technologies described above cover almost every aspect of genome structure. All methods are informative individually and combined analyses will give insight in the mutual relation between different chromatin components. Genome wide studies have shown some general patterns in the distribution of histone modifications<sup>16</sup>, but have also revealed some surprises. For example, in ES cells some gene promoters contain both methylated H3K4, considered a mark of active chromatin, and methylated H3K27, a histone modification associated with silenced genes<sup>23</sup>. During differentiation this bivalency is lost. DamID studies have shown that heterochromatin associated proteins such as HP1 can also be found on active genes<sup>24</sup>.

In addition to these genome-wide views, functional analysis of selected loci will be necessary to understand the exact mechanisms and function of all the different elements of chromatin structure.

### **Genomic techniques to study the sequence of genomes**

Genomic technologies have also greatly influenced genetic research. Before “the genomic revolution” the detection of structural alterations in the genome relied on cytogenetic techniques<sup>25</sup>; visualizing (parts) of chromosomes and searching for changes. In traditional karyotyping metaphase chromosomes are visualized and a trained eye can detect changes in chromosome number or banding patterns. Later developments that improved cytogenetics are spectral karyotyping, in which each chromosome is colored with a unique fluorophore, and fluorescent *in situ* hybridization, which can be used to visualize selected parts of chromosomes.

The detection of deletions and amplifications has been greatly facilitated by Comparative Genome Hybridisation (CGH). In this method patient and control DNA are alternatively labeled and the ratio of the intensities of the two dyes gives a measure for quantitative differences between the two samples. Initially, the labeled DNA was hybridized to intact metaphase chromosomes, but currently microarrays are used (array CGH). Array CGH is

a commonly used tool in genetic studies and has led to the discovery of many disease associated deletions and amplifications<sup>26</sup>.

Only recently, genomic methods were developed that can detect balanced structural rearrangements. In arraypainting, chromosomes are sorted based on size using flow cytometry. The individual chromosomes can then be hybridized to a microarray or sequenced<sup>27, 28</sup>. This will reveal if there was any sequence exchange between chromosomes. However, metaphase chromosomes are needed for this technique, which can not always be obtained. Moreover, not each chromosome can be isolated based on its size and inversions are missed using this method.

Paired end sequencing (PES)<sup>29, 30</sup> is a next generation sequencing based method to detect structural alterations in the genome. In this technique the genome is fragmented in such a way that all the fragments have a comparable size. The ends of the fragments are sequenced and paired. Both sequences are blasted against a reference genome. If the distance between the paired sequences is not the same as the size of the fragment, the analyzed genome is different from the reference genome. Many different types of structural changes can be identified, including small deletions and inversions. The size of the changes that can be seen depends on the sizes of the fragments that are analyzed.

Chapter 6 of this thesis describes the use of 4C to characterize genomic rearrangements within several megabases of a locus of interest (the viewpoint). 4C utilizes the compacted structure of the genome in the nucleus to capture the genomic neighbors of the viewpoint. These neighboring fragments are all PCR amplified in one PCR reaction and identified on a microarray. This way, 4C identifies sequences up to several megabases away. In case of a rearrangement the sequences originate from unexpected positions of the genome. Because many fragments across breakpoints are captured 4C is a very robust method. In Chapter 7 4C technology is further extended by replacing the microarray technology with sequencing technology. In the future, the sequencing strategy could be employed to develop a genome-wide scan for large structural rearrangements.

Very small genetic alterations can also be investigated by genomic approaches. For example, Single nucleotide polymorphisms (SNPs) can be detected using SNP arrays<sup>31</sup> (array technology has advanced so much that effects of single nucleotide changes on the hybridization signals can be measured) or by sequencing of genomic regions of interest or expressed genes<sup>32</sup>.

## References

1. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y)* **270**, 467-470 (1995).
2. Auburn, R.P. *et al.* Robotic spotting of cDNA and oligonucleotide microarrays. *Trends in biotechnology* **23**, 374-379 (2005).
3. Grewal, A., Lambert, P. & Stockton, J. Analysis of expression data: an overview. *Current protocols in bioinformatics / editorial board, Andreas D. Baxeavanis ... [et al]* **Chapter 7**, Unit 7 1 (2007).
4. Page, G.P. *et al.* Microarray analysis. *Methods in molecular biology (Clifton, NJ)* **404**, 409-430 (2007).
5. Steinhoff, C. & Vingron, M. Normalization and quantification of differential expression in gene expression microarrays. *Briefings in bioinformatics* **7**, 166-177 (2006).
6. Mardis, E.R. Anticipating the 1,000 dollar genome. *Genome biology* **7**, 112 (2006).
7. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133-141 (2008).
8. Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8817-8822 (2003).
9. Dehner, M., Hadjihannas, M., Weiske, J., Huber, O. & Behrens, J. Wnt signaling inhibits Forkhead box O3a-induced transcription and apoptosis through up-regulation of serum- and glucocorticoid-inducible kinase 1. *The Journal of biological chemistry* **283**, 19201-19210 (2008).
10. Jura, J. *et al.* Identification of interleukin-1 and interleukin-6-responsive genes in human monocyte-derived macrophages using microarrays. *Biochimica et biophysica acta* **1779**, 383-389 (2008).
11. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell* **9**, 3273-3297 (1998).
12. Bruland, O. *et al.* Gene expression reveals two distinct groups of anal carcinomas with clinical implications. *British journal of cancer* **98**, 1264-1273 (2008).
13. Miller, L.D. & Liu, E.T. Expression genomics in breast cancer research: microarrays at the crossroads of biology and medicine. *Breast Cancer Res* **9**, 206 (2007).

14. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499-509 (2004).
15. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nature biotechnology* **18**, 424-428 (2000).
16. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705 (2007).
17. Mito, Y., Henikoff, J.G. & Henikoff, S. Genome-scale profiling of histone H3.3 replacement patterns. *Nature genetics* **37**, 1090-1097 (2005).
18. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research* **17**, 877-885 (2007).
19. Crawford, G.E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature methods* **3**, 503-509 (2006).
20. Sabo, P.J. *et al.* Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature methods* **3**, 511-518 (2006).
21. Zilberman, D. & Henikoff, S. Genome-wide analysis of DNA methylation patterns. *Development (Cambridge, England)* **134**, 3959-3965 (2007).
22. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* **38**, 1348-1354 (2006).
23. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nature cell biology* **8**, 532-538 (2006).
24. Greil, F. *et al.* Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location. *Genes & development* **17**, 2825-2838 (2003).
25. Morozova, O. & Marra, M.A. From cytogenetics to next-generation sequencing technologies: advances in the detection of genome rearrangements in tumors. *Biochemistry and cell biology* **86**, 81-91 (2008).
26. Kallioniemi, A. CGH microarrays and cancer. *Current opinion in biotechnology* **19**, 36-40 (2008).
27. Fiegler, H. *et al.* Array painting: a method for the rapid analysis of aberrant chromosomes using DNA microarrays. *Journal of medical genetics* **40**, 664-670 (2003).
28. Chen, W. *et al.* Mapping translocation breakpoints by next-generation sequencing. *Genome research* **18**, 1143-1149 (2008).



29. Korbelt, J.O. *et al.* Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 10110-10115 (2007).
30. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature genetics* **37**, 727-732 (2005).
31. Grant, S.F. & Hakonarson, H. Microarray technology and applications in the arena of genome-wide association. *Clinical chemistry* **54**, 1116-1124 (2008).
32. Jones, S. *et al.* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science (New York, N.Y.)* (2008).



# 3

## FISH-eyed and genome-wide views on the spatial organization of gene expression

## FISH-eyed and genome-wide views on the spatial organization of gene expression.

Marieke Simonis<sup>1</sup>, Wouter de Laat<sup>1,2</sup>

<sup>1</sup>Dept of Cell Biology, Erasmus Medical Center, Dr. Molewaterplein 50, 3015 GE, Rotterdam, The Netherlands

<sup>2</sup> Corresponding author

### *Summary*

*Eukaryotic cells store their genome inside a nucleus, a dedicated organelle shielded by a double lipid membrane. Pores in these membranes allow the exchange of molecules between the nucleus and cytoplasm. Inside the mammalian cell nucleus, roughly 2 m of DNA, divided over several tens of chromosomes is packed. In addition, protein and RNA molecules functioning in DNA-metabolic processes such as transcription, replication, repair and the processing of RNA fill the nuclear space. While many of the nuclear proteins freely diffuse and display a more or less homogeneous distribution across the nuclear interior, some appear to preferentially cluster and form foci or bodies. A non-random structure is also observed for DNA: increasing evidence shows that selected parts of the genome preferentially contact each other, sometimes even at specific sites in the nucleus. Currently a lot of research is dedicated to understanding the functional significance of nuclear architecture, in particular with respect to the regulation of gene expression. Here we will evaluate evidence implying that the folding of DNA is important for transcriptional control in mammals and we will discuss novel high-throughput techniques expected to further boost our knowledge on nuclear organization.*

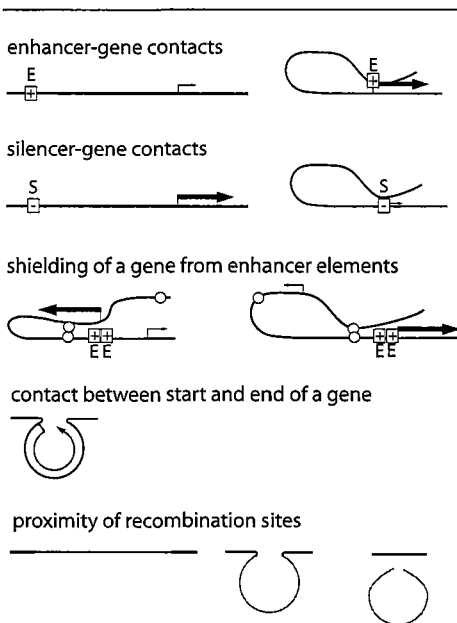
Expression of genes, in particular of tissue-specific genes, is often controlled by regulatory DNA elements like enhancers and silencers that are located away from the gene promoter. In mammals, these DNA elements may be up to 1 megabase apart from the gene [1]. A gene locus is defined as the chromosomal region that carries the gene and its regulatory DNA elements. When evaluating the functional relevance of chromosome folding we consider it important to distinguish between DNA contacts formed within and between gene loci.

### DNA interactions within gene loci

Within gene loci, it is clear that DNA loops are formed which are functionally meaningful for transcription regulation (**Fig. 3.1**). Evidence for the *in vivo* existence of such local chromatin loops was first obtained in the mouse  $\beta$ -globin locus. It relied on the development of two novel techniques, RNA-TRAP and Chromosome Conformation Capture (3C) technology (**Fig. 3.2**). Both techniques independently showed that the  $\beta$ -globin locus control region (LCR), crucial for high  $\beta$ -globin gene expression [2-4], contacts active  $\beta$ -globin genes by looping out the intervening chromatin fibre *in vivo* [5, 6]. 3C technology, in particular, has since become a widely used tool for exploring the functional relevance of DNA interactions. 3C (chromosome conformation capture) is a biochemical method that involves the capture of *in vivo* interacting DNA fragments via formaldehyde crosslinking and ligation. Quantitative PCR across ligation junctions with primers selected for specific DNA fragments subsequently gives a measure for their steady-state interaction frequency in the cell population [7].

Using 3C, it was demonstrated that LCR-gene loops are formed specifically in erythroid cells that express the  $\beta$ -globin genes [6]. It was found that during development, the LCR switches its contacts between different  $\beta$ -globin genes in relation to their switch in expression. Contacts are only established late during erythroid differentiation when the  $\beta$ -globin genes are fully expressed [8] and rely on the transcription factors EKLF and

GATA-1, which both are required for high levels of  $\beta$ -globin gene expression [9, 10]. Collectively, the data show that the LCR increases the transcription rate of the  $\beta$ -globin genes by physically contacting the genes. Similar contacts between enhancer-like DNA elements



**Figure 3.1. Overview of mechanisms in which looping of DNA is involved.** Looping of enhancers or silencers towards gene promoters can influence transcriptional activity [5, 6]. Looped structures can shield a gene from its enhancers, resulting in gene silencing [14, 15]. Contacts between a gene's start and end have been reported and are proposed to facilitate recycling of RNA PolII [17]. DNA sequences that are removed during recombination events in immune receptor loci are looped out, while the recombination sites come in close proximity [19, 23].

and genes have since been demonstrated by 3C technology in the *T helper type 2 cytokine* locus [11], the  $\alpha$ -globin locus [12, 13], the *Kit* locus [14] and many other gene loci.

Local chromatin loops not only serve to promote transcription but appear to also function in gene silencing. Silencing DNA elements have been shown to form chromatin loops in an imprinted gene cluster carrying the *Dlx5* and *Dlx6* genes [15] and to contact gene promoters in the *Kit* locus [14]. CTCF, an insulator protein which can block enhancer activity when bound in between an enhancer and promoter, also forms chromatin loops in gene loci. At the imprinted *Igf2-H19* locus, CTCF-mediated loops formed on the maternal allele shield the *Igf2* gene, causing shared enhancers to exclusively act on the maternal *H19* gene [16, 17]. In the  $\beta$ -globin locus, CTCF mediates the formation of loops between cognate binding sites flanking the locus. These loops are formed only when the locus is active, but they seem to not influence the expression of the  $\beta$ -globin genes [18]. In yeast, loops have been demonstrated between the two ends of actively transcribed genes and a similar observation was recently made for the HIV-1 provirus integrated into human cells [19, 20]. It has been suggested that physical proximity between the end and the start of a transcription unit facilitates the recycling of RNA polymerase II (RNAP II), thus stimulating transcription re-initiation. Interestingly, recent live cell imaging studies in flies confirm local recycling of fluorescently tagged RNAP II at highly transcribed heat-shock loci, however apparently without these genes forming loop structures [21]. Future research therefore should uncover whether gene looping is a general phenomenon and how it influences the transcription process.

Apart from transcription, chromatin loops have been implicated in the rearrangement process that joins the various segments of immunoglobulin loci and T cell receptor loci to assemble a functional antigen receptor gene [22-24]. These loci are very large, spanning hundreds of kilobases or even a few megabases of DNA, and fluorescence in situ hybridization (FISH) could therefore be applied to independently confirm the presence of these chromatin loops in individual cells under the confocal microscope [23, 24]. Ideally, loops detected by 3C in much smaller loci should also be validated by FISH, but this is currently impeded by the limited resolution of microscopes. Novel microscopy techniques such as 4pi [25] may enable the visualization of these smaller chromatin loops in the near future.

### **The significance of local chromatin loops**

How would chromatin loops influence processes like transcription and recombination? In the case of recombination it seems clear: to join two DNA segments they need to



physically meet. An unanswered question still is how two sites separated on the chromosome physically come together. It may involve a deterministic search process, but we would predict contacts are the consequence of the random collisions between DNA sites that occur as consequence of the flexibility of the chromatin fiber. A productive recombination event then takes place as soon as the appropriate sites juxtapose.

The relevance of chromatin loops for transcription regulation seems more difficult to envisage in molecular terms. Looping brings DNA-binding sites for transcription factors in close proximity to the promoter. It has been proposed that this causes a local accumulation of transcription factors, which will reinforce the expression status of the gene [26, 27]. If silencers are involved, the local accumulation of repressor proteins may lock the silenced state of the gene, for example by the deposition of heterochromatin marks onto histones at the promoter. When enhancers loop towards the gene, bound transcription factors may increase transcription rates by modifying the transcription machinery and/or the chromatin, thereby facilitating transcriptional re-initiation or elongation.

Chromatin looping also provides a conceptual framework to understand how genes in the mammalian genome maintain expression profiles that are independent of those of neighboring genes. Mammalian gene loci often do not occupy physically separate domains on the chromosome, but frequently overlap, with unrelated genes showing independent expression patterns located in between regulatory DNA elements and their target genes. In such instances, boundaries or insulators cannot explain how neighboring genes maintain distinct expression profiles. Here, chromatin folding may explain how enhancers specifically act on distant target genes while ignoring more proximal genes. We expect that separated genomic sites meet through random collisions but only form a stable chromatin loop when proteins bound to these sites show affinity. Thus, a promoter needs to be compatible with an enhancer to benefit from its physical proximity [26]. This model explains for example why olfactory receptor genes are not activated by the directly neighboring  $\beta$ -globin LCR in erythroid cells, even not when the insulator protein CTCF no longer binds to the intervening sequence [18].

### **Chromosome folding and gene positioning**

Microscopy studies, in particular FISH-based studies, have shown that DNA inside the cell nucleus is structured also beyond the level of single gene loci. The physically separate parts of the genome, the chromosomes, are organized in territories. When chromosomes are individually stained they appear as distinct areas, rather than being



dispersed throughout the nucleus. Chromosomes show a probabilistic three-dimensional distribution, with small chromosomes preferentially occupying more internal positions and large chromosomes being more peripheral in the nucleus [28]. However, there is intermingling between chromosome territories (CTs) and specific areas found on different chromosomes can find each other in the nuclear space [29, 30]. Well-described examples of this are the nucleoli and chromocenters. In the nucleoli, rDNA clusters present on different chromosomes come together to be transcribed [31]. Chromocenters are the large heterochromatin structures containing pericentromeric regions of different chromosomes [32]. Both nucleoli and chromocenters can differ in size and number between different tissues [33].

Within CTs the chromatin is also organized. When gene density profiles are plotted along the chromosomes, it becomes clear that genes are not distributed randomly, but rather are organized in clusters [34, 35]. These gene-dense regions tend to be localized more towards the nuclear interior than their gene-poor counterparts [36]. This is also true when such regions are physically linked, as was shown by a detailed FISH study that analyzed the folding of a 4.3 Mb chromosomal region containing four gene clusters and four gene-poor regions. It was found that the region had a few favored conformations. In most of the folded structures the gene rich regions were partially or completely clustered in space. The gene poor regions were found to cluster as well, but to a lesser extent. The investigated 4.3 Mb region is often found in the outer shell of the nucleus and in 80% of these peripheral situations, two or more gene poor regions show overlap with Lamin staining, whereas only in 20% of the cases two or more gene rich regions are found there. Lamins are core components of the nuclear lamina, a proteinaceous mesh that coats the inside of the nuclear membrane and connects it to chromatin. Together, the data show that the folding of this 4.3 Mb region is not random and suggest that the region may be “anchored” to the periphery by the gene poor regions [37].

Not only gene density but also transcriptional activity may be linked to nuclear positioning. FISH studies have demonstrated that genes can adopt different nuclear positions upon changes in their expression status. A repositioning of developmentally activated genes away from the nuclear periphery or from pericentromeric heterochromatin has been documented for multiple loci, including immunoglobulin loci [38], *CFTR* [39], the  $\beta$ -globin genes [40] and *Mash-1* [41]. Vice versa, genes have been described that move towards the periphery or towards the chromocenters upon their developmental silencing [42-45]. With respect to the position of a locus versus its chromosome territory (CT), similar observations were made. The *MHCII* cluster genes, *epidermal differentiation complex (EDC)*

and Hox genes are examples of loci that, upon activation, promote a large-scale relocation of the subchromosomal regions that contain them away from the respective CT [46-48]. FISH allows determining whether a locus adopts a different nuclear position in two cell populations, but it does not provide information about dynamics. Transgenic arrays of bacterial Lac or Tet operator sequences that attract and accumulate the cognate DNA-binding protein tagged with a fluorescent protein enable the visualization of chromosomal integration sites in living cells over time [49]. Previously it was reported that during most of the mammalian cell cycle the positions of such tagged chromosomal subregions remain relatively stable, with chromatin moving by constrained diffusion [50]. This implies that in order for a locus to adopt an entirely new position in the nucleus, passage through mitosis is required. Two recent live cell studies that address the impact of the nuclear lamina on gene expression (discussed below) support this idea [51, 52]. They showed that the targeting of nuclear lamina components to a transgenic locus induced its repositioning to the nuclear periphery only after cell division. However, another live cell study reported rapid and directional movement of a locus away from the periphery of interphase nuclei upon the targeting of a transcriptional activator [53]. A similar directional movement in living cells was also observed for a transgenic U2 snRNA array, which moved towards Cajal bodies (CBs) upon transcriptional activation [54]. CBs are nuclear substructures involved in the biogenesis of certain small RNPs. In human cells, they preferentially associate with certain gene loci, notably the histone genes and small nuclear RNA genes. In this study, activation of U2 gene expression induced a directional movement of the transgenic cassette to CBs in roughly 20% of the cells, causing them to stably associate after approximately 6 hours [54]. Intriguingly, in both studies directional movement was dependent on nuclear actin [53, 54]. However, classical actin filaments have not been found in the nucleus [55] and the mechanism employed by nuclear actin is still unclear [56]. It is currently unclear whether these observations are peculiarities of large transgenic arrays that often contain thousands of operator repeats, or may represent a more widespread phenomenon shared also by endogenous loci. In this regard it would be very informative to target operator sequence arrays near endogenous genes and follow (and manipulate) their behavior when the genes undergo their normal developmental expression program.

### **Interactions between gene loci**

The live cell imaging and FISH studies which document that genes adopt a different position in the nucleus upon alterations in their expression status provide support for

models predicting the existence of transcriptional competent zones in the nucleus. Further evidence for non-random positioning of gene loci comes from recent FISH studies which show that selected genes meet in the nucleus. This has been observed for both active and silenced genes. In an RNA FISH study, the positioning of the active  $\beta$ -globin gene was investigated relative to other active genes located on the same chromosome. The selected genes were shown to co-localize frequently when they are active [57]. Two out of the four interacting genes described were erythroid-specific genes, which may be interpreted to suggest that functionally related genes preferentially contact each other in the nuclear space. *Myc* and *IgH*, two frequent translocation partners in lymphomas, are other examples of genes reported to frequently co-localize when they are active [58]. The interactions are reported to occur at nuclear sites dedicated to transcription, so-called transcription factories. They are therefore predicted to be dependent on transcription, an issue that was recently addressed in two independent studies.

Both studies investigated the effect of blocking transcription elongation by DRB (dichloro-beta- D-ribofuranosylbenzimidazole) and the effect of total transcription inhibition either by  $\alpha$ -amanitin [59], which prevents the binding of PolIII to DNA, or by heat-shock treatment [60], which has less well-defined effects. The contacts made within the gene locus, between the gene promoter and the enhancer elements, were analyzed by 3C in both studies. The effect of inhibition of elongation was found to be very minimal, although one study reported a small, but significant drop in the interactions between the  $\beta$ -globin promoter and a part of the LCR. Inhibition of total transcription by  $\alpha$ -amanitin did not affect the structure of the gene locus either, despite a nearly complete depletion of PolIII from the LCR [59]. In contrast, heat-shock treatment resulted in a decrease in the promoter-LCR interactions [60]. The maintenance of long-range interactions between distant gene loci was also investigated. Neither study found an effect of DRB treatment. The effect of heat-shock treatment was measured by investigating two long-range interactions between gene fragments, in a semi-quantitative 3C experimental set-up. The two interactions were found to be lost after heat-shock treatment. The effect of  $\alpha$ -amanitin was investigated by 4C (see below, **Fig. 3.2**), allowing an unbiased simultaneous analysis of many interactions across the genome [59]. Contacts made by the  $\beta$ -globin locus were found to be essentially the same before and after transcription inhibition. The same was true for the inter- and intrachromosomal interactions formed by a gene-rich housekeeping gene locus. Many high-resolution cryo-FISH experiments confirmed these findings. Thus, where Mitchell and Fraser concluded that long-range interactions between gene loci depend on ongoing transcription or transcription initiation, Palstra et al. concluded that

a 4-h block of transcription was not sufficient to disrupt gene interactions and that PolII was not likely to be crucial for keeping distant DNA loci together. The two studies did not exclude that an initial act of transcription is required for the positioning of loci in the nuclear space. Clearly, future research needs to clarify the exact role of the transcription machinery in mediating gene contacts.

Inter-chromosomal interactions between inactive genes have also been reported. These include the contacts made between the *Ifng* and the  $T_H2$  locus in naive CD4<sup>+</sup> T-cells [61]. They are found to co-localize, at least at one allele, in almost 40% of the cells, as measured by 2D FISH. After differentiation to the  $T_H1$ , or  $T_H2$  lineage, *Ifng* or the  $T_H2$  locus becomes activated, respectively. Co-localization decreases to 10-13% at this stage of differentiation. The authors suggest that the loci are kept in a 'poised chromatin hub' in naive CD4<sup>+</sup> T-cells. In the hub the decision can be made which of the loci will be transcribed. The mono-allelic expression of the genes in the  $T_H2$  locus, and the observation that the co-localization is found mostly at one of the alleles, is suggestive for an intricate regulatory mechanism.

Another recently published example of contacts made between silent genes is the co-localization of *GFAP* and *s100 $\beta$*  [62]. Both genes are silent in neural progenitor cells (NPCs) and become expressed when the cells are differentiated into astrocytes. *GFAP* is expressed mono-allelically. In 20% of NPCs, one *GFAP* allele co-localizes with an allele of *s100 $\beta$* . This co-localization decreases to less than 10% after induction of differentiation. Therefore, like the *Ifng*  $T_H2$  interaction, the loci investigated here could be within a 'poised chromatin hub'.

Together, these recent FISH observations fuel a deterministic model for nuclear organization, as they predict that functionally related genes present on the same or on different chromosomes need to come together to coordinate each other's activity. A word of caution needs to be expressed though concerning the term 'co-localization', since its meaning depends on the microscopy technique used and the definition applied by the investigators. As outlined before [63], it has entirely different meanings in different studies, which obviously has an important impact on the interpretation of the results. We and others have previously also argued that the linear distribution of repetitive DNA sequences and of active and inactive DNA regions is important for the folding and relative positioning of chromosomes, implying that the nuclear position of a locus also depends on the properties of neighboring DNA. If true, it would argue more for a stochastic concept of nuclear organization in which functionally relevant interactions between two selected loci present on different chromosomes will be rare [63, 64]. Below we will discuss

a number of recent studies that take into account the properties of neighboring DNA when interpreting the significance of the nuclear position adopted by a given gene.

### **Studying genes in the context of their subchromosomal surrounding**

*Mash1* is one of the loci reported to change position in the nucleus upon transcription activation [41]. It is a neural gene located relatively isolated on the genome, with a liver specific gene *Pah-1* as a close neighbor. In ES cells the locus is present at the periphery of the nucleus. After differentiation into the neural lineage, *Mash1* becomes expressed and located towards the nuclear interior. *Pah-1* as a consequence also moves interiorly, but does not become activated. Other neural genes located in the periphery in ES cells do not relocate after induction [41], showing that genes can be activated also when present at the outer edge of the nucleus. A similar observation was made in a study that followed the radial position of the  $\beta$ -globin locus during erythroid differentiation [40]. This gene locus adopts a more internal position when activated during erythropoiesis, but early during the differentiation process active  $\beta$ -globin genes are also observed at the periphery. The data show that the up-regulation of some, but not all genes coincides with nuclear internalization. Repositioning may facilitate gene activation, but is neither required nor sufficient. A similar conclusion can be drawn with respect to chromosome territories. It has been shown that the erythroid-specific  $\alpha$ -globin genes, which are surrounded in cis by housekeeping genes, are inactive in non-erythroid cells despite their position outside the CT. Vice versa, the  $\beta$ -globin genes remain inside the CT also when highly transcribed in erythroid cells [65]. Thus, genes do not need to move out of their chromosome territories for expression, in agreement with the observation that sites of active transcription are present throughout the nuclear interior [66, 67]. Looping away from the CT also does not necessarily increase transcription rates. *Lnp*, a gene located near the *Hoxd* cluster, was shown to be active and not change its expression level upon looping away from its CT during ES cell differentiation [68]. The relevance of DNA movement relative to the CT was recently further addressed in a study that involved the integration of the  $\beta$ -globin LCR, without  $\beta$ -globin genes, into a gene-dense region of mouse chromosome 8 [69]. Integration was done in both orientations. In erythroid cells taken from transgenic mice, the two oppositely oriented LCRs each caused a repositioning of the locus that carried them away from the CT. Many genes, as far as 150 kilobases away from the integration site, showed higher transcription rates with both LCRs. A second category of genes present in between the activated genes was found that, similar to *Lnp*, did not increase their expression in response to relocation by either of the two LCRs. While chromatin or

gene-intrinsic properties may preclude further activation of these genes, this was clearly not the case for a third category of genes present near the integrated LCRs. This category represented genes that increased their transcriptional activity, but only in response to one orientation of the LCR. Since both LCRs move the region away from its CT, the existence of these genes showed that repositioning cannot be the driving force behind their transcriptional upregulation [69]. Collectively, the data show that the expression status of a gene may correlate with, but does not depend on the position relative to the CT. The studies highlight the importance to also consider surrounding sequences and genes when interpreting the significance of nuclear location.

### High-throughput studies on nuclear architecture

FISH, no matter how revealing, has two major disadvantages: it is biased towards the loci or nuclear structures that were selected for analysis and it can only analyze a limited number of loci simultaneously. It is impossible to know whether observations made by FISH uncover general concepts that also apply to the rest of the genome or reflect a peculiarity of the gene locus investigated. It is also very difficult, if not impossible, to discover gene interaction networks based on testing candidate loci by FISH. In order to understand the general concepts behind nuclear architecture, high-throughput approaches need to be applied that screen the entire genome in an unbiased manner for interactions between genomic sites and with nuclear substructures. In recent years, several of these approaches have been developed.

DamID is a high-throughput strategy that, like ChIP-chip, uncovers the genomic binding sites of selected proteins. It involves the identification of DNA sequences carrying the methylation mark that is deposited by the bacterial methylase *dam*, which is expressed as a fusion to the protein of interest. In a recent DamID experiment, the interaction sites of an important structural component of the nuclear lamina, Lamin B1, were investigated in human lung fibroblasts [36]. This study provided important information about genome organization inside the nucleus and the relationship between the nuclear periphery and gene expression. Lamin B1 was shown to interact with remarkably distinct domains in the human genome, with an average size of 500 kb. The Lamin associated domains (LADs) are characterized as being low in gene density and gene activity, although active genes could be found in LADs and non-LADs also carried silent genes. The profiles of gene density and activity show a steep transition at the LAD borders, illustrating that LADs and non-LADs are functionally different domains within the human genome. The transition area between domains is on the non-LAD side often characterized by CTCF binding sites, CpG islands

and promoters facing away from the LAD. These characteristics could all be involved in creating a boundary, preventing spreading of heterochromatin. However, together these characteristics cover only 30% of the transition sites, suggesting more factors are to be discovered. It is important to mention that FISH experiments showed that LADs could also be found internally in a proportion of cells, illustrating that lamina association of inactive domains is probabilistic [36]. Thus, the nuclear periphery is enriched for inactive domains of the genome. Do these regions move there as a consequence of their silenced state, or is transcription downregulated when genes are positioned at the edge of the nucleus?

The impact of the nuclear lamina on gene expression was recently further explored in three independent live cell imaging studies which all used bacterial operator sequences to target a locus to the periphery [51, 52, 70]. Two groups investigated the effect of repositioning on the expression of a linked transgene. Kumaran et al reported that the relocated transgene could be fully induced, with kinetics similar to that observed at an internal position. They measured expression levels by targeting a fluorescent label to MS2 repeats integrated in the transgenic transcript [51]. Reddy et al however reported downregulation of their transgene upon targeting to the periphery [52]. The third study by Finlan et al involved the targeting of two different endogenous loci to the nuclear periphery, again via Lac operator arrays [70]. At both positions some, but not all genes surrounding the targeted site were downregulated and the same was true for some genes at a large distance *in cis*. Thus, one study found no effect on gene activation [51] while two others observed gene silencing upon targeting to the nuclear periphery [52, 70]. One possible explanation for these apparently contradicting results is that the first study measured transcription activation while the others investigated maintenance of transcription; different activities may be required for these different processes. In addition, the DamID results suggest that the type of promoter, the position in the genome and possibly other variables may also influence the outcome of these experiments. Collectively, the studies show that transcription can occur at the nuclear periphery. However, the outer shell of the nucleus tends to be occupied by the more inactive regions of the genome. Genes can be subject to silencing when targeted to the nuclear lamina, suggesting that the periphery can play an active role in transcriptional repression. In other words, the accumulation of inactive chromatin at the periphery may not only be a consequence of their silenced state and nuclear internalization may indeed facilitate gene expression. Components of the lamina have been shown to interact with HDAC3, a deacetylating enzyme, which may be actively involved in gene silencing at the periphery.

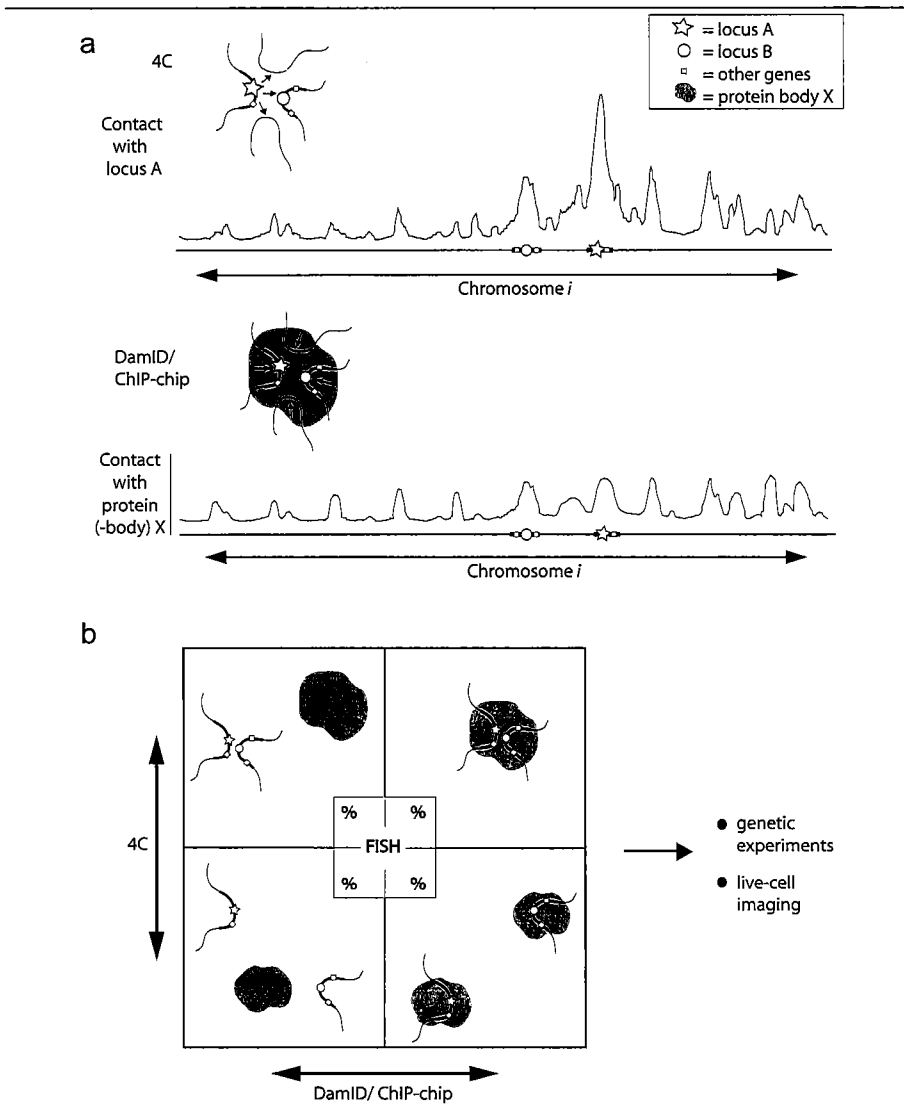
In agreement, down-regulation was accompanied by decreased histone acetylation at the genes [52, 70].

The DamID study for the first time mapped the genomic regions that contact a certain subnuclear compartment. It would be interesting to also investigate the co-localization with proteins found in other nuclear structures, such as PML bodies, CBs and nuclear speckles. Such studies may reveal the function of these bodies, which currently still is rather enigmatic. They may also uncover concepts that shape the genome inside the cell nucleus.

While DamID and ChIP-chip experiments provide genome-wide binding profiles of proteins, several other high-throughput techniques have recently been developed that aim to analyze DNA-DNA interactions across the genome. All these techniques are based on 3C technology [71]. 4C technology (3C-on-chip or circular 3C) involves screening the genome in an unbiased manner for regions that interact with a locus of choice [30, 72].

