Exploring functional elements and genomic variation in the noncoding genome

Sebastiaan August Albert Cornelis van Heesch

Cover art: Reinoud van Vught
Cover design: Sebastiaan van Heesch

ISBN/EAN: 978-94-6108-702-7

Printed by: Gildeprint - Enschede

The research described in this thesis was performed at the Hubrecht Instute for Developmental Biology and Stem Cell Research, within the framework of the Cancer, Stem Cells and Developmental Biology graduate school in Utrecht, The Netherlands.

Financial support by Genzyme, Roche Diagnostics and Life Technologies for printing of this thesis is gratefully acknowledged.

Exploring functional elements and genomic variation in the noncoding genome

Het verkennen van functionele elementen en genomische variatie in het niet-coderende genoom

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 2 juli 2014 des middags te 2.30 uur

door

Sebastiaan August Albert Cornelis van Heesch

geboren op 25 september 1985
te Tilburg

Promotor: Prof.dr. E.P.J.G. Cuppen

The Art of Science

Science and arts are two very alike disciplines, both driven by ambition, creativity and passion, and both pushing the limits of what is technically possible. Centuries ago, science and arts were routinely exercised by the same individual, for example Leonardo da Vinci. However, over time the two disciplines gradually separated, despite their still existing similarities.

With the establishment of the Society of Arts by the Royal Netherlands Academy of Arts and Sciences this year (2014), artists and scientist are again brought closer together, reestablishing the long-standing ties between the two disciplines.

A mutual interest in arts and genetics also brought me closer to Dutch modern artist and family friend Reinoud van Vught. Reinoud's interest in genetics was initially triggered when his son Max and nephew Stijn were diagnosed with Duchenne muscular dystrophy (DMD). DMD is a severe genetic disorder caused by mutations in the dystrophin gene that lead to muscle degeneration. Logically, the illness of Max has had great influence on Reinoud's work.

During my studies and PhD, Reinoud and I frequently discussed newly developed techniques that may one day provide a cure for DMD, such as exon skipping and stem cell replacement. In one of our conversations, the idea rose to join forces and combine Reinoud's art with my scientific work for the realization of this thesis. I explained Reinoud the contents of each chapter, after which he selected details from his paintings that he could associate best with my work. Each selected detail is only a fraction of the complete picture, which helps to put each chapter into perspective and illustrate that much is yet to be discovered. Nine of these details are featured on the first pages of each chapter and a complete work is depicted on the cover.

About Reinoud van Vught

Reinoud van Vught (Goirle, 1960) is a painter to the core. His work is marked by great visual diversity and continuous evolution of applied techniques that may never become a routine. Reinoud's work has been exhibited in multiple Dutch and international cities such as Berlin, Novosibirsk, Paris, Washington and Miami. His work is included in the collections of Museum De Pont (Tilburg), Rijksmuseum (Amsterdam), NoordBrabants Museum (Den Bosch), Frans Hals Museum (Haarlem), Centraal Museum (Utrecht), Vincent van GoghHuis (Zundert), UMC Utrecht, AZL Leiden, the Ministry of Foreign Affairs (Den Haag) and others. For more information: www.vanvught.com

Contents

1

Introduction

Genome function

The human body consists of approximately 3.72 x 10^{13} cells [1]. The majority of cells within the body are diploid; that is, they contain two sets of unique DNA sequence, with each set totaling ~3.2 billion bases of DNA. This DNA is further divided over 23 chromosomes of varying size. Chromosomes are large complexes consisting of DNA and protein, modeled and compacted in such a way that it all is able to fit into a single cell's nucleus (Figure 1A). Taken together, these 46 chromosomes contain the genetic information of ~20,000 protein coding genes [2], providing the "genetic blueprint" of life. Although all cells in our body contain the exact same genetic information, each tissue and organ consists of different cell types exhibiting unique functions. Cell-type specificity and function are established and maintained using highly regulated gene expression programs. Gene expression regulation is a delicate process and the activation and repression of transcriptional programs depends on multiple aspects including genome structure and the presence of transcription factors or functional DNA elements. As the majority of our genome consists mostly of noncoding DNA, with less than 2% actually encoding protein [3, 4], the noncoding genome is crucial in providing the correct context for gene expression.

Nevertheless, it is technically challenging to determine which parts of the genome are functional and under what particular circumstances. Genome sequencing techniques now allow us to detect every base in the genome and its transcribed repertoire, but correct interpretation of all this information is still a major challenge. Furthermore, the mechanisms by which noncoding genomic variants contribute to disease remain largely unclear. Here, I will highlight the advances that have been made towards understanding genome function and the challenges that remain in the correct detection of genomic variation. Also, I will focus on the integration and interpretation of multiple layers of genomic information.

The noncoding genome

A comparison of the genome sequences of humans and closely related species such as the chimpanzee [6], mouse [7] and rat [8] have revealed a high proportion of DNA sequence conservation in coding genes (~98%), but much lower conservation outside of the annotated gene regions [8, 9]. Initially, these large regions were categorized simply as "junk DNA" because of the low conservation of noncoding DNA in combination with a lack of clear function for these regions [10]. Nevertheless, over the past decades it became clear that noncoding DNA is crucial for genome function, and is required for the maintenance of genome structure [11] and gene expression regulation [12]. For example, noncoding chromatin (DNA wrapped around nucleosomes) is continuously remodeled [13, 14] and histone tails continuously covalently modified [15-18] in order to dynamically regulate the state of chromatin and its susceptibility to being transcribed [19, 20]. Combinations of the active parts of chromatin, such as gene promoters or expression silencers and enhancers, can precisely determine

which genes are expressed and to what extent [21, 22]. Specific interactions between distal and proximal regulatory regions are enabled by dynamic chromatin properties such as looping [23], illustrating the importance of genome structure and composition [24] (Figure 1B). However, these continuously changing dynamics make it difficult to determine the particular circumstances, such as tissue type [25, 26] or developmental stage [16, 27], as well as the particular part(s) of the noncoding genome that are important for a functional genome [19]. Furthermore, the majority of genomic differences that exist between individuals reside in the noncoding genome [28-30], frequently resulting in associations with disease [31]. Even though these variants do not directly target protein-coding genes, they may very well affect chromatin state or dynamics and thereby modulate expression regulation of many genes, leading to both phenotypic divergence and disease [32].

DNA and RNA sequencing

The introduction of Next-Generation Sequencing (NGS), a technological revolution introduced in the mid 2000's [33, 34] that followed after the lower throughput Sanger sequencing method, has become a great help in elucidating the complexity of the coding and noncoding genome. Whole genome sequencing (WGS) [35] was developed for the sequencing of complete genomes and exome sequencing [36] was developed for the coding regions, with all of the data being produced within a single sequencing run. Specific applications utilizing DNA-sequencing were developed for the high-throughput assessment of epigenetic modifications and chromatin interactions. For example, by using chromatin immunoprecipitation in combination with DNA sequencing (ChIP-seq) [37], a wide variety of histone tail modifications and transcription factor binding sites could be identified. Epigenetic analyses quickly started to provide evidence that large parts of the noncoding genome were indeed much more complex than previously estimated [15, 38, 39]. This created a boost in the field of epigenomics: the study towards the effects of reversible epigenetic modifications on the complete genome and regulation thereof [40]. Together with whole genome sequencing, these techniques now make the thorough evaluation and integration of genomic, transcriptomic and epigenomic landscapes available for routine molecular biology labs.

Chromatin structure and gene expression regulation

Over the last several years, individual and consortium efforts aiming at decoding all functional DNA elements in the human genome have shed more light on the complexity of chromatin organization and modification [12, 41]. Regulatory functions for regions marked by specific chromatin states could be partially deciphered by determining the combination of post-translational histone modifications present at a genomic locus [16, 19, 42], as was initially suggested nearly a decade earlier in the "histone code hypothesis" [17, 18]. The ENCODE project provides the most comprehensive and high-resolution epigenomics data to date,

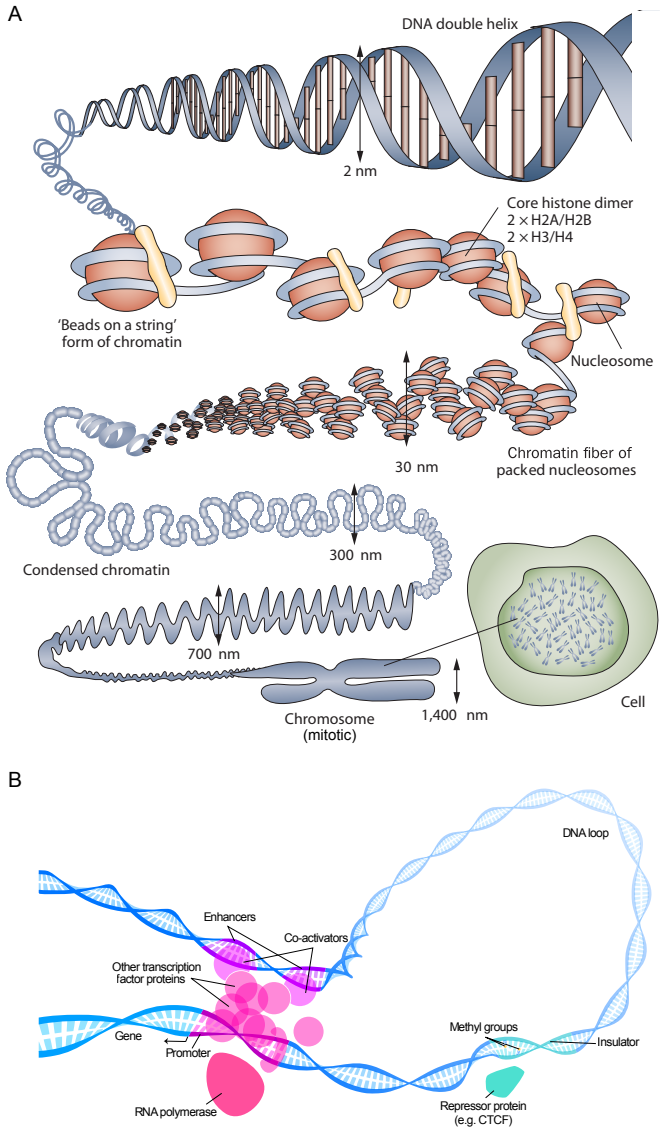


Figure 1 - DNA organization in the nucleus. (**A**) Graphical display of compaction of the two meters of DNA that fit into a single cell's nucleus. DNA is wrapped around nucleosomes consisting of eight histone proteins. Next, the nucleosome-DNA complex ("chromatin") is further organized to reduce the amount of utilized space. In a condensed chromosome state, such as during mitosis, chromosomes display the typical paired sister-chromatid shape connected at the centromeres. Adapted from Tonna et al. Nat. Rev. Nephr. 2010 [5] (**B**) Simplified display of DNA looping between an enhancer and promoter. Noncoding DNA (dark blue) 5' of a gene (light blue) is looped, which allows the physical interaction between transcription (co-) factors at an enhancer and the gene promoter (pink). And example of an insulator element is given as well, which when bound by a repressor transcription factor can inhibit the interaction between the enhancer and the promoter, thereby inhibiting enhanced gene expression. The direction of gene expression via RNA polymerase is indicated with an arrow. Adapted from Kelvinsong (own work), Wikimedia commons (http://commons.wikimedia.org/wiki/File:Transcription_Factors.svg), under a CC-BY-3.0 license).

yielding insight in functionality of DNA elements covering ~80% of the human genome [12]. For a large number of human cell lines, different states of chromatin marking regulatory regions were identified using ChIP-seq. These include, for example, active or poised promoters, enhancers and insulators. Together, this led to a genome-wide, cell type specific map of *in vitro* chromatin states and DNA elements [12].

Chromosome conformation capture (3C) is a different technique developed to provide more insight into the physical interactions between physically proximal DNA elements in the

nucleus [43-45]. 3C-based methods use proximity-based ligation of crosslinked and digested chromatin followed by PCR amplification to study viewpoint-specific reciprocal interactions of DNA elements. Over the last decade, many variations on the 3C technique have been developed to increase resolution via combinations of 3C with DNA arrays or sequencing (4C and 5C) [46-48]. Genome-wide derivations of the 3C method, such as Hi-C [24, 49, 50] and the immunoprecipitation-based ChIA-PET [51-53] can provide interaction information for every locus in the genome, albeit providing lower resolution than targeted methods [50], and requiring higher sequencing depth [44].

Thus far, 3C-based techniques have provided insight on locus-specific enhancer-promoter dynamics, identifying interactions critical for development [54-58] and stem cell pluripotency [59]. Higher throughput methods such as Hi-C and ChIA-PET focused more on the higher architecture organization of the genome, leading to genome-wide chromatin interaction maps [50, 52, 53] and the identification of topologically associating domains (TADs) [11, 60]. TADs contain frequently interacting intra-chromosomal regions limited by strict boundaries (including insulator elements) that constrain the spread of heterochromatin [11]. Disruption of these boundaries results in long-range misregulation of transcription [60]. Topological domains align with coordinately regulated gene clusters and nuclear lamina associated domains (LADs) [60], which in turn can be determined via a different technique termed Dam-ID [61]. Chromatin architecture at such large scale appears fundamental for nuclear chromatin organization [62], with high similarity in these large domains between different cell types and remarkable inter-species conservation [11]. Intra-topological domain interactions, such as enhancer-promoter interactions, are more dynamic and characteristic for the transcriptional program belonging to a specific cell type or state. These intra-topological domain dynamics recently appeared more restricted in dynamics than expected. Jin ***et al*** showed that interactions within topological domains are relatively stable once established in a specific cell type, and functionality of such interactions between DNA regulatory elements depend on the availability of cell-type specific transcription factors [50]. During mitosis, however, the 3D architecture of each genomic region on every chromosome is remodeled to a temporary uniform “linear” landscape with consecutive loops during metaphase, losing the locus-specific composition acquired prior to entering and immediately after cell division [63]. In summary, chromatin modification, dynamics, structure and organization are all associated with the complete genome and not just for the coding DNA, illustrating that the noncoding genome is absolutely vital in genome functioning. However, despite all efforts in determining the role of higher order chromatin dynamics and the function of each unique DNA element, the precise function of the majority of the noncoding regulatory genome remains unknown.

The noncoding genome encodes RNA

As a successor to mRNA expression microarrays, RNA sequencing (RNA-seq) [64] and small RNA-seq [65] were developed for analysis of the total transcriptional landscape of a cell. RNA-

seq not only defines cell-type specific genome-wide transcriptional activity by measuring RNA expression levels, but also distinguishes qualitative RNA differences such as individual transcript isoforms produced via alternative exon usage (alternative splicing) or modifications at the base level (RNA editing) [66]. The introduction of whole transcriptome profiling via RNA-seq for the first time allowed complete analysis of all transcribed regions of the genome, including all the noncoding regions. This is in contrast with microarray analyses that mostly require ***a priori*** knowledge of transcript annotation for RNA array capturing designs. Contrary to the conventional view of the genome, with transcription only taking place at protein-coding regions, many noncoding regions also showed signs of transcription. Estimates on the proportion of the genome that was actually being transcribed began to grow so high that many scientists began to use the term "pervasive" [38, 39, 67, 68], with others claiming that the majority of the transcribed noncoding DNA is only due to transcriptional noise [69, 70]. Of course, some of these noncoding transcribed regions harbored known functional noncoding RNA molecules, such as ribosomal RNAs (ribosome component), snoRNAs (ribosome biogenesis), transfer RNAs (amino acid transport) and small regulatory RNAs (RNA interference). Other RNA molecules were discovered in large intra- and intergenic noncoding regions that had no known function but contained gene-like structures, with signs of transcription and chromatin modification indicating importance [71-74]. Transcripts with a size ≥ 200 nucleotides but lacking a functional open reading frame (ORF) were termed long noncoding RNAs (lncRNAs) [75, 76]. LncRNAs can originate as (intronic) sense transcripts, antisense transcripts on the opposite strand of a protein-coding gene or from intergenic regions with none of the above characteristics [75]. Recent estimates on the number of lncRNAs in the human genome show that there are ~15,000 lncRNAs, expressed in various cell types [75]. For the majority of lncRNAs, their function is still unclear [77]. Some lncRNAs have been implicated to function in nuclear processes such as X-chromosome silencing [78], transcription regulation [79, 80] and telomere maintenance [81]. However, detailed mechanisms of action are often uncertain [75]. The fact that some lncRNAs are critical is illustrated by the fact that aberrant expression of (mutated) lncRNAs can result in (neuro) developmental phenotypes and somatic disease [82-84], including cancer [85-88].

Recently, many lncRNAs were found to localize to the cytosol and associate with ribosomes [89], indicating that lncRNAs might fulfill extra nuclear functions as well. Possibly, lncRNAs contain short ORFs that are translated into short ORF-encoded polypeptides (SEPs) [90]. Otherwise, they might be involved in the regulation of extra-nuclear processes such as translation [91]. It is quite surprising that lncRNAs localize to ribosomes in the cytosol, because according to our current understanding lncRNAs lack the capacity to serve as a template for producing protein. Their role remains unclear, but recent evidence re-establishes that although lncRNAs associate with ribosomes, they do indeed lack protein-coding capacity [92].

These studies show that lncRNA biology adds another layer of complexity to genome regulation and function. Many lncRNAs are indicated to have critical roles in development

and in protecting the cell from a diseased state, but mostly only nuclear roles have been assigned. The association of lncRNAs with ribosomes shows that the role of the noncoding genome likely surpasses that of nuclear functions alone. More research is necessary to get better insight in the diversity of lncRNA localization and all the cellular processes that make use of noncoding RNAs [93]. Questions such as to what extent lncRNAs bind ribosomes, what their cytosolic function is, and if the mechanisms of ribosome binding are identical to those of protein-coding mRNAs need more research.

Accurate detection and interpretation of the functional consequences of genomic variation

Inter-individual genomic variation facilitates phenotypic (population) diversity and evolution [94-96]. The downside, however, is that ***de novo*** (germline or somatic) and inherited variants can also be the cause of disease [97, 98]. A broad range of variation types exist in human genomes, including small variants such as single nucleotide variants (SNVs) and small insertions/deletions (indels), but also larger structural variation (SV) [99, 100]. SVs include balanced events including chromosomal translocations and inversions, but also unbalanced insertions, deletions and (tandem) duplications (copy number changes, or CNVs) [101] (Figure 2). Generally, the contribution of SV to disease is expected to be much greater than that of SNVs [101, 102]. This is exemplified by the fact that the genome-wide number of basepairs affected by SVs is significantly larger than for SNVs, and SVs account for 0.5-1% of all heritable sequence variation between individuals, while SNVs only take up 0.1% [101, 103, 104].

Over recent years, various sequencing and array applications have improved SV detection [101, 102], for which previously low-resolution techniques as karyotyping and fluorescent ***in situ*** hybridization (FISH) were the standard. Since then, several catalogues have been generated to give insight into the diversity and abundance of inherited and somatic SV in the human genome [102-107]. However, the detection of SV is still challenging due to technical limitations and the broad variety of different SV types and sizes [108]. Whereas SNV positions are fairly easily determined by deep sequencing of complete genomes [109], SV detection is less straightforward and challenged by the complexity of genomes, driven by the presence of many repetitive elements such as microsatellite repeats and retrotransposable elements [101, 104]. For unbalanced rearrangements, such as CNVs, detection occurs mostly via quantitative DNA measurements, using either whole genome tiling arrays (e.g. array-comparative genomic hybridization (aCGH)) [110, 111] or sequencing read depth of coverage (DOC) approaches [112]. aCGH in particular is widely used in clinical settings to study rearrangements in congenital disease patients or cancer genomes [113, 114]. Although both aCGH and DOC are widely applied techniques, the actual calling of CNVs from the data is complicated by biases that are likely induced by intrinsic DNA characteristics. The resulting “wave pattern” drives fluctuations in coverage that can lead to false positive CNV calls [115-117].

For balanced rearrangements like inversions or translocations, the detection cannot be based on the quantity of DNA. Paired-read sequencing approaches such as the paired-end (PE) [118] or mate-pair (MP) [119] techniques provide a solution for this problem, making use of read mapping characteristics (e.g. the read location, read orientation and distance between two reads) of pairs of sequencing reads, as compared to a reference genome (Figure 2). PE and MP can also detect CNVs using the distance and orientation of reads, providing an alternative for aCGH and DOC. The main limitation of paired-read approaches is the maximum library insert size, further complicated by differently sized SVs that require different insert size sequencing libraries for optimal detection rates. Small SVs (e.g. 100 bps) cannot be detected with large (2-3 kbs) insert size libraries due to broad insert-size distributions, but are easily detected with insert sizes of 200 bps. Larger SVs driven by, for example, mobile element insertions (LINEs), require inserts of > 8kb to be bridged, which is not easily possible with current sequencing platforms. These examples illustrate that providing an accurate map using individual techniques is almost impossible, which definitely affects the success rate of SV detection in a clinical setting. Eventually, long-read single-molecule sequencing will likely overcome these problems, but until that time technical and analytical optimization is required for accurate detection of SVs.

Mechanisms of DNA repair and SV formation

Mechanisms driving single nucleotide mutations ("point mutations") have been studied extensively, showing that SNVs and indels usually arise during DNA replication, stimulated by mutagens such as radiation and specific chemical compounds [120, 121]. The mechanisms of SV formation are less understood, and require the induction and repair of double-stranded DNA breaks. The current known types of mechanisms that lead to SV formation involve homology- and non-homology (or micro-homology)-based DNA-repair [101, 122-124]. Recurrent SVs in humans are mostly associated with homology-
based mechanisms such as non-allelic homologous DNA recombination (NAHR) [125-127]. SVs are termed recurrent when they are found in multiple non-related individuals and are very similar in location and size of the SV. On the other hand, non-recurrent rearrangements are sporadic and assumed to be more randomly distributed throughout the genome. DNA repair mechanisms associated with this latter type of SV do not make use of sequence homology, but instead make use of other methods, for example non-homologous end joining (NHEJ) to re-anneal broken DNA ends [124, 126-128] (Figure 2). Both simple and more complex types of non-recurrent SV (involving multiple breakpoints) are associated with DNA replication stress leading to genomic breaks. Examples of such break and repair models include fork stalling and template switching (FoSTeS) [122] and the more or less similar microhomology-mediated break-induced repair (MMBIR) [123] (Figure 2). FoSTeS and MMBIR involve the local restoration of damaged DNA after recurrent collapses of replication forks, predicted to occur during initiating mitotic divisions [122].

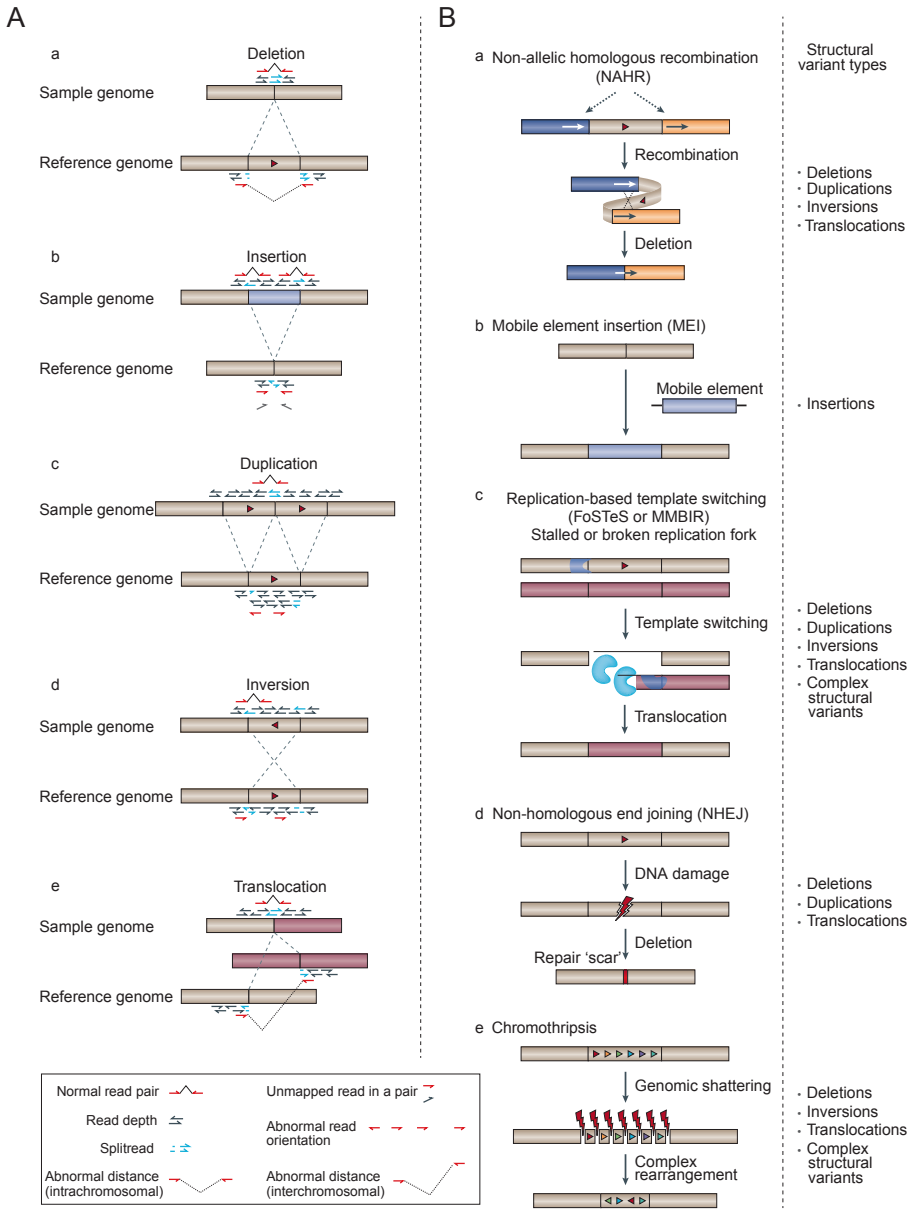


Figure 2 - Classes and mechanisms of structural variation. (**A**) Schematic representation of different types of structural variants and how paired-read sequencing can be used to detect each type. Sample genomes (e.g. patient genome) and reference genome structure are shown for deletions (a), insertions (b), duplications (c), inversions (d) and translocations (e). (**B**) Molecular mechanisms facilitating structural variant formation. Schematic representation of mechanisms driving SV formation: non-allelic homologous recombination (a), mobile element insertions (b), replication-based template switching mechanisms (c), non-homologous end joining (d) and chromothripsis (e). Figure adapted from: Weischenfeldt et al. Nat. Rev. Genet. 2013 [101].

A different class of localized, complex and non-recurrent massive rearrangements can result from random chromosome shattering and re-annealing of DNA fragments, termed "chromothripsis". Chromothripsis was first described in cancer [129] and later shown to drive congenital disease as well [130], reviewed in [131]. In general, chromothripsis is a single catastrophic event, resulting in copy-neutral SVs accompanied by the occasional loss of genomic fragments (Figure 2). The derivative chromosomes that result from chromothripsis can be precisely reconstructed using paired-read sequencing techniques and breakpoint validation by Sanger sequencing. The pattern in which the DNA ends are randomly re-annealed can be precisely followed across all breakpoint junctions, which is characteristic of a one-off event (described as "chromosome walks" in [132]). The listed characteristics of chromothripsis indicate that DNA repair occurs via NHEJ and is not facilitated by mechanisms such as FoSTeS and MMBIR, which have gradual modes of acquiring rearrangements involving multiple copy number states [132, 133]. It is still uncertain what type of stress drives chromothripsis, but due to its highly clustered lesions it is assumed that condensed chromosomes are targeted during mitotic divisions [129, 134]. A different type of localized lesions that have acquired multiple copy number strains has been observed but in other aspects seems to be highly similar to chromothripsis [133]. These events, termed "chromoanasynthesis", may be characteristic of localized, single-event rearrangements that result from MMBIR or FoSTeS, and therefore display multiple copy-number states [133]. However, they may also represent genomically unstable regions initially targeted by chromothripsis, with additional copy number gains occurring during subsequent replication stress.

The increase in resolution of SV detection has resulted in the identification of multiple mechanisms capable of generating highly localized DNA lesions, which could never have been detected by karyotyping or FISH. Nevertheless, sequencing techniques may have allowed double stranded DNA break repair mechanisms to be studied in more detail, but the mechanistic link between the obtained variants and disease (i.e. the phenotypic consequence of the SV) remains largely unclear.

Interpreting variants and defining their causality for disease

Variants determined to be causal for disease can follow Mendelian inheritance patterns [135], originate as ***de novo*** germline variants or can be somatically acquired. For example, developmental syndromes and abnormalities can be caused by inherited and ***de novo*** single point mutations and structural variants [136-142]. Somatic disease such as cancer can originate via point mutations or structural genome changes as well, resulting in malfunctioning tumor suppressor genes or aberrantly activated oncogenes [143, 144]. Of course, not all variation leads to disease and much of the variation in the human genome is described as "common" (e.g. single nucleotide polymorphisms (SNPs)), meaning that the variant is present

in ≥1% of the human population. A major boost in the identification and classification of common and rare variants came from studies using large cohorts of human individuals (e.g. the HapMap project [145] and the 1000 genomes project [9, 35, 104, 146]). Large catalogues of common variation now exist (e.g. dbSNP [147], dbVAR and DGVa [148]) and have been extensively expanded since the introduction of whole-genome DNA sequencing. Filtering against common variation from such databases is useful for finding rare, case-specific disease variants in the haystack of common variants that is generally obtained from sequencing data [149].

Although some rare variants might be directly causal for disease, such as for Mendelian disease [135], defining the exact contribution for the majority of variants associated with disease is difficult [150]. Most common traits are multifactorial and the level of susceptibility to a trait is likely to result from combinations of common variants associated with disease, according to the "common disease / common variant" hypothesis [150-152]. Examples of common complex diseases that have no clear or uniform cause are high blood pressure, obesity, heart failure, and type 2 diabetes.

Using genome-wide association studies (GWAS), each of these diseases has been regularly associated with common SNPs and nearby genes, but showing mechanistic involvement of the proposed candidate genes is still a huge challenge and unmet need [153, 154]. To find significant associations, GWAS often requires large patient cohorts of up to tens of thousands individuals and equal numbers of controls [155]. However, proving causality of an association needs functional follow-up, a necessity that is rarely provided [154, 156, 157]. Rapidly maturing techniques such as (quantitative) mass-spectrometry analyses of proteomes and metabolomes provide novel means in exploring GWAS hits and elucidating mechanisms leading to disease. Complementary to proteome analyses, techniques such as RNA sequencing, ChIP sequencing and whole genome sequencing provide qualitative and quantitative tools for dissecting the molecular basis of disease on multiple mechanistic levels [158]. Integrating genomics and proteomics, so-called proteogenomics [159, 160], allows for systems-level analysis of the previously mentioned multi-factorial traits [161]. For example, RNA editing variants, RNA splice isoforms, rare genomic variants and quantitative RNA levels can all be investigated. Although challenging, integration of multiple data modalities will be necessary to improve our understanding on the effects of disease-associated variation [162].

Genomic variation in the noncoding genome

The effects of genomic variants in genes are relatively easy to predict. These include gene deletions or duplications, nonsynonymous amino acid changes that lead to a malfunctioning protein, or reading frame-shifts that produce premature translational stops. However, the majority of all variants reside in the noncoding genome [28-30]. To date, the potential effects of noncoding variation have remained largely unexplored. The fact that sequence conservation is much lower in regulatory regions than in coding regions implies that this

variation has much less clinical relevance. However, many large complex disease GWAS have pointed to SNP associations in noncoding DNA [29, 32, 163]. Notably, only 7% of the GWAS hits identified over the last decade reside in protein-coding regions [164, 165] and the other 93% map to the noncoding genome [31].

The abundance of noncoding disease-linked variants makes it much more complicated to find mechanistic links to complex disease [30, 32]. Therefore, the mechanisms that couple noncoding variation to disease need further efforts. What (common) noncoding variants affect phenotypes, and which ones do not? Do common noncoding variants contribute to common traits by affecting DNA accessibility, such as through chromatin states or dynamics and thus the nuclear organization of DNA? How easily do our genomes adapt to circumvent ***de novo*** undesired variants? Except for some thorough ***in vitro*** efforts showing that noncoding genomic variation can affect DNA regulatory elements and transcription factor binding sites, the ***in vivo*** effects of ***de novo*** and natural occurring variation and the (combinatorial) contribution of this variation to disease are unclear [166-169]. Several studies, however, have already shown that noncoding variation can have serious consequences, such as contributing to cancer [170]. Genomic variation in lncRNAs has also been implicated in driving neurodevelopmental disability and cancer [83, 171, 172], and comparisons of lncRNA expression levels with noncoding SNPs derived from GWAS studies revealed a number of cis-regulated genotype-lncRNA loci linking noncoding RNAs to disease [173, 174]. These incidental examples only provide a tip of the iceberg, as lncRNAs are present in numbers more or less equal to protein-coding genes [75], thus physically taking up only a small percentage of the noncoding genome. This leaves a large part of all noncoding variation and its functional consequences unstudied.

Noncoding structural variation adds another layer of complexity

Most of the above discussed work towards noncoding variation focuses on single nucleotide variation and not on structural variation. ***De novo*** SVs frequently target noncoding DNA, but defining the contribution of noncoding SVs to disease is difficult. Whereas SNVs may affect transcription factor binding to DNA, SVs are likely to have more profound effects on the regulatory landscape, such as completely abolishing existing enhancer-promoter interactions via deletions or translocations or creating novel interactions via repositioning of regulatory elements [175-178]. Also, they may affect transcription regulation by altering the spatial genome organization, for instance by changing the way the SV-targeted region is associated with the nuclear lamina. The derivative chromosomes, the chromosomes constructed of the chromosomes targeted by an inter-chromosomal translocation, are very likely to be differentially organized in the nucleus as compared to the two original chromosomes. For more complex SVs that involve many more breakpoints, hypothetical combinatorial molecular consequences are numerous and the causal variants are hard if not impossible to pinpoint.

Experimental testing of the consequences of SVs, particularly non-recurrent SVs, is extremely difficult and not only requires insight in changes at the DNA level, but also at the epigenomic, transcriptomic and nuclear organization level. This information is necessary from both the patient carrying the SV and healthy controls that are genetically comparable (e.g. family members). In the clinic, correct implementation of such integrative analyses is difficult, requiring not only knowledge of data integration and interpretation, but also an abundance of patient material and correctly chosen healthy controls. If all are present, the chance of successfully identifying the pathological contribution of a detected genomic alteration is still relatively small. Thus far, most patients with non-recurrent SVs (i.e. not targeting known disease genes) have been left undiagnosed, which is extremely unsatisfying for the patient and family. Better understanding of the mechanisms by which non-recurrent SVs or complex structural rearrangements (e.g. chromothripsis) drive disease will eventually result in improved patient diagnosis.

Summary of thesis

In this thesis, I will discuss technological and conceptual advances in the detection and interpretation of structural variation (**Chapter 2, 3, 6 and 7**) and single nucleotide variation (**Chapter 4,7**), both in a controlled system using the rat as a model organism (**Chapter 2,3,4 and 7**) as well as in large cohorts with patients carrying (complex) SVs (**Chapter 6**). The work presented in chapter 6 provides the first ***in vivo*** molecular analysis of congenital disease patients with chromosomes shattered by germline chromothripsis. I will also explain the use of integrated multi-level "omics" approaches in defining mechanistic links between variation and disease (**Chapter 4,6 and 7**). By applying combinations of multiple sequencing techniques, noncoding genome features such as DNA regulatory elements and genome structure are analyzed (**Chapter 6 and 7**), providing insight into their roles and the need for their integrity in genome functioning. Also, the localization of long noncoding RNAs to ribosomes is further studied using subcellular RNA sequencing, shedding light on the complexity of lncRNA biology, both within and outside the nucleus (**Chapter 5**).

Combined, these chapters provide insight in the complexity of the noncoding genome and all of its facets, including noncoding RNAs and regulatory elements. The chapters show how multiple techniques can be combined efficiently to obtain more insight in biological and disease processes and in predicting the effects of genomic variation.

References

1. Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, Vitale L, Pelleri MC, Tassani S, Piva F, et al: An estimation of the number of cells in the human body. Ann Hum Biol 2013, 40:463-471.

2. Finishing the euchromatic sequence of the human genome. Nature 2004, 431:931-945.

3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: The sequence of the human genome. Science 2001, 291:1304-1351.

5. Tonna S, El-Osta A, Cooper ME, Tikellis C: Metabolic memory and diabetic nephropathy: potential role for epigenetic mechanisms. Nat Rev Nephrol 2010, 6:332-341.

6. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005, 437:69-87.

7. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al: Initial sequencing and comparative analysis of the mouse genome. Nature 2002, 420:520-562.

8. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al: Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 2004, 428:493-521.

9. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: A map of human genome variation from population-scale sequencing. Nature 2010, 467:1061-1073.

10. Ohno S: So much "junk" DNA in our genome. Brookhaven Symp Biol 1972, 23:366-370.

11. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012, 485:376-380.

12. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: An integrated encyclopedia of DNA elements in the human genome. Nature 2012, 489:57-74.

13. Ahmad K, Henikoff S: The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. Mol Cell 2002, 9:1191-1200.

14. Billon P, Cote J: Precise deposition of histone H2A.Z in chromatin for genome expression and maintenance. Biochim Biophys Acta 2012, 1819:290-302.

15. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: High-resolution profiling of histone methylations in the human genome. Cell 2007, 129:823-837.

16. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al: Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 2010, 107:21931-21936.

17. Jenuwein T, Allis CD: Translating the histone code. Science 2001, 293:1074-1080.

18. Strahl BD, Allis CD: The language of covalent histone modifications. Nature 2000, 403:41-45.

19. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al: Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011, 473:43-49.

20. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al: Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009, 459:108-112.

21. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 2007, 39:311-318.

22. Vernimmen D, De Gobbi M, Sloane-Stanley JA, Wood WG, Higgs DR: Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. EMBO J 2007, 26:2041-2051.

23. Bulger M, Groudine M: Looping versus linking: toward a model for long-distance gene activation. Genes Dev 1999, 13:2465-2477.

24. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009, 326:289-293.

25. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al: ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 2009, 457:854-858.

26. Xu J, Watts JA, Pope SD, Gadue P, Kamps M, Plath K, Zaret KS, Smale ST: Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes Dev 2009, 23:2824-2838.

27. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al: A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 2006, 125:315-326.

28. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal Lari R, Lupien M, Markowitz S, Scacheri PC: Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res 2014, 24:1-13.

29. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al: Systematic localization of common disease-associated variation in regulatory DNA. Science 2012, 337:1190-1195.

30. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M: Linking disease associations with regulatory information in the human genome. Genome Res 2012, 22:1748-1759.

31. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009, 106:9362-9367.

32. Ward LD, Kellis M: Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol 2012, 30:1095-1106.

33. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005, 437:376-380.

34. Schuster SC: Next-generation sequencing transforms today's biology. Nat Methods 2008, 5:16-18.

35. Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 2010, 11:415-425.

36. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al: Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009, 461:272-276.

37. Johnson DS, Mortazavi A, Myers RM, Wold B: Genome-wide mapping of in vivo protein-DNA interactions. Science 2007, 316:1497-1502.

38. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007, 447:799-816.

39. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL: Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol 2012, 30:99-104.

40. Bernstein BE, Meissner A, Lander ES: The mammalian epigenome. Cell 2007, 128:669-681.

41. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 2007, 448:553-560.

42. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K: Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 2008, 40:897-903.

43. de Laat W, Dekker J: 3C-based technologies to study the shape of the genome. Methods 2012, 58:189-191.

44. de Wit E, de Laat W: A decade of 3C technologies: insights into nuclear organization. Genes Dev 2012, 26:11-24.

45. Dekker J, Rippe K, Dekker M, Kleckner N: Capturing chromosome conformation. Science 2002, 295:1306-1311.

46. 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). Nat Genet 2006, 38:1348-1354.

47. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, et al: Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet 2006, 38:1341-1347.

48. Dostie J, Dekker J: Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc 2007, 2:988-1002.

49. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J: Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 2012, 58:268-276.

50. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B: A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 2013, 503:290-294.

51. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, et al: CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet 2011, 43:630-638.

52. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al: Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 2012, 148:84-98.

53. Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, Ngan CY, Lim J, Tai E, Poh HM, Wong E, et al: Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature 2013, 504:306-310.

54. Andrey G, Montavon T, Mascrez B, Gonzalez F, Noordermeer D, Leleu M, Trono D, Spitz F, Duboule D: A switch between topological domains underlies HoxD genes collinearity in mouse limbs. Science 2013, 340:1234167.

55. Bau D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA: The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol 2011, 18:107-114.

56. Ferraiuolo MA, Rousseau M, Miyamoto C, Shenker S, Wang XQ, Nadler M, Blanchette M, Dostie J: The three-dimensional architecture of Hox cluster silencing. Nucleic Acids Res 2010, 38:7472-7484.

57. Fraser J, Rousseau M, Shenker S, Ferraiuolo MA, Hayashizaki Y, Blanchette M, Dostie J: Chromatin conformation signatures of cellular differentiation. Genome Biol 2009, 10:R37.

58. Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D: The dynamic architecture of Hox gene clusters. Science 2011, 334:222-225.

59. de Wit E, Bouwman BA, Zhu Y, Klous P, Splinter E, Verstegen MJ, Krijger PH, Festuccia N, Nora EP, Welling M, et al: The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature 2013, 501:227-231.

60. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al: Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 2012, 485:381-385.

61. van Steensel B, Henikoff S: Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat Biotechnol 2000, 18:424-428.

62. Cavalli G, Misteli T: Functional implications of genome topology. Nat Struct Mol Biol 2013, 20:290-299.

63. Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J: Organization of the mitotic chromosome. Science 2013, 342:948-953.

64. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 2008, 320:1344-1349.

65. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP: Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell 2006, 127:1193-1207.

66. Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB: Identifying RNA editing sites using RNA sequencing data alone. Nat Methods 2013, 10:128-132.

67. Hangauer MJ, Vaughn IW, McManus MT: Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet 2013, 9:e1003569.

68. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, et al: The reality of pervasive transcription. PLoS Biol 2011, 9:e1000625; discussion e1001102.

69. Struhl K: Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 2007, 14:103-105.

70. van Bakel H, Nislow C, Blencowe BJ, Hughes TR: Most "dark matter" transcripts are associated with known genes. PLoS Biol 2010, 8:e1000371.

71. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al: Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009, 458:223-227.

72. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al: RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 2007, 316:1484-1488.

73. Mattick JS: The genetic signatures of noncoding RNAs. PLoS Genet 2009, 5:e1000459.

74. Ponting CP, Oliver PL, Reik W: Evolution and functions of long noncoding RNAs. Cell 2009, 136:629-641.

75. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al: The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012, 22:1775-1789.

76. Ulitsky I, Bartel DP: lincRNAs: genomics, evolution, and mechanisms. Cell 2013, 154:26-46.

77. Kowalczyk MS, Higgs DR, Gingeras TR: Molecular biology: RNA discrimination. Nature 2012, 482:310-311.

78. Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N: Requirement for Xist in X chromosome inactivation. Nature 1996, 379:131-137.

79. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, et al: Long noncoding RNAs with enhancer-like function in human cells. Cell 2010, 143:46-58.

80. Tripathi V, Shen Z, Chakraborty A, Giri S, Freier SM, Wu X, Zhang Y, Gorospe M, Prasanth SG, Lal A, Prasanth KV: Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. PLoS Genet 2013, 9:e1003368.

81. Feng J, Funk WD, Wang SS, Weinrich SL, Avilion AA, Chiu CP, Adams RR, Chang E, Allsopp RC, Yu J, et al.: The RNA component of human telomerase. Science 1995, 269:1236-1241.

82. Kretz M, Webster DE, Flockhart RJ, Lee CS, Zehnder A, Lopez-Pajares V, Qu K, Zheng GX, Chow J, Kim GE, et al: Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev 2012, 26:338-343.

83. Talkowski ME, Maussion G, Crapper L, Rosenfeld JA, Blumenthal I, Hanscom C, Chiang C, Lindgren A, Pereira S, Ruderfer D, et al: Disruption of a large intergenic noncoding RNA in subjects with neurodevelopmental disabilities. Am J Hum Genet 2012, 91:1128-1134.

84. Ziats MN, Rennert OM: Aberrant expression of long noncoding RNAs in autistic brain. J Mol Neurosci 2013, 49:589-593.

85. Brunner AL, Beck AH, Edris B, Sweeney RT, Zhu SX, Li R, Montgomery K, Varma S, Gilks T, Guo X, et al: Transcriptional profiling of long non-coding RNAs and novel transcribed regions across a diverse panel of archived human cancers. Genome Biol 2012, 13:R75.

86. Ji P, Diederichs S, Wang W, Boing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, et al: MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 2003, 22:8031-8041.

87. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, et al: Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011, 29:742-749.

88. Yoshimizu T, Miroglio A, Ripoche MA, Gabory A, Vernucci M, Riccio A, Colnot S, Godard C, Terris B, Jammes H, Dandolo L: The H19 locus acts in vivo as a tumor suppressor. Proc Natl Acad Sci U S A 2008, 105:12417-12422.

89. Ingolia NT, Lareau LF, Weissman JS: Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 2011, 147:789-802.

90. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL, Grimmond SM: The abundance of short proteins in the mammalian proteome. PLoS Genet 2006, 2:e52.

91. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A: Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 2013, 9:59-64.

92. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES: Ribosome profiling provides evidence that large noncoding RNAs

do not encode proteins. Cell 2013, 154:240-251.

93. Mercer TR, Mattick JS: Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol 2013, 20:300-307.

94. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L: Natural selection has driven population differentiation in modern humans. Nat Genet 2008, 40:340-345.

95. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al: Identifying recent adaptations in large-scale genomic data. Cell 2013, 152:703-713.

96. Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, et al: Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 2013, 152:691-702.

97. Frazer KA, Murray SS, Schork NJ, Topol EJ: Human genetic variation and its contribution to complex traits. Nat Rev Genet 2009, 10:241-251.

98. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, et al: Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008, 40:1253-1260.

99. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: Detection of large-scale variation in the human genome. Nat Genet 2004, 36:949-951.

100. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al: Large-scale copy number polymorphism in the human genome. Science 2004, 305:525-528.

101. Weischenfeldt J, Symmons O, Spitz F, Korbel JO: Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 2013, 14:125-138.

102. Lee C, Scherer SW: The clinical context of copy number variation in the human genome. Expert Rev Mol Med 2010, 12:e8.

103. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al: Origins and functional impact of copy number variation in the human genome. Nature 2010, 464:704-712.

104. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, et al: Towards a comprehensive structural variation map of an individual human genome. Genome Biol 2010, 11:R52.

105. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE: Diversity of human copy number variation and multicopy genes. Science 2010, 330:641-646.

106. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al: Mapping copy number variation by population-scale genome sequencing. Nature 2011, 470:59-65.

107. Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, Vu TH, Shafer N, Bernier R, Ferrero GB, Silengo M, et al: Relative burden of large CNVs on a range of neurodevelopmental phenotypes. PLoS Genet 2011, 7:e1002334.

108. Alkan C, Coe BP, Eichler EE: Genome structural variation discovery and genotyping. Nat Rev Genet 2011, 12:363-376.

109. Li H, Ruan J, Durbin R: Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008, 18:1851-1858.

110. Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, et al: A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet 2004, 36:299-303.

111. Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, Zhang J, Liu G, Ihara S, Nakamura H, Hurles ME, et al: Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res 2006, 16:1575-1584.

112. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J: Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 2009, 19:1586-1592.

113. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992, 258:818-821.

114. Pinkel D, Albertson DG: Array comparative genomic hybridization and its applications in cancer. Nat Genet 2005, 37 Suppl:S11-17.

115. Benjamini Y, Speed TP: Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 2012, 40:e72.

116. Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET, et al: Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol 2007, 8:R228.

117. van de Wiel MA, Brosens R, Eilers PH, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, Ylstra B: Smoothing waves in array CGH tumor profiles. Bioinformatics 2009, 25:1099-1104.

118. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, et al: Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008, 40:722-729.

119. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al: Paired-end mapping reveals extensive structural variation in the human genome. Science 2007, 318:420-426.

120. Pfeifer GP: Environmental exposures and mutational patterns of cancer genomes. Genome Med 2010, 2:54.

121. Pena-Diaz J, Bregenhorn S, Ghodgaonkar M, Follonier C, Artola-Boran M, Castor D, Lopes M, Sartori AA, Jiricny J: Noncanonical Mismatch Repair as a Source of Genomic Instability in Human Cells. Molecular Cell 2012, 47:669-680.

122. Lee JA, Carvalho CM, Lupski JR: A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 2007, 131:1235-1247.

123. Hastings PJ, Ira G, Lupski JR: A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 2009, 5:e1000327.

124. Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR: The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 2009, 41:849-853.

125. Stankiewicz P, Lupski JR: Genome architecture, rearrangements and genomic disorders. Trends Genet 2002, 18:74-82.

126. Gu W, Zhang F, Lupski JR: Mechanisms for human genomic rearrangements. Pathogenetics 2008, 1:4.

127. Liu P, Carvalho CM, Hastings PJ, Lupski JR: Mechanisms for recurrent and complex human genomic rearrangements. Curr Opin Genet Dev 2012, 22:211-220.

128. Lieber MR: The mechanism of human nonhomologous DNA end joining. J Biol Chem 2008, 283:1-5.

129. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 2011, 144:27-40.

130. Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M, Cuppen E: Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. Hum Mol Genet 2011, 20:1916-1924.

131. Kloosterman WP, Cuppen E: Chromothripsis in congenital disorders and cancer: similarities and differences. Curr Opin Cell Biol 2013, 25:341-348.

132. Korbel JO, Campbell PJ: Criteria for inference of chromothripsis in cancer genomes. Cell 2013, 152:1226-1236.

133. Liu P, Erez A, Nagamani SC, Dhar SU, Kolodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA, et al: Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell 2011, 146:889-903.

134. Crasta K, Ganem NJ, Dagher R, Lantermann AB, Ivanova EV, Pan Y, Nezi L, Protopopov A, Chowdhury D, Pellman D: DNA breaks and chromosome pulverization from errors in mitosis. Nature 2012, 482:53-58.

135. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005, 33:D514-517.

136. Feuk L, Marshall CR, Wintle RF, Scherer SW: Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 2006, 15 Spec No 1:R57-66.

137. Harakalova M, van Harssel JJ, Terhal PA, van Lieshout S, Duran K, Renkens I, Amor DJ, Wilson LC, Kirk EP, Turner CL, et al: Dominant missense mutations in ABCC9 cause Cantu syndrome. Nat Genet 2012, 44:793-796.

138. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al: Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 2010, 42:30-35.

139. Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, Kremer H, van der Burgt I, Crosby AH, Ion A, Jeffery S, et al: Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat Genet 2001, 29:465-468.

140. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al: Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 2008, 320:539-543.

141. Zhang Y, Haraksingh R, Grubert F, Abyzov A, Gerstein M, Weissman S, Urban AE: Child development and structural variation in the human genome. Child Dev 2013, 84:34-48.

142. Bentires-Alj M, Kontaridis MI, Neel BG: Stops along the RAS pathway in human genetic disease. Nat Med 2006, 12:283-285.

143. Croce CM: Oncogenes and cancer. N Engl J Med 2008, 358:502-511.

144. Hahn WC, Weinberg RA: Rules for making human tumor cells. N Engl J Med 2002, 347:1593-1603.

145. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, et al: Integrating common and rare genetic variation in diverse human populations. Nature 2010, 467:52-58.

146. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: An integrated map of genetic variation from 1,092 human genomes. Nature 2012, 491:56-65.

147. Sherry ST, Ward M, Sirotkin K: dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 1999, 9:677-679.

148. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, et al: DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res 2013, 41:D936-941.

149. Cooper GM, Shendure J: Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 2011, 12:628-640.

150. Gibson G: Rare and common variants: twenty arguments. Nat Rev Genet 2011, 13:135-145.

151. Iyengar SK, Elston RC: The genetic basis of complex traits: rare variants or “common gene, common disease”? Methods Mol Biol 2007, 376:71-84.

152. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, et al: Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 2011, 43:519-525.

153. Stranger BE, Stahl EA, Raj T: Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 2011, 187:367-383.

154. Stunnenberg HG, Hubner NC: Genomics meets proteomics: identifying the culprits in disease. Hum Genet 2013.

155. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008, 9:356-369.

156. On beyond GWAS. Nat Genet 2010, 42:551.

157. Liu LY, Fox CS, North TE, Goessling W: Functional validation of GWAS gene candidates for abnormal liver function during zebrafish liver development. Dis Model Mech 2013, 6:1271-1278.

158. Soon WW, Hariharan M, Snyder MP: High-throughput sequencing for biology and medicine. Mol Syst Biol 2013, 9:640.

159. Castellana N, Bafna V: Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteomics 2010, 73:2124-2135.

160. Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, Nguyen N, Ollikainen N, Rodriguez J, Wang J, et al: Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. Genome Res 2008, 18:1133-1142.

161. Altelaar AF, Munoz J, Heck AJ: Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 2013, 14:35-48.

162. Civelek M, Lusis AJ: Systems genetics approaches to understand complex traits. Nat Rev Genet 2014, 15:34-48.

163. Visel A, Rubin EM, Pennacchio LA: Genomic views of distant-acting enhancers. Nature 2009, 461:199-205.

164. Kumar V, Wijmenga C, Withoff S: From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. Semin Immunopathol 2012, 34:567-580.

165. Pennisi E: The Biology of Genomes. Disease risk links to gene regulation. Science 2011, 332:1031.

166. Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK: Effect of natural genetic variation on enhancer selection and function. Nature 2013, 503:487-492.

167. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV, et al: Extensive variation in chromatin states across humans. Science 2013, 342:750-752.

168. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, et al: Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science 2013, 342:744-747.

169. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK: Identification of genetic variants that affect histone modifications in human cells. Science 2013, 342:747-749.

170. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, et al: Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 2013, 342:1235587.

171. Jendrzejewski J, He H, Radomska HS, Li W, Tomsic J, Liyanarachchi S, Davuluri RV, Nagy R, de la Chapelle A: The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. Proc Natl Acad Sci U S A 2012, 109:8646-8651.

172. Martin L, Chang HY: Uncovering the role of genomic “dark matter” in human disease. J Clin Invest 2012, 122:1589-1595.

173. Bhartiya D, Jalali S, Ghosh S, Scaria V: Distinct Patterns of Genetic Variations in Potential Functional Elements in Long Noncoding RNAs. Hum Mutat 2014, 35:192-201.

174. Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, Almeida R, Zhernakova A, Reinmaa E, Vosa U, et al: Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PLoS Genet 2013, 9:e1003201.

175. Kioussis D, Vanin E, deLange T, Flavell RA, Grosveld FG: Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. Nature 1983, 306:662-666.

176. Kleinjan DA, van Heyningen V: Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 2005, 76:8-32.

177. Belloni E, Muenke M, Roessler E, Traverso G, Siegel-Bartelt J, Frumkin A, Mitchell HF, Donis-Keller H, Helms C, Hing AV, et al: Identification of Sonic hedgehog as a candidate gene responsible for holoprosencephaly. Nat Genet 1996, 14:353-356.

178. Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney S, Gordon CT, et al: Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat Genet 2009, 41:359-364.

2

Systematic biases in DNA copy number originate from isolation procedures

Sebastiaan van Heesch[1], Michal Mokry[1,2], Veronika Boskova[1], Wade Junker[3], Rajdeep Mehon[4], Pim Toonen[1], Ewart de Bruijn[1], James D Shull[3,5], Timothy J Aitman[4], Edwin Cuppen[1,6] and Victor Guryev[1,7]

[1] Hubrecht Institute and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands
[2] Current affiliation: Laboratory of Pediatric Gastroenterology, Wilhelmina Children's Hospital, University Medical Centre, Lundlaan 6, 3584 EA Utrecht, The Netherlands
[3] Department of Genetics, Cell Biology and Anatomy, 985805 University of Nebraska Medical Center, Omaha, Nebraska, 68198-5805, USA
[4] Medical Research Council Clinical Sciences Centre, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London, W12 0NN, UK
[5] Current affiliation: McArdle Laboratory for Cancer Research, Department of Oncology, University of Wisconsin, Madison, Wisconsin, 53706-1599, USA
[6] Department of Medical Genetics, UMC Utrecht, Universiteitsweg 100, 3584 GG Utrecht, The Netherlands
[7] Current affiliation: Laboratory of Genome Structure and Ageing; European Research Institute for the Biology of Ageing; RuG and UMC Groningen, Antonius Deusinglaan 1, 9713 AV, Groningen, The Netherlands

Adapted from Genome Biology 2013, 14:R33

Abstract

The ability to accurately detect DNA copy number variation in both a sensitive and quantitative manner is important in many research areas. However, genome-wide DNA copy number analyses are complicated by variations in detection signal. While GC content has been used to correct for this, here we show that coverage biases are tissue-specific and independent of the detection method as demonstrated by next-generation sequencing and array CGH. Moreover, we show that DNA isolation stringency affects the degree of equimolar coverage and that the observed biases coincide with chromatin characteristics like gene expression, genomic isochores, and replication timing. These results indicate that chromatin organization is a main determinant for differential DNA retrieval. These findings are highly relevant for germline and somatic DNA copy number variation analyses.

Introduction

The ability to accurately detect DNA copy number variation (CNV) in both a sensitive and quantitative manner is important in many research areas. While the detection of CNVs previously relied on low resolution techniques like quantitative PCR or MLPA, high-resolution array-based comparative genomic hybridization (aCGH) and next-generation sequencing (NGS)-based depth of read coverage (DOC) approaches [1] now allow for detailed genome-wide analyses. However, both aCGH and DOC are complicated by the presence of 'wave patterns' in the raw data where the measurement deviates systematically from equimolar representation. These regions span up to tens of megabases and pose challenges on CNV calling. To reduce the number of false-positive calls introduced, algorithms were designed to suppress wave effects [2-6]. In these studies, quantity of DNA during hybridization, dye-biases, enzymatic effects, and correlations with GC content were proposed as the main contributors to the wave patterns. However, understanding the source of the observed patterns is important for reliable genome-wide analyses based on aCGH and NGS techniques.