4C applied to the active  $\beta$ -globin locus identified interactions with many gene clusters containing active genes and with some single active gene loci. The erythroid-specific genes previously identified by FISH were among the contacts found by 4C [57]. In the context of all the interactions found, no bias was observed for the  $\beta$ -globin locus to preferentially share nuclear sites with genes that are functionally related. Thus, 4C puts individual interactions measured by FISH in the context of the entire spectrum of DNA contacts. The co-localization of the  $\beta$ -globin locus with active parts of the same chromosome was specific for the active state of the locus in fetal liver cells. When 4C was applied to the inactive  $\beta$ -globin locus in brain cells, contacts were identified with completely different regions of mouse chromosome 7. These regions did not carry genes or only contained inactive genes and located more towards the centromere of mouse chromosome 7. An active housekeeping gene, *Rad23a*, located in a gene dense locus, co-localized with many active, gene dense region elsewhere on the chromosome and on other chromosomes. These interactions were conserved between fetal liver and brain tissue, despite differences in expression of some of the genes located within the gene dense areas. The data demonstrated at the level of DNA contacts that active and inactive chromatin separate in the nucleus. Since active and inactive chromosomal segments each appear to have their own preferred interaction partners, it was proposed that the nuclear environment of a gene is not only determined by the gene itself but also by the surrounding sequences [30]. In case of  $\beta$ -globin, the active state of the ~200 kb locus is dominant over the flanking silent chromatin. It would be interesting to investigate the





**Figure 3.3. Schematic representation of methods to study nuclear organization.** Different methods provide different information on interactions between gene loci and between genes and nuclear protein (-structures). (a) 4C investigates the DNA interactions made with a given gene locus 'A', resulting in a spectrum of interactions across chromosomes. DamID and ChIP-chip generate a genome-wide map of interactions with protein 'X'. (b) FISH studies can determine the frequencies of such interactions, and the relation between protein and DNA contacts made by a locus. Even if a protein binds to two DNA sites that contact each other, it does not need to function in loop formation. For example, NF-E2 binds to the  $\beta$ -globin LCR and to the adult  $\beta$ -globin gene promoter which form a stable chromatin loop, but this loop is maintained also in a NF-E2-null background [76]. The functional relevance of DNA contacts should indeed be determined by genetic studies. Live cell imaging studies can give insight in the dynamics of the interactions.

nuclear environment of a gene that is alternatively expressed between tissues and is located in a gene dense, active area of the genome.

While the 4C study supports the idea that stochastic principles underlie nuclear organisation [63], other studies using similar methodology reported data which suggest that the nucleus is ordered according to more deterministic rules: specific genes present on different chromosomes would come together in the nucleus. Two studies used the *H19/Igf2* locus as their target to screen for DNA interactions [72, 73]. Surprisingly, they identified completely different interactions. Ling et al applied a strategy referred to as the associated chromosome trap (ACT) assay. They found three interacting fragments and focussed on a parent-of-origin specific interaction between the maternal allele of the *H19/Igf2* locus and the paternal allele of the *Wsb1/Nf1* locus [73]. Zhao et al applied 4C technology, which in their case stands for circular 3C, and sequenced 114 captured fragments. They reported interactions with regions on all mouse chromosomes and an overrepresentation of imprinted gene loci, suggesting that epigenetic mechanisms cause their clustering [72]. The number of sequences analyzed in both studies is limited and therefore both data sets may not provide the entire picture of the long-range interactions formed by the *H19/Igf2* locus. This may explain why the results of the two studies do not necessarily agree.

### Future perspectives

The recently developed high-throughput methods to study nuclear architecture have provided exciting new insight into nuclear organization. Genome-wide mapping studies of protein-DNA interactions have proven to be a valuable method to describe the genomic regions that are frequently found near proteinaceous structures in the nucleus. Novel methods that identify all co-localizing sequences of a gene locus put selected interactions measured in FISH studies into perspective. Only an appreciation of the full spectrum of DNA interactions allows defining the concepts of genomic architecture (**Fig. 3.3**). Data obtained with the current strategies do not necessarily always agree though. In part this will be due to the fact that the technologies are new and need to be further developed. It is important to recognize that all the strategies based on 3C involve PCR to enrich for the interactions of interest and that PCR can introduce a bias in the assay. Results therefore always need to be verified by FISH, preferably 3D or cryo-FISH. FISH also allows studying DNA interactions in single cells and determining the percentage of alleles that interact at a given time.

Several reports based on 3C-variants show spectacular, highly specific inter-chromosomal interactions between selected gene loci. These data support a deterministic form of nuclear organization, where gene loci are guided to specific partners located on unrelated chromosomes. Conclusive evidence that such interactions are functionally important needs to come from genetic studies showing that the deletion of genomic parts on the one chromosome affect the expression of genes on the other chromosomes. In this respect it is worth to refer to the studies by Fuss et al [74] who showed that the deletion of an enhancer previously claimed to activate olfactory receptor genes throughout the genome [75] only affected the expression of genes nearby on the chromosome. Based on our 4C data, we have argued that the genome is shaped according to self-organizing principles. In this stochastic concept, specific gene loci will have a very difficult time finding each other, as their nuclear position depends not only on the gene itself but also on the properties of neighboring sequences and, by extrapolation, of the entire chromosome. Clearly, we are only at the beginning of an era dedicated to the uncovering of DNA topology inside the living cell nucleus. Future will tell which principles shape the nucleus and how the conformation of chromatin influences a process like gene expression.

### **Acknowledgements**

This work was supported by grants from the Dutch Scientific Organization (NWO) (912-04-082) and the Netherlands Genomics Initiative (050-71-324).

### **References**

- [1] D.A. Kleinjan, V. van Heyningen, Long-range control of gene expression: emerging mechanisms and disruption in disease, *Am J Hum Genet* 76 (2005) 8-32.
- [2] E. Epner, A. Reik, D. Cimborra, A. Telling, M.A. Bender, S. Fiering, T. Enver, D.I. Martin, M. Kennedy, G. Keller, M. Groudine, The beta-globin LCR is not necessary for an open chromatin structure or developmentally regulated transcription of the native mouse beta-globin locus, *Mol Cell* 2 (1998) 447-455.
- [3] F. Grosveld, G.B. van Assendelft, D.R. Greaves, G. Kollias, Position-independent, high-level expression of the human beta-globin gene in transgenic mice, *Cell* 51 (1987) 975-985.
- [4] A. Reik, A. Telling, G. Zitnik, D. Cimborra, E. Epner, M. Groudine, The locus control region is necessary for gene expression in the human beta-globin locus but not the

maintenance of an open chromatin structure in erythroid cells, *Mol Cell Biol* 18 (1998) 5992-6000.

[5] D. Carter, L. Chakalova, C.S. Osborne, Y.F. Dai, P. Fraser, Long-range chromatin regulatory interactions in vivo, *Nat Genet* 32 (2002) 623-626.

[6] B. Tolhuis, R.J. Palstra, E. Splinter, F. Grosveld, W. de Laat, Looping and interaction between hypersensitive sites in the active beta-globin locus, *Mol Cell* 10 (2002) 1453-1465.

[7] J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing chromosome conformation, *Science* 295 (2002) 1306-1311.

[8] R.J. Palstra, B. Tolhuis, E. Splinter, R. Nijmeijer, F. Grosveld, W. de Laat, The beta-globin nuclear compartment in development and erythroid differentiation, *Nat Genet* 35 (2003) 190-194.

[9] R. Drissen, R.J. Palstra, N. Gillemans, E. Splinter, F. Grosveld, S. Philipsen, W. de Laat, The active spatial organization of the beta-globin locus requires the transcription factor EKLF, *Genes Dev* 18 (2004) 2485-2490.

[10] C.R. Vakoc, D.L. Letting, N. Gheldof, T. Sawado, M.A. Bender, M. Groudine, M.J. Weiss, J. Dekker, G.A. Blobel, Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1, *Mol Cell* 17 (2005) 453-462.

[11] C.G. Spilianakis, R.A. Flavell, Long-range intrachromosomal interactions in the T helper type 2 cytokine locus, *Nat Immunol* 5 (2004) 1017-1027.

[12] D. Vernimmen, M. De Gobbi, J.A. Sloane-Stanley, W.G. Wood, D.R. Higgs, Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression, *Embo J* 26 (2007) 2041-2051.

[13] G.L. Zhou, L. Xin, W. Song, L.J. Di, G. Liu, X.S. Wu, D.P. Liu, C.C. Liang, Active chromatin hub of the mouse alpha-globin locus forms in a transcription factory of clustered housekeeping genes, *Mol Cell Biol* 26 (2006) 5096-5105.

[14] H. Jing, C.R. Vakoc, L. Ying, S. Mandat, H. Wang, X. Zheng, G.A. Blobel, Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus, *Mol Cell* 29 (2008) 232-242.

[15] S. Horike, S. Cai, M. Miyano, J.F. Cheng, T. Kohwi-Shigematsu, Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome, *Nat Genet* 37 (2005) 31-40.

[16] S. Kurukuti, V.K. Tiwari, G. Tavoosidana, E. Pugacheva, A. Murrell, Z. Zhao, V. Lobanenkov, W. Reik, R. Ohlsson, CTCF binding at the H19 imprinting control region

mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to *Igf2*, *Proc Natl Acad Sci U S A* 103 (2006) 10684-10689.

[17] A. Murrell, S. Heeson, W. Reik, Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops, *Nat Genet* 36 (2004) 889-893.

[18] E. Splinter, H. Heath, J. Kooren, R.J. Palstra, P. Klous, F. Grosveld, N. Galjart, W. de Laat, CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus, *Genes Dev* 20 (2006) 2349-2354.

[19] J.M. O'Sullivan, S.M. Tan-Wong, A. Morillon, B. Lee, J. Coles, J. Mellor, N.J. Proudfoot, Gene loops juxtapose promoters and terminators in yeast, *Nat Genet* 36 (2004) 1014-1018.

[20] K.J. Perkins, M. Lusic, I. Mitar, M. Giacca, N.J. Proudfoot, Transcription-dependent gene looping of the HIV-1 provirus is dictated by recognition of pre-mRNA processing signals, *Mol Cell* 29 (2008) 56-68.

[21] J. Yao, M.B. Ardehali, C.J. Fecko, W.W. Webb, J.T. Lis, Intranuclear distribution and local dynamics of RNA polymerase II during transcription activation, *Mol Cell* 28 (2007) 978-990.

[22] K.J. Oestreich, R.M. Cobb, S. Pierce, J. Chen, P. Ferrier, E.M. Oltz, Regulation of TCRbeta gene assembly by a promoter/enhancer holocomplex, *Immunity* 24 (2006) 381-391.

[23] C. Sayegh, S. Jhunjunwala, R. Riblet, C. Murre, Visualization of looping involving the immunoglobulin heavy-chain locus in developing B cells, *Genes Dev* 19 (2005) 322-327.

[24] J.A. Skok, R. Gisler, M. Novatchkova, D. Farmer, W. de Laat, M. Busslinger, Reversible contraction by looping of the *Tcra* and *Tcrb* loci in rearranging thymocytes, *Nat Immunol* 8 (2007) 378-387.

[25] S. Hell, E.H.K. Stelzer Properties of a 4Pi confocal fluorescence microscope, *Journal of the Optical Society of America A* 9 (1992) 2159-.

[26] W. de Laat, F. Grosveld, Spatial organization of gene expression: the active chromatin hub, *Chromosome Res* 11 (2003) 447-459.

[27] P. Droge, B. Muller-Hill, High local protein concentrations at promoters: strategies in prokaryotic and eukaryotic cells, *Bioessays* 23 (2001) 179-183.

[28] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Muller, R. Eils, C. Cremer, M.R. Speicher, T. Cremer, Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes, *PLoS Biol* 3 (2005) e157.

- [29] M.R. Branco, A. Pombo, Intermingling of Chromosome Territories in Interphase Suggests Role in Translocations and Transcription-Dependent Associations, *PLoS Biol* 4 (2006) e138.
- [30] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, W. de Laat, Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C), *Nat Genet* 38 (2006) 1348-1354.
- [31] D. Hernandez-Verdun, The nucleolus: a model for the organization of nuclear functions, *Histochem Cell Biol* 126 (2006) 135-148.
- [32] A.V. Probst, G. Almouzni, Pericentric heterochromatin: dynamic organization during early development in mammals, *Differentiation* 76 (2008) 15-23.
- [33] R. Mayer, A. Brero, J. von Hase, T. Schroeder, T. Cremer, S. Dietzel, Common themes and cell type specific variations of higher order chromatin arrangements in the mouse, *BMC Cell Biol* 6 (2005) 44.
- [34] M.J. Lercher, A.O. Urrutia, L.D. Hurst, Clustering of housekeeping genes provides a unified model of gene order in the human genome, *Nat Genet* 31 (2002) 180-183.
- [35] D. Sproul, N. Gilbert, W.A. Bickmore, The role of chromatin structure in regulating the expression of clustered genes, *Nat Rev Genet* 6 (2005) 775-781.
- [36] L. Guelen, L. Pagie, E. Brassat, W. Meuleman, M.B. Faza, W. Talhout, B.H. Eussen, A. de Klein, L. Wessels, W. de Laat, B. van Steensel, Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions, *Nature* (2008).
- [37] L.S. Shopland, C.R. Lynch, K.A. Peterson, K. Thornton, N. Kepper, J. Hase, S. Stein, S. Vincent, K.R. Molloy, G. Kreth, C. Cremer, C.J. Bult, T.P. O'Brien, Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence, *J Cell Biol* 174 (2006) 27-38.
- [38] S.T. Kosak, J.A. Skok, K.L. Medina, R. Riblet, M.M. Le Beau, A.G. Fisher, H. Singh, Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development, *Science* 296 (2002) 158-162.
- [39] D. Zink, M.D. Amaral, A. Englmann, S. Lang, L.A. Clarke, C. Rudolph, F. Alt, K. Luther, C. Braz, N. Sadoni, J. Rosenecker, D. Schindelbauer, Transcription-dependent spatial arrangements of CFTR and adjacent genes in human cell nuclei, *J Cell Biol* 166 (2004) 815-825.
- [40] T. Ragoczy, A. Telling, T. Sawado, M. Groudine, S.T. Kosak, A genetic analysis of chromosome territory looping: diverse roles for distal regulatory elements, *Chromosome Res* 11 (2003) 513-525.

- [41] R.R. Williams, V. Azuara, P. Perry, S. Sauer, M. Dvorkina, H. Jorgensen, J. Roix, P. McQueen, T. Misteli, M. Merkenschlager, A.G. Fisher, Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus, *J Cell Sci* 119 (2006) 132-140.
- [42] K.E. Brown, J. Baxter, D. Graf, M. Merkenschlager, A.G. Fisher, Dynamic repositioning of genes in the nucleus of lymphocytes preparing for cell division, *Mol Cell* 3 (1999) 207-217.
- [43] K.E. Brown, S.S. Guest, S.T. Smale, K. Hahm, M. Merkenschlager, A.G. Fisher, Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin, *Cell* 91 (1997) 845-854.
- [44] J.L. Grogan, M. Mohrs, B. Harmon, D.A. Lacy, J.W. Sedat, R.M. Locksley, Early transcription and silencing of cytokine genes underlie polarization of T helper cell subsets, *Immunity* 14 (2001) 205-215.
- [45] S.L. Hewitt, F.A. High, S.L. Reiner, A.G. Fisher, M. Merkenschlager, Nuclear repositioning marks the selective exclusion of lineage-inappropriate transcription factor loci during T helper cell differentiation, *Eur J Immunol* 34 (2004) 3604-3613.
- [46] S. Chambeyron, W.A. Bickmore, Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription, *Genes Dev* 18 (2004) 1119-1130.
- [47] E.V. Volpi, E. Chevret, T. Jones, R. Vatcheva, J. Williamson, S. Beck, R.D. Campbell, M. Goldsworthy, S.H. Powis, J. Ragoussis, J. Trowsdale, D. Sheer, Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei, *J Cell Sci* 113 ( Pt 9) (2000) 1565-1576.
- [48] R.R. Williams, S. Broad, D. Sheer, J. Ragoussis, Subchromosomal positioning of the epidermal differentiation complex (EDC) in keratinocyte and lymphoblast interphase nuclei, *Exp Cell Res* 272 (2002) 163-175.
- [49] C.C. Robinett, A. Straight, G. Li, C. Wilhelm, G. Sudlow, A. Murray, A.S. Belmont, In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition, *J Cell Biol* 135 (1996) 1685-1700.
- [50] J.R. Chubb, S. Boyle, P. Perry, W.A. Bickmore, Chromatin motion is constrained by association with nuclear compartments in human cells, *Curr Biol* 12 (2002) 439-445.
- [51] R.I. Kumaran, D.L. Spector, A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence, *J Cell Biol* 180 (2008) 51-65.
- [52] K.L. Reddy, J.M. Zullo, E. Bertolino, H. Singh, Transcriptional repression mediated by repositioning of genes to the nuclear lamina, *Nature* 452 (2008) 243-247.

- [53] C.H. Chuang, A.E. Carpenter, B. Fuchsova, T. Johnson, P. de Lanerolle, A.S. Belmont, Long-range directional movement of an interphase chromosome site, *Curr Biol* 16 (2006) 825-831.
- [54] M. Dunder, J.K. Ospina, M.H. Sung, S. John, M. Upender, T. Ried, G.L. Hager, A.G. Matera, Actin-dependent intranuclear repositioning of an active gene locus in vivo, *J Cell Biol* 179 (2007) 1095-1103.
- [55] S.M. Gonsior, S. Platz, S. Buchmeier, U. Scheer, B.M. Jockusch, H. Hinssen, Conformational difference between nuclear and cytoplasmic actin as detected by a monoclonal antibody, *J Cell Sci* 112 ( Pt 6) (1999) 797-809.
- [56] T. Pederson, U. Aebi, Nuclear actin extends, with no contraction in sight, *Mol Biol Cell* 16 (2005) 5055-5060.
- [57] C.S. Osborne, L. Chakalova, K.E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J.A. Mitchell, S. Lopes, W. Reik, P. Fraser, Active genes dynamically colocalize to shared sites of ongoing transcription, *Nat Genet* 36 (2004) 1065-1071.
- [58] C.S. Osborne, L. Chakalova, J.A. Mitchell, A. Horton, A.L. Wood, D.J. Bolland, A.E. Corcoran, P. Fraser, Myc dynamically and preferentially relocates to a transcription factory occupied by Igh, *PLoS Biol* 5 (2007) e192.
- [59] R.J. Palstra, M. Simonis, P. Klous, E. Brassat, B. Eijkelkamp, W. de Laat, Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription, *PLoS ONE* 3 (2008) e1661.
- [60] J.A. Mitchell, P. Fraser, Transcription factories are nuclear subcompartments that remain in the absence of transcription, *Genes Dev* 22 (2008) 20-25.
- [61] C.G. Spilianakis, M.D. Lalioti, T. Town, G.R. Lee, R.A. Flavell, Interchromosomal associations between alternatively expressed loci, *Nature* 435 (2005) 637-645.
- [62] T. Takizawa, P.R. Gudla, L. Guo, S. Lockett, T. Misteli, Allele-specific nuclear positioning of the monoallelically expressed astrocyte marker GFAP, *Genes Dev* 22 (2008) 489-498.
- [63] W. de Laat, F. Grosveld, Inter-chromosomal gene regulation in the mammalian cell nucleus, *Curr Opin Genet Dev* 17 (2007) 456-464.
- [64] T. Misteli, Beyond the sequence: cellular organization of genome function, *Cell* 128 (2007) 787-800.
- [65] J.M. Brown, J. Leach, J.E. Reittie, A. Atzberger, J. Lee-Prudhoe, W.G. Wood, D.R. Higgs, F.J. Iborra, V.J. Buckle, Coregulated human globin genes are frequently in spatial proximity when active, *J Cell Biol* 172 (2006) 177-187.



- [66] D.A. Jackson, A.B. Hassan, R.J. Errington, P.R. Cook, Visualization of focal sites of transcription within human nuclei, *Embo J* 12 (1993) 1059-1065.
- [67] D.G. Wansink, W. Schul, I. van der Kraan, B. van Steensel, R. van Driel, L. de Jong, Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus, *J Cell Biol* 122 (1993) 283-293.
- [68] C. Morey, N.R. Da Silva, P. Perry, W.A. Bickmore, Nuclear reorganisation and chromatin decondensation are conserved, but distinct, mechanisms linked to Hox gene activation, *Development* 134 (2007) 909-919.
- [69] D. Noordermeer, M.R. Branco, E. Splinter, P. Klous, W. van Ijcken, S. Swagemakers, M. Koutsourakis, P. van der Spek, A. Pombo, W. de Laat, Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region, *PLoS Genet* 4 (2008) e1000016.
- [70] L.E. Finlan, D. Sproul, I. Thomson, S. Boyle, E. Kerr, P. Perry, B. Ylstra, J.R. Chubb, W.A. Bickmore, Recruitment to the nuclear periphery can alter expression of genes in human cells, *PLoS Genet* 4 (2008) e1000039.
- [71] M. Simonis, J. Kooren, W. de Laat, An evaluation of 3C-based methods to capture DNA interactions, *Nat Methods* 4 (2007) 895-901.
- [72] Z. Zhao, G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K.S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, R. Ohlsson, Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions, *Nat Genet* 38 (2006) 1341-1347.
- [73] J.Q. Ling, T. Li, J.F. Hu, T.H. Vu, H.L. Chen, X.W. Qiu, A.M. Cherry, A.R. Hoffman, CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1, *Science* 312 (2006) 269-272.
- [74] S.H. Fuss, M. Omura, P. Mombaerts, Local and cis effects of the H element on expression of odorant receptor genes in mouse, *Cell* 130 (2007) 373-384.
- [75] S. Lomvardas, G. Barnea, D.J. Pisapia, M. Mendelsohn, J. Kirkland, R. Axel, Interchromosomal interactions and olfactory receptor choice, *Cell* 126 (2006) 403-413.
- [76] J. Kooren, R.J. Palstra, P. Klous, E. Splinter, M. von Lindern, F. Grosveld, W. de Laat, Beta-globin active chromatin Hub formation in differentiating erythroid cells and in p45 NF-E2 knock-out mice, *J Biol Chem* 282 (2007) 16544-16552.



# 4

Nuclear organization of active and inactive chromatin domains uncovered by 3C on chip (4C)

## Nuclear organization of active and inactive chromatin domains uncovered by 3C on chip (4C)

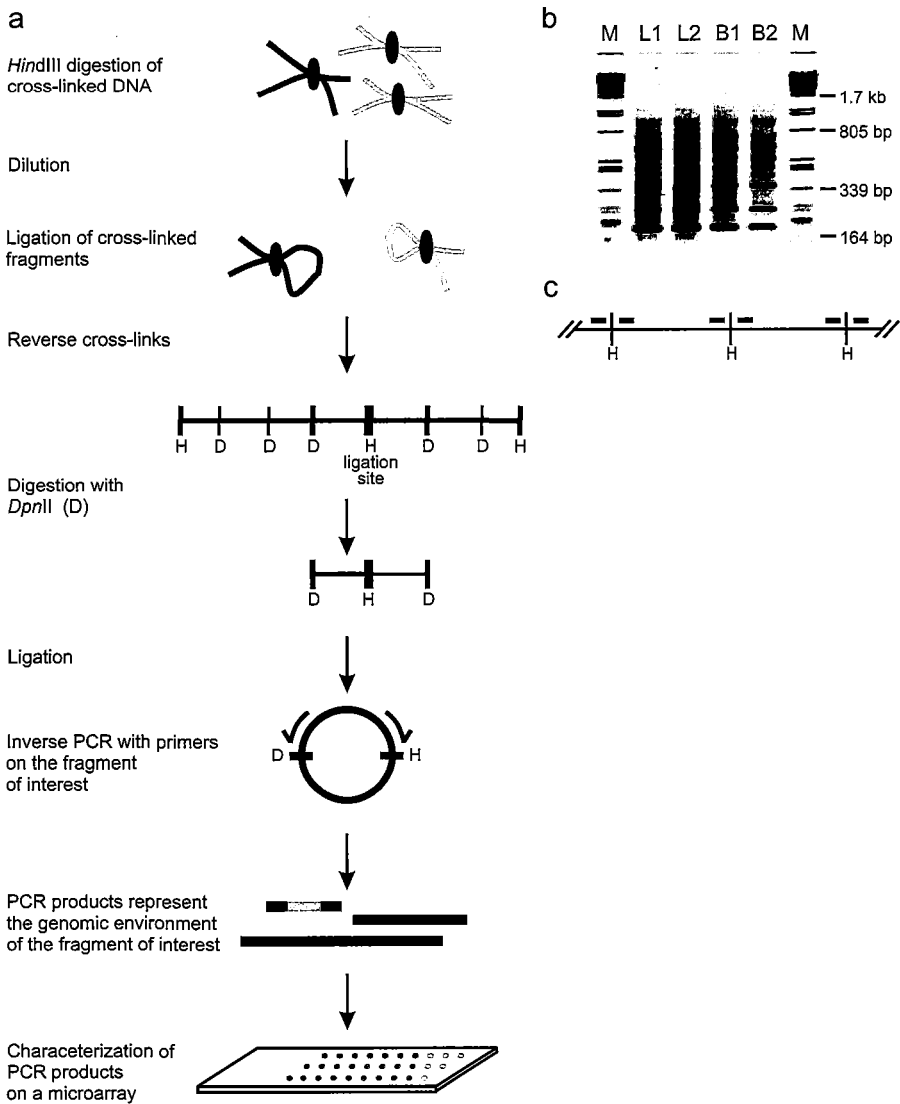
Marieke Simonis<sup>1</sup>, Petra Klous<sup>1</sup>, Erik Splinter<sup>1</sup>, Yuri Moshkin<sup>2</sup>, Rob Willemsen<sup>3</sup>, Elzo de Wit<sup>4</sup>, Bas van Steensel<sup>4</sup> & Wouter de Laat<sup>1</sup>

<sup>1</sup>Department of Cell Biology and Genetics, <sup>2</sup>Department of Biochemistry, <sup>3</sup>Department of Clinical Genetics, Erasmus Medical Centre, PO Box 2040, 3000 CA Rotterdam, The Netherlands. <sup>4</sup>Division of Molecular Biology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. Correspondence should be addressed to W.d.L.

### Summary

*The spatial organization of DNA in the cell nucleus is an emerging key contributor to genomic function<sup>1-12</sup>. We developed 3C on chip technology (4C), which allows for an unbiased genome-wide search for DNA loci that contact a given locus in the nuclear space. We demonstrate here that active and inactive genes are engaged in many long-range intrachromosomal interactions and can also form interchromosomal contacts. The active  $\beta$ -globin locus in fetal liver preferentially contacts transcribed, but not necessarily tissue-specific, loci elsewhere on chromosome 7, whereas the inactive locus in fetal brain contacts different transcriptionally silent loci. A housekeeping gene in a gene-dense region on chromosome 8 forms long-range contacts predominantly with other active gene clusters, both in cis and in trans, and many of these intra- and interchromosomal interactions are conserved between the tissues analyzed. Our data demonstrate that chromosomes fold into areas of active chromatin and areas of inactive chromatin and establish 4C technology as a powerful tool to study nuclear architecture.*

Our understanding of genomic organization in the nuclear space is based mostly on microscopy studies that often use fluorescence in situ hybridization (FISH) to visualize selected parts of the genome. However, FISH, no matter how revealing, can analyze only a limited number of DNA loci simultaneously and therefore produces largely anecdotal observations. In order to get a rigorous picture of nuclear architecture, there is a need for high-throughput technology that can systematically screen the whole genome in an unbiased manner for DNA loci that contact each other in the nuclear space. To this end we



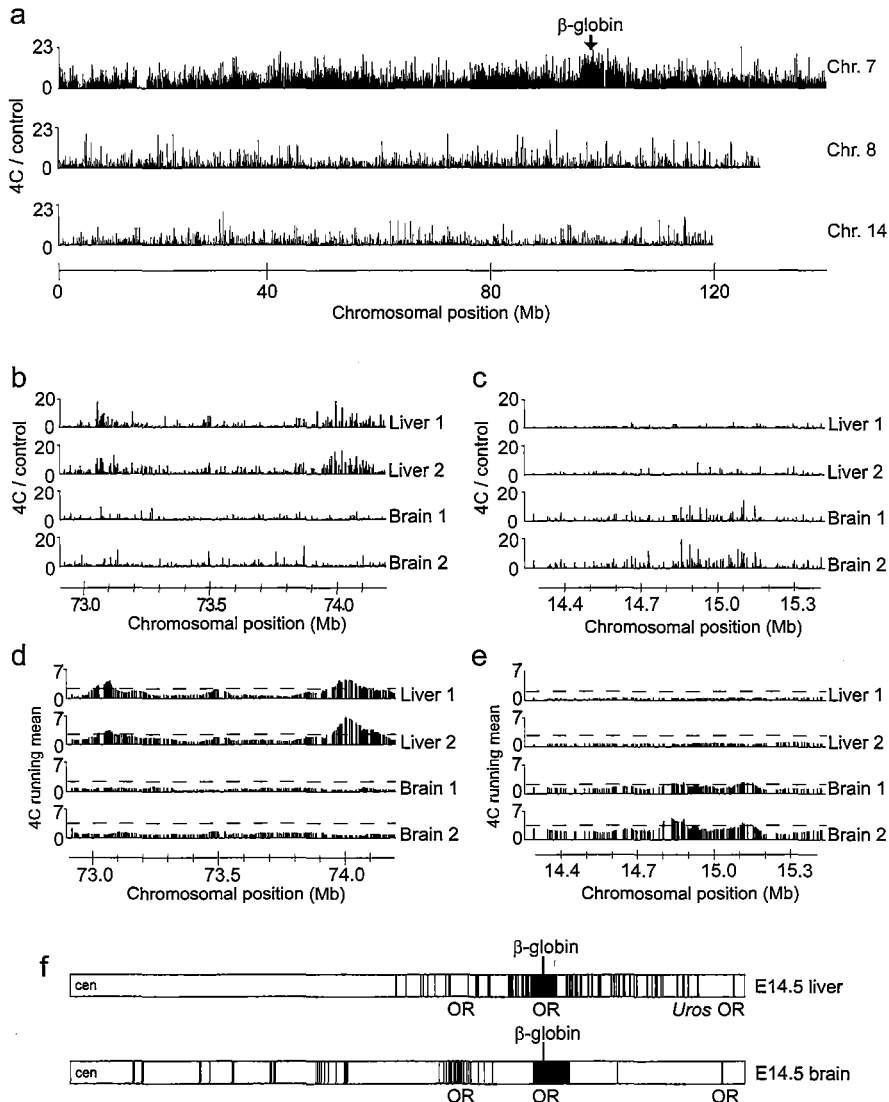
**Figure 4.1 4C technology.** (a) Outline of 4C procedure. Briefly, 3C analysis is performed as usual, but the PCR step is omitted. The 3C template contains bait (for example, a restriction fragment encompassing a gene) ligated to many different fragments (representing this gene's genomic environment). The ligated fragments are cleaved by a frequently cutting secondary restriction enzyme and are subsequently religated to form small DNA circles that are amplified by inverse PCR (30 cycles) using bait-specific primers facing outward. (b) PCR results separated by gel electrophoresis from two independent fetal liver (L1, L2) and brain (B1, B2) samples. (c) Schematic representation of the location of the microarray probes. Probes were designed within 100 bp of *Hind*III sites. Thus, each probe analyzes one possible ligation partner.

have developed 4C technology, which combines chromosome conformation capture (3C) technology<sup>13</sup> with dedicated microarrays.

An outline of the 4C procedure is given in **Figure 4.1a**, and it is explained in detail in the Methods section. In short, 4C involves PCR amplification of DNA fragments cross-linked and ligated to a DNA restriction fragment of choice (here, *HindIII* fragments). Typically, this yields a pattern of PCR fragments specific for a given tissue and highly reproducible between independent PCR reactions (**Fig. 4.1b**). The amplified material, representing the fragment's genomic environment, is labeled and hybridized to a tailored microarray that contains probes each located <100 bp from a different *HindIII* restriction end in the genome (**Fig. 4.1c**). The array used for this study (from Nimblegen Systems) covered seven complete mouse chromosomes.

We applied 4C technology to characterize the genomic environment of the active and inactive mouse  $\beta$ -globin locus, located in a large olfactory receptor gene cluster on chromosome 7. We focused our analysis on a restriction fragment containing hypersensitive site 2 (HS2) of the  $\beta$ -globin locus control region (LCR). Both in embryonic day (E)14.5 liver, where the  $\beta$ -globin genes are highly transcribed, and in E14.5 brain, where the locus is inactive, we found that the great majority of interactions were with sequences on chromosome 7, and we detected very few LCR interactions with six unrelated chromosomes (8, 10, 11, 12, 14 and 15; **Fig. 4.2a**). We found the strongest signals on chromosome 7 within a 5- to 10-Mb region centered around the chromosomal position of  $\beta$ -globin, in agreement with the idea that interaction frequencies are inversely proportional to the distance (in bp) between physically linked DNA sequences<sup>13</sup>. It was not possible to interpret the interactions in this region quantitatively because corresponding probes on the array were saturated (see Methods).

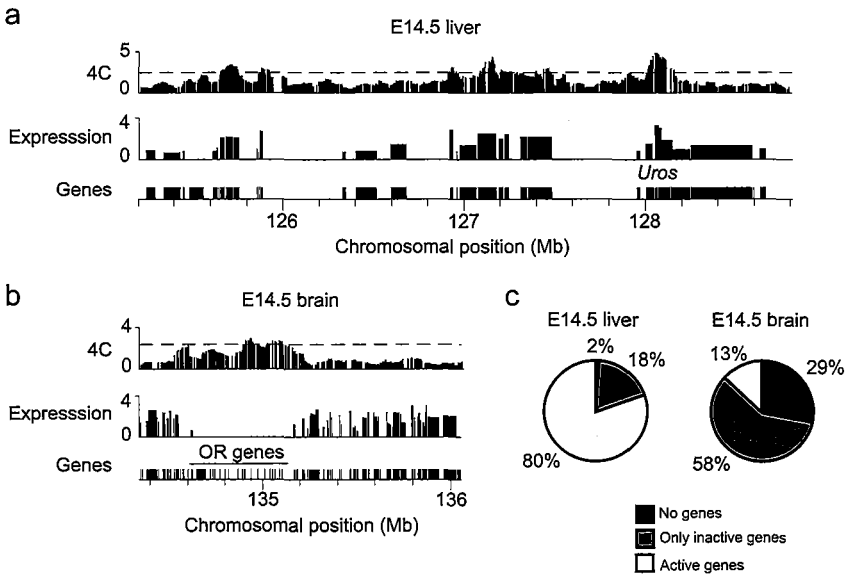
Both in fetal liver and in brain, we identified clusters of 20–50 positive signals juxtaposed on chromosome 7, often at chromosomal locations tens of Mb away from  $\beta$ -globin (**Fig. 4.2b,c**). To determine the statistical significance of these clusters, we ordered data of individual experiments on chromosomal maps and analyzed them using a running mean algorithm with a window size of approximately 60 kb. We then used the running mean distribution of randomly shuffled data to set a threshold value, allowing a false discovery rate of 5%. This analysis identified 66 clusters in fetal liver and 45 in brain that were reproducibly found in duplicate experiments (**Fig. 4.2d-f**). Indeed, high-resolution FISH confirmed that such clusters truly represent loci that interact frequently (see below). Thus, 4C technology identifies long-range interacting loci by the detection of independent ligation events with multiple restriction fragments clustered at a chromosomal position.



**Figure 4.2 Long-range interactions with  $\beta$ -globin, as shown by 4C technology.** (a) Unprocessed ratios of 4C-to-control hybridization signals, showing interactions of  $\beta$ -globin HS2 with chromosome 7 and two unrelated chromosomes (8 and 14). (b,c) Unprocessed data for two independent fetal liver and fetal brain samples plotted along two different 1- to 2-Mb regions on chromosome 7. Highly reproducible clusters of interactions are observed either in the two fetal liver samples (b) or the two brain samples (c). (d,e) Running mean data for the same regions. False discovery rate was set at 5% (dashed line). (f) Schematic representation of regions of interaction with active (fetal liver, top) and inactive (fetal brain, bottom)  $\beta$ -globin on chromosome 7. Note that interacting regions are, on average, 150–200 kb and are not drawn to scale. Chromosomal positions were based on National Center for Biotechnology (NCBI) build m34.

A completely independent series of 4C experiments that focused on a fragment 50 kb downstream containing the  $\beta$ -globin-like gene *Hbb-b1* gave almost identical results (**Supplementary Fig. 4.1**).

A comparison between the two tissues showed that the actively transcribed  $\beta$ -globin locus in fetal liver interacts with a completely different set of loci on chromosome 7 from its transcriptionally silent counterpart in brain ( $\tau = -0.03$ ; Spearman's rank correlation). This excluded that results were influenced by the sequence composition of the probes. In fetal liver, the interacting DNA segments were located within a 70-Mb region centered around the  $\beta$ -globin locus, with the majority (40/66) located toward the telomere of the chromosome. In fetal brain, we found interacting loci at similar or even greater distances from  $\beta$ -globin than in fetal liver, with the great majority of interactions (43/45) located toward the centromere of chromosome 7 (**Fig. 4.2f**).



**Figure 4.3 Active and inactive  $\beta$ -globin interact with active and inactive chromosomal regions, respectively.**

(a) Comparison between  $\beta$ -globin long-range interactions in fetal liver (4C running mean, top), microarray expression analysis in fetal liver (log scale, middle) and the location of genes (bottom) plotted along a 4-Mb region that contains the gene *Uros* (~30 Mb away from  $\beta$ -globin), showing that active  $\beta$ -globin preferentially interacts with other actively transcribed genes. (b) The same comparison in fetal brain around an olfactory receptor gene cluster located ~38 Mb away from  $\beta$ -globin, showing that inactive  $\beta$ -globin preferentially interacts with inactive regions. Chromosomal positions were based on NCBI build m34. (c) Characterization of regions interacting with  $\beta$ -globin in fetal liver and brain in terms of gene content and activity.