Results and discussion

To discover the source of unequal DNA representation in genomic data we performed pairwise aCGH analyses comparing all possible combinations of DNA samples isolated from blood, brain, liver, and testis from two rats from different inbred strains. We observed large-scale tissue-specific variation in hybridization intensities that were reproducible between strains and consistent in dye-swap experiments (Figure 1A). Fold-changes for this variation could computationally be defined as tissue-specific CNVs (within the same strain) and were typically much lower than for germline CNVs (between strains). Even though the amplitude of variation did not exceed 30%, the reproducibility of tissue-specific differences between multiple rat strains was very high, both in terms of pattern and magnitude (Figure 1A). Theoretically, these patterns could reflect somatic copy

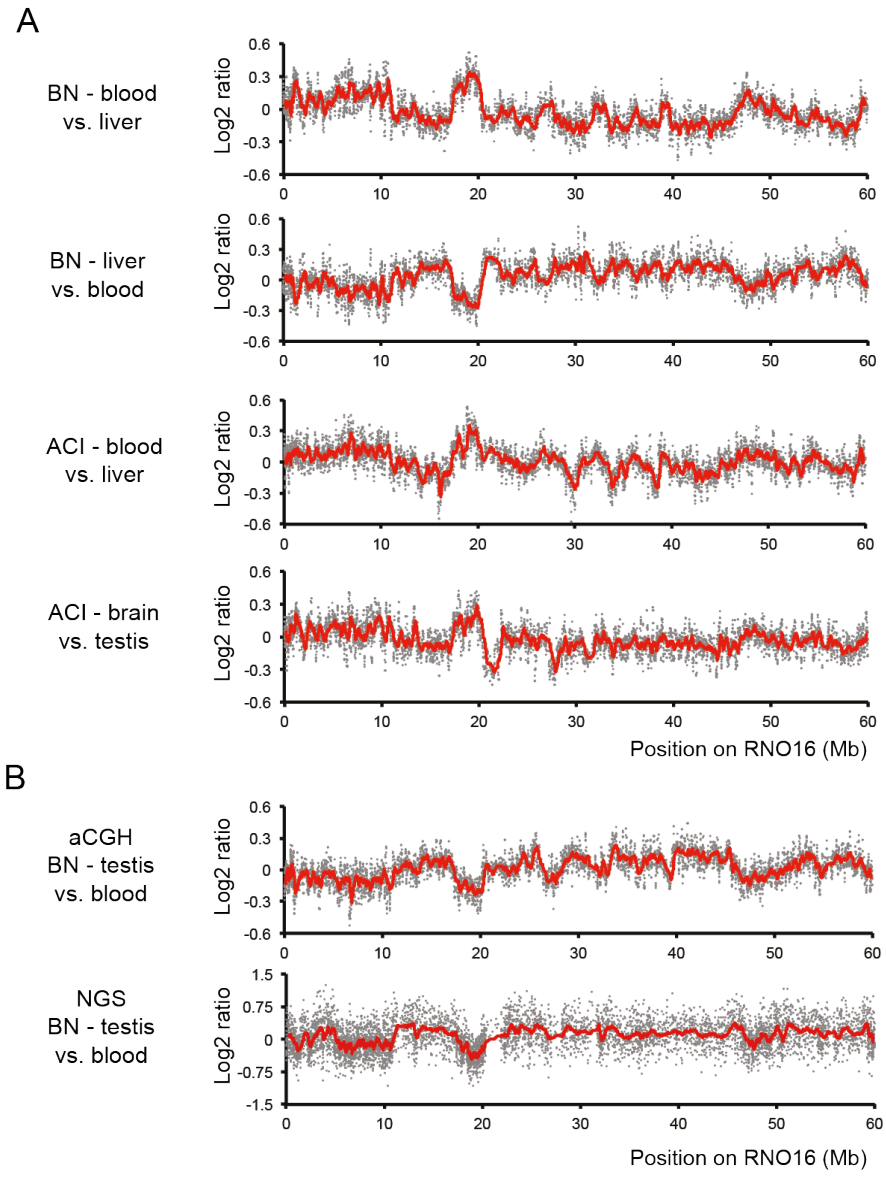


Figure 1 - Reproducible patterns in genome-wide aCGH and NGS data. (**A**) Pairwise aCGH analysis results of blood, liver, brain, and testis samples for rat chromosome 16. For all panels, the variability in log2 ratios is displayed (each dot represents the median value over 100 consecutive probes). The top two panels show dye swap aCGH (Nimblegen) results using blood and liver samples from a single animal (Brown Norway strain). The third panel shows the comparison of blood and liver from an animal from a different inbred strain (ACI). The bottom panel shows the aCGH analysis results between brain and testis of that same ACI animal. (**B**) Comparison of aCGH hybridization signal with NGS depth of coverage analysis results. DNA isolated from the testis and from blood of the same animal was analyzed by aCGH (Nimblegen) and by low-pass next-generation sequencing (6.3-7.2 M reads; 0.075× - 0.086× genome coverage).

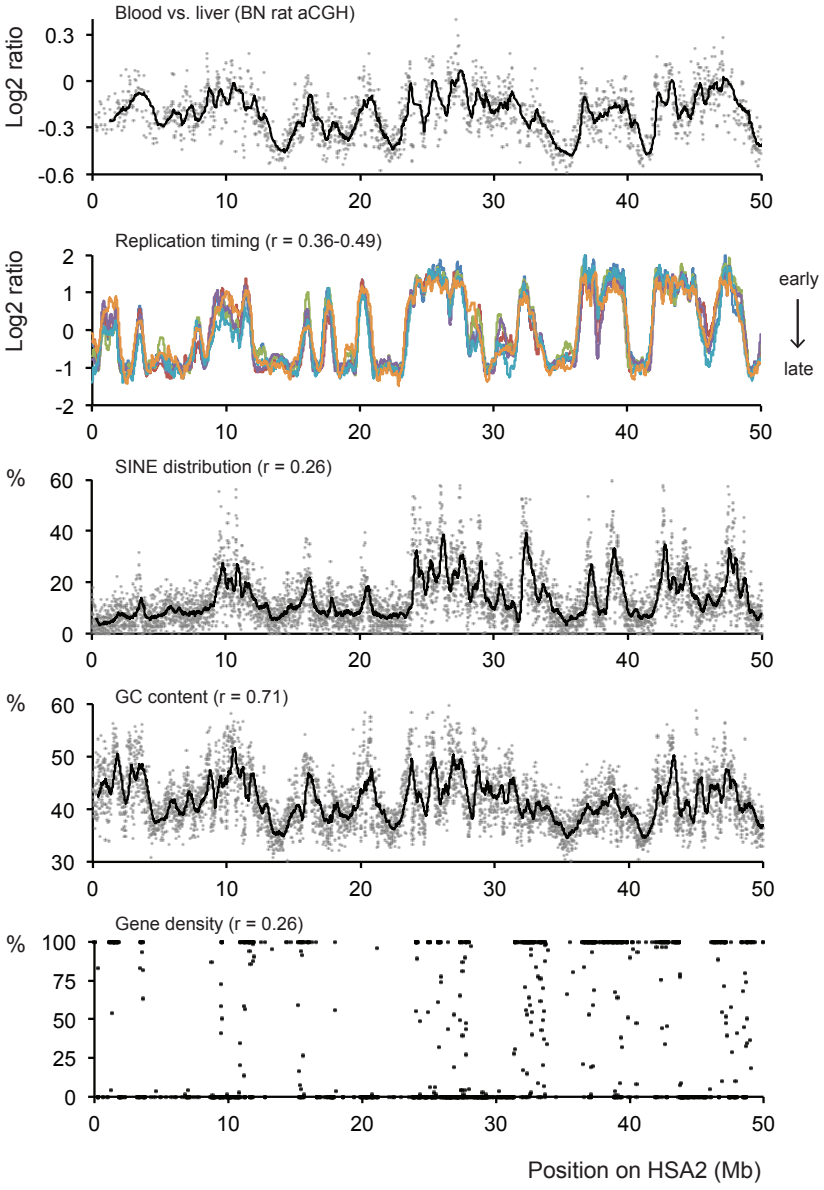


Figure 2 - Correlation of DNA content variability with genome characteristics. The middle panel shows aCGH results comparing BN blood versus BN liver DNA and is aligned with replication timing (early replication is represented by high values, data obtained from Ryba et al. [11]), SINE distribution, GC content (100 kb windows), and gene density. Pearson correlation scores (r) are given per comparison. For each, P values are < 0.001. For this visualization, genomic positions of rat aCGH data were translated to positions on the first 50 Mb of human chromosome 2 (HSA2) to be able to compare rat data with human data on replication timing.

number changes, in line with recently observed somatic heterogeneity [7-9]. Nevertheless, systematic artifacts of the methods used might also underlie such observations. In support of a potential systematic artifact we noted that the genomic regions involved are often megabases in size, while regular CNVs are typically much shorter. Although aCGH analyses using different platforms (Nimblegen and Agilent) and labeling techniques revealed highly similar patterns (not shown), shared artifacts associated with aCGH such as dye or sequence-dependent hybridization effects cannot be excluded. Therefore, we performed a depth of coverage analysis on four tissues from a single animal using NGS-based low-pass whole genome sequencing (5-10 M reads per sample) (Figure 1B). Interestingly, both aCGH and NGS show highly similar DNA content patterns (r2 = 0.71, P <0.001, Additional file 1), excluding previously proposed array-specific artifacts [2-5,10] as the sole basis for the observed patterns and suggesting a common source for the observed variation.

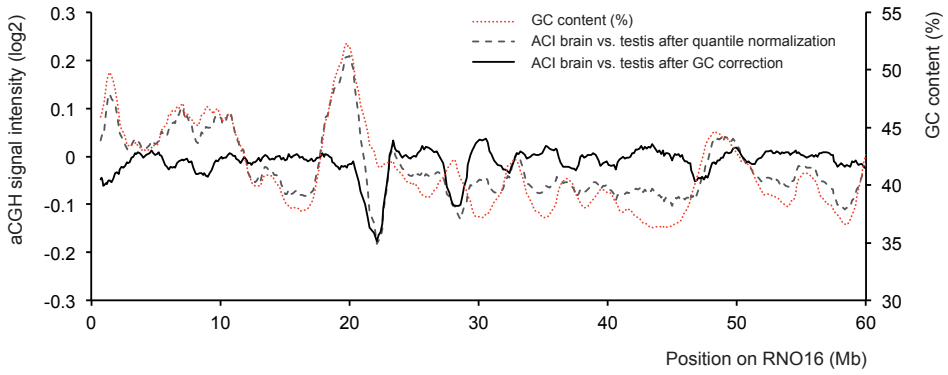


Figure 3 - GC correction reduces, but does not diminish tissue-specificity in aCGH signal intensity. The red line depicts the GC content in percentage (secondary y-axis) across rat chromosome 16. The dashed black line shows the log2 aCGH signal intensity of brain versus testis (ACI rat strain). The black line shows the signal intensity after GC correction. Specific peaks at high GC regions are visibly removed (for example, at 18 Mb), while others are not (for example, at 28 Mb).

Systematic analysis of genomic regions with tissue-specific differences in aCGH hybridization and DOC signal revealed several interesting characteristics. A very clear correlation was found with replication timing [11], gene density, presence of SINE elements, and the relative GC content, which is strongly related to isochores [12, 13] (Figure 2, Additional file 2). GC content has been documented to affect a wide range of molecular biological techniques, including PCR and next-generation sequencing [6] and may thus explain part of the observed patterns. Regional high GC content was recently described to affect the thermostability of DNA, resulting in ultra-fastened regions that affect amplification [14]. However, as DNA content is assumed to be largely the same in every cell, the GC content alone cannot

explain the observed tissue-specific patterns or differences in signal magnitude (Figure 1). When we perform a GC correction on the aCGH data, this flattens out large parts of the pattern, as expected based on the high correlation with GC. However, the GC correction alone is insufficient to flatten the profile between different tissues from the same animal (Figure 3). As we used asynchronous whole tissue samples with only a very small amount of actively proliferating cells, early replicating genomic regions are also unlikely the cause of apparent copy number gains. Intriguingly, replication timing has been shown to correlate with retrotransposon content, genome isochores, and gene expression activity [15], and all of these factors are known to be highly related to chromatin status. Therefore, we hypothesized that tissue-specific chromatin organization may explain the observed correlations and that non-equimolar representation might be due to DNA retrieval artifacts that result in differential representation of euchromatin compared to more densely packed heterochromatin. In support of this, we do observe prominent tissue-specific gene expression in regions with higher apparent tissue-specific copy number status (Additional file 3).

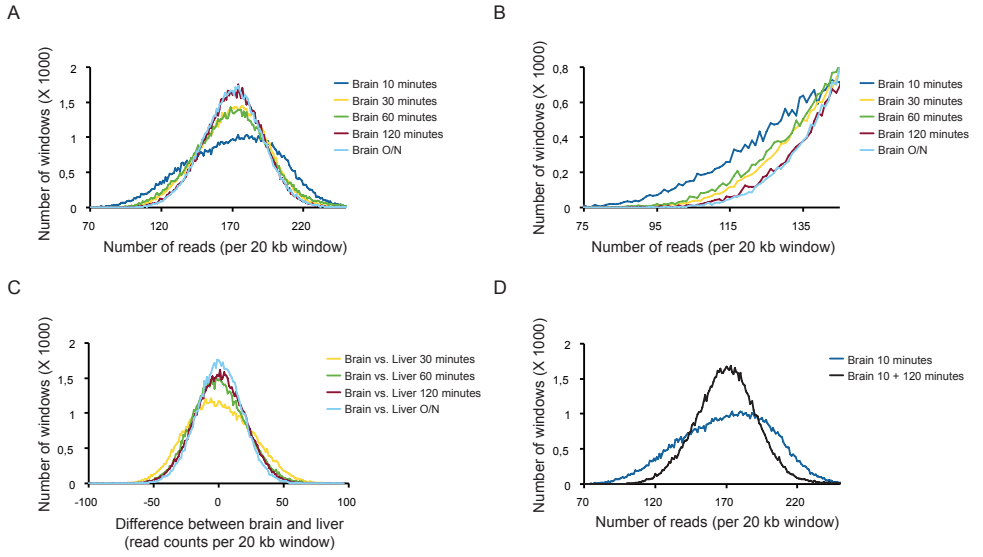


Figure 4 - The duration of proteinase K exposure affects the evenness of genome-wide read distribution. (**A**) For five different durations of lysis (10, 30, 60, 120 min, and overnight (O/N)), the evenness of coverage is determined by calculating the number of 20 kb windows and the number of sequencing reads therein. The y-axis displays the genome-wide number of windows (X 1,000) while the x-axis depicts the number of quantile normalized reads. The width of the curve shows the genome-wide variation in read-depth between all windows. (**B**) Zoomed-in region of (A) to illustrate the read differences in windows with relatively difficult to cover genomic regions. (**C**) Genome-wide tissue-specific differences in read-depth per window are displayed for brain and liver at four different durations of lysis (30, 60, 120 min, and overnight). (**D**) Comparison of the genome-wide read distribution for a sample treated 10 min with proteinase K (blue line), and the exact same DNA sample after 120 min of extra proteinase K treatment (black line).

To study a potential bias resulting from differential chromatin status and introduced during the DNA isolation procedure, we first isolated DNA using standard phenol/chloroform extraction procedures and a commercial DNA isolation kit. aCGH was used to measure potential differences in relative DNA content between the two extraction methods but no significant differences were observed (Additional file 4). Next, we modulated the stringency of extraction by varying proteinase K treatment conditions prior to phenol/chloroform extraction, and used NGS fragment sequencing to determine the DNA recovery patterns across the whole genome. We compared five different lysis durations in the presence of proteinase K and observed that increased duration of treatment improved the evenness of read distribution across the whole genome (thus lowering the wave-amplitude; Figure 4A). Especially in the more difficult to cover regions the increased treatment duration improved coverage (Figure 4B). Next, we determined if the increased duration of the treatment also reduced the tissue-specific differences as depicted in Figure 1. By comparing sequenced DNA from homogenized brain and liver samples of the same animal at four time points, we indeed find that an increased lysis time results in smaller tissue-specific differences (Figure 4C), although it should be noted that biases are not removed. In agreement with our previous observations, the results of copy number profiling of brain and liver samples are affected by proteinase K treatment duration, even after GC correction. While segments totaling to 45 Mb show copy number differences of at least 10% after a 30-min proteinase K treatment, only 1.3 Mb exhibit changes of this scale when treatment is done overnight (Additional file 5).
These results demonstrate that the observed wave patterns are the result of combined tissue-specific DNA isolation biases. As the magnitude, but not the pattern, of the biases decreases with longer proteinase K treatment (Additional file 6), we postulate that DNA retrieval effects are due to differences in degradation of DNA/protein complexes, which subsequently results in depletion of stable aggregates by early precipitation or separation into the phenol phase. Densely packed heterochromatic regions, but also nuclear lamina attached chromatin, are likely to be most affected by such process.

To test whether the DNA in the under-represented genomic regions was simply absent from the sample, or just inaccessible for subsequent applications like sequencing or aCGH, we modulated the DNA isolation experiments even further. First, DNA was extracted after only 10 min of lysis in the presence of proteinase K. After one initial round of phenol/chloroform extraction and precipitation, the sample was divided in half. One part was treated with proteinase K for 2 additional hours, while the other was used as a control and left untreated. We subsequently extracted the DNA from both samples using a second round of phenol/chloroform extraction. Surprisingly, the NGS data show that an additional 2 h of treatment dramatically improves the evenness of the genome-wide coverage as compared to the control sample (Figure 4D), now resembling the read distribution of samples that were treated for a minimum of 2 h. This suggests either that inaccessible DNA was present after the first

phenol/chloroform extraction and made accessible by the additional proteinase K treatment, or that the second phenol/chloroform purification step removed additional protein-bound DNA from the control sample. In any case, these experiments further demonstrate that equimolar DNA representation is affected by differences in DNA isolation conditions.

Conclusions

We demonstrate that DNA isolation procedures can introduce a systematic bias that contributes to the wave effects in aCGH data and the variation in coverage depth in NGS data. We show that extended lysis with proteinase K treatment results in: (1) more even representation of NGS reads across the genome; (2) more similar representation of DNA derived from different tissue sources; and (3) improved DNA content uniformity for a previously undertreated DNA sample. Our data show that the basis for the observed bias is tissue-specific and related to specific chromatin characteristics. Interestingly, from the four tissues that we sampled in this study, brain showed the lowest variation in NGS read coverage. This could reflect the diversity of cell types within this tissue and the associated increased variety of chromatin conformations. More homogeneous tissues like blood and liver exhibited the largest bias in read coverage (Additional file 7), again supporting a cell type-specific origin of the effects rather than primary DNA characteristics. Tissue-specific chromatin characteristics could originate from protein-DNA interactions, 3D organization, and epigenetic modifications.

The observations presented in this study are relevant for a wide range of genomics techniques. Obviously, the described artifacts affect the accuracy of CNV detection [16,17], in particular somatic CNV analyses such as in cancer where sample heterogeneity requires accurate detection of relatively small changes. Furthermore, genome-wide nucleotide variation analyses using next-generation sequencing may also be affected, as depleted regions will have lower sequencing coverage, which results in lower reliability of variant calling. Accurate experimental reflection of the original amounts of DNA is also important for other genomics techniques, as was recently demonstrated for ChIP-seq experiments [18]. As none of the methods or conditions tested could completely remove the signal bias, special care should be taken to control for potential DNA isolation and tissue-specific effects in experiments involving quantitative DNA interpretation. Furthermore, detection of somatic copy number variation will require independent measurements, for example, using allele imbalances [7,8].

Materials and methods

Isolation of genomic DNA. Tissues were collected from BN (BN/Crl, Charles River), WU (HsdCpb:WU), and ACI rats (ACI/Seg/Hsd, Harlan Laboratories B.V., The Netherlands) 6 weeks of age, snap frozen and powdered. A total of 30 mg input material was used for strain and tissue comparisons, 100 mg input material was used for DNA isolation methods comparisons.

DNA was isolated using standard phenol/chloroform extraction (1:1, pH 7.9) or Qiagen DNeasy Blood & Tissue kit (cat.no. 69506). Tissues were lysed in lysis buffer (100 mM Tris-HCl pH 8.0 200 mM NaCl, 0.2% SDS, 5 mM EDTA) using a Kontes Dounce tissue grinder (Kimble and Chase, 885300-0002) and incubated for 2 h at 50°C in the presence of RNase A (50 µg/mL) and proteinase K (100 µg/mL). For WU rat brain and liver, additional proteinase K conditions were tested (10 min, 30 min, 60 min, 120 min, and overnight in lysis buffer). These time series were followed by two rounds of standard phenol/chloroform extraction with in-between precipitation of the DNA (1x phenol, 1x phenol/chloroform, and 1x chloroform). DNA precipitation was done with 3 volumes of pure ethanol in the presence of 1/10 volume sodium acetate (3M). Pelleted DNA was washed with 70% ethanol and dissolved in 10 mM Tris pH 8.0. For the additional proteinase K treatment experiment, 50% of the DNA that was extracted after a 10-min lysis was treated for an additional 120 min of proteinase K (100 µg/mL) in the lysis buffer described above. Next, both samples were cleaned during a second round of phenol/chloroform and ethanol precipitation, similar to the other samples in the time series. For isolation of blood DNA with the Qiagen kit, all steps were performed exactly according to manufacturer's instructions (Qiagen DNeasy Blood & Tissue handbook, 07/2006). DNA quality and quantity of all isolations were measured using NanoDrop ND-1000 (Thermo scientific) and a Qubit Quant-iT™ dsDNA broad range assay (Invitrogen).

Array comparative genomic hybridization (aCGH). NimbleGen whole genome tiling path arrays covering the complete, non-repetitive part of the rat genome were used. The 2.1 M probe arrays had an average probe spacing of 1 probe per 1.3 kb and a GC-content close to 50%. For strain and tissue comparisons, DNA derived from tissues of BN and ACI rats was used for hybridization. The exact quantity of DNA recommended by NimbleGen was used (2 µg input for sonication, 1 µg input for exo- klenow mediated Cy3 and Cy5 labeling, 13 µg for hybridization). DNA labeling (NimbleGen dual-color DNA labeling kit), array hybridization (HX1 mixers, NimbleGen hybridization system 4), washing (NimbleGen wash buffer kit), and scanning were performed exactly according to manufacturer's instructions (NimbleGen Arrays User's Guide - CGH analysis Version 6.0). Image analysis, data normalization, and plotting were performed using NimbleScan 2.4 software using parameters preset by the manufacturer. For platform and extraction method comparisons, Agilent custom designed tiling path arrays (4x44 k, ± 1.5 kb probe spacing) were designed for the complete rat chromosome 14 (RNO14). aCGH DNA preparation steps and array hybridization were performed according to manufacturer's instructions (Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis V4.0).

SOLiD mate-pair sequencing and depth of coverage analysis for the sequencing aCGH comparison. For the mate-pair sequencing data presented in Figure 1, 118 microgram of genomic DNA was fragmented by incomplete digestion during a time series of 15 s to 25

min with the *Alu I* restriction enzyme (Promega, R6281). Time points were pooled and the fragmented DNA was loaded on a 1% agarose gel for gel excision of 1-3 kb fragments. Mate-pair libraries were prepared according to the Applied Biosystems User's Guide (12/2007 4391587 Rev. B) and sequenced on SOLiD V2. Sequencing data were mapped against Rat genome assembly RGSC3.4 by BWA 0.5.9 software [19] (parameters -c -l 25 -n 2 -k 6). We calculated the number of reads using genomic windows containing 100 kb of genome sequence (excluding sequence gaps). The number of reads in each sample was normalized as reads per million sequenced reads. These normalized values were used for the calculation of log ratios and plotting. In total, 6,325,428 reads were used for blood, 6,600,672 for brain, 6,764,588 for liver, and 7,183,059 for testis. These numbers equal a low-pass genome coverage ranging between 0.075x and 0.086x.

SOLiD fragment sequencing and depth of coverage analysis for the time series. For the time series experiments presented in Figure 5, barcoded fragment libraries were produced on an automated system (BioMek), introducing no variation in the library preparation procedure. One microgram of DNA was used as input and libraries were prepared exactly according to manufacturer's instructions for SOLiD 5500XL library preparation. SOLiD libraries were pooled equimolary, quality assessed, and size selected on the Caliper XT system. Sequencing reads were aligned to the Rat reference genome RGSC3.4 using BWA 0.5.9 [19] (parameters -c -k 2 -l 25 -n 10). PCR duplicates were marked in the alignments and were not used in the analysis, resulting in 10 to 35 M unique and unambiguously mapped reads per time point. For tissue comparisons, the coverage of each library was normalized by random removal of reads to 10 M of unambiguously mapped tags (0.2x genome coverage), which corresponds to the liver library with the least amount of mapped reads. For the additional proteinase K treatment comparisons, only brain samples were used and these could thus be normalized to 14.8 M reads (0.3x genome coverage; limited by the brain library with the least number of reads). The genome was partitioned into windows each containing 20 kb of NGS-accessible sequence (excluding repeats and gaps). The read count and GC content were determined for each window and library. GC-correction: read counts were adjusted for each library by normalization against the median read count in 100 genomic windows with most similar GC content using the following formula: $N_{corr}=N_{med}*N_{obs}/N_{medGC}$ where: N_{corr}, GC-corrected number of reads; N_{obs}, observed number of reads; N_{med}, median reads per window for this library; and N_{medGC}, median number of reads in 100 windows with the most similar GC content. After GC correction, potential somatic copy number changes were determined using a dynamic window approach (DWAC-seq).

Correlation of DNA content variability with various genome characteristics and gene expression. GC content, repeat, and gene annotation were extracted from the Ensembl database [20] (v.69). Gene expression data were exported from the UCSC genome browser

[21].

Abbreviations

Array CGH/aCGH: array-comparative genome hybridization; ChIP-seq: chromatin immunoprecipitation sequencing; CNV: copy number variation; DOC: depth of coverage; Mb: megabase; MLPA: multiplex ligation-dependent probe amplification; NGS: next-generation sequencing; PCR: polymerase chain reaction; SINE: short interspersed nuclear element.

Authors' contributions

SvH, MM, VB, WJ, RM, PM, EdB, and EC carried out the molecular biological studies. SvH, MM, VB, and EC performed DNA isolation for next-generation sequencing experiments, SvH, MM, WJ, and RM performed aCGH experiments. PT collected tissues and EdB performed next-generation sequencing. SvH and VG performed data analyses. EC and VG conceived of the study and coordinated experiments. JDS and TJA contributed reagents and data, critically discussed results and directions of the research and contributed to drafting the manuscript. SvH, EC, and VG wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Ewan Birney for his suggestion to explore DNA isolation artifacts as a potential source for the putative somatic CNV patterns observed in our experiments. This work was financially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. HEALTH-F4-2010-241504 (EURATRANS) to EC and TJA and the NWO-CW TOP grant (700.58.303) to EC, the NGI Horizon program grant (93519030) to VG, and the NIH grant R01-CA77876 to JS. TJA acknowledges intramural MRC Programme support.

Data availability

All sequencing data are available from the Sequence Read Archive (SRA) at EBI under accession number (ERP001927). Array CGH data are available from the Gene Expression Omnibus (GEO) database at NCBI under accession number (GSE45308). Whole genome aCGH plots for tissue comparisons are available as Additional file 8 (ACI blood versus liver and brain versus testis; BN blood versus liver and liver versus blood).

Additional files

Additional files can be found online at http://genomebiology.com/2013/14/4/R33 (doi:10.1186/gb-2013-14-4-r33)

References

1. Ceulemans S, van der Ven K, Del-Favero J: Targeted screening and validation of copy number variations. Methods Mol Biol 2012, 838:311-328.

2. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K: Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 2008, 36:e126.

3. Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, Zhang J, Liu G, Ihara S, Nakamura H, Hurles ME, Lee C, Scherer SW, Jones KW, Shapero MH, Huang J, Aburatani H: Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res 2006, 16:1575-1584.

4. Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET, Carter NP, Tavare S, Hurles ME: Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol 2007, 8:R228.

5. van de Wiel MA, Brosens R, Eilers PH, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, Ylstra B: Smoothing waves in array CGH tumor profiles. Bioinformatics 2009, 25:1099-1104.

6. Benjamini Y, Speed TP: Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 2012, 40:e72.

7. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, Hutchinson A, Deng X, Liu C, Horner MJ, Cullen M, Epstein CG, Burdett L, Dean MC, Chatterjee N, Sampson J, Chung CC, Kovaks J, Gapstur SM, Stevens VL, Teras LT, Gaudet MM, Albanes D, Weinstein SJ, Virtamo J, Taylor PR, Freedman ND, Abnet CC, Goldstein AM, Hu N, et al.: Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet 2012, 44:651-658.

8. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, Wei Q, Wang LE, Lee JE, Barnes KC, Hansel NN, Mathias R, Daley D, Beaty TH, Scott AF, Ruczinski I, Scharpf RB, Bierut LJ, Hartz SM, Landi MT, Freedman ND, Goldin LR, Ginsburg D, Li J, Desch KC, Strom SS, et al.: Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet 2012, 44:642-650.

9. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, Faulkner GJ: Somatic retrotransposition alters the genetic landscape of the human brain. Nature 2011, 479:534-537.

10. Liu GE, Hou Y, Robl JM, Kuroiwa Y, Wang Z: Assessment of genome integrity with array CGH in cattle transgenic cell lines produced by homologous recombination and somatic cell cloning. Genome Integr 2011, 2:6.

11. Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM: Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res 2010, 20:761-770.

12. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al.: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

13. Smit AF: Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 1999, 9:657-663.

14. Veal CD, Freeman PJ, Jacobs K, Lancaster O, Jamain S, Leboyer M, Albanes D, Vaghela RR, Gut I, Chanock SJ, Brookes AJ: A mechanistic basis for amplification differences between samples and between genome regions. BMC Genomics 2012, 13:455.

15. Schubeler D, Scalzo D, Kooperberg C, van Steensel B, Delrow J, Groudine M: Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. Nat Genet 2002, 32:438-442.

16. Lee KH, Chiu S, Lee YK, Greenhalgh DG, Cho K: Age-dependent and tissue-specific structural changes in the C57BL/6J mouse genome. Exp Mol Pathol 2012, 93:167-172.

17. Piotrowski A, Bruder CE, Andersson R, Diaz de Stahl T, Menzel U, Sandgren J, Poplawski A, von Tell D, Crasto C, Bogdan A, Bartoszewski R, Bebok Z, Krzyzanowski M, Jankowski Z, Partridge EC, Komorowski J, Dumanski JP: Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat 2008, 29:1118-1124.

18. Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, Liu T, Zhang Y, Kim TK, He HH, Zieba J, Ruan Y, Bickel PJ, Myers RM, Wold BJ, White KP, Lieb JD, Liu XS: Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods 2012, 9:609-614.

19. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754-1760.

20. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, et al.: Ensembl 2012. Nucleic Acids Res 2012, 40:D84-90.

21. Pohl AA, Sugnet CW, Clark TA, Smith K, Fujita PA, Cline MS: Affy exon tissues: exon levels in normal tissues in human, mouse and rat. Bioinformatics 2009, 25:2442-2443.

3

Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing

Sebastiaan van Heesch[1], Wigard P Kloosterman[2], Nico Lansu[1], Frans-Paul Ruzius[1], Elizabeth Levandowsky[3], Clarence C Lee[3], Shiguo Zhou[4], Steve Goldstein[4], David C Schwartz[4], Timothy T Harkins[3], Victor Guryev[1,5] and Edwin Cuppen[1,2]

[1] Hubrecht Institute for Developmental Biology and Stem Cell Research, KNAW and UMC Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands
[2] Department of Medical Genetics, UMC Utrecht, Universiteitsweg 100, 3584 GG Utrecht, The Netherlands
[3] Life Technologies Inc., Advanced Applications Group, 500 Cummings Center, Beverly MA, 01915, USA
[4] Laboratory for Molecular and Computational Genomics, Department of Chemistry, Laboratory of Genetics, UW-Biotechnology Center, University of Wisconsin-Madison, Madison WI, 53706, USA
[5] Present address: Laboratory of Genome Structure and Ageing; European Research Institute for the Biology of Ageing; RuG and UMC Groningen, Antonius Deusinglaan 1, 9713 AV, Groningen, The Netherlands

Adapted from BMC Genomics 2013, 14:257

Abstract

Paired-tag sequencing approaches are commonly used for the analysis of genome structure. However, mammalian genomes have a complex organization with a variety of repetitive elements that complicate comprehensive genome-wide analyses. Here, we systematically assessed the utility of paired-end and mate-pair (MP) next-generation sequencing libraries with insert sizes ranging from 170 bp to 25 kb, for genome coverage and for improving scaffolding of a mammalian genome (Rattus norvegicus). Despite a lower library complexity, large insert MP libraries (20 or 25 kb) provided very high physical genome coverage and were found to efficiently span repeat elements in the genome. Medium-sized (5, 8 or 15 kb) MP libraries were much more efficient for genome structure analysis than the more commonly used shorter insert paired-end and 3 kb MP libraries. Furthermore, the combination of medium- and large insert libraries resulted in a 3-fold increase in N50 in scaffolding processes. Finally, we show that our data can be used to evaluate and improve contig order and orientation in the current rat reference genome assembly. We conclude that applying combinations of mate-pair libraries with insert sizes that match the distributions of repetitive elements improves contig scaffolding and can contribute to the finishing of draft genomes.

Introduction

Genome assemblies consist of kilobase- to megabase-sized contiguous sequences of DNA (contigs) that need to be positioned in a correct order and orientation. This ordering of contigs (scaffolding) requires long-range structural information that reaches beyond the boundaries of contigs. Commonly used reference genome assemblies, like those of human [1,2], rat [3], and mouse [4], were all constructed using long-range structural information obtained by Sanger sequencing based applications. For example, mapped large insert clones (e.g., cosmid, fosmid and bacterial artificial chromosomes) and paired-end whole genome shotgun sequencing of plasmids with variable insert sizes contributed to elucidating the complexity of genomes at the structural level. Despite the high quality of these assemblies, tens to thousands of intercontig gaps still persist [3,5,6].

Currently, genomes are frequently sequenced by cost-effective next-generation sequencing (NGS) technologies. However, long-range structural information is often not available from such efforts and would require more costly and toilsome techniques than routine fragment or paired-end sequencing. The absence of long-range information poses significant challenges for dealing with repetitive sequences that often represent 50% of mammalian genomes [1,7]. Emerging technologies like long-read single-molecule sequencing [8] or single-molecule mapping systems like optical mapping [9-11], may eventually help to overcome many of the challenges put forward here. However, application of methods solely based on current NGS technology would be most optimal because such platforms are maturing fast and are

very broadly available. Current NGS platforms are already capable of producing positional information using paired-end (PE) and mate-pair (MP) templates. PE sequencing involves the generation of pairs of sequencing reads derived from both ends of a contiguous DNA fragment. This sequencing modus is currently standard on most platforms but is limited by technology features (e.g. PCR constraints) that typically only allow for insert sizes of less than 500 bp [12]. MP sequencing, however, can provide much longer distance information [13], but requires several molecular sample processing steps to clone DNA fragment ends through a circularization step, making it a relatively laborious approach. Most commonly used MP approaches span 1 to 3 kilobase pairs (kb) and are therefore capable of spanning many repetitive or low complexity sequence elements. However, common repetitive elements [like LINE (L1) elements] in vertebrate genomes can span as much as 8 kb in size (Additional file 1) [7,14], illustrating the need for longer range information for comprehensive analysis of genome structures. To this end, various bioinformatic algorithms like CREST [15] and ALLPATHS-LG [16] have been developed to increase effective PE read span by systematically merging overlapping sequences. Experimentally, novel methods producing larger insert sizes have also been reported [17,18]. While these techniques clearly demonstrate the power of larger distance information, most do have limitations that could interfere with comprehensive analysis (e.g., maximum insert sizes of ~10 kb [17,19], potential biases introduced by enzymatic digestions [18], and relatively laborious or costly approaches that can only produce single fixed insert size libraries [20,21]). Furthermore, a systematic assessment of the utility and combination of different library insert sizes for resolving existing assembly difficulties in complex regions of genomes is currently lacking.

Here, we modified existing MP library construction protocols to allow for the generation of a wide range of small, medium and large insert size mate-pair libraries (3 kb up to 25 kb) and present a systematic comparison of their individual and combined utility for exploring mammalian genome structure. Our results show that two of the medium-sized MP libraries (8 kb and 15 kb) are most efficient for bridging repeats in the rat genome as well as for contig scaffolding. Furthermore, combining the medium-sized MPs with large insert (20-kb and 25-kb) libraries reduces the number of scaffolds by another 25% and results in a 3-fold increase in N50. Our results are useful to define the most optimal experimental paired-read approach to support the de novo assembly of mammalian genomes.

Results

Large insert mate-pair library generation

We constructed MP libraries through modification of the standard SOLiD protocol for mate-pair library construction (Additional file 2), to allow construction of MPs with insert sizes up to 25 kb. We used ~100 μg high-molecular-weight genomic DNA isolated from tissue of a

single Brown Norway rat as starting material for all libraries. Sheared DNA was size-separated by pulsed-field gel electrophoresis [22,23] followed by excision of various fragment sizes from a single lane and conversion into mate-pair libraries. In total, we generated seven different library insert sizes, including six libraries produced with this adapted MP protocol and one PE fragment library that was prepared in a separate experiment (Table 1). Based on paired read mapping, the libraries showed median insert sizes of 170 bp (PE), 3 kb, 5 kb, 8 kb, 15 kb, 20 kb, and 25 kb.

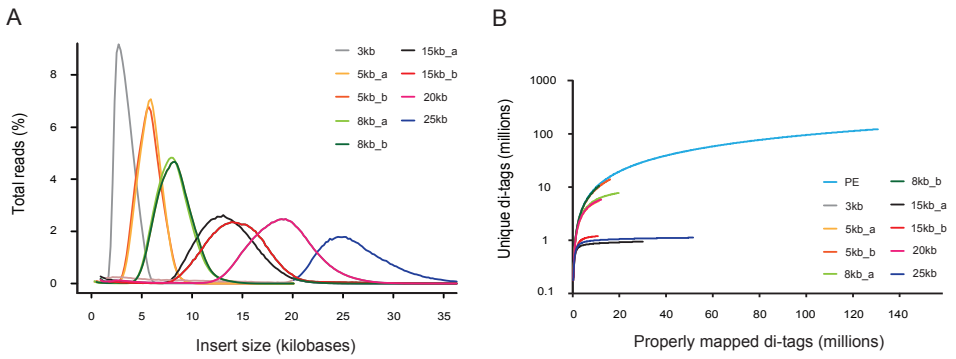


Figure 1 - **MP insert size distribution and library complexity.** (**A**) Insert size distribution of all mate-paired libraries and biological duplicates. Data have been filtered for non-clonal pairs. (**B**) Complexity of each library is depicted by the number of unique read-pairs versus the number of properly mapped read-pairs. On the x-axis, increasing sequencing depth is represented based on actual sequencing data versus the amount of unique information obtained on the y-axis. A plateau indicates that a library has been sequenced to saturation.

To assess library complexity by determining the maximum number of unique reads obtainable from a MP library, two of the large-insert libraries (20-kb and 25-kb) were sequenced to a higher depth in an additional sequencing run. To assess reproducibility of the adapted MP protocol, three libraries (5-kb, 8-kb, and 15-kb) were generated in duplicate from independently isolated, sheared, and separated genomic DNA samples. Insert size distributions of the individually produced replicates were highly consistent (Figure 1A and Table 1). In total, 192.4 million pairs of MP reads and 160 million pairs of PE reads were generated. A total of 62.3 million non-duplicate MP read pairs and 131 million non-duplicate PE reads were consistently mapped against the rat reference genome, resulting in a genome-wide physical coverage of 228.5× (220× for MP libraries and 8.5× for 170-bp PE). Less than 1% of the paired reads were inverted (one of the reads in other orientation than expected) or everted (both reads in other orientation resulting in wrong order of tags) and approximately 10% were mapped remotely (i.e., to a distant genomic position, significantly deviating from what is expected based on the insert-size distribution). Remote, inverted, or everted events represent a mixture of 1) library construction artifacts due to chimeric molecules, 2) errors in the reference genome assembly (misassemblies) and 3) real structural differences between the reference strain

Table 1: Sequencing and coverage statistics for all paired-read libraries

Library name*	Sequenced pairs	Non-duplicate pairs	Unique, consistently mapped	Median insert size (bp)	Physical coverage (x-fold)	5 M reads coverage (x-fold)	PCR cycles **	Relative complexity ***	Inconsistent pairs forming clusters	Inconsistent inter-chromosomal pairs forming clusters
PE	160 M	151 M (95%)	131 M (87%)	166	8.5	0.34	5	>131 M	9.5%	2.0%
3 kb	17.7 M	16.3 M (92%)	15.2 M (93%)	3,208	50	6	14	4.6 M	24.0%	3.2%
5 kb_a	11.9 M	6.0 M (51%)	5.5 M (92%)	5,695	11	10.1	18	4.7 M	46.5%	3.7%
5 kb_b*	16.7 M	14.7 M (88%)	14.0 M (95%)	5,811	28	10.1	13	>16.7 M	47.5%	4.6%
8 kb_a	20.8 M	8.6 M (41%)	7.8 M (91%)	8,298	25	16.2	14	5.7 M	58.4%	6.9%
8 kb_b*	11.8 M	11.2 M (95%)	10.6 M (95%)	8,160	34	16.2	13	>11.8 M	50.6%	5.5%
15 kb_a	31.7 M	1.7 M (5%)	1.0 M (60%)	14,561	6	30.3	21	0.6 M	60.8%	7.6%
15 kb_b*	11.6 M	1.6 M (14%)	1.2 M (73%)	13,556	7	30.3	21	0.7 M	23.2%	3.2%
20 kb	13.3 M	6.7 M (51%)	5.9 M (87%)	19,375	48	40.5	14	4.9 M	41.8%	4.7%
25 kb	56.9 M	2.3 M (4%)	1.1 M (49%)	25,871	11	50.6	17	0.7 M	51.9%	5.4%
TOTAL	352.4 M	220.1 M (62%)	193.3 M (88%)		228.5					

* The _b samples are retrieved from a replicate experiment using an independent DNA isolate from the same animal
** Number of PCR cycles required to retrieve sufficient library molecules in the final adapter-mediated PCR
*** Complexity is defined as minimal sequencing depth (in million clones) at which over half of the pairs are clonal.

and the substrain tested here. The first category typically involves stochastic events that are supported by a single read pair and that are filtered out by requiring multiple independent supporting read pairs for calling.

Library sequencing and quality assessment

Sequencing libraries may suffer from low complexity due to library amplification steps in the protocol. When the proportion of unique library molecules is low due to inefficient molecular reactions or low amounts of input material, sequencing more reads of that same library would not yield any additional information, but only extra copies of previously sequenced molecules (duplicate reads). We assessed the complexity of each library by plotting the number of read-pairs with unique genome coordinates against the total number of all mapped pairs (Figure 1B). In general, the complexity of the small-insert libraries is higher than that of the large-insert libraries, which more quickly saturate to the level where deeper sequencing delivers predominantly non-informative duplicate reads. Duplicate reads do not necessarily affect the utility of the libraries, because these reads are filtered out as a first step in the analysis procedure; however, low complexity does decrease the capacity to obtain sufficient physical genome coverage. Three sample groups can be distinguished in Figure 1B: (1) high-complexity libraries that deliver approximately 100 million unique pairs (PE, 3kb, 5kb_b and 8kb_b), (2) medium-complexity libraries that result in about 10 million unique pairs (5kb_a, 8kb_a and 20kb), and (3) low-complexity libraries resulting in approximately 1 million unique pairs (15kb_a, 15kb_b, 25kb). Several of the low-complexity libraries show a plateau in the curve, indicating that these have been sequenced to saturation (25kb, 15kb_a). For others (5kb_b, 8kb_b), deeper sequencing would be informative.

Library complexity may be influenced by several experimental conditions. When starting with an equal quantity of genomic DNA, fragmentation for a standard PE library provides approximately 140-fold more unique molecules than for a 25-kb library. Furthermore, MP library preparation involves a circularization step (Additional file 2) that becomes less efficient as the size of the molecule increases. Quantification of DNA before and after circularization (and removal of non-circularized molecules) showed a circularization efficiency of up to 37% for libraries below 10 kb and 5–10% for libraries above 10 kb (Additional file 3). Each of these library generation steps has a negative impact on the recovery of material; for example, an input of 10 µg 25-kb size-selected DNA would result in approximately 6 ng (>4,000-fold reduction) of DNA for adapter ligation and subsequent adapter-mediated PCR. As a consequence, more PCR cycles are required for larger insert libraries to obtain sufficient amounts of library DNA for NGS (Table 1). Although the 3- and 5-kb insert libraries could routinely be generated at high complexity, we observed more technical variation for the large insert libraries. For example, the 20-kb library required only 14 PCR cycles during the library preparation procedure and performed well in the complexity analysis (comparable to

5- and 8-kb libraries). The 15- and 25-kb libraries required 21 and 17 cycles, respectively, and resulted in libraries of lower complexity (Figure 1B). These results indicate that the number of required PCR cycles is a very good predictive parameter for library complexity.

The 5-, 8-, and 15-kb libraries were generated in duplicate using DNA isolated from two different tissues of the same animal. The insert size distribution was found to be highly reproducible (Figure 1A), but the library complexity was much more variable between duplicates (Figure 1B). These differences might have been due to differences in DNA quality (e.g. amount of single strand breaks) or purity (e.g. associated protein or small molecule contaminants) of the DNA and subsequent differences in shearing efficiency. Indeed, DNA yields after size fragmentation were as much as 2.5-fold lower for the duplicate DNA sample (data not shown), which systematically resulted in less complex libraries. Most importantly, however, statistics for the amount of consistently mapped read pairs were comparable for all replicates (Table 1), indicating that the mapped unique read pairs were similar in quality (e.g., low chimerism) and insert size. Low complexity in libraries could be circumvented by using larger amounts of input DNA and/or by optimization of shearing conditions to concentrate DNA in the desired size range. In our experiments we aimed for a broad size distribution to be able to simultaneously extract DNA for a range of different sizes. Although the larger insert libraries come with more duplicate reads, far fewer sequencing pairs are required to physically cover the complete genome. It should be noted, that for all MP libraries in the experiments described here more than 10x physical coverage was obtained, including 48x coverage for 20 kb inserts. To assess the value of the various insert size libraries for genome structure analysis, we determined the ability of each library to (1) physically cover the reference genome and overlap various repeat elements, (2) drive contig scaffolding, and (3) fix contig assembly issues in the current genome assembly (errors in contig order and orientation).

Spanning repeats and physical genome coverage

The ability to physically cover a complete genome by sequencing is not only determined by the length of the read, the insert size of the library, and the number of paired reads, but also depends on genome-specific characteristics, like the composition and distribution of repetitive elements. The rat genome is representative for other mammalian genomes and contains 1.24 Gb of repetitive sequences, which is over 49% of the 2.51 Gb in the current reference genome assembly (RGSC 3.4, v.66 [24]). Retrotransposable LINE (L1) elements are the largest class of repeats with a total length of 474.6 Mb (18.9% of the genome), followed by retrotransposons that are flanked by long terminal repeats (LTRs; 220.9 Mb; Figure 2A and B). To evaluate the effect of library insert size on the degree of physical genome coverage, we merged data from duplicate libraries with the same insert size. Although the MP libraries had

far more physical genome coverage, all datasets were normalized to an equal physical genome coverage based on properly mapped and oriented read pairs, which was limited to 8.5x by the available amount of data for the PE library. Next, we determined per library the fraction of bases per contig of the rat reference genome that is physically covered, specifically focusing on repetitive elements. Despite the same physical genome coverage and much higher base coverage, short-insert libraries (PE and 3 kb MP) were much less efficient in spanning long repetitive elements, such as LINEs or LTRs, than larger insert MP libraries (≥ 15 kb) (Figure 2C and D respectively). As expected, PE pairs overlapped hardly any of these elements but also the most widely used 3-kb MP libraries were found to only span approximately half of the 3-kb repeat elements, and only a few elements with sizes above 4 kb. Slightly improved results were observed for the 5-kb and 8-kb MP libraries, where approximately half of the repeats with a matched size could be spanned by at least one mate-pair. The 15-, 20-, and 25-kb libraries spanned over 90% of the repeat elements across the whole size spectrum and all displayed a very similar performance, indicating that there is limited added value for even larger insert sizes.

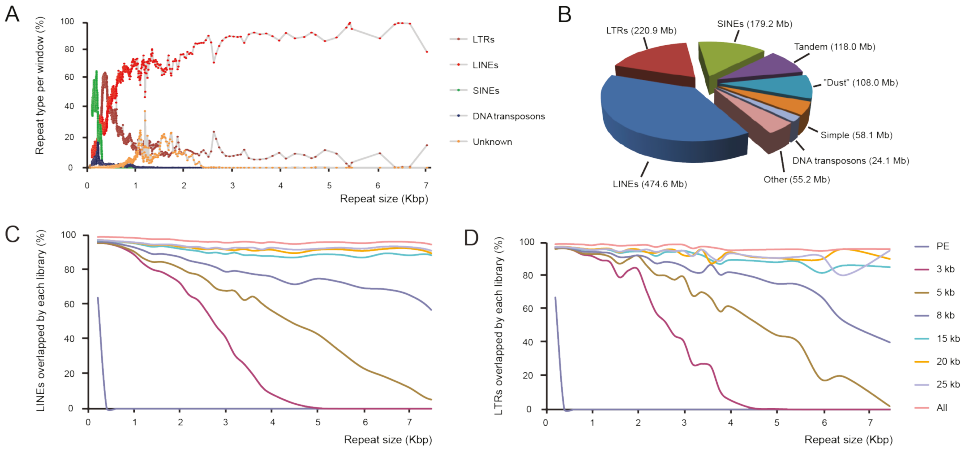


Figure 2 - **Bridging of repeat elements by paired read libraries.** (**A**) The percentage of each repeat type per window of 1000 repeats (y-axis) is shown, relative to the size of each repeat on the x-axis. A higher density of dots indicates the presence of more repeats in the indicated size bin. (**B**) Pie chart of the largest classes of repetitive elements based on their total length (Mb) in the rat genome. Satellite repeats, RNA repeats, and low-complexity repeats are listed as "Other". (**C + D**) Bridging by paired-tag libraries of all annotated LINEs (**C**) and LTRs (**D**) within contigs of RGSC 3.4. The size of LINE elements or LTRs (x-axis) is plotted against the percentage of elements of that specific size that were bridged by one or more read-pairs from each of the libraries. All single library datasets were normalized to 8.5x physical genome coverage.

Contig scaffolding

To evaluate the utility of the various libraries for guiding genome assembly, we simulated the scaffolding step of such process by using the 137,257 contigs from the current rat genome build and the different MP data sets as input for the SSPACE 2.0 [25] software. To allow for library insert-size comparison, we again used the normalized datasets at 8.5x physical

coverage and determined the N50 (at least half of the genome bases are in scaffolds that are equal or exceeding the N50 value) and the number of scaffolds (segments of the genome reference consisting of contigs in known order, separated by gaps) for each individual library and combinations thereof, using the output of the SSPACE software (Figure 3A, Table 2 and Additional file 4). When we consider only the utility of single libraries, the N50 increases from ~38 kb for the PE data to 140–163 kb for the MP libraries of 5 kb and up. PE libraries are not effective in reducing the number of scaffolds as compared to the capillary sequencing-based contigs: a reduction of only 15 scaffolds is obtained (from 137,256 scaffolds in RGSC3.4 to 137,241 scaffolds using the PE data). In contrast, individual MP libraries decreased the number of scaffolds by up to more than 50% (~67,000 for the 5 kb library, which performs best of all individual MP libraries). When considering two insert size libraries, combination of 5 or 8 kb and 20 or 25 kb are most optimal with N50's of ~0.5 Mb. Intriguingly, 3 kb mate-pair libraries, which are most commonly used, showed the worst performance from all MP libraries, also when combined with other libraries. Including all libraries in the scaffolding process results in a further decrease of scaffolds (36,348) with an N50 increase up to 1.3 Mb. Increasing the physical coverage for a single insert library shows to be far less effective than combining libraries with different insert size (Additional file 5 and Additional file 6). For example, when we increase the coverage of the 5-kb insert library to 34x physical coverage,

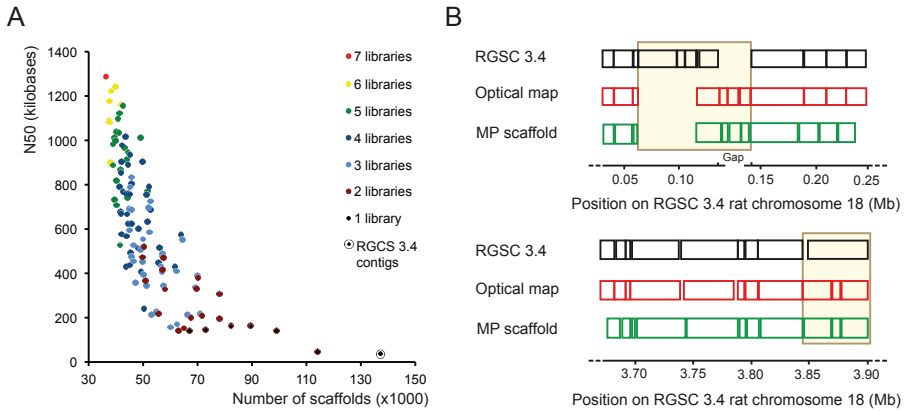


Figure 3 - **Combinations of libraries with different insert sizes improve contig scaffolding.** (**A**) All library data sets were normalized to 8.5x non-clonal physical genome coverage resulting in the use of approximately 130 million pairs for the PE library to several million pairs for the MPs. The scaffold N50 (y-axis) as determined by SSPACE is plotted against the total number of scaffolds (x-axis) for each individual library and for all combinations of libraries. Scaffolding results for the current genome reference (RGSC 3.4) are displayed as well. (**B**) Representative examples of the genomic loci on rat chromosome 18 that show major discordance between optical map and the RGSC 3.4 reference genome. MP-assisted scaffolding restored concordance between sequence scaffolds and optical maps. The top panel (black) represents the reference genome assembly with the vertical lines indicating predicted SwaI sites; the middle panel (red) represents optical map data obtained using SwaI digests; the lower panel represents the rescaffolded genome using the MP data. The indicated positions on chromosome 18 are according to the current RGSC 3.4 assembly. A large region of approximately 75 kb (top panel) that shows low concordance with the predicted path of the optical map (0.065 Mb–0.14 Mb), increased significantly after MP-scaffolding. The bottom panel shows another example of increased resemblance to optical mapping data (3.85 Mb–3.90 Mb). Order and placement of contigs was shifted in the new scaffold resulting in SwaI sites identical to the optical map.

the N50 increases from 141 kb to 262 kb. However, combining 8.5x coverage data for the 5 kb insert library with similar coverage of any other MP library (up to 34x) results in much higher N50 values ranging from ~431 kb up to ~1 Mb.

Reference genome improvement

Finally, we evaluated the value of various MP insert sizes for improving the existing rat genome assembly. To this end, we compared de novo scaffolds constructed using MP data with independently obtained genome-wide optical mapping data. Optical mapping is an integrated system that provides long-range genome structural information by the construction and analysis of genome-wide, ordered restriction maps [9,10,26,27]. We limited the analysis of concordance between sequence scaffolds and optical maps to one of the small rat chromosomes (RNO18), because the fine level optical structural alterations (OSAs) that were automatically called by the optical mapping pipeline [10] were manually curated between sequence scaffolds and optical maps, which required exploration on a case-by-case basis for mediation at the nucleotide-level. We divided the chromosome into 872 100-kb windows and found that 96 out of 872 of such bins harbored structural changes within scaffolds of RNO18 when comparing the MP-updated genome structure with the original genome. The 96 bins contained a total of 199 unique inconsistent connections between contigs within scaffolds. Next, we looked at structural differences between scaffolds of RNO18, based on the comparison of the MP-based scaffolding and the reference genome and observed many more bins to be affected (166/ 872 bins containing a total of 1374 inconsistent links between scaffolds). In total, 236 bins showed one or both types of inconsistent connections. Of these 236 bins, only 106 showed concordance with the reference genome. 130 bins were found to contain OSAs including absence or discordance of alignment between the optical maps and RGSC 3.4 genome assembly (detailed description in Additional file 7 and Additional file 8). Because the optical mapping system constructs ordered restriction maps and does not evaluate genome structure at the nucleotide level, not all discordances detected by the mate-pair analysis are revealed through optical mapping data. For example, small contigs or changes that do not overlap with a SwaI restriction site will not be identified. We explored two of the largest segments with long-range disagreement between optical maps and RGSC3.4 assembly and conclude that MP-assisted re-scaffolding can recover concordance with the independently generated optical maps (Figure 3B). The complete MP data described here has therefore also been used for building the new genome reference of the rat (Rnor5.0, GenBank ID GCA_000001895.3, unpublished results).

Discussion

Here, we show that large insert MP sequencing is a versatile tool for analyzing genomes at the structural level and providing long-range information for genome scaffolding. Our results

Table 2: Scaffolding value of different paired-read library combinations				
No. of libraries*	Most efficient	Scaffold N50	Least efficient	Scaffold N50
1	15 kb	163,475	PE	37,694
2	5 kb + 25 kb	522,027	PE + 3 kb	46,699
2	5 kb + 20 kb	474,308	PE + 5 kb	141,403
2	8 kb + 25 kb	470,890	PE + 25 kb	142,007
3	5 kb + 20 kb + 25 kb	834,964	PE + 3 kb + 5 kb	158,525
3	5 kb + 15 kb + 25 kb	789,954	PE + 3 kb + 8 kb	171,253
3	8 kb + 20 kb + 25 kb	726,289	PE + 3 kb + 25 kb	198,696
7	ALL	1,287,609	N/A	N/A
*All libraries were normalized to 8.5x physical genome coverage, limited by the amount of available data for the paired-end (PE) library.				

show that the addition of MP sequencing can dramatically increase contingency of mammalian genome references. In all analyses, insert sizes of >8 kb were shown to be essential because of their ability to bridge the longer and more abundant LINE and LTR elements. The analysis where the fraction of long repeats that is spanned by each MP library is determined shows that large insert MPs are capable of spanning ~90% of the annotated long repeats. The remaining approximately 10% of elements that could not be bridged by any of the MP reads can likely be explained by a highly repetitive nucleotide context around the repeat elements themselves. When a repeat element is surrounded by other repeats (mostly at centromeric or telomeric regions) one or both reads of the pair that would span such region can not be mapped uniquely to the genome and can thus not be included in the analysis. In agreement with this, our data show that even a combination of all libraries in this study fails to span 4–5% of repetitive elements larger than 3 kb in size (Figure 2C and D). Because rat and other vertebrate genomes contain tens of thousands of repeat elements that exceed the routinely used paired-end insert sizes (up to 500 bp), but include the very common LINE elements, we conclude that the inclusion of mate-pair libraries with insert sizes of 8 kb and above are instrumental for comprehensive reconstruction of genome structures.

The largest insert libraries (20–25 kb) were instrumental for increasing the N50 of scaffolds to megabase levels. Because the draft rat genome is already of relatively high quality, the improvements presented here have only mild effects. However, we anticipate that large insert MP sequencing will be very useful for finalizing low-pass capillary sequenced or NGS-based genomes like those of most primates as well as many of the vertebrate genomes. Genomes with large fractions or large segments of repeats, like that of the zebrafish or certain plants, might benefit even more from large insert mate-pair data as their genomes have a

very high repeat content in combination with recently duplicated sequences. Furthermore, most ongoing genome sequencing projects employ next-generation sequencing techniques, and because de novo genome assembly based on short-reads is still in its infancy, contig sizes for vertebrate genomes are typically in the kilobase range [19,28,29]. Although paired-end data with insert sizes up to 500 bp are now commonly included in these processes, our results demonstrate that longer-range information as provided by the large insert MPs described here is essential for comprehensive genome assembly. It should be stressed that the structure of every genome of interest is unique and variable in complexity. Therefore, the optimal combination of MP insert sizes will vary as well. A quick examination of the repeat size and distribution could aid in determining which MP insert size combination is expected to be optimal, but experimental optimization or a broad range of libraries such as used here might be required.

In the analyses presented here, we focused on the application of large insert MPs for genome sequencing efforts, but the findings could be extrapolated to the detection of structural variation. Previous analyses of whole human genomes have shown that SVs affect more base pairs than single point mutations, yet the field has struggled to find a suitable approach for comprehensive detection of such events [30]. Hillmer et al. concluded that the most optimal insert size for SV detection is approximately 10 kb, although a thorough examination of the value of insert sizes above 10 kb was not described [17]. In unraveling the structure and organization of ultra-complex clustered mutation events, like the recently described chromothripsis, larger insert sizes (20–25 kb) may extend the detection limit and help to complete the overall picture [31-34]. It should be noted, however, that a “mate-pair only” approach also comes with disadvantages: small insertions, inversions, duplications, and deletions may be missed due to the broad size distribution and relatively low coverage at the base level.