Although the average size of interacting areas in fetal liver and brain was comparable (183 kb and 159 kb, respectively), we observed marked differences in their gene content and activity, the latter being determined by Affymetrix expression array analysis. In fetal liver, 80% of the  $\beta$ -globin interacting loci contained one or more actively transcribed genes, whereas in fetal brain, the great majority (87%) did not show any detectable gene activity (**Fig. 4.3**). Thus, the  $\beta$ -globin locus contacts different types of genomic regions in the two tissues (**Supplementary Table 4.1** online). Notably, 4C technology identified *Uros*, *Eraf* and *Kcnq1* (all ~30 Mb away from  $\beta$ -globin) as genes interacting with the active  $\beta$ -globin locus in fetal liver, in agreement with previous observations made by FISH<sup>8</sup> (**Supplementary Fig. 4.2**). Notably, in brain, we observed contacts with two other olfactory receptor gene clusters present on chromosome 7 that were located at each side of, and 17 and 37 Mb away from,  $\beta$ -globin.

*Uros* and *Eraf* and the genes encoding  $\beta$ -globin are all erythroid-specific genes that may be regulated by the same set of transcription factors, and it is an attractive idea that these factors coordinate the expression of their target genes in the nuclear space. We compared Affymetrix expression array data from E14.5 liver with that of brain to identify genes expressed preferentially (that is, showing more than fivefold greater expression) in fetal liver. Of the 560 active genes on chromosome 7, 15% were 'fetal liver-specific'; this was true for 13% of the 156 active genes within the colocalizing areas. More notably, 49 out of 66 (74%) interacting regions did not contain a 'fetal liver-specific' gene. Thus, we find no evidence for the intrachromosomal clustering of tissue-specific genes.

As the  $\beta$ -globin genes are transcribed at exceptionally high rates, we subsequently asked whether the locus preferentially interacts with other regions of high transcriptional activity. Using Affymetrix counts as a measure for gene activity, we performed a running sum algorithm to measure overall transcriptional activity within 200-kb regions around actively transcribed genes. This analysis showed that transcriptional activity around interacting genes was not higher than around noninteracting active genes on chromosome 7 ( $P = 0.9867$ ; Wilcoxon rank sum).

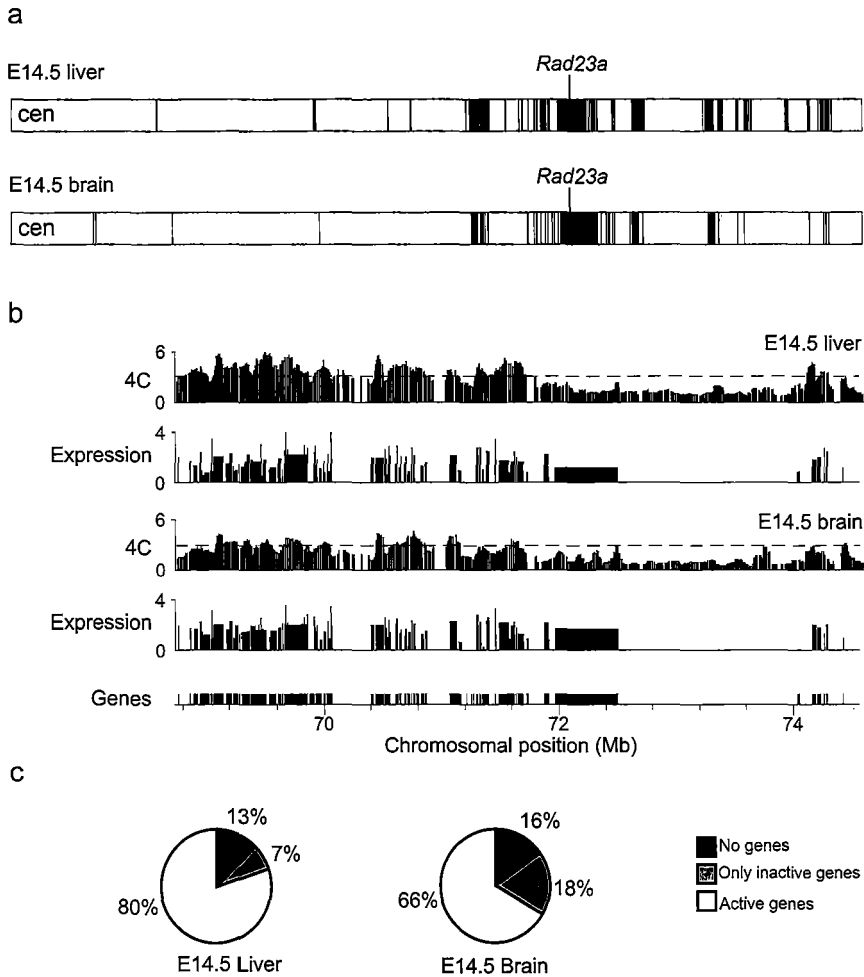
We next investigated whether a gene that is expressed similarly in both tissues also switches its genomic environment. *Rad23a* is a ubiquitously expressed DNA repair gene that resides in a gene-dense cluster of predominantly housekeeping genes on chromosome 8. Both in E14.5 liver and in brain, this gene and many of its direct neighbors are active. We performed 4C analysis and identified many long-range interactions with loci up to 70 Mb away from *Rad23a*. Notably, interactions with *Rad23a* were highly correlated between fetal liver and brain ( $\tau = 0.73$ ; Spearman's rank correlation) (**Fig. 4.4a**), providing

evidence for a general chromosomal folding pattern that is conserved between different cell types. Again, a shared hallmark of the interacting loci was that they contained actively transcribed genes. In both tissues, roughly 70% contained at least one active gene (**Fig. 4.4b,c**). Regions around interacting genes showed statistically significant higher levels of gene activity than for active genes elsewhere on the chromosome, as determined by a running sum algorithm ( $P < 0.001$  for both tissues). Thus, the *Rad23a* gene, which is located in a gene-rich region, preferentially interacts over a distance with other chromosomal regions of increased transcriptional activity.

To validate the results obtained by 4C technology, we performed cryo-FISH experiments. Cryo-FISH is a recently developed microscopy technique that has an advantage over current three-dimensional FISH in that it better preserves the nuclear ultrastructure while offering improved resolution in the z axis by the preparation of ultrathin cryosections<sup>10</sup>. Notably, 4C technology measures interaction frequencies rather than (average) distances between loci. Therefore, we verified 4C data by measuring how frequently  $\beta$ -globin or *Rad23a* alleles (always  $n > 250$ ) colocalized with selected chromosomal regions in 200-nm ultrathin sections prepared from E14.5 liver and brain. Notably, colocalization frequencies measured for loci positively identified by 4C technology were all significantly higher than frequencies measured for background loci ( $P < 0.05$ ; G test) (**Supplementary Table 4.2** online). For example, distant regions that we found to interact with  $\beta$ -globin by 4C technology colocalized more frequently than intervening areas not detected by 4C (7.4% and 9.7% versus 3.6% and 3.5%, respectively). Also, the two distant olfactory receptor gene clusters found (by 4C) to interact with  $\beta$ -globin in fetal brain but not liver scored colocalization frequencies of 12.9% and 7%, respectively, in brain, versus 3.6% and 1.9% in liver sections (**Fig. 4.5**). We concluded that 4C technology faithfully identified interacting DNA loci. Next, we used cryo-FISH to demonstrate that loci identified to interact with  $\beta$ -globin also frequently contacted each other. This was true for two active regions separated over a large chromosomal distance in fetal liver (**Supplementary Fig. 4.3**) as well as for two inactive olfactory receptor gene clusters far apart on the chromosome in brain (**Fig. 4.5**). Notably, we also found frequent contacts between these two distant olfactory receptor gene clusters in fetal liver, where they did not interact with the olfactory receptor gene cluster that contained the actively transcribed  $\beta$ -globin locus. These data provided further evidence for spatial interactions between distant olfactory receptor gene clusters<sup>14</sup>.

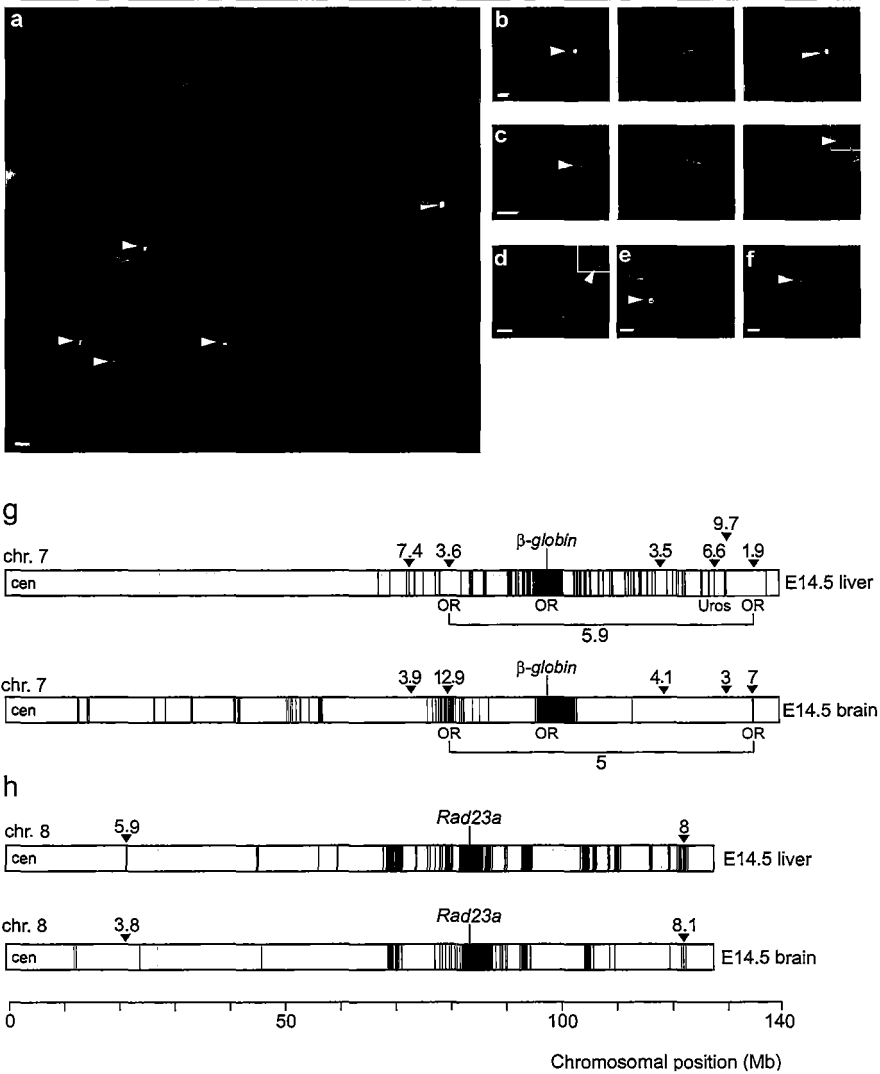
FISH analysis showed that the gene-dense chromosomal region containing *Rad23a* resides mostly at the edge of (82%) or outside (14%) the territory of chromosome 8 (D.

Noordermeer, M.R. Branco, A. Pombo and W.d.L., unpublished data) and we considered the possibility that *Rad23a* also interacted with regions on other chromosomes. Six unrelated chromosomes (7, 10, 11, 12, 13 and 14) were represented on our microarrays. Typically, these chromosomes showed very low 4C signals, with a few strong signals that



**Figure 4.4 Ubiquitously expressed *Rad23a* interacts with very similar active regions in fetal liver and brain.**

(a) Schematic representation of regions on chromosome 8 interacting with active *Rad23a* in fetal liver and brain. Note that interacting regions are on average 150–200 kb and are not drawn to scale. (b) Comparison between *Rad23a* long-range interactions (4C running mean) and microarray expression analysis (log scale) in fetal liver and fetal brain. Location of genes is plotted (bottom) along a 3 Mb region of chromosome 8. Chromosomal positions were based on NCBI build m34. (c) Characterization of regions interacting with *Rad23a* in fetal liver and brain in terms of gene content and activity.

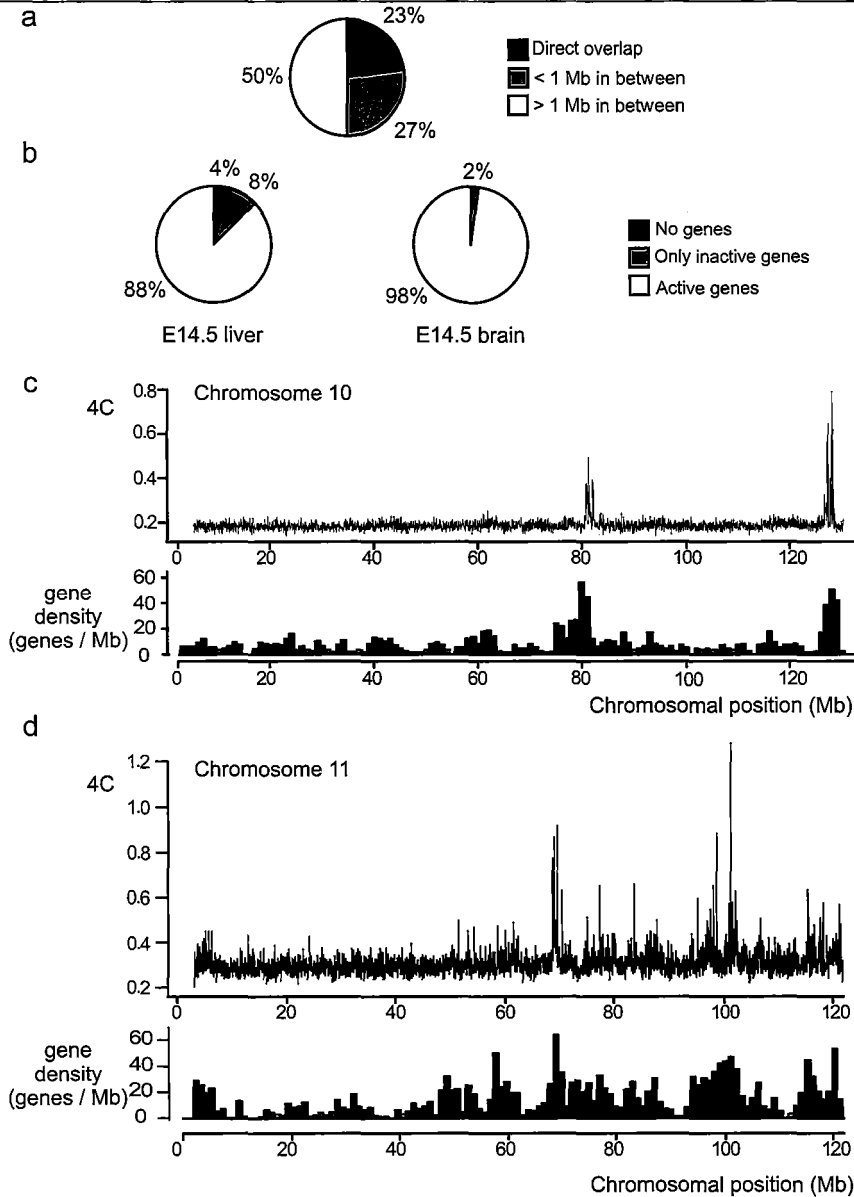


**Figure 4.5 Cryo-FISH confirms that 4C technology truly identifies interacting regions.** (a) example of part of a 200-nm cryosection showing more than ten nuclei, some of which contain the  $\beta$ -globin locus (white arrowhead) and/or *Uros* (grey arrowhead). Owing to sectioning, many nuclei do not show signals for these two loci. (b,c) Examples of completely (b) and partially (c) overlapping signals, which were all scored as positive for interaction. (d-f) Examples of nuclei containing non-contacting alleles (d,e) and a nucleus containing only  $\beta$ -globin (f), which were all scored as negative for interaction. Scale bars in a-f: 1  $\mu$ m. (g-h) Schematic representation of cryo-FISH results. Percentages of interaction with  $\beta$ -globin (g) and *Rad23a* (h) are indicated above the chromosomes for regions positively identified (black arrowhead) and negatively identified (grey arrowhead) by 4C technology. The same BACs were used for the two tissues. Interaction frequencies measured by cryo-FISH between two distant olfactory receptor gene clusters in fetal liver and brain are indicated below the chromosomes. Interacting regions are on average 150–200 kb and are not drawn to scale.

often appeared isolated on the linear DNA template. Indeed, when we analyzed each chromosome separately by applying running mean algorithms, we identified mostly regions that contained one such very strong hybridization signal. When tested by cryo-FISH, these regions scored negative for interaction (**Supplementary Table 4.3** online). To better identify clusters of possibly weaker, but positive, signals on the unrelated chromosomes, we applied a running median algorithm that ignores the isolated strong signals and scores only regions containing multiple positive signals. In fetal liver and brain, 24 and 44 of these interchromosomal regions were reproducibly identified in duplicate experiments (false discovery rate of 0%). We tested two of these regions by cryo-FISH, and both showed significant colocalization with *Rad23a* ( $P < 0.05$ ; **Supplementary Table 4.3**). We identified *trans*-interacting regions on all six chromosomes analyzed by 4C (**Supplementary Table 4.4** online). A comparison between fetal liver and brain showed that the *trans*-interacting regions were often located at the same parts of the chromosomes, suggesting that the interchromosomal environment of the gene-dense region containing *Rad23a* is conserved to some extent between the tissues (**Fig. 4.6a**). Similar to what we had observed for the intrachromosomal contacts, interchromosomal interactions with *Rad23a* seemed biased toward regions with increased gene density (**Fig. 4.6b–d** and **Supplementary Fig. 4.4**). These data suggested that active gene-dense regions preferentially contact each other in *cis* and in *trans*.

The  $\beta$ -globin locus has been shown to preferentially locate inside its own chromosome territory<sup>15</sup>. In agreement with this, we found only four interacting regions in *trans* in each tissue (data not shown). In fetal liver, all four regions contained actively transcribed genes, whereas in fetal brain, where the locus is inactive, three out of the four regions did not contain an active gene. Cryo-FISH on two loci in fetal liver, however, did not show colocalization frequencies significantly above background (data not shown). Thus, either the method detects some false positive interchromosomal interactions or is so sensitive that it identifies interactions too infrequent to be discerned from background under the microscope.

Our observations demonstrate that not only active, but also inactive, genomic regions can transiently interact over large distances with many loci in the nuclear space. The data strongly suggest that each DNA segment has its own preferred set of interactions. This implies that it is impossible to predict the long-range interaction partners of a given DNA locus without knowing the characteristics of its neighboring segments and, by extrapolation, the whole chromosome. The fact that a housekeeping gene shows interchromosomal contacts provides further evidence for extensive chromosome



**Figure 4.6 Interchromosomal interactions with *Rad23a*.** (a) Interchromosomal interactions with *Rad23a* are conserved between tissues. Indicated are the percentages of trans-interacting regions observed in fetal liver that are identical, close to (that is, <1 Mb) or unrelated to the interchromosomal interactions detected in fetal brain. (b) The great majority of regions contacting *Rad23a* in trans contain active genes, both in fetal liver (left) and brain (right). (c,d) High running median values of the 4C data (top) are found at chromosomal locations with a high gene density (bottom). Chromosomes 10 (c) and 11 (d) from a fetal brain sample are shown as examples. Data from fetal liver have a similar profile. Chromosomal positions were based on NCBI build m34.

intermingling<sup>10</sup> and shows that individual loci have preferred neighbors in the nuclear space. We propose that the extensive network of long-range interactions that we have identified both between inactive and between active genomic loci reflects cell-to-cell differences in chromosome conformations in combination with dynamic movements during interphase.

## **Methods**

### *4C technology*

The initial steps of the 3C technology procedure were performed as described previously<sup>16</sup>, yielding ligation products between HindIII fragments. This HindIII-ligated 3C template (~50 µg) was digested overnight at a concentration of 100 ng/µl with 50 units (U) of a secondary, frequently cutting restriction enzyme (either DpnII (used for HS2 and Rad23a) or NlaIII (used for Hbb-b1)). Other combinations of restriction enzymes will also work, provided that the secondary, but not the primary, restriction enzyme is a frequent DNA cutter. To avoid constraints in DNA circle formation<sup>17</sup>, we chose a secondary restriction enzyme that did not cut within 350–400 bp of the HindIII restriction site that demarcates the restriction fragment of interest (that is, the 'bait'). After secondary restriction enzyme digestion, DNA was phenol extracted, ethanol precipitated and subsequently ligated at low concentration (50 µg sample in 14 ml using 200 U ligase (Roche)); incubated for 4 h at 16 °C) to promote DpnII circle formation. Ligation products were phenol extracted and ethanol precipitated using glycogen (Roche) as a carrier (20 µg/ml). We linearized the circles of interest by digesting overnight with 50 U of a tertiary restriction enzyme that cuts the bait between the primary and secondary restriction enzyme recognition sites, using the following restriction enzymes: SpeI (for HS2), PstI (for Rad23a) and Pflml (for Hbb-b1). This linearization step was performed to facilitate subsequent primer hybridization during the first rounds of PCR amplification. Digested products were purified using a QIAquick nucleotide removal (250) column (Qiagen).

PCR reactions were performed using the Expand Long Template PCR system (Roche) using conditions carefully optimized to assure linear amplification of fragments sized up to 1.2 kb (80% of 4C PCR fragments are <600 bp). PCR conditions were as follows: 94 °C for 2 min; 30 cycles of 94 °C for 15 s, 55 °C for 1 min and 68 °C for 3 min; followed by a final step of 68 °C for 7 min. We also carefully determined the maximum amount of template that still showed a linear range of amplification. For this, serial dilutions of template were

added to PCR reactions; amplified DNA material was separated on an agarose gel and PCR products were quantified using ImageQuant software. Typically, 100–200 ng of template per 50  $\mu$ l PCR reaction gave products in the linear range of amplification. We pooled 16 to 32 PCR reactions and purified this 4C template using the QIAquick nucleotide removal (250) system (Qiagen). Primer sequences are listed in **Supplementary Table 4.5** online.

Purified 4C template was labeled and hybridized to arrays according to standard chromatin immunoprecipitation (ChIP)-chip protocols (Nimblegen Systems). Differentially labeled genomic DNA, digested with the same primary and secondary enzyme, served as a control template to correct for differences in hybridization efficiencies. For each experiment, two independently processed samples were labeled with alternate dye orientations. Data were highly reproducible between independent experiments ( $t > 0.99$ ; Spearman's rank correlation). Under the conditions described, sequences nearby on the chromosome template were together with the 'bait' so frequently that their large overrepresentation in our hybridization samples saturated the corresponding probes. This was confirmed when we performed hybridizations with samples diluted 1:10 and 1:100 and found that signal intensity was reduced at probes outside and at the edge of, but not inside, this region (data not shown).

#### *4C arrays*

An important aspect of our strategy is the use of dedicated microarrays. Recently, others have used techniques based on the 3C method to screen for colocalizing sequences<sup>18,19</sup>. However, these approaches were based on the sequencing of a limited number of captured fragments and, hence, did not provide a comprehensive overview of long-range interacting DNA segments. For our arrays, we designed probes (60-mers) that each represent a different *HindIII* restriction end in the genome. The advantages of such a design are that (i) each probe is informative, as each analyzes an independent ligation event, greatly facilitating the interpretation of the results, and (ii) a large representation of the genome can be spotted on a single array (which is cost effective). The array used for this study (Nimblegen Systems) covered seven complete mouse chromosomes. Arrays and analyses were based on NCBI build m34. Probes (60-mers) were selected from the sequences 100 bp up- and downstream of *HindIII* sites. To prevent cross-hybridization, probes that had any similarity with highly abundant repeats (as judged by RepBase 10.09)<sup>20</sup> were removed from the probe set. In addition, probes that gave more than two BLAST hits in the genome were also removed from the probe set. Sequence alignments



were performed using MegaBLAST<sup>21</sup> using the standard settings. A hit was defined as an alignment of 30 nt or longer.

#### *4C data analysis*

The ratio of sample-to-genomic DNA 4C signal was calculated for each probe, and the data were visualized with SignalMap software provided by Nimblegen Systems. Data were analyzed using the R package, Spotfire and Excel. Unprocessed hybridization ratios showed clusters of 20–50 strong signals along the chromosome template. It is important to realize that each probe on the array analyzes an independent ligation event. Moreover, only two copies of a given restriction fragment are present per cell, and each can ligate only to one other restriction fragment. Therefore, the detection of independent ligation events with 20 or more neighboring restriction fragments strongly indicates that the corresponding locus contacts the 4C bait in multiple cells. To define the clusters *in cis*, we applied a running mean or running median. We used various window sizes, ranging from 9–39 probes, which all identified the same clusters. Results shown are based on a window size of 29 probes (on average 60 kb) and were compared with the running mean performed across randomized data. This was done for each array separately. Consequently, all measurements were judged relative to the amplitude and noise of that specific array. The false discovery rate (FDR), defined as  $(\text{number of false positives}) / (\text{number of false positives} + \text{number of true positives})$ , was determined as follows:  $(\text{number of positives in the randomized set}) / (\text{number of positives in the data})$ . The threshold level was determined using a top-down approach to establish the minimal value for which  $\text{FDR} < 0.05$ .

Next, biological duplicate experiments were compared. Windows that met the threshold in both duplicates were considered positive. When comparing randomized data, no windows were above threshold in both duplicates. Positive windows directly adjacent on the chromosome template were joined (no gaps allowed), creating positive areas.

In defining interacting regions *in trans*, we took a similar approach. However, there we applied a running median, again with a window size of 29 probes. The threshold was set at an FDR of 0%. Thus, a region was called interacting when, in both duplicates, the median signal was higher than any signal found in the respective randomized data sets.

#### *Expression analysis*

For each tissue, three independent microarrays were performed according to Affymetrix protocol (mouse 430\_2 arrays). Data were normalized using Bioconductor RMA ca-tools,

and for each probe set, the measurements of the three microarrays were averaged. In addition, when multiple probe sets represented the same gene, they were also averaged. Mas5calls was used to establish 'present', 'absent' and 'marginal' calls. Genes with a 'present' call in all three arrays and an expression value  $>50$  were called expressed. 'Fetal liver-specific genes' were classified as genes that met our criteria of being expressed in fetal liver and having over fivefold higher expression than in fetal brain. To provide a measure of overall transcriptional activity around each gene, a running sum was applied. For this, we used log-transformed expression values. For each gene, we calculated the sum of the expression of all genes found in a window 100 kb upstream of the start and 100 kb downstream of the end of the gene, including the gene itself. We compared the resulting values for active genes found inside 4C-positive regions ( $n = 124, 123$  and  $208$ , respectively, for HS2 in liver, *Rad23a* in brain and *Rad23a* in liver) with the values obtained for active genes outside 4C-positive areas ( $n = 153, 301$  and  $186$ , respectively, where  $n = 153$  corresponds to the number of active, noninteracting genes present between the most centromeric interacting region and the telomere of chromosome 7; we compared the two groups using a one-tailed Wilcoxon rank sum test.

#### *FISH probes*

The following BAC clones (BACPAC Resources Centre) were used: RP23-370E12 for *Hbb-b1*, RP23-317H16 for chromosome 7 at 80.1 Mb (olfactory receptor gene cluster), RP23-334E9 for *Uros*, RP23-32C19 for chromosome 7 at 118.3 Mb, RP23-143F10 for chromosome 7 at 130.1 Mb, RP23-470N5 for chromosome 7 at 73.1 Mb, RP23-247L11 for chromosome 7 at 135.0 Mb (olfactory receptor gene cluster), RP24-136A15 for *Rad23a*, RP23-307P24 for chromosome 8 at 21.8 Mb, RP23-460F21 for chromosome 8 at 122.4 Mb, RP24-130O14 for chromosome 10 at 74.3 Mb, RP23-153N12 for chromosome 11 at 68.7 Mb, RP23-311P1 for chromosome 11 at 102.2 Mb, RP24-331N11 for chromosome 14 at 65.1 Mb and RP23-236O12 for chromosome 14 at 73.7 Mb.

Random primer-labeled probes were prepared using BioPrime Array CGH Genomic Labeling System (Invitrogen). Before labeling, DNA was digested with *DpnII* and purified with a DNA Clean-up and Concentrator 5 Kit (Zymo Research). Digested DNA (300 ng) was labeled with SpectrumGreen dUTP (Vysis) or Alexa Fluor 594 dUTP (Molecular Probes) and purified through a GFX PCR DNA and Gel Band Purification kit (Amersham Biosciences) to remove unincorporated nucleotides. The specificity of labeled probes was tested on metaphase spreads prepared from mouse embryonic stem (ES) cells.

### Cryo-FISH

Cryo-FISH was performed as described before<sup>10</sup>. Briefly, E14.5 liver and brain were fixed for 20 min in 4% paraformaldehyde (vol/vol)/250 mM HEPES (pH 7.5) and were cut into small tissue blocks, followed by another fixation step of 2 h in 8% paraformaldehyde at 4 °C. Fixed tissue blocks were immersed in 2.3 M sucrose for 20 min at room temperature (18–22 °C), mounted on a specimen holder and snap-frozen in liquid nitrogen. Tissue blocks were stored in liquid nitrogen until sectioning. Ultrathin cryosections of approximately 200 nm were cut using a Reichert Ultramicrotome E equipped with a cryo-attachment (Leica). Using a loop filled with sucrose, sections were transferred to coverslips and stored at –20 °C. For hybridization, sections were washed with PBS to remove sucrose, treated with 250 ng/ml RNase in 2× SSC for 1 h at 37 °C, incubated for 10 min in 0.1 M HCl, dehydrated in a series of ethanol washes and denatured for 8 min at 80 °C in 70% formamide/2× SSC, pH 7.5. Sections were again dehydrated directly before probe hybridization. We coprecipitated 500 ng labeled probe with 5 µg of mouse Cot1 DNA (Invitrogen) and dissolved it in hybridization mix (50% formamide, 10% dextran sulfate, 2× SSC, 50 mM phosphate buffer, pH 7.5). Probes were denatured for 5 min at 95 °C, reannealed for 30 min at 37 °C and hybridized for at least 40 h at 37 °C. After posthybridization washes, nuclei were counterstained with 20 ng/ml DAPI (Sigma) in PBS/0.05% Tween-20 and mounted in Prolong Gold antifade reagent (Molecular Probes).

Images were collected with a Zeiss Axio Imager Z1 epifluorescence microscope (100× plan apochromat, 1.4× oil objective), equipped with a charge-coupled device (CCD) camera and Isis FISH Imaging System software (Metasystems). A minimum of 250  $\beta$ -globin or *Rad23a* alleles were analyzed and scored (by a person not knowing the probe combination applied to the sections) as overlapping or nonoverlapping with BACs located elsewhere in the genome. Replicated goodness-of-fit tests (*G* statistic)<sup>22</sup> were performed to assess significance of differences between values measured for 4C-positive versus 4C-negative regions. An overview of the results is provided in **Supplementary Table 4.3**. The frequencies measured by cryo-FISH were considerably lower than those reported by others based on two-dimensional and three-dimensional FISH<sup>8,9</sup>. Although cryo-FISH may slightly underestimate actual interaction frequencies owing to sectioning, we expect that its increased resolution will provide more accurate measurements.

## URLs

The R package can be downloaded from <http://www.r-project.org>. RMA ca-tools and Mas5calls for the analysis of Affymetrix microarray expression data can be found at <http://www.bioconductor.org>. The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) and are accessible through GEO Series accession number GSE5891.

*Note: Supplementary information is available on the Nature Genetics website.*

## Acknowledgements

We thank F. Grosveld for support and discussion and S. van Baal, M. Branco, A. Pombo, P. Verrijzer, J. Hou, B. Eussen, A. de Klein, T. de Vries Lentsch, D. Noordermeer and R.-J. Palstra for assistance.

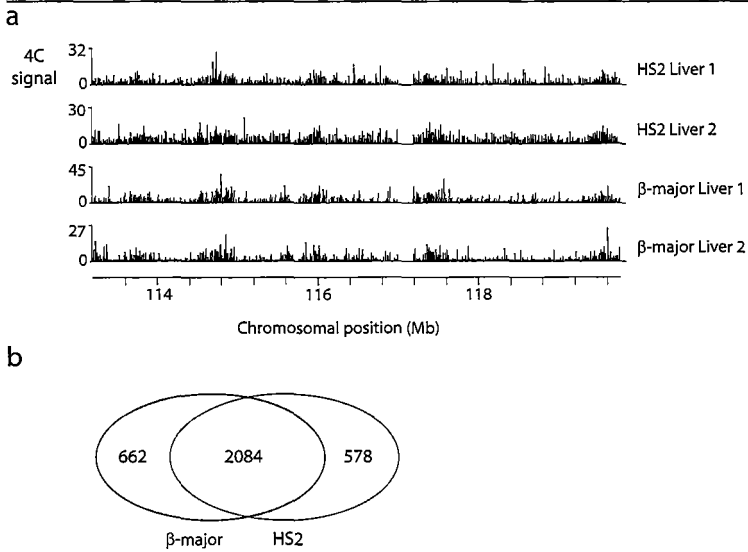
## References

1. Misteli, T. Concepts in nuclear architecture. *Bioessays* **27**, 477–487 (2005).
2. Sproul, D., Gilbert, N. & Bickmore, W.A. The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.* **6**, 775–781 (2005).
3. Chakalova, L., Debrand, E., Mitchell, J.A., Osborne, C.S. & Fraser, P. Replication and transcription: shaping the landscape of the genome. *Nat. Rev. Genet.* **6**, 669–677 (2005).
4. Volpi, E.V. *et al.* Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *J. Cell Sci.* **113**, 1565–1576 (2000).
5. Chambeyron, S. & Bickmore, W.A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev.* **18**, 1119–1130 (2004).
6. Brown, K.E. *et al.* Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. *Cell* **91**, 845–854 (1997).
7. Grogan, J.L. *et al.* Early transcription and silencing of cytokine genes underlie polarization of T helper cell subsets. *Immunity* **14**, 205–215 (2001).
8. Osborne, C.S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36**, 1065–1071 (2004).
9. Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R. & Flavell, R.A. Interchromosomal associations between alternatively expressed loci. *Nature* **435**, 637–645 (2005).

10. Branco, M.R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* **4**, e138 (2006).
11. Roix, J.J., McQueen, P.G., Munson, P.J., Parada, L.A. & Misteli, T. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat. Genet.* **34**, 287–291 (2003).
12. Lemaitre, J.M., Danis, E., Pasero, P., Vassetzky, Y. & Mechali, M. Mitotic remodeling of the replicon and chromosome structure. *Cell* **123**, 787–801 (2005).
13. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
14. Lomvardas, S. *et al.* Interchromosomal interactions and olfactory receptor choice. *Cell* **126**, 403–413 (2006).
15. Brown, J.M. *et al.* Coregulated human globin genes are frequently in spatial proximity when active. *J. Cell Biol.* **172**, 177–187 (2006).
16. Splinter, E., Grosveld, F. & de Laat, W. 3C technology: analyzing the spatial organization of genomic loci in vivo. *Methods Enzymol.* **375**, 493–507 (2004).
17. Rippe, K., von Hippel, P.H. & Langowski, J. Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem. Sci.* **20**, 500–506 (1995).
18. Ling, J.Q. *et al.* CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science* **312**, 207–208 (2006).
19. Wurtele, H. & Chartrand, P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended chromosome conformation capture methodology. *Chromosome Res.* **14**, 477–495 (2006).
20. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
21. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).
22. Sokal, R.R. & Rohlf, F.J. *Biometry: the Principles and Practice of Statistics in Biological Research* 3rd edn. (W.H. Freeman, New York, 1995).

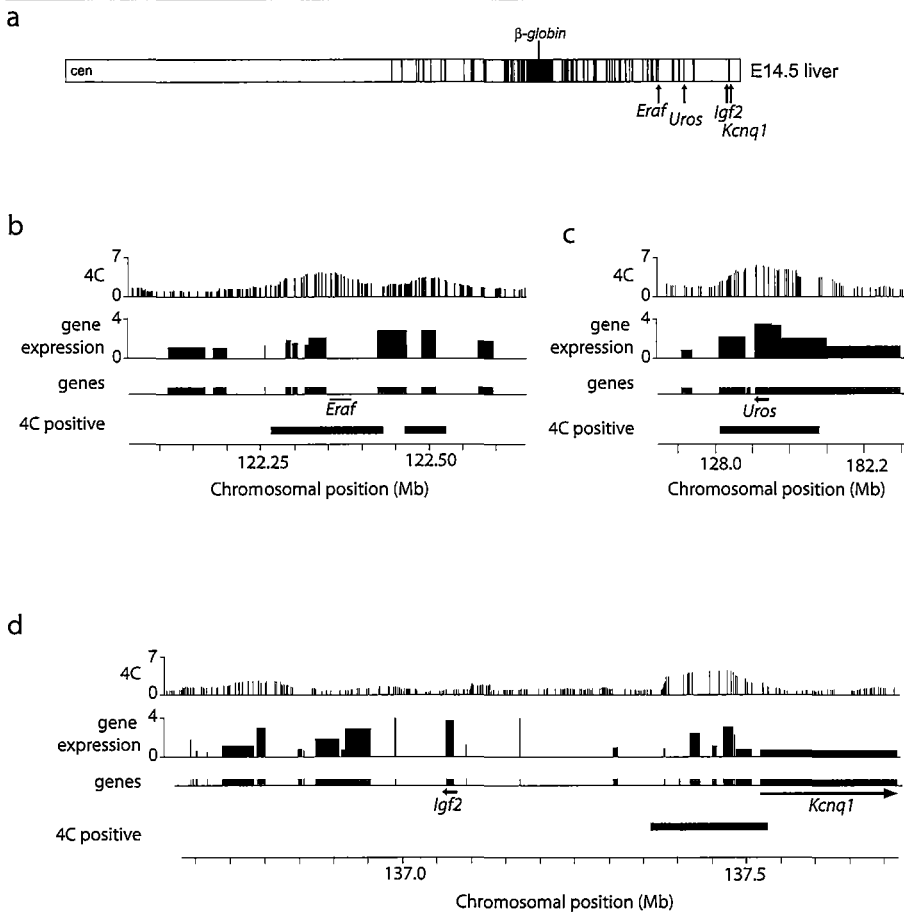
Supplementary figures

Supplementary figure 4.1



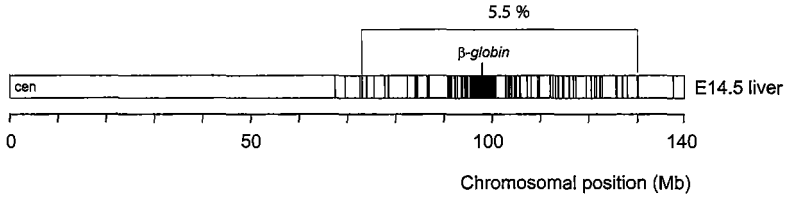
**Supplementary Figure 4.1. 4C analysis of HS2 and *Hbb-b1* give highly similar results.** (a) Unprocessed 4C data of four independent E14.5 liver samples show a very similar pattern of interaction with HS2 (top) and *Hbb-b1* (bottom). Positions are based on NCBI build m34. (b) A large overlap exists between probes scored positive for interaction in the HS2 experiment and probes that scored positive for interaction in the *Hbb-b1* experiment.

Supplementary figure 4.2



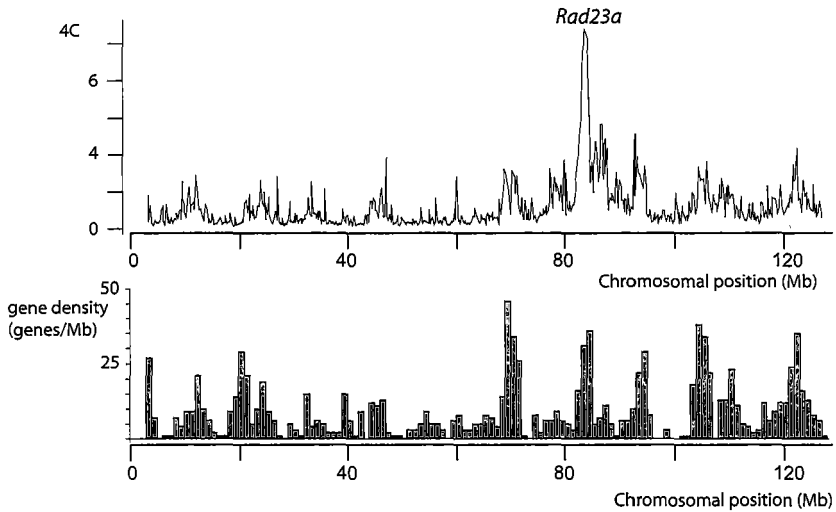
**Supplementary Figure 4.2. 4C detects interactions previously identified by FISH.** (a) The position of the four genes previously shown by Osborne et al. [12] to interact with  $\beta$ -globin, relative to interacting areas found by 4C in fetal liver (grey bars). (b-d) 4C (running mean, top), microarray expression data (log scale, black), location of genes (grey) and location of areas that were significant in 4C analysis (bottom horizontal bars). The position and transcriptional direction of the four genes is indicated (black arrows). Interactions with *Eraf*, *Uros* and *Kcnq1* are found with 4C. *Igf2* is located very close to a 4C positive area. Positions are based on NCBI build m34. The location of *Eraf* was not found directly in the Ensembl database, but blasting the sequence of the gene resulted in the indicated position.

Supplementary figure 4.3



**Supplementary Figure 4.3. Regions that interact with β-globin also frequently contact each other.** Two regions on chromosome 7 (almost 60 Mb apart), containing actively transcribed genes and identified by 4C technology to interact with β-globin in fetal liver, showed co-localization frequencies by cryo-FISH of 5.5%, which was significantly more than background co-localization frequencies.

Supplementary figure 4.4



**Supplementary Figure 4.4. *Rad23a* co-localizes with gene-dense regions in cis.** High 4C signals (running mean, top) are found on chromosomal locations of high gene density (bottom). A fetal brain sample is shown as example, 4C data of fetal liver showed a similar profile. Chromosomal positions are based on NCBI build m34.



# 5

## An evaluation of 3C-based methods to capture DNA interactions

## An evaluation of 3C-based methods to capture DNA interactions

Marieke Simonis#, Jurgen Kooren# and Wouter de Laat\*

Department of Cell Biology and Genetics, Erasmus MC, Rotterdam, The Netherlands

# These authors contributed equally

\* To whom correspondence should be sent

### Summary

*The shape of the genome is thought to play an important role in the coordination of transcription and other DNA-metabolic processes. Chromosome conformation capture (3C) technology allows analyzing the folding of chromatin in the native cellular state at a resolution beyond that provided by current microscopy techniques. It has been used for example to demonstrate that regulatory DNA elements communicate with distant target genes via direct physical interactions that loop out the intervening chromatin fiber. Here, we will discuss the intricacies of 3C and novel 3C-based methods like 4C, 5C and the ChIP-loop assay.*

3C technology was originally developed to study the conformation of a complete chromosome in yeast<sup>1</sup> and was subsequently adapted to investigate the folding of complex gene loci in mammalian cells<sup>2</sup>. It has now become a standard research tool for studying the relationship between nuclear organization and transcription in the native cellular state. Other technologies based on the 3C principle have been developed that aim to increase the throughput. 4C technology allows for an unbiased genome-wide screen for interactions with a locus of choice, while 5C technology enables parallel analysis of interactions between many selected DNA fragments. ChIP-loop combines 3C with chromatin immuno-precipitation to analyze interactions between specific protein-bound DNA sequences. Detailed protocols that should help researchers setting up 3C<sup>3-5</sup> and 5C<sup>6</sup> technology in their own laboratory and an excellent review<sup>7</sup> explaining the controls necessary for correct interpretation of 3C results have been published. Here we present a detailed 4C procedure as Supplementary Protocol.

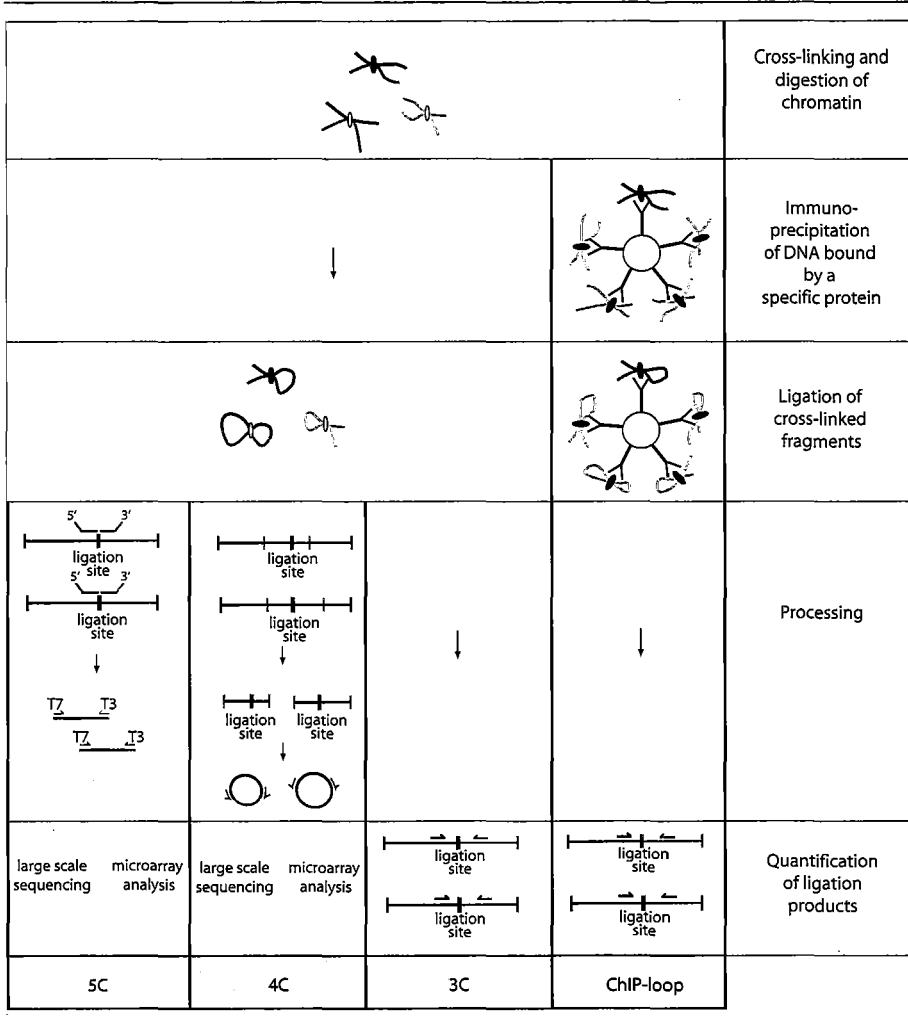
### **Common principles**

In short, the 3C procedure involves five experimental steps (**Fig. 5.1**). First, cells are fixed with formaldehyde, which cross-links proteins to other proteins and to DNA segments that are in close proximity in the nuclear space. Second, the cross-linked chromatin is digested with an excess of restriction enzyme, separating cross-linked from non-cross-linked DNA fragments. Third, DNA ends are ligated under conditions that favor junctions between cross-linked DNA fragments. In a fourth step, cross-links are reversed. And finally, ligation events between selected pairs of restriction fragments are quantified by PCR, using primers specific for the fragments being studied.

The technique enables the identification of physical interactions between distant DNA segments and of chromatin loops that are formed as a consequence of these interactions, for example between transcriptional regulatory elements and distant target genes<sup>2,8-11</sup>. 3C technology is particularly suited to study the conformation of genomic regions that are roughly between five to several hundreds of kilobases (kb) in size. To our knowledge, the smallest region studied so far by 3C technology spans 6700 basepairs (bp)<sup>12</sup>, while the largest region analyzed spans ~600 kb<sup>13</sup>. It is important to note that due to flexibility of the chromatin fiber, DNA segments on the same fiber are engaged in random collisions, with a frequency inversely proportional to the genomic distance between them. Therefore, the mere detection of a ligation product does not necessarily reveal a specific interaction. To ascertain that an interaction is specific requires the demonstration that two DNA sites interact more frequently with each other than with neighboring DNA sequences. Thus, 3C technology is a quantitative assay and a meaningful analysis critically relies on an accurate comparison of interaction frequencies between multiple DNA segments.

3C and 3C-based technologies provide information about the frequency, but not the functionality, of DNA interactions. Thus, additional, often genetic, experiments are required to address whether an interaction identified by 3C-based technologies is functionally meaningful. For example, many of the interactions identified by 4C technology<sup>14</sup> between genomic regions far apart on the same chromosome or on other chromosomes may well be non-functional and merely the consequence of general folding patterns of chromosomes<sup>15</sup>.

During most of the cell cycle one mammalian cell provides maximally two events for 3C analysis, as it contains only two copies of a given restriction fragment, each end of which can be ligated to maximally one other restriction site during the 3C procedure. This implies that a meaningful (i.e. quantitative) 3C PCR analysis must be performed on a DNA



**Figure 5.1. Schematic representation of 3C based methods.** In all 3C-based methods DNA interactions are captured by formaldehyde treatment followed by DNA digestion with a restriction enzyme. Cross-linked fragments are ligated to each other and ligation frequencies are measured. In the ChIP-loop assay, immunoprecipitation enriches the sample for fragments bound by a specific protein and restriction fragments are ligated to each other on the beads. In ChIP-loop and 3C, ligation frequencies are measured by quantitative PCR, using a unique primers set for each ligation junction analyzed. In 5C, oligonucleotides are annealed and ligated in a multiplex setting, and contain either a 5' T7 primer extension or a 3' T3 primer extension allowing massive parallel PCR amplification of different ligation events, which are analyzed by large-scale sequencing or microarray analysis. In 4C, ligation junctions are first trimmed by a frequently cutting secondary restriction enzyme, followed by ligation to form circles and inverse PCR to amplify captured fragments. If a frequent cutting enzyme is used in the first digestion the second digestion can be omitted (see Fig. 3). The 4C PCR product is analyzed by large-scale sequencing or microarray analysis.

template that represents many genome equivalents. It also implies that DNA interactions can only be quantified accurately if they occur in a substantial proportion of the cells. Sites separated over large genomic distances (i.e. hundreds of kilobases or more) or present on other chromosomes often form not enough ligation products for accurate quantification, even if microscopy studies suggest that they come together in a substantial proportion of cells. To study such long-range interactions we recommend using high-throughput 4C technology.

Below, a more detailed outline of the experimental steps involved in all 3C-based technologies will be presented in order to allow a better appreciation of the potentials and limitations of these methods.

### **Common experimental steps in 3C-based technologies**

#### *Step 1: formaldehyde cross-linking*

The method uses formaldehyde to crosslink protein-protein and protein-DNA interactions via their amino and imino groups. Advantage of this cross-linking agent is that it works over a relatively short distance (2 Å) and that cross-links can be reversed at higher temperatures<sup>16-18</sup>. Although cross-linking is sometimes performed on isolated nuclei, it is preferentially done on living cells, since this better guarantees taking a faithful snapshot of the chromatin conformation. Routinely, cells are cross-linked at room temperature for ten minutes, using a formaldehyde concentration of 1-2%, but optimal fixation conditions depend on the frequency and stability of the interactions analyzed and need to be redefined for every new 3C experiment. Many 3C experiments demonstrate preferential interactions between transcription regulatory DNA elements. These sites are known to carry transcription factors and often contain less histone proteins, hence their hypersensitivity to nuclease digestion. A concern often raised is that the 3C assay may be biased due to better cross-link ability of these sites. However, evidence that the contrary may be true comes from recently developed FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements)<sup>19,20</sup>. FAIRE involves phenol-chloroform extraction of formaldehyde cross-linked and sonicated chromatin and isolates regulatory DNA sequences based on the fact that they tend to end up in the aqueous phase more than other genomic regions (hence they are less cross-linkable to proteins).

Formaldehyde is also used under similar experimental conditions in chromatin immunoprecipitation (ChIP) experiments as the cross-linking agent that captures protein-

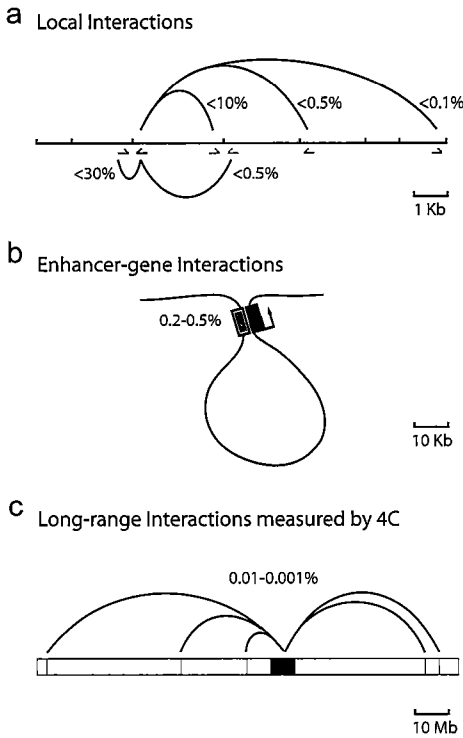
DNA interactions. It is conceivable that formaldehyde often produces complex aggregates containing more than two DNA fragments. In support of this, it was found that a single restriction fragment frequently captures two or more other restriction fragments together in a 4C experiment<sup>21</sup>. This notion would imply that both ChIP and 3C-like technologies also pick up indirect interactions.

*Step 2: Restriction enzyme digestion*

After cross-linking, nuclei are isolated and digested with a restriction enzyme. The choice of restriction enzyme will mainly depend on the locus to be analyzed. The restriction enzyme should dissect the locus such that it allows for the separate analysis of the relevant regulatory elements (gene bodies, promoters, enhancers, insulators, etc.). Analyzing the topology of small loci (< 10-20 kb) requires the use of frequently cutting restriction enzymes such as DpnII or NlaIII (4-base cutters). When analyzing larger loci, 6-base cutters can also be used. Not all enzymes digest cross-linked DNA equally well and we prefer to use EcoRI, BglII or HindIII<sup>3</sup>. When we digest overnight with a large excess of one of these restriction enzymes, we do not observe notable preferential digestion for specific regions of the genome such as open chromatin. This may be different with different enzymes and conditions though and we recommend for each new 3C experiment to exclude that there is a bias in the assay due to preferential digestion of some sites over others. Digestion efficiency also decreases with increasing cross-linking stringency<sup>3</sup>. We recommend that at least 60-70% of the DNA, but preferably 80% or more, is digested before continuing with the ligation step.

*Step 3: Ligation*

A critical selective step in the procedure is the ligation step carried out under conditions that favor intra-molecular ligation events between cross-linked DNA fragments. This step creates the actual 3C library that is enriched for ligated junctions between DNA fragments that originally were close together in the nuclear space. It is relevant to know how frequently a given ligation occurs. We have carefully quantified the abundance of the most frequently formed ligation products (**Fig. 5.2**). Independently of the restriction site analyzed, two types of junction are always over-represented. The first most abundant junction is with the neighboring DNA sequence. This junction is the result of incomplete restriction enzyme digestion and can constitute up to 20-30% of all the junctions; this number drops when less stringent fixation conditions are used. The second most abundant junction is with the other end of the same restriction fragment, as a consequence of



**Figure 5.2. Ligation events measured at the  $\beta$ -globin locus.** Schematic presentation of the relative abundance of frequently formed ligation products at the mouse  $\beta$ -globin locus. Typical values for ligation frequencies (in % alleles) of a 'bait' restriction fragment end with a given other restriction fragment end are indicated for (a) Local interactions with directly neighboring restriction fragments; data measured using 3C-qPCR. Arrows below the restriction fragments indicate the location and direction of 3C primers ('bait' primer indicated in dark grey) (b) Enhancer-gene interactions over 30-100 Kb; data measured using 3C-qPCR (c) Long-range interactions in cis and in trans ( $> 1$  Mb); frequencies estimated from 4C data.

restriction fragment circularization. This product can be formed independent of the cross-linking step and can account for up to 5-10% of all the junctions formed. Interestingly, this percentage goes up when less stringent cross-linking conditions are used (data not shown), suggesting that under such

conditions less restriction fragments are cross-linked together. The formation of other junctions is much less efficient. For example, ligation to ends of directly neighboring restriction fragments (which will always be close together in the nuclear space and therefore should also ligate relatively efficiently) already only occurs 0.2-0.5% of the time. This percentage quickly drops down to  $< 0.1\%$  with increasing genomic site separation, unless two sites are engaged in a specific interaction. However, even sites thought to frequently interact with each other, such as sites in the  $\beta$ -globin locus control region and the active  $\beta$ -globin genes 30-50 kb away, only account for 0.2-0.5% of the junctions formed between them. It should be clear that in order to accurately quantify such rare events that often occur in less than 1/1000 cells, many genome equivalents need to be included in a PCR reaction.

#### PCR in 3C

After reversal of the cross-links, ligation frequencies of restriction fragments are analyzed by PCR, using primers specific for the restriction fragments of interest. We

routinely use 50-200 ng of 3C template, or  $\sim 8 \times 10^4 - 3 \times 10^5$  genome equivalents, per PCR reaction. A meaningful 3C analysis critically relies on the accurate quantification of the different ligation products and measurements therefore need to be taken when each DNA amplification reaction is in the linear range. The standard 3C PCR protocol uses a standard number of PCR cycles and a standard amount of DNA template for the analysis of all different ligation products. This approach is only semi-quantitative and prone to inaccuracies since measurements may be taken outside the linear range of the amplification reaction. To overcome this limitation, a real-time PCR approach using TaqMan® probes, called 3C-qPCR, was developed<sup>22,23</sup>. A single probe and fixed PCR primer are used that hybridize to opposite strands of the restriction fragment of interest and work in combination with a series of test PCR primers hybridizing to other restriction fragments. This configuration ensures that the fluorescent signal provided by the probe is strictly specific to the amplification of the ligation product selected for analysis<sup>5</sup>.

Different primer pairs will have different amplification efficiencies and in order to account for this, these efficiencies need to be assessed. This is done on a control template containing all ligation products in equimolar amounts<sup>1-3,7,8</sup>, mixed with the same amount of genomic DNA as is present in the 3C PCR reaction. To account for possible differences in quality and quantity between 3C templates, interaction frequencies are analyzed between segments in a control locus that is expected to adopt a similar conformation in the different cell-types of interest<sup>2,3,7</sup>.

### **The ChIP-loop assay**

It is often found by 3C technology that, in the population of cells analyzed, a single DNA site interacts with multiple other sites. In many cases, this is likely to reflect cell-to-cell differences in chromatin conformation and it is very well possible that in different subpopulations of cells distinct proteins bind to such a given site and mediate the different DNA interactions. The chromatin immunoprecipitation-combined loop (ChIP-loop) assay was developed to investigate this<sup>24-26</sup>. The method involves formaldehyde cross-linking of cells, restriction enzyme digestion and urea gradient purification of cross-linked chromatin, immunoprecipitation using an antibody against the protein of interest, ligation of precipitated DNA fragments (still coupled to the beads) and PCR analysis of the junctions (**Fig. 5.1**).

In our opinion, a number of technical aspects complicate the analysis of results obtained by ChIP-loop. First, it is not clear why current protocols ligate the fragments when they are bound and concentrated to the beads. Concentrating the DNA on the beads prior to



ligation is expected to facilitate the formation of junctions between bead-associated, but not necessarily formaldehyde-crosslinked, DNA fragments, hence also producing results that reflect loops formed on the beads rather than in nuclear space. Unless the user can demonstrate that such undesired events do not take place, we would argue it is better to carry out the precipitation after the ligation step, which needs to be performed under conditions as described in the 3C procedure.

Second, accurate quantification of ligation products, already very challenging in standard 3C, is even more complicated in ChIP-loop assays because it must take into account the relative enrichment of each site on the beads. For example, we would argue that ChIP-loop assays should only be directed to the analysis of fragments that are both enriched by ChIP. Indeed, we tend to question the relevance of analyzing, via ChIP-loop, interactions between DNA segments that are not bound by the protein of interest, or between DNA segments of which only one is enriched by the antibody. After all, if a sequence had been co-precipitated because it was cross-linked to a target sequence of the protein of interest, it should also be found enriched in the ChIP assay.

It may be possible to obtain unique information, not obtainable from ChIP or 3C only, when studying loops formed between sites that are both precipitated because of their association to a protein of interest. But like in 3C, the mere detection of a ligation product may reflect a random collision rather than a specific interaction between ligated fragments and interaction frequencies need to be quantified accurately and compared to other interactions. This requires taking into account the genomic site separation between each pair of segments and the relative enrichment of each site on the beads, which seems to make the interpretation of ChIP-loop results very difficult. ChIP-loop assays can be useful though to identify proteins participating in long-range interactions in *cis* (i.e. over hundreds of kilobases or more) or in *trans*, as interpretation of these results will not be complicated by frequent random collisions.

### **5C technology**

Large-scale mapping of for example several hundreds of chromatin interactions using standard 3C is time consuming and difficult. The introduction of the 3C-Carbon Copy (5C) method generates the possibility of such large-scale locus-wide analysis<sup>6,27</sup>. The method uses a multiplex ligation-mediated amplification (LMA) step to amplify selected ligation junctions, thereby generating a quantitative carbon copy of a part of the initial 3C library, which is subsequently analyzed via microarray detection or high throughput sequencing (**Fig. 5.1**). LMA involves using a combination of test and fixed 5C primers that hybridize

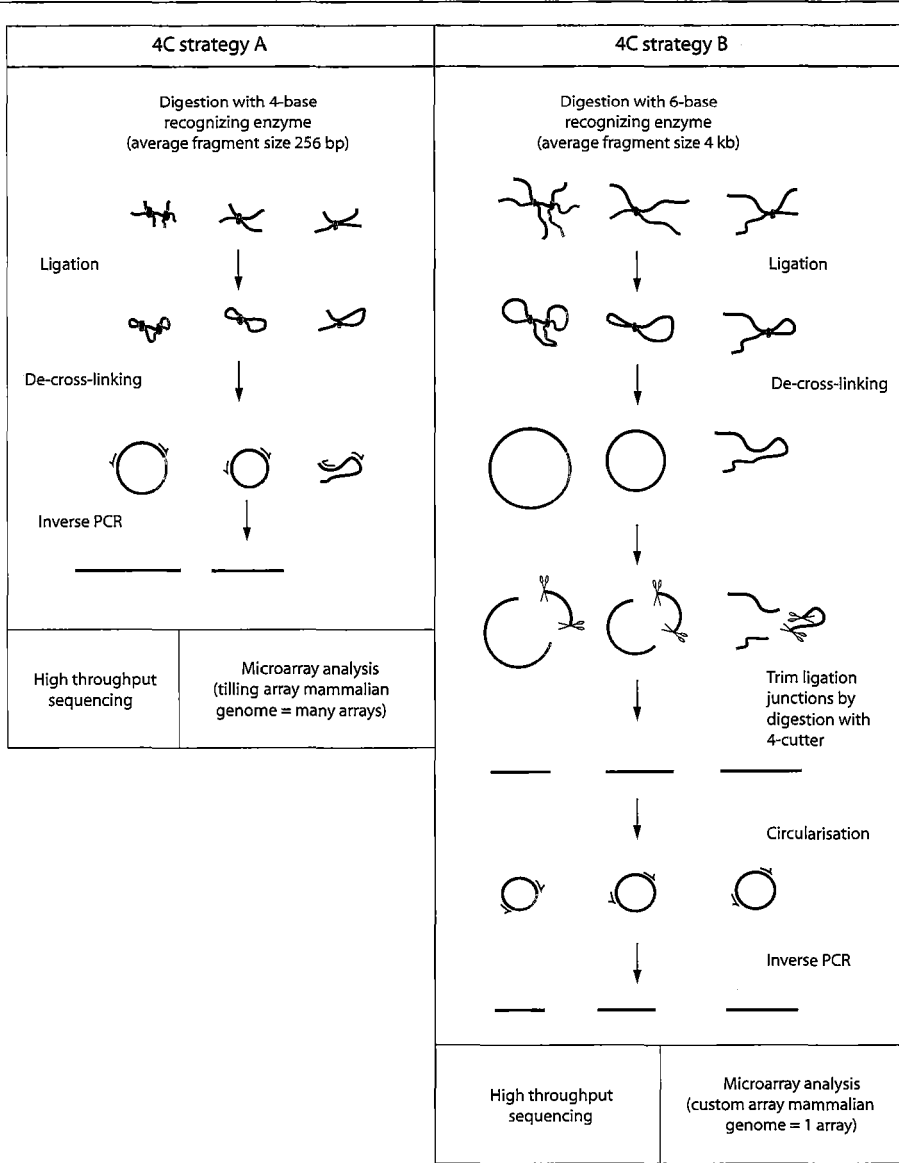
to the sense and antisense strand, respectively, of the restriction fragment ends analyzed. Fixed and test primers will be directly juxtaposed when a ligation junction is formed between the corresponding restriction sites, allowing subsequent primer-primer ligation. Universal tails protruding from the test and fixed primers, such as T7 and complementary T3 promoter sequences, subsequently enable massive parallel quantitative amplification of all investigated ligation products. Interestingly, 5C technology has the opportunity of analyzing a locus from a single or multiple fixed points. It can generate a complex matrix of interaction frequencies for a given genomic region, which can be used to reconstruct the intricate topology of this region. However, the size of the genomic region that can be studied is limited by the number of 5C primers that can be used simultaneously. Scanning hundreds of megabases of the genome will require using tens of thousands of 5C primers, which makes the technology less suitable to genome-wide scans<sup>6</sup>.

#### 4C technology

3C and 5C technology have been developed to identify interacting elements between selected parts of the genome and both techniques require the design of primers for all restriction fragments analyzed. Recently, new related strategies, collectively referred to as '4C technology', have been developed that allow screening the entire genome in an unbiased manner for DNA segments that physically interact with a DNA fragment of choice<sup>14, 21,23,28</sup>.

An outline of 4C technology is provided in (**Fig. 5.3**). Just like 3C, 4C technology depends on the selective ligation of cross-linked DNA fragments to a restriction fragment of choice (the 'bait'). In 4C technology, all the DNA fragments captured by the bait are simultaneously amplified via inverse PCR, using two bait-specific primers that amplify from circularized ligation products. Essentially two strategies can be pursued to obtain these DNA circles (**Fig. 5.3**). One strategy relies on the formation of circles during the standard 3C ligation step, i.e. while the DNA is still cross-linked<sup>21,28</sup>. Here, circle formation requires both ends of the bait fragment to be ligated to both ends of a captured restriction fragment. After de-crosslinking, captured DNA fragments are directly amplified by inverse PCR, using bait-specific primers facing outwards. Four-cutters are preferred in this method<sup>21</sup>, since they produce smaller restriction fragments (average size 256 bp, versus ~4 kb for six-cutters) and linear PCR amplification of the captured DNA fragments requires that the average product size is small.

The second strategy relies on the formation of DNA circles after the chromatin has been de-cross-linked. Here, the standard 3C procedure is followed, using a six-cutter as the



**Figure 5.3 Outline of the two basic 4C strategies.** Strategy A uses a 4-cutter enzyme, resulting in a resolution of 256 bp, and relies on the circular ligation of DNA fragments while they are still cross-linked. In strategy B, the cross-linked material is digested with a 6-cutter enzyme resulting in a resolution of 4 kb. The ligation junctions are trimmed and circularized after de-cross-linking. In both strategies PCR products can be analyzed by either large-scale sequencing or microarray analysis, in which strategy A requires (many) tilling arrays and strategy B one custom designed array, for a whole genome analysis.

restriction enzyme and yielding a de-cross-linked 3C template. The ligation junctions are then trimmed using a frequently cutting secondary restriction enzyme and re-ligated under conditions that favor the formation of selfligated circles. Inverse PCR primers hybridizing to the bait are used to linearly amplify (the small outer ends of) captured DNA fragments<sup>14,23</sup>. Since the two strategies have not been worked out in similar detail yet, it is currently difficult to compare them. Theoretically though, each strategy will have its own advantages and disadvantages.

The first approach presents the advantage of requiring less processing steps and the use of four-cutters provides a higher resolution (256 bp versus 4 kb for a four-cutter versus six-cutter, respectively). This should allow for a better definition of the site of interaction, which is expected to be particularly useful for identifying cis-regulatory DNA elements that locate away from a gene of interest. A potential issue of concern exists if formaldehyde cross-links multiple DNA fragments together. As a consequence of this, circles formed between cross-linked DNA fragments may contain more than two captured fragments, which will often be too large to be amplified in a linear fashion. This, in turn, may affect the reproducibility of the approach. It is also not clear how efficient circle formation is between DNA fragments that reside in cross-linked chromatin aggregates. Analysis of fragments captured by this approach so far has been limited to the sequencing of relatively small numbers of clones, 114<sup>21</sup> and 320<sup>28</sup>, respectively. Although these studies identified interesting DNA fragments, it is not clear whether such small number of clones provides a fair representation of the complex library of ligation junctions.

The advantage of the second approach is that it depends on the ligation of only one end of the bait to one end of a cross-linked DNA fragment, which will be more efficient than forming a circle between cross-linked DNA fragments. Circle formation takes place when the DNA is naked, which will also be more efficient than when the DNA is cross-linked. Products to be amplified will generally be smaller, as the circles will not contain more than one captured fragment, and therefore easier to amplify in a linear fashion. Since the strategy selectively amplifies the ligated outer ends of the restriction fragments created by the six-cutter, the complexity of the genomic library to be analyzed is strongly reduced. One can take advantage of this by designing tailored microarrays containing only probes located directly adjacent (within 100 bp) to each recognition site of a given six-cutter (e.g. HindIII) in the genome to analyze the captured DNA fragments<sup>14</sup>. This design allows for a large representation of the genome to be spotted on a single array. In fact, current designs cover the complete human or mouse genome on a single Nimblegen microarray (400,000

probes), enabling the identification of interactions at a resolution of ~7 kb (unpublished data).

Tailored microarrays were used to simultaneously analyze hundred thousands of fragments captured by the second approach. Replicate experiments performed on biologically independent samples demonstrated this strategy to be highly reproducible. No matter the bait chosen for analysis (we have now analyzed interactions with more than 15 different baits), it is always found that sequences physically close on the linear chromosome template are largely over-represented (**Fig. 5.4a**). In fact, restriction fragments within 5 to 10 megabases (Mb) from the bait are always captured so efficiently that they saturate every corresponding probe present on the array, precluding a quantitative analysis of local signal intensities. Further away from the bait and on other chromosomes, clusters of 20-50 neighbouring restriction fragments can be identified that all show increased hybridisation signals (**Fig. 5.4b**). Since each probe analyses an independent ligation event and only two fragments can be captured per cell, such clustering of interacting DNA fragments strongly indicates that this genomic region contacts the bait in multiple cells. Importantly, high-resolution cryo-FISH confirmed in an independent manner for more than 20 of these regions that they truly represent interacting regions in cis and in trans<sup>14</sup>. These experiments also showed that 4C technology identifies trans- and cis-interacting regions even if they are together in only 4% and 6% of the cells, respectively (cryo-FISH background: < 2% in trans and < 4% in cis).

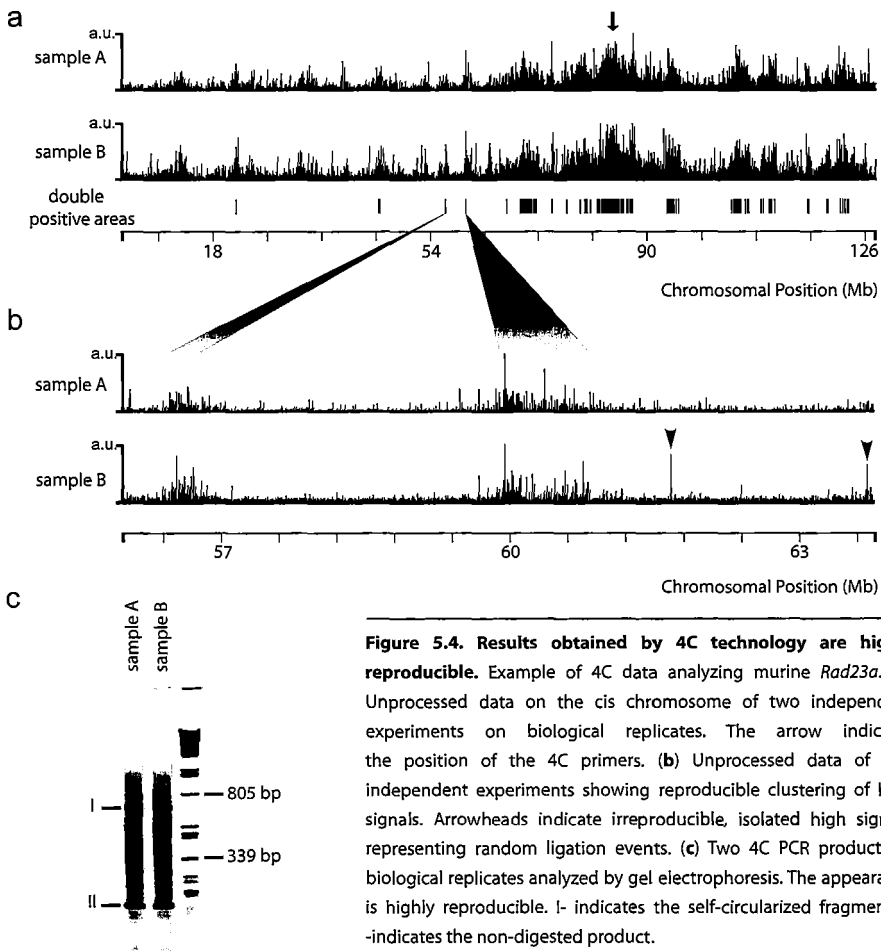
## **Potential pitfalls of 4C technology**

### *Number of cells*

No matter which strategy is followed, a number of critical steps need to be considered. First, the analysis needs to be performed on a relatively large population of cells. Even frequent interactions between fragments close on the linear chromosome template are captured often in less than 1/500 cells and we think that the trans and long-range cis-interactions that we identify are captured in only 1/10,000 or even 1/100,000 cells. We routinely process 10 million cells and perform 16 inverse PCR reactions on 200ng template, which we subsequently pool and label for microarray hybridization. Hence, we analyse an equivalent of approximately one million interactions on a single microarray.