Large insert MP sequencing represents a good alternative for the more traditional bacterial artificial chromosome-end sequencing because the sequencing libraries can be produced by relatively simple and scalable procedures without the need for laborious cloning and colony picking. Furthermore, the protocol can be fitted to all existing NGS platforms by changing the oligonucleotide adapters that are used. The mate-pair library construction protocol is relatively laborious compared to standard fragment library construction protocols, but with the latest improvements of the mate-pair protocol (SOLiD 5500 version), the procedure takes ~14 hours of hands-on work. More importantly, robustness of the protocol has been increased and the required input genomic DNA was reduced to only 1–5 μg for a standard≤ 3-kb library, compared to 5–20 μg for the SOLiD V4 protocol (Additional file 9). The removal of column-based cleanup steps and the increased circularization efficiency (via the implementation of intra-molecular hybridization instead of circularization to an internal adaptor) are the main

factors that allow for a reduced amount of input DNA. Nevertheless, our results show that limiting the amount of input DNA can strongly affect the complexity of the resulting library. For larger insert libraries it is therefore recommended to start with maximized amounts of DNA (>20 µg).

Although large insert MP libraries must be sufficiently complex, high physical genome-wide coverage is readily obtained at relatively low sequencing depth of tens of million read pairs. Alternative large insert approaches, like fosmid di-tag sequencing [20], have been documented to suffer from low library complexity, which may be overcome by using larger amounts of input material, but they have an additional disadvantage as they are restricted to a fixed insert size of approximately 40 kb [16,20,35,36]. Our data clearly demonstrate the added value of medium-sized insert libraries for genome structure analysis, a conclusion that was supported by Hampton et al. [20], who had to use supporting 4–6 kb mate-pair data to obtain essential long-range information that could not be obtained by fosmid di-tags alone. Using the MP protocol presented here, small, medium and large insert MP libraries can be generated in one go. Nevertheless, we did not generate libraries of equal size to 40-kb fosmid clones, so we could not determine if inserts of 25 kb are sufficient to fully replace 40 kb fosmid clones or if 40 kb pairs would span the last 4-5% of repeats that could not be covered by any of the MPs used here.

Materials and methods

Generation of MP libraries and mapping. To allow for the construction of large insert MP libraries, we modified the standard SOLiD 4 mate-pair library preparation protocol (Additional file 2). In short, genomic DNA was isolated from Brown Norway (BN/RijHsd) rat brain and testis tissue. DNA (100 µg) was sheared under mild conditions using HydroShear (JHSH204007, 20 cycles, SC15) and subsequently end-repaired (Epicentre
End-It™ DNA-end repair kit) in 1 mL End-It mix per 100 µg input DNA. CAP adapters were ligated in 500-µL reaction volumes of New England Biolabs Quick Ligase reaction mix. The amount of ligated adapter was determined based on the DNA content (100 pmol CAP adapter/pmol DNA). Following CAP-adapter ligation, DNA fragments were purified with phe nol:chloroform:isoamylalcohol (PCI, pH7.9) by gentle mixing and centrifugation in MaXtract high-density tubes (QIAgen, 1.5mL, #129046). The fragmented DNA was separated via pulsed-field gel electrophoresis (PFGE; Bio-Rad CHEF Mapper XA system). PFGE conditions and settings were: 1% low melt agarose gel (Invitrogen, #16520100), 0.5x TBE, 14°C, 19 hours, forward current: 9.0 V/cm, switch time 0.08 s–0.46 s, reverse current 6.0 V/cm, switch time 0.08 s–0.46 s. Multiple size ranges (<7 kb, 7–10 kb, 10–14 kb, 14–18 kb, 18–24 kb, 24–33 kb, and >33 kb) were selected from the gel using a 1 kb extension ladder (Invitrogen, #10511-012). For unknown reasons, actual library insert sizes after library construction and

mapping tend to be lower than their initial appearance on gel. A probable explanation for this is that most library construction steps may show a small bias towards smaller molecules (e.g. during circularization). Further on in the manuscript, each library will be referred to as the actual insert size as determined after data analysis. Following gel excision, DNA was carefully recovered using GELaseTM (Epicentre, #G09200). DNA fragments were circularized with a biotinylated internal adapter at a final DNA concentration of 1 nanogram per microliter (ng/µl). For every 40-µl reaction volume, 1 µl T4 DNA ligase was used. The reaction mixture was purified, and non-circularized fragments were removed by a plasmid-safe DNase treatment (Epicentre, #E3101K). Following linear DNA removal, DNA polymerase I-directed nick translation "pushed" the nick from the adapters into the circularized target DNA to generate sufficient tag length for sequencing (~100 bp for each tag, 13 minutes on ice water [0°C], inactivated with PCI). T7 exonuclease and S1 nuclease treatment were used to digest the circles at the position of the nick. The digested fragments of approximately 300 bp in size were end-repaired and bound to MyOne C1 streptavidin beads via the biotinylated internal adapter. Standard SOLiD P1 and P2 adapters were ligated to the blunt ends of the library molecules (a-tailing and alternative adapters should be used at this step to make the library compatible with the other sequencing platforms like Illumina, see Additional file 10), followed by another round of nick translation to remove the nick introduced by adapter ligation. Mate-pair libraries were amplified by PCR for 13–21 cycles, depending on the library. Amplification of the 14–18-kb size range samples (with an estimated final insert size of 10–12 kb) did not result in sufficient material (most likely because of unsuccessful adaptor ligation) and were not further included in the process. For all other insert size libraries we continued with templated bead preparation and libraries were successively sequenced on the SOLiD 4 system. For all libraries together, 192.4 million pairs of MP reads (AB/SOLiD V4, two slides) were sequenced. Paired reads were mapped against rat reference genome RGSC 3.4 using BWA v0.5.9 [37], and non-unique (based on identical read start sites for the forward and reverse read) and ambiguously mapped read pairs were removed from the data set.

Generation of the paired-end library (170-bp) and mapping. For construction of the paired-end library (PE; 170-bp insert), the SOLiD 3 protocol for fragment library preparation was used (SOLiD™ 3 System Library Preparation Guide; Section 2.1). DNA (3 µg) derived from the Brown Norway rat was used for shearing using Covaris S2 (10 cycles of 60 s, intensity 5, 100 cycles/burst, 4°C). Sheared DNA was end-repaired using the Epicentre End-ItTM DNA-end repair kit. P1 and P2 adapters were ligated to the DNA fragments, and the library molecules were selected based on size (220–300 bp, including 90 bp for both adapters). PCR amplification was done for 5 cycles with primers specific to the adapters to obtain sufficient library molecules for ePCR and sequencing. Sequencing was done in paired-end mode (forward and reverse tag; 50 bp and 35 bp, respectively) on the SOLiD 3 system (1 slide AB/SOLiD V3). Approximately 1.6 x 108 paired-end reads were sequenced (95% non-clonal) and

mapped with BWA, resulting in a data set with a median insert size of 170 bp.

Calculation of library insert size and contig scaffolding. Forward and reverse reads were mapped independently against contigs of rat RGSC3.4 genome assembly using BWA 0.5.9. Only read pairs with a single best hit for each tag (X0 flag equal to 1) were taken into consideration for estimate of insert size distribution. Analysis of library complexity was done by randomly sampling of reads from a library and determining the number of non-clonal pairs. Next, read pairs with exactly the same mapping coordinates of forward and reverse tags were marked as clonal and excluded from further analysis. Distribution of insert sizes was estimated from read pairs with proper orientation and distance between tags (below 100 kb). To allow comparison of different library insert sizes for contig scaffolding, we created a subset of data for each library that corresponds to 8.5x non-clonal physical coverage of rat genome. We randomly sampled read pairs from each library, computing physical coverage represented by non-clonal pairs with expected orientation and distance between tags on chromosomal level (skipping pairs corresponding to first and last percentiles of insert size distribution). Read pairs from the normalized datasets with forward and reverse tags mapped to different contigs were selected for scaffolding analysis. Scaffolding was performed using SSPACE v2.0 software [25] with default parameters. The order in which the libraries were used by SSPACE was as recommended by the SSPACE manual - always from the smallest to largest insert size.

Repeat analysis. Repeat annotation of the rat genome reference was obtained from Ensembl database [24] (v.66) and was used for calculation of size distribution and abundance of different repeat types. This annotation was used to determine the percentage of repeat elements spanned by mapped fragments from each mate-paired library. We used the normalized dataset where every library had 8.5x physical coverage (as described above). We considered only unambiguously mapped read pairs that had a proper orientation of tags and did not exceed 99th percentile of fragment size distribution. Since individual copies of a mobile element or repeat class differ in size, we used 500 bp windows for calculation of percentage of mobile elements overlapped by fragments from each library.

Comparison to optical maps. Original RGSC3.4 scaffolds and those obtained after NGS-assisted re-scaffolding were compared to each other by nucleotide BLAST search. Nearly identical (>99%) super-kb segments were plotted as Harr-plot visualization graphs. Most evident discordant regions were manually selected and the corresponding genomic segments were compared to optical maps [10] of Brown Norway rats. Optical maps are generated from large, randomly sheared high-molecular weight genomic DNA molecules that are stretched on a microscope slide. After stretching, the DNA molecule is digested with a SwaI restriction enzyme. While the DNA remains attached to the slide, the cuts become visible under the microscope as small gaps and the sizes of the stained DNA fragments can be measured.

Multiple optical maps are then combined to form a comprehensive reference optical map for the Brown Norway rat genome. The optical maps used in this study are produced in the lab of David C. Schwartz and are available upon request. To compare our assembly with existing Brown Norway rat optical maps, nucleotide sequences of MP-enhanced scaffolds were digested in silico with a SwaI restriction enzyme. The in silico digested fragments were plotted next to optical maps, originally aligned to the RGSC 3.4 assembly and visually inspected for concordance.

Authors' contributions

SvH, WK, NL, and EL, carried out the molecular experiments; FR, CL, SZ and VG performed bioinformatics analyses. SvH, WP SZ, SG, DS and VG performed and analyzed optical mapping experiments, SvH, WK, and EL generated large insert mate pair sequencing libraries, NL performed next-generation sequencing. SvH, WK, TH, VC and EC conceived of the study, and participated in its design and coordination and drafted the manuscript. All authors contributed to the final version of the manuscript and have read and approved it.

Competing interests

EL, CL and TTH are employees of Life Technologies, manufacturer of SOLiD sequencing systems used in this study.

Acknowledgments

This work was supported by funding from the European Union framework 7 integrated project EURATRANS (HEALTH-F4- 2010–241504 to E.C.); the NWO-CW TOP grant (700.58.303 to EC); and the National Institutes of Health (R01 HG000225 to D.C.S). Funding for open access charge: primary funding from the Hubrecht Institute.

Availability of data

The data sets supporting the results of this article are available in the ArrayExpress repository under accession number E-MTAB-1082.

Additional files

Additional files can be found online at http://www.biomedcentral.com/1471-2164/14/257 (doi:10.1186/1471-2164-14-257)

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al.: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al.: The sequence of

the human genome. Science 2001, 291:1304-1351.

3. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE et al.: Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 2004, 428:493-521.

4. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P et al.: Initial sequencing and comparative analysis of the mouse genome. Nature 2002, 420:520-562.

5. Finishing the euchromatic sequence of the human genome. Nature 2004, 431:931-945.

6. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M et al.: Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 2009, 7:e1000112.

7. Cordaux R, Batzer MA: The impact of retrotransposons on human genome evolution. Nat Rev Genet 2009, 10:691-703.

8. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D et al.: Origins of the E. coli Strain Causing an Outbreak of Hemolytic-Uremic Syndrome in Germany. N Engl J Med 2011, 365:709-717.

9. Dimalanta ET, Lim A, Runnheim R, Lamers C, Churas C, Forrest DK, de Pablo JJ, Graham MD, Coppersmith SN, Goldstein S et al.: A microfluidic system for large DNA molecule arrays. Anal Chem 2004, 76:5293-5301.

10. Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM et al.: High-resolution human genome structure by single-molecule analysis. Proc Natl Acad Sci U S A 2010, 107:10848-10853.

11. Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S et al.: A single molecule scaffold for the maize genome. PLoS Genet 2009, 5:e1000711.

12. Ajay SS, Parker SC, Ozel Abaan H, Fuentes Fajardo KV, Margulies EH: Accurate and comprehensive sequencing of personal genomes. Genome Res 2011, 21:1498-1505.

13. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L et al.: Paired-end mapping reveals extensive structural variation in the human genome. Science 2007, 318:420-426.

14. Treangen TJ, Salzberg SL: Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 2012, 13:36-46.

15. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L et al.: CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods 2011, 8:652-654.

16. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S et al.: High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 2011, 108:1513-1518.

17. Hillmer AM, Yao F, Inaki K, Lee WH, Ariyaratne PN, Teo AS, Woo XY, Zhang Z, Zhao H, Ukil L et al.: Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome Res 2011, 21:665-675.

18. Peng Z, Zhao Z, Nath N, Froula JL, Clum A, Zhang T, Cheng JF, Copeland AC, Pennacchio LA, Chen F: Generation of long insert pairs using a Cre-LoxP Inverse PCR approach. PLoS One 2012, 7:e29437.

19. Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J et al.: Complete Khoisan and Bantu genomes from southern Africa. Nature 2010, 463:943-947.

20. Hampton OA, Miller CA, Koriabine M, Li J, Den Hollander P, Carbone L, Nefedov M, Ten Hallers BF, Lee AV, De Jong PJ et al.: Long-range massively parallel mate pair sequencing detects distinct mutations and similar patterns of structural mutability in two breast cancer cell lines. Cancer Genet 2011, 204:447-457.

21. Williams LJ, Tabbaa DG, Li N, Berlin AM, Shea TP, Maccallum I, Lawrence MS, Drier Y, Getz G, Young SK et al.: Paired-end sequencing of Fosmid libraries by Illumina. Genome Res 2012, 22: 2241-2249.

22. Herschleb J, Ananiev G, Schwartz DC: Pulsed-field gel electrophoresis. Nat Protoc 2007, 2:677-684.

23. Schwartz DC, Cantor CR: Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell 1984, 37:67-75.

24. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S et al.: Ensembl 2012. Nucleic Acids Res 2012, 40(Database issue):D84-90.

25. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 2011, 27:578-579.

26. Sarkar D, Goldstein S, Schwartz DC, Newton MA: Statistical significance of optical map alignments. J Comput Biol 2012, 19:478-492.

27. Valouev A, Schwartz DC, Zhou S, Waterman MS: An algorithm for assembly of ordered restriction maps from single DNA molecules. Proc Natl Acad Sci U S A 2006, 103:15770-15775.

28. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al.: De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 2010, 20:265-272.

29. Ye L, Hillier LW, Minx P, Thane N, Locke DP, Martin JC, Chen L, Mitreva M, Miller JR, Haub KV et al.: A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol 2011, 12:R31.

30. Alkan C, Coe BP, Eichler EE: Genome structural variation discovery and genotyping. Nat Rev Genet 2011, 12:363-376.

31. Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M et al.: Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. Hum Mol Genet 2011, 20:1916-1924.

32. Kloosterman WP, Hoogstraat M, Paling O, Tavakoli-Yaraki M, Renkens I, Vermaat JS, van Roosmalen MJ, van Lieshout S, Nijman IJ, Roessingh W et al.: Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biol 2011, 12:R103.

33. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA et al.: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 2011, 144:27-40.

34. Molenaar JJ, Koster J, Zwijnenburg DA, van Sluis P, Valentijn LJ, van der Ploeg I, Hamdi M, van Nes J, Westerman BA, van Arkel J et al.: Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature 2012, 483:589-593.

35. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F et al.: Mapping and sequencing of structural variation from eight human genomes. Nature 2008, 453:56-64.

36. Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G et al.: Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Methods 2010, 7:365-371.

37. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754-1760.

4

Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis

Teck Yew Low[1,2*], Sebastiaan van Heesch[3*], Henk van den Toorn[1,2*], Piero Giansanti[1,2], Alba Cristobal[1,2], Pim Toonen[3], Sebastian Schafer[4], Norbert Hübner[4,5], Bas van Breukelen[1,2], Shabaz Mohammed[1,2,6] Edwin Cuppen[3], Albert J. R. Heck[1,2], Victor Guryev[3,7]

* These authors contributed equally to this work

[1] Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
[2] Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, The Netherlands
[3] Hubrecht Institute-KNAW & University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands
[4] Max-Delbruck-Center for Molecular Medicine (MDC), Robert-Rossle-Str. 10, 13125 Berlin, Germany
[5] DZHK (German Centre for Cardiovascular Research), partner site Berlin, Berlin, Germany
[6] Current address: Departments of Chemistry and Biochemistry, University of Oxford, Physical & Theoretical Chemistry Laboratory, South Parks Road, OX1 3QZ Oxford, UK
[7] Current address: European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Antonius Deusinglaan 1, Building 3226, 9713 AV Groningen, The Netherlands

Adapted from Cell Reports 2013 Dec 12;5(5):1469-78.

Abstract

Quantitative and qualitative protein characteristics are regulated at genomic, transcriptomic, and posttranscriptional levels. Here, we integrated in-depth transcriptome and proteome analyses of liver tissues from two rat strains to unravel the interactions within and between these layers. We obtained peptide evidence for 26,463 rat liver proteins. We validated 1,195 gene predictions, 83 splice events, 126 proteins with nonsynonymous variants, and 20 isoforms with nonsynonymous RNA editing. Quantitative RNA sequencing and proteomics data correlate highly between strains but poorly among each other, indicating extensive nongenetic regulation. Our multilevel analysis identified a genomic variant in the promoter of the most differentially expressed gene ***Cyp17a1***, a previously reported top hit in genome-wide association studies for human hypertension, as a potential contributor to the hypertension phenotype in SHR rats. These results demonstrate the power of and need for integrative analysis for understanding genetic control of molecular dynamics and phenotypic diversity in a system-wide manner.

Introduction

Mass spectrometry (MS)-based proteomics and next-generation sequencing (NGS) are rapidly maturing techniques, each enabling comprehensive measurements of gene products at a system level [1-3]. Although MS and NGS are highly complementary, they are still rarely applied integrated in large-scale studies [4]. State-of-the-art MS approaches can currently identify over 10,000 proteins in a single experiment [5, 6], which brings the analysis of complete proteomes within reach [2, 7]. However, as long as noncustomary protein databases that are derived from (typically incomplete) reference genome assemblies and annotations remain the sole source used for MS spectra matching, true completeness will not be reached. For example, protein isoforms arising from genetic polymorphisms, posttranscriptional events such as RNA-editing and posttranslational modifications are largely missed [8, 9].

Recent advances in NGS techniques, including whole genome sequencing (WGS) and total RNA sequencing (RNA-seq) allow for the generation of near-complete inventories of genetic variation in a system and its transcribed repertoire [10]. However, from such analyses, the effects on the proteins cannot be predicted with high confidence. For example, the consequence of a single nucleotide variant (SNV) on the coding capacity of a transcript can be predicted accurately, but not the potential effect on the stability of the corresponding protein. Systematic comparison of RNA-seq data with genomic data reveals another layer of complexity. It has now been convincingly demonstrated that certain transcripts are modified by posttranscriptional editing, primarily by targeted A to I deamination [11-14]. Most likely, all these types of variation will not only affect the composition and function of a protein, but also influence expression levels. However, additional layers of translation control may

dampen or completely abolish such effects.

An integrative analysis of different data modalities, ideally derived from samples of a single source, is required for correctly deciphering the effects of genomic and transcriptomic variation on molecular processes and cellular functioning. An example of such data integration is the use of proteomic data derived from MS in combination with complete genome data to improve gene annotation [15, 16]. This approach has so far been sparsely performed and mainly in organisms with smaller genomes [17, 18]. On the other hand, integrative investigations of messenger RNA levels and the proteins they encode reveal only modest correlations, implying an unresolved level of complexity in regulation of expression [4, 19-22].

For this study, we selected two rat inbred strains BN-*Lx*/Cub (BN-*Lx*) and SHR/OlaIpcv (SHR) [23], representing widely studied, renewable, and genetically homogeneous resources. Both strains have previously been extensively characterized at the genomic [24, 25] and phenotypic level [26-28]. The BN-*Lx* strain is derived from, and thus very closely related to, the Brown Norway (BN) strain. The latter strain was used for creating the rat reference genome assembly [29] and is commonly used as the protein reference data set in rat proteomics studies. The spontaneously hypertensive rat (SHR) is more diverged from BN and is a widely used disease model for hypertension studies. Whereas several blood pressure quantitative trait loci (QTLs) have been mapped to the SHR genome, no functional variants driving elevated blood pressure have been validated to date. Here, we combine in-depth genomic, transcriptomic, and proteomic analyses from inbred rats of two different genetic backgrounds using the same sets of rat liver tissues (Figure 1A). The liver is a large and relatively homogeneous tissue source that is well known to be involved in both hypertension and metabolic syndrome - the phenotypes associated with the SHR strain. We determine quantitative and qualitative molecular dynamics at different functional levels and achieve a level of proteome completeness by adding variation information derived from WGS and RNA-seq data. These data allow us to apply a genome-wide genetic-genomics approach [30] to start understanding multilevel systems regulation and to identify candidate genes that are potentially involved in the hypertension phenotype of the SHR rat.

Results

Extension of the rat protein database

In proteomics, tandem mass spectra are typically annotated by searching against in silico-generated spectra based on a publicly available protein database. For rat, such a database is derived from the reference genome assembly of the BN rat [29]. To create a sample-specific database for MS peptide searching, we extended the existing RefSeq-based peptide database by incorporating strain-specific peptides and predicted peptides. We first obtained all strain-

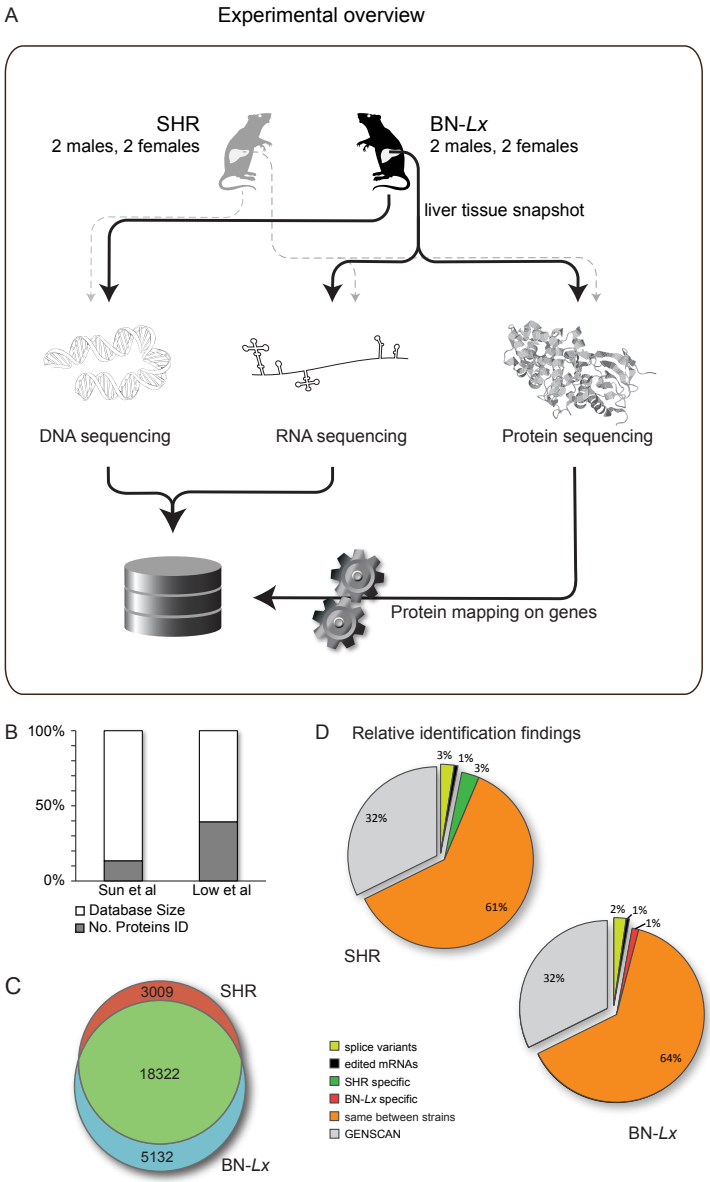


Figure 1 - **Integrated Proteomics, Genomics, and Transcriptomics to Improve Sample-Specific Protein Identification**. (**A**) Schematic representation of the integrated genome and proteome analysis of BN-*Lx* and SHR rat liver using NGS and deep-proteome profiling. (**B**) Bar plot showing the percentage of the current reference database that is covered by the experimentally derived proteomes, with respect to recent other proteomics efforts [31]. For BN-*Lx* and SHR, 39.7% of the Ensembl database is represented (13,088 out of 32,971 entries; release 3.4.63). The human liver proteome generated by the Chinese Human Liver Proteome Profiling Consortium cover 13.5% of the IPI human database (version 3.07; 7,050 out of 50,225 entries). (**C**) Diagram displaying identified proteins specific to BN-*Lx* (blue), SHR-specific proteins (red), and proteins shared between both strains (green). (**D**) Relative contribution (%) to the BN-*Lx* and SHR rat proteomes (containing unique peptides) of each additional layer of genomics- and transcriptomics-derived protein variants.

specific genetic variation of the BN-*Lx* and SHR strains including single nucleotide variants (SNVs) and in-frame indels. Most genomic SNVs are located in the noncoding sequences, and only the less frequent nonsynonymous variants in coding regions give rise to altered amino acid sequences [32-34]. We collected 10,493 nonsynonymous variants from recently generated high-coverage WGS data of the BN-*Lx* and SHR genomes [24, 25], which are predicted to affect 6,187 protein isoforms derived from 4,566 genes. Furthermore, to be able to detect in silico gene predictions using the proteomics data as evidence [35], we added 44,993 GENSCAN gene predictions to our rat database [36].

Next, we performed RNA-seq (Table S1) on RNA extracted from liver tissue of both rat strains (two males and two females per strain). To this end, paired-end sequencing data were generated to construct de novo transcriptome assemblies for each strain, In total, we found expression evidence for 18,116 known genes (12,052 with fragments per kilobase of exon per million fragments mapped [FPKM] >1), of which 2,612 (1,820 with FPKM >1) overlap the nonsynonymous variants previously detected by genome resequencing. Also, we identified 2,545 transcript splicing events affecting 1,015 genes. Although the majority of the identified splice events (1,687) were detected in both strains, 220 and 638 events were specific to BN-*Lx* and SHR rats, respectively (Table S2). Independent RT-PCR-based Sanger sequencing confirmed 74.1% (43 out of the 58 successful PCR assays) of a randomly sampled subset as true transcript isoforms (Table S3A). In addition, the same transcriptome assembly data provided expression evidence for 2,903 GENSCAN predictions (Table S4). The de novo assembled transcriptome data also allow for characterization of transcriptomes at nucleotide resolution. Because both BN-*Lx* and SHR strains are fully inbred, observed changes at the transcript level are unlikely to be allele-specific variation and can thus be attributed to technical artifacts (introduced during sequencing or mapping) or to RNA editing [37]. We find a total of 799 canonical (A to I or C to T) RNA-editing variants (Table S5) of which 176 and 354 are specifically observed in BN-*Lx* and SHR, respectively. As expected, a large proportion of edits resides in the noncoding UTR parts of transcripts or do not change the coding capacity of a transcript. Yet, they might be affecting RNA secondary structure, stability, or miRNA binding. Only 196 edits were nonsynonymous and therefore included in our protein database as potentially detectable by MS. Of a subset of 169 candidate editing events tested by independent RT-PCR-based amplicon resequencing, most (104) showed reads corresponding to expected edited transcripts, and another 12 likely represent germline variants that missed detection during genome resequencing (Table S3B).
All peptide variants and isoforms derived from genome and transcriptome variation and all newly predicted peptides based on GENSCAN and de novo transcriptome assembly data were appended to the Ensembl rat database (3.4.63) to create our customized RAT_COMBINED database, which was used for all subsequent proteomic analyses.

Proteomics analysis

We generated proteomics data with the same liver tissues used for RNA-seq. Each lysate was proteolyzed with five orthogonal proteases, and the resulting 36 SCX fractions per digest were analyzed with LC-MS/MS, cumulating in 180 runs per strain, yielding ~12 million tandem MS spectra. By using multiple proteases, not only the identification and sequence coverage of each protein increase, but also the chance of capturing evidence for predicted peptides/proteins and consequences of RNA editing [38-40]. To ensure comprehensive coverage, two different but complementary algorithms for spectra-to-peptide assignment were applied. First, Mascot search engine was used for database searching. Next, the remaining unassigned spectra were processed with PEAKS Studio 6.0, which incorporates a proprietary de novo sequencing algorithm. The large amount of data allowed us to apply a false discovery rate (FDR) filter of 0% (q = 0) and still identify over 2 million peptide-spectral matches (PSMs), corresponding to ~175,000 nonredundant peptides (Tables S6A and S6B). By performing a merged BN-Lx and SHR data set search against our custom RAT_COMBINED database, we obtained peptide evidence for 26,463 database entries. Of these, 18,322 are shared between BN-***Lx*** and SHR (Figure 1C; Table S7) whereas 3,009 and 5,132 appear strain specific for SHR and BN-***Lx***, respectively. For comparison, we counted the number of identifications matching entries in the Ensembl database (3.4.63), disregarding the variants. Out of the 32,971 original database entries, 13,088 were matched, representing 39.7% of database entries. In contrast, the most extensive liver proteome so far (the human liver proteome generated by the Chinese Human Liver Proteome Profiling Consortium) covers only 13.5% of the human IPI database (version 3.07; 50,225 entries), illustrating the depth of our data (Figure 1B; Table S6D). Over 86.5% of all proteins could be supported by evidence of gene expression in the RNA-seq data. As expected, identified peptides are evenly distributed over the rat chromosomes, concordant with the distribution of genes and transcripts (Figure S1A). The median coverage of all proteins is 15.6% with roughly equal contributions from each protease data set (Figure S1B).

Identification of predicted proteins and protein isoforms

Approximately 5,700 unique peptides (Table S8A) provide experimental evidence for 1,195 in silico predicted GENSCAN proteins (Tables S7 and S8B). For 1,187 (99%) of those, RNA-seq data support the observed expression. Fifty of them show best reciprocal hits with known mouse proteins, and another 32 with known human proteins (Table S8C). Furthermore, we detect N-terminally acetylated peptides for 69 of these 1,187 proteins, with A, M, S, and T as their N-terminal residues (Table S8D) [41, 42]. N-terminal peptides validate these putative genes by confirming their translational start sites. A different class of proteins with largely uncertain existence is the short expressed proteins (SEPs) encoded by short open reading frames [43]. Of all peptides in our data set, 0.25% could be assigned to 86 known SEPs and 37

identified SEPs (Figure S2; Tables S7and S8B).
The proteomics data also provide support for 83 transcript splicing events (0% FDR) that were previously not annotated (Figure 1D; Tables S6C, S7, S10A, and S10D). From all predicted proteins and splice isoform identifications, 309 and 15 respectively were unique for BN-***Lx***, and 193 and 13 were specific to SHR (Table S6C).

Detection of nonsynonymous protein variants

Next, we explored to what extent the addition of strain-specific variants affected protein detectability and stability. Of the uniquely assigned spectra, 3.5% did discriminate between allele-specific protein isoforms. We detected 126 nonsynonymous variants in our proteomic data, 38 for BN-***Lx***, and 88 for SHR (Table S10A and S10B). By applying a 0% FDR cutoff, we reassuringly did not find any BN-***Lx*** variants in the SHR samples, and vice versa (Table S6C). The fact that only a portion of nonsynonymous variants was confirmed by peptide-based evidence can be explained by our experimental design in which only genes expressed in the liver could be detected. Clearly, the inclusion of allele-specific variants has a measurable impact on protein discovery and results in more balanced peptide count per strain. The latter is most notable for the SHR rat because its genome is more diverged from the reference strain (BN). We used SIFT and Polyphen2 to predict if nonsynonymous SNVs could affect protein stability (Tables S9 and S11). Potentially damaging mutations were clearly overrepresented in differentially expressed proteins with nondifferential transcript levels ($p < 0.002$) (Table S11). This illustrates that nonconservative and structural missense variants may have limited influence on the abundance of a transcript yet can show a pronounced effect on protein stability.

Peptide-based evidence for RNA editing

To identify functional RNA-editing events, we mapped our peptide spectra to the set of potential RNA-editing events. In total, 20 out of the 196 nonsynonymous editing events could be confirmed by unique peptide-based evidence (Tables S6C, S9, S10A, and S10C). Because unique peptide evidence needs to overlap with the predicted editing site, many of the remaining 176 edits are likely missed because of incomplete coverage or redundancy in peptide data. Whereas limitations in the MS technology obviously result in an underrepresentation of identified RNA edits, MS still provides the best means to confirm the presence of such posttranscriptional modifications in the expressed proteins. On the other hand, we cannot rule out a possibility that the relatively low percentage of confirmed events is a true representation of the actual level of posttranscriptional modifications that make it to mature proteins. This may be due to negative selection against modified mRNA molecules. The high level of RNA sequencing coverage and the strict calling settings used to define editing events make it unlikely that an overestimation of editing events is introduced

during the RNA sequencing procedure and analysis.

It is worth noting that our comparison of de novo assembled and the annotated transcriptomes may not only reveal genetic differences, transcript isoforms, and common edited sites. Sequence and annotation imperfections within the current assembly and gene build can also be detected because the proteogenomics approach used in this study accounts for differences between observed and annotated transcriptome that originate from both biological and technical sources. Also, we emphasize that the de novo transcriptome assembly approach should be supplemented by regular transcriptome profiling if one aims to discover transcript variants that correspond to low-abundance transcripts and low-frequency events. To this end, we performed direct alignment of RNA-seq data to the rat transcriptome (known proteins and GENSCAN predictions) and predicted additional modifications of annotated transcripts (Table S5).

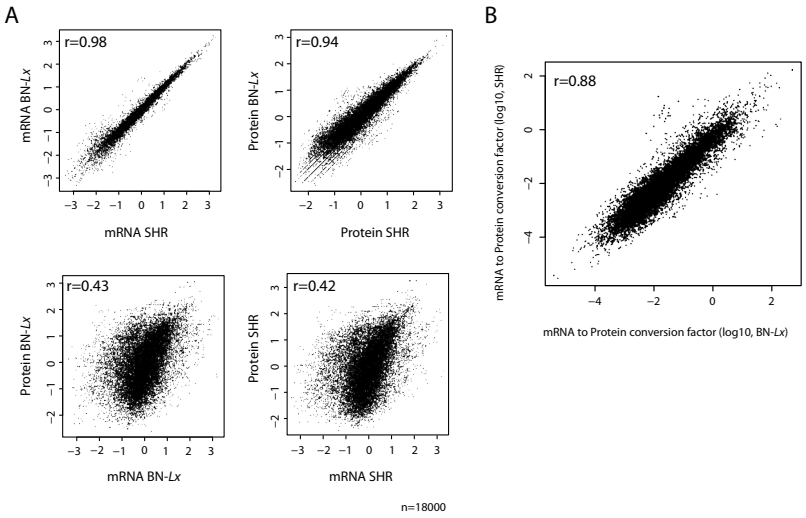


Figure 2 - **Global Correlation Plots Displaying the Complexity of mRNA and Protein Abundance**. (**A**) The top two panels display the high correlations between BN-Lx and SHR mRNA (left, r = 0.98) and protein levels (right, r = 0.94), estimated using log10 normalized spectral counts (Log10SAF) and normalized RNA seq counts (Log10FPKM). The bottom two panels show the poor correlations between mRNA and protein abundance for BN-Lx (r = 0.43) and SHR (r = 0.42), respectively. (**B**) Scatterplot depicting the correlation between experimentally determined gene-specific mRNA to protein abundance conversion factors as calculated for both BN-Lx and SHR (r = 0.88).

Relation between transcriptome and proteome levels

Next, we studied quantitative aspects by investigating the abundance of mRNA and protein levels. Although being derived from two different strains of rats, we observed a very high correlation of liver mRNA between BN-***Lx*** and SHR (r = 0.98). Similarly, the correlation coefficient for protein expression between BN-***Lx*** and SHR is also remarkably high (r = 0.94)

(Figure 2), providing confidence in our quantification strategy based on spectral counts.

Next, we sought to define a correlation between mRNA and protein expression levels in our data. Making a direct correlation between mRNA and protein levels is hampered by the fact that in peptide-based proteomics many proteins contain similar peptide sequences. It is therefore hard to assign any of the shared peptides unambiguously to a protein, the so-called protein-inference problem [44, 45]. Consequently, it is hard to integrate the quantitative measurements, which are necessarily restricted to peptides, to a protein measurement. Still, numerous studies conclude that the global correlation between mRNA and protein is certainly not linear and often an r of 0.4–0.5 is reported [4, 19, 22]. Such findings are corroborated by results that show that indeed only part of the variation in the protein levels can be explained by mRNA levels [21]. Here, we use a spectra-count method for quantification of protein levels. We use data derived from five different proteolytic enzymes, which is sufficient to exclude a proteolytic digest-specific bias [39]. Although we did identify unique peptides per protein (Table S12), we chose to take the total number of PSMs for every peptide matching a protein as a measurement of its abundance to increase the quantitative resolution per protein. Subsequently, we determined the proteome-transcriptome correlation for BN-*Lx* and SHR to be r = 0.43 and 0.42, respectively (Figure 2A). This correlation is thus weak, albeit in line with the previous studies in other systems. Based on these quantitative comparisons, we also found that 3' UTR expression levels correlate increasingly better with protein levels (r = 0.54) than do 5' UTR levels (r = 0.43) or reads derived from the coding sequence (r = 0.47) (Figures S3C–S3E). We speculate that the abundance of 3' UTR reads depends on transcript integrity and reflects both transcript count and stability. Transcript levels could also be reproducibly converted to predicted protein levels using a gene-specific conversion factor, which showed high correlation between the two strains (r = 0.88) (Figure 2B; Table S15). The high correlation between strains for this gene-specific factor illustrates the conservation of quantitative mRNA levels in relation to protein levels, independent of intermediate (less understood) levels of expression regulation. Although the conversion factor cannot be analyzed in-depth within the scope of this article, we postulate that translation efficiency, RNA, and protein degradation (and thus stability) are likely to play an important role. The top 100 proteins with the lowest and highest conversion factors were subjected to gene ontology (GO) overrepresentation analysis. We observed a trend in cellular localization toward cytoskeleton (highest 100) or the membrane (lowest 100), although the observations were not significant (Figure S3B; Table S15). We can only speculate that the conversion factor appears to be protein specific and conserved between strains. This factor combines the aforementioned levels of gene expression regulation in one value. One particular group of proteins appears to behave differently, representing the family of α_{2u}-globulins (known as rat major urinary proteins; Figures S3A and S3B). Unfortunately, none of these proteins could be identified by unique peptides due to high protein sequence homology within this class of

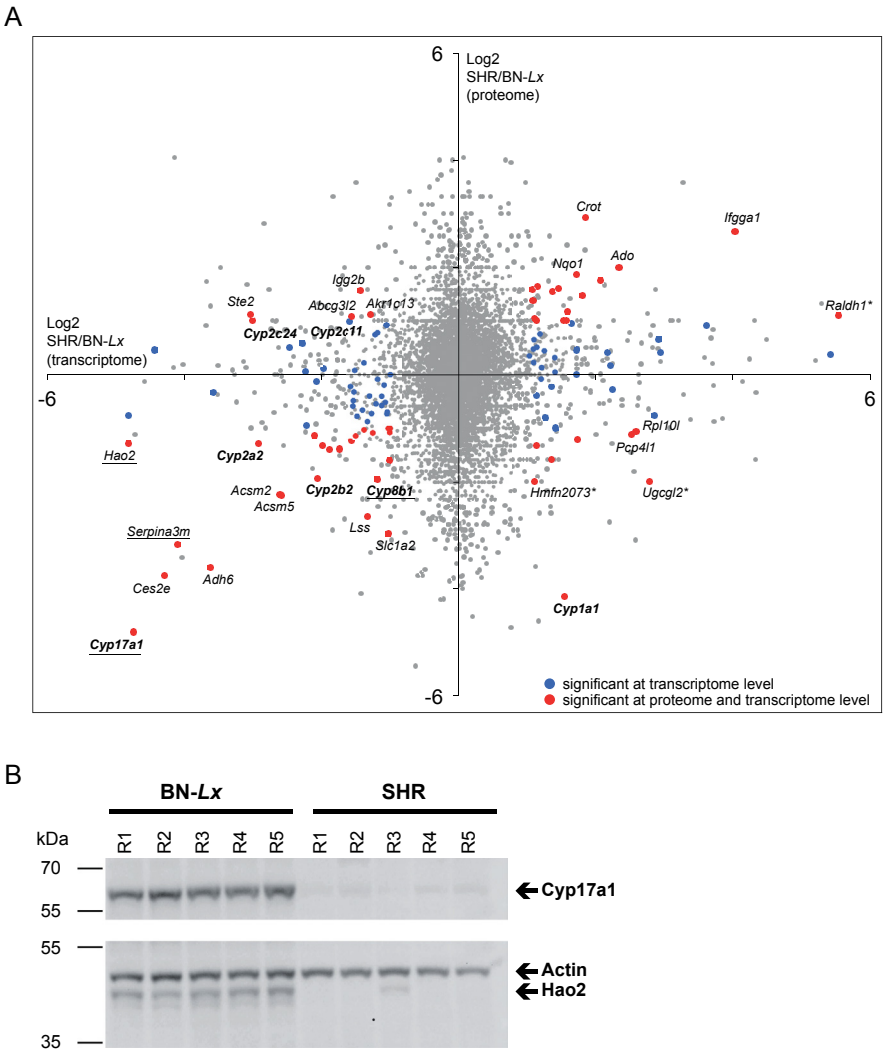


Figure 3 - **Gene-Centric Strain-to-Strain Comparison of Significantly Differentially Expressed Genes**. (**A**) Genes in BN-Lx and SHR with significantly deviating mRNA levels (blue dots; n = 59) or mRNA and protein levels (red dots; n = 54) are highlighted. Gene names marked by an asterisk are based on GENSCAN blast predictions derived from the closest predicted homology to human and mouse genes. Genes belonging to the CYP450 superfamily of catalytic enzymes are in bold and genes associated with hypertension in human or rat literature (*Hao2, Serpina3m, Cyp8b1, and Cyp17a1*) are underscored. (**B**) Western blot performed with liver tissues from five animals each for BN-*Lx* and SHR. Both Cyp17a1 and Hao2 are downregulated in all the biological replicates in the SHR strain compared to BN-*Lx*, consistent with the proteomics data. Actin was used as a loading control.

genes. Therefore, although this class stands out in the quantitative comparisons, absolute differences between BN-***Lx*** and SHR cannot be determined with high confidence at this point.

Genetic control of quantitative proteome characteristics

To determine the effects of genetic variation on quantitative transcriptome and proteome characteristics, we compared the difference of mRNA and protein expression between the two rat strains. First, we filtered our quantitative data with more stringent criteria retaining only genes quantifiable at both the protein and the RNA level (reliable expression level estimates by Cuffdiff and nonzero SAF counts). This allowed us to compare 6,743 genes (Figure 3A; Table S13), 113 of which were differentially expressed at the RNA level (at least 2-fold change in expression; and $q < 0.01$) and 205 at the protein level (at least 2-fold change and $q = 0$). The majority of the differentially expressed transcripts (59/113) do not show comparable changes at proteomics level. This group of proteins likely acquires stable expression through regulation of at the level of translation or through proteostasis. A small proportion of the genes (13/113) shows discordant behavior with opposite expression patterns for transcripts and proteins. Both groups do not show any overrepresentation in GO terms or pathways. The limited number of genes with significantly altered expression indicates the high global genome and proteome similarity between the two inbred rat strains. However, it also illustrates that interindividual differences may be in the details, such as represented by changes in posttranslational protein modifications and protein networks [1, 46]. Finally, 41 out of the 113 differential genes show strain-specific expression changes that are consistent between transcriptome and proteome (Figure 3A; Table S13). The products of these 41 genes relate to catalytic activity (28 genes, GO-term enrichment p value 1.4e-5) and metabolic pathways (13 genes, $p = 2.6e\text{-}4$).

A germline promoter variant deregulates Cyp17a1 expression in Spontaneously Hypertensive Rats

This set of 41 genes likely underlies some of the phenotypic differences known to exist between BN-***Lx*** and SHR rats, like spontaneous hypertension [48] and metabolic syndrome [49, 50]. We therefore investigated which genes were previously reported to be associated with hypertension in human or rat. First, three out of the 41 genes that are differential at both the mRNA and protein level were found to be associated with hypertension in the rat. Those three genes, ***Hao2*** [51], ***Serpina3m*** and ***Cyp8b1*** [52], came out as top hits when studying SHR (-related) strains or a panel of congenic rat strains to define candidates for hypertension. All three genes also overlap known blood pressure QTLs in the rat [53], and two of them (***Serpina3m*** and ***Hao2***) show a very strong connection to the SHR genotype based on eQTL data derived from the BXH/HXB recombinant inbred panel (founded by the BN-***Lx*** and SHR strains) (Figure S4A–S4C). This implies that the gene expression regulation of Serpina3m and

Hao2 is regulated in ***cis*** and thus strongly related to the genotype of the SHR strain. A fourth gene, ***Cyp17a1***, was identified as a top hit in relation to blood pressure and hypertension in human genome-wide association studies on European, Japanese, and Chinese individuals [54-57] (Table S14A). ***Cyp17a1*** also overlaps a blood pressure QTL in the rat and shows the most extreme downregulation in SHR compared to BN-***Lx*** in our analysis (Figure 3A). The differential expression of Hao2 and Cyp17a1 was verified independently by western blot, using liver samples of five animals from each strain (Figure 3B). Like Cyp8b1, Cyp17a1 is a member of the cytochrome P450 (CYP450) superfamily [58] of catalytic enzymes that mediate monooxygenase reactions and regulate drug metabolism. Interestingly, mutations in human ***CYP17A1*** are known to lead to congenital adrenal hyperplasia due to 17 alpha-hydroxylase deficiency, which results in hypogonadism, pseudohermaphroditism, and severe hypertension [59-62]. To determine the genetic basis of the ***Cyp17a1*** expression differences

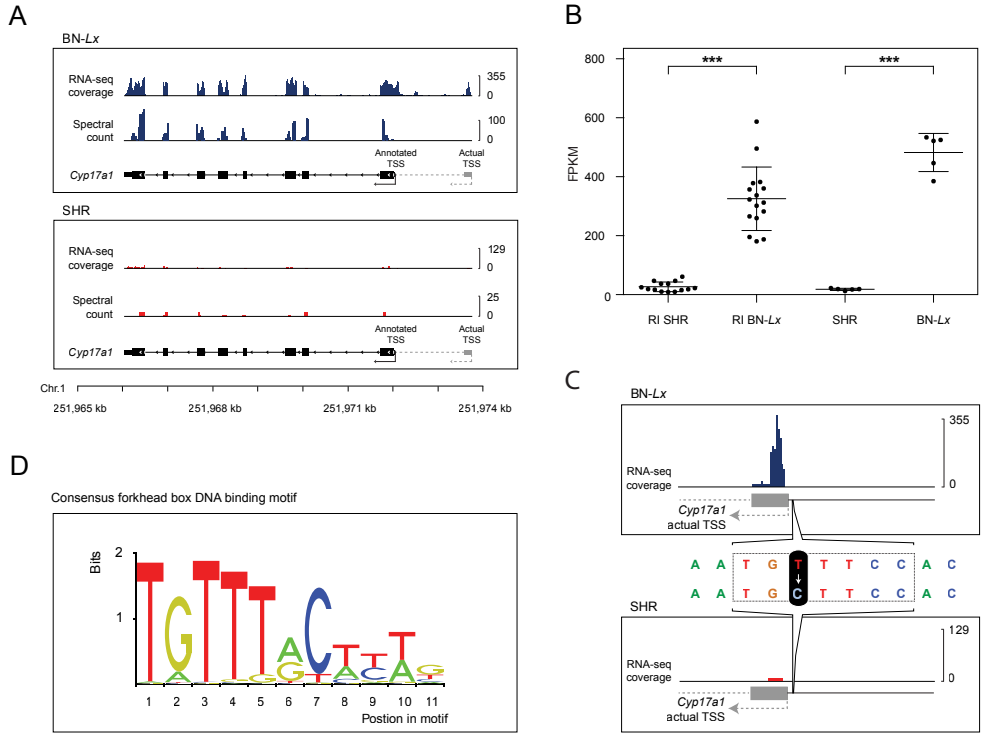


Figure 4 - **A Germline Promoter Variant Deregulates Cyp17a1 Expression in Spontaneously Hypertensive Rats**. (**A**) Experimental evidence covering this part of the genome from RNA sequencing and the proteomics data (spectral counts) are plotted along the gene body of Cyp17a1 for BN-*Lx* (blue) and SHR (red). The transcript is positioned on the reverse strand. Both the annotated transcription start site (TSS, black arrow) and the actual TSS (gray arrow) are shown. (**B**) Expression QTL analysis of *Cyp17a1* expression in the HXB/BXH recombinant inbred panel. Gene expression is plotted based on RNA-seq for the ancestral strains (n = 5) and the RI strains split by ancestral haplotype at the *Cyp17a1* locus (n = 16 for BN-*Lx* and n = 14 for SHR). (**C**) Zoomed-in view of the actual TSS, with the position of the germline T/C SNV shown. The dashed box (gray) shows the core part of the forkhead box DNA binding motif. (**D**) Consensus forkhead box DNA binding motif, obtained from the JASPAR database FOXA1 motif [47].

between BN-*Lx* and SHR, we sought for genetic variants in the annotated exons and flanking regulatory sequences, but none were present. Exploration of eQTL data, however, revealed a very strong *cis*-effect (Figure 4B;Table S14B), indicating that the measured expression difference is due to genetic variants in the gene itself or in neighboring regulatory elements. Upon closer inspection of the RNA-seq data, we found that the transcriptional start site (TSS) of the *Cyp17a1* gene was incorrectly annotated and resides approximately 2 kb upstream of the currently annotated most 5′ exon (Figure 4A). The true location of the promoter could be confirmed by H3K4me3 ChIP data that show specific enrichment of this active promoter mark surrounding the nucleosome-free region of the unannotated TSS (Figure S4D). Interestingly, this promoter does harbor a germline variant in SHR that disrupts the core part of an evolutionary conserved forkhead-box DNA binding domain (Figures 4C and 4D) [47], specifically deregulating transcription in SHR (Figure 4A). Because this expression trait is regulated in *cis* and this SNV is the only germline variant in the vicinity of the gene, our integrated genomics, transcriptomics, and proteomics approach has most likely identified the source of expression variation. The overlap with the RGD blood pressure QTL (http://rgd.mcw.edu/), top GWAS loci in humans, and known link to hypertension as a result of renal hyperplasia in patients carrying *CYP17A1* mutations are good indications that this promoter mutation in the SHR *Cyp17a1* gene contributes to the observed hypertensive phenotype of SHR rats.

Conclusions

Technological advances in both the proteomics and the sequencing community now provide the ability to discriminate genetic and posttranscriptional polymorphisms at the proteome level. These advances also allow improved quantitation of gene expression, which is generally restricted by the imprecise proxy of transcriptome data alone. We here show that the synergistic use of genomic, transcriptomic, and proteomic technologies significantly improves the information load that can be gained from proteomics as well as genomics efforts. By matching deep MS-based proteomics to a personalized database built from a sample-specific genome and transcriptome, we identify thousands of peptides that would otherwise escape identification. We believe that future efforts on both platforms benefit largely from our proof-of-concept approach, which brings integrated proteogenomics to a higher level. To highlight the strength of this approach, we present a link of a genomic variant in the *Cyp17a1* gene promoter and associate it with the hypertension phenotype of the extensively studied SHR rats.

Materials and methods

Custom rat protein database construction. We modified and appended the Ensembl [63-65] rat protein FASTA (build 3.4.63), which was derived from the reference (BN) genome

assembly, with DNA resequencing and RNA-seq data of the BN-*Lx* and SHR strains. Single nucleotide variants and indels were obtained from previous genome sequencing efforts [24, 25]. RNA was isolated from liver tissues of 6-week-old inbred BN-*Lx*/Cub and SHR/OlaIpcv rats. SOLiD RNA-seq libraries were prepared with ribosomal RNA depleted RNA and sequenced on the SOLiD V4 system. Next, we used CLC assembly cell version 4 (CLC Bio) to de novo assemble each rat liver transcriptome. Merged BN-*Lx* and SHR transcriptomes were mapped against the reference genome assembly using BLAT software [66]. Splicing and RNA-editing events were detected using alignments between the assembled transcriptome and genome and compared to their corresponding proteins. For all nonsynonymous genomic and transcriptome variants, individual entries were added to the extended protein search database. Also, we included 44,993 GENSCAN gene predictions of which 17,100 (FPKM >0.1) or 4,998 (FPKM >1.0) show evidence of expression.

Quantification of transcriptome data and identification of eQTLs. To quantify expression differences, RNA-seq data for each sample were aligned to reference genome using TopHat2. Expressed and differentially expressed genes were defined by Cuffdiff using all four transcriptomes per strain [67]. Determination of eQTLs in the HXB/BXH recombinant inbred panel consisting of 30 rat strains was performed exactly as previously described [68].

Strong cation exchange chromatography. After sonication and centrifugation, liver tissue lysates (300 µg each) were proteolyzed using trypsin, LysC, GluC, AspN, and chymotrypsin. After desalting, peptides were fractionated using a strong cation exchange (SCX) column (Zorbax BioSCX-Series II; 0.8 mm inner diameter × 50 mm length, 3.5 µm), and 36 fractions were collected per digest.

Mass spectrometry analysis. The first 26 fractions were analyzed with an Agilent 1290 Infinity (Agilent Technologies) LC, operating in reverse-phase (C18) mode, coupled to a TripleTOF 5600 (AB Sciex). MS spectra (350–1,250 m/z) were acquired in high-resolution mode (R > 30,000), whereas MS2 in high-sensitivity mode (R > 15 000). The next ten fractions were analyzed with a Proxeon EASY-nLC 1000 (Thermo Scientific) operating in reverse phase (C18) and connected to an LTQ-Orbitrap Velos (Thermo Fisher Scientific). For MS analysis, MS spectra (350–1,500 m/z) were acquired at a resolution of 30,000 and for MS2, R = 7,500.

Protein database searching. Peak lists (MGFs) were submitted to the Mascot (version 2.3) via Proteome Discoverer version 1.3 (Thermo Fisher Scientific) and searched against RAT_COMBINED with the respective proteases chosen. Peptide tolerance was 50 ppm, and MS/MS tolerance was 0.1 Da (TOF), 0.02 Da (Orbitrap), and 0.5 Da (ion trap). All PSMs were validated with Percolator [69] based on q = 0 (0% FDR). Only PSMs ranked first by the search engine with at least six amino acids were kept. Unmatched spectra were exported for

analysis with PEAKS Studio (version 6.0). Peak lists were filtered with a quality value of 0.65, followed by a tag database search. The maximum allowed variable PTM per peptide was set to 3. De novo interpreted PSMs were submitted to PEAKS DB database matching, allowing semienzymatic specificity and a maximum cleavage per peptide of 2. The FDR was estimated using a concatenated decoy database and according to a threshold of 0.0%.

Quantitative comparison of proteome and transcriptome data. To combine quantitative data from all methods, we developed a relational database schema (Figure S8) for data storage. The database schema was converted to Java (Java SE 7, Oracle) entities, using Java Persistence API (JPA version 2) implemented in EclipseLink version 2.3.2 (http://www.eclipse.org/eclipselink), with the tools provided in Netbeans IDE 7.3 (http://www.netbeans.org). The database used was MySQL version 5.5 (Oracle).

An extended version of the experimental procedures can be found in the Supplemental Information at the online version of this article (http://www.cell.com/cell-reports/fulltext/S2211-1247(13)00640-2 (doi: 10.1016/j.celrep.2013.10.041)).

Authors' contributions

T.Y.L designed, performed, and analyzed the proteomics experiments; S.v.H. designed, performed, and analyzed the RNA-seq experiments. V.G. performed bioinformatics analysis on genomics, transcriptomics, and proteomics data and is responsible for generating the protein database. H.v.d.T. and V.G. performed qualitative, quantitative, and bioinformatics analysis on both transcriptomics and proteomics data. P.G. and A.C. performed MS and data analysis. B.v.B. and S.M. provided consultation and support for bioinformatics and MS. S.v.H, P.T., and V.G. performed and analyzed RNA-seq validation experiments. S.S. and N.H. provided eQTL data and interpretation. T.Y.L., S.v.H., H.v.d.T., S.M., V.G., E.C., and A.J.R.H. contributed to conceptual design and scientific discussions. T.Y.L., S.v.H., H.v.d.T., B.v.B., S.M., A.J.R.H., E.C., and V.G. wrote the manuscript. E.C. and A.J.R.H. supervised the study.

Acknowledgments

This work was supported by the Netherlands Proteomics Centre, which is part of the Netherlands Genomics Initiative, and a TOP grant from NWO-CW (N_ 700.58.303) to E.C. This work was financially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. HEALTH-F4-2010-241504 (EURATRANS) to E.C. and N.H. and the PRIME-XS project grant agreement number 262067 to A.J.R.H. We would like to thank Dr. Vincentius A. Halim for technical assistance. We would also like to thank the PRIDE Team for assistance.

Data availability

The ProteomeXchange [70] accession number for the MS data reported in this paper is PXD000131. The Sequence Read Archive accession numbers for the DNA data are ERP001355 (BN-Lx genome), ERP001371 (SHR genome), and ERP000510 (BN reference genome). RNA sequencing data were stored in ArrayExpress under the accession number E-MTAB-1666.

Supplementary files

Supplementary files can be found online at http://www.cell.com/cell-reports/fulltext/S2211-1247(13)00640-2 (doi: 10.1016/j.celrep.2013.10.041)

References

1. Altelaar AFM, Munoz J, Heck AJR: Next-generation proteomics: towards an integrative view of proteome dynamics. Nature reviews Genetics 2013, 14:35-48.
2. Cox J, Mann M: Quantitative, high-resolution proteomics for data-driven systems biology. Annu Rev Biochem 2011, 80:273-299.
3. Soon WW, Hariharan M, Snyder MP: High-throughput sequencing for biology and medicine. Mol Syst Biol 2013, 9:640.
4. Ning K, Fermin D, Nesvizhskii AI: Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. J Proteome Res 2012, 11:2261-2271.
5. Munoz J, Low TY, Kok YJ, Chin A, Frese CK, Ding V, Choo A, Heck AJ: The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol Syst Biol 2011, 7:550.
6. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, Mann M: Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 2011, 7:548.
7. Ahrens CH, Brunner E, Qeli E, Basler K, Aebersold R: Generating and navigating proteome maps using mass spectrometry. Nat Rev Mol Cell Biol 2010, 11:789-801.
8. Jensen ON: Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr Opin Chem Biol 2004, 8:33-41.
9. Uhlen M, Ponten F: Antibody-based proteomics for human tissue profiling. Mol Cell Proteomics 2005, 4:384-393.
10. Ozsolak F, Milos PM: RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 2011, 12:87-98.
11. Kleinman CL, Majewski J: Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. Science 2012, 335:1302; author reply 1302.
12. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG: Widespread RNA and DNA sequence differences in the human transcriptome. Science 2011, 333:53-58.
13. Lin W, Piskol R, Tan MH, Li JB: Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. Science 2012, 335:1302; author reply 1302.
14. Pickrell JK, Gilad Y, Pritchard JK: Comment on “Widespread RNA and DNA sequence differences in the human transcriptome”. Science 2012, 335:1302; author reply 1302.
15. Jaffe JD, Berg HC, Church GM: Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 2004, 4:59-77.
16. Renuse S, Chaerkady R, Pandey A: Proteogenomics. Proteomics 2011, 11:620-630.
17. Merrihew GE, Davis C, Ewing B, Williams G, Kall L, Frewen BE, Noble WS, Green P, Thomas JH, MacCoss MJ: Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res 2008, 18:1660-1669.