PCR

The advantage of 4C (and 5C) over 3C is that only 2 primers are required to amplify all products, circumventing the problem of differences in primer pair efficiencies. All PCR based methods suffer from the fact that different amplicons amplify with a different efficiency. By performing the same PCR on a control template containing all ligation products in equimolar amounts one can correct for these differences in 3C and 5C, but not in 4C. It is absolutely critical to optimize the 4C-PCR step, as this step will select the DNA fragments for analysis, which need to correctly represent the fragments captured by the bait. Typically, 80% of the DNA fragments are smaller than 600 bp when samples are processed first with a six cutter and then with a four cutter, but one also wants larger



**Figure 5.4. Results obtained by 4C technology are highly reproducible.** Example of 4C data analyzing murine *Rad23a*. (a) Unprocessed data on the cis chromosome of two independent experiments on biological replicates. The arrow indicates the position of the 4C primers. (b) Unprocessed data of two independent experiments showing reproducible clustering of high signals. Arrowheads indicate irreproducible, isolated high signals, representing random ligation events. (c) Two 4C PCR products of biological replicates analyzed by gel electrophoresis. The appearance is highly reproducible. I- indicates the self-circularized fragment. II- indicates the non-digested product.

fragments to be amplified in a linear fashion. Different polymerases will perform this task with different levels of success (data not shown). One can use 3C primers and real-time PCR to test if the abundance of different sized products is similar before and after the inverse PCR step in 4C. We have used this strategy to define conditions that allow fragments of at least 1.2 kb to be amplified at very similar efficiencies (less than two-fold bias after 30 cycles of PCR; see supplementary protocol). When separated by gel electrophoresis, biological replicates should give a similar smear of PCR products and a number of more prominent bands that are reproducible between the samples (**Fig. 5.4c**). One should also check if the theoretically most abundant products that originate from the non-digested template and from the self-ligated circle are prominently present, which also confirms that the inverse PCR works (**Fig. 5.4c**).

#### *High-throughput analysis*

While sequencing of even hundreds of clones may reveal potentially interesting DNA fragments, we strongly recommend high-throughput analysis of captured DNA fragments, using either microarrays or large-scale sequencing, to exclude that analysis is focused on a misrepresentation of the actual library of captured fragments. Indeed, no matter the bait chosen for analysis and no matter the 4C strategy used, the great majority of captured fragments will always be located close to the bait on the linear chromosome template<sup>14,23</sup>.

#### *Analyzing 4C data*

High-throughput microarray analysis shows that probes with high signals are found across the chromosome and to a lesser extent also on other chromosomes. Many of these captures are random though, as they are not reproducible between independent duplicate experiments. Thus, highly specific long-range intra- and interchromosomal interactions with single restriction fragments may exist, but it is very difficult to discriminate them from random captures. The presence of genomic clusters of restriction fragments that show increased hybridization signals in biologically replicate experiments reveal interacting regions, as explained above (**Fig. 5.4b**). These regions can be identified by the application of a sliding window approach that provides a measure for the relative abundance of ligated fragments per genomic area<sup>14</sup>.

#### *Verification of 4C data*

3C technology may be used as a first verification of data obtained by 4C technology. However, they are not independent technologies and long-range interactions identified by

4C technology should therefore always be verified by completely independent methods such as FISH. Preferably this is done by high-resolution FISH studies, such as 3D-FISH or cryo-FISH<sup>29</sup> that use fixation conditions, which well preserve the nuclear ultra-structure. It needs to be demonstrated that two regions identified to interact by 4C technology indeed come together more frequently in the population of cells than two randomly chosen loci.

### **Concluding remarks and perspectives**

The development of 3C technology has contributed enormously to our understanding of the intricate folding of gene loci and revealed for example that transcriptional regulatory DNA elements loop towards their target genes to regulate the expression. Based on 3C technology, a number of new approaches have recently been developed. The ChIP-loop assay may direct structure analysis to specific protein-bound DNA sequences, but correct interpretation is currently still complicated, as it requires a quantitative comparison between ChIP-loop, ChIP and 3C data. 5C technology is expected to provide unprecedented insight into the conformational fine-structure of selected regions in the genome. Like 4C, it may help screening a genomic region for DNA elements that interact with a DNA segment of choice, being either a gene (promoter), an insulator sequence, an enhancer, an origin of replication, etc. 4C technology is expected to contribute importantly to a comprehensive understanding of nuclear architecture<sup>15</sup>, picking up interactions not previously anticipated and putting the relative frequency of interactions in perspective. Current 4C microarray studies allow identifying long-range interactions in cis, over tens of megabases and in trans, between chromosomes. The large over-representation of fragments closer to the bait precludes a quantitative analysis of local interactions, but it is to be expected that 4C can be modified to also identify loops formed in smaller genomic regions. In the near future more novel 3C-based methods may be expected. Their potential should be evaluated not so much on the possibly exciting nature of the interactions identified but on the independent evidence, obtained e.g. by FISH, that is provided to demonstrate that interactions are real.

### **Supplementary Information**

Supplementary Protocol '4C technology' available online.

### **Competing Financial Interests**

WdL holds a patent application no. PCT/IB2006/002268 on 4C technology.



## **Acknowledgements**

We thank Frank Grosveld and the members of our lab for discussion. This work was supported by grants from the Dutch Scientific Organization (NWO) (016-006-026) and (912-04-082) to W.d.L.

## **References**

1. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* 295, 1306-11 (2002).
2. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* 10, 1453-65 (2002).
3. Splinter, E., Grosveld, F. & de Laat, W. 3C technology: analyzing the spatial organization of genomic loci in vivo. *Methods Enzymol* 375, 493-507 (2004).
4. Miele, A., Gheldof, N., Tabuchi, T.M., Dostie, J. & Dekker, J. Mapping chromatin interactions by chromosome conformation capture (3C). In *Current protocols in molecular biology* (eds. Ausubel, F.M. et al.), pp. 21.11-20. John Wiley & Sons, NewYork. 2006.
5. Hagege, H. et al. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat Protoc* 2, 1722-33 (2007).
6. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc* 2, 988-1002 (2007).
7. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods* 3, 17-21 (2006).
8. Palstra, R.J. et al. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* 35, 190-4 (2003).
9. Murrell, A., Heeson, S. & Reik, W. Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent specific chromatin loops. *Nat Genet* 36, 889-93 (2004).
10. Spilianakis, C.G. & Flavell, R.A. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nat Immunol* 5, 1017-27 (2004).
11. Liu, Z. & Garrard, W.T. Long-Range Interactions between Three Transcriptional Enhancers, Active  $V\{\kappa\}$  Gene Promoters, and a 3' Boundary Sequence Spanning 46 Kilobases. *Mol Cell Biol* 25, 3220-31 (2005).
12. O'Sullivan, J.M. et al. Gene loops juxtapose promoters and terminators in yeast. *Nat Genet* 36, 1014-8 (2004).
13. Skok, J.A. et al. Reversible contraction by looping of the *Tcra* and *Tcrb* loci in rearranging thymocytes. *Nat Immunol* 8, 378-87 (2007).

14. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38, 1348-54 (2006).
15. de Laat, W. & Grosveld, F. Inter-chromosomal gene regulation in the mammalian cell nucleus. *Curr Opin Genet Dev* 18; [Epub ahead of print] (2007).
16. Solomon, M.J. & Varshavsky, A. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A* 82, 6470-4 (1985).
17. Orlando, V., Strutt, H. & Paro, R. Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods* 11, 205-14 (1997).
18. Jackson, V. Formaldehyde cross-linking for studying nucleosomal dynamics. *Methods* 17, 125-39 (1999).
19. Hogan, G.J., Lee, C.K. & Lieb, J.D. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet* 2, e158 (2006).
20. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 17, 877-85 (2007).
21. Zhao, Z. et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38, 1341-7 (2006).
22. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* 20, 2349-54 (2006).
23. Wurtele, H. & Chartrand, P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Res* 14, 477-95 (2006).
24. Horike, S., Cai, S., Miyano, M., Cheng, J.F. & Kohwi-Shigematsu, T. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat Genet* 37, 31-40 (2005).
25. Cai, S., Lee, C.C. & Kohwi-Shigematsu, T. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nat Genet* 38, 1278-88 (2006).
26. Kumar, P.P. et al. Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. *Nat Cell Biol* 9, 45-56 (2007).
27. Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16, 1299-309 (2006).

28. Lomvardas, S. et al. Interchromosomal interactions and olfactory receptor choice. *Cell* 126, 403-13 (2006).
29. Branco, M.R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* 4, e138 (2006).



# 6

## High-resolution identification of chromosomal rearrangements by 4C technology

Submitted

## High-resolution identification of chromosomal rearrangements by 4C technology

Marieke Simonis<sup>1,5</sup>, Petra Klous<sup>1,5</sup>, Irene Homminga<sup>4</sup>, Robert-Jan Galjaard<sup>2</sup>, Erik-Jan Rijkers<sup>3</sup>, Frank Grosveld<sup>1</sup>, Jules P.P. Meijerink<sup>4</sup>, Wouter de Laat<sup>1,5,6</sup>

<sup>1</sup> Dept of Cell Biology, <sup>2</sup> Dept of Clinical Genetics, <sup>3</sup> Dept of Biochemistry, <sup>4</sup> Dept of Pediatric Oncology, Erasmus Medical Center, Dr. Molewaterplein 50, 3015 GE, Rotterdam, The Netherlands. <sup>5</sup>New address: Hubrecht Institute, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands. <sup>6</sup> Corresponding author: E-mail: w.delaat@niob.knaw.nl

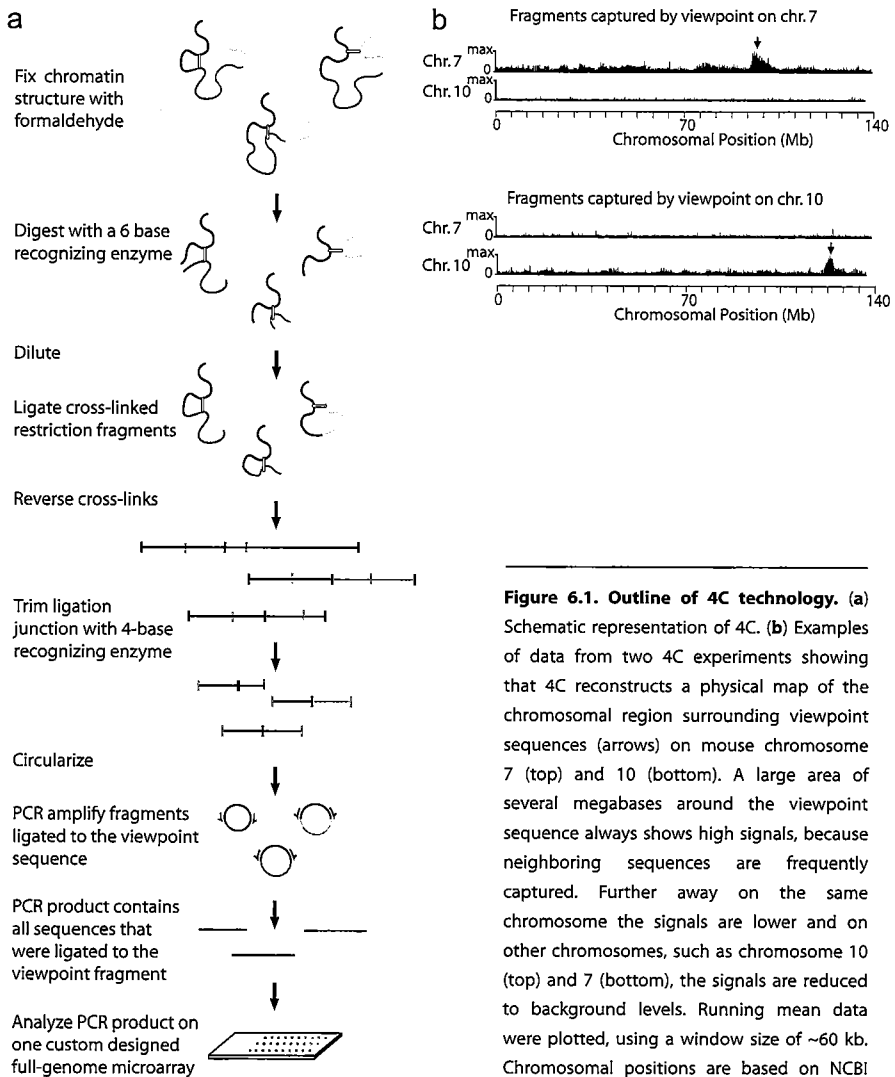
### **Summary**

*Balanced chromosomal rearrangements (inversions, translocations) can cause disease, but techniques for their rapid and accurate identification are missing. Here we demonstrate that 4C technology detects balanced and complex inversions and translocations at high resolution. 4C is robust as it uniquely relies on the capture and identification of many genomic fragments across the breakpoint. Importantly, it is little hampered by repetitive DNA. Using 4C, the LMO3 gene is uncovered as a novel potentially oncogenic translocation partner of the T cell receptor  $\beta$  gene (TCRB) in leukemia. Multiplex 4C is developed in combination with a single tailored microarray to rapidly screen for known and new translocation partners of loci that are frequently rearranged in leukemia. Unsuspected novel translocations and complex rearrangements are identified. Furthermore, we show that 4C can detect translocations even in small sub-populations of cells. This work opens avenues for the efficient screening of balanced rearrangements, the discovery of genes associated with cancer and other genetic diseases and the detection of rearrangements present in mixtures of cell types.*

### **Introduction**

Chromosomal rearrangements (deletions, amplifications, inversions, translocations) occur naturally in the genome[1-7] and often cause disease. Their accurate analysis is important both for understanding the mechanism of disease and for optimal diagnosis and

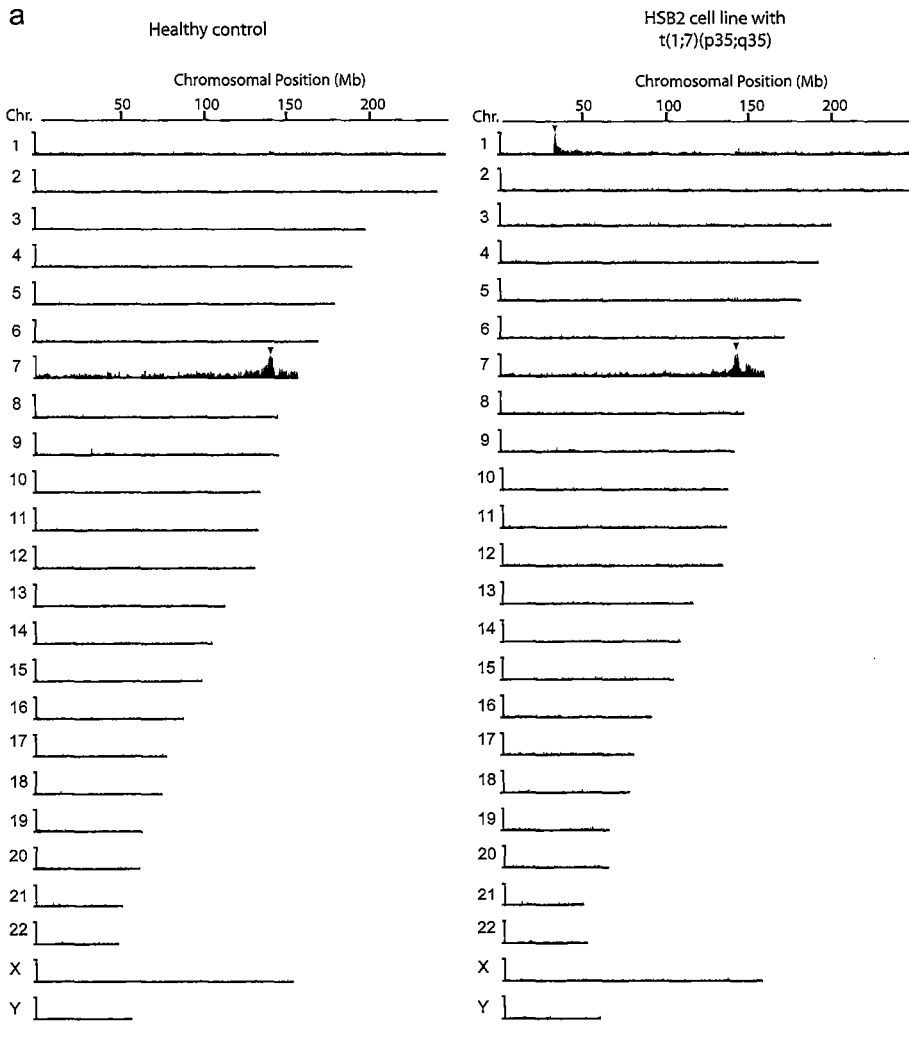
treatment decisions. Deletions and amplifications causing copy number variation can be detected at high resolution using microarray-based comparative genomic hybridization (array-CGH). However, high resolution mapping of translocations and inversions not accompanied by loss or gain of DNA cannot be done on a routine basis[8]. Here, we demonstrate that novel Chromatin Conformation Capture on Chip (4C) technology[9] rapidly identifies balanced and unbalanced genomic rearrangements at high resolution.



**Figure 6.1. Outline of 4C technology.** (a) Schematic representation of 4C. (b) Examples of data from two 4C experiments showing that 4C reconstructs a physical map of the chromosomal region surrounding viewpoint sequences (arrows) on mouse chromosome 7 (top) and 10 (bottom). A large area of several megabases around the viewpoint sequence always shows high signals, because neighboring sequences are frequently captured. Further away on the same chromosome the signals are lower and on other chromosomes, such as chromosome 10 (top) and 7 (bottom), the signals are reduced to background levels. Running mean data were plotted, using a window size of ~60 kb. Chromosomal positions are based on NCBI m36.

**Results**

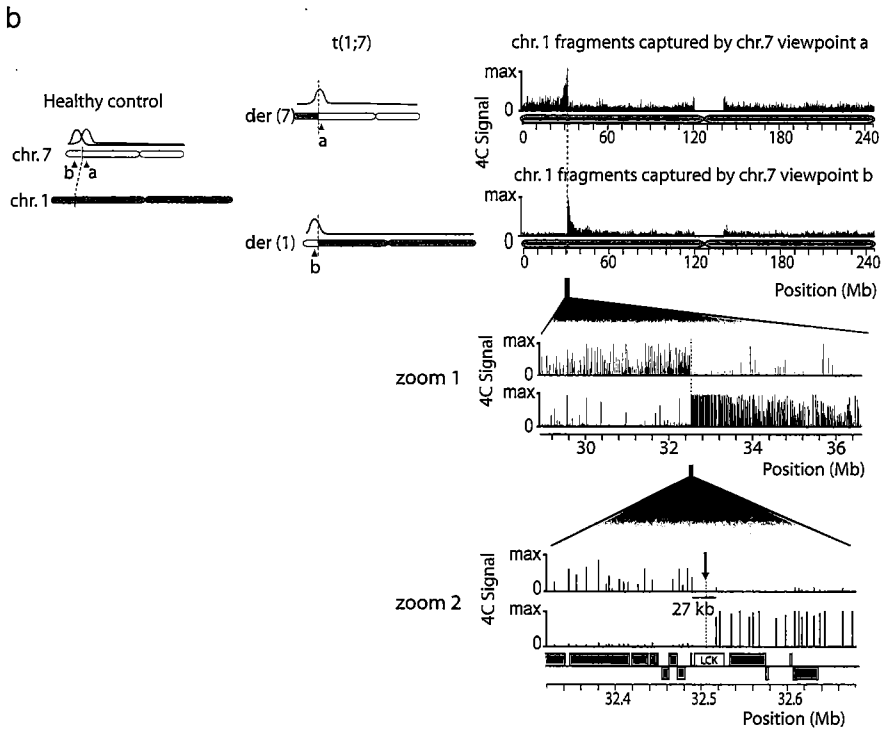
4C technology[10] (**Fig. 6.1a**) involves treatment of cells with formaldehyde to cross-link parts of the genome that are physically close in the nucleus. The DNA is subsequently digested with a restriction enzyme (here: HindIII) and cross-linked DNA fragments are ligated. Consequently, DNA fragments are ligated to fragments that are physically close in the nucleus. Inverse PCR with primers specific for a selected locus (the 'viewpoint') subsequently allows amplifying its interacting partners. When analyzed on a 4C-tailored microarray (385K probes) that analyzes the entire human genome at an average resolution of 7 kb[9], the highest hybridization signals always map to a region of several megabases



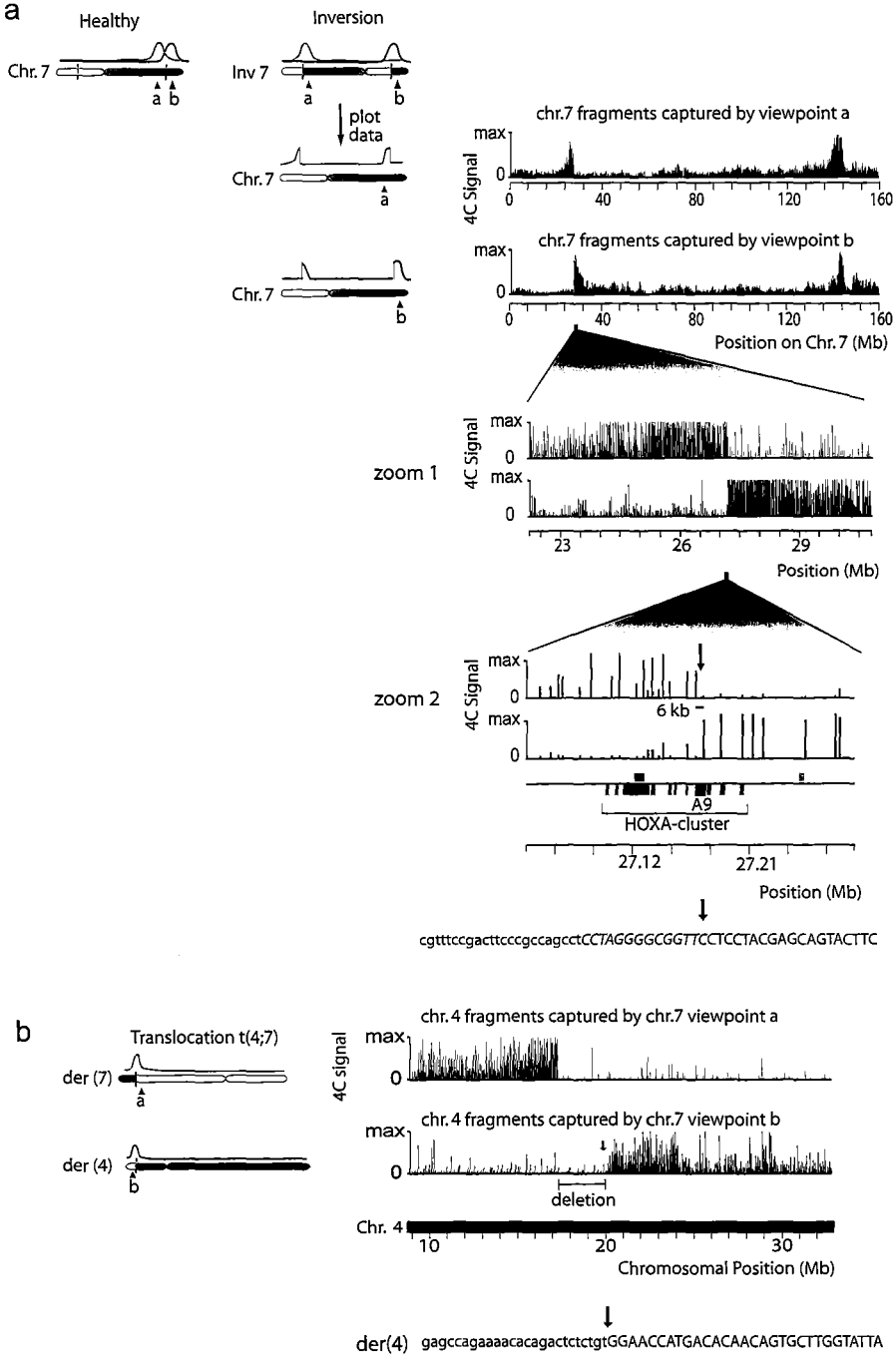


surrounding the viewpoint sequence (**Fig. 6.1b**). Thus, 4C technology predominantly identifies flanking sequences of the viewpoint and we reasoned it should therefore detect genomic rearrangements present in this chromosomal area.

To test this hypothesis, 4C technology was applied to the HSB-2 T-ALL cell line, containing a reciprocal translocation  $t(1;7)(p35;q35)$  between the *T cell receptor  $\beta$*  (*TCRB*) locus on 7q35 and the *LCK* locus on 1p35[11]. Two independent 4C experiments were performed, analyzing DNA interactions with viewpoint sequences located 462 kb centromeric and 239 kb telomeric of the breakpoint in *TCRB*. With both viewpoint sequences, strong hybridization signals were observed not only around the *TCRB* locus on chromosome 7 but



**Figure 6.2. Whole genome 4C accurately detects a balanced translocation.** (a) 4C signals across all chromosomes in a healthy control and the HSB-2 cell line carrying  $t(1;7)(p35;q35)$ . The black arrowheads indicate position of viewpoint sequences. The grey arrowhead indicates the position of the translocation site. Running mean data were plotted, using a window size of ~60 kb. Scale on Y-axis (arbitrary units) is identical for all chromosomes. (b) 4C signals on chromosome 1, using viewpoint fragments a (grey) and b (black) located at opposite sides of the locus on chromosome 7. The regions on chromosome 1 captured by viewpoint fragments on chromosome 7 directly neighbor each other and flank the previously cloned breakpoint (arrow). -axes represent raw intensities of the microarray signals, max=65535. For whole chromosome views a running mean with a window size of 29 probes was applied. Chromosomal positions are based on NCBI m36.



also across a megabase region on 1p35, specifically in HSB-2 cells (**Fig. 6.2**). These signals represented restriction fragments captured by the viewpoint sequences on chromosome 7 as an effect of their close physical proximity. Importantly, the first restriction fragments captured on chromosome 1 in both experiments directly flank the previously identified chromosomal breakpoint. Thus, 4C locates translocation breakpoints to a position in between the pair of probes that represents the transition from non-captured to captured restriction fragments.

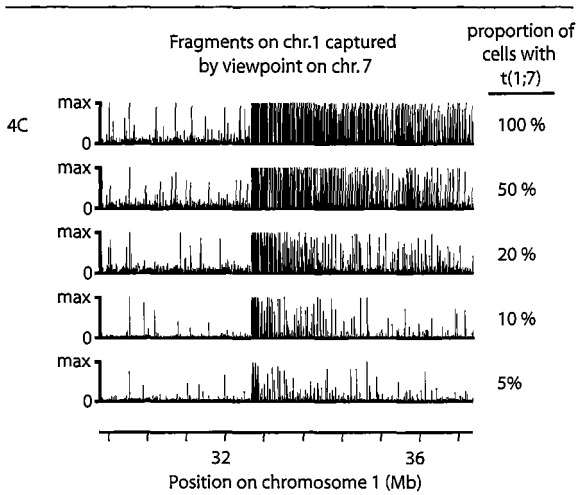
Next, we tested whether 4C can also identify inversions by applying it to a pediatric T-ALL sample that, based on FISH, carries an inversion on chromosome 7, *inv(7)(p15q35)*. This abnormality leads to the rearrangement of the *TCRB* locus into the *HOXA* gene cluster and activation of the *HOXA* genes [12, 13]. The same set of *TCRB* viewpoint sequences described above was used. The telomeric *TCRB* viewpoint sequence captured centromeric *HOXA* fragments and the centromeric *TCRB* viewpoint fragment captured telomeric *HOXA* fragments, thus revealing an inversion between the loci in the patient sample (**Fig. 6.3a**). The two captured regions directly neighbor each other and locate the breakpoint to a 6 kb region. Restriction-fragment-paired-end sequencing of the breakpoint (**Supplementary Fig. 6.1**) confirmed the location of this breakpoint (**Fig. 6.3a**). Thus 4C technology can detect balanced translocations and inversions at high resolution.

The potential of 4C technology was further explored by applying it to an EBV transformed cell line derived from a patient with Postaxial Polydactyly (PAP). PAP is an autosomal dominant heritable disorder characterized by extra ulnar of fibular digits. The patient cells were previously characterized by karyotyping and FISH to contain an unbalanced translocation between chromosomes 4 and 7, *t(4;7)(p15.2;q35)*, with a micro-deletion of unknown size [14]. Two 4C experiments were performed, each analyzing DNA interactions with a viewpoint fragment located on another side of the rearranged part of chromosome 7. Importantly and in contrast to what was found for the balanced translocation, the chromosome 4 fragments captured by the two viewpoint sequences on chromosome 7 did not directly flank each other but were 2.8 Mb apart (**Fig. 6.3b** and **Supplementary**

---

**Figure 6.3. 4C accurately detects a balanced inversion and an unbalanced translocation.** (a) Balanced inversion detected in a T-ALL patient sample using viewpoint fragments a (grey) and b (black) located at opposite sides of the *TCRB* locus on chromosome 7. The breakpoint (arrow) was sequenced. (b) 4C accurately detects an unbalanced translocation *t(4;7)*. Viewpoint fragments a (grey) and b (black) were located on opposite sides of the breakpoints on chromosome 7 and both captured fragments on chromosome 4. The two captured regions do not directly neighbor each other, demonstrating a deletion on chromosome 4. The breakpoint of derivate chromosome 4 (arrow) was sequenced. Y-axes represent raw intensities of the microarray signals, max=65535. For whole chromosome views a running mean with a window size of 29 probes was applied. Chromosomal positions are based on NCBI m36.

---



**Figure 6.4. 4C finds translocations in small subpopulations of cells.** 4C was applied to mixtures of K562 cells and decreasing amounts of HSB-2 cells, the latter containing translocation  $t(1;7)$ . Using a viewpoint neighboring the breakpoint on chromosome 7, sequences on chromosome 1 were captured efficiently in all mixtures, even when only 5% of the cells carried the translocation.

**Fig. 6.2).** One of the breakpoints was cloned and sequenced and confirmed to locate at 20.08 Mb (**Fig. 6.3b**). The sequence showed that the breakpoint on chromosome 7 was 4 Mb away from the viewpoint sequence. Thus, 4C viewpoint sequences can capture DNA fragments and characterize rearrangements even when the breakpoints are several megabases away. The data also show that 4C technology is very suitable to fine-map poorly characterized rearrangements identified by chromosomal

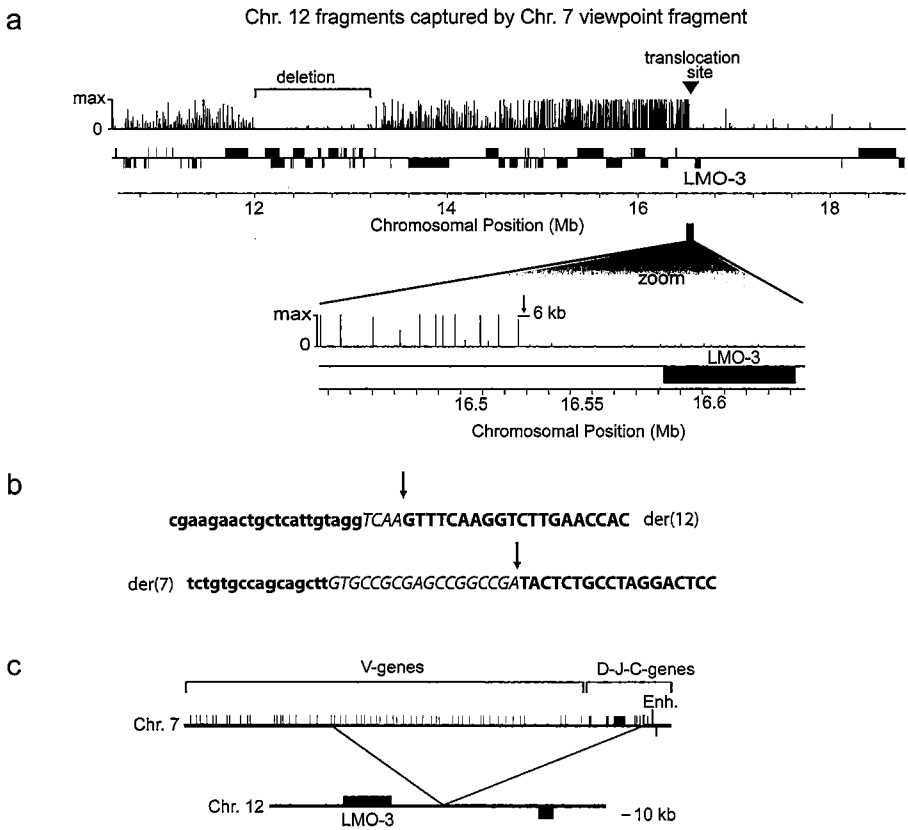
karyotyping. When directed to both sides of a genomic breakpoint, 4C can immediately identify whether a translocation or inversion is balanced or accompanied by additional rearrangements such as a deletion (i.e. unbalanced).

We next investigated whether 4C technology can identify a deletion not associated with a translocation. For this, 4C was applied to a pediatric T-ALL patient sample previously characterized by array-CGH to contain a homozygous deletion of the *p15/p16* loci on chromosome 9p21. Using a viewpoint fragment located ~2 Mb away from the nearest breakpoint, a region lacking probe signals was observed, demarcating the deleted area (**Supplementary Fig. 6.3a**). Importantly, increased hybridization signals are observed for the region immediately downstream of the deletion. This is expected since the deletion brings it in closer physical proximity to the viewpoint fragment. PCR across the ~2Mb deleted region confirmed the positions of the two breakpoints (**Supplementary Fig. 6.3b**). This shows that 4C technology identifies homozygous deletions as regions containing reduced hybridization signals across the deleted area in combination with increased hybridization signals on the other side of the deletion.

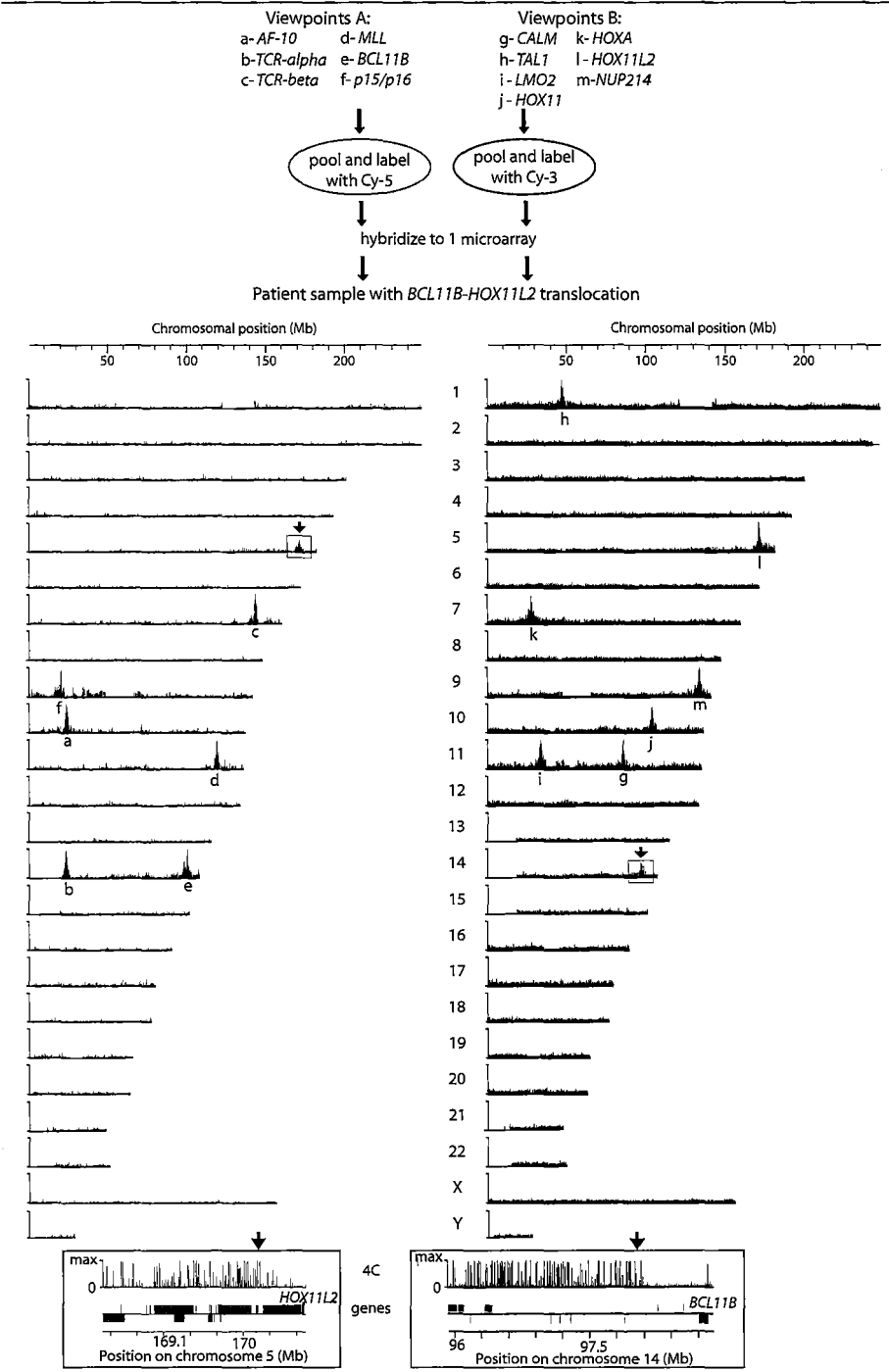
Tumors and tumor samples are often mosaic and current high-throughput techniques fail to detect rearrangements present in small sub-populations of cells. We applied 4C to cell

mixtures containing control cells (K562) and decreasing amounts of HSB-2 cells carrying the t(1;7) described above. We found that even if only 5 % of the analyzed pool of cells carried the translocation, this rearrangement could still be detected efficiently (**Fig. 6.4**). Thus, 4C can be used to identify balanced rearrangements in non-homogeneous samples, enabling early detection of rearrangements in small tumors.