18. Venter E, Smith RD, Payne SH: Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One 2011, 6:e27587.

19. de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C: Global signatures of protein and mRNA expression levels. Mol Biosyst 2009, 5:1512-1526.

20. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R: Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics 2006, 5:652-670.

21. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: Global quantification of mammalian gene expression control. Nature 2011, 473:337-342.

22. Vogel C, Marcotte EM: Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature reviews Genetics 2012, 13:227-232.

23. Printz MP, Jirout M, Jaworski R, Alemayehu A, Kren V: Genetic Models in Applied Physiology. HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics. J Appl Physiol (1985) 2003, 94:2510-2522.

24. Atanur SS, Birol I, Guryev V, Hirst M, Hummel O, Morrissey C, Behmoaras J, Fernandez-Suarez XM, Johnson MD, McLaren WM, et al: The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Res 2010, 20:791-803.

25. Simonis M, Atanur SS, Linsen S, Guryev V, Ruzius FP, Game L, Lansu N, de Bruijn E, van Heesch S, Jones SJ, et al: Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. Genome Biol 2012, 13:r31.

26. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, et al: Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet 2005, 37:243-253.

27. Johnson MD, He L, Herman D, Wakimoto H, Wallace CA, Zidek V, Mlejnek P, Musilova A, Simakova M, Vorlicek J, et al: Dissection of chromosome 18 blood pressure and salt-sensitivity quantitative trait loci in the spontaneously hypertensive rat. Hypertension 2009, 54:639-645.

28. Pravenec M, Kurtz TW: Recent advances in genetics of the spontaneously hypertensive rat. Curr Hypertens Rep 2010, 12:5-9.

29. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al: Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 2004, 428:493-521.

30. Jansen RC, Nap JP: Genetical genomics: the added value from segregation. Trends Genet 2001, 17:388-391.

31. Sun A, Jiang Y, Wang X, Liu Q, Zhong F, He Q, Guan W, Li H, Sun Y, Shi L, et al: Liverbase: a comprehensive view of human liver biology. J Proteome Res 2010, 9:50-58.

32. Hurst JM, McMillan LE, Porter CT, Allen J, Fakorede A, Martin AC: The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum Mutat 2009, 30:616-624.

33. Su ZD, Sun L, Yu DX, Li RX, Li HX, Yu ZJ, Sheng QH, Lin X, Zeng R, Wu JR: Quantitative detection of single amino acid polymorphisms by targeted proteomics. J Mol Cell Biol 2011, 3:309-315.

34. Valentine SJ, Sevugarajan S, Kurulugama RT, Koeniger SL, Merenbloom SI, Bohrer BC, Clemmer DE: Split-field drift tube/mass spectrometry and isotopic labeling techniques for determination of single amino acid polymorphisms. J Proteome Res 2006:5,1879-1887.

35. Volkening JD, Bailey DJ, Rose CM, Grimsrud PA, Howes-Podoll M, Venkateshwaran M, Westphall MS, Ane JM, Coon JJ, Sussman MR: A proteogenomic survey of the Medicago truncatula genome. Mol Cell Proteomics 2012, 11:933-944.

36. Burge C, Karlin S: Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997, 268:78-94.

37. Farajollahi S, Maas S: Molecular diversity through RNA editing: a balancing act. Trends Genet 2010, 26:221-230.

38. Mohammed S, Lorenzen K, Kerkhoven R, van Breukelen B, Vannini A, Cramer P, Heck AJ: Multiplexed proteomics mapping of yeast RNA polymerase II and III allows near-complete sequence coverage and reveals several novel phosphorylation sites. Anal Chem 2008, 80:3584-3592.

39. Peng M, Taouatas N, Cappadona S, van Breukelen B, Mohammed S, Scholten A, Heck AJ: Protease bias in absolute protein quantitation. Nat Methods 2012, 9:524-525.

40. Swaney DL, Wenger CD, Coon JJ: Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res 2010, 9:1323-1329.

41. Dormeyer W, Mohammed S, Breukelen B, Krijgsveld J, Heck AJ: Targeted analysis of protein termini. J Proteome Res 2007, 6:4634-4645.

42. Starheim KK, Gevaert K, Arnesen T: Protein N-terminal acetyltransferases: when the start matters. Trends Biochem Sci 2012, 37:152-161.

43. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A: Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 2013, 9:59-64.

44. Grobei MA, Qeli E, Brunner E, Rehrauer H, Zhang R, Roschitzki B, Basler K, Ahrens CH, Grossniklaus U: Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. Genome Res 2009, 19:1786-1800.

45. Nesvizhskii AI, Aebersold R: Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 2005, 4:1419-1440.

46. Bensimon A, Heck AJ, Aebersold R: Mass spectrometry-based proteomics and network biology. Annu Rev Biochem 2012, 81:379-405.

47. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 2004, 32:D91-94.

48. Okamoto K, Aoki K: Development of a strain of spontaneously hypertensive rats. Jpn Circ J 1963, 27:282-293.

49. Aitman TJ, Glazier AM, Wallace CA, Cooper LD, Norsworthy PJ, Wahid FN, Al-Majali KM, Trembling PM, Mann CJ, Shoulders CC, et al: Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. Nat Genet 1999, 21:76-83.

50. Aitman TJ, Gotoda T, Evans AL, Imrie H, Heath KE, Trembling PM, Truman H, Wallace CA, Rahman A, Dore C, et al: Quantitative trait loci for cellular defects in glucose and fatty acid metabolism in hypertensive rats. Nat Genet 1997, 16:197-201.

51. Lee SJ, Liu J, Qi N, Guarnera RA, Lee SY, Cicila GT: Use of a panel of congenic strains to evaluate differentially expressed genes as candidate genes for blood pressure quantitative trait loci. Hypertens Res 2003, 26:75-87.

52. Kinoshita K, Ashenagar MS, Tabuchi M, Higashino H: Whole rat DNA array survey for candidate genes related to hypertension in kidneys from three spontaneously hypertensive rat substrains at two stages of age and with hypotensive induction caused by hydralazine hydrochloride. Exp Ther Med 2011, 2:201-212.

53. Dwinell MR, Worthey EA, Shimoyama M, Bakir-Gungor B, DePons J, Laulederkind S, Lowry T, Nigram R, Petri V, Smith J, et al: The Rat Genome Database 2009: variation, ontologies and pathways. Nucleic Acids Res 2009, 37:D744-749.

54. Li X, Ling Y, Lu D, Lu Z, Liu Y, Chen H, Gao X: Common polymorphism rs11191548 near the CYP17A1 gene is associated with hypertension and systolic blood pressure in the Han Chinese population. Am J Hypertens 2013, 26:465-472.

55. Liu C, Li H, Qi Q, Lu L, Gan W, Loos RJ, Lin X: Common variants in or near FGF5, CYP17A1 and MTHFR genes are associated with blood pressure and hypertension in Chinese Hans. J Hypertens 2011, 29:70-75.

56. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, Najjar SS, Zhao JH, Heath SC, Eyheramendy S, et al: Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet 2009, 41:666-676.

57. Takeuchi F, Isono M, Katsuya T, Yamamoto K, Yokota M, Sugiyama T, Nabika T, Fujioka A, Ohnaka K, Asano H, et al: Blood pressure and hypertension are associated with 7 loci in the Japanese population. Circulation 2010, 121:2302-2309.

58. Danielson PB: The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. Curr Drug Metab 2002, 3:561-597.

59. Biglieri EG: 17 alpha-Hydroxylase deficiency: 1963-1966. J Clin Endocrinol Metab 1997, 82:48-50.

60. Biglieri EG, Herron MA, Brust N: 17-hydroxylation deficiency in man. J Clin Invest 1966, 45:1946-1954.

61. Geller DH, Auchus RJ, Mendonca BB, Miller WL: The genetic and functional basis of isolated 17,20-lyase deficiency. Nat Genet 1997, 17:201-205.

62. Goldsmith O, Solomon DH, Horton R: Hypogonadism and mineralocorticoid excess. The 17-hydroxylase deficiency syndrome.

N Engl J Med 1967, 277:673-677.

63. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, et al: An overview of Ensembl. Genome Res 2004, 14:925-928.

64. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: The Ensembl automatic gene annotation system. Genome Res 2004, 14:942-950.

65. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, et al: The Ensembl genome database project. Nucleic Acids Res 2002, 30:38-41.

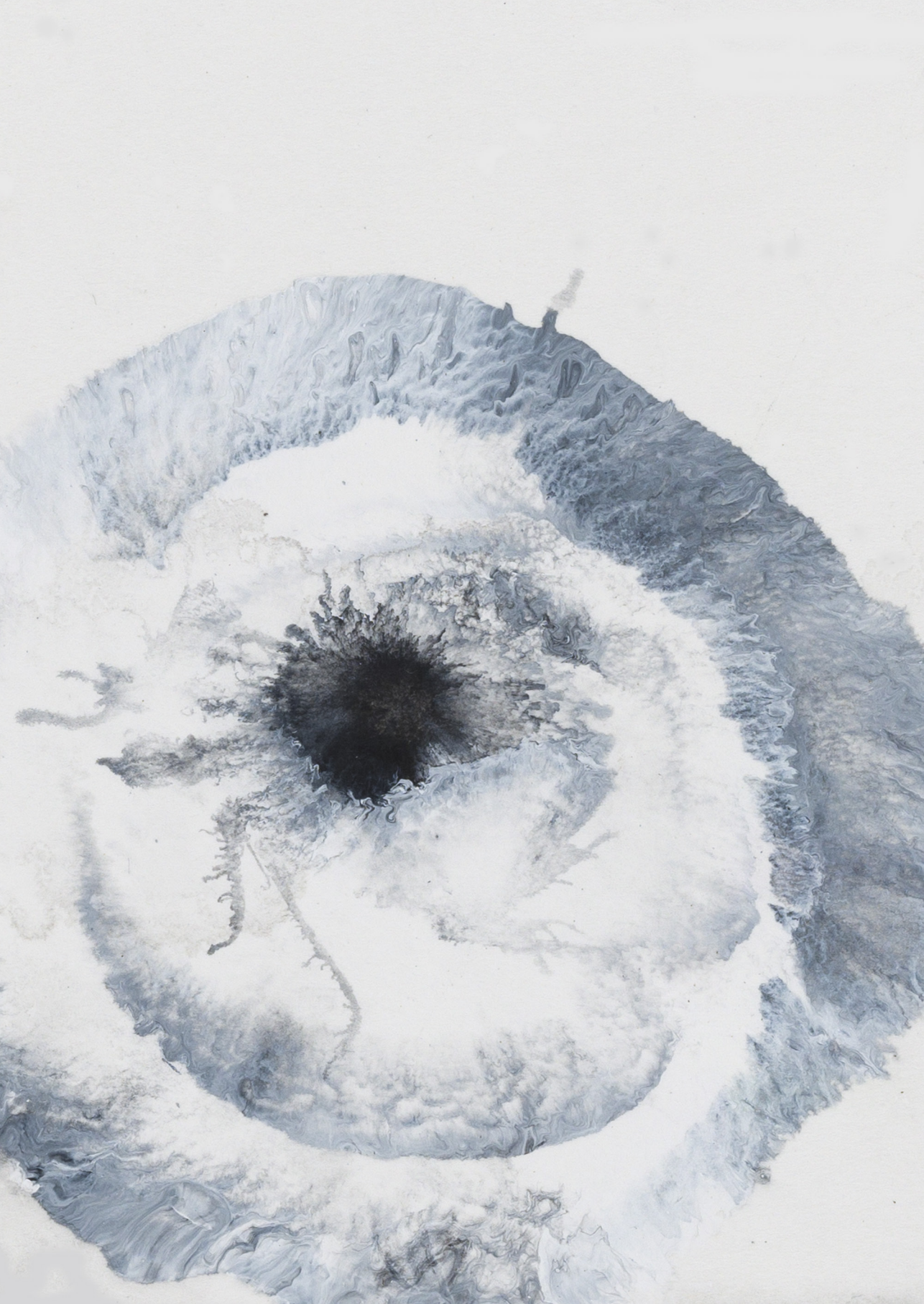
66. Kent WJ: BLAT--the BLAST-like alignment tool. Genome Res 2002, 12:656-664.

67. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012, 7:562-578.

68. Heinig M, Petretto E, Wallace C, Bottolo L, Rotival M, Lu H, Li Y, Sarwar R, Langley SR, Bauerfeind A, et al: A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature 2010, 467:460-464.

69. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ: Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 2007, 4:923-925.

70. Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, et al: The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res 2013, 41:D1063-1069.



5

Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes

Sebastiaan van Heesch[1], Maarten van Iterson[1], Jetse Jacobi[1], Sander Boymans[1], Paul B Essers[2], Ewart de Bruijn[1], Wensi Hao[1], Alyson W MacInnes[2], Edwin Cuppen[1,3] and Marieke Simonis[1]

[1]Genome Biology Group, Hubrecht Institute-KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands
[2]Ribosome Biogenesis and Disease Group, Hubrecht Institute-KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands
[3]Department of Medical Genetics, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands

Adapted from Genome Biology 2014, 15:R6

Abstract

Long noncoding RNAs (lncRNAs) form an abundant class of transcripts, but the function of the majority of them remains elusive. While it has been shown that some lncRNAs are bound by ribosomes, it has also been convincingly demonstrated that these transcripts do not code for proteins. To obtain a comprehensive understanding of the extent to which lncRNAs bind ribosomes, we performed systematic RNA sequencing on ribosome-associated RNA pools obtained through ribosomal fractionation and compared the RNA content with nuclear and (non-ribosome bound) cytosolic RNA pools. The RNA composition of the subcellular fractions differs significantly from each other, but lncRNAs are found in all locations. A subset of specific lncRNAs is enriched in the nucleus but surprisingly the majority is enriched in the cytosol and in ribosomal fractions. The ribosomal enriched lncRNAs include ***H19*** and ***TUG1***. Most studies on lncRNAs have focused on the regulatory function of these transcripts in the nucleus. We demonstrate that only a minority of all lncRNAs are nuclear enriched. Our findings suggest that many lncRNAs may have a function in cytoplasmic processes, and in particular in ribosome complexes.

Introduction

The importance of noncoding RNA transcripts for key cellular functions has been well established by studies on for example ***XIST*** [1], which acts in X-chromosome silencing, and ***TERC*** [2], which functions in telomeric maintenance. Genomic studies performed in the last decade have shown that these are likely not isolated examples as many more long non protein-coding transcripts were identified [3-5]. Although it remains to be demonstrated that all of these transcripts have specific functions [6], functional studies showing the importance of long noncoding RNAs (lncRNAs) as regulators in cellular pathways are accumulating rapidly (for example, [7-12]). However, the function and the mechanisms of action of the majority of lncRNAs are still unexplored [13].

Cellular location is an important determinant in understanding the functional roles of lncRNAs. Subcellular RNA sequencing (RNA-seq) has been performed to explore the differences between nuclear, chromatin-associated and cytoplasmic transcript content in several cell lines [14] and macrophages [15]. Derrien ***et al***. [3] specifically estimated the relative abundance of lncRNAs in the nucleus versus the cytosol and concluded that 17% of the tested lncRNAs were enriched in the nucleus and 4% in the cytoplasm. This is in line with the function of some individual lncRNAs, such as ***NEAT1*** and ***MALAT1***, which were shown to be involved in nuclear structure formation and gene expression regulation [7,8]. However, it has been argued that relative enrichment does not mean that the absolute number of transcripts for each lncRNA is also higher in the nucleus [13]. Some lncRNAs were enriched in the cytoplasm and ribosome profiling demonstrated that part of the cytoplasmic lncRNAs is bound by ribosomes [16]. More detailed characterization of the ribosome profiling data showed that

ribosomal occupation of lncRNAs does not match with specific marks of translation [17].
While these results suggest diverse roles of lncRNAs in different cellular compartments and biological processes, comprehensive knowledge on the relative abundances of lncRNAs in ribosomes, the cytosol and the nucleus is currently still lacking. Moreover, as ribosomal profiling measures single sites in RNA molecules that are occupied by ribosomes, this technique does not yield information on the number of ribosomes that are present per single (physical) lncRNA transcript [18]. In a different method, named ribosomal fractionation, a cytosolic size separation is performed that results in the isolation of translation complexes based on the number of ribosomes associated per transcript [19]. This method has been used in combination with microarrays to analyze ribosomal density on protein-coding transcripts [20-22] but not on lncRNAs.
Here we perform subcellular RNA-seq on nuclei, cytosol and mono- and polyribosomes separated by ribosomal fractionation. Our data show relative enrichment of specific lncRNAs in the nucleus, but also demonstrate that most lncRNAs are strongly enriched in the cytosol and in ribosomal fractions.

Results

Nuclear, cytosolic and ribosomal fractions differ in transcript content

Different subcellular RNA fractions were isolated from the human cell line LS-174T-pTER-*β*-catenin [23] (Figure 1). The cells were first subjected to a mild lysis after which the nuclei were separated from the cytosol and other organelles by centrifugation. Microscopic inspection and nuclear staining confirmed the presence of clean nuclei in the pellet and thus the co-sedimentation of the rough endoplasmic reticulum-derived ribosomes with the cytosolic supernatant (Additional file 1). The cytosolic sample was fractionated further using a sucrose gradient and ultracentrifugation, which sediments the sample components based on size and molecular weight. UV was used to measure the RNA content of the fractions and the number of ribosomes in each of the fractions was established based on the resulting distinct peak pattern. We isolated each of the fractions containing one, two, three, four, five and six ribosomes and the fraction containing seven or more ribosomes. In addition, we isolated the fraction that contained the cytosolic part without ribosomes, which we will refer to as the 'free cytosolic' sample. RNA molecules in the free cytosolic fraction are, however, associated with various other types of smaller protein complexes that reside in the cytosol. The fractions containing 40S and 60S ribosomal subunits were also extracted and these two samples were pooled for further analysis. The RNA of three ribosomal fractionation experiments was pooled to level out single experimental outliers. Through this experimental setup we obtained a complete set of subcellular samples from which RNA was extracted.

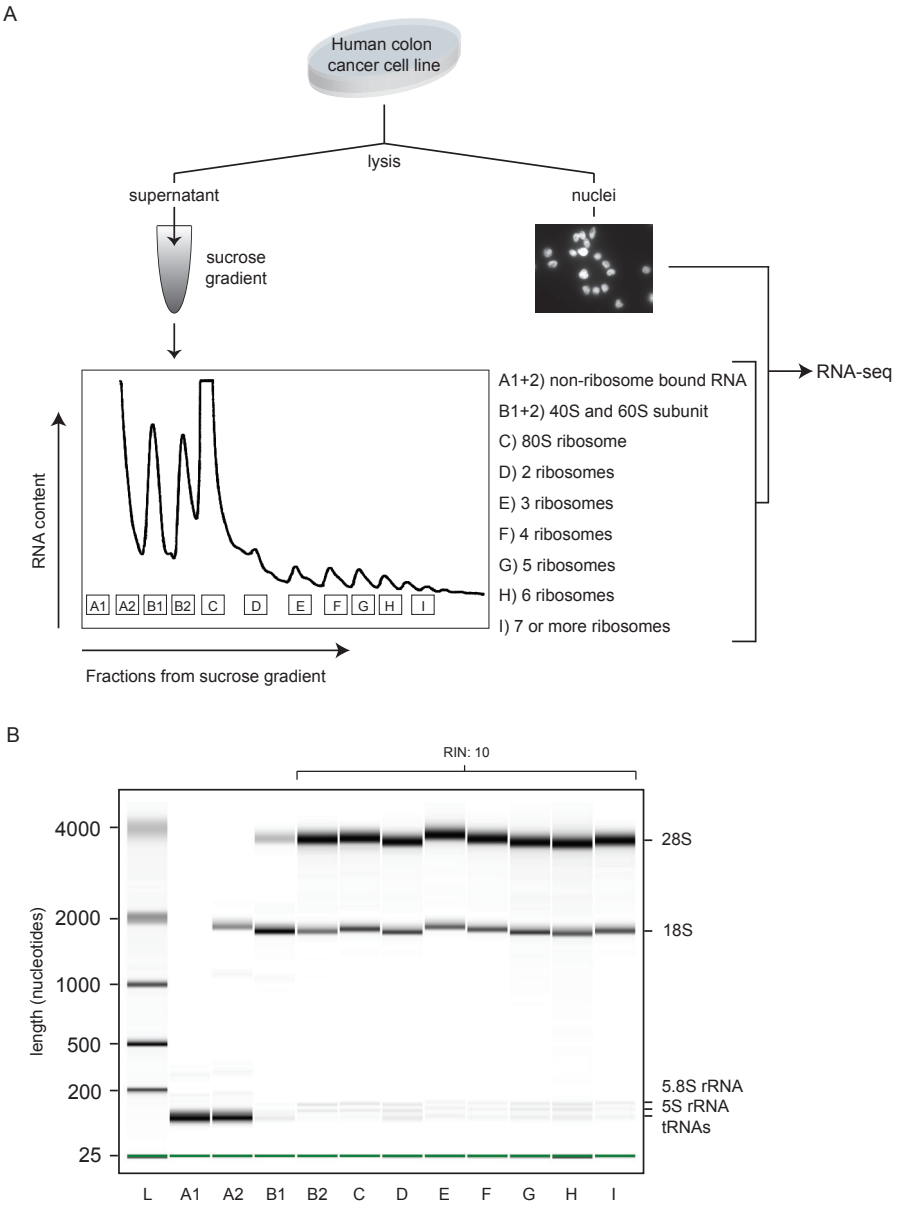


Figure 1 - **Experimental workflow and quality control.** (**A**) Cells were lysed and the complete cytosolic fraction was used for ribosomal fractionation. Pelleted nuclei and nine fractions (indicated A to I) derived from the ribosomal fractionation were subsequently used for RNA isolation and strand-specific RNA-seq. Fractions A1 and A2 as well as B1 and B2 were merged prior to the RNA-seq. (**B**) 2100 Bioanalyzer RNA 6000 Pico results showing the integrity of the collected RNA samples obtained by ribosomal fractionation. Each ribosomal fraction has an RNA integrity (RIN) value of 10. These results also show the sample-specific content of tRNAs, 5S, 5.8S, 18S and 28S rRNA, which nicely indicate the purity of the fractionation.

Strand-specific RNA-seq was performed after rRNA depletion on all the subcellular samples and for each we obtained at least six million aligned reads. The GENCODE annotation [24] of coding and noncoding transcripts was used to establish the read counts per gene (Additional file 2). In our data analyses, we considered three types of transcripts: protein-coding transcripts; small noncoding RNAs (sncRNAs), which included small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs); and lncRNAs, which included antisense transcripts, long intergenic noncoding RNAs and processed transcripts (these were transcripts that did not contain an open reading frame (ORF) and could not be placed in any of the other categories) [3]. We left out some small RNAs such as miRNAs, because these were not captured in our experimental setup. Also, to prevent false assignments of sequencing reads to noncoding transcripts, we did not consider lncRNAs in which the annotation partially overlapped with protein-coding transcripts on the same strand. We selected expressed transcripts using a stringent threshold to allow us to reliably detect quantitative differences. Our expressed transcript set contained 7,734 genes including 7,206 protein-coding genes, 152 lncRNAs (46 antisense transcripts, 71 long intergenic noncoding transcripts and 35 processed transcripts) and 376 sncRNAs (134 snoRNAs and 242 snRNAs).

To determine the similarity of the RNA content of the different subcellular samples we analyzed the correlations between each sample pair (Figure 2A). The highest correlations were seen between ribosomal fractions, ranging from 0.60 to 0.97. By contrast, the correlations between the different ribosomal fractions and the nuclear sample ranged from 0.35 to 0.53. We investigated the source of the variable correlation between subcellular RNA samples by comparing the origin of the RNA reads from each fraction (Figure 2B). This analysis showed that more than half of the reads in the nuclear sample aligned to sncRNAs and this group of small RNAs was visible as a distinct cloud in the comparative scatter plots (Figure 2A and Additional file 3). The ribosomal fractions primarily consisted of protein-coding genes as expected, but highly expressed lncRNAs were also clearly present. Because these read count distributions did not directly translate into transcript composition of the different samples, we also analyzed the sample composition based on reads per kilobase per million. This resulted in essentially the same distribution among the samples, but the relative contribution of sncRNAs was larger (Additional file 4).
Combined, these analyses show that subcellular RNA samples have very different compositions and that lncRNAs are found in each of the subcellular RNA samples.

Long noncoding RNAs are primarily enriched in the cytosol and in the ribosomal fractions

The clear difference in composition of the subcellular RNA samples raises the question how individual transcripts are distributed among the samples and in particular how lncRNAs behave compared to protein-coding transcripts. Therefore we investigated the distribution

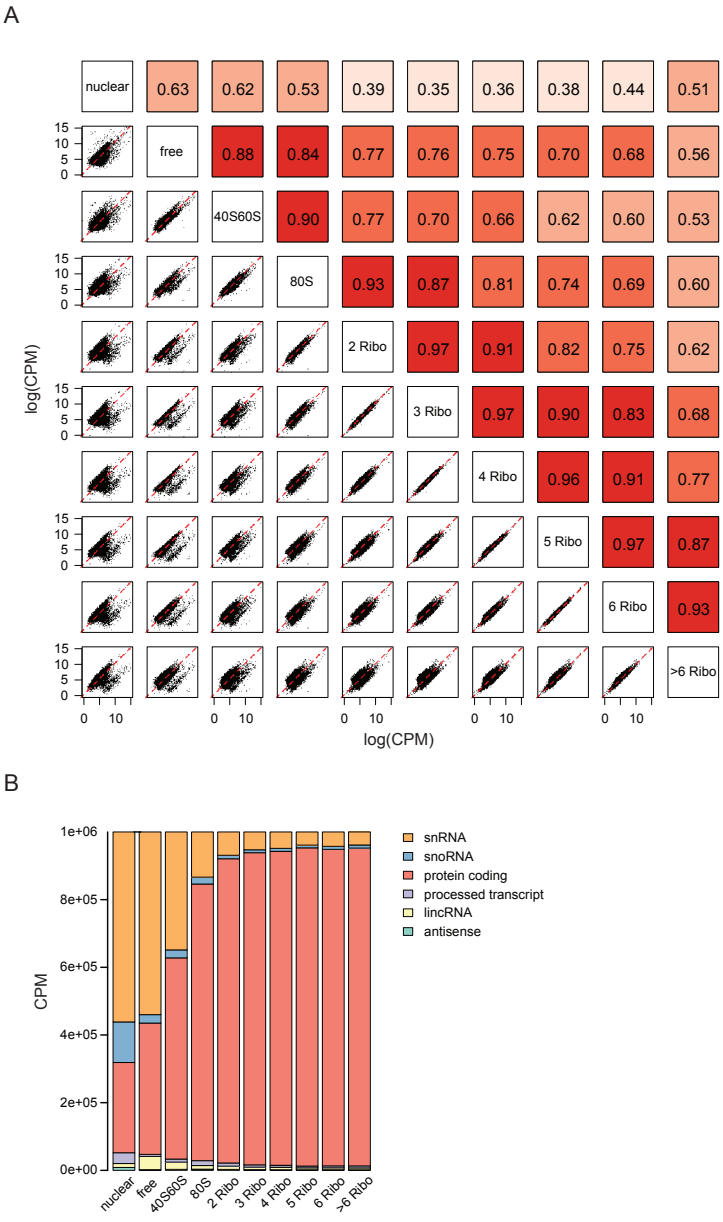


Figure 2 - **Subcellular RNA fractions have a different transcript composition**. (**A**) Scatter plot and correlation matrix of all sequenced samples. The color intensity of the correlation boxes (r values) depicts the relative strength of the correlation, ranging between 0.39 and 0.97. (**B**) RNA species content of each sequenced fraction in counts per million (CPM).

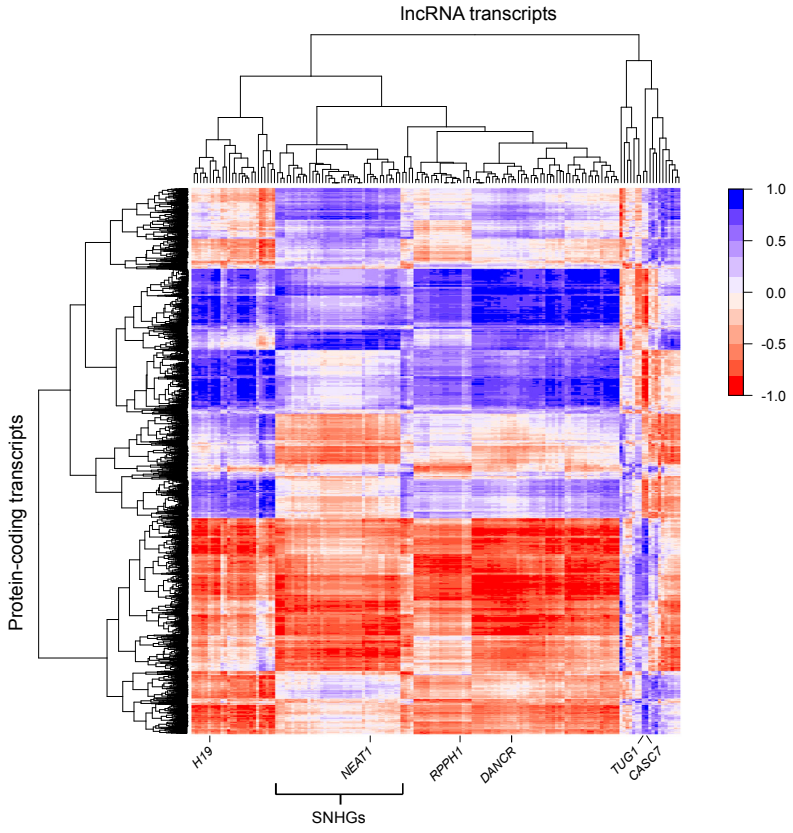


Figure 3 - **Long noncoding RNAs show a subcellular distribution similar to specific groups of protein-coding transcripts.** Heatmap of the Spearman-Rank correlation between the each of the 152 expressed lncRNAs and 7,206 expressed protein-coding transcripts across the subcellular RNA samples. Strong correlations are shown in blue, anti-correlations are shown in red. Six frequently studied lncRNAs with varying correlations to protein-coding transcripts are highlighted at the bottom together with a large cluster that harbors the majority of expressed snoRNA host genes.

of each lncRNA across the cellular fractions versus the distribution of each protein-coding transcript (Figure 3). The correlation between each protein-coding transcript-lncRNA pair was calculated and the obtained scores depicted in a clustered heatmap (Figure 3). A high correlation between two transcripts in this heatmap meant that the two showed a very similar distribution across all different subcellular samples. This analysis showed that there are several different groups of lncRNAs that can be distinguished based on their correlation with protein-coding transcripts. Each group of lncRNAs had specific sets of positively correlated and negatively correlated protein-coding transcripts. Examples of such groups are the noncoding snoRNA host genes, that all showed very similar correlation profiles (Figure 3). A few lncRNAs, including ***TUG1*** and ***CASC7***, had a more specific correlation profile.

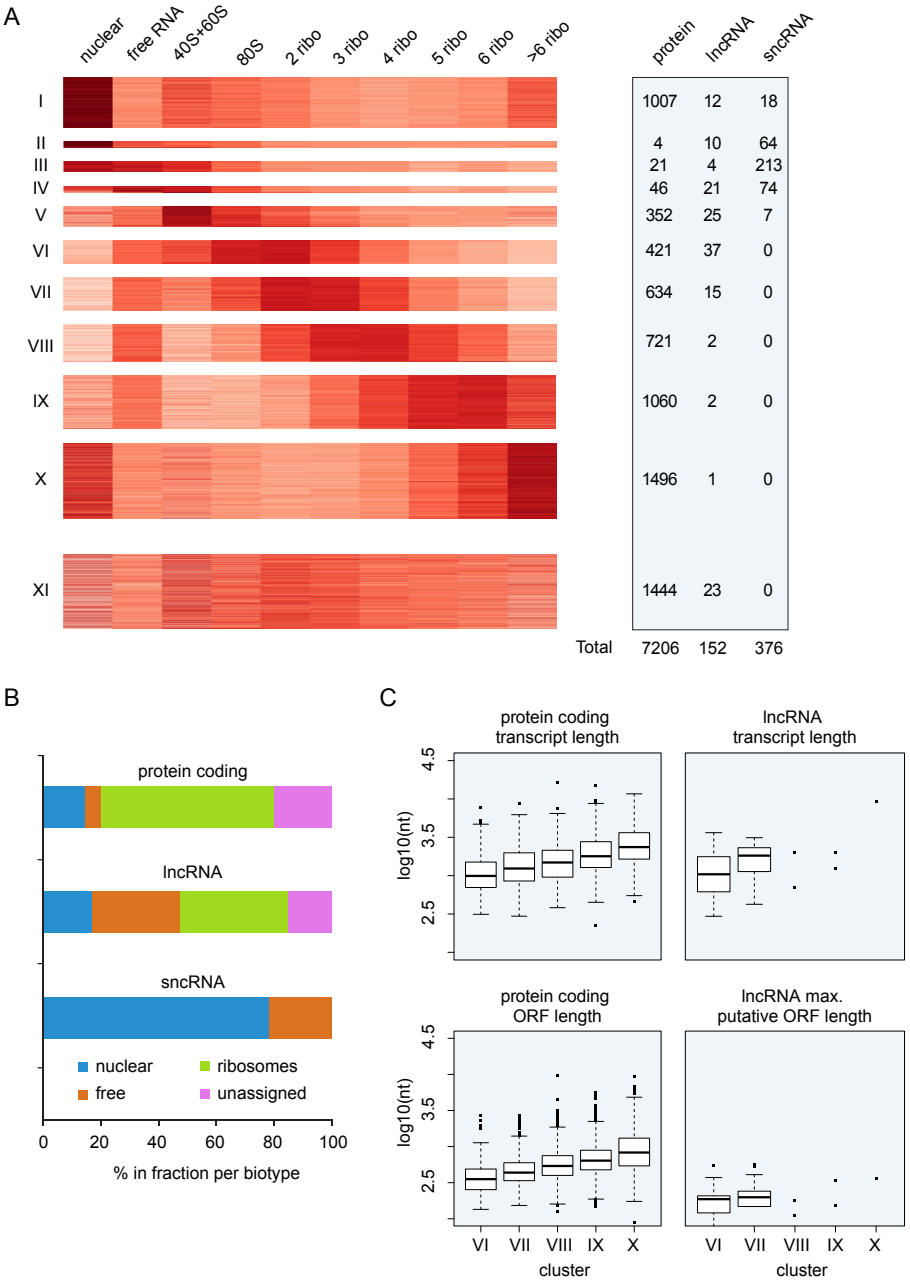


Figure 4 - **RNA species show specific distributions across the subcellular RNA samples.** (**A**) Heatmap display of the 11 clusters and the number of protein-coding, lncRNA and sncRNA transcripts present in each cluster. (**B**) Summarizing plot showing the distribution of the three types of transcripts over the four major types of clusters that could be derived from the analysis in (**A**). (**C**) Boxplots of the total transcript length and the maximum (potential) open reading frame of protein-coding transcripts and lncRNAs in clusters VI to X.

These results show that there is no general negative correlation between cellular localization of lncRNAs and protein-coding transcripts, but that the relationships are complex.

To reduce this complexity and to focus on the distribution of protein-coding transcripts and non-protein-coding RNAs across the subcellular fractions we applied model-based clustering on the normalized read counts per transcript [25]. We applied the clustering algorithm using variable amounts of clusters and found that a separation in 11 clusters best describes the data (Figure 4A and Additional files 5 and 6). All RNA-seq transcript levels were normalized to the total number of sequencing reads produced per sample. Therefore, the normalized value of a transcript depended on the complexity of the sample (number of different transcripts) and the expression level of all other transcripts. Because of the large fraction of reads that arose from sncRNAs, we tested the effect of omitting these RNAs from the dataset and found that this did not affect the clustering results (Additional file 7). The final set of 11 clusters included one cluster (XI) containing transcripts that did not show an obvious enrichment in any of the samples, and 10 clusters (I to X) containing genes that did show a specific cellular localization. Clusters I, II and III all contained transcripts enriched in the nucleus and depleted from the ribosomal fractions, but the clusters differed from each other based on the relative transcript levels in the free cytosolic and the 40S/60S sample. Cluster IV and V contained transcripts enriched in the free cytosolic sample and transcripts enriched in the 40S/60S sample, respectively. Clusters VI through X contained transcripts enriched in specific ribosomal fractions. Each of these ribosomal-enriched clusters also showed mild enrichment in the free cytosolic sample, except for cluster X, which was higher in the nucleus than in the free cytosol.

Overall, we consider clusters I, II and III as enriched in the nucleus; IV and V as enriched in the ribosome-free cytosol; and VI, VII, VIII, IX and X as enriched in the ribosomes. The distribution of protein-coding genes and sncRNAs among the clusters was largely as expected (Figure 4B). Protein-coding transcripts were present in all of the clusters, but the majority (60%) was found in the ribosomal-enriched clusters. Nonetheless, 14% of the protein-coding transcripts were found in the nuclear clusters and depleted from ribosomes, suggesting that this large part of the protein-coding transcripts is not actively translated or has a rapid turn-over rate in the cytosol. sncRNAs were found only in the nuclear and ribosome-free cytosolic clusters and not in the ribosomal clusters, which matched expectations and thus demonstrated the effectiveness of the fractionation. The majority of the sncRNAs could be found in cluster III, showing high levels both in the nucleus and free in the cytosol, suggesting that many of these small RNAs shuttle between nucleus and cytoplasm.

The most notable result was the distribution of the lncRNAs among the different clusters. In line with previous analyses [3], 17% of the lncRNAs were found in one of the nuclear clusters

(Figure 4B). However, in contrast to previous studies, a relatively large part of the lncRNAs (30%) was located in clusters enriched in the ribosome-free cytosol and a striking 38% was present in ribosome-enriched clusters. As noted above, the transcript levels determined by RNA-seq represent which part of the total RNA samples can be assigned to each specific transcript. These results thus show that many individual lncRNAs (38% of the expressed lncRNAs) make up a larger part of specific ribosomal fractions than of the nuclear sample.

Although the correlations between ribosomal fractions were high (Figure 2A), these clustering results highlight the transcripts that are differential across the ribosomal samples. Previous studies have shown that many protein-coding transcripts are not evenly distributed among the ribosomal fractions, but rather show enrichment for a specific number of ribosomes [20,21]. The coding sequence length was shown to be a major determinant of the modular number of ribosomes per transcript. In our data, the total transcript length of protein-coding transcripts in the five ribosomal clusters also increased with increasing numbers of ribosomes present (Figure 4C). For lncRNAs, we could determine such a relationship only between cluster VI (80S and two ribosomes) and VII (three and four ribosomes), because the number of lncRNAs in the clusters with a higher number of ribosomes was too low (Figure 4A). lncRNAs in cluster VII (three and four ribosomes) had a longer transcript length, longer maximum putative ORF length and more start codons than the lncRNAs in cluster VI (80S and two ribosomes) (Figure 4C and Additional file 8). However, the maximum ORF lengths of the lncRNAs were much shorter than the coding sequence length of the protein-coding genes in the same cluster, so these ORF lengths likely do not determine the number of ribosomes associated with a lncRNA.

Combined, these analyses showed that many lncRNAs were enriched in specific subcellular fractions. Although some lncRNAs were enriched in the nucleus, many more were enriched in the cytosolic and ribosomal fractions.

Known long noncoding RNAs are enriched in different ribosomal fractions

The cellular localization of some lncRNAs was established previously and our results were largely in agreement with earlier findings. For example, ***MALAT1*** and ***NEAT1***, which are known to regulate nuclear processes such as gene expression [8] and the formation and maintenance of nuclear speckles and paraspeckles [7,26] respectively, were located in nuclear cluster I (Figure 5). Another lncRNA with a known nuclear function is ***TUG1*** (Figure 5), which is involved in the upregulation of growth-control genes [27]. We indeed found high levels of ***TUG1*** in the nucleus, but the transcript also showed a clear enrichment in the fractions containing five or six ribosomes. The association of ***TUG1*** with polysomes has not been described previously and suggests mechanisms of action in regulation of translation at the ribosome in addition to the previously described function in the nucleus.

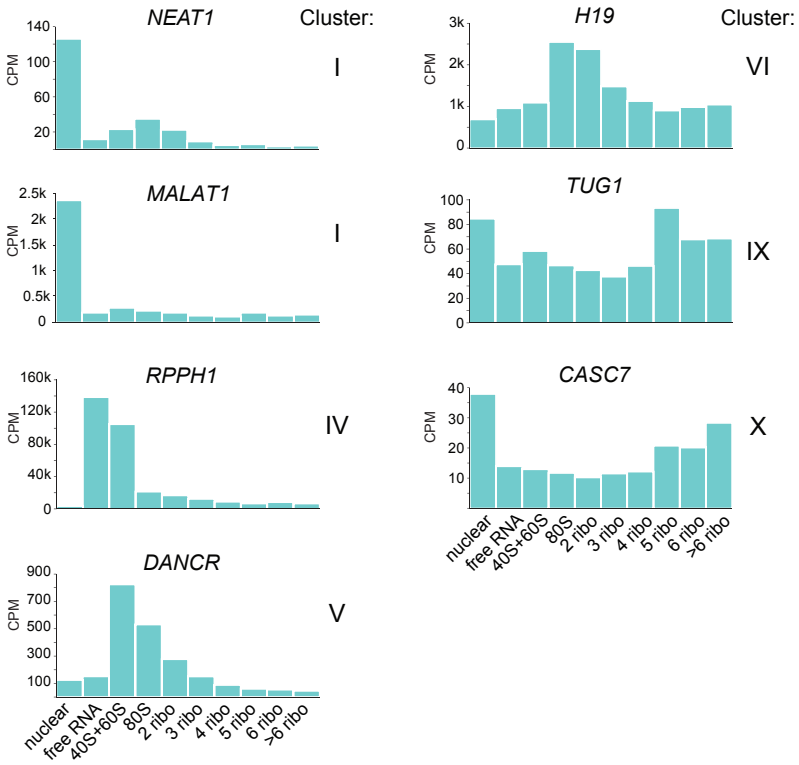


Figure 5 - **Individual long noncoding RNAs are differentially distributed across subcellular samples**. The normalized read counts of seven lncRNAs that are found in different clusters in Figure 4.

In the ribosome-free cytosolic sample we found enrichment of lncRNAs that are known components of cytosolic complexes, for example ***RPPH1*** and ***RN7SL1. RPPH1*** is part of ribonuclease P [28] and ***RN7SL1*** is part of the signal recognition particle that mediates co-translational insertion of secretory proteins into the lumen of the endoplasmic reticulum [29,30]. In addition, we also found many unstudied lncRNAs in the free cytosolic fraction. In cluster V, which showed enrichment in the 40S/60S sample, we found the lncRNA ***DANCR*** (Figure 5). ***DANCR*** was recently shown to be involved in retaining an undifferentiated progenitor state in somatic tissue cells [10] and osteoblast differentiation [31]. The exact mechanisms through which ***DANCR*** acts are unknown, but our data suggest a role for ***DANCR*** predominantly outside of the nucleus. One of the most abundant lncRNAs in our data was the evolutionary conserved and imprinted ***H19***. This transcript is a strong regulator of cellular growth and overexpression of ***H19*** contributes to tumor initiation as well as progression, making it a frequently studied noncoding RNA in cancer [9,32]. An enrichment of ***H19*** in the cytoplasm over the nucleus has previously been observed [3]. Here, we found only moderate

levels of *H19* RNA in the nucleus and ribosome-free cytosol, but very high levels of *H19* RNA associated with ribosomes (Figure 5). This predominant association with ribosomes suggests a possible role for *H19* in the regulation of the translation machinery and, more specifically, in polysomal complexes.

CASC7 was the only lncRNA that was enriched in the sample with seven or more ribosomes. Even though *CASC7* has been identified as a cancer susceptibility candidate, not much is known about this transcript. Our data indicate that it is sequestered to large polysomal complexes and it may thus function in regulation of translation.

Using quantitative PCR, we confirmed the enrichment of *NEAT1* and *MALAT1* in the nucleus and the enrichment of *TUG1* and *H19* in ribosomes (Additional file 9).

These results reveal the subcellular enrichment of known and unknown lncRNAs and suggest that many lncRNAs function primarily outside the nucleus.

Discussion

We performed transcriptome analyses on subcellular samples of the human cell line LS-174T-pTER-*β*-catenin and found that the lncRNAs that were expressed in these cells were present in all subcellular fractions, but the majority of the expressed lncRNAs were enriched in the cytosol and in ribosomes. Our data partially contradict an earlier study in which most lncRNAs were found enriched in the nucleus, compared to the cytoplasm [3]. This discrepancy could have resulted from the use of different cell types, but may also have partially resulted from measuring and comparing relative enrichments between multiple samples. Measuring the whole cytoplasm would thus result in different enrichment values compared to analysis of a specific subset of the cytoplasm, such as the ribosomes.

We are not the first to find lncRNAs associated with ribosomes. Ribosome profiling in mouse embryonic stem cells also showed examples of these interactions and our results overlap with the results from that study [16]. For example, both our work and work from Ingolia *et al*. pinpoint the lncRNA *NEAT1* as not highly associated with ribosomes. The results for *MALAT1* are more intricate, as we found that *MALAT1* was strongly enriched in the nucleus, but previous work showed binding of ribosomes to the 5'-part of this lncRNA [16,33]. It is possible that a small proportion of the *MALAT1* transcripts is bound by ribosomes. It is also likely hat ribosomal association with lncRNAs is specific to cell type, growth condition and organism.

Our data add significant insight into ribosomal association of lncRNAs, because ribosomal profiling and ribosomal fractionation provide different, yet complementary, information. In ribosome profiling, specific binding sites of ribosomes are measured and the amount of binding is estimated based on the total number of reads in the ribosome-bound versus the total RNA sample. By applying ribosomal fractionation we can directly measure the number of ribosomes associated per lncRNA. Moreover, we measured the full range of subcellular samples including free cytosolic and nuclear RNA in one analysis. From our data we can

conclude that many lncRNAs are found in complexes that contain multiple ribosomes. In addition, the enrichment of lncRNAs in ribosomal fractions shows that many lncRNAs make up a relatively larger part of the ribosomal samples than of the nuclear sample. This did not change when sncRNAs were excluded from the analyses. It should be noted that the identification of the ribosomes was based on size fractionation and RNA content. We can therefore not fully exclude that the lncRNAs are associating with protein complexes of sizes similar to the specific number of ribosomes [34]. However, these thus far unknown complexes would have to be present in such high quantities that the result is an enrichment of the associated transcripts equal to the enrichment of protein-coding transcripts. Moreover, we found lncRNAs in different ribosomal fractions, so the alternative explanation would require the involvement of multiple different protein complexes.

So why do lncRNAs associate with ribosomes? The possibility that these lncRNAs all code for proteins was recently eliminated by in-depth comparison of ribosome occupancy around translation termination codons [17]. lncRNAs did not show a steep drop in ribosomal binding after the translation termination codons (determined by the ribosome release score), as was seen for protein-coding genes. However, that does not exclude the possibility that ribosomes spuriously bind initiation codons in lncRNAs. In our data, the number of ribosomes per lncRNA correlates with lncRNA length, maximum ORF length and the number of ORFs present per lncRNA, but those three factors are not independent of each other.

It is possible that one of the processes that keep lncRNAs at ribosomes is nonsense-mediated decay (NMD). NMD functions via ribosomal binding and has previously been described as a possible breakdown route of the noncoding RNA *GAS5* [35]. However, if NMD of a transcript results in such strong enrichment in the ribosomal fractions as observed in our experiments, it would mean that under standard culturing conditions a very significant portion of transcripts at ribosomes are engaged in NMD and not in active translation.

Arguably the most attractive hypothesis is that lncRNAs have functional roles in regulating translation. This could be a general phenomenon in which the lncRNAs occupy the ribosomes to keep them in a poised state and inhibit the energetically expensive process of translation until specific stimulatory cues are received. Alternatively, lncRNAs could regulate translation of specific protein-coding transcripts, for example by sequence-specific pairing. Indeed, recent data show that at least some lncRNAs associate with ribosomes to exert such a function [36]. For another class of noncoding RNAs, the microRNAs, similar roles have also been described [34]. One specific lncRNA, the antisense lncRNA of Uchl1, has been shown to regulate the association of sense Uchl1 with active polysomes in mice [36]. This regulatory function was partially established via the sequence homology between the lncRNA and the target mRNA. Translation regulatory mechanisms based on sequence homology have also been found for noncoding transcripts in bacteria [37]. Of the 25 antisense lncRNAs expressed in our data, only three pairs had both partners expressed and showed subcellular co-localization: ***DYNLL1***

and ***DYNLL1-AS1***, ***PCBP1*** and ***PCBP1-AS1,*** and ***WAC*** and ***WAC-AS1*** (Additional file 10). The fact that we found so few co-localizing sense-antisense pairs makes it unlikely that a similar mechanism is abundant in the human system studied here.

Conclusions

Our data show that different subcellular compartments differ significantly in RNA content, especially when the nucleus is compared to the ribosomal fractions. The lncRNAs expressed in this cell line are found in all subcellular samples and show an intricate correlation profile to protein-coding transcripts. Most lncRNAs are enriched in the cytosolic (free and the 40S/60S) samples and in the subcellular samples containing one, two or three ribosomes. The fact that lncRNAs show enrichment in diverse subcellular fractions and not only the nucleus suggests that lncRNAs may have a wider range of functions than currently anticipated. Our study provides insight into this diversity and our data can serve as a valuable resource for the functional characterization of individual lncRNAs.

Materials and methods

Cell culture and media. Human colon cancer cells carrying a doxycycline-inducible short hairpin RNA against B-catenin (LS-174T-pTER-*β*-catenin [23]) were cultured in 1X DMEM + GIBCO GlutaMAX™ (Life Technologies, Carlsbad, CA, USA) supplemented with 10% fetal calf serum and penicillin streptomycin. Cells were harvested during the exponential growth phase.

Ribosome fractionation. All steps of the mono- and polyribosome profiling protocol were performed at 4°C or on ice. Gradients of 17% to 50% sucrose (11 mL) in gradient buffer (110 mM KAc, 20 mM MgAc and 10 mM HEPES pH 7.6) were poured the evening before use. Three replicates of 15 cm dishes with LS-174T-pTER-*β*-catenin cells were lysed in polyribosome lysis buffer (110 mM KAc, 20 mM MgAc, 10 mM HEPES, pH 7.6, 100 mM KCl, 10 mM MgCl, 0.1% NP-40, freshly added 2 mM DTT and 40 U/mL RNasin (Promega, Madison, WI, USA)) with help of a Dounce tissue grinder (Wheaton Science Products, Millville, NJ, USA). Lysed samples were centrifuged at 1200 g for 10 min to remove debris and loaded onto the sucrose gradients. The gradients were ultra-centrifuged for 2 h at 120,565 g in an SW41 Ti rotor (Beckman Coulter, Indianapolis, IN, USA). The gradients were displaced into a UA6 absorbance reader (Teledyne ISCO, Lincoln, NE, USA) using a syringe pump (Brandel, Gaithersburg, MD, USA) containing 60% sucrose. Absorbance was recorded at an optical density of 254 nm. Fractions were collected using a Foxy Jr Fraction Collector (Teledyne ISCO). Corresponding fractions from each of the three replicates were merged prior to RNA isolation.

Nuclei isolation. Pelleted nuclei of LS-174T-pTER-*β*-catenin cells were obtained by

centrifugation at 1200 g after whole-cell lysis prior to ribosome fractionation (see previous section). To exclude the presence of rough endoplasmic reticulum and thus validate the purity of the isolated nuclei, nuclear staining and imaging were performed (Additional file 1).

RNA sequencing library preparation. Total RNA was isolated from purified nuclei using the TRIzol® reagent (#15596-026, Invitrogen, Life Technologies). RNA derived from triplicate mono- and polyribosome fractionation experiments was purified using TRIzol® LS reagent (#10296-028, Invitrogen, Life Technologies). Isolated RNA from the pooled triplicate fractions corresponded to the (A1 + 2) non-ribosome bound RNA, (B1) 40S subunit, (B2) 60S subunit, (C) 80S ribosome, (D) 2 ribosomes, (E) 3 ribosomes, (F) 4 ribosomes, (G) 5 ribosomes and (H) 6 ribosomes and (I) more than 6 ribosomes (Figure 1). For RNA-seq, RNA derived from A1 + 2 (non-ribosome bound RNA) and B1 + B2 (individual ribosomal subunits) was pooled prior to library preparation. RNA-seq libraries were prepared from rRNA-depleted RNA (Ribo-Zero™ Magnetic Gold Kit for Human/Mouse/Rat (MRZG12324, Epicentre®, Madison, WI, USA) using the SOLiD™ Total RNA-seq kit (#4445374, Life Technologies). All libraries were sequenced on the SOLiD™ 5500 Wildfire system (40 bp fragment reads).

Data analysis. RNA-seq reads were mapped using Burrows-Wheeler Aligner [38] (BWA-0.5.9) (settings: -c -l 25 -k 2 -n 10) onto the human reference genome hg19. Only uniquely mapped, non-duplicate reads were considered for further analyses. Reads that mapped to exons were used to determine the total read counts per gene. Exon positions were based on the GENCODE v18 annotation [24]. The polyribosomal samples (from two to seven or more associated ribosomes) yielded 13 to 32 million reads. For the non-polyribosomal samples (nuclear, free cytosolic, combined 40S and 60S, and 80S (monosomes)), data from three sequencing lanes (technical replicates) were merged yielding 6 to 64 million reads. Data analysis was performed on the genes with GENCODE gene_type: protein coding, antisense, processed transcript, long intergenic noncoding RNA and snRNA/snoRNAs. Filtering was performed on the read count per gene over all samples combined. The per transcript sum of the sequencing reads in all samples showed a bimodal distribution (Additional file 11). Based on these data we used a total read count threshold of 2,500 per transcript to select the expressed genes. Genes with total read count below 2,500 were filtered out, leaving 7,734 genes for further analysis. Subsequently, normalization was performed using the DEseq [39] to correct for library size and technical biases. Gene clustering was performed using a model-based clustering approach with the R package HTSCluster [25]. The protein coding-lncRNA correlation matrix (Figure 3) was calculated using Spearman rank correlation. The matrix was visualized after hierarchical clustering using Euclidean distance with complete linkage. Median transcript length and coding sequence length were calculated for the protein-coding genes using annotation from Ensembl. The maximum lncRNA ORFs were predicted using a custom Perl script aimed at finding reading frames with in-frame START and STOP codons,

without intervening in-frame STOP codons.

Quantitative PCR analysis. Quantitative PCR analysis was performed on cDNA derived from total RNA of cytosolic, nuclear and pooled polyribosomal RNA. The RT reaction was performed on 1 µg of total RNA using oligo d(T) primers and the high capacity cDNA reverse transcription kit (Life Technologies, #4368814). Three primer sets were designed per lncRNA. Quantitative PCR reactions were performed in 20 µl reactions using 2 ng of cDNA and iQ™ SYBR® Green Supermix (Bio-Rad, Hercules, CA, USA, #170-8880) on a MyIQ2 Real-time PCR detection system (Bio-Rad).

Abbreviations

bp, base pairs; CPM, counts per million; lncRNA, long noncoding RNA; NMD, nonsense mediated decay; ORF, open reading frame; PCR, polymerase chain reaction; RNA-seq, RNA-sequencing; rRNA, ribosomal RNA; RT, reverse transcription; sncRNA, small noncoding RNA; snoRNA, small nucleolar RNA; snRNA, small nuclear RNA.

Authors' contributions

SvH, MvI, EC and MS wrote the manuscript. MvI, MS, SB and JJ performed RNA-Seq data analyses. SvH performed cell culture, nuclei isolation and RNA-Seq experiments. SvH and PBE performed polysomal fractionation experiments. WH and EdB performed next generation sequencing. SvH, EC, MS and AWM designed the experiments. All authors contributed to scientific discussions, and read and approved the final version of the manuscript.

Acknowledgments

EC acknowledges funding from NWO TOP grant (700.58.303), the NGI-NBIC program and the NGI-NCSB program. MS acknowledges funding from the NWO Vernieuwingsimpuls program (grant number 863.10.007).

Data availability

All next generation sequencing data used in this study can be downloaded from EMBL European Nucleotide Archive [PRJEB5049].

Additional files

Additional files can be found online at http://genomebiology.com/2014/15/1/R6 (doi: 10.1186/gb-2014-15-1-r6)

References

1. Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N: Requirement for Xist in X chromosome inactivation. Nature 1996, 379:131–137.

2. Feng J, Funk WD, Wang SS, Weinrich SL, Avilion AA, Chiu CP, Adams RR, Chang E, Allsopp RC, Yu J, Le S, West MD, Harley CB, Andrews WH, Greider CW, Villeponteau B: The RNA component of human telomerase. Science 1995, 269:1236–1241.

3. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R: The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012, 22:1775–1789.

4. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES: Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009, 458:223–227.

5. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest ARR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, et al: The transcriptional landscape of the mammalian genome. Science 2005, 309:1559–1563.

6. Kowalczyk MS, Higgs DR, Gingeras TR: Molecular biology: RNA discrimination. Nature 2012, 482:310–311.

7. Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB: An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 2009, 33:717–726.

8. Tripathi V, Shen Z, Chakraborty A, Giri S, Freier SM, Wu X, Zhang Y, Gorospe M, Prasanth SG, Lal A, Prasanth KV: Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. PLoS Genet 2013, 9:e1003368.

9. Yoshimizu T, Miroglio A, Ripoche MA, Gabory A, Vernucci M, Riccio A, Colnot S, Godard C, Terris B, Jammes H, Dandolo L: The H19 locus acts in vivo as a tumor suppressor. Proc Natl Acad Sci USA 2008, 105:12417–12422.

10. Kretz M, Webster DE, Flockhart RJ, Lee CS, Zehnder A, Lopez-Pajares V, Qu K, Zheng GX, Chow J, Kim GE, Rinn JL, Chang HY, Siprashvili Z, Khavari PA: Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev 2012, 26:338–343.

11. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigó R, Shiekhattar R: Long noncoding RNAs with enhancer-like function in human cells. Cell 2010, 143:46–58.

12. Geisler S, Coller J: RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol 2013, 11:699–712.

13. Ulitsky I, Bartel DP: lincRNAs: genomics, evolution, and mechanisms. Cell 2013, 154:26–46.

14. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT Bar NS, Balut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, et al: Landscape of transcription in human cells. Nature 2012, 489:101–108.

15. Bhatt DM, Pandya-Jones A, Tong AJ, Barozzi I, Lissner MM, Natoli G, Black DL, Smale ST: Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell 2012, 150:279–290.

16. Ingolia NT, Lareau LF, Weissman JS: Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 2011, 147:789–802.

17. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES: Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell 2013, 154:240–251.

18. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS: Genome-wide annotation and quantitation of translation by ribosome profiling. Curr Protoc Mol Biol 2013, Chapter 4:Unit 4 18.

19. Lecocq RE, Cantraine F, Keyhani E, Claude A, Delcroix C, Dumont JE: Quantitative evaluation of polysomes and ribosomes by density gradient centrifugation and electron microscopy. Anal Biochem 1971, 43:71–79.

20. Arava Y, Boas FE, Brown PO, Herschlag D: Dissecting eukaryotic translation and its control by ribosome density mapping. Nucleic Acids Res 2005, 33:2421–2432.

21. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D: Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci USA 2003, 100:3889–3894.

22. Arribere JA, Doudna JA, Gilbert WV: Reconsidering movement of eukaryotic mRNAs between polysomes and P bodies. Mol Cell 2011, 44:745–758.

23. van de Wetering M, Oving I, Muncan V, Pon Fong MT, Brantjes H, van Leenen D, Holstege FC, Brummelkamp TR, Agami R, Clevers H: Specific inhibition of gene expression using a stably integrated, inducible small-interfering-RNA vector. EMBO Rep 2003, 4:609–615.

24. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, et al: GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012, 22:1760–1774.

25. Rau A, Celeux G, Martin-Magniette M, Maugis-Rabusseau C: Clustering high-throughput sequencing data with Poisson mixture models. Orsay: Inria; 2011. Technical Report RR-7786.

26. Sasaki YT, Ideue T, Sano M, Mituyama T, Hirose T: MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. Proc Natl Acad Sci USA 2009, 106:2525–2530.

27. Yang L, Lin C, Liu W, Zhang J, Ohgi KA, Grinstein JD, Dorrestein PC, Rosenfeld MG: ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. Cell 2011, 147:773–788.

28. Bartkiewicz M, Gold H, Altman S: Identification and characterization of an RNA molecule that copurifies with RNase P activity from HeLa cells. Genes Dev 1989, 3:488–499.

29. Ullu E, Murphy S, Melli M: Human 7SL RNA consists of a 140 nucleotide middle-repetitive sequence inserted in an alu sequence. Cell 1982, 29:195–202.

30. Walter P, Blobel G: Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. Nature 1982, 299:691–698.

31. Zhu L, Xu PC: Downregulated LncRNA-ANCR promotes osteoblast differentiation by targeting EZH2 and regulating Runx2 expression. Biochem Biophys Res Commun 2013, 432:612–617.

32. Berteaux N, Lottin S, Monte D, Pinte S, Quatannens B, Coll J, Hondermarck H, Curgy JJ, Dugimont T, Adriaenssens E: H19 mRNA-like noncoding RNA promotes breast cancer cell proliferation through positive control by E2F1. J Biol Chem 2005, 280:29625–29636.

33. Wilusz JE, JnBaptiste CK, Lu LY, Kuhn CD, Joshua-Tor L, Sharp PA: A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(A) tails. Genes Dev 2012, 26:2392–2407.

34. Thermann R, Hentze MW: Drosophila miR2 induces pseudo-polysomes and inhibits translation initiation. Nature 2007, 447:875–878.

35. Tani H, Torimura M, Akimitsu N: The RNA degradation pathway regulates the function of GAS5 a non-coding RNA in mammalian cells. PLoS One 2013, 8:e55684.

36. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, Forrest AR, Carninci P, Biffo S, Stupka E, Gustincich: Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature 2012, 491:454–457.

37. Darfeuille F, Unoson C, Vogel J, Wagner EG: An antisense RNA inhibits translation by competing with standby ribosomes. Mol Cell 2007, 26:381–392.

38. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754–1760.

39. Anders S, Huber W: Differential expression analysis for sequence count data. Genome Biol 2010, 11:R106.

6

Recurrent genome breakage in cancer and germline is marked by late replication timing

Sebastiaan van Heesch[1†], Marieke Simonis[1†], Markus J. van Roosmalen[2], Vamsee Pillalamarri[3], Harrison Brand[3], Kim L. de Luca[1], Ewart W. Kuijk[1], Nico Lansu[1], A. Koen Braat[4], Androniki Menelaou[2], Wensi Hao[1], Jeroen Korving[1], Simone Snijder[5], Lars T.J. van der Veken[2], Ron Hochstenbach[2], Alida C. Knegt[5], Karen Duran[2], Ivo Renkens[2], Evelien Kruisselbrink[4], Najla Alekozai[2], Myrthe Jager[2], Sarah Vergult[6], Björn Menten[6], Ewart de Bruijn[1], Sander Boymans[1], Elly Ippel[2], Ellen van Binsbergen[2], Michael E. Talkowski[3], Klaske Lichtenbelt[2], Edwin Cuppen[1,2] and Wigard P. Kloosterman[2]

†These authors contributed equally to this work

[1] Hubrecht Institute-KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands.
[2] Department of Medical Genetics, Institute for Molecular Medicine, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands.
[3] Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA.
[4] Department of Cell Biology, Institute for Molecular Medicine, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands.
[5] Department of Clinical Genetics, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands.
[6] Center for Medical Genetics, Ghent University Hospital, De Pintelaan 185, B-9000 Ghent, Belgium.

Abstract

Genomic rearrangements are a common cause of human congenital abnormalities. However, in the majority of cases a mechanism linking rearranged chromosomes to disease phenotype is lacking. Here, we studied the molecular effects of ***de novo*** structural variations in patients with congenital disorders. In one patient, a complex chromothripsis rearrangement resulted in gene fusions involving ***ETV1*** and ***FOXP1***, which are both involved in recurrent gene fusions caused by somatic breakpoints in cancer. In a second patient, a tandem duplication caused activation of the miRNA cluster C19MC, which is highly upregulated as a consequence of genomic rearrangements in thyroid adenomas and hepatocellular carcinoma. We show that expression of C19MC miRNAs results in severe defects in brain morphogenesis during embryonic development. Driven by this striking resemblance between cancer and constitutional genomic rearrangements, we analyzed a large set of 550 high-resolution genomic breakpoints involved in congenital disorders, which are mainly (73%) derived from complex genomic rearrangements. We observed that constitutional breakpoints can target cancer genes and overlap with breakpoints from cancer rearrangements. The overlapping breakpoints do not associate with common fragile sites, but coincide with late-replicating regions, which are prone to replication stress. These findings indicate the presence of recurrent genomic breakpoints involved in cancer and congenital disease, which may result in comparable effects on gene function.

Introduction

Constitutional genomic rearrangements are a common cause of congenital disease, including mental retardation, neurodevelopmental delay and a broad spectrum of morphological malformations [1, 2]. Constitutional rearrangements can be classified in two major categories. One category arises through non-allelic homologous recombination via genomic repeats [1-4] primarily resulting in copy number changes (CNVs). Most of these CNVs are recurrent and give rise to recognizable phenotypes known as microdeletion and microduplication syndromes, which can result from dosage effects of one or more genes within the CNV interval [5]. The second category contains sporadic (non-recurrent) genomic rearrangements and comprises more diverse rearrangement types including CNVs, translocations, inversions and complex events. These rearrangements are primarily caused by non-homologous modes of DNA repair, such as direct end-joining of free DNA ends [6] or template-switching following replication fork stalling [7]. Also, ultra-complex rearrangements resulting from the shattering of chromosomes in a single event, termed chromothripsis, arise through non-homologous DNA repair [8-10].

For the majority of patients with non-recurrent (sporadic) rearrangements the actual cause of disease remains unclear. The uniqueness of the breakpoints of such rearrangements and the multiple possible effects on gene function make it difficult to establish a direct causal

relationship with a congenital phenotype [11]. This is even more difficult for complex genomic rearrangements, such as those caused by chromothripsis or chromoanasynthesis, where one or several breakpoints could cause disease [8, 9, 12]. A first step towards understanding disease mechanisms resulting from chromosome rearrangements is the detection of breakpoints at nucleotide resolution, including information on orientation of breakpoint junctions. We and others have used genome-wide paired-end sequencing to pinpoint genomic rearrangements in patients with congenital disease, revealing insight into underlying gene defects [8, 9, 13-15].

Here, we gained deep insight into the effects of ***de novo*** genomic rearrangements in patients with congenital disease by performing integrated transcriptome, small RNA and ChIP profiling in patient blood. We observed that ***de novo*** constitutional breakpoints occasionally target cancer genes. In one patient, a constitutional chromothripsis rearrangement caused the formation of novel fusion genes involving the cancer genes ***ETV1*** and ***FOXP1***. In a second case, we identified strong upregulation of ~50 oncogenic C19MC microRNAs through a ***de novo*** tandem duplication. Besides a known role for these microRNAs as drivers of several cancers, including embryonal brain tumors [16], we here report that their overexpression can drive brain morphogenesis defects in zebrafish as well. Furthermore, a comprehensive analysis of 550 constitutional breakpoints identified in patients with congenital disease revealed additional breakpoints in cancer genes, significantly more than would be expected by chance. Also, we show that ~10% of the constitutional breakpoints resides in close vicinity to somatic breakpoints identified in cancer genomes. These overlapping breaks do not associate with known common fragile sites (CFSs), but are specifically marked by late replication timing, suggesting that replicative mechanisms of genome rearrangement may govern their formation in cancer and congenital disease.

Results

Family-based transcriptome analysis reveals fusion transcripts involving the cancer genes ETV1 and FOXP1 as a result of constitutional chromothripsis

We employed an ***in vivo*** family-based molecular profiling approach to characterize the effects of ***de novo*** structural genomic rearrangements in two independent patients with multiple congenital abnormalities and intellectual disability (MCA/ID), both their parents, and one unaffected sibling (Supplementary Table 1).

In one patient with speech delay, psychomotor retardation, dysmorphic facial appearance and doubling of one of the thumbs we identified a ***de novo*** chromothripsis rearrangement. This constitutional chromothripsis rearrangement involves 17 breakpoints divided over four chromosomes (1,3,7 and 12) (Fig. 1A, Supplementary Fig. 1A, Supplementary Table 2) [8].

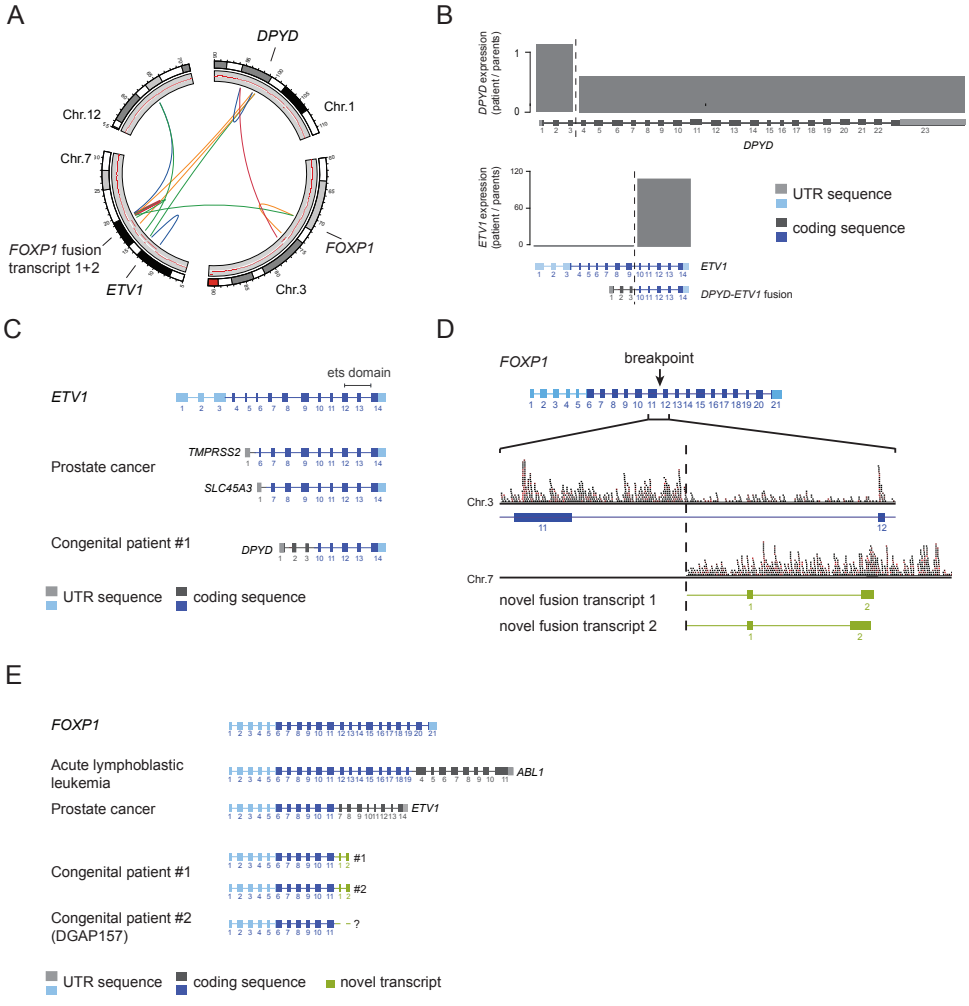


Figure 1 - **A de novo chromothripsis rearrangement results in *ETV1* and *FOXP1* fusion transcripts**. (**A**) Circos plot of the 13 breakpoint junctions forming the chromothripsis rearrangement. The outer circle displays the chromosome ideogram. The inner circle represents the copy number profile as based on read-depth measurements relative to the parents. The colored lines indicate breakpoint junctions. Blue, tail-to-head; green, head-to-tail; red, head-to-head inverted; yellow, tail-to-tail inverted. The locations of relevant genes are indicated. Chromosome coordinates are in Mb. (**B**) Visualization of the *DPYD-ETV1* fusion gene and the transcriptional consequences thereof. RNA-seq reads within the genomic intervals from the start of the genes to the breakpoint and from the breakpoint to the end of the gene were normalized for the total amount of reads per sample. The plot visualizes the ratios of normalized reads in the patient versus the average of the parents. (**C**) Diagram showing the full-length *ETV1* gene, examples of *ETV1* fusion genes as observed in cancer [19, 22] and the *DPYD-ETV1* fusion in our patient. (**D**) Raw RNA-Seq reads depicting *FOXP1* transcription across the chromosome 3 - chromosome 7 breakpoint, resulting in two novel fusion transcripts (green). (**E**) Diagram showing the full-length *FOXP1* gene, examples of *FOXP1* fusion genes as observed in cancer and the *FOXP1* breakpoints in two patients in our dataset [9, 22, 23].

To study the effects of chromothripsis on gene-expression we performed RNA-seq on peripheral blood mononuclear cells (PBMCs) of this patient and both parents. First, we examined the expression levels of 11 genes that reside within three large ***de novo*** genomic deletions caused by the chromothripsis. Four of these genes were expressed in PBMCs and showed a clear decrease in expression levels relative to the parents (Supplementary Fig. 1B). In addition to these deleted genes, six genomic breakpoints were located within a gene, thus splitting up the coding sequence (Supplementary Table 2). Of the six genes disrupted by breakpoints, three are transcriptionally active in PBMCs. Two of them (***DPYD*** and ***FOXP1***) showed a decrease in expression following the breakpoint (Fig. 1B and Supplementary Fig. 1C). In contrast, for the third gene (***ETV1***) the C-terminal part showed elevated expression in the patient relative to the parents, while the N-terminal part was not expressed. Examination of the breakpoint junctions involving ***ETV1*** revealed a genomic fusion between the first three exons of the ***DPYD*** gene and exons 10-14 of the ***ETV1*** gene (Fig. 1B). ***DPYD*** encodes for dihydropyrimidine dehydrogenase, an essential factor for uracil and thymidine catabolism that is ubiquitously expressed. As a result, the genomic fusion between ***DPYD*** and ***ETV1*** leads to high expression of the 3' part of ***ETV1*** in patient blood, while the parents do not express ***ETV1***.

ETV1 is a member of the ETS (E-twenty six) family of transcription factors that modulate target genes involved in various biological processes, such as cell differentiation, proliferation, migration and apoptosis [17]. ***ETV1*** gene fusions are frequently found in Ewing sarcoma and prostate cancer, but have not been described as drivers of congenital disease [18-22]. Remarkably, the topology of the ***DPYD-ETV1*** fusion in this patient resembles that of the ***ETV1*** fusion genes found in cancer (Fig. 1C). Both in this patient and in cancer the 3' part of ***ETV1***, containing the ETS transcription activation domain, becomes ectopically expressed by fusion to the 5' part of an actively transcribed gene [20]. We constructed a cDNA gene mimicking the ***DPYD-ETV1*** fusion and overexpressed it in HEK293 cells to determine functionality of the protein. Although we can detect sporadic protein product in these cells (~1/50, Supplementary Fig. 1D), we could not detect a stable DPYD-ETV1 fusion protein on western blot analysis, suggesting that this product is only stable and/or translated under specific conditions.

Next, we studied the transcriptional consequences of the breakpoint in ***FOXP1***, a genomic region that is also frequently involved in translocations in cancer [22, 23]. RNA-seq analysis showed read-through transcription from exon 11 of the ***FOXP1*** gene to a genomic segment on chromosome 7 (Fig. 1D). No annotated coding gene was present as 3' fusion partner of ***FOXP1***, but cDNA analysis of the read-through transcripts showed two differentially spliced novel transcripts fused to the 11th exon of ***FOXP1*** (Fig. 1D and Supplementary Fig. 1E). These novel fusion transcripts resemble ***FOXP1*** cancer gene fusions (Fig. 1E), with the intron targeted for translocation being identical to the breakpoint in many ***FOXP1*** gene fusions in

cancer. Furthermore, we identified a second patient with a constitutional breakpoint in the same intron in ***FOXP1***, indicating that this is a recurrently rearranged region in both cancer and germline [24]. The two transcript isoforms identified in the patient with chromothripsis add 24 and 46 amino acids respectively to the ***FOXP1*** open reading frame. Upon expression in HEK293 cells both transcripts result in stable protein products (Supplementary Fig. 1F and 1G).

The contribution of both the ***ETV1*** and ***FOXP1*** fusions to the patient's phenotype is difficult to assess from the current data. ***FOXP1*** is often hit by breakpoints in cancer and has been associated with neurodevelopmental disorders by ***de novo*** translocation breakpoints, copy number variants, and point mutations [24, 25]. It is very well possible that the loss of one functional allele and not to the ubiquitous gain of FOXP1 or ETV1 expression drives the patient's phenotype. Our data demonstrate that (novel) spliced transcripts resulting in stable proteins that can be formed through constitutional chromothripsis breakpoints. These mechanisms of gene activation are very similar to those of somatic rearrangements in cancer genomes. Furthermore, the results provide a first insight into the molecular effects of constitutional chromothripsis rearrangements and show that chromosome shattering can lead to transcriptional activation in addition to gene disruption.

A *de novo* congenital duplication activates a cluster of oncogenic microRNAs

The second patient with congenital defects analyzed using molecular phenotyping carries a ***de novo*** 424.5-kb tandem duplication on chromosome 19, resulting in macrocephaly and severe psychomotor retardation (Supplementary Fig. 2A, Supplementary Table 1). The most predictable effect of a genomic duplication is elevated gene expression due to an increase in gene copy number. Indeed, RNA-seq analysis performed on peripheral blood mononuclear cells (PBMCs) demonstrates that three duplicated genes are expressed significantly higher in the patient compared to both the parents and an unaffected sibling (Supplementary Fig. 2B). Closer examination of the breakpoints of the duplication revealed unexpected additional molecular effects. The 5' breakpoint of the duplicated region is located within the chromosome 19 microRNA cluster (C19MC) and the 3' breakpoint disrupts the ***NDUFA3*** gene. The tandem duplication repositioned a major part of the C19MC miRNA cluster immediately downstream of the promoter of ***NDUFA3***. This prompted us to investigate the presence of active promoter elements in the rearranged locus by H3K4me3 chromatin immunoprecipitation sequencing (ChIP-seq). The ***NDUFA3*** promoter was found to have high H3K4me3 levels in all samples and this H3K4me3 signal was found to extend into the C19MC cluster downstream of the 5' duplication breakpoint in the patient (Fig. 2A). In addition, small RNA-seq revealed that the C19MC miRNAs positioned downstream of the ***NDUFA3*** promoter were highly expressed (Fig. 2B), whereas they are non-expressed in both parental samples and the unaffected

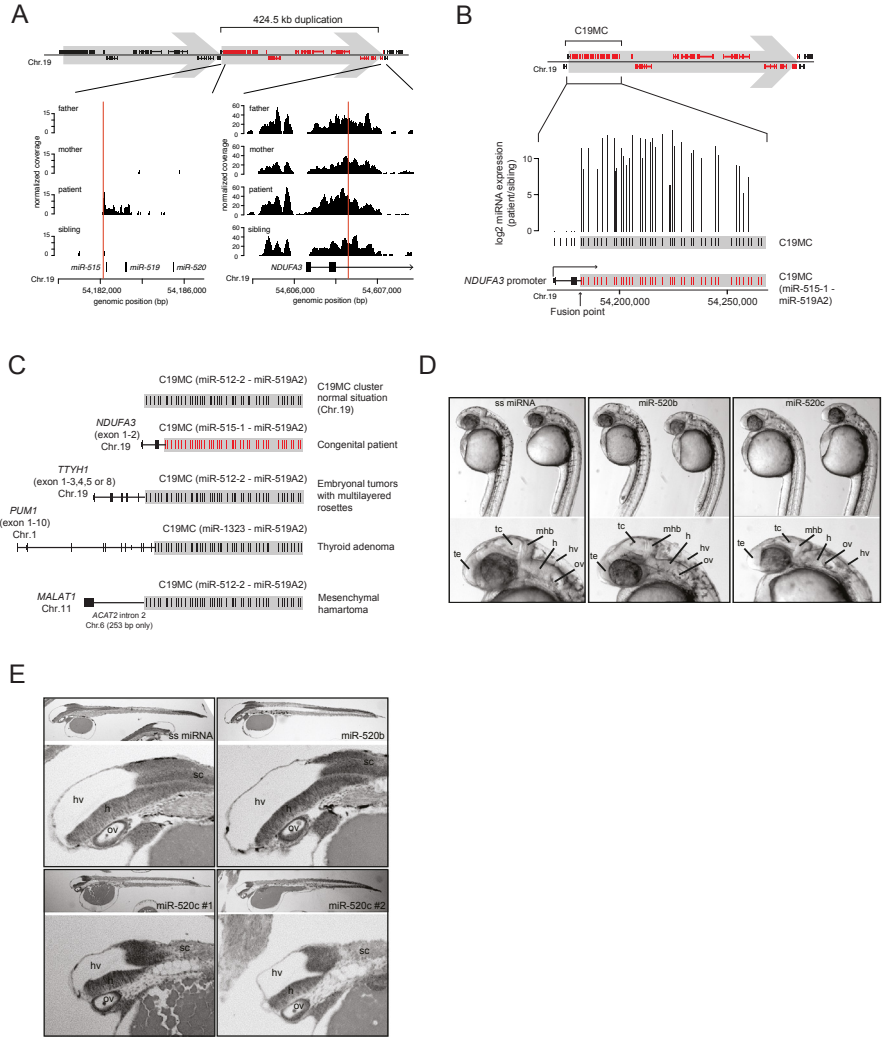


Figure 2 - **A *de novo* 424.5-kb tandem duplication activates C19MC expression.** (**A**) H3K4me3 ChIP-seq results for the promoter of the *NDUFA3* gene in the father, mother, patient and healthy sibling. The upper panel shows a schematic representation of the duplication. The grey arrows show which part of the chromosome is duplicated and (parts of) genes within the duplication are depicted in red. The lower left panel shows H3K4me3 signals for a region surrounding the 5' breakpoint of the tandem duplication. The lower right panel shows H3K4me3 signals for the *NDUFA3* promoter region. The vertical red lines indicate the position of the duplication breakpoints. (**B**) Normalized log2 expression ratios of microRNAs in C19MC for the patient versus healthy sibling (control). The duplicated fraction of C19MC is colored red. An arrow depicts the breakpoint junction that fuses exon 2 of *NDUFA3* to C19MC. (**C**) Examples of chromosomal rearrangements activating C19MC in cancer. The constitutional rearrangement activating C19MC is depicted followed by three previously described rearrangements in embryonal brain tumors [16], thyroid adenoma [29] and mesenchymal hamartoma [30]. (**D**) Whole mount bright-field images of 24 hours post fertilization (hpf) zebrafish embryos derived from control injections (single stranded miRNAs) and injections with miR-520b and miR-520c duplexes. Zoomed views of the zebrafish are displayed in the lower panel, with annotation of the hindbrain (h), hindbrain ventricle (hv), otic vesicle (ov), telencephalon (te), midbrain-hindbrain boundary (mhb), and tectum (tc). (**E**) Sagittal sections of embryos derived from the same experiments as under (a-b), with two examples of miR-520c injected embryos. Hindbrain (h), hindbrain ventricle (hv), otic vesicle (ov) and spinal cord (sc) are annotated.

sibling. The part of the C19MC cluster that is not repositioned by the duplication was not expressed in the patient (Fig. 2B) and miRNAs elsewhere on the genome were also unaffected (Supplementary Fig. 2C). The miRNA encoded by the ***MIR371*** gene, which is also located in the duplication but is driven by its own promoter, also shows no upregulation (Supplementary Fig. 2C). Endogenous expression of C19MC is exclusive to embryonic stem cells and tumors, which suggests that normal differentiation and development could be disturbed upon ectopic expression of this cluster [26, 27]. The ***NDUFA3*** gene, which encodes a subunit of a mitochondrial protein complex, is broadly expressed and therefore expected to drive C19MC miRNA expression in many tissues in the patient.

Expression of C19MC microRNAs drives cancer and defects in embryonic development

Genomic rearrangements in the 150-kb common breakpoint cluster on the long arm of chromosome 19 are known to affect C19MC expression in thyroid adenomas, epithelial tumors and embryonal brain tumors [16, 28, 29]. Previous reports have shown aberrant expression of part of the C19MC cluster in cancer resulting from the repositioning of an active promoter [16, 29, 30] (Fig. 2C). Other studies have shown that expression of the C19MC cluster in cancer cells is an important driver of tumorigenesis, tumor invasion and metastasis [31, 32], with eight C19MC members directly targeting ***p21*** (***CDKN1A***) and C19MC being a transcriptional target of TP53 [26, 33, 34].

We selected two of these cancer related miRNAs, miR-520b and miR-520c, to study the effects of overexpression of the mature miRNA duplex on zebrafish embryonic development (Supplementary Fig. 2D-F) [35]. The miRNAs were selected based on homology with zebrafish miRNAs, presence of the same miRNA seed sequence among several other C19MC miRNAs, oncogenic potential as shown by previous studies [31, 32] and ***de novo*** expression in the patient. Injection of single-stranded miRNA controls and miR-520b duplexes in the 1-cell stage embryo did not result in a noticeable phenotype at 24 hours post fertilization (the miRNA is stable up to ~ 30h after injection). However, for miR-520c, we detected specific developmental malformations (Fig. 2D). Particularly the head of miR-520c injected embryos is smaller, displays a reduced fore- and hindbrain ventricle and an altered morphology of the midbrain-hindbrain boundary (Fig. 2E). These results demonstrate that overexpression of specific oncogenic C19MC miRNAs can disturb normal embryonic development, and are therefore likely to have contributed to the neurodevelopmental defects in the patient.

Constitutional structural variation breakpoints from patients with congenital disease overlap cancer genes

Triggered by the two cases described above we hypothesized that cancer fusion genes or other

oncogenic changes could more often be associated with congenital disease. To pursue this concept we systematically analyzed a larger set of chromosomal rearrangements identified in 96 sporadic patients with congenital phenotypes, including neurodevelopmental delay, mental retardation and morphological abnormalities. Breakpoint positions and junctions from these rearrangements were obtained from published and in-house resources [8, 9, 14, 24, 36-38]. The dataset comprises 550 genomic breakpoints identified using whole-genome read pair sequencing and/or breakpoint cloning. Most breakpoints (83%) are defined at the individual nucleotide level and therefore this dataset represents the largest collection of congenital breakpoints at this resolution published to date. For all breakpoints in this set, break-repair occurred through non-homologous mechanisms such as non-homologous end joining (NHEJ) or microhomology-mediated mechanisms [4, 6]. Of the 550 breakpoints, 402 breakpoints in 33 patients arose from complex genome rearrangements, including those resulting from chromothripsis or multiple template-switching events [8, 9, 12]. Furthermore, the set contains 114 translocation breakpoints, 16 deletion breakpoints, 16 tandem duplication breakpoints and 2 inversion breakpoints.

We performed a systematic survey of the overlap of all 550 constitutional breakpoints with genes targeted by translocation breakpoints in cancer as listed in the Cancer Gene Census (CGC) database. This revealed 12 overlapping breakpoints (Supplementary Table 3), which is a small but significant increase compared to what would be expected based on matched randomly simulated breakpoint sets (Fig. 3A, permutation test, $p = 0.0114$, Methods) [39]. This enrichment for cancer genes does not result from a general enrichment for protein-coding genes, because in line with the genome-wide gene density, only 43% of the constitutional breakpoints overlap a protein-coding gene (Supplementary Fig. 3A). The overlapping CGC genes include the ***FOXP1 and ETV1*** breakpoints described above, but for example also a reciprocal fusion involving ***TNS3*** and ***FGFR1***. Recently, transforming activity of recurrent FGFR1 fusions was reported for glioblastoma, where activity of the FGFR1 kinase domain at the mitotic spindle leads to aneuploidy [40].

Constitutional breakpoints overlap with somatic breakpoints in cancer genomes

Because of the enrichment of cancer genes from the Cancer Gene Census among genes hit by constitutional rearrangement breakpoints, we hypothesized that regions in the genome prone to constitutional breaks in patients with congenital disease could also be prone to somatic breaks in cancer. To test this, we collected a large set of somatic genomic breakpoints from cancer genomes based on published resources. The dataset contains 68,018 genomic breakpoints derived from 19 cancer types, including breast (46%), prostate (18%), lung (8%), ovarian (5%) and colorectal cancer (3%). The breakpoints in this set are all defined at nucleotide-resolution.

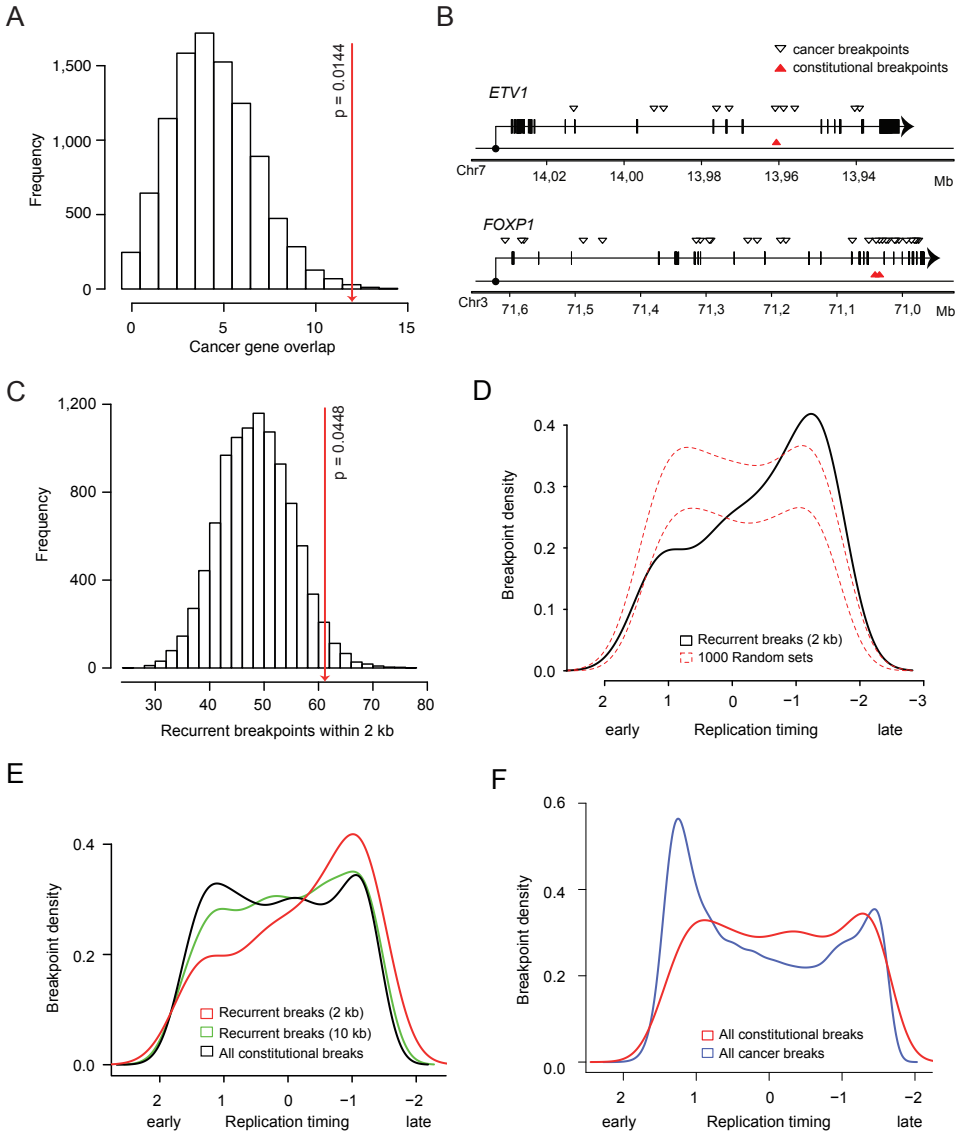


Figure 3 - **Cancer and congenital breakpoints overlap and associate with late replication timing**. (**A**) Overlap of congenital breakpoints with recurrently translocated genes from the Cancer Gene Census (CGC) database. The overlap was compared with 10,000 random control sets equal in size to the congenital breakpoint set. (**B**) Positions of cancer (white triangles) and constitutional breakpoints (red triangles) in the *ETV1* and *FOXP1* genes, respectively. (**C**) Permutation testing to determine the number of breakpoints that overlap between the cancer and constitutional breakpoint set. The empirical p-value was calculated using a maximum distance between overlapping breakpoints of 2,000 bp. (**D**) Replication timing analysis of all recurrent breakpoints (within 2kb) as compared to random sets of simulated breakpoints. The area between de dotted red lines represents the mean replication timing +/- 1 SD computed over 1000 simulation sets. (**E**) Replication timing analysis of all overlapping breakpoints with 2 kb (red) and 10 kb (green) inter-breakpoint distances. The black line shows the replication distribution of all constitutional breakpoints. For each breakpoint, 100-bp flanks were analyzed for replication timing scores. If no score within this window was available, the score closest to the breakpoint was used. (**F**) Similar to (E), but depicting all cancer (blue) and constitutional (red) breakpoints.

We next calculated the overlap between the 550 constitutional breakpoints and the 68,018 cancer breakpoints and found that a total of 61 (out of 550) congenital breaks in 34 patients overlapped a cancer break within a window of 2 kb. These 61 constitutional breakpoints include 43 breakpoints from complex rearrangements, such as the breakpoints in ***ETV1*** and ***FOXP1*** (Fig. 3B). However, they also include 15 translocation and 3 deletion breakpoints, resulting in a rearrangement type distribution that matches that of the complete dataset. This suggests that there is no specific type of rearrangement that preferentially contributes to the overlap with cancer breakpoints. Hypothetically, the highly localized clustering of chromothripsis breakpoints could contribute to the high degree of overlapping breakpoints we observe between both disease types. However, for only two out of 61 pairs of overlapping constitutional and cancer breakpoints a second pair was found within the same cancer and congenital patient samples. This indicates that clustering of chromothripsis breaks is likely not a cause of the overlapping breakpoint pairs. The overlap between the cancer and congenital breakpoint sets is higher than expected by chance as determined based on matched randomly simulated datasets using a maximum inter-breakpoint distance of 2 kb (permutation testing, $p = 0.0448$) (Fig. 3C). Also, the distribution of constitutional breakpoints positioned within 10 kb of a cancer breakpoints shows that in general these breakpoints are located closer to cancer breakpoints than simulated breakpoints (Supplementary Fig. 3B).

Altogether, these data provide evidence for mild though significant recurrence in genomic breakpoint positions between constitutional and somatic rearrangements.

Overlapping breakpoints are a specific subset of breakpoints enriched in late replicating regions

Sites of genome rearrangements in cancer are known to be associated with specific DNA sequence characteristics, including GC content and DNA secondary structure [41]. In addition, late-replicating common fragile sites (CFSs) were identified in cancer [42], and early replicating fragile sites (ERFSs) have been shown to underlie local genome fragility in B-cell lymphoma [43].

We determined what genomic features are associated with the overlapping breakpoints, as compared to genomic background sites and relative to the complete set of cancer and constitutional breakpoints. To do so, we used the above-described set of 61 breakpoints that show overlap with cancer breakpoints (i.e. with a maximum distance of 2 kb). This revealed that recurrent breakpoint locations are not significantly associated with the presence of retrotransposons (LINEs, SINEs, LTRs), G-quadruplexes (G4s), CpG islands, common fragile sites ((aphidicolin-induced) CFSs), segmental duplications and early-replicating fragile sites (ERFSs) [43] (Supplementary Fig. 3C). However, we did find that the recurrent breakpoints are specifically enriched in late-replicating genomic regions (Kruskal-Wallis test, $p = 0.02721$) (Fig. 3D). The late replication nature of the recurrent breakpoints decreases for pairs of breakpoints that have a larger inter-breakpoint distance (Fig. 3E). Also, the replication timing

distribution of the recurrent breakpoints strongly deviates from the distributions observed for the complete sets of constitutional and cancer breakpoints (Fig. 3F). Late replication timing is characteristic of common fragile sites, but we find no significant overlap with CFSs. Also, as expected based on the late-replicating nature of the overlapping rearrangements, no significant association was observed with ERFSs (Supplementary Fig. 3C).

Discussion

In this study, we describe two patients with congenital disease caused by ***de novo*** constitutional breakpoints. Molecular analysis of these patients revealed breakpoints in cancer genes and the formation of fusion genes similar to those described in cancer. Extension of this observation to a large set of 550 constitutional breakpoints identified additional breakpoints that overlap with cancer genes. This overlap cannot be fully explained by chance alone, suggesting a propensity of cancer genes to breakage in the germline. For example, ***FOXP1*** and ***RUNX1*** harbored breaks in two independent congenital patients. Intersection of the 550 constitutional breakpoints with a large set of 68,018 somatic breakpoints from cancer genomes further substantiated the overlap in breakpoint positions. Our findings imply a mechanistic link between both disease types, which is reflected in the two patients phenotyped at the molecular level. We propose a model whereby endogenous or exogenous stimuli may cause breakage at the same genomic positions in the germline and in cancer cells. The timing (germline or soma) and context (additional mutations) of the breakage determines whether the breaks promote developmental defects or cancer (Figure 4).

Recurrence driven by sensitivity to genomic breakage

The existence of sites in the genome that are prone to rearrangements has long been recognized, especially in cancer genomes. Common fragile sites (CFSs) are replication stress-induced and associated with genomic properties such as large genes [42, 44], late-replicating regions [45], AT-rich segments and condensed chromatin [45, 46]. In contrast to CFSs, early replication fragile sites (ERFSs) are gene-rich, GC-rich and euchromatic [43]. We find that recurrent breakpoints are specifically associated with late replication timing, but do not coincide with known CFSs or any of the other tested genomic features. We also did not observe an association of the recurrent breakpoints with the recently described ERFSs [43], as identified in B-cell lymphomas in mouse. In general, constitutional breakpoints that reside in early-replicating regions appear limited in number compared to the complete cancer dataset (Fig. 3F). This suggests that ERFSs are not likely to contribute to genome instability in congenital disease.

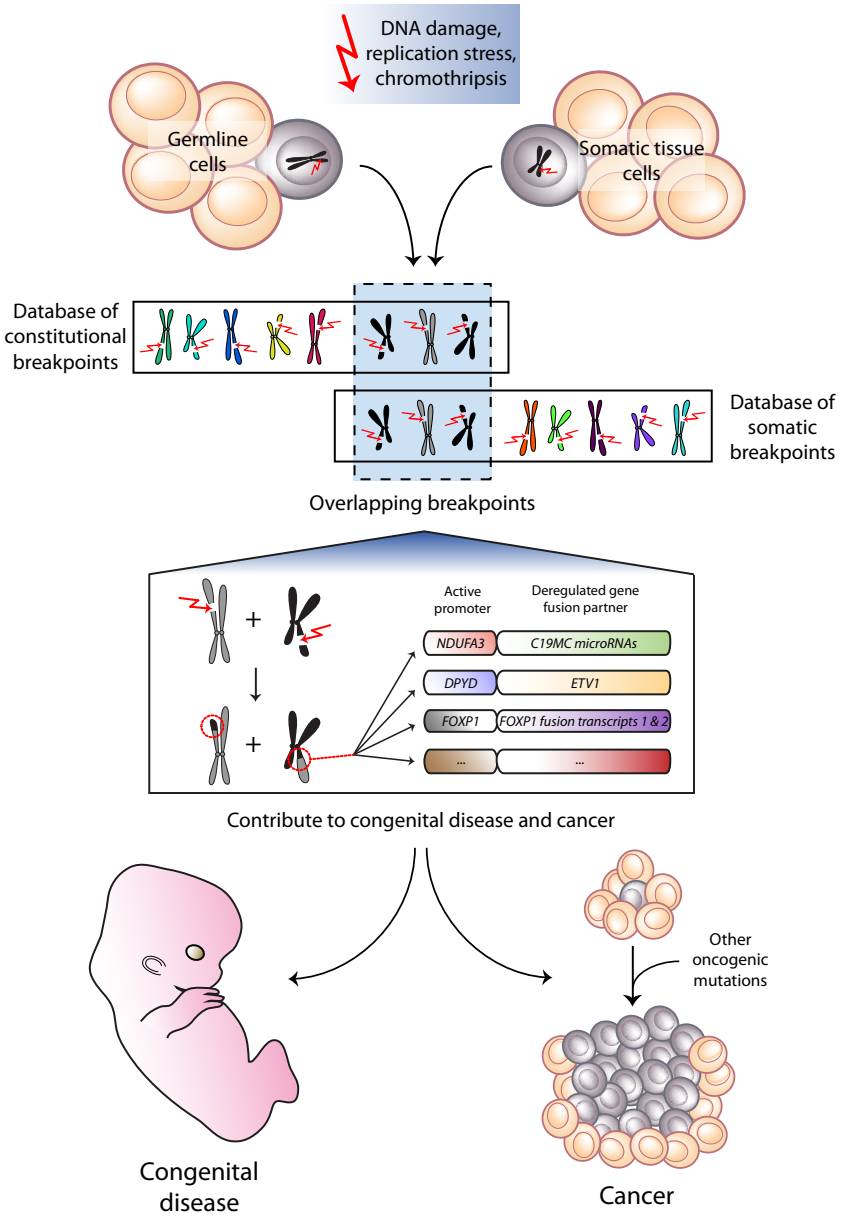


Figure 4 - **Schematic overview of the origin and consequences of recurrent breakpoints between the germline and cancer**. The breakpoint datasets collected and analyzed in this study originate in either the germline (or very early development), resulting in congenital disease, or in somatic tissue cells, leading to cancer. The data presented here demonstrate the presence of overlapping breakpoints, which might underlie both types of disease. These recurrent breakpoints originate specifically in late replicating regions. The consequences of these breakpoints can be highly similar in cancer and congenital disease, such as aberrant transcriptional activation of disease-promoting genes, for example via formation of gene fusions, of which a few examples are depicted as described in this paper. Depending on cellular context, timing and presence of mutations, the breakpoints result in either congenital disease or cancer.

Recurrence driven by function

Apart from local sensitivity to chromosome breakage and rearrangement formation, the biological function of affected genes may also explain the observed overlap between breakpoints in cancer and congenital disease. Deregulated oncogenes can perturb developmental programs, thereby leading to congenital disease. For example, Noonan syndrome is caused by deregulated RAS/MAPK signaling due to germline mutations in several pathway members [47]. Also, germline and postzygotic mutations of the PI3K pathway have been associated with megalencephaly syndromes [48].

Here, we highlight the activation of known oncogenic miRNAs in a patient with congenital disease and we show that overexpression of the miRNAs results in brain morphogenesis defects in zebrafish, pinpointing a phenotypic effect of this ***de novo*** genomic rearrangement. Thus, it is possible that gene function contributes to the recurrent breakpoints at these locations. However, only 24/61 recurrent breakpoints with an inter-breakpoint distance of less than 2kb overlap genes, so functional selection does not appear to be the only determinant and, even more, the occurrence of recurrent breaks due to genomic features or functional selection are not mutually exclusive.

Pediatric cancer and congenital disease

The observation of a genomic parallel between cancer and constitutional rearrangements raises the question of cancer predisposition among individuals carrying constitutional chromosome rearrangements. Although the two patients described here (chromothripsis case aged 25 and tandem duplication case aged 8) do not suffer from cancer at this point, we cannot rule out a predisposition for developing cancer later in life. The fusion partner that activates the expression of the oncogene in specific cells may be an important determinant for cancer susceptibility. In the case of the C19MC activation, of which somatic rearrangements were recently shown to drive embryonal tumors with multilayered rosettes (ETMR) [16], the fusion partner is different from the one that drives ETMR formation (***TTYH1***). Also, 5 microRNAs in the beginning of the C19MC cluster are not activated in the patient described here, but are activated in ETMRs. Likely, the identified oncogenic rearrangements may require additional driver mutations for cancer development (Figure 4), which are not known to be present in these patients described. In contrast, constitutional rearrangements that do directly result in cancer formation have been described. For example, rearrangements of the ***MYCN*** locus have been found to underlie childhood neuroblastoma [49], germline rearrangement of ***RUNX1*** caused acute myeloid leukemia [50] and 7q22 rearrangements are associated with myeloproliferative disorder [51]. This is further illustrated by two patients within our dataset who contain rearrangements of the ***RUNX1*** gene. Both suffered from AML possibly as a result of the constitutional ***RUNX1*** rearrangement. The high incidence of

morphological abnormalities among patients with childhood cancer further underscores a potential genetic and mechanistic link between cancer and congenital disease, which could partly be driven by recurrent genomic breakage as outlined here [52].

In summary, through a combination of genomic, transcriptomic, epigenomic and functional studies, we find a genomic and functional overlap between somatic genome rearrangements underlying cancer and constitutional rearrangements driving congenital disease. The recurrent breaks show a strong association with late replicating regions and we demonstrate that these breaks can induce similar molecular and functional effects. We show that a molecular approach to phenotype patients can be valuable for diagnostics and for predicting the molecular causes of congenital disease. For example, the existence of two novel FOXP1 products in the chromothripsis patient and the enormous increase in C19MC miRNA expression in the patient with a tandem duplication could not have been predicted accurately based on genomic information alone. Our results set the stage for further efforts to characterize genome rearrangement mechanisms in human development and disease and show that applying multiple sequencing approaches (including (small) RNA-seq and ChIP-seq) to analyze the ***in vivo*** molecular phenotypes provides novel insights in congenital disease etiology

Materials and methods

Patient Material and informed consent. We obtained informed consent for the analysis of DNA and RNA from each patient and their parents. The genetic analysis was performed according to the guidelines of the Medical Ethics Committee of the University Medical Center Utrecht.

Isolation of PBMCs from fresh blood. Approximately 30 mL of fresh blood was obtained from each individual of the two families of the patients (family 1: chromothripsis rearrangement, family 2: chr.19 duplication) (Supplementary Table 1). The blood was diluted 4x with PBS. PBMCs were isolated using 13 mL Histopaque®-1077 (family 1; Sigma-Aldrich 10771-500ML) or Ficoll-Paque™ PLUS (family 2; GE Healthcare, 17-1440-02) per 35 mL of diluted blood. After centrifugation at RT (20°C), for 20 minutes at 2,000 rpm (no brake), blood plasma was discarded and the PBMC layer recovered. PBMCs were washed twice using 12 mL of PBS, centrifuged at 1800 rpm for 5 minutes and collected in 1 mL PBS.

RNA sequencing and analysis. Total RNA was isolated from ±5M PBMCs using TRIzol (Life Technologies, 15596-018) and subsequent isopropanol precipitation (1:1). RNA concentrations were measured using a Qubit RNA assay (Invitrogen™, Q32852) and RNA was checked for quality using a Bioanalyzer 2100 RNA 6000 nano assay (Agilent Technologies,

5067-1511). For family 1, RNA was purified prior to library preparation using the RiboMinus™ Eukaryote Kit for RNA-Seq (Life Technologies, A10837-08). For family 2, RNA was purified using the Ambion® Poly(A)Purist™ (Life Technologies, AM1916) and mRNA only™ Eukaryote mRNA Isolation Kit (Epicentre®). Whole transcriptome library preparation was done for both families using the SOLiD® Total RNA-Seq Kit (Life Technologies, 4445374), exactly according to manufacturers instructions (Life Technologies protocol, 4452437 Rev. B). Following templated bead preparation, paired-end libraries of family 1 were sequenced in a multiplexed manner on the SOLiD V4 platform. Fragment libraries of family 2 were sequenced in a multiplexed manner on SOLiD 5500xl. RNA sequencing reads from SOLiD were mapped using BWA (settings: -c -l 25 -k 2 -n 10). The number of reads mapping to the total gene sequence (family 1) or the coding gene sequence (family 2) was used as a measure for expression (gene annotation Ensembl 67). The difference in the analyses of the families was chosen because of the different RNA purification procedures that were used. Per gene the number of reads was normalized to the total number of reads mapping within coding sequences. Gene expression changes were identified using DEGSeq v1.10.0 [53]. Construction of the schematic representation of fusion genes is based on the gene annotation in Ensembl 67. Transcripts with name 001 were selected for the genes that have multiple known transcripts.

Small RNA library preparation and sequencing. Small RNA libraries were prepared from the same RNA isolates of family 2 (described above). Small RNA Library preparation was done using the SOLiD® Total RNA-Seq Kit, exactly according to manufacturer's instructions (Life Technologies protocol, 4452437 Rev. B). Following templated bead preparation small RNA libraries were sequenced in a multiplexed manner on SOLiD 5500xl. Reads were trimmed to 23 bp and mapped using BWA (settings: -c -l 18 -k 1 -n 5). The number of reads mapping to each miRNA was used as a measure of expression. Significant changes of miRNA expression were identified using DEGSeq v1.10.0 [53].

Chromatin immunoprecipitations, library preparation and sequencing. ±20M cells were cross-linked with 2% formaldehyde in 10 mL PBS/10%FCS for 10 minutes rotating at RT (20°C). 0.125 M glycine was added to quench the reaction and cells were stored on ice. Following the cross-linking procedure, samples were centrifuged for 8 minutes at 1300 rpm (4°C). Pelleted cells were washed with 1 mL cold PBS and centrifuged at 1300 rpm, 4°C for 5 minutes again. After removal of the supernatant, the cell pellet was dissolved in 1mL freshly prepared lysis buffer to recover cross-linked nuclei (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM EDTA, 0.5% NP-40, 1% Triton X-100 and 1x Complete protease inhibitors (Roche)). Cells were lysed for an initial 10 minutes on ice and cell lysis was determined to be complete using Methyl Green - Pyronin staining. After completion of lysis, nuclei were washed twice in cold PBS. After the final wash, ±5M nuclei were dissolved in 100µL lysis buffer (MAGnify system, Invitrogen™) and 1x protease inhibitors (MAGnify system, Invitrogen™) and sheared in

microtubes (AFA Fiber Pre-Slit Snap-Cap 6x16mm, 520045) using the Covaris S2 sonicator (6 cycles of 60 seconds; duty cycle: 20%, intensity: 3, cycles per burst: 200, frequency sweeping). The remaining cross-linked nuclei were stored at -80°C for later use. Soluble chromatin in a size range of 150 - 400 bp was stored at -80°C and the equivalent of 1M nuclei was used per immunoprecipitation (IP). H3K4me3 IPs were carried out using the MAGnify system (Invitrogen™, 49-2024) following manufacturers instructions (Invitrogen manual A11261). Per IP, 1 µg of antibody was used (Millipore, 07-473 LOT# JBC1863338).
For library preparation, chromatin immunoprecipitated DNA was sheared to ± 100 bp in size using Covaris S2 (microtubes, 6 cycles of 60 seconds with duty cycle: 10%, intensity: 5, cycles/burst: 100, frequency sweeping). Following fragmentation, the End-It™ DNA end-repair kit (Epicentre, ER81050) was used to make the fragments blunt-ended and phosphorylate the 5' end of each molecule. Ligation of double stranded P1 and barcoded P2 adapters was done using the Quick Ligation kit (New England Biolabs, M2200). Samples were purified using Agencourt AMPure XP beads (A63882), and PCR amplified for 11 cycles using Platinum PCR SuperMix (Invitrogen™, 11306-016) and primers matching the P1 and P2 adapters. The PCR amplification included an initial 20 minutes nick translation step at 72 °C to remove the nick introduced at the 3' end of each molecule during adapter ligation. Samples were purified using AMPure XP beads and quality checked using an Agilent Technologies 2100 Bioanalyzer high-sensitivity DNA assay. DNA concentration for each sample was determined using a Qubit® Fluorometer high sensitivity DNA assay (Invitrogen™), to allow equimolar pooling of the barcoded libraries and subsequent templated bead preparation. Libraries were sequenced on one lane of SOLiD 5500 and reads were mapped using BWA.

Zebrafish injections. RNA oligonucleotides matching hsa-miR-520c-5p and hsa-miR-520b were ordered from IDT integrated DNA Technologies. Complementary strands (miRNA star sequence) contained two mismatches opposing the two 5' terminal bases of the miRNA strands. This improves miRNA effectiveness by facilitating loading in the RNA-induced silencing complex [35]. Single-stranded miRNA injection controls and mature miRNA duplexes were injected in the zebrafish zygote at a concentration of 5 µM (1 nL per zygote).

Overexpression of DPYD-ETV1 and FOXP1 fusions in HEK293 cells. Synthetically produced FOXP1 transcript variants were obtained from IDT (Leuven, Belgium). FOXP1 inserts were amplified using a forward primer (containing an NheI restriction site and the HA-tag-sequence; 5'- GCTAGCCCACCATGTACCCATACGATGTTCCAGATTACGCTATGCAAGAATCTGGGACTG-3') and reverse primers containing a BamHI restriction site (FOXP1_1: 5'-AATGGATCCTTAAAAAGTTAATTTTGAGGCCTTAGAGGG-3'; FOXP1_2: 5'-AATGGATCCTTAATTTTGAGGCCTTAGAGGGCTCATGTCC-3'). PCR products were subcloned into the mammalian expression vector Phage2-EF1alpha-IRES-Puro (Westburg, Leusden, The Netherlands) using the NheI and BamHI restriction sites. The resultant plasmids were

transfected into HEK293FT cells using calcium phosphate precipitation and 1µg plasmid DNA per 100-mm dish.

For immunofluorescence, cells were cultured on glass coverslips and fixed with 4% paraformaldehyde two days after transfection. Fixed cells were permeabilized in 0.5% Triton X-100 in PBS with 10% Fetal Calf Serum followed by blocking in 0.1% Triton X-100 in PBS with 10% Fetal Calf Serum (blocking buffer). The cells were then incubated with a rabbit polyclonal anti-HA tag antibody (Abcam, Cambridge, UK) diluted in blocking solution to a concentration of 2 µg/mL. After three washes in blocking solution, the cells were incubated with goat anti rabbit secondary antibody conjugated with Alexa Fluor 488 (Life Technologies, Bleiswijk, The Netherlands), followed by three additional washes in blocking solution. Coverslips with stained cells were mounted in Vectashield with DAPI (Brunschwig Chemie, Amsterdam, The Netherlands) and imaged using a Leica SPE confocal microscope. Acquired images were merged using ImageJ software.

For immunoblot analysis HEK293FT cells were lysed in NP-40 buffer (150 mM sodium chloride, 1.0% Triton X-100, and 50 mM Tris, pH 8.0) 2 days post-transfection. Protein lysates were diluted in 2X Laemmli buffer, boiled for 5 minutes and subjected to electrophoresis in 10% poly acrylamide gels followed by transfer to a PVDF-membrane. The resulting blot was blocked in PBS with 0.2% Tween and 5% milk powder (blocking buffer), followed by incubation with a rabbit polyclonal anti-HA tag antibody (Abcam), washing, and incubation in a Horseradish Peroxidase-conjugated secondary antibody (VWR, Amsterdam, The Netherlands). Blots were extensively washed in PBS with 0.2% Tween and antibody binding to the membrane was visualized using ECL (GE Healthcare, Diegem, Belgium) after which the signals were captured on light sensitive film.

Breakpoint data from cancer and patients with congenital disease. We obtained breakpoint data for genomic rearrangements in 96 patients with congenital disorders from published studies and from our own genome sequencing efforts. Cancer breakpoints were derived from published studies. If coordinates were in hg18, we used the UCSC lift over tool to convert the coordinates to hg19. We only included cancer breakpoints that were defined at nucleotide resolution. Per sample, the breakpoints were ordered by genomic position and breaks from the same sample occurring within a genomic interval of 2 kb were merged, because these may represent the same double-stranded DNA break [8]. In those cases the average of the two break-junction coordinates was used as the breakpoint coordinate.

Analysis of the overlap between cancer and constitutional breakpoint sets. To test the overlap between cancer and constitutional breakpoint sets, we generated random breakpoint sets equal in size to the constitutional breakpoint set. The random breakpoints in these sets were only taken from positions in the genome that could be covered by next-generation mate-pair sequencing datasets [8]. Therefore, we compiled a BAM file from six high-quality

datasets and required at least 300 uniquely, unambiguously and perfectly mapped (no mismatches) sequence reads with high mapping quality in the region of 1 kb flanking each side of each breakpoint in the random control set. This eliminated more than 5% of the random breakpoints that mostly covered repetitive regions such as the centromeres. Next, we matched the sample and chromosomal distribution and the size of the constitutional breakpoint set. We used 10,000 permutated constitutional breakpoint datasets and calculated the overlap with cancer breakpoints based on an inter-breakpoint distance of 2,000 bp. The empirical p-value was derived based on a comparison of the 10,000 random sets and the true constitutional breakpoint set. We chose 2,000 bp as the maximum distance between breakpoints because it provided us with sufficient breakpoints for subsequent feature overlaps while maintaining significance in overlap between cancer and constitutional breakpoints. To test the overlap with the cancer gene census (CGC) database genes we used the same 10,000 random breakpoint sets.

Determining the association of recurrent breakpoints with genomic features. We retrieved sets of genomic features from a variety of resources to determine overlap with recurrent breakpoints using permutation testing: CFS [54], aCFS [55], ERFS [43], DNase hypersensitive sites (UCSC), CpG islands (UCSC), G4 structures [56], SINE (UCSC), LINE (UCSC), LTR (UCSC), recombination hotspots [57], replication timing [58] and segmental duplications (UCSC). For ERFSs we used the UCSC LiftOver tool to convert mouse ERFS coordinates to hg19. Random control breakpoints were selected as described under the section "Analysis of the overlap between cancer and constitutional breakpoint sets". Permutation testing was performed for the 61 recurrent breakpoints using 10,000 sets of matched random control breakpoints. Overlaps were determined using the intersect function of BEDTools [59] and counting the unique number of breakpoints overlapping a feature. P-values were determined and corrected for multiple testing using Bonferroni correction. The window sizes used for the feature overlap were determined based on the distribution of the control sets. For calculation of replication timing, we took the score closest to the breakpoint based on recently published replication timing data [58]. The significance of the enrichment of late replication timing for recurrent breakpoints was determined using a Kruskal-Wallis test and 1000 sets of matched control breakpoints.

Author contributions

S.v.H., M.S., E.C. and W.K. designed the study. S.v.H., W.K., E.d.B. and N.L. performed RNA-seq. W.K., K.D., E.d.B., M.J., N.L., I.R., V.P., H.B., and M.E.T. performed mate-pair seq and breakpoint validations. S.v.H., K.d.L. and N.L. performed H3K4me3 ChIP-seq. W.H., S.v.H. and W.K. performed fusion gene cloning. E.W.K. performed overexpression in cell lines, immunoblotting, and confocal microscopy. W.K. and N.A. performed zebrafish experiments. JK performed sectioning and staining of zebrafish embryos. K.B. and E.K. isolated PBMCs.