We next asked the question whether 4C can easily identify novel rearrangements. *TCR* loci are frequently involved in chromosomal rearrangements in T-ALL, because translocations



**Figure 6.5. 4C identifies novel translocation partners.** (a) Five uncharacterized T-ALL patient samples were screened with 4C, using a viewpoint fragment near the *TCRB* locus on chromosome 7. In one sample high signals appeared specifically on chromosome 12, revealing a translocation, t(7;12)(q35;p12.3) (see Supplementary Fig. 5 for other chromosomes). A deletion is present several megabases from the translocation site (arrow) on chromosome 12 (zoom 1). The translocation site is present in a 6 kb region close to the *LMO3* gene (zoom 2). (b) Sequences of both breakpoints of t(7;12)(q35;p12.3); nucleotides in upper case are from 12, in lower case from 7 and in italics are from unknown origin. (c) Schematic representation of the translocation site of t(7;12)(q35;p12.3). The enhancer of *TCRB* is positioned 70 kb downstream of the *LMO3* gene. Running mean data were plotted, using a window size of ~60 kb. Zooms show unprocessed signal intensities. Chromosomal positions are based on NCBI m36.



can arise during the process of VDJ recombination. We screened five T-ALL patient samples for novel genetic rearrangements associated with the *TCRB* locus. One patient sample was found to carry a translocation between *TCRB* and the p arm of chromosome 12, plus an additional deletion ~3 Mb away from the translocation breakpoint on chromosome 12 (**Fig. 6.5** and **Supplementary Fig. 6.4**). The translocation  $t(7;12)(q35;p12.3)$  is new in T-ALL. The breakpoint on chromosome 12 was mapped at 6 kb resolution, cloned and sequenced (**Fig. 6.5b**). The translocation positions the enhancer of *TCRB* 70 kb downstream of the Lim-domain-only gene *LMO3* (**Fig. 6.5c**). Microarray expression data showed that *LMO3*, normally silenced in T cells, is highly active in this T-ALL sample (**Supplementary Fig. 6.5**). The protein family members *LMO-1* and *LMO-2*, but not *LMO3*, have previously been found as oncogenic translocation partners of the *TCR* loci in T-ALL. Interestingly, *LMO3* was recently found to act as an oncogene in neuroblastoma[15]. Thus, 4C technology can discover new oncogenes rearranged with frequently modified loci. Finally, we aimed to develop a 4C strategy that would simultaneously identify all recurrent rearrangements associated with a given disease using a single microarray (**Fig. 6.6**, **Supplementary Fig. 6.6**, **6.7**). In T-ALL, a set of loci, in particular *TCRA*, *TCRB*, *Bcl11B* and *MLL*, is frequently found to recombine with various other genes[16, 17]. We included these 4 loci, together with 9 other loci described either as their translocation partner or to be rearranged otherwise in T-ALL[16, 17], in a 4C-multiplex strategy. The genomic sites interacting with each of the 13 viewpoints were PCR amplified separately and pooled in two mixes representing the chromosomal neighborhoods of 6 and 7 viewpoints, respectively. These mixes were differentially labeled and hybridized to the same microarray. The data show that multi-view 4C accurately identifies rearrangements in each of 11 T-ALL samples analyzed. We identified translocations between *TCRA-HOX11* and *TCRA-LMO2*, translocations between *Bcl11B-Nkx-2.5* and *Bcl11B-HOX11L2*, the translocation between *CALM-AF10*, the common *SET-Nup214* deletion, a *TCRA-c-myc* translocation and an *MLL-AF-6* translocation (**Fig. 6.6**, **Supplementary Fig. 6.6**, **6.7**). In one patient sample we found a novel *LMO1-TCRB* translocation (**Supplementary Fig. 6.7**); *LMO1* previously was only found to translocate to *TCRA*[16, 17]. Retrospectively, all these rearrangements

---

**Figure 6.6. 4C can analyze multiple sites frequently involved in rearrangements in T-ALL on one microarray.** 4C is performed on 13 viewpoints separately, PCR products are combined in two pools, labeled, and hybridized to a microarray. The 4C data show high signals on all the viewpoints included in the pools (small letters). Additional signals in the Cy-5 pool (square) on chromosome 5, demonstrate *HOX11L2* is a translocation partner of one of the Cy-5 viewpoints. In the Cy-3 pool *BCL11B* is found as a translocation partner (square). Together the data demonstrate the tested sample carries a *BCL11B-HOX11L2* translocation. The data of 8 other rearrangements detected by this method are shown in Supplementary figure 6 and Supplementary figure 7.

---

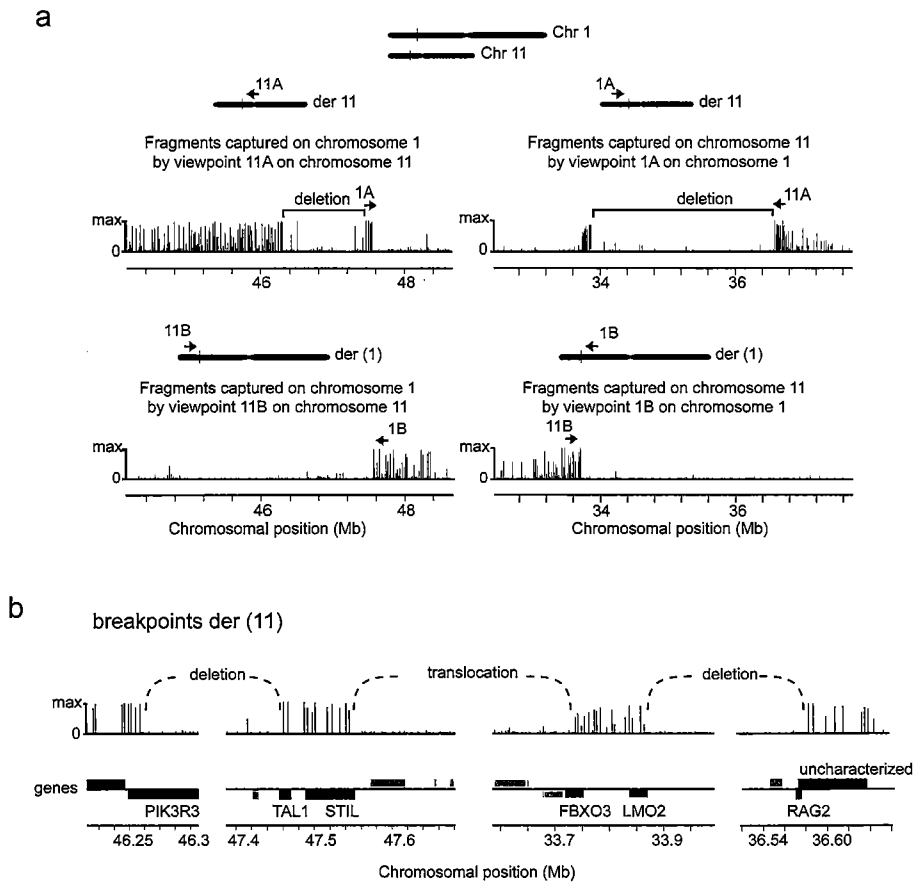
were confirmed by chromosomal karyotyping and/or FISH. In another patient (10110) we initially failed to identify the deletion previously mapped by array-CGH to locate between the *LMO2* gene and the *RAG1/RAG2* genes on chromosome 11[18]. Instead, using a target sequence ~500 kb telomeric of *LMO2*, we surprisingly found the region to be fused to an area on chromosome 1 that contained the *STIL* and *TAL1* gene, both also implicated in T-ALL (**Fig. 6.7a**). To further map this rearrangement, we applied 4C to sequences on either side of the breakpoints of the two chromosomes. Together, the data revealed a complex chromosomal rearrangement involving a translocation  $t(1;11)$  (**Fig. 6.7a**). The breakpoint on derivate chromosome 11 was flanked by two small (100-200 kb) regions carrying respectively the T-ALL genes *LMO2* and *STIL* and *TAL1*, followed on either side by a large deleted region spanning 1-3 Mb (**Fig. 6.7b**). Thus, the deletion identified by array-CGH[18] was shown by 4C to be part of a more complex chromosomal rearrangement involving a translocation. Interestingly, oncogenes like *LMO2* and *TAL1* are known to translocate to the *TCR* loci in T-ALL, but have not been documented previously to rearrange with each other. The fact that they also recombine may support the idea that nuclear co-localization of all these loci at some stage of T-cell development is an important mechanism behind translocation partner selection in T-ALL. Collectively, the data show that multi-view 4C with a single microarray identifies all recurrent, but also novel translocations and complex chromosomal rearrangements associated with T-ALL. Clearly, this strategy can also be adapted to the detection of such rearrangements in other types of disease.

## Discussion

4C was originally designed to study the folding structure of chromosomes by identifying regions that frequently contact a viewpoint (or "bait"). Folding of chromosomes is indeed not random and some regions of the genome are captured and identified by 4C due to the 3D structure of the chromosomes [9]. When interpreting 4C data, chromosomal rearrangements can easily be discerned from looped chromatin structures. First, signal intensities seen for genomic regions that are physically close on the linear chromosome template are much higher than those observed for distant regions that loop in 3D to the genomic viewpoint. Second, signal profiles from translocations, inversions and deletions each have a very distinct shape that is different from the more Gaussian curves seen for looped regions. Most distinctly, chromosomal rearrangements will show a sharp transition in signal intensities at the probes that surround the breakpoint. Third, the number of probes with positive signals near genomic breakpoints is usually much higher than observed for a region which loops towards the viewpoint. Thus, it is not difficult to



discriminate structural variation in the genome from three-dimensional configurations detected by 4C. However, it is important that sufficient numbers of ligation products are analyzed simultaneously. Each diploid cell donates a maximum of two ligation products per viewpoint. For singleplex 4C, we therefore routinely analyze ~0.5 million cells simultaneously on a microarray, or ~1 million interactions with the viewpoint. Under such conditions, translocations and other rearrangements are readily identified (Fig. 6.2, 6.3 and 6.5), even when the breakpoint is 3 Mb away from the viewpoint (Fig.



**Figure 6.7. 4C demonstrates an unexpected complex rearrangement.** 4C was applied to a T-ALL patient sample known to carry a large deletion on chromosome 11. The deletion was found to be accompanied by an unsuspected translocation to chromosome 1. (a) 4C was performed on both sides of the breakpoints on chromosome 1 and 11. The 4C data show that a deletion is found on both chromosomes. Moreover, the deletions do not directly neighbor the translocation sites. (b) Zoomed in view of the 4C data. The complex chromosomal rearrangement at der(11) preserves two small (100-200 kb) regions carrying the oncogenes LMO2 and TAL1 but deletes large (1-3Mb) regions further downstream of the translocation breakpoint.

**6.3b**) or present in small subpopulations of cells (**Fig. 6.4**). In the 4C-multiplex strategy, aimed to simultaneously screen at several sites for structural rearrangements, less PCRs were performed per viewpoint and PCR products from different viewpoints were mixed, reducing the net amount of PCR material per viewpoint hybridized to the array. As a result, signal intensities and the number of positive probes identified at the breakpoint dropped, but in each case we could still identify the underlying rearrangement, as confirmed by FISH and by additional 4C experiments.

The diagnostic method presented here is based on a novel concept: the capture and identification of all DNA fragments that are physically close in the nucleus, to reconstruct linear chromosome maps. Unlike other high-throughput methods like genomic re-sequencing or paired-end-sequencing[19, 20], it therefore does not depend on finding the single DNA fragment carrying the genomic breakpoint for the detection of balanced rearrangements. Instead, 4C identifies rearrangements based on the capture of many genomic fragments across the breakpoint. This implies that 4C is particularly powerful to uncover balanced rearrangements also when breakpoints are present in, or surrounded by, repetitive sequences. Deletions associated with (unbalanced) translocations are readily identified as well (**Fig. 6.3b** and **Fig. 6.7**). The same is true for large (balanced) inversions, but we expect that inversions smaller than ~1 Mb will be more difficult to detect with 4C. 4C requires the selection of one or more genomic viewpoints: for the detection of chromosomal breakpoints they need to be located within several megabases of these viewpoints. As such, the technique is very suitable for the rapid fine-mapping of balanced and complex rearrangements found with low-resolution techniques like FISH and chromosomal karyotyping. It is also very useful for the characterization of translocations of which only one of the involved chromosomes is known, and for screening of frequently rearranged sites to identify novel oncogenes. Examples of such sites are the T cell receptor loci in leukemia and the immunoglobulin loci in lymphomas. Moreover, we have shown that 4C can uncover complex structural rearrangements that accompany copy number changes identified with array CGH. We predict that genes that are aberrantly expressed in patient samples without apparent variation in their DNA copy number are interesting targets for 4C analysis, which is expected to lead to the identification of novel balanced rearrangements. In all such instances, 4C offers a cost-effective alternative to genomic re-sequencing, with the additional advantage that it will also identify translocations when present in small sub-populations of cells. Array painting is another recently developed technique for fine-mapping of translocation breakpoints. It involves isolating chromosomes based on size using flow-sorting and characterization of a selected

(derivative) chromosome by hybridization to a microarray or by large-scale sequencing [21, 22]. However, metaphase chromosomes often cannot be isolated (e.g. from solid tumors), not all chromosomes and chromosome derivatives can be isolated based on size and inversions cannot be detected using this technique.

The fact that 4C can detect translocations present in small subpopulations of cells makes it a potent technique to study rearrangements in mixtures of cells and mosaic tumors and creates the prospect of identifying tumor cells in early stages of metastasis.

## Materials and Methods

### *Sample preparation*

T-ALL patient samples and healthy control T-cell samples were processed as described [9, 23]. The EBV-transformed cell line derived from the PAP patient was cultured and handled as described before [9, 14].

### *4C array design*

The 60 bp probes were designed within 100 bp from a HindIII site, using criteria described previously, for example the selection of only unique DNA sequences [9]. To be able to cover the entire genome with the 400,000 probes that fit on the Nimblegen microarray, a selection was made. Probe numbers were first reduced by keeping only one probe per HindIII fragment, instead of one on each side. Secondly, probes were selected such that the spacing of probes was as equal as possible across the genome.

### *4C analysis*

4C analysis was performed as described previously [9], using the following primer sequences:

5'-end of <i>TCRB</i>	CATGAAGAAACGAGCACCC CCTTGATGTTTCTCCCTTACC
3'-end of <i>TCRB</i>	TGTCAGGCTCTTCTCTACAC GTCGTCCAGAACTCACC
Centromeric t(4;7)	AATCCAGGGCTACTTCCAG CCGTGATGCTATCTGCCA
Telomeric t(4;7)	TGTTGGAAGACCAGGTGAAG TGTCGTGGAAAGCGAGTG

## Chapter 6

Deletion 9            CAATCCCAGATACATTCTCAT  
                          ACAAATACTTTCCAAGACTGGAC  
3' of TCRA            GAATATGTTATGCTTGATCC  
                          TTCCATGAGAGAAGTCTAG

4C data was visualized using SignalMap software. To create whole chromosome view pictures of the 4C data, a running mean with a window size of 29 probes was calculated using the R package (<http://www.r-project.org>).

### *Restriction-fragment-paired-end-sequencing*

10 µg of genomic DNA was first digested in 500 µl with 10 U of an enzyme that recognizes 6 bases (HindIII, BglII or EcoRI) (37 °C for 2 hours). Samples were purified by phenol-chloroform extraction and ethanol precipitation. Subsequently, samples were ligated in 2 ml with 40 U of ligase (Roche) for 4 hours at 16 °C and 30 minutes at 20 °C. Ligated samples were purified by phenol-chloroform extraction and ethanol precipitation. A second digestion was performed with a restriction enzyme that recognizes 4-bases (e.g. NlaIII or DpnII) under the same conditions as described for the 6-base recognizing enzyme. Subsequent ligation was also as described above. Samples were purified by phenol-chloroform extraction and ethanol precipitation. Selected fragments were PCR amplified from 50-100 ng of DNA, using the following conditions: 94 °C for 3 minutes, followed by 30 cycles of 15 seconds at 94 °C, 1 minute at 55 °C and 2 minutes at 72 °C and one final step of 7 minutes at 72 °C.

### **Acknowledgements**

We would like to thank Annelies de Klein, Wilfred van IJcken, Jessica Gladdines, Joris Veltman, Alexander Hoischen and Erik Splinter for assistance. This work was supported by grants from the Dutch Scientific Organization (NWO) (912-04-082), the Netherlands Genomics Initiative (050-71-324) and an Erasmus MC grant to W.L and CGC, and the EU to FG.

### **References**

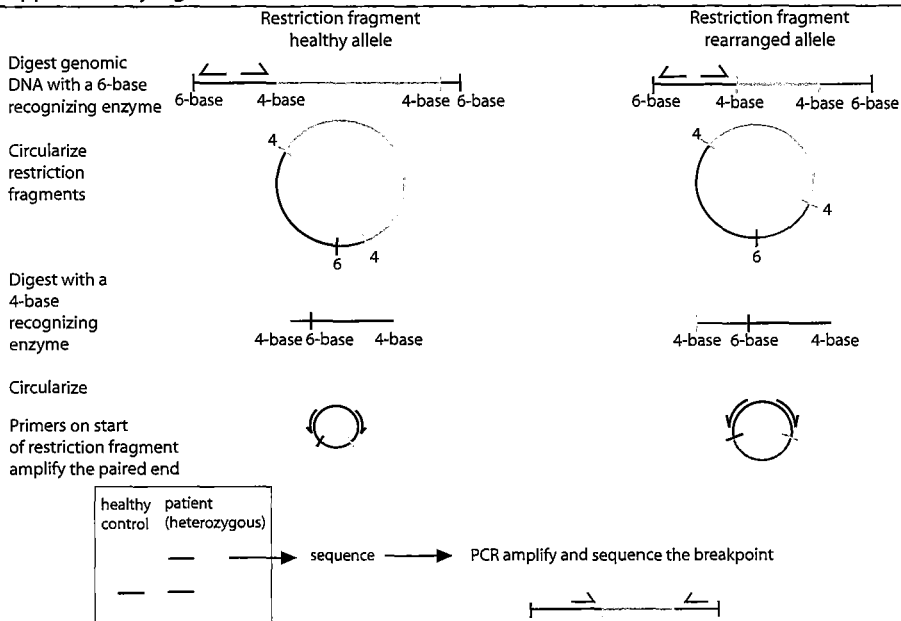
1.        *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
2.        Easton, D.F., et al., *Genome-wide association study identifies novel breast cancer susceptibility loci.* Nature, 2007. **447**(7148): p. 1087-93.

3. Eichler, E.E., et al., *Completing the map of human genetic variation*. *Nature*, 2007. **447**(7141): p. 161-5.
4. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome*. *Nat Rev Genet*, 2006. **7**(2): p. 85-97.
5. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome*. *Nat Genet*, 2004. **36**(9): p. 949-51.
6. Mehan, M.R., N.B. Freimer, and R.A. Ophoff, *A genome-wide survey of segmental duplications that mediate common human genetic variation of chromosomal architecture*. *Hum Genomics*, 2004. **1**(5): p. 335-44.
7. Sharp, A.J., Z. Cheng, and E.E. Eichler, *Structural variation of the human genome*. *Annu Rev Genomics Hum Genet*, 2006. **7**: p. 407-42.
8. Higgins, R.A., S.R. Gunn, and R.S. Robetorye, *Clinical application of array-based comparative genomic hybridization for the identification of prognostically important genetic alterations in chronic lymphocytic leukemia*. *Mol Diagn Ther*, 2008. **12**(5): p. 271-80.
9. Simonis, M., et al., *Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)*. *Nat Genet*, 2006. **38**(11): p. 1348-54.
10. Simonis, M., J. Kooren, and W. de Laat, *An evaluation of 3C-based methods to capture DNA interactions*. *Nat Methods*, 2007. **4**(11): p. 895-901.
11. Burnett, R.C., et al., *Molecular analysis of the T-cell acute lymphoblastic leukemia-associated t(1;7)(p34;q34) that fuses LCK and TCRB*. *Blood*, 1994. **84**(4): p. 1232-6.
12. Soulier, J., et al., *HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL)*. *Blood*, 2005. **106**(1): p. 274-86.
13. Speleman, F., et al., *A new recurrent inversion, inv(7)(p15q34), leads to transcriptional activation of HOXA10 and HOXA11 in a subset of T-cell acute lymphoblastic leukemias*. *Leukemia*, 2005. **19**(3): p. 358-66.
14. Galjaard, R.J., et al., *Isolated postaxial polydactyly type B with mosaicism of a submicroscopic unbalanced translocation leading to an extended phenotype in offspring*. *Am J Med Genet A*, 2003. **121**(2): p. 168-73.
15. Aoyama, M., et al., *LMO3 interacts with neuronal transcription factor, HEN2, and acts as an oncogene in neuroblastoma*. *Cancer Res*, 2005. **65**(11): p. 4587-97.
16. Armstrong, S.A. and A.T. Look, *Molecular genetics of acute lymphoblastic leukemia*. *J Clin Oncol*, 2005. **23**(26): p. 6306-15.
17. Graux, C., et al., *Cytogenetics and molecular genetics of T-cell acute lymphoblastic leukemia: from thymocyte to lymphoblast*. *Leukemia*, 2006. **20**(9): p. 1496-510.

18. Van Vlierberghe, P., et al., *The cryptic chromosomal deletion del(11)(p12p13) as a new activation mechanism of LMO2 in pediatric T-cell acute lymphoblastic leukemia*. *Blood*, 2006. **108**(10): p. 3520-9.
19. Korbelt, J.O., et al., *Paired-end mapping reveals extensive structural variation in the human genome*. *Science*, 2007. **318**(5849): p. 420-426.
20. Tuzun, E., et al., *Fine-scale structural variation of the human genome*. *Nat Genet*, 2005. **37**(7): p. 727-32.
21. Chen, W., et al., *Mapping translocation breakpoints by next-generation sequencing*. *Genome Res*, 2008. **18**(7): p. 1143-9.
22. Fiegler, H., et al., *Array painting: a method for the rapid analysis of aberrant chromosomes using DNA microarrays*. *J Med Genet*, 2003. **40**(9): p. 664-70.
23. van Vlierberghe, P., et al., *A new recurrent 9q34 duplication in pediatric T-cell acute lymphoblastic leukemia*. *Leukemia*, 2006. **20**(7): p. 1245-53.

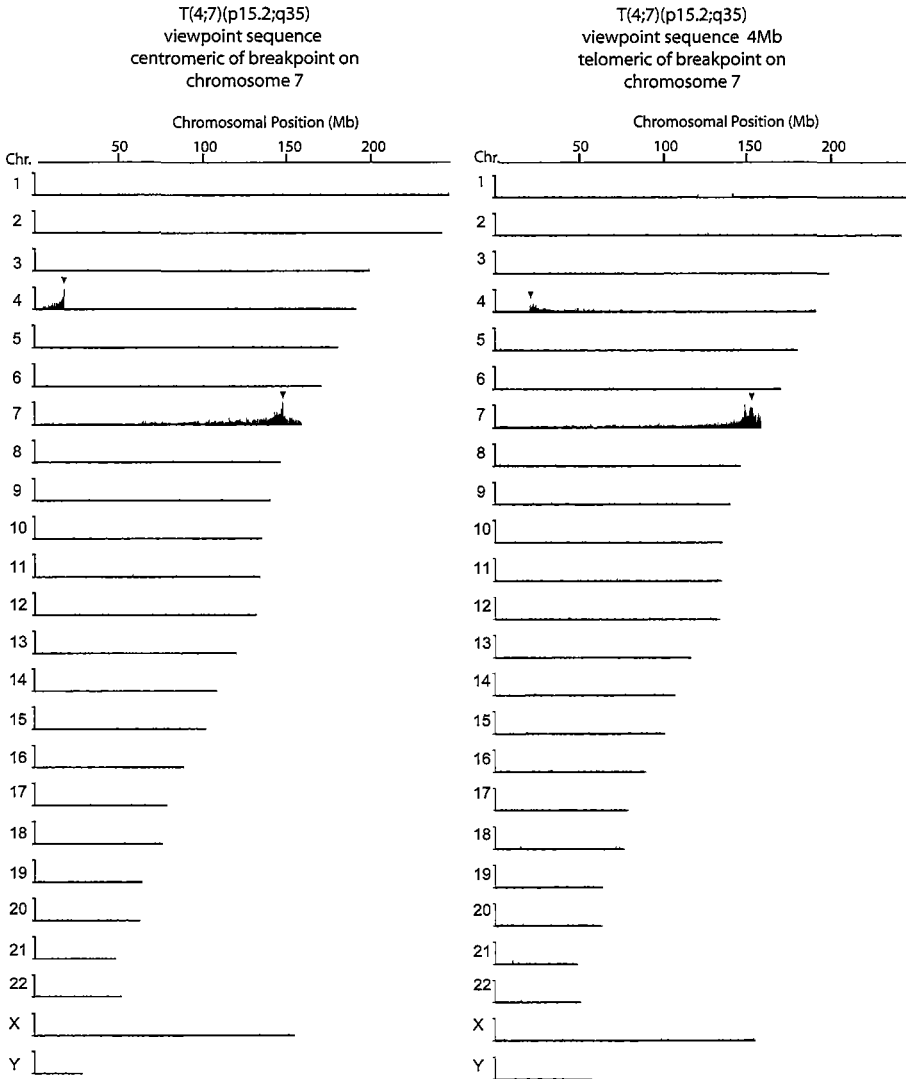
**Supplementary Figures**

**Supplementary Figure 6.1**



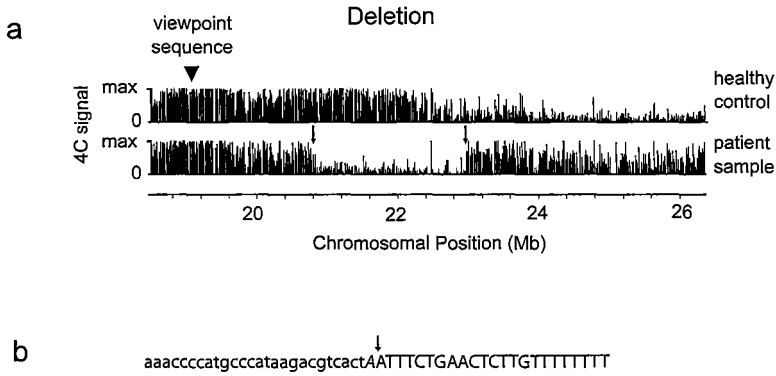
**Supplementary Figure 6.1. Restriction-fragment-paired-end-sequencing.** Schematic representation of restriction-fragment-paired-end-sequencing.

Supplementary Figure 6.2



**Supplementary Figure 6.2.** 4C signals across all chromosomes obtained with two different chromosome 7 viewpoint fragments in a sample carrying  $t(4;7)(p15.2;q35)$ . The black arrowheads indicate position of viewpoint sequences. The grey arrowheads indicate the position of the translocation sites. Running mean data were plotted, using a window size of ~60kb. Scale on Y-axis (arbitrary units) is identical for all chromosomes.

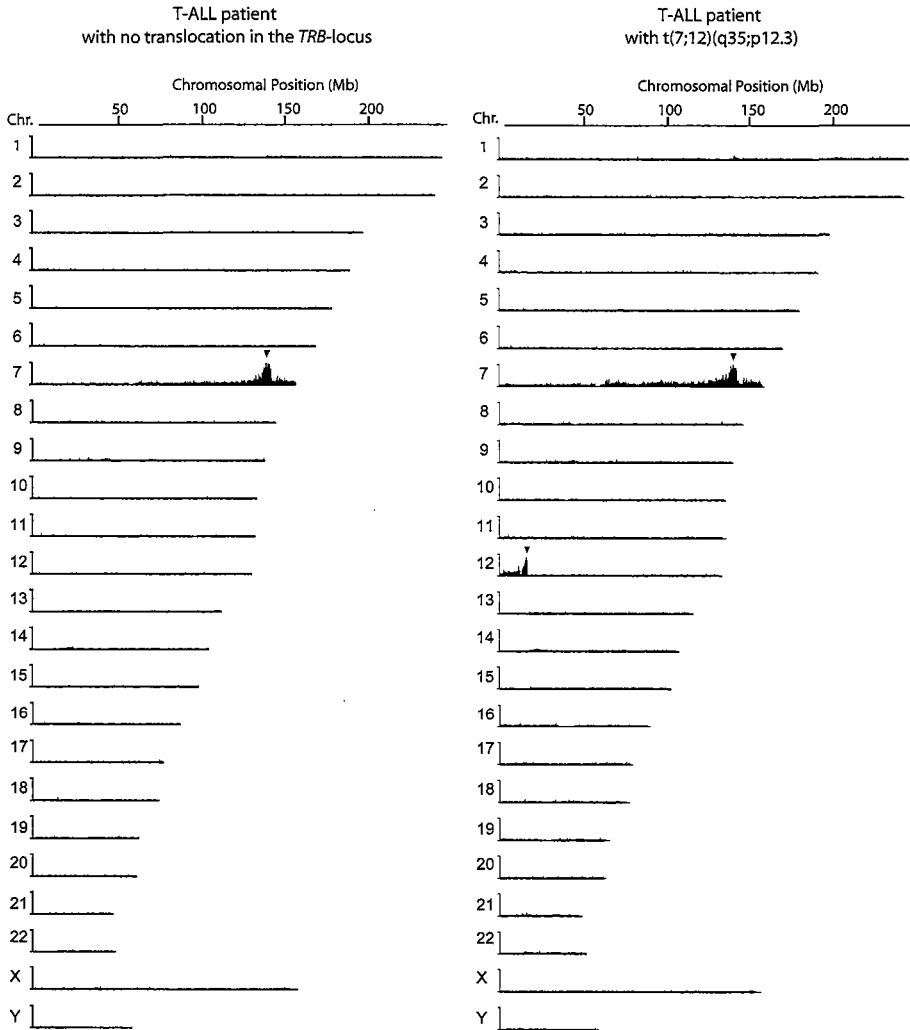
Supplementary Figure 6.3



**Supplementary Figure 6.3. 4C detects a homozygous deletion.** (a) A viewpoint fragment located on chromosome 9 at 19.3 Mb identifies a region (between arrows) lacking high signals in a T-ALL patient sample (bottom) compared to control (top), demonstrating a deletion. Signals downstream of the deletion are higher in patient versus control, because these sequences are closer to the viewpoint. The signals in the deleted area are higher than background. This is because both samples were hybridized to the same microarray. The strong Cy-3 fluorescence of the healthy control on the probes in the deleted area leaks through Cy-5 filter, resulting in elevated signals in the patient sample data. This can be prevented by hybridizing the control and the patient on separate microarrays. 4C can detect deletions, but array CGH is more powerful in detecting copy number changes. (b) Sequence identifying the breakpoints that flank the deletion.



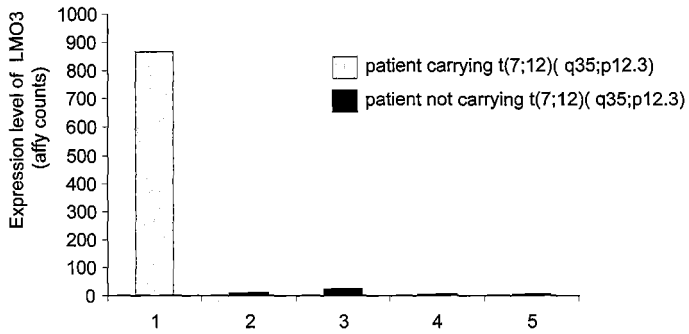
Supplementary Figure 6.4



**Supplementary Figure 6.4.** 4C signals across all chromosomes obtained with a viewpoint sequence near the *TCRB* locus on chromosome 7 in two T-ALL patient samples, one of which carrying a t(7:12) translocation. The black arrowheads indicate position of viewpoint sequences. The grey arrowhead indicates the position of the translocation site. Running mean data were plotted, using a window size of ~60kb. Scale on Y-axis (arbitrary units) is identical for all chromosomes.

Supplementary Figure 6.5

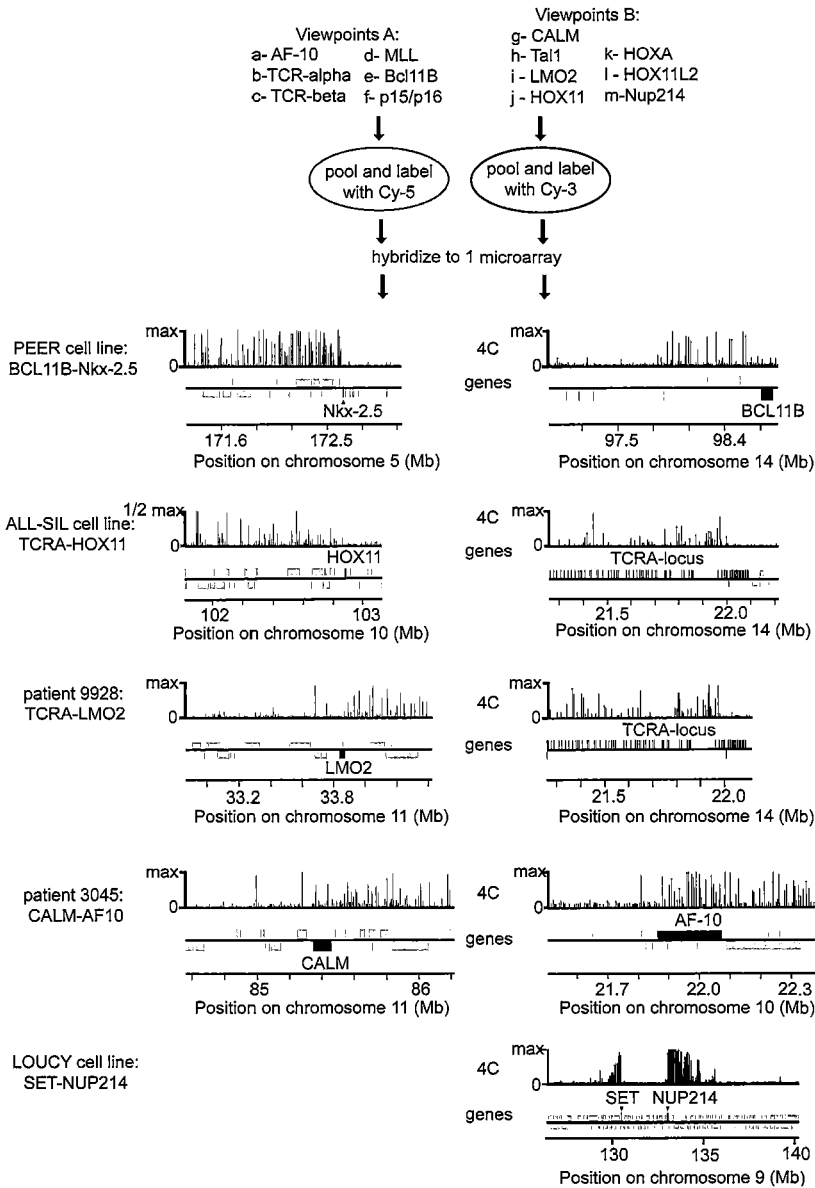
---



**Supplementary Figure 6.5. LMO3 expression in T-ALL patient samples.** Gene expression was measured on affymetrix gene expression arrays. LMO3 is expressed in the patient carrying t(7;12)(q35;p12.3), but not in the other patients.