M.S., W.K., S.v.H., M.v.R., A.M. and S.B. analyzed data. B.M., S.V., S.S., A.K., E.I., K.L., V.P., M.T., RH, L.v.d.V and E.v.B. provided patient material. S.v.H., M.S., W.K. and E.C. wrote the manuscript. All authors contributed to the final version of the manuscript.

Acknowledgments

We would like to thank Anko de Graaff and the Hubrecht Imaging Center for supporting the imaging. This work was funded by the Cancer Genomics Center and Netherlands Center for Systems Biology programs of the Netherlands Genomics Initiative, the Child Health Priority Program of the University Medical Center Utrecht, and the National Institutes of Health (GM061354, HD065286, MH095867). All authors declare no conflict of interest.

Data availability

The mate-pair, RNA and ChIP sequencing data are available from the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) under accession numbers: PRJEB5063 and ERP001438.

References

1. Stankiewicz P, Lupski JR: Structural variation in the human genome and its role in disease. Annu Rev Med 2010, 61:437–455.

2. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE: A copy number variation morbidity map of developmental delay. Nat Genet 2011, 43:838–846.

3. Vissers LELM, Stankiewicz P: Microdeletion and microduplication syndromes. Methods Mol Biol 2012, 838:29–75.

4. Hastings PJ, Lupski JR, Rosenberg SM, Ira G: Mechanisms of change in gene copy number. Nat Rev Genet 2009, 10:551–564.

5. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, Reymond A, Sun M, Sawa A, Gusella JF, Kamiya A, Beckmann JS, Katsanis N: KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature 2012, 485:363–367.

6. Lieber MR: The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annu Rev Biochem 2010, 79:181–211.

7. Hastings PJ, Ira G, Lupski JR: A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 2009, 5:e1000327.

8. Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ, van Binsbergen E, Renkens I, Duran K, Ballarati L, Vergult S, Giardino D, Hansson K, Ruivenkamp CAL, Jager M, van Haeringen A, Ippel EF, Haaf T, Passarge E, Hochstenbach R, Menten B, Larizza L, Guryev V, Poot M, Cuppen E: Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. Cell Rep 2012, 1:648–655.

9. Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR, McLaughlan CJ, Bawden CS, Reid SJ, Faull RLM, Snell RG, Hall IM, Shen Y, Ohsumi TK, Borowsky ML, Daly MJ, Lee C, Morton CC, Macdonald ME, Gusella JF, Talkowski ME: Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat Genet 2012.

10. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin M-L, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, et al.: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 2011, 144:27–40.

11. Weischenfeldt J, Symmons O, Spitz F, Korbel JO: Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 2013, 14:125–138.

12. Liu P, Erez A, Nagamani SCS, Dhar SU, Kołodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA, Bacino CA, Campos-Acevedo LD, Delgado MR, Freedenberg D, Garnica A, Grebe TA, Hernández-Almaguer D, Immken L, Lalani SR, McLean SD, Northrup H, Scaglia F, Strathearn L, Trapane P, Kang S-HL, Patel A, Cheung SW, Hastings PJ, Stankiewicz P, Lupski JR, et al.: Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell 2011, 146:889–903.

13. Talkowski ME, Ernst C, Heilbut A, Chiang C, Hanscom C, Lindgren A, Kirby A, Liu S, Muddukrishna B, Ohsumi TK, Shen Y, Borowsky M, Daly MJ, Morton CC, Gusella JF: Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. Am J Hum Genet 2011, 88:469–481.

14. Vergult S, van Binsbergen E, Sante T, Nowak S, Vanakker O, Claes K, Poppe B, Van der Aa N, van Roosmalen MJ, Duran K, Tavakoli-Yaraki M, Swinkels M, van den Boogaard M-J, van Haelst M, Roelens F, Speleman F, Cuppen E, Mortier G, Kloosterman WP, Menten B: Mate pair sequencing for the detection of chromosomal aberrations in patients with intellectual disability and congenital malformations. European journal of human genetics : EJHG 2013.

15. Chen W, Ullmann R, Langnick C, Menzel C, Wotschofsky Z, Hu H, Döring A, Hu Y, Kang H, Tzschach A, Hoeltzenbein M, Neitzel H, Markus S, Wiedersberg E, Kistner G, van Ravenswaaij-Arts CMA, Kleefstra T, Kalscheuer VM, Ropers H-H: Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. European journal of human genetics : EJHG 2010, 18:539–543.

16. Kleinman CL, Gerges N, Papillon-Cavanagh S, Sin-Chan P, Pramatarova A, Quang D-AK, Adoue V, Busche S, Caron M, Djambazian H, Bemmo A, Fontebasso AM, Spence T, Schwartzentruber J, Albrecht S, Hauser P, Garami M, Klekner A, Bognár L, Montes J-L, Staffa A, Montpetit A, Berube P, Zakrzewska M, Zakrzewski K, Liberski PP, Dong Z, Siegel PM, Duchaine T, Perotti C, et al.: Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR. Nat Genet 2014, 46:39–44.

17. Oh S, Shin S, Janknecht R: ETV1, 4 and 5: An oncogenic subfamily of ETS transcription factors. Biochim Biophys Acta 2012, 1826:1–12.

18. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM: Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005, 310:644–648.

19. Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, Menon A, Jing X, Cao Q, Han B, Yu J, Wang L, Montie JE, Rubin MA, Pienta KJ, Roulston D, Shah RB, Varambally S, Mehra R, Chinnaiyan AM: Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 2007, 448:595–599.

20. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM: Recurrent gene fusions in prostate cancer. Nat Rev Cancer 2008, 8:497–511.

21. Jeon IS, Davis JN, Braun BS, Sublett JE, Roussel MF, Denny CT, Shapiro DN: A variant Ewing's sarcoma translocation (7;22) fuses the EWS gene to the ETS gene ETV1. Oncogene 1995, 10:1229–1234.

22. Hermans KG, van der Korput HA, van Marion R, van de Wijngaart DJ, Ziel-van der Made A, Dits NF, Boormans JL, van der Kwast TH, van Dekken H, Bangma CH, Korsten H, Kraaij R, Jenster G, Trapman J: Truncated ETV1, fused to novel tissue-specific genes, and full-length ETV1 in prostate cancer. Cancer Res 2008, 68:7541–7549.

23. Ernst T, Score J, Deininger M, Hidalgo-Curtis C, Lackie P, Ershler WB, Goldman JM, Cross NCP, Grand F: Identification of FOXP1 and SNX2 as novel ABL1 fusion partners in acute lymphoblastic leukaemia. Br J Haematol 2011, 153:43–46.

24. Talkowski ME, Rosenfeld JA, Blumenthal I, Pillalamarri V, Chiang C, Heilbut A, Ernst C, Hanscom C, Rossin E, Lindgren AM, Pereira S, Ruderfer D, Kirby A, Ripke S, Harris DJ, Lee J-H, Ha K, Kim H-G, Solomon BD, Gropman AL, Lucente D, Sims K, Ohsumi TK, Borowsky ML, Loranger S, Quade B, Lage K, Miles J, Wu B-L, Shen Y, et al.: Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell 2012, 149:525–537.

25. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, Rieder MJ, Nickerson DA, Bernier R, Fisher SE, Shendure J, Eichler EE: Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet 2011, 43:585–589.

26. Flor I, Bullerdiek J: The dark side of a success story: microRNAs of the C19MC cluster in human tumours. J Pathol 2012, 227:270–274.

27. Bar M, Wyman SK, Fritz BR, Qi J, Garg KS, Parkin RK, Kroh EM, Bendoraite A, Mitchell PS, Nelson AM, Ruzzo WL, Ware C, Radich JP, Gentleman R, Ruohola-Baker H, Tewari M: MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. Stem Cells 2008, 26:2496–2505.

28. Belge G, Rippe V, Meiboom M, Drieschner N, Garcia E, Bullerdiek J: Delineation of a 150-kb breakpoint cluster in benign thyroid tumors with 19q13.4 aberrations. Cytogenet Cell Genet 2001, 93:48–51.

29. Rippe V, Dittberner L, Lorenz VN, Drieschner N, Nimzyk R, Sendt W, Junker K, Belge G, Bullerdiek J: The two stem cell microRNA gene clusters C19MC and miR-371-3 are activated by specific chromosomal rearrangements in a subgroup of

thyroid adenomas. PLoS ONE 2010, 5:e9485.

30. Rajaram V, Knezevich S, Bove KE, Perry A, Pfeifer JD: DNA sequence of the translocation breakpoints in undifferentiated embryonal sarcoma arising in mesenchymal hamartoma of the liver harboring the t(11;19)(q11;q13.4) translocation. Genes Chromosomes Cancer 2007, 46:508–513.

31. Huang Q, Gumireddy K, Schrier M, le Sage C, Nagel R, Nair S, Egan DA, Li A, Huang G, Klein-Szanto AJ, Gimotty PA, Katsaros D, Coukos G, Zhang L, Puré E, Agami R: The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis. Nat Cell Biol 2008, 10:202–210.

32. Hu N, Zhang J, Cui W, Kong G, Zhang S, Yue L, Bai X, Zhang Z, Zhang W, Zhang X, Ye L: miR-520b regulates migration of breast cancer cells by targeting hepatitis B X-interacting protein and interleukin-8. J Biol Chem 2011, 286:13714–13722.

33. Fornari F, Milazzo M, Chieco P, Negrini M, Marasco E, Capranico G, Mantovani V, Marinello J, Sabbioni S, Callegari E, Cescon M, Ravaioli M, Croce CM, Bolondi L, Gramantieri L: In hepatocellular carcinoma miR-519d is up-regulated by p53 and DNA hypomethylation and targets CDKN1A/p21, PTEN, AKT3 and TIMP2. J Pathol 2012, 227:275–285.

34. Wu S, Huang S, Ding J, Zhao Y, Liang L, Liu T, Zhan R, He X: Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3' untranslated region. Oncogene 2010, 29:2302–2308.

35. Kloosterman WP, Wienholds E, Ketting RF, Plasterk RHA: Substrate requirements for let-7 function in the developing zebrafish embryo. Nucleic Acids Res 2004, 32:6284–6291.

36. Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SCM, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M, Cuppen E: Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. Hum Mol Genet 2011, 20:1916–1924.

37. Nazaryan L, Stefanou EG, Hansen C, Kosyakova N, Bak M, Sharkey FH, Mantziou T, Papanastasiou AD, Velissariou V, Liehr T, Syrrou M, Tommerup N: The strength of combined cytogenetic and mate-pair sequencing techniques illustrated by a germline chromothripsis rearrangement involving FOXP2. European journal of human genetics : EJHG 2013.

38. Granot-Hershkovitz E, Raas-Rothschild A, Frumkin A, Granot D, Silverstein S, Abeliovich D: Complex chromosomal rearrangement in a girl with psychomotor-retardation and a de novo inversion: inv(2)(p15;q24.2). Am J Med Genet A 2011, 155A:1825–1832.

39. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: A census of human cancer genes. Nat Rev Cancer 2004, 4:177–183.

40. Singh D, Chan JM, Zoppoli P, Niola F, Sullivan R, Castano A, Liu EM, Reichel J, Porrati P, Pellegatta S, Qiu K, Gao Z, Ceccarelli M, Riccardi R, Brat DJ, Guha A, Aldape K, Golfinos JG, Zagzag D, Mikkelsen T, Finocchiaro G, Lasorella A, Rabadan R, Iavarone A: Transforming fusions of FGFR and TACC genes in human glioblastoma. Science 2012, 337:1231–1235.

41. Drier Y, Lawrence MS, Carter SL, Stewart C, Gabriel SB, Lander ES, Meyerson M, Beroukhim R, Getz G: Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res 2013, 23:228–235.

42. Durkin SG, Glover TW: Chromosome fragile sites. Annu Rev Genet 2007, 41:169–192.

43. Barlow JH, Faryabi RB, Callén E, Wong N, Malhowski A, Chen HT, Gutierrez-Cruz G, Sun H-W, McKinnon P, Wright G, Casellas R, Robbiani DF, Staudt L, Fernandez-Capetillo O, Nussenzweig A: Identification of early replicating fragile sites that contribute to genome instability. Cell 2013, 152:620–632.

44. Le Tallec B, Millot GA, Blin ME, Brison O, Dutrillaux B, Debatisse M: Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. Cell Rep 2013, 4:420–428.

45. Letessier A, Millot GA, Koundrioukoff S, Lachages A-M, Vogt N, Hansen RS, Malfoy B, Brison O, Debatisse M: Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature 2011, 470:120–123.

46. Jiang Y, Lucas I, Young DJ, Davis EM, Karrison T, Rest JS, Le Beau MM: Common fragile sites are characterized by histone hypoacetylation. Hum Mol Genet 2009, 18:4501–4512.

47. Cirstea IC, Kutsche K, Dvorsky R, Gremer L, Carta C, Horn D, Roberts AE, Lepri F, Merbitz-Zahradnik T, König R, Kratz CP, Pantaleoni F, Dentici ML, Joshi VA, Kucherlapati RS, Mazzanti L, Mundlos S, Patton MA, Silengo MC, Rossi C, Zampino G, Digilio C, Stuppia L, Seemanova E, Pennacchio LA, Gelb BD, Dallapiccola B, Wittinghofer A, Ahmadian MR, Tartaglia M, et al.: A restricted spectrum of NRAS mutations causes Noonan syndrome. Nat Genet 2010, 42:27–29.

48. Fam HK: Caught in the AKT: identification of a de novo pathway in MCAP and MPPH and its therapeutic implications. Clin Genet 2012.

49. Lipska BS, Koczkowska M, Wierzba J, Ploszynska A, Iliszko M, Izycka-Swieszewska E, Adamkiewicz-Drozynska E, Limon J: On the significance of germline cytogenetic rearrangements at MYCN locus in neuroblastoma. Mol Cytogenet 2013, 6:43.

50. Buijs A, Poot M, van der Crabben S, van der Zwaag B, van Binsbergen E, van Roosmalen MJ, Tavakoli-Yaraki M, de Weerdt O, Nieuwenhuis HK, van Gijn M, Kloosterman WP: Elucidation of a novel pathogenomic mechanism using genome-wide long mate-pair sequencing of a congenital t(16;21) in a series of three RUNX1-mutated FPD/AML pedigrees. Leukemia 2012, 26:2151–2154.

51. Forrest DL, Lee CLY: Constitutional rearrangements of 7q22 in hematologic malignancies. a new case report. Cancer Genet Cytogenet 2002, 139:75–77.

52. Merks JHM, Ozgen HM, Koster J, Zwinderman AH, Caron HN, Hennekam RCM: Prevalence and patterns of morphological abnormalities in patients with childhood cancer. JAMA 2008, 299:61–69.

53. Wang L, Feng Z, Wang X, Wang X, Zhang X: DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 2010, 26:136–138.

54. Debacker K, Kooy RF: Fragile sites and human disease. Hum Mol Genet 2007, 16 Spec No. 2:R150–8.

55. Fungtammasan A, Walsh E, Chiaromonte F, Eckert KA, Makova KD: A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? Genome Res 2012, 22:993–1005.

56. Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT, Bacolla A, Collins JR, Stephens RM: Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Res 2013, 41(Database issue):D94–D100.

57. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: A fine-scale map of recombination rates and hotspots across the human genome. Science 2005, 310:321–324.

58. Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA: Differential relationship of DNA replication timing to different forms of human mutation and variation. Am J Hum Genet 2012, 91:1033–1040.

59. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010, 26:841–842.

Supplementary Information

Supplementary Table 1. Karyotype and phenotypic description of patients with chromosomal aberrations used for molecular phenotyping.

Patient	Karyotype	Phenotypic description
Patient 1 (chromothripsis signature)	46,XX,t(1;12;7)(p21;q14;p21).arr 7p21.3p21.1(12,725,574- 15,056,327)x1 dn	Psychomotor retardation, facial dysmorphisms, doubling of thumb
Patient 1 mother	46,XX	
Patient 1 father	46,XY	
Patient 2 (tandem duplication)	46,XX.arr 19q13.4 1q13.42(58,878,634-59,294,958)x3 dn	Severe macrocephaly and psychomotor retardation
Patient 2 mother	46,XX	
Patient 2 father	46,XY	
Patient 2 sibling	46,XY	

Supplementary Table 2. Breakpoint junctions of de novo chromothripsis in the patient with congenital disease subjected to molecular phenotyping.

For each of the 13 junctions, coordinates, orientation, number of supporting mate-pair reads and disrupted genes are given.

ID	chr_bp1	co_bp1	chr_bp2	co_bp2	ori[a]	#reads	sequence at breakpoint	gene_bp1	gene_bp2
1	1	95146001	1	97683903	TH	19	Insertion ATTT		*DPYD*(-)
2	1	95146004	3	76406934	HH	7	Blunt		*ROBO2*(+)
3	1	97683903	7	19114083	TT	16	Microhomology A	*DPYD*(-)	
4	1	98219840	7	17140783	TT	7	Insertion TA	*DPYD*(-)	
5	1	98219835	7	13962633	HT	24	Microhomology CT	*DPYD*(-)	*ETV1*(-)
6	3	71078624	3	75208212	TT	24	Blunt	*FOXP1*(-)	
7	3	71078621	7	19003340	HT	14	Microhomology T	*FOXP1*(-)	*HDAC9*(+)
8	7	11404421	7	13612031	TH	21	Blunt		
9	7	15423145	12	63505335	HT	8	Blunt	*AGMO*(-)	
10	7	18519891	7	19320228	HT	28	Microhomology T	*HDAC9*(+)	
11	7	19003351	7	19114085	HH	26	Microhomology CT	*HDAC9*(+)	
12	7	19512505	12	63505338	TH	12	Microhomology TTC		
13	7	19320226	7	19512655	HH	7	Blunt		

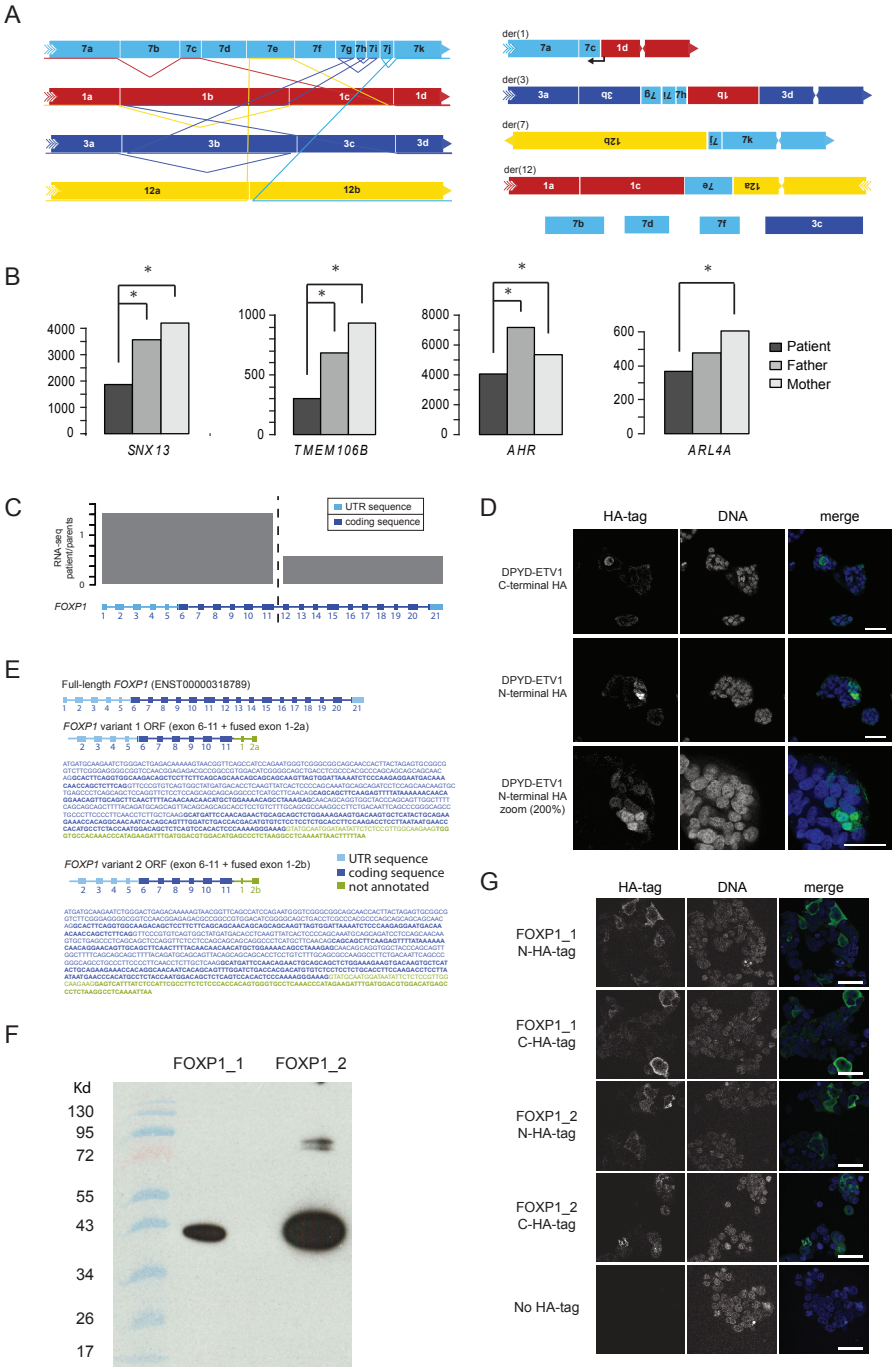
a) Orientation of breakpoint junction, H=head, T=tail. The first letter corresponds to chr_bp1 and co_bp1. The second letter corresponds to chr_bp2 and co_bp2.

Supplementary Table 3. List of CGC genes hit by chromosomal rearrangements in patients with congenital disease.

For each broken gene, chromosome, location, patient ID and the study they were derived from are given.

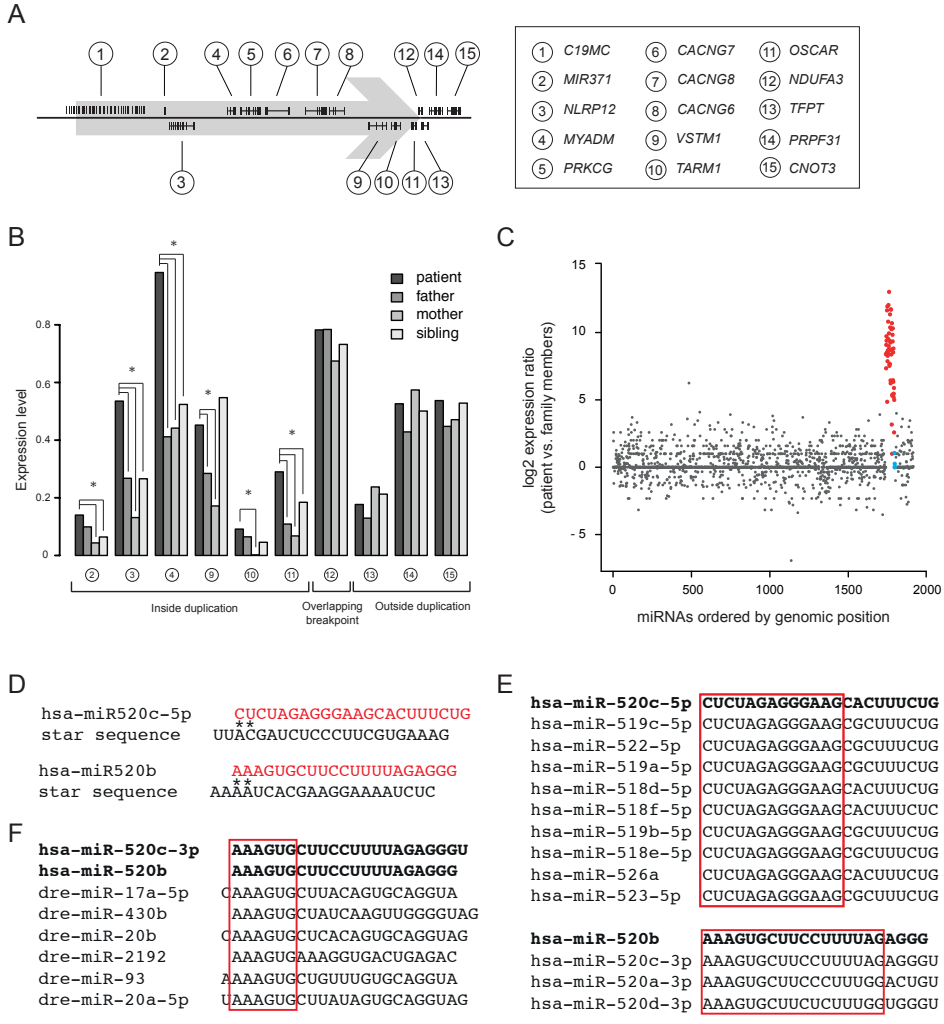
#	Chr	Breakpoint	Patient ID	Study	Ensembl_ID	Genename
1	3	71072121	DGAP157	Chiang et al, Nature Genetics 2012 Mar 4;44(4):390-7	ENSG00000114861	*FOXP1*
2	3	71078713	Patient 3	Kloosterman et al, Cell Reports, Volume 1, Issue 6, 648-655, 2012	ENSG00000114861	*FOXP1*
3	3	155627600	DGAP203/AC-05-0118	Chiang et al, Nature Genetics 2012 Mar 4;44(4):390-7	ENSG00000163655	*GMPS*
4	5	158400207	Patient 8	Kloosterman et al, Cell Reports, Volume 1, Issue 6, 648-655, 2012	ENSG00000164330	*EBF1*
5	5	158516037	Patient 8	Kloosterman et al, Cell Reports, Volume 1, Issue 6, 648-655, 2012	ENSG00000164330	*EBF1*
6	7	13962541	Patient 3	Kloosterman et al, Cell Reports, Volume 1, Issue 6, 648-655, 2012	ENSG00000006468	*ETV1*
7	8	38292459	DGAP011	Chiang et al, Nature Genetics 2012 Mar 4;44(4):390-7	ENSG00000077782	*FGFR1*
8	9	14228567	Case12	This work	ENSG00000147862	*NFIB*
9	10	114339379	Patient 6	Kloosterman et al, Cell Reports, Volume 1, Issue 6, 648-655, 2012	ENSG00000151532	*VTI1A*
10	12	66354057	11D2515V	This work	ENSG00000149948	*HMGA2*
11	21	36293422	09D5727_09D0921	This work	ENSG00000159216	*RUNX1*
12	21	36346553	09D5727_09D0921	This work	ENSG00000159216	*RUNX1*

Supplementary figure 1



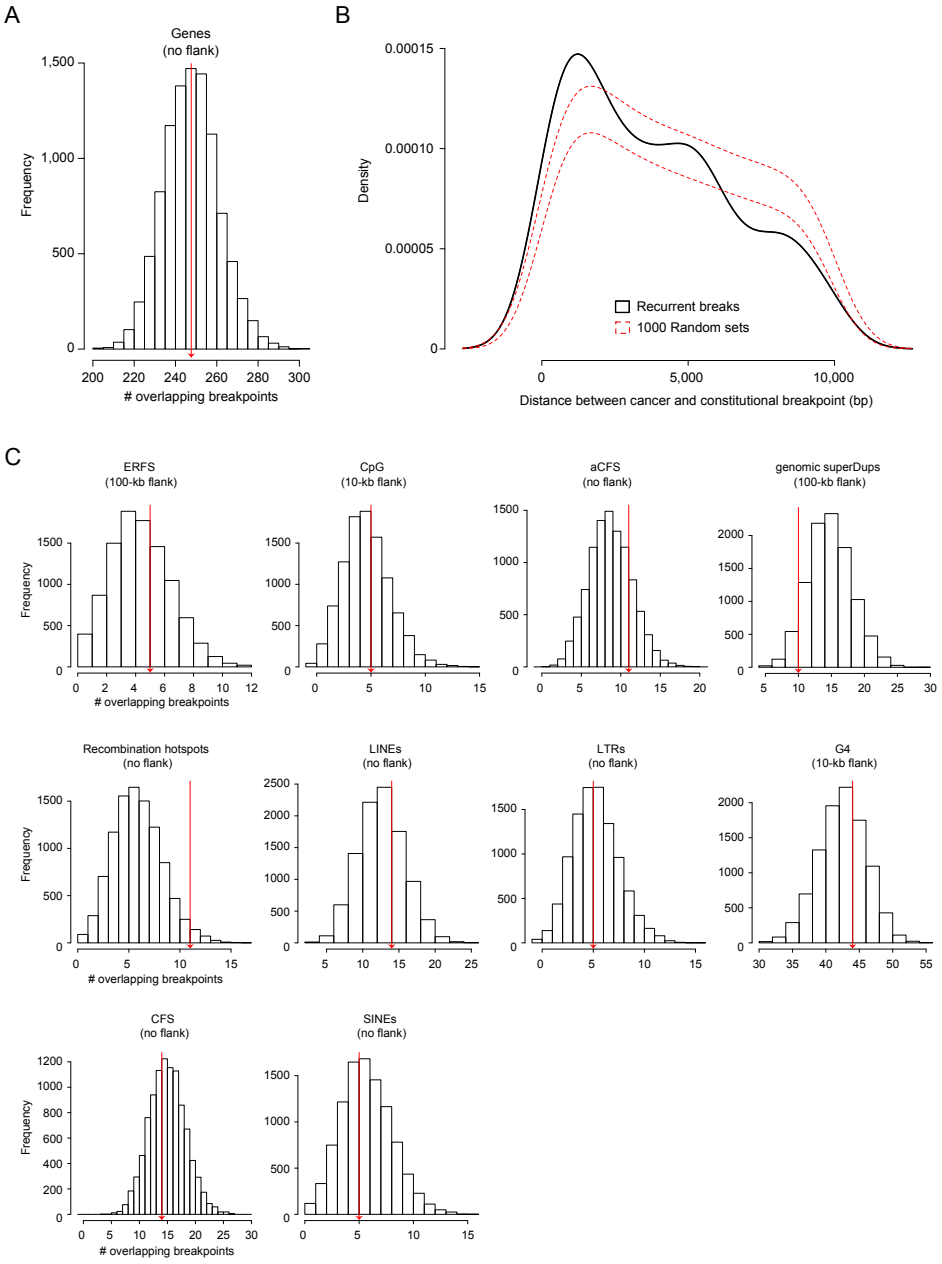
Supplementary figure 1 - **Result of molecular phenotyping of a patient with constitutional chromothripsis.** (**A**) Schematic representation of breakpoints and derivative (der) chromosomes resulting from chromothripsis. The lines connecting the individual fragments indicate the junctions. The derivative chromosomes frequently involve inverted fragments (shown upside down). Four deleted fragments on chromosomes 3 and 7 are depicted below the derivative chromosomes. The black arrow at the breakpoint junction between chromosome 1 and 7 indicates the DPYD (chr1) - ETV1 (chr7) gene fusion. To allow comprehensive visualization, the fragments are not displayed to scale. (**B**) Bar plots showing the normalized gene expression levels for 4 genes that were located in deleted genome regions. Significant expression level changes ($p < 0.0001$) between the chromothripsis patient, the father and the mother are marked with an asterisk. (**C**) Ratio of the FOXP1 gene expression level difference between patient and both parents, across the breakpoint on chromosome 3. (**D**) Expression of the DPYD-ETV1 fusion in HEK293 cells. Immunostainings showing the expression and cellular localization of the DPYD-ETV1 fusion. Nuclear DNA (middle panel) is stained using DAPI. Both N- and C-terminal HA-tagged products show expression in a minority of the cells suggesting that only under specific conditions DPYD-ETV1 expression results in protein. (**E**) Formation of novel FOXP1 fusion transcripts resulting from chromothripsis. RNA-seq and RT-PCR revealed the formation of two novel FOXP1 fusion transcripts. The fusion transcripts were formed because of read-through transcription across a de novo breakpoint junction into a fragment of chromosome 7. (**F**) Western blot analysis showing bands corresponding to both FOXP1 products using antibodies against a N-terminal HA-tag. (**G**) Immunostainings showing the expression and cellular localization of both FOXP1 products. Nuclear DNA (middle panel) is stained using DAPI. Both N- and C-terminal HA-tagged products show identical localization and expression, indicative of a stable protein.

Supplementary figure 2



Supplementary figure 2 - **Expression of genes and microRNAs in a 424.5 kb de novo genomic duplication in the second patient used for molecular phenotyping**. (**A**) Overview of duplication region. (**B**) Normalized expression levels for six genes positioned within the duplication (NLRP12, MYADM, VSTM1, TARM1, OSCAR and MIR371), one gene overlapping the 3' breakpoint (NDUFA3) and three genes outside the duplication (TFPT, PRPF31 and CNOT3). Significant changes in gene expression are indicated by an asterisk ($p < 0.0001$). The x-axis shows normalized exonic read counts. (**C**) Genome-wide miRNA expression analysis in the patient with a chromosome 19 tandem duplication. miRNA expression is based on miRNA-Seq. Expression levels are normalized for the total amount of reads in microRNAs and represent log2 ratios of expression in the patient versus the average of the father, mother and healthy sibling. miRNAs (dark grey) are ordered by their genomic location. miRNAs residing in the C19MC cluster are indicated in red; miRNAs from the MIR371 gene that is expressed via its own promoter are shown in light blue. (**D**) Structure of miRNA duplexes injected in zebrafish embryos. The asterisk indicates as mismatch in the duplex. (**E**) Homology of miR-520b and miR-520c among C19MC miRNAs. (**F**) Similarity of miR-520b and miR-520c miRNA seed sequences with zebrafish miRNAs. Identical sequences are indicated with red square boxes.

Supplementary figure 3



Supplementary figure 3 - **Genomic feature overlap of recurrent breakpoints.** (**A**) Histogram depicting the overlap of all 550 constitutional breakpoints with protein-coding genes. The red line shows the number of overlapping breakpoints, the bars show permutation testing results of 10,000 matched random control breakpoints. (**B**) Density plot showing the absolute distance between constitutional breakpoints and cancer breakpoints within 10 kb (black line). The area between de dotted red lines represents the mean distance between breakpoints +/- 1 SD computed over 1000 simulation sets. This plot highlights a skewed distribution for constitutional breakpoints towards short inter-breakpoint distances (< 2 kb) as compared to the random breakpoint sets. (**C**) Histograms showing genomic feature overlaps of 10 features with the set of breakpoints recurrent within 2 kb (n = 61). The x-axis shows the number of breakpoints overlapping features and the y-axis shows the frequency of random breakpoint sets with X number of overlaps. Flanks were determined based on the minimum flank to create sufficient overlap for a normal distribution. The red arrow shows the actual number of recurrent breakpoints overlapping the genomic features. After Bonferroni correction, none of these features are significantly associated ($p < 0.05$) with the recurrent breakpoints.

7

Multi-level effects of noncoding single nucleotide and structural variation on genome function

Sebastiaan van Heesch[1*], Roel Hermsen[1*], Nico Lansu[1], Kim de Luca[1], Wim Spee[1], Sander Boymans[1], Ewart de Bruijn[1], Mark Verheul[1], Geert Geeven[1], Peter Krijger[1], David Thybert[2], Paul Flicek[2], Wouter de Laat[1], Edwin Cuppen[1,3] & Marieke Simonis[1]

*These authors contributed equally to this work

[1] Hubrecht Institute and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands
[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[3] Department of Medical Genetics, UMC Utrecht, Universiteitsweg 100, 3584 GG Utrecht, The Netherlands

Abstract

The importance of the noncoding genome in gene expression regulation and disease has become pivotal over recent years, yet the mechanisms via which noncoding genomic variants contribute to phenotypic differences are poorly understood. Here, we systematically investigate the functional consequences of natural occurring noncoding variation on regulatory elements and chromatin organization in liver samples from ten inbred rat strains. We combine data on genome-wide nucleotide and structural variation (WGS), epigenetic marks for promoters and enhancers (ChIP-seq), transcription (RNA-seq) and genome-wide chromatin conformation (Hi-C) to dissect multiple levels of genome function and variation therein.

We find that both single nucleotide variants and structural variants directly modulate the activity of regulatory elements such as enhancers, primarily by impairing transcription factor binding. Also, we find that structural variants that do not directly overlap regulatory elements or genes drive long-range gene expression and epigenetic changes by affecting the higher-order chromatin organization. For example, deletions that target boundaries separating active from inactive chromatin domains drive the extension of these domains, which results in the activation or repression of enhancers and subsequent changes in gene expression.

This study is the first showing the diversity and relevance of mechanisms by which noncoding genomic variants modulate genome function ***in vivo***. While these finding re-emphasize the extraordinary complexity of genomic regulation, the identified principles are valuable for interpretation of whole genome sequencing data and the identification of pathogenic variation and molecular mechanisms driving disease.

Introduction

Genetic studies aiming for the identification of causal disease variants typically prioritize single nucleotide variants (SNVs) in protein-coding genes, with high potential to affect gene function [1, 2]. However, of all disease-associated variants detected by human genome-wide association studies (GWAS), the vast majority (93%) resides in the noncoding genome [3-5]. This is not unexpected, because only ~2-3% of the human genome encodes protein [6, 7] and for an estimated 80% of the genome a regulatory function has been proposed [8, 9], suggesting a mechanistic link between the noncoding genome and disease risk [10].

Variants in noncoding genomic regions can contribute to disease formation by affecting regulatory elements. These regulatory elements, such as enhancers and promoters, are crucial for gene expression regulation and marked by epigenetic modifications. Trimethylation of the lysine 4 residue of histone H3 (H3K4me3) is generally associated with active promoters, indicative of start sites of actively transcribed genes [11]. Acetylation of the lysine 27 residue of histone H3 (H3K27Ac) is associated with both distal and proximal active enhancer elements as well as active promoters, thus co-localizing with H3K4me3 at gene promoters

but not enhancers [12]. Regulatory element activity depends on the binding of transcription factors (TFs) [13], resulting in precise regulation of cell-type specific transcriptional networks. TFs bind both promoters and enhancers at proximal and distal sites and are involved in the physical interactions between these regulatory regions [14]. TFs can play general roles in transcription regulation of every gene, or be dedicated to tissue-specific genes. Examples of the latter are FOXA1, HNF4A and CEBPA, three factors that regulate the transcription of liver-specific genes, but also depend on each other for chromatin binding [15].
In addition to these local modes of gene expression regulation, the genome is also organized in nuclear space. The nuclear 3D organization divides chromosomes into large independent units of transcriptional activity or inactivity and interactions between regulatory elements occur largely within these domains [16]. These so-called topological domains are separated by boundaries and disruption of these boundaries can result in long-range epigenetic misregulation [17]. Genome-wide adaptations of 3C-based chromatin conformation capture techniques (e.g. Hi-C [18] or ChIA-PET [19]) provide the means to study chromatin organization in great detail and determine the location of topological domains. However, these techniques have not been applied yet to catalogue the effects of natural genetic variation on chromatin organization ***in vivo***.

Recently, a series of ***in vitro*** studies [20-23] revealed a direct effect of noncoding SNVs on histone modification and TF binding; two successive key events in the regulation of gene expression. These studies, as well as all genome-wide association studies, primarily focus on the effects of single nucleotide changes. SNVs are the most studied type of genomic variation because of their high detection efficiency and their mostly predictable effects, for example on coding genes [24, 25]. However, another category of genomic variation is structural genome variation. Even though structural variants are less frequent than SNVs, they affect many more bases per genome, for example via large genomic deletions or duplications [26]. A thorough evaluation of the effects of structural changes in the noncoding genome is required to improve our understanding of the functional consequences of all types of noncoding variation on genome function.

Here, we study the effects of genomic variants on the chromatin state and nuclear organization of rat liver tissue from ten inbred strains, making use of their homozygous genomes to avoid allele-specific effects. Integration of structural and single nucleotide genomic variant information from whole genome sequencing, regulatory element profiling obtained by ChIP-seq (enhancers, promoters) and genome-wide chromatin organization maps obtained by Hi-C allows us to dissect the effects of naturally occurring single nucleotide and structural variation on genome function ***in vivo***.

Results

In vivo characterization of the epigenetic regulatory landscape in ten rats

We used chromatin immunoprecipitation (ChIP) targeting the trimethylated lysine 4 residue of histone H3 (H3K4me3) and the acetylated lysine 27 residue of histone H3 (H3K27Ac) in liver tissue from ten rat strains (ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WN/N, WKY/N, BN-***Lx***/Cub and SHR/OlaIpcv) to gain insight in the ***in vivo*** distribution of regulatory elements (Fig. 1A). This resulted in the identification of 52,240 active enhancers and 14,511 active promoters in all ten rats combined (Table 1). Publicly available rat liver ChIP data of liver-specific TFs (HNF4A, CEBPA and FOXA1) [15] were used to get a general impression of the binding of TFs to the identified enhancers and promoters. We determined the overlap of TF binding profiles with regulatory elements and find that the three liver-specific TFs show very frequent co-localization with active regulatory elements, with over 80% of all TF peaks overlapping an enhancer or promoter (Fig. 1B and 1C). These data indicate that the genome-wide binding profiles of tissue-specific TFs are highly restricted to active regulatory elements. Also, approximately half of the liver TF-bound enhancers (n = 9,356) bind all three liver-specific TFs, indicative of high co-localization between tissue-specific TFs (Supplementary fig. 1). In total, the TFs bind 35.9% of all active enhancers and 47.9% of all active promoters. KEGG pathway analyses of genes associated with regulatory elements bound by the liver-specific TFs are enriched for pathways involving liver-specific processes such as drug metabolism ($p = 1.86*10^{-7}$), retinol metabolism ($p = 3.42*10^{-6}$) and metabolism of xenobiotics by cytochrome P450 ($p = 5.18*10^{-9}$). Genes associated with regulatory elements not bound by any of the liver-specific TFs showed enrichment for more general KEGG pathways such as WNT signaling ($p = 1.75*10^{-4}$), cell cycle ($p = 1.92*10^{-4}$) and ubiquitin mediated proteolysis ($p = 5.93*10^{-4}$).

Table 1. Epigenetic and genomic variation in ten rat strains

Strains	Enhancers	Promoters	SNVs	Deletions	Duplications
ACI/N	15,926	10,024	2,514,743	569	115
BN/SsN	24,667	12,773	52,147	21	1
BN-*Lx*/Cub	22,989	10,981	91,139	9	24
BUF/N	23,270	12,001	2,312,032	485	113
F344/N	26,907	12,323	2,370,007	610	180
M520/N	19,065	11,779	2,365,896	521	187
MR/N	27,611	12,898	2,353,340	600	317
SHR/OlaIpcv	22,124	11,122	3,628,566	461	384
WKY/N	23,380	12,108	2,582,142	557	291
WN/N	21,616	12,477	2,367,931	551	108
Total unique	52,240	14,511	6,642,812	1,780	1,176

This confirms that liver-specific TFs regulate gene expression of liver-specific genes. Taken together, we have identified ~15,000 promoters and ~52,000 enhancers that collectively hold the binding sites of 80% of three liver-specific TFs. Binding of these liver-specific TFs is correlated with liver gene function.

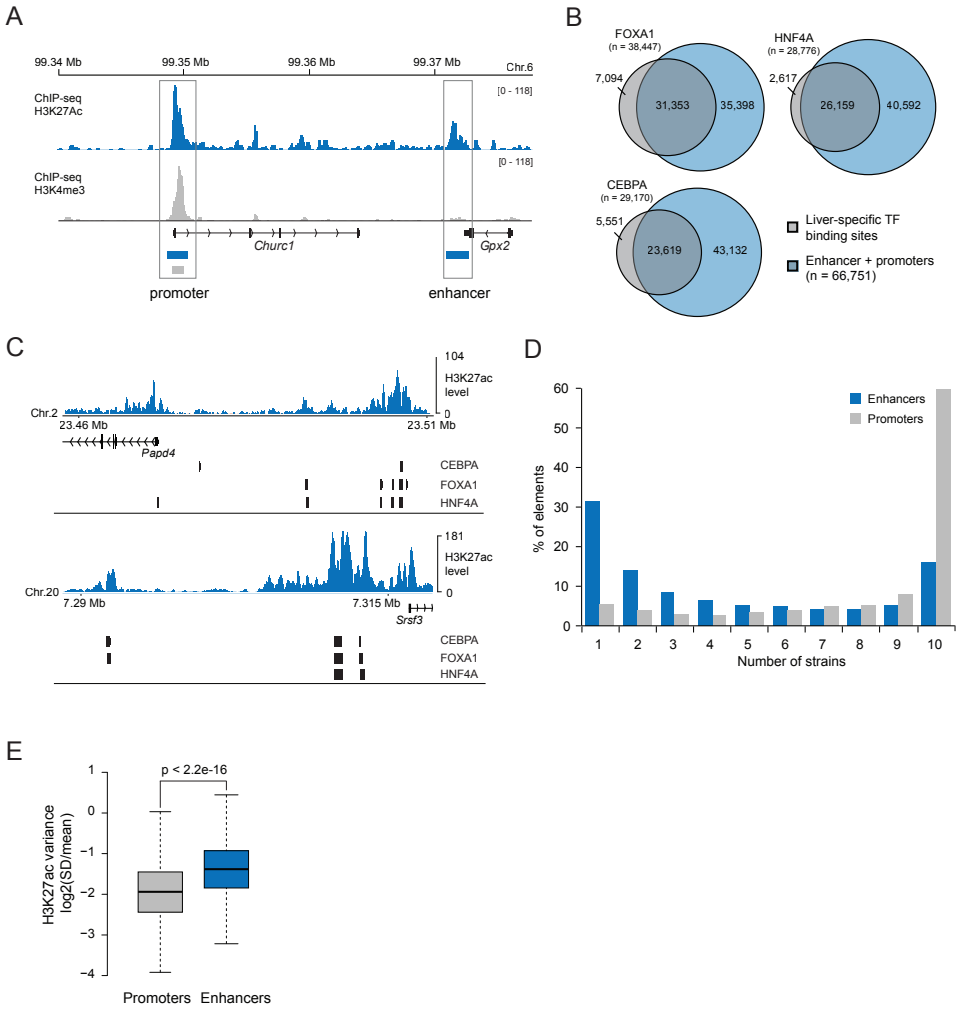


Figure 1 - **Epigenetic characterization of liver tissue from ten rat strains**. (**A**) Identification of active enhancers and promoters. Promoters show enrichment for H3K4me3 (grey) and H3K27Ac (blue) modifications, as defined by peak calling. Enhancers are solely marked by H3K27Ac and lack H3K4me3 enrichment. (**B**) Venn diagrams displaying the overlap of liver-specific transcription factor binding positions of FOXA1, HNF4A and CEBPA with active enhancers and promoters. (**C**) Examples of liver-specific TF binding localized to gene promoters and enhancers. (**D**) Bar plot showing the distribution of enhancers and promoters among the ten rat strains. (**E**) Boxplot showing the relative variance in H3K27Ac level in promoters and enhancers.

Enhancers are highly diverse between strains

Next, we studied the amount of variation that is found between strains at the epigenetic level by determining the presence or absence of regulatory elements. We first examined in how many strains each regulatory element could be identified. The vast majority (~60%) of all promoters is present in all ten rat livers, ~35% is shared between 2-9 strains and only 5% is strain specific. Enhancer elements appear more variable in presence between the 10 strains. Only 16% is present in all 10 strains, 53% in 2-9 strains and 31% is strain specific (Fig. 1D). To substantiate this finding, we determined the relative variance in H3K27Ac level for both types of elements, between all ten strains. In agreement with the previous results (Fig. 1D) enhancers show significantly higher relative variance in H3K27Ac signal strength than promoters ($p < 2.2e-16$, Fig. 1E). Although methodological factors may influence our observations, these results undoubtedly show that promoters appear much less variable between the ten strains than enhancers.

Genomic variation contributes directly to epigenetic diversity *in vivo*

We next asked what the contribution of genetic variability was on epigenetic diversity between the ten strains. Recently, we have sequenced the genomes of the 10 rat strains studied here, providing detailed information on all the detected strain-specific genetic variants and all the variation shared between multiple strains [27, 28]. To determine the genetic component underlying enhancer conservation and TF binding, we first determined the distribution of SNVs over the coding and noncoding genome regions of the ten rats (Table 1). For the SNVs of all ten genomes combined, we find densities of 2.8 SNVs/kb genome wide, 1.7 SNVs/kb in coding regions, 2.2 SNVs/kb in promoters and 2.6 SNVs/kb in active enhancers. Thus, active enhancers show a level of genetic variation that is intermediate between that found genome wide and in coding regions.

Next, we determined the segregation of SNVs with quantitative H3K27Ac level differences throughout the ten rat genomes to investigate a causal link between specific variants and the activity of the element. We defined the strain distribution pattern for each SNV that is located in an enhancer (n = 97,030) and selected the SNVs that were found in at least 3 strains (and no more than 7 strains) to allow statistical analyses. We then for each SNV compared the H3K27Ac levels of the enhancer in strains with the SNVs to the levels in strains without the variant using student's t-test. We calculated the number of positive SNV-H3K27Ac level correlations with different p-value thresholds (Fig. 2A). To value each individual association, multiple testing would be required, but here we investigated the associations as a whole. To assess the significance of associations with small p-values, we repeated the analysis for a dataset in which we randomized the SNV distribution over the strains, for each SNV. We find a clear enrichment for the actual significant SNV distribution compared to random ***in silico***

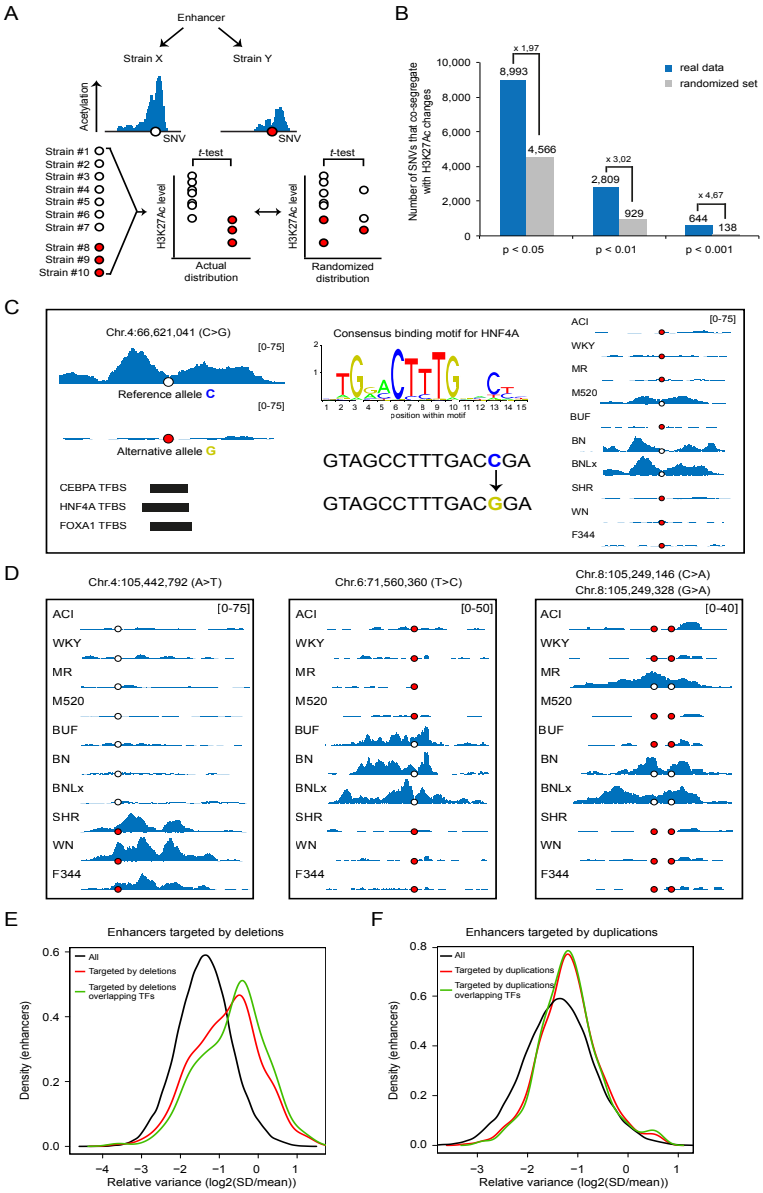


Figure 2 - **Direct effects of genomic variation on epigenetic diversity**. (**A**) Schematic representation of the analysis used to test the association of SNVs with H3K27Ac levels. For each enhancer locus, H3K27Ac levels of strains with and without the SNV were correlated and compared with a random strain distribution using Student's t-test. (**B**) Bar plot showing the number of SNVs positively correlating with H3K27Ac levels at p-value cut-offs of 0.05, 0.01 and 0.001, for both the in vivo and in silico results of the Student's t-test. (**C**) Example of an SNV perturbing the HNF4A consensus DNA binding motif, likely driving the loss of H3K27Ac enrichment (blue) for 7 strains carrying the mutation. (**D**) Three examples of SNVs that co-segregate with H3K27Ac levels of enhancers without overlapping a liver-specific TF peak. (**E+F**) Density plots depicting the variance in H3K27Ac level across the ten strains for enhancers targeted by deletions (**E**) or duplications (**F**) (red lines). The green line shows the H3K27Ac level variance for deletions or duplications targeting a TF binding site within the enhancer.

sampling, which increases with decreasing p-values (8,993 versus 4,566 with $p < 0.05$; 644 versus 138 with $p < 0.001$) (Fig. 2B). The co-segregation of SNVs with enhancer levels can involve SNVs that overlap (and potentially disrupt or initiate) binding sites of the liver-specific TFs (n = 8,796), or not overlap with TF binding positions at all (n = 88,234). We find that SNVs that overlap one of the three tested TF binding sites show a slightly increased correlation with H3K27Ac levels compared to random sets, than enhancers that do not possess an SNV in a TF peak (Table 2). For example, we find examples of SNVs that change the consensus binding motif a TF, resulting in decreased H3K27Ac levels in the strains that carry the variant (Fig. 2C and Supplementary fig. 2). However, we also find many examples of SNVs that do not directly overlap with the binding sites of the three liver-specific TFs, but that do co-segregate with a change in H3K27Ac level of the enhancer (Fig. 2D). Possibly, these SNVs do affect TF binding of other factors that were not measured in this study, but do define the activity of the element. These results reconfirm the previous ***in vitro*** finding that SNVs can directly influence enhancer activity [20-23].

Table 2. SNV - H3K27Ac level correlation analysis with multiple p-value thresholds

p-value threshold	Actual data	Randomized distribution	Fold-change (actual vs random)
	SNVs in enhancers		
$p < 0.05$	8993	4566	1.97
$p < 0.01$	2809	929	3.02
$p < 0.001$	644	138	4.67
	SNVs in enhancers in a TF peak		
$p < 0.05$	1067	437	2.44
$p < 0.01$	342	100	3.42
$p < 0.001$	77	12	6.42
	SNVs in enhancers not in a TF peak		
$p < 0.05$	7926	4129	1.92
$p < 0.01$	2467	829	2.98
$p < 0.001$	567	126	4.50

Next to the effects of single nucleotide changes, the much larger structural genomic variants are also likely to affect enhancer function. We previously determined CNVs for all ten strains investigated here using whole genome sequencing (WGS) read depth of coverage analysis (DOC) [27, 28]. With DOC, structural copy number changes (deletions and duplications)

are detected based on quantitative differences in the number of WGS reads. In total, 1,780 deletions and 1,176 duplications are present in one or more of the ten strains (Table 1). Because the identification of CNVs is less sensitive than the identification of SNVs, the false negative rate does not allow a similar association analysis as was performed for SNVs. Therefore, we took another approach to define the direct effects of CNVs on enhancer functionality. We analyzed the level of variation in H3K27Ac level among the strains for each enhancer, independent of the strain distribution. We find that 827 enhancers overlap 398 CNVs (266 enhancers overlap 156 deletions and 561 enhancers overlap 242 duplications). The CNVs that overlap enhancers result in an increased variance in H3K27Ac level across the ten strains, and this effect is more profound for deletions than for duplications (mean of all is -1.38, of deletions -8.83 (***t***-test; p-value $2.2*10^{-16}$) and duplications -1.17 (***t***-test; p-value $1.8*10^{-14}$)) (Fig. 2E and 2F). Similar to the analysis performed for single nucleotide variants, we also compared variance in H3K27Ac level for enhancers with CNVs that overlap TF binding sites, to enhancers with CNVs that do not target a TF binding site. For deletions that overlap enhancers and also TF sites, the change in H3K27Ac level is even more profound (mean -0.66, ***t***-test; p-value $2.2*10^{-16}$) (Fig. 2E and 2F). This shows that changes in TF binding due to CNVs affect the H3K27Ac state of the designated regulatory region more than for enhancers where TF binding is not affected.

Combined, we provide insight in the contribution of natural occurring genomic variation to epigenetic diversity between ten rat strains ***in vivo***. We find that not only SNVs, but also CNVs have a clear effect on the epigenetic differences between the strains, especially when they target TF binding sites. These data indicate that DNA sequence variants in regulatory elements are at least partially responsible for the observed epigenetic variation between inbred rat strains. However, it is difficult to assess the exact contribution of these genomic variants to the observed epigenetic variation and to determine what number of variable regulatory elements is not driven by a direct overlap with genomic variants.

Hi-C reveals a high degree of 3D chromatin organization conservation in different genetic backgrounds

We next investigated how genetic and epigenetic variation is reflected in the higher order 3D chromatin structure. To determine the contribution of chromatin organization to the differential distribution of regulatory elements between the studied rat strains, we performed Hi-C on liver tissue of two of the genetically most distinct rat strains, BN-***Lx*** and SHR. BN-***Lx*** and SHR are the founders of the HXB/BXH recombinant inbred panel and have been used for many genetic and eQTL studies [28-30]. The Hi-C experiments resulted in genome-wide information on chromatin conformation and interaction, showing many intra-chromosomal and few inter-chromosomal interactions (Fig. 3A).

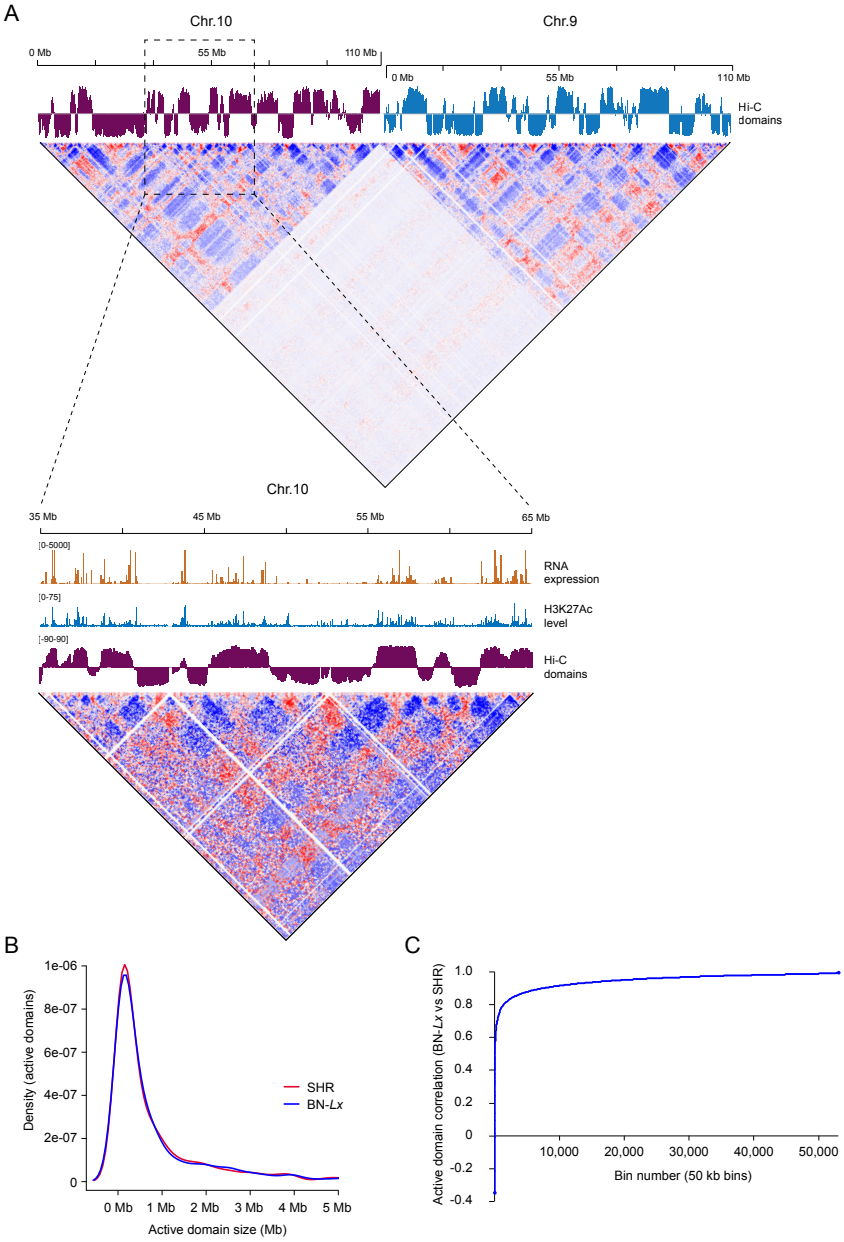


Figure 3 - **Hi-C results showing the similarities in higher order chromatin structure between BN-Lx and SHR**. (**A**) An example of active and inactive transcriptional domains determined by principal component analysis of chromosome 9 (right; blue) and 10 (left; purple) in the BN-Lx rat. Below the domains, inter- and intra-chromosomal interactions are displayed as occurring more (red) or less frequent (blue) than expected using a 100kb background model. The zoomed panel displays active and inactive chromatin domains (purple track), H3K27Ac levels (blue track) and RNA-seq expression data (orange track) for a 30 Mb region of chromosome 10. (**B**) Density plot showing the size distribution of active chromatin domains in BN-Lx (blue line) and SHR (red line). (**C**) Plot showing ranked correlation scores for 52,000 50-kb bins derived from the two-sample Hi-C comparison between BN-Lx and SHR.

Previous 4C and Hi-C studies have shown that the genome is spatially organized in active and inactive domains, and that these can be detected using principle component (PCA) analyses. We could identify 940 (BN-*Lx*) and 991 (SHR) active domains at a 50-kb resolution (Materials and methods). In total, the active domains make up 1,280 Mb of the BN-*Lx* genome and 1,282 Mb of the SHR genome, both equaling approximately half of the rat genome. As expected the active domains contain the majority of active genes and enhancers (83% and 80% respectively). Between both strains, we not only observe high overlap in the number of active domains (940 vs. 991) and the total genome that they cover (~50%), but also in the size distribution of the domains (median = 350 kb for both strains) (Fig. 3B). Nonetheless we do find differences between the strains. We find 93 active domains larger than 100kb that are unique to BN-*Lx* and 90 that are unique to SHR. Also, domains that are present in both strains can still be different in size. In total, we find 704 regions that are shared between BN-*Lx* and SHR. When we analyze the 1,408 edges of these overlapping regions (start and end sites combined), we find that 44 and 39 edges have shifted in SHR and BN-*Lx* respectively, requiring a domain increase of at least 100 kb.

In addition to separating the genome in domains, Hi-C profiles between two strains can be compared directly for local interactions per window of 50 kb. This results in correlation scores independent of the previous principle component analysis that separates active from inactive domains. Based on these correlations, we again find a high degree of similarity between both strains; 95% correlates with over 0.84, and only 3% drops below 0.80 (totaling 71.4 Mb with $r \leq 0.8$ or 2.55 Mb with $r \leq 0.5$) (Fig. 3C).
The high resemblance in 3D chromatin organization between the two strains is in line with high domain conservation found in other organisms and between species [31]. Nevertheless, not all regions correlate well and the genomic regions that do deviate could be related to epigenetic differences between BN-*Lx* and SHR.

3D chromatin changes correspond with epigenetically differential regions

To investigate to what extent differences at the epigenetic level correlate with the 3D chromatin level, we first selected the most differential enhancers between BN-*Lx* and SHR and then investigated how these corresponded with differences at the 3D level. An enhancer was termed differential based on the H3K27Ac level variation between all ten rat strains (FDR < 0.05; log(average CPM) > 2), resulting in 208 enhancers that are differential between BN-*Lx* and SHR.
For these 208 differential enhancers, we first determined how many can be directly connected to genomic variants. In total, the genomes of BN-*Lx* and SHR possess 116,996 SNVs and 130 CNVs that overlap with an enhancer. Only 13 differential enhancers show direct overlap with deletions and 2 overlap with duplications. The effects of these CNVs on H3K27Ac levels are

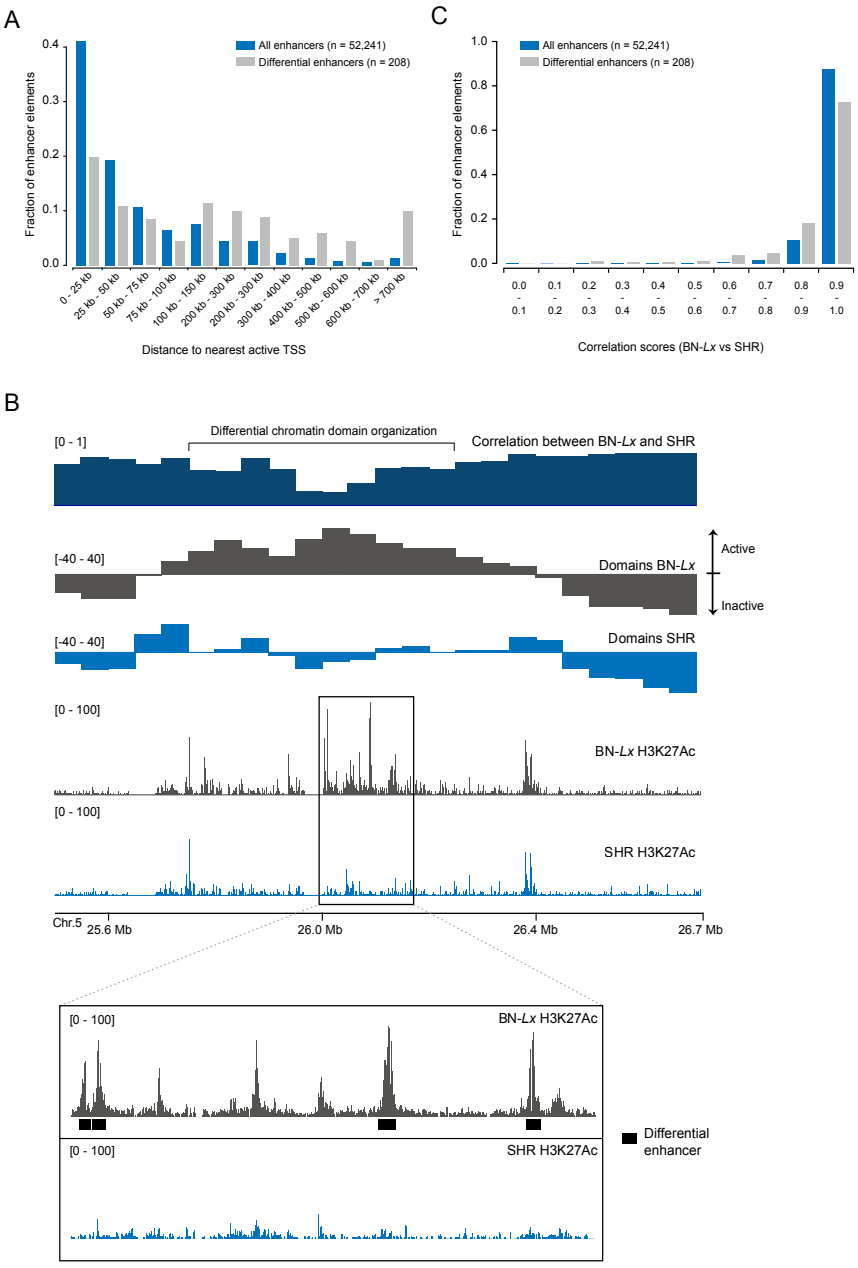


Figure 4 - **Differential chromatin domain organization results in differential enhancers between BN-Lx and SHR**. (**A**) Bar plot showing the distance of all enhancers (blue bars) and the 208 differential enhancers (grey bars) to the nearest active TSS as determined by RNA-seq. (**B**) Example of differential enhancers that co-cluster in a chromatin domain that is differentially active between BN-Lx and SHR. The zoomed region shows 4 enhancers that are called differential between the two strains. Also, enhancers are visible that show quantitative differences but are not significantly differential between the two rats. (**C**) Bar plot with correlation scores between BN-Lx and SHR for the 50 kb bins that contain differential (grey bars) or all enhancers (blue bars).

clear, with decreasing H3K27Ac levels for deletions and increasing levels for duplications (Supplementary fig. 3). Furthermore, we find that 164 differential enhancers possess one or more SNVs. However, only 47 show a significant correlation between the SNV and H3K27Ac variation ($p < 0.05$), as tested in the SNV-H3K27Ac analysis performed on all ten rats (Fig. 2A and B). In total, we can thus associate 62 out of the 208 differential enhancers with genomic variants.

Strikingly, of all 208 differential enhancers, we find almost half (n = 97) back in inactive chromatin domains, whereas we find only one-fifth of the complete set of enhancers back in inactive domains. This suggests that differential enhancers are more often located in inactive domains. In line with this, we mostly find differential enhancers at sites distal from the nearest active gene promoter, residing in relatively gene-poor regions (Fig. 4A). The depletion of differential enhancers from gene-dense regions could reflect different selection pressure on regulatory elements in gene poor regions, resulting in an enrichment of differential enhancers. Interestingly, differential enhancers also frequently co-cluster in domains (Fig. 4B). 42 out of the 111 differential enhancers in active domains (38%) map to only 18 out of the 941 active BN-***Lx*** domains (2%).

We next studied if the differential chromatin folding as determined by Hi-C could be connected to the presence of differential enhancers. By first focusing on the strain-specific domains, we find that four of the BN-***Lx*** specific domains and three of the SHR-specific domains contain a differential enhancer. Next, we sought for differential enhancers at the previously determined shifted domain edges and find two examples (e.g. Supplementary fig. 4A), plus many more enhancers with quantitative differences that do not meet the threshold for being termed differential.

Finally, we inspected the genomic regions that correlate the least between BN-***Lx*** and SHR for differential enhancers. In general, we find that the genomic regions that possess differential enhancers have much lower correlation scores compared to the genome-wide distribution of all bins that possess enhancers (Fig. 4C). Many of those regions display a reversed domain activity (i.e. from active to inactive or vice versa) (Supplementary fig. 4B). In total, 24 out of the 208 differential enhancers are located in chromatin regions with correlation scores below 0.8, meaning that they are in the top 3% of most differentially interacting regions when comparing BN-***Lx*** and SHR. The 2.55 megabases in the rat genome with a correlation below 0.5 consist of 35 of such regions, of which 2 contain multiple differential enhancers (e.g. Fig. 4B). On top of that, out of the 11 low-correlating regions that contain enhancers, 4 possess enhancers that are at least two-fold different between BN-***Lx*** and SHR but do not meet the threshold for being significantly differential (e.g. Fig. 4B). Low correlations indicate that chromatin interactions have changed in regions with differential enhancers, which can indeed be observed as differences in interaction profiles (Supplementary fig. 5).

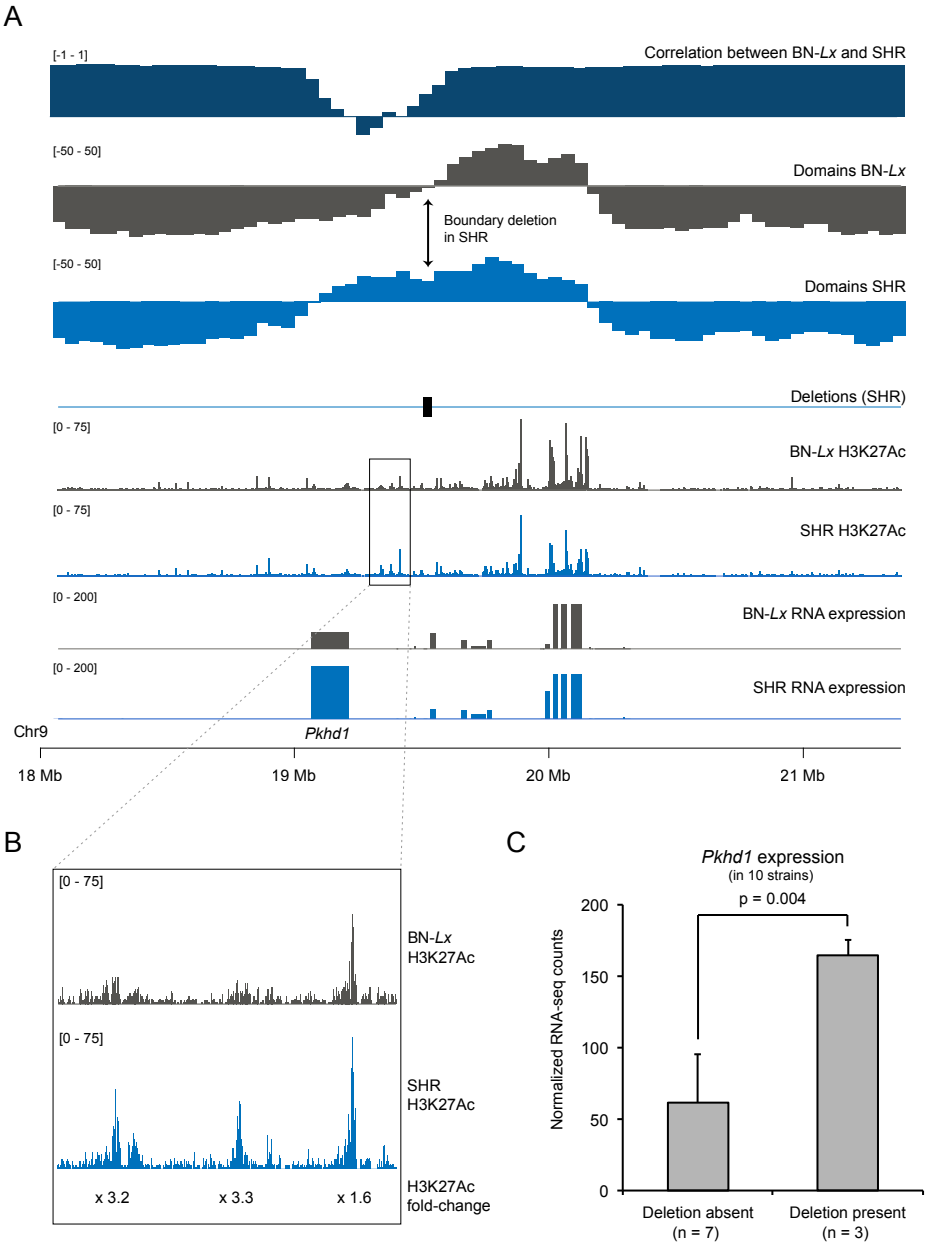


Figure 5 - **Extended domain activity and increased Pkhd1 expression due to a deletion in SHR**. (**A**) Illustrative example of deletion in SHR overlapping what is domain boundary in BN-Lx, resulting in the extension of an active domain in SHR. In the differential part of this domain we observe a ~3-fold increase in the H3K27Ac level of three enhancers and a corresponding ~3-fold increase in expression of the nearby gene Pkhd1. (**B**) Zoomed display of the three elevated enhancers in SHR (blue) compared to BN-Lx. Fold-change is determined based on the normalized H3K27Ac read counts per enhancer. (**C**) Bar plot showing the gene expression levels of the Pkhd1 gene in all ten rat strains that do (n = 3) or do not carry the deletion (n = 7). Error bars represent the standard error of the mean.

Combined, we find a strong correlation between chromatin domain organization and the distribution of regulatory elements, providing a connection to the differential enhancers. Domains with low correlation between BN-*Lx* and SHR possess many (frequently co-localizing) differential enhancers indicating that chromatin organization is of major influence to regulatory element activity, as measured by H3K27Ac levels.

Cause or effect: A genomic basis for differential chromatin organization

Finally, we determined the relationship between genomic variation and higher order chromatin structure. Effects of genomic variation on 3D folding could be established via altered enhancer or transcriptional activity, resulting in a different domain structure. However, genomic variants may also influence chromatin folding directly. For example, the disruption of domain boundaries can drive ectopic chromosomal contacts resulting in long-range deregulation of gene expression [17]. The most likely candidate events to disrupt domains are genomic deletions of elements that define domain boundaries.

To test the contribution of genomic variation to the 3D organization, we first determined the distribution of genomic variants from BN-*Lx* and SHR over the identified chromatin domains. For both SNVs and CNVs we find a varying distribution over active and inactive domains. In active domains, SNVs show densities of 0.88 SNVs/kb whereas this density is 1.20 SNVs/kb in inactive domains. We thus observe lower numbers of SNVs in inactive chromatin domains. The CNV distribution is even more shifted towards inactive domains. Of all 470 deletions, only 128 (27%) map to active domains and 342 (73%) map to inactive domains. For the 408 duplications, these ratios are identical: only 109 duplications (27%) map to active domains and 299 duplications (73%) map to inactive domains. If we focus on the identified strain-specific domains of at least 100 kb (93 for BN-*Lx* and 90 for SHR), we see that three of the SHR domains contain deletions but none contain genomic duplications. In BN-*Lx*, 2 strain-specific domains contain deletions and 4 possess duplications. Also, in the 35 regions that have low interaction correlations ($r \leq 0.5$), six contain deletions and 3 contain duplications.

Of the 83 shifted domain ends previously identified between BN-*Lx* (n = 39) and SHR (n = 44), 5 overlap a deletion and 3 overlap a duplication. If we zoom into these regions, we find active domains that are extended or restricted (Supplementary fig. 6). In one example, we see that the range of an active domain in SHR is extended by 450 kb, most likely as a result of a deletion (Fig. 5A). This extended active domain in SHR corresponds to an increase in H3K27Ac level of three enhancers present in the extended domain (Fig. 5B). Also, we observe higher expression of the *Pkhd1* gene, which is the only expressed gene in the differential part of this domain. To define if this particular deletion, which is at a 155 kb distance of the gene, could be causal for the increased expression of *Pkhd1* in the SHR strain, we determined the strain

distribution of the deletion and the expression of ***Pkhd1*** in all ten rats initially examined in this study. This shows that every strain that caries the deletion allele, also displays increased expression of ***Pkhd1*** (Fig. 5C). This makes it very likely that the deletion is indirectly causal for the increased expression of ***Pkhd1***.
These results suggest that natural occurring structural variations can indeed influence higher order chromatin domain organization by disturbing chromatin domain boundaries. In addition to previously described effects of CNVs and SNVs on that directly overlap regulatory elements, we now also show that CNVs that do not directly overlap regulatory elements can modulate gene expression regulation.

Discussion

In this study we analyze the epigenetic diversity between ten rat strains and search for the basis of this diversity in the genetic background and the 3D chromatin organization. Using ChIP sequencing, we profile enhancer and promoter elements and study the conservation of enhancers and promoters across the ten strains. We find that levels of enhancers are more variable than promoters and show that the genetic background contributes to differences in enhancer activity as determined by quantitative differences in H3K27Ac levels. We find enrichment for SNVs that associate with differential H3K27 acetylation, but also clearly demonstrate the poorly studied effects of CNVs on enhancers. Transcription factor binding of the liver-specific transcription factors FOXA1, CEBPA and HNF4A occurs mainly at the identified elements and genomic variants that target these binding positions affect enhancers more heavily than variants that do not overlap TF binding sites.

Next, we employed Hi-C on the BN-***Lx*** and SHR strains to define the effect of chromatin organization on epigenetic differences such as the presence or absence of enhancers. In agreement with previous reports [31], we generally find high conservation in 3D organization. We observe only few regional differences in chromatin domains, but these regions do appear to be related to the differential enhancers. More often than expected, these enhancers are located in single domains that are frequently different in activity between BN-***Lx*** and SHR.
Interestingly, the Hi-C data also allows us to test the effect of noncoding genomic variation on chromatin domain organization. In particular, we observe that CNVs, which are generally depleted from active domains, can affect domain boundaries resulting in the extension or reduction of domain activity. For example, extended domains can alter enhancer activity and gene expression, as is the case for the ***Pkhd1*** gene and three enhancers in an extended domain present in SHR but not BN-***Lx*** (Figure 5).

From the many GWAS efforts published to date, we know that the majority of disease-linked alleles map to the noncoding genome [3-5], and our data provide insight in the potential mechanisms that noncoding variants can use to contribute to common disease. We not only

find SNVs and CNVs to directly influence regulatory element activity (e.g. enhancer gains or losses), but we also find variants that likely display long-range deregulating effects by changing the 3D chromatin morphology. This indicates that genomic variants in the noncoding genome can indirectly contribute to molecular phenotypic changes by affecting chromatin architecture. Such genomic variants are frequently being overlooked in studies aiming to find the basis of disease, because they do not directly overlap protein-coding genes making it difficult to assess the consequences.

Conclusions

The findings in this paper highlight the diversity of mechanisms that noncoding single nucleotide and structural genomic variants use to affect genome function ***in vivo***. Using an integrative approach with information from epigenetic profiling and whole genome sequencing, we assess the molecular effects of these variants on multiple levels of regulation and illustrate how noncoding variants could contribute to complex disease. Our results provide insight in the complexity of the noncoding genome and the widespread phenotypic consequences of structural and single nucleotide variation.

Materials and methods

Liver tissue collection. We obtained snap frozen liver tissues of 6 weeks old animals from 10 rat strains. Liver tissues from six out of ten strains was kindly provided to us by: dr. James D. Shull (University of Wisconsin, Madison): ACI/SegHsd; dr. Myrna Mandel (NIH - Office of Research Services): M520/N, MR/N and WN/N and dr. Michal Pravenec (Charles University, Prague): BN-***Lx***/Cub and SHR/OlaIpcv. Tissues from F344/NHsd, WKY/NHsd, BN/SsNOlaHsd and BUF/SimRijHsd were purchased from Harlan Laboratories. If tissue material from the original founder animals was no longer available, we chose the most closely related substrain based on similarity in genome sequence.

Preparation of cross-linked cell nuclei for ChIP and Hi-C. ~40 mg snap frozen and powdered rat liver tissue was resuspended in 2 mL cold PBS-10% FCS and dissociated using a 40 µm Nylon Cell Strainer (BD Biosciences). The cell suspension was cross-linked with 2% formaldehyde in a total volume 10 mL PBS/10%FCS for 10 minutes rotating at RT (20°C). 0.125 M glycine was added to quench the reaction and cells were stored on ice. Following the cross-linking procedure, samples were centrifuged for 8 minutes at 400 g (4°C). Pelleted cells were washed with 1 mL cold PBS and centrifuged at 400 g, 4°C for 5 minutes again. After removal of the supernatant, the cell pellet was dissolved in 1mL freshly prepared lysis buffer to recover cross-linked nuclei (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM EDTA, 0.5% NP-40, 1% Triton X-100 and 1x cOmplete, EDTA-free Protease Inhibitor Cocktail (#11873580001, Roche Applied Sciences, Indianapolis, IN, USA). Cells were lysed for an initial 10 minutes on

ice and cell lysis was determined to be complete using Methyl Green - Pyronin staining. After completion of lysis, nuclei were washed twice in cold PBS.

Chromatin Immunoprecipitation, library preparation and sequencing. Cross-linked nuclei were dissolved in 100µL lysis buffer (MAGnify system, Invitrogen™) with 1x protease inhibitors (MAGnify system, Invitrogen™) and sheared in microtubes (AFA Fiber Pre-Slit Snap-Cap 6x16mm, 520045) using the Covaris S2 sonicator (6 cycles of 60 seconds; duty cycle: 20%, intensity: 3, cycles per burst: 200, frequency sweeping). Soluble chromatin equivalent to ±20 mg input liver tissue, in a size range of 150 - 300 bp, was used for immunoprecipitation (IP). H3K4me3 and H3K27Ac IPs were carried out using the MAGnify system (Invitrogen™, 49-2024) following manufacturers instructions (Invitrogen manual A11261). Per IP, 1 µg of antibody was used (H3K4me3: Millipore, 07-473 LOT# JBC1863338 - H3K27Ac: Abcam, ab4729 LOT# 1415784).
For library preparation, chromatin immunoprecipitated DNA was sheared to ± 100 bp in size using Covaris S2 (microtubes, 6 cycles of 60 seconds with duty cycle: 10%, intensity: 5, cycles/burst: 100, frequency sweeping). SOLiD 5500XL Wildfire fragment library preparation was done according to manufacturer's instructions. Libraries were sequenced on SOLiD 5500XL Wildfire resulting in 40-bp reads.

Calling enhancer and promoter regions. Sequencing reads were mapped using Burrows-Wheeler Aligner (BWA-0.5.8c) (settings: –c –l 25 –k 2 –n 10) [32]. This resulted in 67-71 million mapped read per sample for the H3K27Ac ChIP and 14-21 million mapped reads per H3K4me3 ChIP. Peak calling was done using MACS [33] (version 1.4, settings: -g 2718897334 -B -S --to-small -p 1e-10 bandwidth=300 model=TRUE shiftsize=100), using chromatin input DNA as control.
Called peaks for H3K4me3 and H3K27Ac were first processed separately. Peak regions of all strains were merged using the mergeBed command from the BEDtools suite [34], resulting in one peak set per histone mark. Overlap between H3K27Ac and H3K4me3 peaks was determined and H3K27Ac peaks that did not overlap an H3K4me3 peak were assigned as enhancers, those that did overlap an H3K4me3 peak were assigned as promoters. Strain specificity for each enhancer and promoter was determined by checking which strain possessed an H3K27Ac or H3K4me3 peak overlapping an element in the final merged set.
H3K27Ac levels were determined by counting the number of sequencing reads per strain that overlapped the enhancers and promoters in the final set, using the coverageBed command of the BEDtools suite [34]. Read counts were normalized to the total number of reads that mapped to enhancers or promoters (see expression analyses). Enhancer and promoter regions were normalized separately.

SNV calling. To make a comprehensive comparison between the rat strains we applied

multi-sample SNVs calling using GATK [35]. SNVs were determined using raw data first described in [27, 28]. Nine of the ten strains (ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WN/N, WKY/N and BN-*Lx*/Cub) were sequenced on SOLiD™ 5500, and they were analyzed simultaneously, using multi-sample SNV calling by the haplotype caller in GATK. The LE rat strain was included to determine specificity, and SNP array data was used to determine sensitivity, as described previously [27]. Based on BAC sequences available for the LE strain, SNV calls are 99% specific. Based on SNP-array data, SNVs are 99.5% sensitive.
SHR/OlaIpcv could not be included in the multi-sample calling, because this is the only strain sequenced on the Illumina platform. SNVs in SHR/OlaIpcv were called using GATK UnifiedGenotyper. Based on SNP-array data SNVs were filtered for being 99% sensitive. This is slightly less than the other nine strains, because we cannot take advantage of multi-sample calling here to make the calls more sensitive. SNVs for all ten strains were then merged into one VCF file.

Association of H3K27Ac levels with SNVs and CNVs. The association analysis was performed using SNVs that had a homozygous SNV or a reference call from the GATK SNV analysis (either a '0/0' or '1/1' in the SNV list), for each of the strains. SNVs that were noisy in at least one of the strains (a '0/1' or '.' in the SNV list produced with GATK) were removed from the set. Only SNVs that were present in at least three and at most seven strains were analyzed, because these separate the ten strains in two groups of at least three strains, such that a student's ***t***-test could be performed. The final set used for all SNV-related analyses in this study consisted of 2,084,592 SNVs. Student's ***t***-tests were performed using the t.test function in R; two-sided testing.
Intersections between CNVs and enhancers were determined using the intersectBed command form the BEDtools suite [34].

Hi-C: massive parallel sequencing of proximity-based ligation products. Isolated and cross-linked liver nuclei of three BN-*Lx* and SHR animals were digested with the DpnII restriction enzyme (#R0543, NEB, Ipswich, MA, USA). Subsequently, proximity ligation of interacting fragments was performed using T4 DNA ligase (#10799009001, Roche Applied Sciences) to produce 3C template according to Simonis et al [36]. After reverse cross-linking and precipitation, 10 µg template was sheared in microtubes (AFA Fiber Pre-Slit Snap-Cap 6x16mm, 520045) using the Covaris S2 sonicator (1 cycle of 25 seconds; duty cycle: 5%, intensity: 3, cycles per burst: 200, frequency sweeping). Fragments between 500-1500bp were selected using a 2% agarose gel. 1.1 µg of size-selected fragments was used as input for the TruSeq DNA Low Sample (LS) protocol (Illumina). Constructed libraries were size-selected using a LabChip XT DNA 750 Assay Kit (Caliper), resulting in libraries between 800-950 bp. These libraries were sequenced in a paired-end manner on the Illumina HiSeq 2500, resulting in 2x100-bp reads. Sequenced read pairs were mapped using Burrows-Wheeler Aligner

(BWA-0.5.8c) (settings: –c –l 25 –k 2 –n 10) [32], yielding 70 million (M) mapped reads per animal (totaling 210 M mapped reads per strain).

Exploring 3D chromatin organization in two rat strains using Hi-C. Hi-C data were analyzed using Homer [37]. Sequenced read-paired were filtered to have minimum distance of 1.5 kb to omit self-ligated fragments. Full filter settings using Homer makeTagdirectory are: -update -removePEbg -fragLength 1500 -removeSpikes 10000 5. This resulted in 60 M usable read pairs per strain. PCA analyses and correlation differences between the strains were calculated using window-sizes of 100 kb, with a step size of 50 kb (a 100kb super-resolution and a 50 kb resolution in Homer). Background models were generated to normalize the data, using the same window sizes.

RNA library preparation, sequencing and analysis. ~40 mg of the same snap frozen and powdered liver tissue was used for total RNA isolation from purified nuclei using the TRIzol® reagent (#15596-026, Invitrogen, Life Technologies). RNA-seq libraries were prepared from rRNA-depleted RNA (Ribo-Zero™ Magnetic Gold Kit for Human/Mouse/Rat (MRZG12324, Epicentre®, Madison, WI, USA)) using the SOLiD™ Total RNA-seq kit (#4445374, Life Technologies). All libraries were sequenced on the SOLiD™ 5500 Wildfire system (40 bp fragment reads). RNA-seq reads were mapped using Burrows-Wheeler Aligner (BWA-0.5.9) (settings: -c -l 25 -k 2 -n 10) onto the rat reference genome RGSC3.4. Only uniquely mapped, non-duplicate reads were considered for further analyses. Reads that mapped to exons were used to determine the total read counts per gene. Exon positions were based on the Ensembl 56 annotation. Read counts per gene (K) for each sample (X) were normalized to the dataset with the lowest number of reads (sample Y), in the following manner: normalized_read_counts_geneK_sampleX=int(read_counts_geneK_sampleX*(total_number_of_reads_mapped_to_exons_in_sampleY/total_number_of_reads_mapped_to_exons_in_sampleX)).

Authors' contributions

SvH, RH, NL and KdL performed wet lab experiments (ChIP-seq, RNA-seq and Hi-C). MS and SvH analyzed the ChIP-seq, RNA-seq and Hi-C data. EdB and MV performed next-generation sequencing. WS and SB performed next-generation sequencing mapping and SNP calling. PK, GG and WdL contributed to the Hi-C library construction and data analysis. DT and PF contributed to scientific discussions regarding the ChIP-seq experiments and transcription factor overlaps. SvH, RH, EC and MS conceptually designed the study, coordinated experiments, critically discussed results and wrote the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgments

This work was financially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. HEALTH-F4-2010-241504 (EURATRANS and the NWO-CW TOP grant (700.58.303) to EC. MS acknowledges funding from the NWO Vernieuwingsimpuls program (grant number 863.10.007). We are grateful to dr. James D. Shull (University of Wisconsin, Madison), dr. Myrna Mandel (NIH - Office of Research Services) and dr. Michal Pravenec (Charles University, Prague) for kindly providing us the liver tissues.

References

1. McCarthy MI, Hirschhorn JN: Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 2008, 17:R156-165.

2. Stranger BE, Stahl EA, Raj T: Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 2011, 187:367-383.

3. Pennisi E: The Biology of Genomes. Disease risk links to gene regulation. Science 2011, 332:1031.

4. Kumar V, Wijmenga C, Withoff S: From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. Semin Immunopathol 2012, 34:567-580.

5. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009, 106:9362-9367.

6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

7. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: The sequence of the human genome. Science 2001, 291:1304-1351.

8. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: An integrated encyclopedia of DNA elements in the human genome. Nature 2012, 489:57-74.

9. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al: Unlocking the secrets of the genome. Nature 2009, 459:927-930.

10. Ward LD, Kellis M: Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol 2012, 30:1095-1106.

11. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 2007, 39:311-318.

12. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al: Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 2010, 107:21931-21936.

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

14. Nolis IK, McKay DJ, Mantouvalou E, Lomvardas S, Merika M, Thanos D: Transcription factors mediate long-range enhancer-promoter interactions. Proc Natl Acad Sci U S A 2009, 106:20222-20227.

15. Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, et al: Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell 2013, 154:530-540.

16. Voss TC, Hager GL: Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat Rev Genet 2014, 15:69-81.

17. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, et al: Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 2012, 485:381-385.

18. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009, 326:289-293.

19. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, et al: CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet 2011, 43:630-638.

20. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, et al: Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science 2013, 342:744-747.

21. Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK: Effect of natural genetic variation on enhancer selection and function. Nature 2013, 503:487-492.

22. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV, et al: Extensive variation in chromatin states across humans. Science 2013, 342:750-752.

23. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK: Identification of genetic variants that affect histone modifications in human cells. Science 2013, 342:747-749.

24. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: A method and server for predicting damaging missense mutations. Nat Methods 2010, 7:248-249.

25. Ng PC, Henikoff S: SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003, 31:3812-3814.

26. Weischenfeldt J, Symmons O, Spitz F, Korbel JO: Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 2013, 14:125-138.

27. Baud A, Hermsen R, Guryev V, Stridh P, Graham D, McBride MW, Foroud T, Calderari S, Diez M, Ockinger J, et al: Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nat Genet 2013, 45:767-775.

28. Simonis M, Atanur SS, Linsen S, Guryev V, Ruzius FP, Game L, Lansu N, de Bruijn E, van Heesch S, Jones SJ, et al: Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. Genome Biol 2012, 13:r31.

29. Heinig M, Petretto E, Wallace C, Bottolo L, Rotival M, Lu H, Li Y, Sarwar R, Langley SR, Bauerfeind A, et al: A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature 2010, 467:460-464.

30. Low TY, van Heesch S, van den Toorn H, Giansanti P, Cristobal A, Toonen P, Schafer S, Hubner N, van Breukelen B, Mohammed S, et al: Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep 2013, 5:1469-1478.

31. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012, 485:376-380.

32. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754-1760.

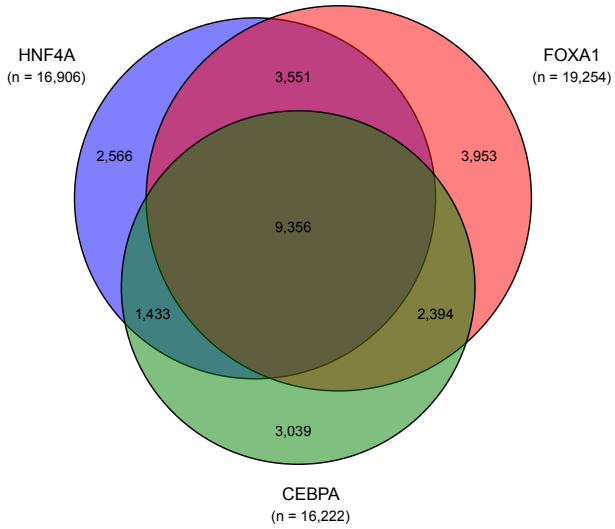
33. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008, 9:R137.

34. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010, 26:841-842.

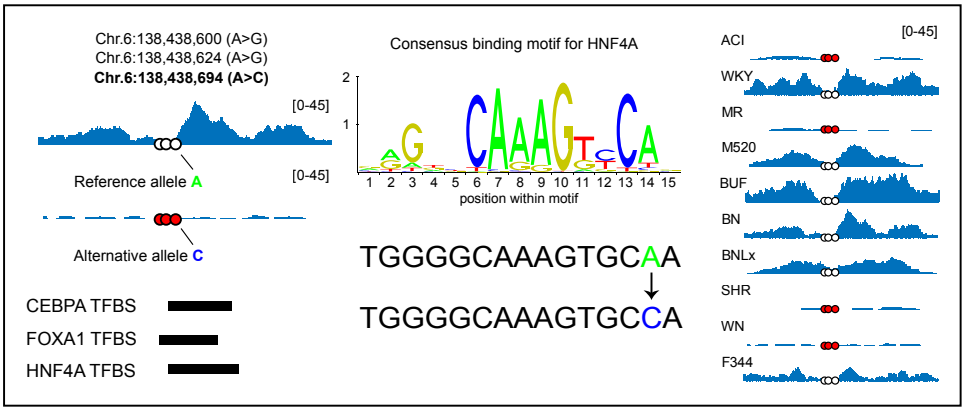
35. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010, 20:1297-1303.

36. 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). Nat Genet 2006, 38:1348-1354.

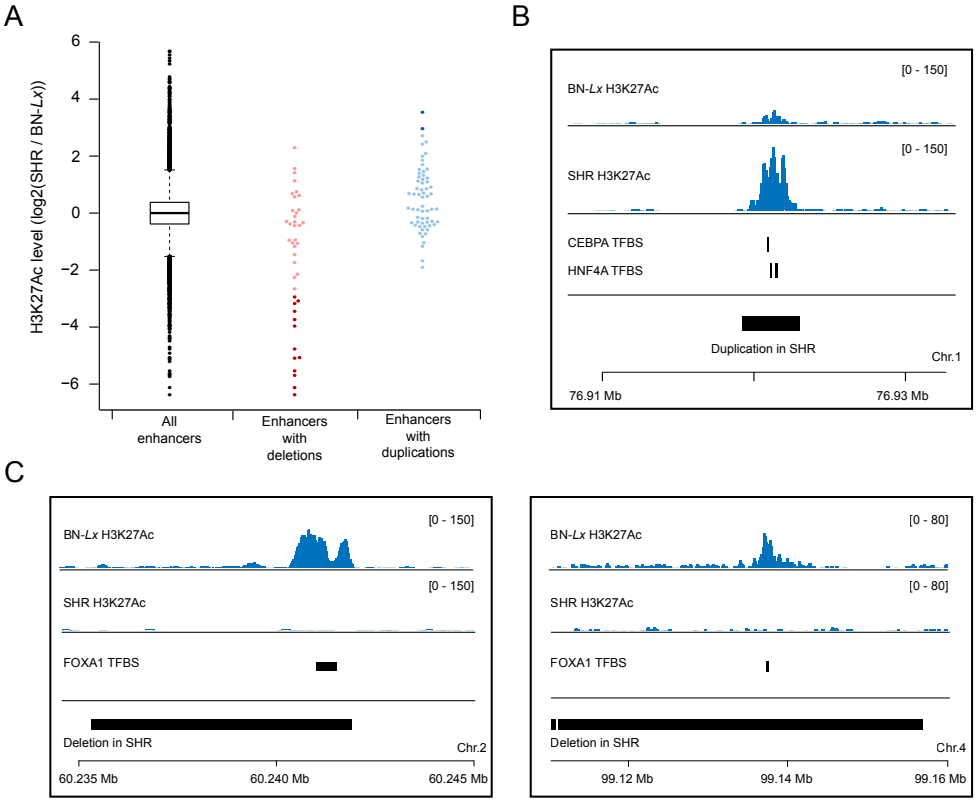
37. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell 2010, 38:576-589.



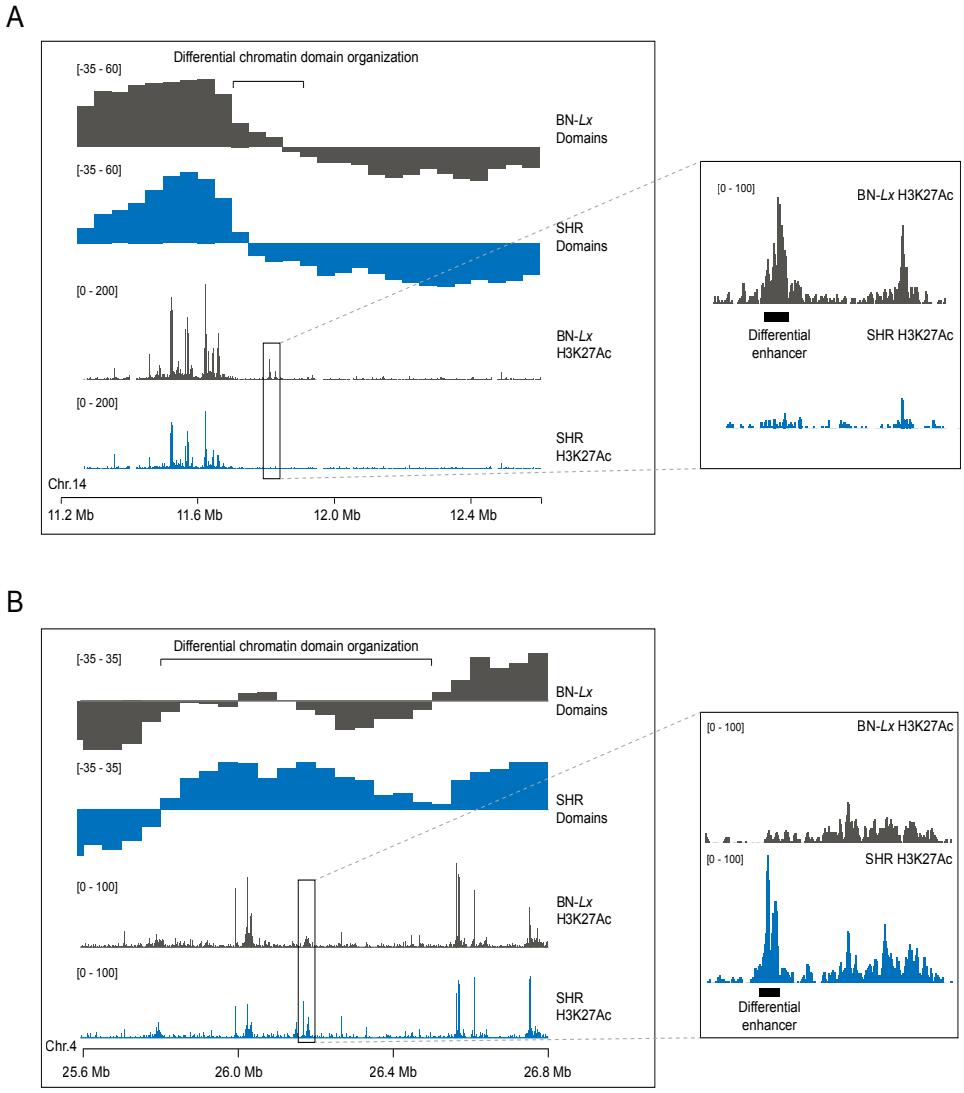
Supplementary figure 1 - **Three-way Venn diagram showing the overlap of liver TF-bound enhancers.** Approximately half of the liver TF-bound enhancers (n = 9,356) bind HNF4A (blue), FOXA1 (red) and CEBPA (green).



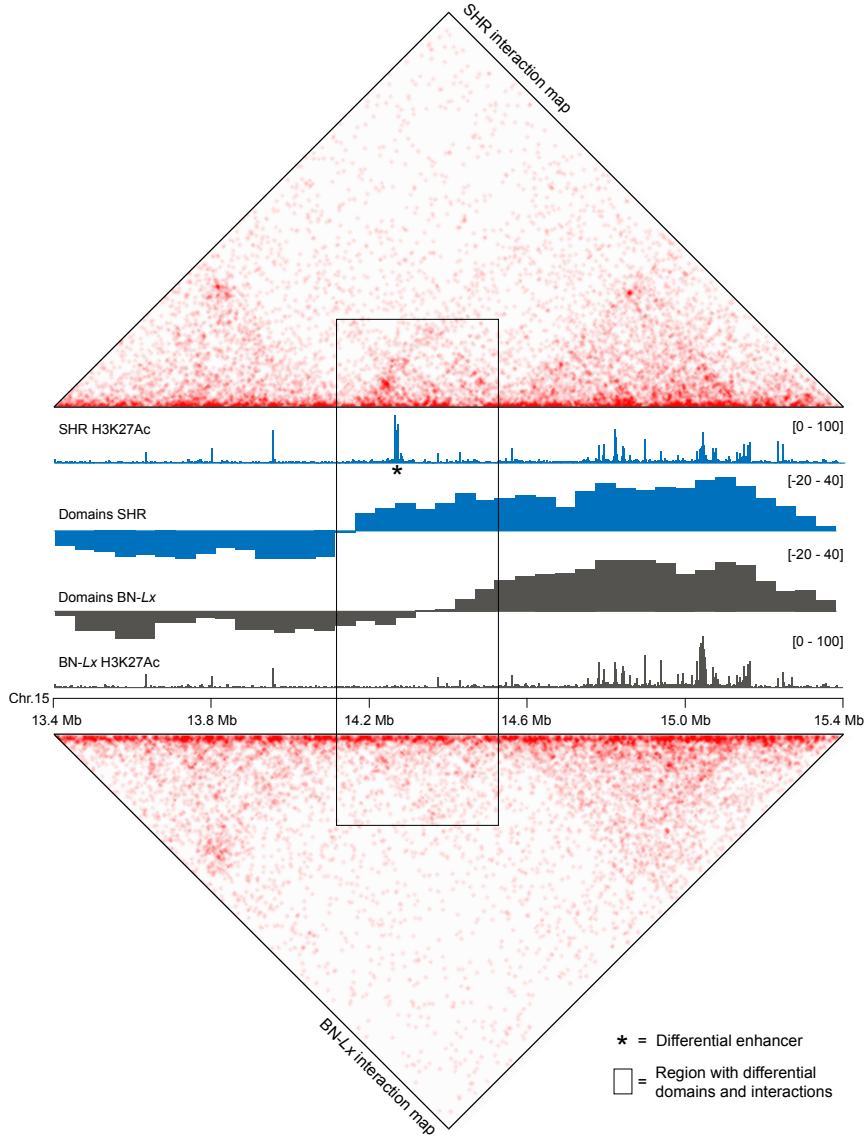
Supplementary figure 2 - **SNVs affect H3K27Ac levels of enhancers by disrupting TF binding sites**. Example of an SNV on chromosome 6 which perturbs consensus binding motif for HNF4A, resulting in decreased H3K27Ac levels (blue).



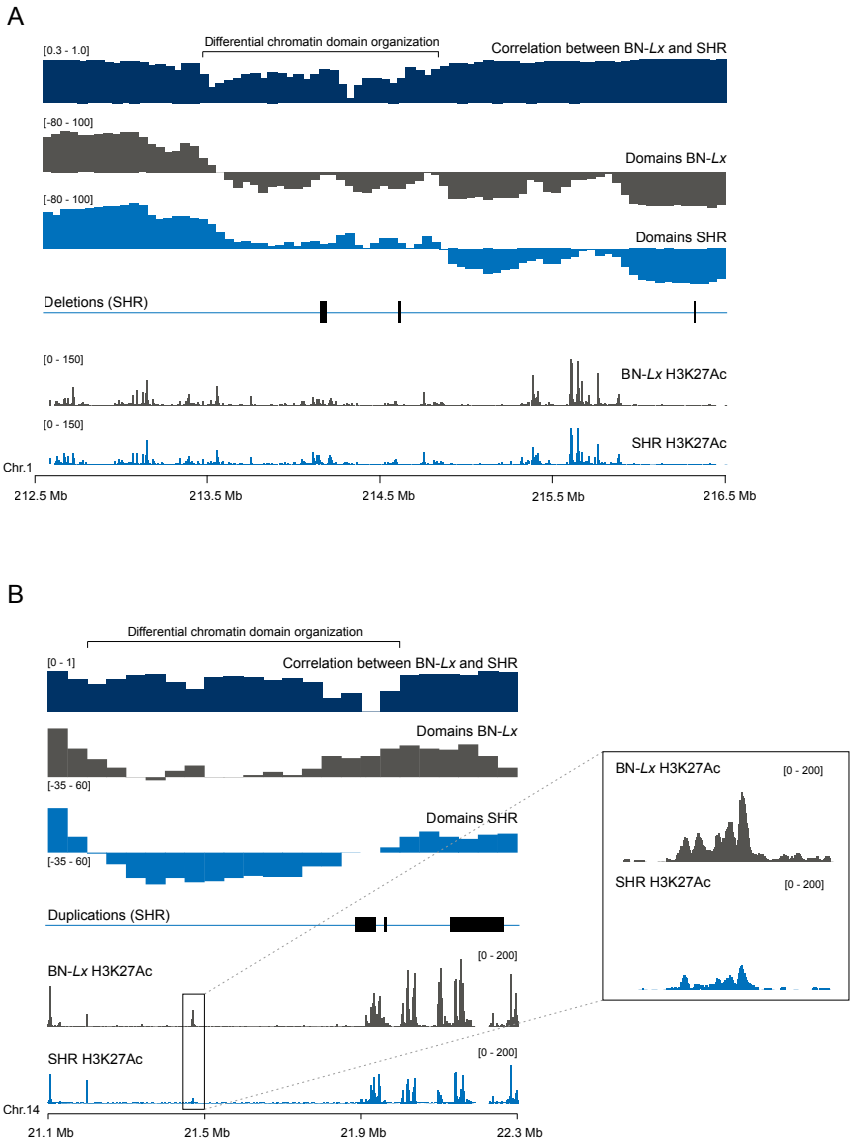
Supplementary figure 3 - **Structural genomic variants modulate enhancers in BN-*Lx* and SHR rats.** (**A**) Boxplot with H3K27Ac levels of enhancers overlapping a deletion or duplication in BN-Lx or SHR. Differential enhancers overlapping deletions or duplications are highlighted in dark red (deletions) and dark blue (duplications). The mean log2(SHR/BN-*Lx*) of H3K27Ac is -0.007 when all enhancers are taken into account. This decreases to -2.25 in enhancers with a deletion in SHR (*t*-test, p-value<2.8 *10e-7) and increases to 0.56 in enhancers with a duplication in SHR (*t*-test, p-value < 7.5*10e-5). (**B**) Example of a duplication in SHR overlapping a differential enhancer containing a CEPBA and HNF4A binding site. (**C**) Two examples of deletions in SHR that overlap a differential enhancer that is bound by FOXA1.



Supplementary figure 4 - **Differential chromatin organization result in differential enhancers**. (**A+B**) Two examples of differential enhancers located in differentially organized chromatin domains between BN-*Lx* (dark grey) and SHR (blue). The zoomed regions highlight the differential enhancer in both strains.



Supplementary figure 5 - **A differential enhancer in an extended SHR domain on chromosome 15 creates novel cis-interactions captured by Hi-C**. Hi-C derived interaction plots for SHR (top) and BN-*Lx* (bottom) at 20 kb resolution. The higher intensity (darker red) in the interaction map represents an increase in interactions, which is most apparent within the rectangle near the newly gained enhancer (marked by an asterisk).



Supplementary figure 6 - **Deletions and duplications that overlap domain boundaries appear to affect the higher order chromatin organization**. (**A**) Example of a region with multiple deletions that appear to change the chromatin domain landscape. The deletion positioned most to the left overlaps a domain boundary in the SHR strain and appears to alter it. Two other SHR deletions in the region appear to have less effect, although the middle one also resides in a domain that is opposite between the strains. The general correlation between the two Hi-C experiments is depicted on top (dark blue bars), showing interaction correlations that decrease to 0.30. (**B**) Identical to (**A**), but this time showing duplications in SHR (blue), of which one appears to initiate an inactive domain in what is active chromatin in BN-*Lx*. This results in reduced enhancer activity in SHR (lower H3K27Ac levels), highlighted in the zoomed panel on the right. Again, the independent correlation scores are displayed on top (dark blue bars), showing correlation drops to almost 0 at the position of the duplication.

8

Summarizing discussion

Interpreting genomes

Personalized genomics and medicine, the thousand-dollar genome and population scale genomics are all goals introduced nearly a decade ago that now are or will soon become reality. Technological improvements in mainly the DNA sequencing industry have resulted in a dramatic increase in sequencing throughput and possibilities, together with a steep decrease in costs and man-hours required to produce enormous amounts of sequencing data. Together with these almost unlimited sequencing possibilities also came along the challenges of large-scale data interpretation and integration. With most technological hurdles overcome, reliable interpretation of sequencing information remains challenging and one can wonder if we are currently sufficiently capable of correctly interpreting genomics data and are really ready for responsible use of such information for personalized genomics. Of the challenges that remain in the genomics field, most can be placed in one of the following categories: (i) high quality data generation / technological issues, (ii) data interpretation (iii) (multidisciplinary) data integration and (iv) determining genotype-phenotype relations. In this thesis, I have countered several of those challenges and I will further elaborate on them in this summarizing discussion.

Technological challenges in genomics approaches

For whole genome DNA sequencing and interpretation, it is crucial to obtain equal sequencing read coverage throughout the genome. Especially for reliable quantitative interpretation of DNA to localize genomic variants, a DNA sequencer needs to read a genome equally well at every genomic location. In **Chapter 2** of this thesis, I describe variation in genome-wide coverage between the different tissues of a single individual, which could be interpreted as tissue-specific somatic copy number variation (CNV). However, we find that the source material for these analyses systematically affects the genome-wide coverage. By varying the proteinase treatment conditions prior to DNA extraction, we discovered that genome coverage patterns could be tweaked, suggesting that chromatin state influences DNA isolation efficiency and thereby contributes to the false discovery of CNVs.

Technically, there are several possibilities that can explain why reduced protein removal can affect the evenness in coverage of DNA sequencing reads. For example, phase-separation techniques (e.g. phenol-chloroform extraction) can influence the recovery of DNA because they require DNA to move to the aqueous phase while proteins end up in a separate fraction. Tightly associated DNA-protein complexes could be depleted from the aqueous phase and therefore no longer be available for subsequent DNA extraction. An alternative option is that DNA-protein complexes are still present in the isolated DNA sample, but are simply not ‘available’ for subsequent applications such as library preparation for sequencing. Interestingly, we find that if we re-treat poorly isolated DNA with proteinase K, we achieve

much higher evenness of coverage. This indicates that inaccessible DNA can be made accessible by additional (more thorough) removal of protein. In unpublished results, we also observe that the measured DNA yield varies between poorly isolated and re-isolated DNA (after additional proteinase K treatment). By splitting the initially isolated sample exactly in two halves, we find that the sample that was treated with proteinase twice showed a two-fold increase in DNA, measured by multiple independent techniques (Nanodrop and Qubit). Hypothetically, the DNA content of a sample containing poorly isolated DNA could thus be rescued or at least improved by additional treatment with proteinase K.

Our findings show that despite the impressive technological progress in DNA sequencing over recent years, biases introduced during the initial (basic) steps of each DNA-related experiment should not be overlooked. With increased resolution and sensitivity of DNA sequencing, each bias that is introduced in early steps of a protocol is amplified. Especially for clinical and diagnostic purposes, this could have major implications. Patient material is scarce and DNA is mostly obtained from single tissues (e.g. blood). When DNA obtained from different sources (biopsies, buccal swabs, etc) needs to be compared, for example in a diseased versus healthy tissue comparison, isolation effects need to be taken into account. Especially if the genomics analysis focuses on heterogeneous tissue samples such as cancer biopsies, false-positive CNV calls can arise and hamper the identification of true driver variants (e.g. if gains or losses of oncogenes/tumor suppressors). To date, several studies have described differences in DNA content between tissues such as in **Chapter 2** as somatic mosaicism and proclaimed the extensiveness of this variation [1-5]. However, in line with our work others have shown an effect of local protein-DNA interactions on the efficiency of DNA isolation, such as in the architecture of promoters [6, 7]. Somatic copy number changes were also described to indeed exist in healthy cells, albeit on a much smaller scale [8, 9].

Over recent years, most clinical and research environments have developed, optimized and standardized in-house DNA isolation and sequencing protocols. Variation in the efficiency of these protocols becomes evident in consortium efforts, where sequencing data from multiple centers is combined and compared. For example, when DNA isolations are carried out by the individual centers, but library preparation and sequencing are done at one sequencing company, variation in the evenness of coverage that can be linked to different DNA-providing centers becomes apparent (unpublished data, Figure 1). These results do not imply that data with a less than optimal genome-wide coverage cannot be used, but they do nicely illustrate how a procedure as straightforward as the isolation of DNA can have serious impact on DNA sequencing data. If for example the copy number states in DNA samples from multiple centers need to be compared, this bias should be taken into account and corrected for.

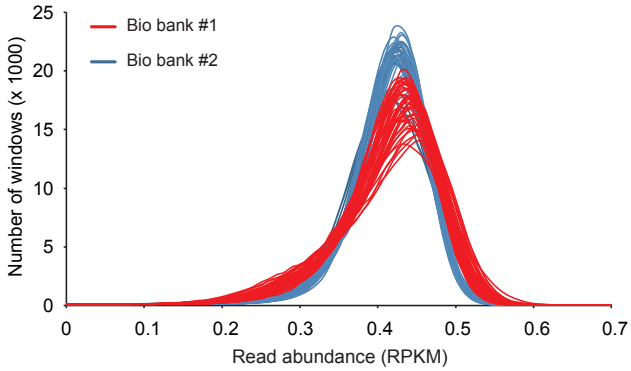


Figure 1 - **Evenness of genome-wide coverage differs between bio banks that use different isolation methods**. The x-axis shows the normalized numbers of reads (RPKM) calculated per 10 kb of sequenceable genome (meaning that parts that cannot be covered are filtered out), the y-axis shows the genome-wide number of 10-kb bins with x RPKM. The sharper the distribution, the more even the genome-wide coverage is. DNA samples of each bio bank were sequenced at the same sequencing center, leaving the isolation of DNA as the only variable responsible for the observed differences.

Genomic variation and predicting phenotypic consequences

Chapter 2 provides insight in the interpretation of genomics data and more specifically the detection of genomic variation. However, once such variants have been reliably identified, determining their phenotypic (molecular) effects and potential link to disease forms a whole new challenge. These effects can be diverse, ranging from single nucleotide changes that affect the function of protein-coding genes to large structural variants that remodel megabases of DNA. In **Chapter 6**, we study non-recurrent structural genomic rearrangements in patients with complex congenital disease. For these patients, it is challenging to provide a correct diagnosis, because the effects of non-recurrent genomic variants on genome function and development are unclear and differ per patient. We show that, by applying multiple molecular (sequencing) analyses on blood of these patients, we can study the consequences of these variants. We find that transcription is deregulated across many of the genomic breakpoint junctions, resulting in for example fusion genes and deregulated expression of genes. Genomic rearrangements thus not only disrupt functional alleles, but also drive the ectopic activation of the fusion gene partner, which is often undesired and damaging. Interestingly, we find that certain germline rearrangements that drive congenital disease show remarkable resemblance to rearrangements that drive cancer. The molecular consequences of these rearrangements reflect the cancer situation as well. For example, we detect gene fusions in congenital disease patients that involve genes that are also recurrently rearranged in multiple tumor types. Also, we find that a miRNA cluster with known oncogenic properties becomes ectopically expressed in a second patient, although this cluster is normally only active in human trophoblast cells or cancer. Functional follow up studies in zebrafish

revealed that overexpression of some of these oncogenic microRNAs can indeed result in brain morphogenesis defects in zebrafish. We find a possible explanation for the overlap in breakpoints in the fact that the recurrent breaks are specifically localized to late-replicating regions. Although late replication is a known characteristic of common fragile sites, we do not observe enrichment for known common fragile sites.