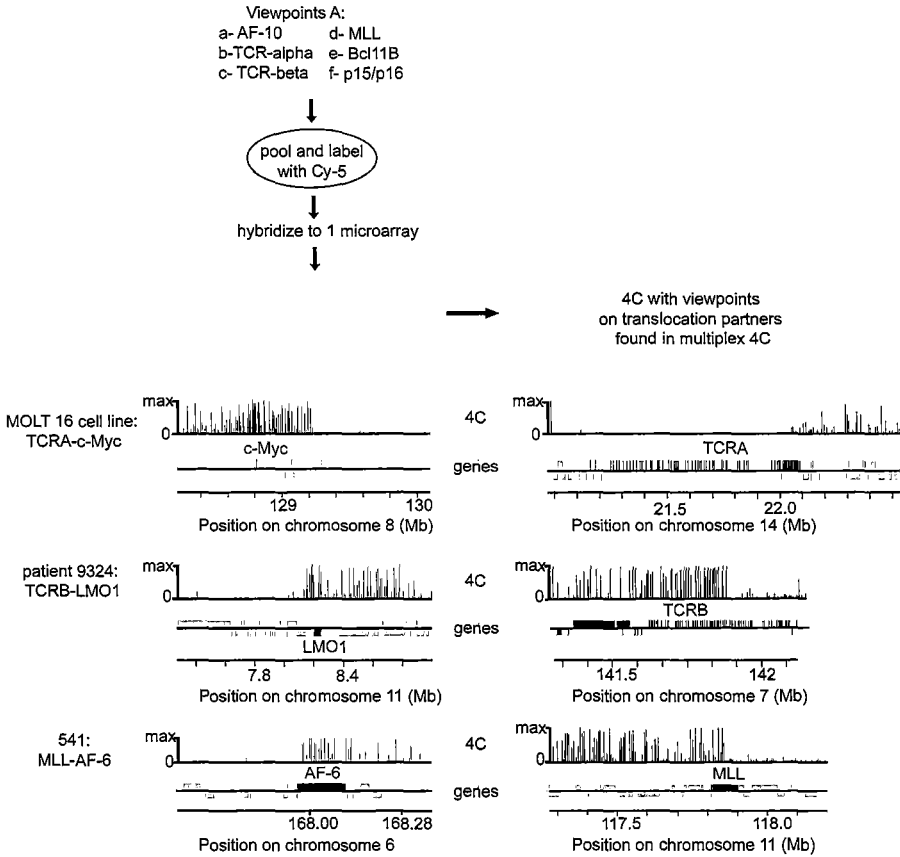
---

Supplementary Figure 6.6



**Supplementary Figure 6.6. Multi-view 4C experiments detect frequently occurring translocations and a common deletion.** 13 viewpoints located near frequent rearrangement sites were combined in one microarray experiment. This multi-view 4C was applied to different T-ALL patient samples. The translocations indicated and the common *SET-Nup214* deletion were detected.

Supplementary Figure 6.7



**Supplementary Figure 6.7. Multi-view 4C experiments detect uncharacterized and novel translocations.** In three multi-view 4C experiments only the data of the Cy-5 pool showed high signals outside the viewpoint areas. In these three cases the detected locus was not included in the Cy-3 pool. 4C using the newly detected translocation partner as a viewpoint subsequently revealed which of the viewpoints in the Cy-5 pool was connected to it.

# 7

## A pilot study for Chromosome Conformation Capture *Se*-Quencing (4C-Q)

Work in progress.

## A pilot study for Chromosome Conformation Capture Sequencing (4C-Q)

Marieke Simonis, Petra Klous, Frank Grosveld and Wouter de Laat

### **Summary**

*4C is a recently developed technique that can be used to study the folded structure of chromosomes. In addition, 4C can be applied to characterize genomic rearrangements within megabases of a selected viewpoint fragment. Here we adapt the microarray based 4C into sequencing based 4C (4C-Q). 4C-Q is designed such that multiple viewpoints can be analyzed in one experiment. In the future, 4C-Q could be extended to a genome wide method for the identification of genomic rearrangements. Moreover 4C-Q is expected to allow a more quantitative measurement of (functional) spatial interaction between genomic fragments than was achieved with 4C.*

### **Introduction**

Eukaryotic genomes are extensively folded to fit in the cell nucleus. The folded structure of chromosomes can be studied with a recently developed technique, Chromosome Conformation Capture on Chip (4C)<sup>1</sup>. In 4C, the chromatin structure in the nucleus is first fixed, then fractionated by restriction enzyme digestion and for a selected DNA fragment (the viewpoint) all the restriction fragments that are in spatial vicinity in the nucleus are captured and identified using a dedicated microarray.

The frequency with which different fragments are caught by the viewpoint can differ significantly. Segments of the genome directly adjacent to the viewpoint are per definition spatially close in all cells and thus captured most frequently. Preferential contacts made between distant regions, megabases away on the same chromosome or on other chromosomes, occur in a smaller subpopulation of the cells; ~5-15 % as measured in cryo-FISH experiments<sup>1</sup>. Consequently, fragments in these distant regions are captured less frequently than those adjacent to the viewpoint.

4C applied on the  $\beta$ -globin showed that the technology is well suited to measure the less frequent, distant contacts in the genome. The local folded structure of the  $\beta$ -globin locus had previously been analyzed in detail and it is known that the promoter of the active gene interacts with regulatory elements spread over 200 kb. In the 4C experiment

quantitative interpretation of these local interactions between functional elements close to the viewpoint was not possible, which was at least partially due to saturation of microarray signals<sup>1</sup>. Sequencing methods provide a more quantitative analysis than the microarray approach. A sequencing based 4C strategy is therefore possibly more suited for the detection of contacts between specific functional elements relatively close to each other on the genome (<1 Mb).

4C can also be used as a diagnostic method for the identification of rearrangements near selected genomic viewpoints (Chapter 6). This approach takes advantage of the fact that interactions with local DNA fragments are captured much more frequently than others. If as a consequence of rearrangements new DNA sequences are brought to a position on the genome close to an investigated viewpoint, they will be captured more frequently than expected and can easily be recognized in the 4C data. Genomic rearrangements (e.g. translocation and inversions) located up to several megabases away from the viewpoint can be identified with 4C. Detailed analysis of the breakpoints of genomic rearrangements is important, because it can lead to the identification of disease associated genes<sup>2</sup> and specific rearrangements were shown to have prognostic value<sup>3</sup>. The detection of (balanced) genomic rearrangements traditionally depended on laborious cytogenetic (banding) techniques that only provided limited resolution. 4C can detect balanced and complex rearrangements more rapidly and at high resolution. The disadvantage of 4C is that only rearrangements within a few megabases of selected viewpoints are found.

Here we adapted the strategy of 4C from a microarray based to a sequencing based approach to identify co-localizing sequences. The sequencing variant is named 4C-Q (Chromosome Conformation Capture *Se*-Quencing). The sequencing strategy may prove beneficial for the identification of specific interaction between functional elements on the genome. Moreover, the strategy allows the analysis of multiple sites of the genome, such that larger screens for genomic rearrangements are feasible, possibly even a genome wide screen.

## Results

### *Experimental adaptation of 4C*

The first steps of 4C are not altered for the sequencing based set-up (**Fig 7.1a**). 4C involves the cross-linking of co-localizing DNA, followed by digestion and subsequent ligation of cross-linked fragments. To identify the fragments that were ligated to the viewpoint fragment (i.e. the fragments that co-localized with it), ligation sites are trimmed and circularized, after which an inverse PCR on the viewpoint fragments is performed to

a

Fix chromatin structure with formaldehyde



Digest with a 6 base recognizing enzyme *Hind*III



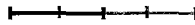
Dilute



Ligate cross-linked restriction fragments



Reverse cross-links



Trim ligation junction with 4-base recognizing enzyme *Dpn*II



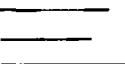
Circularize



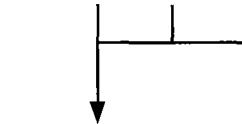
PCR amplify with primers on viewpoint extended with sequencing adaptors



PCR product contains all sequences that were ligated to the viewpoint fragment



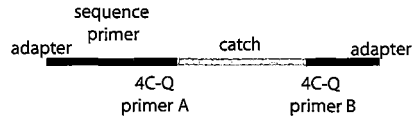
Pool viewpoints



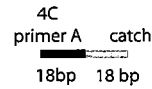
Massive parallel sequencing

b

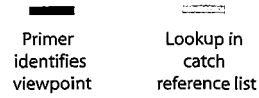
PCR



Sequencing



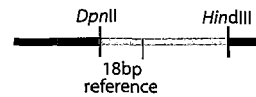
Split sequence



Analyze catches for each viewpoint

c

Catch sequence





amplify all the ligation partners (“catches”) in the sample simultaneously. Previously these catches were analyzed on a microarray. Here, massive parallel sequencing (illumina) is applied to characterize the sequences that co-localize with the viewpoint fragment (**Fig. 7.1a**).

In illumina sequencing technology, adapters are attached to the DNA fragments that are analyzed, a different adapter on each side. The sequences are then hybridized to a glass slide that contains oligos complementary to the adapter probes. The sequences will form bridges as each end hybridizes to a different oligo. Subsequently a PCR is performed using the adapter sequences as primers, such that each bridged DNA fragment becomes a spot containing many copies of the same sequence. The spots are sequenced, using a primer that was coded in one of the adapters (**Fig. 7.1b**).

In our strategy, the adapters are not ligated to the DNA fragments, as is customary, but incorporated in the primers used in the 4C-Q PCR reaction (**Fig. 7.1a, b**). One of the primers thus consists of adaptor sequence A, the sequencing primer and 4C primer A (from 5’ to 3’). The other primer consists of adaptor sequence B and 4C primer B. Due to this design each DNA fragment amplified in the PCR will be flanked by the adapters and can be sequenced. Notably, background genetic material that is not amplified in the 4C-Q PCR reaction but is still abundantly present in the sample will not carry the adapters and should therefore not be found back in the sequencing data. The long 4C-Q primers amplified the input material with approximately the same efficiency as primers without attached adapter sequences (**Supplementary Fig. 7.1**).

The analysis of the generated sequencing data is different from many other massive parallel sequencing experiments<sup>4</sup>. The 36 bp 4C-Q sequences will consist of two defined parts (**Fig. 7.1b**). The first 18 sequenced bases should always be 4C primer A, because this sequence is directly attached to the sequencing primer. Importantly, primer A also serves as a barcode for the viewpoint, allowing 4C-Q PCR products of different viewpoints to be mixed in one experiment.

---

**Figure 7.1. Schematic representation of 4C-Q.** (a) 4C-Q sample preparation. For each viewpoint, fragments that co-localize with it in the nucleus (catches) are amplified in one PCR reaction. PCR products of different viewpoints are pooled and sequenced (see text). (b) Schematic overview of the 4C-Q data analysis. PCR products all have the sequencing adapters at their outer ends. The sequencing primer is also incorporated. Sequencing of these products results in 36 bp sequences consisting of a 4C-Q primer sequence, which is a barcode for the viewpoint, and 18 bp of the catch. (c) The 18 bp that will be sequenced can be extracted from the genome sequence. This is done for all possible catches in the genome to create a catch reference list. The 4C-Q catches are compared to this list to determine which genomic fragment they derived from.

---

4C-Q primer A is designed on top of the restriction site of the viewpoint fragment that is analyzed and will therefore only amplify the catches and not any sequence of the viewpoint fragment itself. Thus, the second part of the 4C-Q sequence reads always consists of the first 18 bp of one of the catches of the viewpoint (**Fig 7.1c**). The catches always start with the restriction site used in the preparation, thus for each possible catch the expected 18 bp of sequence that will be read can be extracted from the genome sequence. These 18 bp sequences are references to the catches and can be stored in a relatively simple, local database. Each sequence read can be identified by direct comparison to this database, making extensive blast analysis unnecessary. Summarizing, all sequence reads can be split into a primer, which serves as a barcode for the viewpoint, and 18 bp of a catch which can be identified by looking it up in the catch-reference database.

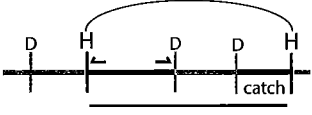
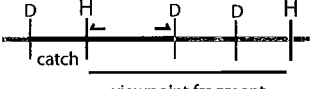
#### *Identity of the 4C-Q sequence reads*

To test the feasibility of 4C-Q, a pilot experiment was performed in which three viewpoints on chromosome 7 were analyzed in the human leukemia derived HSB-2 cell line. The PCR products of the three viewpoints were pooled and analyzed by massive parallel sequencing in a single illumina lane, which resulted in 4.9 million successful sequence reads (**Fig. 7.2**). Of 93% of the sequences the first 18 sequenced basepairs were identical to one of the 4C-Q primer sequences as expected. Most of the remaining 7% of sequences were similar to one of the 4C-Q primers but carried a mismatch likely as a consequence of sequencing mistakes. 1-2 million sequence reads were obtained for each viewpoint. The second 18 base pairs were identical to one of the catch reference in about half of the sequences. The sequences that were unexpectedly not identical to one of the catch reference could have resulted from sequencing errors, SNPs, random events in the second ligation step or other sources of background, such as unspecific annealing of the PCR primers.

The data show that the experimental set-up of 4C-Q results in the expected type of sequence reads. The number of sequence reads that contains a perfect match to a catch reference is already high, but can possibly be optimized.

#### *4C-Q data distribution*

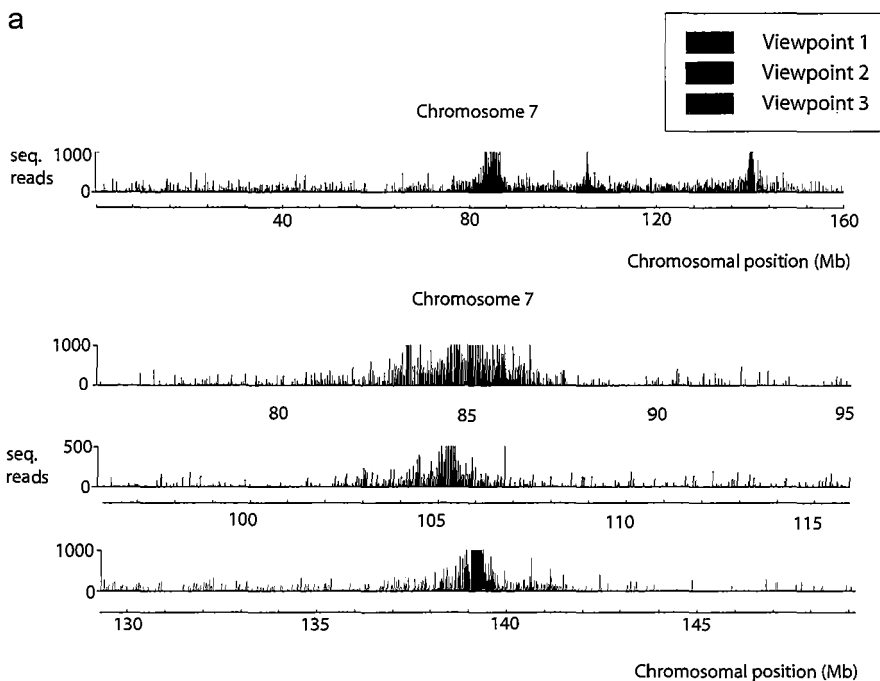
The sequence reads that matched a catch reference were analyzed further for each viewpoint. The genomic areas directly surrounding the viewpoints are clearly

	4.9 *10e6 sequences in the total data set								
	4.6 * 10e6 sequences start with one of the expected primer sequences (93%)								
	viewpoint 1			viewpoint 2:			viewpoint 3:		
total number of sequences	2.2 *10e6			1.2 *10e6			1.1 *10e6		
perfect matches to catch reference list	match to unique	match to non- unique	no match	match to unique	match to non- unique	no match	match to unique	match to non- unique	no match
	9.7 *10e5 (43%)	1.3 *10e5 (5.6%)	1.1 *10e6 (51%)	3.0 *10e5 (24%)	3.6 *10e5 (29%)	5.7 *10e5 (47%)	4.0 *10e5 (37%)	1.2 *10e5 (11%)	5.6 *10e5 (52%)
number of different catches	8.2 *10e3	4.7 *10e3	5.0 *10e5	6.0 *10e3	3.8 *10e3	2.7 *10e5	8.7 *10e3	4.0 *10e3	2.9 *10e5
viewpoint fragment self-ligation	2.0 *10e5 (18% of matched catches)			1.1 *10e5 (17% of matched catches)				0.75 *10e2 (0.1% of matched catches)	
									
viewpoint restriction site undigested	2.3 *10e5 (21% of matched catches)				2.3 *10e5 (35% of matched catches)		8.1 *10e3 (1.6% of matched catches)		
									

**Figure 7.2. Characterization of the 4C-Q sequences of three pooled viewpoints.** 4.9 mln sequences were obtained by illumina sequencing. Primer sequences and catch sequences were identified as represented in the diagram. Two catches were found to be most abundant in previous 4C experiments, the circularised and the undigested viewpoint (Chapter 4, 5 and 6). The abundance of these catches was also high in the sequencing data, except for viewpoint 3.

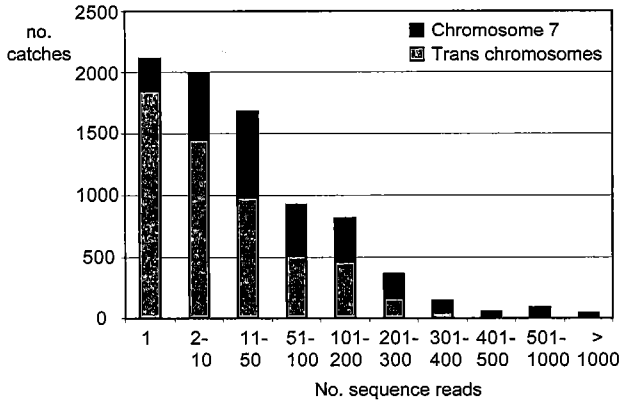
distinguished in the data by their high number of catches, as expected based on previous 4C analyses<sup>1</sup> (**Fig. 7.3a**).

In 4C and 4C-Q experiments, two catches are always most prominent when the PCR products are analyzed on an agarose gel<sup>1,5</sup>. One of them is due to incomplete digestion of the analyzed *Hind*III site of the viewpoint and the second derives from self-ligation of the analyzed *Hind*III fragment (**Fig. 7.2**). Indeed these products were found frequently in the catches, together constituting 39% and 42% of the total amount of identified catch references of viewpoint 1 and 2 respectively (**Fig 7.2**). However, in the sequence reads of viewpoint 3 they made up only 1.7 % of the total. The catch reference of the product of the *Hind*III fragment self-ligation was found only 753 times (0.1 % of all viewpoint 3 sequences). Possibly a SNP destroyed the *Hind*III site of the catch or properties of the chromatin fiber prevented circularization. The catch reference of the undigested viewpoint

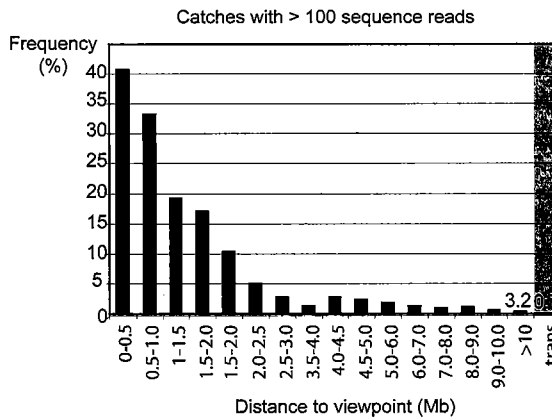
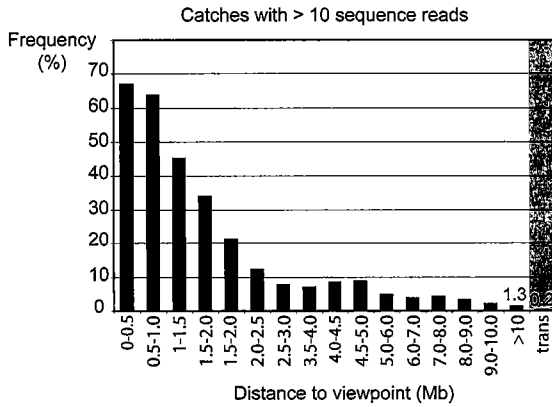


**Figure 7.3. Distribution of 4C-Q data.** (a) The viewpoints are distinguished in the data, by their high density in catches with many sequence reads, as expected from previous 4C data. (b) Most catches that are found are identified in 1-200 sequence reads. Catches on chromosome 7 are overrepresented, especially in the categories with a high amount of sequence reads. (c) The percentage of all possible catches that is found is high in the area close to the viewpoint. Even more than 10 Mb away from the viewpoint more different catches are found than on average on trans chromosomes.

b



c



restriction site only constituted 1.6 % of the identified references. Possibly, this number is so low because the size of the fragment that is amplified in the PCR reaction for this catch is 1700 bp, while the median size of the fragments that are PCR amplified is 256 bp. We hypothesize that the product is selected against due its large size, which may compromise PCR amplification or cluster formation on the illumina slide.

Previous 4C microarray analysis only allowed minimal quantitative data analysis. To gain insight in the distribution of the sequenced reads, viewpoint 1, which had the highest number of successful reads, was analyzed in more detail. The  $1.1 * 10^6$  sequence reads matched to 8208 unique catches (**Fig. 7.2**). A third of the unique catches (2715) were located on chromosome 7, which is more than expected based on a random distribution of the data, because chromosome 7 only constitutes ~5% of the genome. This overrepresentation of catches in cis was also seen in previous 4C analysis and can be explained by the finding that chromosomes are located in distinct territories in the nucleus<sup>6</sup>.

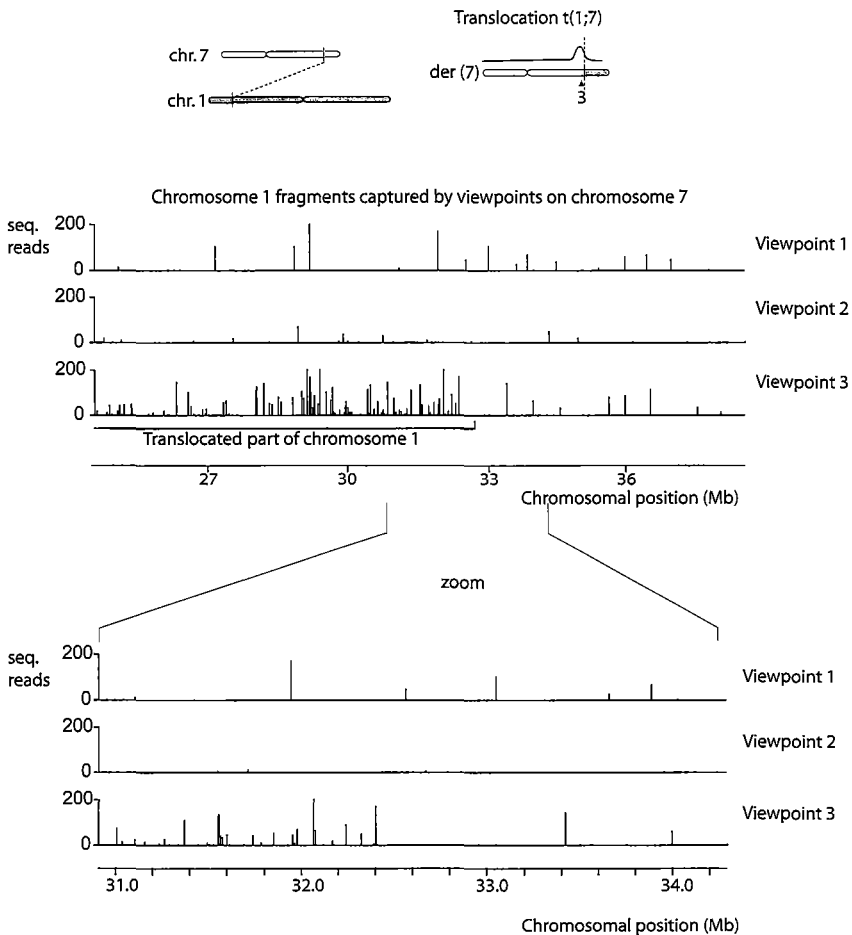
The number of times each unique catch was found in the sequence reads ranges from 1 to 227,919. The overrepresentation of sequences on chromosome 7 is most obvious for catches with a high number of sequence reads (**Fig. 7.3b**). Only the two most frequently captured fragments, the self-ligation and the incomplete digestion of the viewpoint, were found more than 200,000 times. The remainder of catches had a much lower number of sequence reads. The 36 unique sequences that were found between 1000 and 10,000 times all located within 2Mb of the viewpoint (**Fig. 7.3a**). Some, but not all represented direct neighbors. The more distantly located, high frequent catches may have resulted from specific folded structures in the genomic region. Fragments on other chromosomes than chromosome 7 (trans) that were captured by a viewpoint had only 1 to 550 sequence reads (**Fig. 7.3b**).

The proportion of fragments captured per genomic segment is also highest in the area surrounding the viewpoint (**Fig. 7.3c**). The data suggest that if a genomic rearrangement occurs within 2.5 Mb of the viewpoint, it would be identified by this strategy (**Fig. 7.3c**).

#### *A translocation is detected with 4C-Q*

The HSB-2 cell line that is analyzed here carries a translocation t(1;7). Due to this rearrangement a part of chromosome 1 was positioned 2.5 Mb telomeric of viewpoint 3 on chromosome 7. Consequently, viewpoint 3 had captured many fragments in this part of chromosome 1 (**Fig. 7.4**). Each viewpoint has catches on all chromosomes, but the translocated area of chromosome 1 is clearly different, because the density of captured

fragments is much higher (**Fig. 7.4**). The position of the breakpoint could be estimated with an accuracy of  $\sim 100\text{kb}$ . The density of the captures resulting from genomic proximity to viewpoint 3 determines the resolution at which the translocation is detected. Thus, the resolution can be improved by increasing the number of different captures in the genomic area of the breakpoint. This can be achieved by decreasing the distance between viewpoint and breakpoint, which was relatively large in this experiment (2.5 Mb). Closer to the viewpoint more fragments are captured (chapter 6). A second factor that will influence



**Figure 7.4 A translocation located 2.5 Mb from viewpoint 3 is detected.** Viewpoint 3 has multiple frequent catches on chromosome one, in contrast to viewpoint 1 and 2. These catches were brought into spatial vicinity of viewpoint 3 due to the t1:7 translocation that is present in the analysed HSB-2 cells. The breakpoint on chromosome one is indicated by the arrowhead. Viewpoint 3 was located 2.5 Mb from the breakpoint on chromosome 7.

the resolution is the number of cells that is analyzed. The more cells are analyzed, the more different catches are included in the sample, which enlarges the density of captured fragments and the distance over which catches are found. Here around 100,000 cells were analyzed, five times less than in previous experiments (chapter 6).

The data show that translocations can be detected using 4C-Q, even at a distance of 2.5 Mb from the viewpoint.

### Discussion

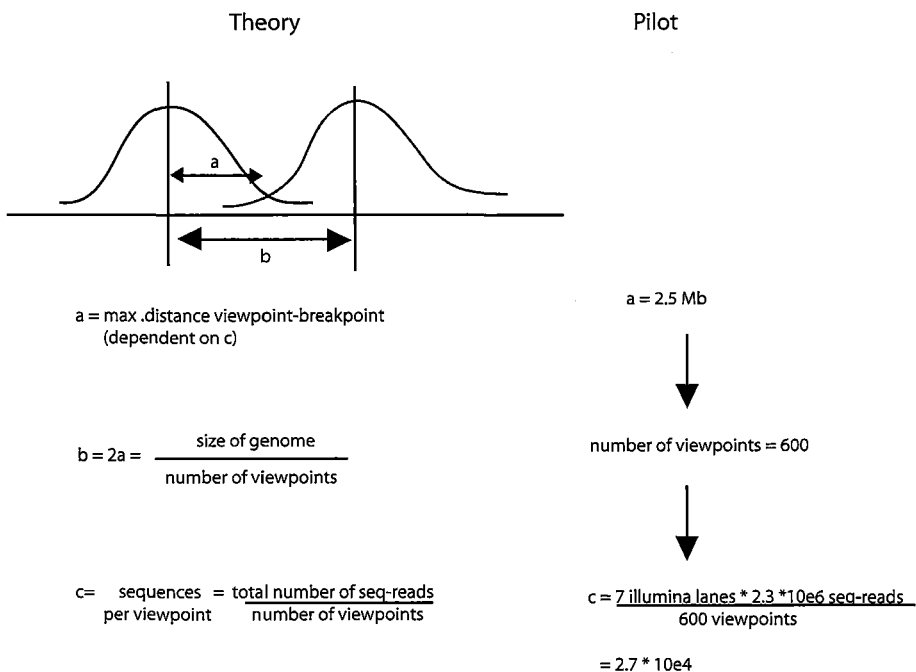
4C was successfully adapted from a microarray based technique to the sequencing based method, 4C-Q. In the 4C-Q pilot experiment described here, the distribution of the data was as expected. 93% of the sequences contained one of the 4C-Q primers and the areas surrounding the viewpoints contained the most frequently found catches and had the highest density of catches. Differences were seen in the number of sequence reads found per catch, also in the region directly neighboring the viewpoints. However, the capability of 4C-Q to detect specific spatial interactions between functional elements can not be assessed from the current data. Future 4C-Q analysis of a gene locus with a well described structure, for example *β-globin*, will be informative in this respect.

Several aspects of 4C-Q can be optimized to obtain more informative catches per experiment. Massive parallel sequencing is a novel technique and it is likely that the technical aspects of the method will improve, resulting in less sequencing errors. The experimental design of 4C-Q can also be improved. We found that the two most frequent catches (the *HindIII* fragment self-ligation and the non-digested *HindIII* site) represented 40% of the total amount of unique catches in the data of two of the viewpoints. These catches are non-informative. If these frequent catches could be filtered out, before the sequencing analysis, the proportion of informative data would increase. The results of viewpoint 3 suggested that the abundant products can be selected against if the DNA stretch that is amplified in the 4C-Q PCR is relatively large. Thus, using viewpoints of which the two most frequent catches have long PCR fragments may result in a reduction of these catch products in the sequence data.

The pilot study demonstrated that multiple viewpoints can be pooled and analyzed in one experiment. Is a genome-wide 4C-Q scan for genomic rearrangements feasible? In a genome wide 4C-Q experiment all the chromosomes would be covered with viewpoints, which combined would detect all the large structural rearrangements in the genome. In the pilot experiment described here, a translocation was detected at 2.5 Mb from the viewpoint, based on  $5.2 \cdot 10^5$  identified catches (including non-unique catches). If the



pilot would be extended by simply pooling more viewpoints the number of sequences obtained for each viewpoint would be lower. If a complete illumina sequence run of 7 lanes is used  $2.7 \cdot 10^4$  sequence reads would be found for each viewpoint, (see **Fig. 7.5** for calculation), which is only a twenty fold reduction. Estimating from the data distribution of viewpoint 1 (**Fig. 7.3c**), rearrangements within 2.5 Mb of the viewpoint would still be detected. Thus, theoretically a genome wide 4C-Q scan in one illumina sequencing experiment is achievable.



**Figure 7.5. The feasibility of a genome-wide 4C-Q scan for genomic rearrangements.** Important for the feasibility of 4C-Q is the number of sequences that is needed per viewpoint (c) to detect rearrangements at least as far as half the distance to the next viewpoint (a). This number depends on the average distance between two given viewpoints (b), and related to that, the total number of viewpoints on the genome. The translocation detected with viewpoint 1 was located 2.5 Mb from the viewpoint ( $a = 2.5$  Mb). If the genome is covered with viewpoints spaced on average 5 Mb, 600 viewpoints are needed. The current pilot experiment resulted in  $2.3 \cdot 10^6$  sequence reads with an identified 4C-Q primer and a match to a catch reference. The pilot was performed in one illumina sequencing lane. A complete illumina sequencing analysis consists of 7 lanes. If 7 illumina lanes are filled with 600 viewpoints,  $2.7 \cdot 10^4$  sequences would be obtained per viewpoint. Based on the distribution of the sequencing reads (Fig. 7.3c), this 20 fold reduction (or 40 fold for viewpoint 1) in sequence reads could still allow detection of the translocation.

To extend 4C-Q to a genome-wide method, other aspects need to be tested and optimized as well. Here, we used an equivalent of 100,000 cells as input for each viewpoint. For a genome wide analysis 60 mln cells would be needed. Setting up the PCRs in a multiplex format would decrease this number substantially. It needs to be assessed if it is feasible to obtain a comparable number of sequences for each viewpoint, if many viewpoints are mixed. The data processing will also require some adaptation to allow a comprehensible analysis and visualization of 600 viewpoints.

If all obstacles are overcome, 4C-Q is a valuable new tool to search for structural rearrangements. Other genomic techniques exist that can detect structural alterations in genomes, most notably, paired end sequencing approaches<sup>4,7</sup>. In these methods the genome is fractionated into fragments with limited variation in size and the outer ends of the fragments are sequenced. If both ends can be mapped and their relative position is not as expected, this indicates a rearrangement is present. This method allows the detection of a variety of structural changes at high resolution. However the detection critically relies on successful sequencing of the single fragment that contains the breakpoint of the rearrangement.

4C/4C-Q is only suited to detect large structural rearrangements, but may well be more robust than paired end sequencing, because it does not rely on the identification of the single informative DNA fragment. Instead, rearrangements are identified by multiple catches across the breakpoint. In addition, 4C has been shown to be able to detect rearrangements even when present in only a small subpopulation of the cells (chapter 6). 4C-Q would therefore also be suited to characterize changes in heterogeneous samples.

### **Acknowledgements**

We would like to thank Yavuz Ariyurek, Harmen van de Werken and the department of Biomics of the Erasmus MC for assistance.

### **Materials and methods**

#### *Sample preparation*

4C preparation of the HSB-2 cells and 4C-Q PCR conditions were as described in Chapter 4. Three PCR reactions on 200 ng 4C template each were performed for each viewpoint. The

three PCR products of each viewpoint were pooled and purified using the GFX PCR DNA and Gel Band Purification kit (GE Healthcare) to remove unused primers. The concentration of the purified samples was measured using a NanoDrop spectrophotometer. Equal amounts of the products of the three viewpoints were pooled and sequenced in one lane of the illumina sequencing system. Sequences of the adapters are available on request. The sequences of the 4C-Q primers were:

Viewpoint 1-primer A	5'-ATGTGACTCCTCTAGATC-3'
Viewpoint 1-primer B	5'-CCCTGAACCTCTTGAAGCT-3'
Viewpoint 2-primer A	5'-CGGCCTCCAATTGTGATC-3'
Viewpoint 2-primer B	5'-GAATTGCTTTTGGTAAGCTT-3'
Viewpoint 3-primer A	5'-TTTTAGCCCTGACAGATC-3'
Viewpoint 3-primer B	5'-AGTCAAACATAAGCCTAAGC-3'

#### *Data analysis*

A reference catch dataset was created containing the expected 18 bp sequences (see text and **Fig. 7.1c**) from the human genome, NCBI build 37. The reference catches were split into unique and non-unique catches. Unique was defined here as occurring once in the total list of catches, not as occurring once in the whole genome. Of the total 1.46 mln catches on the genome, 1.17 mln were unique. The reference table consisted of the unique reference catches and their genomic positions.

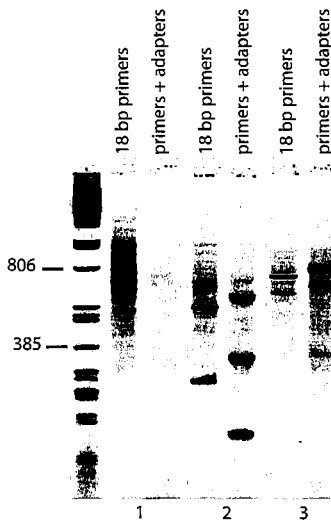
The illumina sequencing resulted in  $4.9 * 10^6$  sequences in total. The data-sheet consisted of a list of unique sequences and the number of times each was found (counts or sequence reads). The sequences were split into two 18 bp sequences. Sequences that contained one of the primer A sequences in the first 18 bp were extracted and sequences for each viewpoint were treated separately in downstream analysis. The catches of each viewpoint were compared to the reference catch list and for perfect matches the genomic position of the reference was linked to the frequency with which the catch was found

in the sequencing data (counts). Subsequently the counts of all the found catches were plotted on the genome, using the SignalMap programme obtained from Nimblegen. Data processing was performed using perl scripts, available on request.