Our findings could imply that congenital disease patients that show genomic rearrangement overlap with cancer are more susceptible to develop (pediatric) cancer. However, the two patients with germline rearrangements that we analyzed at the molecular level did not develop cancer at age 25 (chromothripsis patient with ETV1 fusion) and age 8 (patient with tandem duplication that activate C19MC). This could depend on the gene promoters that activate these genes (***DPYD*** and ***NDUFA3***) and their tissue and developmental stage-specific gene expression profiles. Also, cancer development likely relies on other (additional) mutations, and the mutations found in these patients would be oncogenic in a different genetic context and/or may not be the primary oncogenic driver.

In **Chapter** 6, we show that integrating molecular profiling techniques (a so-called "multi-omics" approach), including RNA-seq, small RNA-seq and ChIP-seq, with genomic information obtained by DNA sequencing improves our understanding of the mechanisms by which rearrangements drive disease. Nevertheless, for most variants, especially the ones that do not directly affect genes, the effects remain mostly unclear. In **Chapter 7**, we apply a systematic integrative approach to evaluate the diversity of phenotypic consequences driven by naturally occurring genomic variants. We again use multiple NGS applications such as RNA-seq, ChIP-seq and Hi-C (chromatin structure and organization) to study the phenotypic consequences of genomic variants on these regulatory layers. We do not only focus on single nucleotide variation (SNVs), but also on structural genome variation (SV). The effects of structural variants, such as large deletions or duplications, are less well understood than those of SNVs. In this chapter we show that especially in noncoding genome regions, effects can be diverse and hard to predict. We find that genomic variants not only affect regulatory elements via direct overlap, but may also modify the higher order chromatin organization, thereby exhibiting a long-range effect on gene expression regulation. For example, we find a deletion over 150 kb away from the ***Pkhd1*** gene, that increases the expression of this gene by deleting the boundary of a chromatin domain, thereby extending its activity. This chapter illustrates that heritable genomic variation, which is present in every genome, affects genome function at multiple levels. Many of those levels are currently not being considered while studying the genetic basis of disease, where mostly only SNVs that perturb coding regions are subjected to follow-up studies.

Complex diseases are common in the human population, are polygenic and multifactorial. This

means that genetic predisposition for these diseases exists, but there is not one genetic factor that is fully responsible for disease development. In contrast to most ***de novo*** (e.g. germline) genomic variants such as discussed in **Chapter 6**, heritable variants are likely to have more subtle effects. Individually, each variant may not be capable of triggering disease, but when present in the right combinations they might be causal or predispose to common disease. Our work demonstrates that it will be important to take into account the combinatorial effects of noncoding SNVs and SVs along with protein-coding DNA polymorphisms for dissection of causal variation contributing to complex disease, e.g. as identified in large-cohort genome-wide association studies (GWAS).

Integrating genomics approaches to study complex disease

To better understand the genetic component in complex disease, genetically homozygous animal models for disease provide a good alternative to the outbred heterozygous human genome. To define the genetic basis of disease, not only the genomes but also the transcriptional and translational output of cells should be considered. Although techniques that can measure differences in RNA and protein have been around for a while, integration of these different layers of data is challenging. In **Chapter 4**, we describe a proof-of-concept multi-omics approach for the integration of genomics, transcriptomics and (mass spectrometry-based) proteomics data. We show that when executed correctly multidisciplinary genomics approaches provide better insight in gene expression regulation and disease mechanisms. For example, one of the disease models that we study, the SHR rat strain, is a frequently studied model for hypertension. We find several genes deregulated in SHR that were previously identified as important candidates for human hypertension. Among the deregulated genes is ***Cyp17a1***, a top hit from a previous genome-wide association study (GWAS) and a gene known to lead to congenital adrenal hyperplasia and early-onset hypertension when mutated [10, 11]. Using ChIP sequencing, we find that the promoter of ***Cyp17a1*** is much less active in SHR than in BN-***Lx*** due to a noncoding promoter mutation that disrupts a transcription factor binding site, likely being responsible for the low mRNA and protein levels that we observe in SHR. The functional involvement of Cyp17a1 needs further studying, preferably not only in liver but also in rat kidney (and development), because the kidney is the only tissue in human that expresses Cyp17a1. Using the recently developed targeted genome editing technologies (e.g. CRISPR/Cas9 or TALENs [12, 13]), rescue experiments could be carried out to restore the expression levels of Cyp17a1, which might reduce the chance for these rats to develop hypertension.

The results described in **Chapter 4** not only provide a proteogenomics proof-of-concept approach, revealing the effects of genetic variation at multiple levels, but also reveals that damaging mutations are not necessarily located within the coding regions of genes. In the

case of ***Cyp17a1***, we find a promoter variant that deregulates the expression of the gene and not a nonsynonymous mutation that damages the protein product. Especially in the context of common complex diseases, such as heart failure, diabetes or hypertension, combinatorial effects of widespread and frequently occurring genomic variants in noncoding regulatory elements could predispose to disease. An integrative approach as presented here, including molecular techniques like ChIP-seq, whole genome sequencing, RNA sequencing and mass-spectrometry, can help in pinpointing causal noncoding variants that are now frequently being overlooked.

Transcription in the noncoding genome

As becomes clear from the chapters discussed above, many genomic variants map to the noncoding genome, like the promoter variant in ***Cyp17a1*** and the deletion that deregulates the expression of ***Pkhd1***. However, the noncoding genome not only has the regulatory function discussed above, but also encodes genes that are transcribed but do not code for proteins. These noncoding RNAs are an abundant class of RNAs but for most their function remains unclear [14]. Because noncoding RNAs do not produce protein, their function is most likely linked to the secondary structure of the RNA molecule. For example, noncoding RNAs form the RNA component of large RNA-protein complexes (ribonucleoproteins) such as ribosomes and the signal recognition particle (SRP). Although the above-mentioned ribonucleoproteins function in the cytosol, for long noncoding RNAs (lncRNAs) in particular mostly nuclear roles have been described. Precise mechanisms of action are often uncertain, but many have been implicated to be vital for nuclear structure and function [14-16]. Recently, lncRNAs were surprisingly shown to associate with ribosomes using ribosome profiling [17], an unexpected finding since lncRNAs are not translated into peptides [18]. In **Chapter 5**, we further explore the biology behind lncRNA binding to ribosomes. We do that by applying RNA sequencing to subcellular fractions including polysomal fractionated RNA and isolated nuclei. The advantage of this approach is that we are able to sequence complete RNA transcripts and not only the fragments protected by ribosomes, which is the case for the earlier studies that showed ribosome association of lncRNAs [17-19]. With polysome fractionation we can thus physically distinguish transcripts that are bound by single ribosomes from transcripts that associate with 2, 3, 4 or even up to 7 ribosomes. We use this information to study the subcellular behavior of noncoding transcripts compared to protein-coding transcripts and show that the (size) distribution of lncRNAs across all measured cellular compartments is very diverse and highly similar in complexity to that of protein-coding transcripts. This suggests that lncRNAs actually behave very similar to mRNAs, which is unexpected when these lncRNAs are mainly involved in nuclear functions. Interestingly, we not only confirm ribosomal association and show extensive polysomal binding, we actually find preferential localization of the majority of highly expressed lncRNAs to ribosomes and the cytosol, and not to the nucleus as was previously estimated [20].

So what are lncRNAs doing at (poly-)ribosomes? Theoretically, ribosome association could be the result of random background binding that occurs by chance to each RNA molecule in the vicinity of a ribosome. Also, the possibility that lncRNAs bind ribosomes as part of an mRNA degradation pathway, triggered by the lack of a functional open reading frame (e.g. nonsense-mediated decay (NMD)), cannot be excluded. However, the enrichment and diversity of lncRNA binding to ribosomes that we describe in **Chapter 5** would be unexpected in the case of background binding or a decay mechanism. Such mechanisms would imply that the majority of ribosomes would be occupied by energy consuming mechanisms that are not functional or could have been dealt with otherwise (e.g. reducing transcription). A decay mechanism, such as NMD, is also not expected since we find multiple (up to 7) ribosomes associated, whereas NMD is believed to be triggered after one initial round of translation.
Two possible other explanations for the observed association are that (i) lncRNAs are translated into small (potentially noncoding) peptides, such as discussed previously [17, 18, 21] or (ii) lncRNAs have regulatory roles during translation that they exhibit at ribosomal sites. For some lncRNAs, functional roles at ribosomal sites have been proposed. For example, high levels of the snoRNA host gene transcript *GAS5* have been described to trigger degradation by ribosome binding after snoRNAs are processed out [22] and in mouse the noncoding antisense transcript of Uchl1 binds its sense counterpart to regulate the association with polysomes [23]. In our data, we could not find widespread evidence for such functions. For example, most sense-antisense pairs did not co-localize in the same ribosomal configurations and NMD is unlikely because most lncRNAs are polysomal (3-4 ribosomes). Because polysomal fractionation and ribosome profiling are gradient-based approaches, we cannot be absolutely certain that the lncRNA are situated in the ribosomal binding pocket. Therefore, it could be possible that lncRNAs reside in ribonucleoproteins that stably associate with polysomes during gradient centrifugation. This seems unlikely, because no such ribonucleoproteins are known and this association has to be extremely stable to achieve such high levels of transcripts. Also, ribosomal profiling has previously shown footprints of binding that are highly reminiscent of ribosomal sites (protecting 28/29 nucleotides). A ribosome-independent complex that binds lncRNAs and moves through the gradient in a similar fashion as mono or polyribosomal complexes also seems unlikely, because such large ribonucleoproteins have not been discovered and would be hard to miss. Also, for specific enrichment of lncRNAs in one of the fractions, this would require a wide variety of these complexes, in different sizes.

Another possibility is that lncRNAs bind ribosomes independent of ribonucleoproteins and function as scaffolds that keep ribosomes intact but poised until translation of mRNA molecules is needed again, for example in situations when the degradation and re-assembly ribosomes would consume too much energy or a quick switch in protein expression is required. A function as a ribosome scaffold seems to be a likely possibility and would also explain why we observe a similar length-dependency for the number of ribosomes that lncRNAs can bind.

Also, the structure of the lncRNA would be the most important property to function as a scaffold, which would explain the relatively low sequence conservation of most lncRNAs [14].

Hypothetically, individual lncRNAs could be responsible of regulating translation of single mRNAs or groups of lncRNAs, for example multiple mRNAs involved in a single biological process. The translational regulation of a process could be why many lncRNAs have been implicated in contributing to diseases such as cancer, though exact disease-linked mechanisms are largely unclear [14, 24-26]. Also, it would be interesting to target lncRNAs via knockouts, modify the subcellular localization of lncRNAs or target their secondary structure formation, to study effects on the translational landscape, for example via a proteomics approach.

The findings we describe in **Chapter 5** again highlight the complexity of the noncoding genome, and show that lncRNAs are likely to possess more widespread roles in biological processes than currently anticipated. The extra-nuclear localization suggests that lncRNAs primarily function outside the nucleus and the association with ribosomes suggests a role in the regulation of translation. Although our work does not elaborate on further cytosolic functions of lncRNAs, we do show that most lncRNAs have unique subcellular distributions, which makes their functions likely to be diverse.

Future applications of systematic "omics" approaches

The field of genomics has experienced a major boost over the last decade. New technologies have emerged, that have shed light on the complexity and diversity of our genomes and the somatic or germline genetic component of many diseases.
Despite these technological improvements and the knowledge gained, many challenges in the generation, interpretation and integration of genomics data have arisen and need to be addressed. So could the combinatorial use of omics technologies, such as described in this thesis, add to better patient diagnosis and improved personalized genomics in the near future? I would argue that assessing multiple levels of data will always provide more detail than a single layer of data and that eventually systematic omics analyses will be a quick and cost-efficient method for everyday patient diagnosis. More information results in less room for interpretation and more accurate diagnosis, something that is almost impossible based on a single layer of (genomic) information. Logically, multi-level omics integration needs further streamlining at both the experimental and the computational side. This not only requires that (clinical) labs should be capable of generating different types of data, but also that specific software should be made accessible for correct integration and interpretation of the data. Although implementing these techniques will take time, I believe that it will only be a matter of years before whole genome, proteome, transcriptome and epigenome analyses are carried out side-by-side for detailed patient diagnosis. In the near future, the knowledge gained from these integrated personalized approaches might allow more accurate

prediction of (late-onset) diseases based on genetic material. Until then, the genome and all its regulatory facets need to be explored in more detail and all the various mechanisms via which genomic variants can trigger disease should be carefully assessed.

References

1. O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP: Extensive genetic variation in somatic human tissues. Proc Natl Acad Sci U S A 2012, 109:18018-18023.

2. Piotrowski A, Bruder CE, Andersson R, Diaz de Stahl T, Menzel U, Sandgren J, Poplawski A, von Tell D, Crasto C, Bogdan A, et al: Somatic mosaicism for copy number variation in differentiated human tissues. Hum Mutat 2008, 29:1118-1124.

3. Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz de Stahl T, Menzel U, Sandgren J, von Tell D, Poplawski A, et al: Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 2008, 82:763-771.

4. Mkrtchyan H, Gross M, Hinreiner S, Polytiko A, Manvelyan M, Mrasek K, Kosyakova N, Ewers E, Nelle H, Liehr T, et al: The human genome puzzle - the role of copy number variation in somatic mosaicism. Curr Genomics 2010, 11:426-431.

5. O'Huallachain M, Weissman SM, Snyder MP: The variable somatic genome. Cell Cycle 2013, 12:5-6.

6. Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, Liu T, Zhang Y, Kim TK, He HH, Zieba J, et al: Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods 2012, 9:609-614.

7. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB: Characterizing and measuring bias in sequence data. Genome Biol 2013, 14:R51.

8. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, Hutchinson A, Deng X, Liu C, Horner MJ, et al: Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet 2012, 44:651-658.

9. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, et al: Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet 2012, 44:642-650.

10. Geller DH, Auchus RJ, Mendonca BB, Miller WL: The genetic and functional basis of isolated 17,20-lyase deficiency. Nat Genet 1997, 17:201-205.

11. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, Najjar SS, Zhao JH, Heath SC, Eyheramendy S, et al: Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet 2009, 41:666-676.

12. Miller JC, Tan SY, Qiao GJ, Barlow KA, Wang JB, Xia DF, Meng XD, Paschon DE, Leung E, Hinkley SJ, et al: A TALE nuclease architecture for efficient genome editing. Nat Biotechnol 2011, 29:143-U149.

13. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F: Genome engineering using the CRISPR-Cas9 system. Nature Protocols 2013, 8:2281-2308.

14. Ulitsky I, Bartel DP: lincRNAs: genomics, evolution, and mechanisms. Cell 2013, 154:26-46.

15. Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB: An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 2009, 33:717-726.

16. Sasaki YT, Ideue T, Sano M, Mituyama T, Hirose T: MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. Proc Natl Acad Sci U S A 2009, 106:2525-2530.

17. Ingolia NT, Lareau LF, Weissman JS: Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 2011, 147:789-802.

18. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES: Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell 2013, 154:240-251.

19. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS: Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 2009, 324:218-223.

20. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al: The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012, 22:1775-1789.

21. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A: Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 2013, 9:59-64.

22. Tani H, Torimura M, Akimitsu N: The RNA degradation pathway regulates the function of GAS5 a non-coding RNA in mammalian cells. PLoS One 2013, 8:e55684.

23. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, et al: Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. Nature 2012, 491:454-457.

24. Berteaux N, Lottin S, Monte D, Pinte S, Quatannens B, Coll J, Hondermarck H, Curgy JJ, Dugimont T, Adriaenssens E: H19 mRNA-like noncoding RNA promotes breast cancer cell proliferation through positive control by E2F1. J Biol Chem 2005, 280:29625-29636.

25. Tripathi V, Shen Z, Chakraborty A, Giri S, Freier SM, Wu X, Zhang Y, Gorospe M, Prasanth SG, Lal A, Prasanth KV: Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. PLoS Genet 2013, 9:e1003368.

26. Yoshimizu T, Miroglio A, Ripoche MA, Gabory A, Vernucci M, Riccio A, Colnot S, Godard C, Terris B, Jammes H, Dandolo L: The H19 locus acts in vivo as a tumor suppressor. Proc Natl Acad Sci U S A 2008, 105:12417-12422.

&

Addendum

Nederlandse samenvatting
Dankwoord
List of publications
Curriculum vitae

Nederlandse samenvatting

DNA... hoe zat dat ook alweer?

Ons lichaam bestaat uit miljarden cellen die vrijwel allemaal twee kopieën DNA in zich dragen. Dit DNA is bij elkaar ongeveer twee meter lang en bevat 6.4 miljard individuele bouwstenen (nucleotiden), waarvan 50% afkomstig is van je vader en 50% van je moeder. Tezamen noemen wij dit DNA het "genoom", wat alle informatie bevat die nodig is om een bevruchte eicel uit te laten groeien tot een volwassen mens.
Voor alle processen die plaatsvinden in ons lichaam tijdens deze ontwikkeling maar ook daarna zijn eiwitten nodig. De informatie om eiwit te kunnen maken zit opgeslagen in het DNA in de vorm van genen, waarvan het genoom er in totaal ongeveer 20.000 bevat. Genen zijn kleine stukjes van het DNA die kunnen worden afgelezen tot een boodschapper molecuul genaamd mRNA, wat op zijn beurt weer vertaald wordt tot eiwit. Eiwitten zorgen ervoor dat een grote variëteit aan processen kan plaatsvinden, zoals bijvoorbeeld het aflezen van DNA, het delen van cellen, het verteren van voedsel of de afbraak van schadelijke stoffen in de lever. Het is van groot belang dat eiwitten hun werk goed kunnen doen, anders kan dit leiden tot ziekten. Veranderingen (mutaties) in de DNA volgorde van een gen kunnen zorgen voor beschadigde eiwitten die hun werk niet meer kunnen doen, of niet meer op de juiste manier. Eiwitten zijn dus ontzettend belangrijk, echter, de totale hoeveelheid DNA die codeert voor eiwit beslaat slechts 2% van ons complete genoom (vier centimeter van de twee meter aan DNA die in iedere cel aanwezig is). Hoewel men inmiddels weet dat de overige 98% ook van belang is voor een goed functionerende cel, is nog veel onduidelijk over de gevolgen van DNA mutaties die niet in genen liggen maar in dit zogenaamde "niet-coderende genoom".

De laatste jaren zijn verschillende technieken om DNA en RNA af te lezen ("sequencen") sterk verbeterd. Daardoor is het nu mogelijk veranderingen in het DNA op te sporen én te onderzoeken wat de effecten van deze veranderingen zijn op RNA en de eiwitten die daarvan gemaakt worden. In dit proefschrift staan het niet-coderende genoom en verschillende sequencing technieken centraal. Ik beschrijf verbeteringen in deze technieken, maar ook manieren waarop meerdere typen informatie met elkaar verweven kunnen worden om beter zicht te krijgen op processen die leiden tot ziekten. In **hoofdstuk 1** introduceer ik de huidige stand van zaken in de genoom biologie. In deze introductie komen verschillende sequencing technieken aan de orde, evenals de reeds bekende functies van het niet-coderende genoom en andere onderwerpen besproken in hoofdstukken 2-7.

Het isoleren van DNA kan onverwachte consequenties hebben

Hoofdstuk 2 gaat in op de ongewenste effecten van niet-optimaal geïsoleerd DNA op DNA sequencing en de daaropvolgende data-interpretatie. In dit hoofdstuk laten we zien dat een

van de meest basale stappen die nodig is voor DNA sequencing, namelijk het isoleren van DNA, veel nadelige gevolgen kan hebben in de praktijk.
Omdat iedere celkern twee meter aan DNA bevat, wordt DNA erg compact opgevouwen aan de hand van eiwitcomplexen. Tijdens de isolatie van DNA uit cellen of weefsels wordt dit eiwit van het DNA gescheiden zodat je idealiter zuiver DNA overhoudt. Gebeurt dit niet efficiënt genoeg, met als gevolg dat bepaalde stukken DNA gebonden blijven door eiwit, dan heeft dat als gevolg dat de gedeeltes die gebonden blijven door eiwit niet goed afgelezen kunnen worden door een DNA sequencer. Zo ontstaat variatie tussen gebieden in je DNA, waarbij bepaalde stukken veel minder goed afgelezen worden dan anderen. Kwantitatieve DNA informatie (hoe vaak kan ik een bepaald stuk DNA aflezen?) wordt gebruikt om gedupliceerde of verloren gegane stukken DNA op te sporen. Het opsporen van zogenaamde DNA deleties of duplicaties is een routine in de diagnostiek, omdat bekend is dat sommigen van dit soort deleties en duplicaties kunnen bijdragen aan de ontwikkeling van ziekten. De kwantitatieve interpretatie van DNA sequencing data helpt dus bij het stellen van de juist patiënt diagnose. We beschrijven dat suboptimale DNA isolatie als gevolg kan hebben dat een geneticus denkt dat bepaalde delen DNA gedupliceerd zijn of verloren zijn gegaan, terwijl dit niet het geval is. We laten vervolgens zien dat verschillende DNA isolatie condities de kwantitatieve interpretatie van het DNA kunnen verslechteren of verbeteren. Hoe beter de isolatie, hoe minder valse ontdekkingen van deleties of duplicaties.

De grote DNA puzzel: hoe lossen we hem op?

In **hoofdstuk 3** beschrijven we een verbeterde methode voor het bepalen van langeafstand informatie in het DNA, het zogenaamde “mate-pair sequencing”. Standaard DNA sequencing technieken bepalen de volgorde van een klein stukje DNA dat maximaal een paar honderd bouwstenen groot is. Ons DNA bevat echter veel stukken die we heel moeilijk kunnen aflezen omdat ze erg veel op elkaar lijken. Los van elkaar hebben we eigenlijk te maken met een ontzettend ingewikkelde puzzel van duizenden stukjes DNA waarvan we niet weten hoe ze met elkaar verbonden zijn. Deze stukken moeten in de juiste volgorde komen te liggen, en om deze volgorde te bepalen is het erg belangrijk dat we langeafstand informatie verzamelen. Deze informatie kan als het ware bruggen slaan tussen de opeenvolgende puzzelstukjes. Op die manier koppelen we de juiste stukjes DNA aan elkaar en weten we bijvoorbeeld precies welk stuk DNA in welk chromosoom thuis hoort.
Door middel van mate-pair sequencing kan een DNA sequencing apparaat informatie leveren voor dit soort bruggen. Dat doet het door beide uiteindes van een circulair DNA fragment lezen, terwijl deze normaal gezien een paar duizend DNA nucleotiden (bouwstenen) uit elkaar liggen. Helaas bevat ons genoom veel moeilijk af te lezen gebieden die zo lang zijn dat ze met de huidige mate-pair techniek niet aan elkaar gekoppeld kunnen worden.
De verbetering van de “mate-pair sequencing” techniek die we beschrijven in **hoofdstuk 3** vergroot de grootte van bruggen tussen puzzelstukjes van de huidige standaard (2,000 -

3,000 nucleotiden) tot een afstand van wel 25,000 nucleotiden. We laten zien dat we door het toevoegen van deze extra lange bruggen het genoom van de rat veel completer kunnen maken dan het was. Veel meer puzzelstukjes krijgen we op de juiste plaats, omdat de meeste moeilijk af te lezen gebieden kleiner zijn dan 25,000 nucleotiden (meestal ~6,000). Deze techniekverbetering heeft bijgedragen aan de opbouw van de meest recente versie van het referentie genoom van de rat (RNOR versie 5.0), welke gebruikt wordt als gouden standaard voor al het DNA, RNA en eiwit onderzoek dat op dit moment gedaan wordt in de rat.

Hoe verbinden we DNA, RNA en eiwit informatie met elkaar?

Er bestaan verschillende methodes om de status en activiteit te meten van een bepaalde cel of weefsel. Men kan bijvoorbeeld meten welke genen afgelezen worden en hoe frequent ze afgelezen worden via het sequencen van RNA. Door middel van eiwit sequencing kan bepaald worden welke RNA moleculen vertaald zijn tot eiwit en hoeveel eiwit in totaal aangemaakt is. Veranderingen in het DNA die mogelijk leiden tot ziekte kunnen effect hebben op het RNA en daarvan gemaakt eiwit. Om te bepalen hoe veranderingen in het DNA of RNA effect hebben op de aanwezigheid en kwantiteit van eiwit, is het belangrijk om van ieder niveau (DNA, RNA, eiwit) informatie te verzamelen en met elkaar te integreren. In **hoofdstuk 4** beschrijven we een nieuwe methode om deze verschillende datatypen te integreren. Op die manier vergelijken we twee soorten ratten, die model staan voor veel voorkomende menselijke aandoeningen zoals stofwisselingsziekten en een (te) hoge bloeddruk.

We lossen met deze aanpak twee problemen op. Ten eerste gebruiken we de DNA volgorde van iedere rat, inclusief alle variatie die daarin zit, om eiwitten beter te kunnen identificeren. Hiertoe gebruiken we ook informatie van variatie op het RNA niveau. Het identificeren van eiwitten gebeurde altijd met een standaard database, die geen gebruik maakt van dit soort sample-specifieke informatie uit zowel DNA als RNA. Ten tweede interpreteren we de RNA data ook op een kwantitatieve manier, zodat we een betere schatting kunnen maken van de hoeveelheid eiwit die van dit RNA gemaakt wordt. We gebruiken hiervoor data van een dusdanig hoge kwaliteit dat we heel precies de vergelijking tussen RNA en eiwit kunnen maken. Door beide ratten die we onderzocht hebben op deze manier te vergelijken, vinden we opzienbarende verschillen. Eén gen, genaamd ***Cyp17a1***, komt zowel op RNA als eiwit niveau veel minder vaak voor in de rat die last heeft van een te hoge bloeddruk. In studies in de mens werd dit gen al eerder in verband gebracht met het hebben van een te hoge bloeddruk, maar een directe oorzaak werd nooit aangetoond. Door de DNA variatie hier weer bij te betrekken konden we bepalen dat een specifieke verandering net buiten het gen, in een stukje DNA met een regulerende rol, verantwoordelijk is voor het verschil tussen beide ratten. Dit laat zien dat in mensen met een hoge bloeddruk, ook in dit soort regulerende stukjes DNA gezocht moet worden naar de genetische basis van een aandoening.

RNA moleculen die niet voor eiwit coderen zitten toch in eiwit fabriekjes

De hier geïdentificeerde verandering zat dus niet in het gen zelf, maar op een plek die gebonden wordt door eiwitten om zo het aflezen van ***Cyp17a1*** te kunnen reguleren. Het niet-coderende genoom zit vol met dit soort plekken, die tezamen zorgen voor dynamiek in genexpressie per cel en weefseltype. Naast deze regulerende stukjes DNA bevat het ook stukken DNA die wel degelijk afgelezen worden en RNA produceren, maar niet coderen voor eiwit. Dit zijn dus wel degelijk genen, maar ze verschillen van genen die coderen voor eiwit. De RNA moleculen die van deze genen gemaakt worden noemt men niet-coderende RNAs. Niet-coderende RNAs bestaan in alle soorten en maten, variërend van microRNAs (miRNAs; slechts 22 nucleotiden lang) tot lange niet-coderende RNAs (long noncoding RNAs; lncRNAs - tot wel duizenden nucleotiden lang). Van deze laatste soort is niet veel bekend, behalve dat ze ontzettend belangrijk voor het functioneren van een cel. Zo bestaan voorbeelden van lncRNAs die kanker of aangeboren afwijkingen kunnen veroorzaken. Ondanks het gebrek aan een eiwit-coderende functie, werd recent beschreven dat lncRNAs verrassend genoeg binden aan ribosomen, de eiwit fabriekjes van iedere cel. Dit is natuurlijk onverwacht, aangezien geen eiwit gemaakt kan worden van lncRNAs. In hoofdstuk 5 gaan we verder in op de biologie van lncRNA binding aan ribosomen. Zo was nog niet duidelijk hoe frequent lncRNAs aan ribosomen binden en hoeveel ribosomen überhaupt per lncRNA kunnen binden. Een normaal verschijnsel voor eiwit-coderende RNAs is namelijk dat meerdere ribosomen tegelijkertijd aan een RNA molecuul kunnen gaan zitten om zo nog efficiënter eiwitten te kunnen produceren. Wij beschrijven in dit hoofdstuk dat de manier waarop lncRNAs aan ribosomen binden erg veel lijkt op hoe eiwit-coderende RNAs dit doen. Ook vinden we dat verschillende aantallen ribosomen tegelijkertijd gebonden kunnen zijn en dat iedere lncRNA een specifieke voorkeur heeft voor een bepaald aantal ribosomen. In tegenstelling tot eerdere bevindingen, observeerden we ook dat de meerderheid van de lncRNAs liever aan ribosomen bindt dan vrij in de celkern te blijven, wat de plek is waar ze geproduceerd worden. Onze bevindingen laten zien dat lncRNAs waarschijnlijk veel meer functies hebben dan tot op heden werd gedacht, waaronder wellicht een regulerende rol die zij uitvoeren door te binden aan ribosomen.

Verbeterde patiënt diagnose door gecombineerd gebruik van meerdere sequencing technieken

In **hoofdstuk 6** brengen we de allernieuwste sequencing technieken wat dichter naar de patiënt. Hierbij focussen we specifiek op patiënten met zeer complexe aangeboren afwijkingen, vaak in combinatie met een ernstige intellectuele achterstand. Deze patiënten zijn niet te plaatsen binnen een bepaald syndroom omdat de genetische veranderingen die aan de basis liggen van de afwijkingen uniek zijn en niet vaker voorkomen. Voor artsen is het erg moeilijk diagnoses te stellen bij dit soort patiënten. Een ontbrekende diagnose heeft

weer tot gevolg dat geen duidelijkheid gecreëerd kan worden richting de (gezonde) ouders van de patiënt, wat er in resulteert dat zij nooit precies zullen weten wat hun kind mankeert. In dit hoofdstuk laten we zien dat we door een combinatie van (moleculaire) analyses die we toepassen op niet alleen de patiënt, maar ook beide gezonde ouders, we veel beter kunnen voorspellen welke DNA veranderingen hebben bijgedragen aan de ziekte van de patiënt. Veel van deze effecten zijn niet of nauwelijks te voorspellen op basis van DNA informatie alleen maar worden pas echt zichtbaar als we naar RNA en eiwit kijken.
Zo vonden we dat veel structurele DNA veranderingen in patiënten met aangeboren afwijkingen overeenkomen met veranderingen die we vinden in kankercellen. Ook de effecten van de veranderingen, zoals het ontstaan van nieuwe RNA moleculen of het hoger aanzetten van bepaalde genen, kwamen overeen met wat we vaker zien in kankercellen. Vervolgens laten we zien dat bepaalde posities in ons DNA inderdaad vaker betrokken zijn bij veranderingen in zowel kankercellen als in patiënten met aangeboren afwijkingen. Deze resultaten laten zien dat de situatie en omgeving waarin DNA veranderingen plaatsvinden bepalend zijn voor het type ziekte dat veroorzaakt wordt. In samenloop met bijvoorbeeld andere kankerverwekkende mutaties, zorgen deze veranderingen misschien wel voor kanker, terwijl dat in onze bestudeerde patiënten gelukkig (nog) niet het geval is.

Niet-coderende DNA mutaties beïnvloeden de functie van het genoom op meerdere manieren

Het laatst hoofdstuk in dit proefschrift, **hoofdstuk 7**, beschrijft onderzoek naar DNA verschillen in het niet-coderende deel van ons DNA. Ook kijken we naar de uitwerking van deze verschillen op de organisatie van DNA in de celkern. Zoals ik hierboven al beschreef, wordt DNA erg compact gemaakt om in een celkern te passen. Echter, niet ieder deel is even compact en sommige toegankelijke delen DNA komen met elkaar in aanraking om zo de efficiëntie waarmee genen worden afgelezen te reguleren. Deze regio's, die we regulerende elementen noemen, binden allerlei eiwitten die betrokken zijn bij het aflezen van genen en de productie van RNA. Het DNA wordt gevormd in een specifieke "3D structuur" om interacties tussen regulerende stukjes DNA toe te staan of juist te voorkomen. In deze studie gebruiken we tien verschillende ratten stammen die allemaal kleine genetische verschillen hebben ten opzichte van elkaar. We gebruiken deze DNA verschillen om te kijken hoe ze van invloed zijn op regulerende processen en het aflezen van genen. Door met zeer nieuwe technieken te kijken naar zowel de positie van regulerende elementen als de 3D organisatie van het DNA, zien we precies in welke gebieden DNA verschillen invloed hebben op DNA organisatie of interacties van regulerende elementen.
In dit hoofdstuk geven we een goed overzicht van allerlei mogelijke verschillende effecten van DNA veranderingen, zowel direct (bijvoorbeeld het verlies van een DNA regio die een gen reguleert) als indirect (bijvoorbeeld verlies van een regio die belangrijk is voor DNA organisatie, en daardoor meerdere genen of regulerende elementen beïnvloedt). Het was

nog onduidelijk hoe de meeste DNA veranderingen in het niet-coderende genoom van invloed kunnen zijn op het aflezen van genen. Wij laten hier zien dat er veel diversiteit bestaat in de verschillende manieren waarop veranderingen effect hebben, en dat vooral de indirecte effecten op de 3D organisatie van het DNA hierbij niet over het hoofd gezien moeten worden.

Hoofdstuk 8 vat vervolgens de bevindingen uit dit proefschrift samen in een algemene discussie en gaat verder in op de potentiele impact van de gepresenteerde resultaten.

Dankwoord

U heeft het gehaald - het dankwoord.

Na vier jaar promotieonderzoek is dan eindelijk mijn proefschrift af. Misschien is dit het eerste proefschrift dat u voor u hebt, misschien het tiende, misschien het honderdste. De afgelopen vier jaar heb ik verschillende proefschriften de revue zien passeren, de een nog indrukwekkender dan de ander. Regelmatig dacht ik: “Waar begin ik aan?! Hoe krijg ik dat boekje vol?!”

Vier jaar werken, boekje. Het lijkt bijna een vanzelfsprekendheid. Dat is het zeker niet!
Maar het is gelukt en het bewijs ligt voor u.

Ik ben niet alleen blij dat het er op zit, maar ook dankbaar voor iedereen die geholpen heeft bij de totstandkoming van dit proefschrift. Dankbaar voor de mensen die er voor gezorgd hebben dat de afgelopen vier jaar voorbij schoten; ik heb een geweldige tijd gehad op het Hubrecht. Iedere dag ging ik met plezier naar werk, of naar een congres ergens ter wereld, een borrel, feest, bbq, labstapdag, concert, (kerst)diner, nieuwjaarsuitje, Cuppen groep retraite, CGDB/CSnD retraite, PhD Masterclass, Olympos, floorball toernooi, Kafé Els, etc etc.

Dit alles was natuurlijk niet mogelijk geweest als ik niet aangenomen was door **Edwin**. Edwin, bedankt voor je begeleiding van de afgelopen jaren (en ook al gedurende mijn stage). Ik ben je dankbaar dat je me de kans hebt gegeven promotie onderzoek te doen in de Cuppen groep. Ik vond het ontzettend fijn om voor je te mogen werken, je bent altijd bereikbaar en reageert meteen wanneer je hulp nodig is. Input en correcties op manuscripten kreeg ik vrijwel altijd binnen een paar dagen (soms zelfs dezelfde dag nog!), ik heb inmiddels geleerd dat dat niet overal vanzelfsprekend is. Ook tijdens de wat lastigere fases van de projecten wist jij de paper de juiste richting te geven en er “doorheen te slepen”. Met twee onderzoeksgroepen op twee locaties, plus een hoop nevenfuncties en een gezin, heb ik veel bewondering voor de manier waarop je de Cuppen groepen en al je AIOs leiding geeft!

Verder gaat misschien wel mijn meeste waardering uit naar **Marieke**, **Victor** en **Wigard**. Victor, I know your Dutch is fine so here it goes: Met jullie, als senior postdocs/beginnend groepsleiders, heb ik op dagelijkse basis het meest samengewerkt. Op vijf van de zes wetenschappelijke hoofdstukken in mijn proefschrift staan jullie als senior author genoteerd en dat is niet voor niks. Jullie waren er voor de dagelijkse begeleiding en ik vond het erg prettig om veel met jullie te kunnen overleggen en samen te werken. Ik weet dat ik soms wat ongeduldig kan zijn en echt kan zeuren om dingen gedaan te krijgen (right, Victor?),

maar ik hoop dat jullie ook trots zijn op het werk dat we samen gepubliceerd hebben of snel gaan publiceren! Heel veel succes met jullie toekomstige carrières in de wetenschap en (een beetje) daarbuiten. Bedankt voor alles!

Dan de **Cuppen groep**: legendarisch binnen het Hubrecht en ver buiten de Utrechtse stadsgrenzen. Na ieder internationaal congresbezoek ook berucht "op locatie", dankzij de borrels, het nacht/ochtend zwemmen in hotels (of daarbuiten), het whisky drinken, maar vooral ook het maken van sfeer waar we/jullie ook komen. De beste borrels organiseren (met of zonder steeldrum en cocktailbar) en standaard als eerste aanwezig op borrels van andere groepen; klasse!

En aan wie heeft de Cuppen groep die reputatie te danken? Om te beginnen aan de vaste garde aan analisten/labmanagers/bioinformatici; **Pim**, **Ewart de 1e**, **Sander, Mark V**, **Nico**, **Esther**, recentelijk **Lisanne** en voorheen **Henk, Maarten, Wensi, Frans Paul en Wim.** Ontzettend bedankt voor de hulp in het lab, met de computer, maar ook daarbuiten. Jullie bijdrage aan mijn papers was cruciaal, ik hoop dat jullie dat beseffen! Dank! Ook de legendarische avonden bij de Rex (indien open) zal ik nooit vergeten. Vooral de dinsdag middag/avonden aan de bar met Ewart waren top! **Wensi**, a special thanks to you for directly working for me for about a year! I hope you enjoy your new job and still get as wasted as only you can get.

Ruben, Ewart de 2e en **Joep**, ook jullie bedankt voor de gezelligheid (ook in Duitsland.. eh. Nijmegen) en de interessante discussies en input voor mijn projecten. Alle drie getrouwd, kinderen en iedere dag een hele reis maken om op het Hubrecht te komen. Respect! Vooral met drie of vier van die bengels (Ruben, Ewart).... maar misschien (waarschijnlijk?) komt het voor Joep ook ooit zo ver. Bedankt ook voor jullie enthousiasme over niet-wetenschappelijke dingen, en dan met name over boeken, films en vooral muziek! Ewart, de concerten die we bezocht hebben waren super en de muzikale tips die je me gegeven hebt goud waard! Vooral Chk Chk Chk (!!!) staat iedere zonnige dag met vol volume op! En Hausmagger mogen we ook niet vergeten natuurlijk. Ruben, laat je kinderen nog lang dansen op Tool!

Paranimf en collega-AIO **Roel** Hermsen... Oh nee Schelland. We zijn ongeveer tegelijkertijd begonnen en hebben dus een vergelijkbare vier jaar achter de rug. Uitgezonderd van je huwelijk met Merel en de geboorte van Tom, en je verhuizing niet te vergeten, dat doe je er nog allemaal naast. Onze tripjes naar Tutzing, Stresa, Stockholm, New York en Malaga (die laatste komt nog op moment van schrijven) waren allemaal legendarisch, en dan met name 's nachts. We kenden elkaar niet voordat we bij de Cuppen groep gingen werken en zijn qua persoonlijkheid misschien verschillend, maar toch klopte het wel tussen ons. We hadden ieder onze eigen projecten, en gelukkig nog een mooie samenwerking aan het einde waarvan ik hoop dat hij goed gepubliceerd gaat worden! Ik ben je nog eeuwig dankbaar voor het feit

dat je onze appartement sleutel in New York op wonderbaarlijke wijze terug gevonden hebt... Hoewel, je was hem natuurlijk ook zelf kwijtgeraakt. Heel veel succes met de afronding van je PhD en je verdere carrière. Welke keuze je ook gaat maken, het zal vast de juiste blijken te zijn!

Huidige AIOs **Myrthe** en **Francis**, allebei bezig aan een goede start van je promotie traject. Roel en ik kunnen met een gerust gevoel het stokje aan jullie doorgeven. Het percentage vrouwen groeit zo langzaam tot ongekende hoogte (althans, voor de Cuppen groep), maar dat kan volgens mij geen kwaad. Nu wordt dat alpha-male gedrag van de rest een beetje gecompenseerd. Maak je niet druk over alles wat komen gaat, als Roel en ik het kunnen, geldt dat zeker voor jullie!

Mijn dank gaat natuurlijk ook naar de studenten die ik begeleid heb tijdens de afgelopen jaren **Dennis** en **Kim**. Dennis ik hoop dat je het naar je zin hebt op het NKI, bedankt voor je hulp tijdens je stage! **Kim**, je kwam zelfs twee keer stage lopen bij onze groep, dan moet het wel naar je zin geweest zijn! Je was een ontzettend goede student, slim en met veel potentie! Tegen de tijd dat je dit proefschrift onder ogen krijgt heb je een AIO plek gekozen. Ik vertrouw er in dat je de juiste keuze hebt gemaakt en hoop dat je een succesvolle AIO periode door gaat maken! Succes!

Cuppen groep studenten **Robin**, **Rutger** en **Silvia:** Succes met jullie verdere (wetenschappelijke carrière), ik denk dat jullie alle drie de potentie hebben om het ver te schoppen! Silvia, dan moet je wel even Nederlands onder de knie krijgen om dit te begrijpen, maar daar helpen Robin en Rutger je vast wel mee. Rutger, voor mij blijf je altijd Mental Rudy. Die taxi rit vanuit Nijmegen was legendarisch, die beelden krijg ik nooit meer uit mijn hoofd (ook al zou ik dat willen). Ook voormalig studenten **Maryvonne**, **Martijn**, **Jetse**, **Dennis**, **Peter**, **Yannick** en **Friso** bedankt!!

De voormalig AIOs **Ruben** (wederom), **Mul**, **Michal**, **Sam** en **Jos**. Ik heb ontzettend veel van jullie geleerd in mijn eerste jaar als AIO. Bedankt voor de hulp in het lab, de goede discussies in de AIO kamer en de adviezen die ik van jullie kreeg. Ik dacht dat vier jaar een eeuwigheid was en begreep jullie stress richting het einde van de vier jaar niet altijd. Nu wel! Jos, Yeti Sports Flamingo Drive doet het helaas niet meer op mijn Mac, heb jij nog een goede versie? Dan kan die mee naar Berlijn!

Ook de **UMC** Cuppen groep / afdeling humane genetica bedankt voor alles. Tijdens de retraites was het altijd feest en dankzij de combinatie UMC / Hubrecht kregen wij toegang tot veel interessant patiënt materiaal waar we veel fundamentele vraagstukken mee konden onderzoeken. De werkbesprekingen kregen door jullie kritische blik een heel andere inslag,

wat erg verfrissend werkt en veel nieuwe ideeën opgeleverd heeft. Ik kwam altijd graag op het UMC, maar volgens mij kwamen jullie ook allemaal wel graag een op zijn tijd een biertje drinken op het Hubrecht ;-). Dus **Mark**, **Marlous**, **Naja**, **Kirsten**, **Gijs**, **Ivo**, **Ies**, **Pjotr**, **Martin E**, **Martin P**, **Karen**, **Marco**, **Nicolle**, **Mirjam**, **Petra**, **Glen**, **Glenn**, **Magdalena**, **Terry** en **Monique** en wie ik ook vergeten ben ontzettend bedankt! Special thanks to Monique voor de organisatie van alle uitjes en de hulp bij administratieve dingen tijdens mijn promotie. Glen thanks for being a native English speaker and being willing to check my introduction for errors! Met name Mark en Ivo bedankt voor de hulp bij de totstandkoming van hoofdstuk 6!

Voor de gezelligheid binnen het Hubrecht is de Cuppen groep niet alleen verantwoordelijk. Ook de andere groepen zorgen hiervoor, dus dank aan de **Creyghton**, **Geijsen**, **Berezikov**, **Knipscheer**, **de Laat**, **Korswagen, Robin, van Rooij** en **van Oudenaarden** groep. En wat ooit de **Ketting** groep was natuurlijk niet te vergeten. Met name **Menno** bedankt voor de goede gesprekken toen je net op het Hubrecht kwam (en het mij belachelijk maken bij de introductie van mijn lunchmeetings:-)).
Maartje V, **Pieterjan**, **Peter**, **Charles**, **Rick**, **Dan**, **Axel**, **Bas**, **Elke**, **Lucas**, **Manda**, **Oliver**, **Maaike**, **Nico**, **Javier**, **Kay**, **Mauro**, **Leon**, **Lennart**, **Britta**, **Eirinn, Maartje L**, **Teije**, **Remco**, **Reinoud**, **Alex**, **Nune**, **Els** en wie ik nog vergeet, bedankt voor de top tijd! Dankzij jullie hebben sfeersponzen nooit de macht over kunnen nemen.

De meeste hoofdstukken in dit proefschrift zijn tot stand gekomen door middel van samenwerkingen, zowel binnen Utrecht als internationaal. Hiervan wil ik in ieder geval de groep van **Albert Heck** (**Teck, Henk, Bas, Shabaz**) bedanken voor de totstandkoming van hoofdstuk 4. Ik vond het prettig en vooral ook erg leerzaam om met jullie samen te mogen werken. Proteomics en Genomics liggen nog steeds een beetje uit elkaar, maar dankzij onze samenwerking is de integratie van beide onderzoekvelden er heel wat op vooruit gegaan. Ik ga ervanuit dat iemand dit stuk wel even vertaald voor Teck en Shabaz!
Dan de **MacInnes** groep, **Alyson** en **Paul** bedankt voor jullie hulp bij hoofdstuk 5. Paul, je hebt er even voor in de koude kamer moeten staan maar ik hoop dat je net als ons tevreden bent met de paper!

Also a big thanks to all the colleagues from the **EURATRANS** consortium for all the collaborations and inspiring annual meetings, mainly organized by **Erik**. The EURATRANS consortium resulted in collaborations that were crucial for chapters 2,4 and 7. Special thanks to **Michel Werner** and his group (mostly **Helen** and **Camille**) for having me over at the CEA in Paris, for showing me the beauty of Paris and helping me with the ChIP experiments. I had a great time!
Thanks as well to **Norbert Hübner** and his group for help with Chapter 4 and of course for offering me a postdoc position at the MDC in Berlin. I am really looking forward to continue

my scientific career at the MDC!

Ook zeker het vermelden waard zijn de vaste medewerkers van het Hubrecht, waaronder de facilitaire dienst, PZ, de IT afdeling, financiën, dames van de Albron en de receptionistes waaronder **Thea**! Altijd een glimlach 's ochtends en 's middags, top! Ook **Richard**, **Elroy**, **Jules**, **Romke**, **Peter**-**Erik**, **Jimmy** en **Arjan** bedankt voor alle hulp! Romke, ik zat net mijn e-mail nog door te nemen en zag dat vrijwel alle mailtjes van jou aan mij gericht gaan over niet opruimen na borrels of bierflesjes en kratten die rondslingeren door het gebouw of op de derde. Excuses daarvoor! Goed dat jullie alles zo netjes houden en ik begrijp de frustratie :-).

Voordat het erop gaat lijken dat ik buiten het Hubrecht geen leven had, hierbij een dankwoord gericht aan familie en vrienden buiten het werk.

Om te beginnen met mijn hockeyteam van **Goirle Heren 2** (en daarna **3**). Mannen, ik moest er jaren voor op en neer naar Goirle maar ik heb er geen moment spijt van gehad. Het reizen was niet altijd handig, maar de vrijdagavonden, zondagmiddagen, teamuitjes, teamweekenden, bbqs, feestjes en vakanties maakten het meer dan waard. Die momenten zorgden voor de ontspanning na een drukke werkweek. Scheidsrechters en tegenstanders moesten het soms ontgelden, ik hoop dat jullie na het lezen van dit proefschrift begrijpen waar die frustratie soms vandaan kwam ;-). Naast teamgenoten zijn jullie ook mijn vrienden, en het feit dat jullie me nog regelmatig bellen / sms'en - ondanks dat ik afgelopen jaar gestopt ben met hockeyen - benadrukt dat nog maar eens! Zodra ik in Berlijn zit gaan we daar snel een weekend de boel onveilig maken! Jullie zijn altijd welkom. Bedankt!

Zeker niet minder belangrijk mijn vrienden en bandgenoten van **Seedorf**. Het was even zoeken naar een goede bandnaam, maar dit is wat mij betreft een voltreffer Koen! **Bart**, **Thijs**, **Koen** en **Jan Willem** ontzettend bedankt voor de tijd die we met Seedorf gehad hebben (en nog hebben op moment van schrijven). Niet alleen de muziek was belangrijk voor me, maar ook de gesprekken en biertjes/whiskey tijdens de dinsdag avonden in Vuurland. De etentjes, concerten, optredens en andere uitjes vond ik stuk voor stuk super, net als het praten over en uitwisselen van muziek. Ik ben ook erg blij dat de combinatie van vrienden uit Goirle (Bart en ik) en Twente (de rest) zo goed samen ging. Ik weet dat ik het met mijn dronken kop op koningsdag al 10 keer gezegd heb, maar ik ga jullie echt ontzettend missen als ik in Berlijn zit. Jullie moeten snel een keer op bezoek komen en sowieso doorgaan met muziek maken!

Thijs, paranimf, jij hebt dit hele riedeltje al eens meegemaakt en toen had ik de eer om jouw paranimf te mogen zijn. Ik ben blij dat je ook nu achter mij staat; we gaan er een topdag van maken! Samen hebben we veel koffie, bier en whisky gedronken en veel besproken tijdens

onze studie en PhD. Die gesprekken waren (en zijn!) ontzettend waardevol voor me. Bedankt daarvoor!

Dan mijn **familie**, die altijd veel interesse hebben getoond en met veel bewondering geluisterd hebben naar mijn onbegrijpelijke uitleg als antwoord op de vraag "Wat doe je nou precies, Sebas?". In dit proefschrift staat een Nederlandse samenvatting, ik hoop van harte dat jullie nu eindelijk een beter beeld krijgen van wat me de afgelopen jaren bezig heeft gehouden! **Opa Kees**, **Oma Tiny** en **Opa Albert**, jullie interesse was ontzettend lief en jullie hebben me altijd op de voet gevolgd. Opa heeft zelfs via Google Translate al mijn papers vertaald en uitgeprint (met extra kopie voor mijzelf) om mijn werk beter te kunnen begrijpen. Daarnaast waren er altijd de telefoontjes wanneer er iets over genetica op TV te zien was ("Schat, die meneer Clevers is weer op TV! Nederland 3, zet maar snel op!). Ik hoop (en weet) dat jullie trots op me zijn, maar dat ben ik ook op jullie. Het is toch 50% van Heesch en 50% van der Sanden die ervoor gezorgd hebben dat ik zover gekomen ben.
Mijn ooms, tantes, neefjes en nichtjes, ook jullie bedankt voor de belangstelling en het altijd op de hoogte willen blijven van mijn promotie traject. **Hans** & **Seetje**, **Lauran** & **Marjolijn**, **Mark** & **Detje**, **Daan**, **Madelon**, **Tess**, **Bart**, **Bo**, **Anke**, **Laut**, **Peggy**, **Judith**, **Terni, Foppe, Wybe, Melle** allemaal bedankt!! Mark en Detje ook bedankt voor de tijd in Oman en New Orleans en de manier waarop jullie Jenn en mij ontvangen hebben. We hebben in Oman (in de woestijn nota bene) uitvoerig gesproken over het wel of niet starten van een promotietraject en mede dankzij die gesprekken heb ik ervoor gekozen toch deze weg in te slaan. Bedankt daarvoor! En Wybe, met jouw interesse in de biologie en genetica moet je echt de wetenschap in!

Mijn schoonfamilie **Wilma**, **Bert**, **Max**, **Marja**, **Charlaine**, **Reza** en **Melvin**, ook jullie bedankt voor de interesse in mijn werk de afgelopen jaren, en natuurlijk voor het feit dat jullie zo'n fantastische dochter/zus aan mij toevertrouwen. Berlijn is helaas niet naast de deur, maar ook niet zo ver als Amerika. Ik beloof jullie goed op haar te passen en over een paar jaartjes wonen we vast weer een stuk dichterbij!

Mijn lieve zusje **Benthe** en vrouw **Celine** (en **Baco**), jullie begrepen misschien ook niet altijd waar ik in godsnaam mee bezig was, en waarom het nu weer belangrijk kon zijn in wat voor tijdschrift iets gepubliceerd werd, maar toch altijd oprecht geïnteresseerd in jullie (schoon) broer! Ook bedankt voor alle ontspanning tijdens de etentjes, hockey, feestjes en avondjes drinken. Ik ben blij dat ons contact zo goed is Bent!

Lieve **Papa** en **Mama**, jullie hebben me altijd gezegd te doen wat ik het leukst vond zolang ik er maar uit haalde wat er in zat. Mijn enthousiasme voor de biologie en wetenschap hebben jullie aangewakkerd en van jongs af aan gestimuleerd. Dat ik deze weg ben ingeslagen heb ik grotendeels aan jullie te danken! Zowel tijdens mijn studie als mijn promotietraject volgden

jullie me op de voet en waren jullie trots wanneer het goed met me ging, geïnteresseerd als ik naar het buitenland mocht voor een congres en altijd betrokken. Jullie hebben me vaak geholpen, in de weekenden kon ik altijd bij jullie terecht en ik heb ook altijd alles met jullie kunnen delen. Ik weet dat jullie er altijd voor Jenn en mij zullen zijn, ook als we in Berlijn zitten. Dankjewel voor alles!

Last but definitely not least, mijn lieve vriendin **Jenn**. Toen ik begon met promoveren heb ik je vaak gewaarschuwd dat het nog wel eens druk kon gaan worden zo in het derde en vierde jaar. Dat zien we dan wel, zei je... Ik weet dat het niet altijd makkelijk voor je is geweest dat ik vaak weg was, vaak 's avonds zat te werken (ongezellig!) en continu met mijn e-mail en telefoon bezig was. Ik ben je ontzettend dankbaar dat je al die tijd geduldig gewacht hebt totdat de oude Sebas weer boven water kwam. Samen gaan we naar Duitsland en ik weet dat je dat daar naar uit kijkt maar dat je dat ook doet omdat je weet dat het voor mij belangrijk is. Ik denk dat we daar een top tijd gaan hebben en ik ben blij dat je met me mee gaat! Je bent er altijd voor me, en ik hoop dat je er altijd voor me zult blijven. Ik hou van je!

Sebas

List of publications

van Heesch S, van Iterson M, Jacobi J, Boymans S, Essers PB, de Bruijn E, Hao W, Macinnes AW, Cuppen E, Simonis M: Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. **Genome Biol** 2014, 15:R6.

Kemmeren P, Sameith K, van de Pasch LA, Benschop JJ, Lenstra TL, Margaritis T, O'Duibhir E, Apweiler E, van Wageningen S, Ko CW, van Heesch S, Kashani MM, Ampatziadis-Michailidis G, Brok MO, Brabers NA, Miles AJ, Bouwmeester D, van Hooff SR, van Bakel H, Sluiters E, Bakker LV, Snel B, Lijnzaad P, van Leenen D, Groot Koerkamp MJ, Holstege FC: Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. **Cell** 2014, 157:740-752.

van Nuland R, van Schaik FM, Simonis M, van Heesch S, Cuppen E, Boelens R, Timmers HM, van Ingen H: Nucleosomal DNA binding drives the recognition of H3K36-methylated nucleosomes by the PSIP1-PWWP domain. **Epigenetics Chromatin** 2013, 6:12.

van Heesch S, Mokry M, Boskova V, Junker W, Mehon R, Toonen P, de Bruijn E, Shull JD, Aitman TJ, Cuppen E, Guryev V: Systematic biases in DNA copy number originate from isolation procedures. **Genome Biol** 2013, 14:R33.

van Heesch S, Kloosterman WP, Lansu N, Ruzius FP, Levandowsky E, Lee CC, Zhou S, Goldstein S, Schwartz DC, Harkins TT, Guryev V, Cuppen E: Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. **BMC Genomics** 2013, 14:257.

Low TY*, van Heesch S*, van den Toorn H*, Giansanti P, Cristobal A, Toonen P, Schafer S, Hubner N, van Breukelen B, Mohammed S, Cuppen E, Heck AJ, Guryev V: Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. **Cell Rep** 2013, 5:1469-1478. *Contributed equally

Simonis M, Atanur SS, Linsen S, Guryev V, Ruzius FP, Game L, Lansu N, de Bruijn E, van Heesch S, Jones SJ, Pravenec M, Aitman TJ, Cuppen E: Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. **Genome Biol** 2012, 13:r31.

van Boxtel R, Kuiper RV, Toonen PW, van Heesch S, Hermsen R, de Bruin A, Cuppen E: Homozygous and heterozygous p53 knockout rats develop metastasizing sarcomas with high frequency. **Am J Pathol** 2011, 179:1616-1622.

Lenstra TL, Benschop JJ, Kim T, Schulze JM, Brabers NA, Margaritis T, van de Pasch LA, van Heesch SA, Brok MO, Groot Koerkamp MJ, Ko CW, van Leenen D, Sameith K, van Hooff SR, Lijnzaad P, Kemmeren P, Hentrich T, Kobor MS, Buratowski S, Holstege FC: The specificity and topology of chromatin interaction pathways in yeast. **Mol Cell** 2011, 42:536-549.

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, Hubner N, Cuppen E: Distribution and functional impact of DNA copy number variation in the rat. **Nat Genet** 2008, 40:538-545.

Curriculum vitae

Sebastiaan van Heesch was born on September 25, 1985 in Tilburg, The Netherlands. In 2003, he received his Athenaeum diploma from the Mill-Hillcollege in Goirle, after which he started to study Biomedical Sciences at Utrecht University that same year. In 2007 he received his Bachelor's degree and continued his education with the genetics-oriented Master's program 'Cancer Genomics and Developmental Biology', also at Utrecht University.

During his Master's, he successfully completed internships in the labs of Prof. dr. Edwin Cuppen at the Hubrecht Institute ("Somatic copy number variation in the rat") and Prof. dr. Frank Holstege at the UMC Utrecht ("Functional analysis of the interplay between transcriptional regulators using high-throughput gene expression profiling in yeast"). The work performed during these internships led to publications in the high-ranked journals Nature Genetics, Molecular Cell and Cell. After writing his Master's thesis under supervision of Dr. Suzanne Lens at the UMC Utrecht ("Aurora-B kinase as a possible sensor of tension in meiosis") he received his Master's degree (***cum laude***) in March 2010.

In April 2010 Sebastiaan started his PhD research in the Genome Biology group of Prof. dr. Edwin Cuppen at the Hubrecht Institute for Developmental Biology and Stem Cell Research (Utrecht). The results of this research are presented in this thesis. In September 2014, Sebastiaan will continue his scientific carreer as a postdoctoral researcher in the lab of Prof. dr. Norbert Hübner at the Max Delbrück Center (MDC) for Molecular Medicine in Berlin.