## References

- 1 M. Simonis, P. Klous, E. Splinter et al., *Nature genetics* **38** (11), 1348 (2006).
- 2 J. K. Millar, S. Christie, C. A. Semple et al., *Genomics* **67** (1), 69 (2000)
- 3 A. A. Ferrando, D. S. Neuberg, R. K. Dodge et al., *Lancet* **363** (9408), 535 (2004).
- 4 J. O. Korb, A. E. Urban, J. P. Affourtit et al., *Science (New York, N.Y)* **318** (5849), 420 (2007).
- 5 M. Simonis, J. Kooren, and W. de Laat, *Nature methods* **4** (11), 895 (2007).
- 6 P. Lichter, T. Cremer, J. Borden et al., *Human genetics* **80** (3), 224 (1988).
- 7 E. Tuzun, A. J. Sharp, J. A. Bailey et al., *Nature genetics* **37** (7), 727 (2005).

Supplementary Figure 7.1



**Supplementary Figure 7.1. The efficiency of the 4C-Q PCR reaction is not affected by the addition of adapter sequences to the primers.** 4C-Q PCR using standard sized primers (18 bp) resulted in the same amount of product as a PCR with primers elongated with adapter sequences (total primer length 71 bp and 40 bp) analysed on a 2 % agarose gel. Due to the incorporation of the adapters each amplified sequence is longer and runs higher in the agarose gel.

# 8

## General discussion and future directions

## General discussion and future directions

Eukaryotic genomes are extensively folded to fit into the small volume of the cell nucleus. Despite this compacted structure of the DNA the right set of genes is transcribed in every healthy cell. The spatial organization of the genome has been linked to gene expression in different ways. Fluorescent in situ hybridization (FISH) experiments have shown that some genes move to a different position in the nucleus when their expression status changes<sup>1</sup>. In these studies the nuclear location of gene loci is described relative to large nuclear structures or to other selected loci.

In a different type of experiments, the local folded structure of small parts of the genome around specific genes has been analyzed in detail and has been correlated with gene expression. For this type of analyses Chromosome Conformation Capture (3C)<sup>2,3</sup> is used, but this technique can not be applied on a genome-wide scale.

We developed 3C-on-chip (4C)<sup>4</sup>, which allows an unbiased genome wide search for DNA sequences that are found in spatial proximity of a selected part of the genome. 4C can thus be used to describe the position of a gene locus in the nucleus. In contrast to FISH experiments, this nuclear location is not described relative to one other element in the nucleus. Instead, all parts of the genome found in spatial proximity to the investigated locus are identified.

### 4C versus 3C

Although 4C is derived from 3C, the data that is produced by the technology and the information that can be obtained with it is very different, which can sometimes lead to confusion. Technical details of 3C, 4C and other methods derived from 3C were discussed in detail in chapter 5<sup>5</sup>. The main differences between 3C and 4C are caused by additional processing steps performed in 4C and the method of quantification.

In 3C technology nuclei are fixed such that elements in the genome spatially close to each other are cross-linked. The chromatin is subsequently digested, diluted and ligated, such that co-localizing fragments will become ligated to each other. The number of times two fragments form a ligation product is a measure for their interaction frequency and is analyzed by quantitative PCR across the ligation junctions, using a specific primer for each fragment.

The disadvantage of 3C is that each restriction fragment that is measured requires a specific primer and the quantification of each different ligation product requires a separate PCR (and control PCRs). The enzymes generally used in 3C generate fragments of around 4 kb. Thus, a gene locus of 200 kb contains about 50 restriction sites and digestion results in 100 restriction fragment ends. When the entire gene locus is analyzed the number of PCRs that needs to be performed is very large. Therefore 3C is mainly suited for gene loci of small size or well described gene loci that allow an educated guess about which sites should be analyzed. For example in the  $\beta$ -globin locus the position of the genes and the main regulatory elements was known, thus a 3C experiment was achievable. 3C experiments on this locus showed that the regulatory elements spread over 200 kb frequently looped towards the promoter of the active gene<sup>3</sup>.

To overcome the need for an independent PCR reaction for each ligation product, 3C was adapted into 4C (3C on Chip). In 4C one restriction fragment is selected (the "bait" or "viewpoint") and all the sequences that were ligated this viewpoint fragment are characterized in one experiment. This is achieved by circularization of the ligated fragments after which an inverse PCR with viewpoint specific primers amplifies all fragments ligated to the viewpoint. To allow an efficient PCR the ligation products are shortened before circularization in an additional restriction step. The 4C PCR product contains all the fragments that co-localized with the viewpoint fragment and is analyzed on a dedicated microarray. Each probe on the microarray measures the frequency with which one specific restriction fragment was ligated to the viewpoint.

The first 4C experiment was performed on the  $\beta$ -globin locus (chapter 4). Both a gene promoter and a regulatory element were analyzed, but the loops measured in the 3C experiments could not be found back in the 4C data. Interactions between two small sequence elements such as a promoter and an enhancer only result in an increased ligation frequency of one or a couple of restriction fragments. The enrichment of these interacting fragments is about 5-8 fold over non interacting fragments. In 4C the measurement made for each individual restriction fragment is less quantitative than it is in a well performed 3C analysis and as a consequence these specific interactions are missed. This loss in quantitative accuracy is likely due to PCR amplification biases and limited detection range of the microarray. In 4C, co-localization is measured by the number of different restriction fragments in a genomic area that was ligated to the selected fragment, as judged by the high microarray signal on the fragments. Thus, whereas 3C is used to study the local folded structure of DNA, 4C analyzes the position of a gene locus in the nucleus, by finding all other parts of the genome that are frequently in its spatial

surroundings. Generally many genomic regions of ~50 kb and larger that co-localize with the viewpoint are detected.

### **Studying nuclear organization with 4C**

A detailed discussion on studies addressing the spatial organization of gene expression and the impact of 4C is given in chapter 3 of this thesis<sup>1</sup>.

Using 4C we showed that the active murine  $\beta$ -globin locus in fetal liver tissue co-localizes with other transcriptionally active regions, mainly located on the same chromosome (chapter 4)<sup>4</sup>. In contrast, in a tissue in which the same locus is inactive it co-localizes with completely different areas on the same chromosome, containing gene poor and transcriptionally inactive regions. *Rad23a*, located in a gene rich region on chromosome 8, largely has the same nuclear environment in two analyzed tissues in which it is transcriptionally active. The sequences surrounding *Rad23a* in the nucleus consist of gene dense and transcriptionally active regions on chromosome 8 and also on other chromosomes.

The results demonstrate that chromosomes do not fold randomly, but that genomic regions have many preferred co-localization partners, which consist of chromatin that has the same gene density and transcriptional activity. What causes this co-localization and is it functionally relevant?

### *Discriminating factors of gene locus positioning*

In the regions that co-localize with the active  $\beta$ -globin locus and those that co-localize with *Rad23a* not every gene is actively transcribed. This suggests that the nuclear environment alone is not sufficient to determine expression levels and that the nuclear position of a gene is largely influenced by neighboring parts of the genome. Both conclusions were also drawn from localization studies of other loci, as discussed in chapter 3<sup>1</sup>. It could be that the inactive genes are identified due to a lack of resolution of 4C, therefore it would be interesting to investigate the nuclear environment of a gene that is alternatively expressed between tissues and is located in a gene dense, active area of the genome.

The position of the  $\beta$ -globin locus is not determined by the neighboring parts of the genome. The genes are located within olfactory receptor cluster that is silenced in both tissues investigated with 4C. In fetal liver cells the active state of the ~200 kb  $\beta$ -globin locus is dominant over the flanking silent chromatin in determining the position of the locus. This dominance could be the result of the size of the locus, although the 200 kb does not exist entirely of active chromatin. The high transcription level of the  $\beta$ -globin



genes could also play a role, the genes have one of the highest expression levels measured in this cell type.  $\beta$ -globin genes are tissue specific genes and it could be that their regulation requires special mechanisms that include repositioning in the nucleus. Lastly, it is possible that transcriptional activity is dominant over a silent chromatin status in determining the position a gene locus. In support of this the parts of the genome that co-localize with the  $\beta$ -globin locus in its inactive state contain almost no active genes.

To understand the properties that dictate nuclear positioning different types of gene loci need to be analyzed with 4C. Genes with different expression levels and different function, tissue specific and ubiquitously expressed genes, should be analyzed. The genes should also be selected based on differences in gene density of flanking DNA.

The 4C data were compared mainly to gene density and gene activity, but other genome characteristics could contribute to 3D organization as well. 4C data could be aligned with genome wide measurements of chromatin components, such as specific histone modifications, DNA methylation and chromatin proteins like HP1.

#### *Functional relevance of the nuclear environment*

Identifying discriminating factors in nuclear positioning is important, but it will not necessarily reveal how a position is established nor will it show the functionality of nuclear repositioning. How is nuclear architecture established? During mitosis chromosomes are condensed and in early G1 of the cell cycle they unfold and each part of the genome moves to a favored position. In G1 the chromatin is most dynamic and it is likely that the main organization of the nucleus is established at this stage. Roughly two theories exist on what determines nuclear organization. The first suggests that co-localization is the result of affinities between different parts of the chromatin fiber and the stochastic process of chromosome unfolding. This can be considered a process of self-organization<sup>6</sup>. In this model active transcription could dictate nuclear organization by the decondensation of chromatin that occurs in the process<sup>7</sup>. Contrasting models suggest that the folding of chromosomes is dictated by specific contacts made in the DNA strands. One theory is that specific parts of the genome are anchored to an underlying structure, the nuclear matrix<sup>8,9</sup>. It has also been suggested that local accumulations of RNA polymerase II (RNA polII), called transcription factories, determine the shape of the genome<sup>10</sup>. In addition, organization could be dominated by specific contacts made between structural proteins that create loops in the DNA, of which the main candidate is CCTC-binding factor (CTCF). This protein was recently found to co-localize with cohesin, a component that supports the structure of mitotic chromosomes<sup>11,12</sup>. This CTCF-cohesin relationship is interesting,

because it could be the basis of a regulated decondensation after mitosis. Moreover, CTCF is involved in the regulation of gene expression levels, thus, it possibly links nuclear organization to transcription.

Hybrids of all theories are of course also possible.

Our 4C experiments support the theory of self organization, because the regions that co-localize have similar chromatin compositions, as estimated from their gene density and gene activity. Each analyzed locus has a large number of genomic regions it frequently contacts, which could not physically contact each other all at once. This suggests the genome can be folded in multiple preferred configurations.

4C data are an average of around half a million cells. To gain insight in cell to cell differences FISH experiments are required, preferably multicolor FISH, which allows localization of multiple areas in one cell. Loci can also be visualized in live cells by genomic integration of a series of repeat elements for which a specific binding partner is known. The binding protein can be linked with a fluorescent protein, such as GFP to visualize the locus and follow its location through time. Studying contacts between chromosomal regions in live cells will provide insight in the dynamics of chromosomal organization. This should reveal whether different conformations are a consequence of the dynamic movement of loci in all cells or the reflection of a more static nuclear position of loci that differs from cell to cell.

Because active regions co-localize on a large scale, the 4C data do not directly argue against a role of RNA polII in shaping the genome. The influence of RNA polII binding was recently addressed in two studies<sup>13,14</sup>. One 4C study in which transcription was inhibited for 4 h showed that the binding of RNA PolII is not necessary for the *maintenance* of the localization of the active  $\beta$ -globin and *Rad23a* loci<sup>14</sup>. The other study described partially conflicting results (see also chapter 3)<sup>13</sup>. It could be that RNA polII binding is important for *establishing* the position of the locus. However RNA polII can not be the only determinant because silenced regions also co-localize.

The possible large scale formation of CTCF loops is interesting and is preferentially investigated in a 4C-like genome wide study. However to measure interactions between the CTCF sites accurately, the resolution of 4C may need to be increased.

#### **Increasing 4C resolution**

Although 4C is informative in its current set-up, it would be a valuable improvement if the resolution with which co-localization partners are detected is increased. This can only be achieved if ligation frequencies are measured more quantitatively.

The microarray analysis was hampered by the large difference in signal intensities. To measure the infrequent ligation events with restriction fragments tens of megabases away on the genome, the scanner settings were set at a level at which the highest signal intensities were saturated and could not be quantified. In addition, due to the 30 cycles of PCR in the 4C protocol the measured DNA fragments are (theoretically) enriched  $2^{30} = 1$  billion times. This is a much higher enrichment than is achieved for example in immunoprecipitation based methods, which the microarrays were originally designed for. The amount of probe that is spotted is an excess for this latter type of enrichment. Possibly the large overrepresentation of fragments near the viewpoint in 4C prevented linear detection and thus accurate quantification of these fragments. The microarray analysis could be developed further to be better suited for quantification of relatively frequent contacts made in the genome. However, the sequencing based 4C described in chapter 7 provides even better opportunities. In sequencing analysis all different frequencies can be measured, provided that enough sequence reads are obtained to also include all the infrequent ligation products.

But the limited detection range of microarrays is possibly not the only thing that limits the quantitative measurement of individual restriction fragments. In the 4C PCR DNA fragments of different length and relative abundance are amplified in a single reaction. The reaction was optimized to minimize the bias for small fragments and the linearity of amplification was checked by performing a titration series (chapter 5). However, the accuracy of a Q-PCR as is used in 3C may not be reached in the 4C experimental set-up. The PCR was originally incorporated in the protocol to enrich for fragments that were ligated to the viewpoint and keep the hybridization signal on the microarray of other parts of the genome low. In the sequencing strategy only DNA fragments that are amplified in the PCR are analyzed and not all the other parts of the genome present in the original sample (chapter 7). Perhaps linear amplification of the DNA products will yield an enrichment that is high enough for sequencing analysis.

#### **Using 4C to detect genomic rearrangements**

The finding that several megabases of the genome flanking an investigated locus always produce high microarray signals in 4C sparked the idea to investigate loci of which the neighboring sequences had changed, due to genomic rearrangements in the area. Chapter 6 described the studies that showed the high potential of this novel method.

Before 4C, it was difficult to advance from a rearrangement found in low resolution cytogenetic experiments to describing the precise breakpoint. Narrowing down

a breakpoint by FISH experiments is laborious and sometimes impossible due to misalignment of BACs or large repetitive areas. The task is even more complex if only one side (one of the chromosomes) of a rearrangement is known. Many breakpoint fine-mapping projects have stopped, because of these limitations. Moreover, cytogenetic rearrangement detection methods (and others such as array painting) depend on obtaining metaphase chromosomes. For this reason the detection of rearrangements in certain tissues, for example solid tumors, is very difficult. 4C will undoubtedly contribute to the characterization of many rearrangements and this will likely lead to the identification of novel disease associated genes.

A disadvantage of 4C is that it only detects rearrangements located within several megabases of a selected viewpoint. However, by pooling multiple viewpoints, several genomic positions can be analyzed on one microarray and with the sequencing based 4C described in chapter 7 even a genome wide application of 4C is feasible.

A different disadvantage of 4C is that only large rearrangements can be measured and that it is not quantitative enough to measure inversions or any heterozygous change close to the viewpoint. Using sequencing analysis in stead of microarrays may improve the possibilities of 4C in this area.

Other genomic techniques were developed recently that analyze changes in the genome, most notably paired-end-sequencing (PES) approaches<sup>15,16</sup>. In these methods the genome is fragmented in such a way that all the fragments have a comparable size. The ends of the fragments are sequenced and paired. Both sequences are blasted against a reference genome. If the genomic distance between the paired sequences is not the same as the size of the fragment, the analyzed genome is different from the reference genome. The power of these assays is that many different types of structural changes can be identified in one experiment, including small deletions and inversions. The disadvantage is that the detection of genomic rearrangements critically depends on capturing the one DNA fragment that covers the breakpoint of the rearrangement and contains an unexpected combination of ends. Deep sequencing is required to achieve a high enough coverage of the genome, especially in heterogeneous samples. Moreover, if one or both ends of the informative fragment consist of a non-unique sequence in the genome, it can not be analysed.

As described in chapter 2 sequencing technology rapidly advances and affordable whole genome sequencing methods may become available in the not too distant future.

Despite these developments 4C will still have a reason to exist, because it uniquely acts as a pair of high resolution genomic binoculars. Because rearrangement detection

depends on capturing many fragments in stead of one, the method is more robust than any other sequencing method. Within the region of high signals no large rearrangement can be missed. This property is also relevant for possible use of 4C as a diagnostic tool. Theoretically, paired end sequencing of a large library of fosmid clones, can result in detection of many fragments across a breakpoint, but this is a very laborious method<sup>16</sup>. Sequencing based 4C can also be used in unfinished whole genome sequencing projects to help establish chromosomal maps.

## References

1. M. Simonis and W. de Laat, *Biochimica et biophysica acta* (2008).
2. J. Dekker, K. Rippe, M. Dekker et al., *Science (New York, N.Y)* **295** (5558), 1306 (2002)
3. B. Tolhuis, R. J. Palstra, E. Splinter et al., *Molecular cell* **10** (6), 1453 (2002).
4. M. Simonis, P. Klous, E. Splinter et al., *Nature genetics* **38** (11), 1348 (2006).
5. M. Simonis, J. Kooren, and W. de Laat, *Nature methods* **4** (11), 895 (2007).
6. T. Misteli, *The Journal of cell biology* **155** (2), 181 (2001).
7. E. V. Volpi, E. Chevret, T. Jones et al., *Journal of cell science* **113** ( Pt 9), 1565 (2000).
8. J. Nickerson, *Journal of cell science* **114** (Pt 3), 463 (2001)
9. T. Pederson, *Journal of molecular biology* **277** (2), 147 (1998).
10. P. R. Cook, *Nature genetics* **32** (3), 347 (2002).
11. V. Parelho, S. Hadjur, M. Spivakov et al., *Cell* **132** (3), 422 (2008)
12. K. S. Wendt, K. Yoshida, T. Itoh et al., *Nature* **451** (7180), 796 (2008).
13. J. A. Mitchell and P. Fraser, *Genes & development* **22** (1), 20 (2008).
14. R. J. Palstra, M. Simonis, P. Klous et al., *PLoS ONE* **3** (2), e1661 (2008).
15. J. O. Korbil, A. E. Urban, J. P. Affourtit et al., *Science (New York, N.Y)* **318** (5849), 420 (2007)
16. E. Tuzun, A. J. Sharp, J. A. Bailey et al., *Nature genetics* **37** (7), 727 (2005).



Summary

Samenvatting

Samenvatting voor niet ingewijden

## Summary

Eukaryotic genomes are extensively folded to fit into the small volume of the cell nucleus. Despite this compacted structure of the DNA the right set of genes is transcribed in every healthy cell. The relation between gene expression and the spatial organization of the genome has been studied in different ways. The local folded structure small parts of the genome containing specific genes has been analysed in detail. In addition, fluorescent in situ hybridization (FISH) studies have investigated the position in the nucleus of several gene loci in relation to the expression status of the genes. The nuclear location of gene loci is described relative to large nuclear structures or to other selected loci.

The gene locus of which the folded structure was first analysed in detail is the  $\beta$ -globin locus. Chromosome Conformation Capture (3C) experiments showed that the regulatory elements spread over 200 kb frequently loop towards the promoter of the active gene. In 3C technology nuclei are fixed such that elements in the genome spatially close to each other are cross-linked. The chromatin is subsequently digested, diluted and ligated, such that co-localizing fragments will become ligated to each other. The number of times two fragments form a ligation product is a measure for their interaction frequency and is analyzed by (semi-) quantitative PCR across the ligation junctions, using a specific primer for each fragment. 3C is well suited for the analysis of a gene locus, but an unbiased search for DNA interactions across the genome would require a staggering amount of PCR experiments. To overcome the need for an independent PCR reaction for each ligation product, 3C was adapted into 4C (Chromosome Conformation Capture on Chip). In 4C one restriction fragment is selected (the "bait" or "viewpoint") and all the sequences that were ligated this viewpoint fragment are characterized in one experiment. This is achieved by circularization of the ligated fragments after which an inverse PCR with viewpoint specific primers amplifies all fragments ligated to the viewpoint. To allow an efficient PCR the ligation products are shortened in an additional restriction step before circularization. The 4C PCR product contains all the fragments that co-localized with the viewpoint fragment and is analysed on a dedicated microarray. Each probe on the microarray measures the frequency with which one specific restriction fragment was captured by the viewpoint fragment.

Although 4C is based on 3C, the data that is generated and therefore the information that can be obtained from it is very different (chapter 5). In 4C generally a million ligation products of the selected viewpoint fragment are amplified and analyzed in one experiment. Mainly due to PCR amplification biases and limited detection range of



the microarray, the measurement made for each individual restriction fragment is far less quantitative in 4C than it is in a well performed 3C analysis. In stead, co-localization is measured in 4C experiments by the number of different restriction fragments in a genomic area that was ligated to the selected fragment, as judged by the high microarray signal on the fragments. Thus, whereas 3C is used to study the local folded structure of DNA, 4C analyzes the position of a gene locus in the nucleus, by finding all other parts of the genome that are frequently in its spatial surroundings. Generally many genomic regions of ~50 kb and larger that co-localize with the viewpoint are detected.

Using 4C we showed that the active murine  $\beta$ -globin locus in fetal liver tissue co-localizes with other transcriptionally active regions, mainly located on the same chromosome (chapter 4). In contrast, in a tissue in which the same locus is inactive it co-localizes with completely different areas on the same chromosome, containing gene poor and transcriptionally inactive regions. *Rad23a* located in a gene rich region on chromosome 8 largely has the same nuclear environment in two analyzed tissues in which it is transcriptionally active. The sequences surrounding *Rad23a* in the nucleus consist of gene dense and transcriptionally active regions on chromosome 8 and also on other chromosomes. These results demonstrate that chromosomes do not fold randomly, but that genomic regions have many preferred co-localization partners, which consist of chromatin that has the same gene density and transcriptional activity. By describing all frequent contacts made by a gene locus, the significance of the co-localization of selected parts of the genome, as measured in FISH experiments can be better appreciated.

4C can also be used as a diagnostic method for the identification of rearrangements near selected genomic viewpoints (chapter 6). Detailed analysis of the breakpoints of genomic rearrangements is important, because it can lead to the identification of disease associated genes and specific rearrangements were shown to have prognostic value. The detection of (balanced) genomic rearrangements traditionally depended on laborious cytogenetic techniques that only provided limited resolution. 4C can detect balanced and complex rearrangements rapidly and at high resolution. This application of 4C takes advantage of the fact that the viewpoint most frequently contacts the fragments flanking it on the genome, resulting in high microarray signals in this area. If as a consequence of rearrangements new DNA sequences are brought to a genomic position close to viewpoint, they will have unexpectedly high 4C microarray signals. Rearrangements located up to several megabases away from the viewpoint can be identified this way.

A second type of 4C based on sequencing rather than on microarray analysis was also developed (chapter 7). The sequencing strategy allows the analysis of multiple sites of

### *Summary*

the genome, making larger screens for genomic rearrangements feasible, possibly even a genome wide screen. Moreover, the sequencing strategy is more quantitative than microarray analysis and may prove beneficial for the identification of specific interactions between small functional elements on the genome.

## Samenvatting

Eukaryote genomen zijn sterk gevouwen om in het kleine volume van de celkern te passen. Ondanks de compacte structuur van het DNA komt in elke gezonde cel de juiste set genen tot expressie. De relatie tussen genexpressie en de ruimtelijke organisatie van het genoom is bestudeerd in verschillende onderzoekslijnen. De lokale vouwing van het genoom rondom bepaalde genen is bijvoorbeeld in detail bestudeerd. Daarnaast is in fluorescente in situ hybridisatie (FISH) experimenten voor verschillende genen de positie in de kern onderzocht in relatie tot de expressie status van de genen. De locatie in de kern wordt in deze studies beschreven ten opzichte van nucleaire structuren of andere geselecteerde loci.

Het gen locus waarvan voor het eerst de DNA vouwing in detail geanalyseerd is, is het  $\beta$ -globine locus. Chromosome Conformation Capture (3C) experimenten hebben aangetoond dat de regulatieve elementen die zich uitspreiden over 200 kb frequent naar de promotor van het actieve  $\beta$ -globine gen buigen, lussen vormend in het DNA. In de 3C techniek worden celkernen gefixeerd, zodat stukken DNA die in de celkern bij elkaar in de buurt zitten aan elkaar geplakt worden. Het DNA wordt vervolgens geknipt met een restrictie enzym, sterk verdund en geligeerd zodat aan elkaar geplakte fragmenten ligatie producten vormen. Het aantal ligatie producten dat twee fragmenten samen vormen is een maat voor de frequentie waarmee ze bij elkaar in de buurt komen in de kern. Ligatie producten worden gemeten in een kwantitatieve PCR over het ligatie punt met een specifieke primer voor elk restrictie fragment.

3C is een geschikte methode om de ruimtelijke structuur van een gen locus op te helderen, maar voor een zoektocht naar DNA interacties zonder enige voorkennis zijn al snel een onmogelijk aantal PCR reacties nodig. Om de noodzaak van een aparte PCR voor elk verschillend ligatie product te omzeilen is 3C omgezet naar 3C on chip (4C). In 4C worden de ligatie producten tot cirkels gemaakt waarna alle ligatie partners van een geselecteerd fragment (het uitzicht punt) vermenigvuldigd worden in een inverse PCR reactie met primers specifiek voor dat gekozen restrictie fragment. Omdat de restrictie fragmenten te groot zijn voor efficiënte amplificatie, worden de ligatie producten korter gemaakt voordat het cirkels worden. Het product van de inverse PCR bevat alle DNA fragmenten die in de buurt zaten van het uitzicht punt en wordt geanalyseerd op een microarray.

Ondanks dat 4C gebaseerd is op 3C is de data die eruit komt en dus ook de informatie die ermee verkregen kan worden heel anders (hoofdstuk 5). In een 4C experiment worden over het algemeen ongeveer een miljoen ligatie producten van het uitzicht punt ge-amplificeerd en gemeten. Voornamelijk door verschillen in PCR efficiëntie en een beperkt bereik van de microarray is de meting van elke individuele ligatie partner minder kwantitatief dan in 3C. In plaats daarvan wordt in 4C gelet op het aantal verschillende fragmenten per gebied van het genoom dat een ligatie product vormt met het uitzicht punt. Dus, hoewel 3C geschikt is om de lokale vouwing van een gen locus in detail te bestuderen, kan met 4C de positie van een gen locus in de kern bestudeerd worden, door alle delen van het genoom te beschrijven die zich frequent in de directe omgeving van het uitzicht punt bevinden. Over het algemeen worden vele gebieden van 50 kb en groter gevonden in een 4C analyse.

Gebruik makend van 4C (hoofdstuk 4) is laten zien dat de ruimtelijke omgeving van het transcriptioneel actieve  $\beta$ -globine locus voornamelijk bestaat uit andere actieve gebieden van hetzelfde chromosoom. Dit werd bestudeerd in foetale lever cellen van de muis. In foetale hersencellen waarin hetzelfde locus niet actief is maakt het contact met hele ander gebieden op het chromosoom die heel weinig genen of inactieve genen bevatten. *Rad23a*, een gen dat actief is in de twee geanalyseerde weefsels en zich op chromosoom 8 bevindt in een gen rijk, transcriptioneel actief gebied, heeft vrijwel dezelfde kern omgeving in beide weefsels. Deze omgeving bestaat uit gebieden met hoge gen dichtheid en veel gen expressie. Deze resultaten laten zien dat chromosomen niet willekeurig gevouwen zijn, maar dat gebieden in het genoom vele preferentiële contacten hebben met andere stukken van het genoom met dezelfde eigenschappen.

Een tweede, heel andere toepassing van 4C is het karakteriseren van veranderingen in de lineaire structuur van het genoom ("rearrangements") (hoofdstuk 6). Gedetailleerde beschrijving van breekpunten in het genoom is belangrijk omdat zo nieuwe aan ziekte gerelateerde genen kunnen worden gevonden. Bovendien hebben sommige specifieke rearrangements prognostische waarde. De detectie van bepaalde rearrangements was eerst afhankelijk van arbeidsintensieve methoden met een lage resolutie. 4C kan rearrangements binnen relatief korte tijd detecteren met hoge resolutie. Deze toepassing van 4C maakt gebruik van het feit dat het uitzicht punt het vaakst in de buurt komt van de stukken DNA waar het naast gepositioneerd is op het genoom. Dit resulteert in hoge microarray signalen in dit gebied. Als er als gevolg van een rearrangement een ander stuk DNA naast het uitzicht punt gepositioneerd wordt, zal dit gebied onverwacht hoge

microarray signalen bevatten. Rearrangements enkele megabasen van het uitzicht punt kunnen gedetecteerd worden met 4C.

Een tweede type 4C is ontwikkeld die gebaseerd is op sequentie analyse in plaats van microarrays. Met deze strategie kunnen meerdere uitzicht punten geanalyseerd worden in één experiment, zodat grotere zoektochten naar rearrangements gedaan kunnen worden, mogelijk zelfs een genoom breed onderzoek. Bovendien is sequentie analyse meer kwantitatief dan microarray analyse wat voordelen kan hebben voor het bestuderen van de lokale vouwing van het DNA.

## Samenvatting voor niet ingewijden

Genen kunnen beschouwd worden als de recepten voor belangrijke elementen in de cel, zoals enzymen en onderdelen van de structuur van de cel. Genen zijn gecodeerd in lang gerekte moleculen, het DNA. Zoogdiercellen bevatten elk ongeveer twee meter DNA, waarop ongeveer 30.000 genen te vinden zijn en wat opgevouwen zit in een kleine celkern van gemiddeld 10 micrometer in doorsnede. Recentelijk is gebleken dat de vouwing van het DNA een belangrijke invloed heeft op het aflezen van de genen. Daarom is het belangrijk om systematisch de vouwing van het DNA te kunnen analyseren, maar technieken daarvoor ontbreken. In dit promotieonderzoek is een nieuwe moleculaire methode ontwikkeld genaamd 4C (Chromosome Conformation Capture on Chip). Met de 4C techniek kan voor een geselecteerd stukje DNA onderzocht worden met welke andere stukken DNA het fysiek contact maakt in de ruimte van de celkern.

4C is toegepast om de positie van de beta-globine genen in de celkern te bestuderen. Daarbij werd gevonden dat wanneer de globine genen afgelezen worden ("aanstaan") in rode bloedcellen, zij contact maken met andere actieve genen die elders op hetzelfde chromosoom liggen. Echter, wanneer de genen uitstaan (in ander weefsel), verandert hun ruimtelijke omgeving en maken de genen contact met chromosoom gebieden die ook uitstaan. Ook werd er laten zien dat er contacten worden gevormd niet alleen binnen, maar ook tussen chromosomen.

De 4C techniek is verder ontwikkeld voor een heel andere toepassing; het opsporen van genetische veranderingen in cellen van patiënten die bijvoorbeeld lijden aan ziekten zoals kanker. De toegevoegde waarde van 4C is dat deze techniek bepaalde, veel voorkomende, DNA veranderingen vele malen sneller en veel nauwkeuriger in kaart kan brengen dan de bestaande methoden.

Ten slotte is de 4C techniek verder ontwikkeld om geanalyseerd te worden met de nieuwste sequentie-methodes die tientallen miljoenen stukjes DNA tegelijkertijd kunnen aflezen. Deze nieuwe ontwikkeling zal naar verwachting het onderzoek naar DNA vouwing in de celkern nog nauwkeuriger maken en de mogelijkheden van 4C als diagnostische methode voor het opsporen van DNA veranderingen verder verbreden.



## Curriculum Vitae

Name: Marieke Simonis  
Date of birth: 03-11-1980, Utrecht

### Education

1993-1999 VWO.  
Niels Stensen College, Utrecht and GSG Lingecollege, Tiel

1999-2002 Bachelor  
Biomedical Sciences, University of Utrecht

2002-2004 Master  
Genomics and Bioinformatics, University of Utrecht

Nine months research project:  
Functional Genomics group, Hubrecht Laboratory, Utrecht  
*Retrieval and analysis of a Prg-1 mutant in C. elegans.*  
(Dr. M. Tijsterman)

Six months research project:  
Central Microarray Facility, NKI, Amsterdam  
*Microarray analysis of mouse embryonic fibroblasts after release from a cell-cycle arrest in G2.*  
(Dr. R. Kerkhoven)

2004-2008 PhD research  
Department of Cell Biology, Erasmus MC, Rotterdam.  
*Chromosome Conformation Capture on Chip (4C)*  
Promotor: Prof.dr. Frank Grosveld, copromotor: Dr. Wouter de Laat.



## List of Publications

**Simonis M**, deLaat W. *FISH-eyed and genome-wide views on the spatial organisation of gene expression*. Biochim Biophys Acta. 2008 Nov;1783(11):2052-60.

Palstra RJ, **Simonis M**, Klous P, Brasset E, Eijkelkamp B, de Laat W. *Maintenance of Long-Range DNA Interactions after Inhibition of Ongoing RNA Polymerase II Transcription*. PLoS ONE. 2008 Feb 20;3(2):e1661.

Foijer F, **Simonis M**, van Vliet M, Wessels L, Kerkhoven R, Sorger PK, Te Riele H. *Oncogenic pathways impinging on the G2-restriction point*. Oncogene 2008 Feb 14;27(8):1142-54.

**Simonis M\***, Kooren J\*, de Laat W. *An evaluation of 3C-based methods to capture DNA interactions*. Nature Methods 2007 Nov;4(11):895-901.

**Simonis M**, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. *Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)*. Nature Genetics 2006 Nov;38(11):1348-54.

\* These authors contributed equally

## Dankwoord

Een van de mooie dingen van een promotie tijd is dat het wordt afgesloten met zoiets tastbaars als dit boekje. Bovendien kan ik mijn dank voor de mensen die het tot stand komen hiervan mogelijk gemaakt hebben vereeuwigen in dit schrijven.

Beste Frank, zonder jouw brede wetenschappelijk inzicht en je grote interesse voor het toepassen van nieuwe technieken was dit werk zeker niet tot stand gekomen.

Wouter, jouw wetenschappelijke creativiteit vormde de stabiele basis voor het werk in dit boekje. Je serieuze manier van begeleiden en de goede sfeer die je in je lab creëert hebben in belangrijke mate bijgedragen aan het slagen van de projecten. Bedankt voor de goede start die je me meegeeft.

Dat mijn AIO periode een hele leuke tijd was heb ik te danken aan alle prettige collega's. Petra, zonder jou was dit boekje zeker nooit tot stand gekomen. Heel erg bedankt voor al je goede werk en je gezelligheid. Ik ben blij dat je ook tijdens mijn verdediging nog even aan mijn zijde staat.

Daan, ik heb genoten van je gevoel voor humor en heb veel bewondering voor je positieve instelling. Ik hoop dat mijn cynisme een beetje dragelijk voor je was.

Erik, mijn co-co-promoter. Ik heb op verschillende vlakken bijzonder veel van je geleerd en je was ook nog eens een heel gezellige buurman tijdens mijn "excuus PCR-etjes"! Heel veel succes met je eigen promotie traject.

Robert-Jan, als nestor was jij de rots in de branding voor al dat (ietsjes) jongere grut, bedankt voor je belangstelling voor mijn project.

En wat zouden de spelletjes avonden geweest zijn zonder de spelletjes koning. Jurgen, jouw fanatisme heeft veel indruk gemaakt. Veel succes met het afronden van je nieuwe leertraject.

Emily, unfortunately you were only with us for a year. I really enjoyed your presence and wish you all the best for you and your family.

Harmen bedankt voor je geduld tijdens mijn last-minute toevoegingen aan (/weglatingen uit) dit boekje, en bedankt voor het zeilmoment in Loosdrecht. Heb je die handleiding nou al uit?

Leonie, al was het relatief kort, ik vond je een prettige buurvrouw in het 702-hok. Alle goeds voor de toekomst op je werk en thuis.

Haider bedankt voor al je inzet tijdens je stage periode en heel veel succes in de toekomst (in Dubai ofzo).

Sanja and Athina good luck with finishing your PhD's and moving to Amsterdam or other places somewhat prettier than Rotterdam. I wish you all the best.

Het werken aan 4C bracht ook veel samenwerking met zich mee.

Bas en Elzo, bedankt voor jullie input en enthousiasme in de vroege dagen van 4C. Yuri, thanks for your view on the data analysis and the introduction to "R". Erikjan, bedankt voor je werk aan onze probes en je "getting (re-)started with Perl" cursus.

Jules, Irene en Jessica bedankt voor de prettige en vruchtbare samenwerking bij het "rearrangement project". Ik hoop dat jullie nog veel nieuwe spannende dingen gaan vinden met 4C.

Rudi and Claudia, unfortunately our joint work was not far enough advanced to be included in this thesis. I enjoyed working with you and I am sure you and Sjoerd will make the project successful.

Zonder goede fundering kun je geen fatsoenlijk huis bouwen. Lieve ouders, lieve schoonouders, bedankt voor alle steun en het vertrouwen. Hard werken is makkelijker als je weet dat er mensen trots op je zijn.

Hansje, bedankt voor de goed in elkaar gezette InDesign cursus.

Lieve zus, bedankt voor al die keren schooltje spelen en dat jij altijd de lerares wilde zijn. Een goed begin is het halve werk.

Rowland, dat dit boekje af is zal voor jou een even grote opluchting zijn als voor mij. Bedankt voor je begrip en je geduld. Nu eindelijk tijd voor lange boswandelingen en avonden voor de openhaard.

*Marieke*

